Belajar Statistika dengan R

(disertai beberapa contoh perhitungan manual)

Prana Ugiana Gio
Dasapta Erwin Irawan

USU press
2016

USU Press
Art Design, Publishing & Printing
Gedung F, Pusat Sistem Informasi (PSI) Kampus USU Jl. Universitas No. 9 Medan 20155, Indonesia

Telp. 061-8213737; Fax 061-8213737

usupress.usu.ac.id

ISBN 979 458 ...

Perpustakaan Nasional: Katalog Dalam Terbitan (KDT)

Belajar Statistika dengan R / Prana Ugiana Gio [dan] Dasapta Erwin Irawan -- Medan: USU Press 2016.

vi, 253 p. ; ilus.: 29 cm

Bibliografi
ISBN: 979-458-..

KATA PENGANTAR

Alhamdulillah, puji syukur atas kehadirat Allah SWT, karena atas izin-Nya, penulis dapat terus mempertahankan semangat untuk menulis, dan akhirnya dapat menyelesaikan buku ini. Hadirnya buku ini, tidak semata-mata atas usaha penulis sendiri, melainkan atas izin-Nya. Sungguh suatu kebahagiaan bagi penulis bisa berbagi sebagian kecil ilmu pengetahuan milik-Nya melalui buku yang berjudul "**Belajar Statistika dengan R**".

Ucapan terima kasih penulis sampaikan kepada semua pihak yang telah membantu dalam rangka penyelesaian buku ini. Penulis menyadari bahwa buku ini tentunya masih perlu perbaikan, sehingga penulis mengharapkan kritik dan saran yang membangun dari para pembaca agar buku ini dapat menjadi lebih baik. Kritik dan saran dapat ditujukan ke alamat email gioprana89@gmail.com atau website www.olahdatamedan.com.

Medan, 23 Februari 2016

Prana Ugiana Gio
Dasapta Erwin Irawan

DAFTAR ISI

BAB 1

PENDAHULUAN

Sekilas Sejarah R

"R" sebenarnya bukan bahasa pemrograman yang baru. Setidaknya R telah dikembangkan secara intensif sejak 10 tahun yang lalu, sebagai pengembangan bahasa pemrograman "S" di Bell Laboratories. Tepatnya R adalah bahasa pemrograman yang telah didisain ulang untuk memudahkan analisis statistika. Menurut situs **R project**, R adalah bahasa dan lingkungan untuk komputasi statistik dan grafis. R adalah proyek berjenis *open source* GNU. Entah apa yang dipikirkan oleh sang pembuat dengan memberi nama karyanya hanya dengan satu huruf. Tapi apalah arti sebuah nama.

Walaupun awalnya dikembangkan untuk analisis statistik, namun saat ini telah berkembang aplikasinya hingga dapat melakukan manipulasi data spasial serta menampilkannya secara dinamis dalam situs web. Ditambah lagi dengan *era data analysis* atau akrab disebut *big data*, maka perkembangan R menjadi tidak terbendung lagi.

Perintah dasar dalam bahasa R telah menyediakan berbagai *tool* untuk pemodelan statistik linear dan nonlinear, analisis *time-series*, klasifikasi, analisis klaster, dan analisis grafis. Kemampuan ini terus berkembang dengan adanya ribuan paket tambahan yang diunggah ke server CRAN tiap tahunnya.

R dan Markdown

Dari pemaparan ringkas di atas, sudah jelas apa itu R. Sekarang apakah "*Markdown*" itu? Nama ini diberikan oleh kreatornya, karena itu, John Gruber seorang programmer mengembangkan *markup language* "*Markdown*". Ia menyederhanakan berbagai perintah LaTeX agar dapat lebih mudah dipahami pemakai bagi pengguna yang bukan programmer dan bukan ahli matematika. Salah satu contohnya adalah *R markdown*. Bahasa *markup* (*markup language*) yang lebih mudah dari LaTeX atau html sekalipun. Dengan menggunakan *R markdown*, saat ini bisa digunakan untuk membuat *blogpost* atau naskah buku dengan R, seperti halnya naskah yang sedang anda baca saat ini.

Karakter R

Beberapa karakter R di antaranya:

- **R gratis, *Open Source,* dan *Cross Platform***. Karena gratis dan *open source*, maka kita dapat mengembangkan R sesuai kebutuhan kita, misalnya dengan membuat *add on package*. Karena bersifat *cross platform*, maka para pengguna yang menggunakan sistem operasi (OS) Linux, Mac dan Windows dapat saling bekerjasama. Peningkatan versi R akan selalu dilakukan bersamaan. Oleh karenanya menggunakan OS apapun, kita akan memiliki versi R yang setara.

- **R Mendukung Prinsip *Reproducibility***. R adalah aplikasi berbasis *command line*, artinya setiap perintah harus diketik sebagai baris perintah, yang dapat diulang oleh orang lain hanya dengan meng-*copy-paste* kode perintahnya. Prinsip ini disebut sebagai *reproducibility*. Bila anda melakukan hal ini dengan SPSS, Statistica, atau Minitab, yang berbasis *point and click* serta *drag and drop*, maka anda harus menangkap (*screen captur*e) untuk menggambarkan urutan langkah analisis yang anda lakukan.

- **R Menghasilkan Visualisasi yang Berkualitas Tinggi**. R memiliki kemampuan plot yang tinggi. Plot sangat diperlukan untuk memvisualisasikan hasil analisis anda. Bentuknya sudah bukan lagi hanya *scatter plot* dan histogram, tapi R sudah dapat membuat peta *chloropleth* dalam *format spasial*.

Komunitas pengguna R

R seperti halnya piranti lunak *open source* lainnya memiliki basis komunitas pengguna yang sangat banyak. Daftar berbagai komunitas R sebagian dapat dilihat di Situs *R-evolution*. Mereka berkumpul secara rutin dalam pertemuan pengguna R (*R meet up*) di berbagai negara. Kegiatan tersebut saat ini telah diadakan 127 kota di 31 negara, menurut situs ***R user group***.

Kebiasaan yang Dianjurkan

Sebagai pengguna R yang sampai saat ini masih belajar, maka kami menganjurkan tiga hal berikut ini:

- **Belajar dan Berbagi**. Pada hari anda memutuskan untuk menggunakan R, maka di hari itulah anda berkomitmen untuk berkontribusi kepada para pengguna lainnya. Caranya mudah sekali, bagilah pengetahuan baru yang anda pelajari, posting kode anda di blog atau media sosial anda. Bila anda memiliki akun Twitter gunakan **hashtag#rstats** pada tweet anda tentang R. Bila anda memiliki akun Google Plus, Statistics dan R adalah komunitas pengguna R yang dapat diikuti. Kebiasaan berbagi kode juga dapat dilakukan melalui akun **Github**. *Plaform* ini adalah semacam media sosial khusus untuk para programmer. Uniknya semua materi yang diunggah seluruhnya berlisensi bebas untuk dibagikan. Biasanya lisensi yang digunakan adalah *Creative Commons Attribution* (CC-BY) atau *Creative Commons Zero* (CC-0). Anda dapat mengkopi (*cloning*), membuat varian (*forking*) dari kode atau materi lainnya dari para pengguna dan memodifikasinya tanpa khawatir dituduh melakukan plagiarisme. Riwayat penyuntingannya pun dapat dilacak (*file versioning*) dan diketahui oleh penulis aslinya. Demikian pula pengguna lainnya dapat melakukan hal yang sama.

- **Belajar *Markdown Syntax***. Secara umum kode R dan umumnya Github akan bekerja baik bila anda menggunakan format teks dalam dokumen anda. Anda boleh tidak percaya, bahwa sekarang anda dapat menulis satu buku lengkap dengan *syntax Markdown*. *Syntax* ini adalah penyederhanaan dari *syntax* LaTeX tapi dengan format perintah yang lebih sederhana dan mudah diikuti. Oleh karenanya selain harus menginstalasi R dan R Studio IDE, anda harus menginstalasi distribusi LaTeX. Berikut tautannya untuk masing-masing OS dari **Situs LaTeX project**: **LaTeX for Linux**, **LaTeX for Mac**, **LaTeX for Windows**. Bila anda masih banyak berhubungan dengan file format doc atau docx, maka anda perlu menginstalasi Pandoc. Dengan

dapat melakukan konversi format dokumen apa saja, misalnya: Markdown (**.md**) atau **.html** ke format **.doc**/**.docx**, begitu pula sebaliknya. Dengan Pandoc, maka anda dapat menulis apa saja dalam format Markdown langsung dari jendela R atau R Studio anda. Menarik bukan.

⇨ **Pantau Package Terbaru**. Anda perlu memantau keberadaan *package* terbaru, karena sangat mungkin 10 baris perintah menggunakan fungsi dasar R dapat digantikan oleh satu baris perintah menggunakan *package* tersebut. Ingat bahwa R adalah *open source*, oleh karenanya **pasti ada setidaknya satu orang di belahan dunia yang lain** yang membuat *package* untuk **satu kebutuhan yang belum terpikirkan oleh orang lain**.

Referensi

1. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

2. Github site, url: www.github.com, diakses 14 Feb 2016

3. John Gruber Wikipedia site, url: https://en.wikipedia.org/wiki/John_Gruber, diakses 14 Feb 2016

4. Markdown syntax site, url: https://daringfireball.net/projects/markdown/syntax, diakses 14 Feb 2016

5. LaTeX project official site, url: http://latex-project.org/ftp.html, diakses 14 Feb 2016

6. R-evolution Analytics site, url: http://www.revolutionanalytics.com/, diakses 14 Feb 2016

7. R user group site, url: http://blog.revolutionanalytics.com/local-r-groups.html, diakses 14 Feb 2016

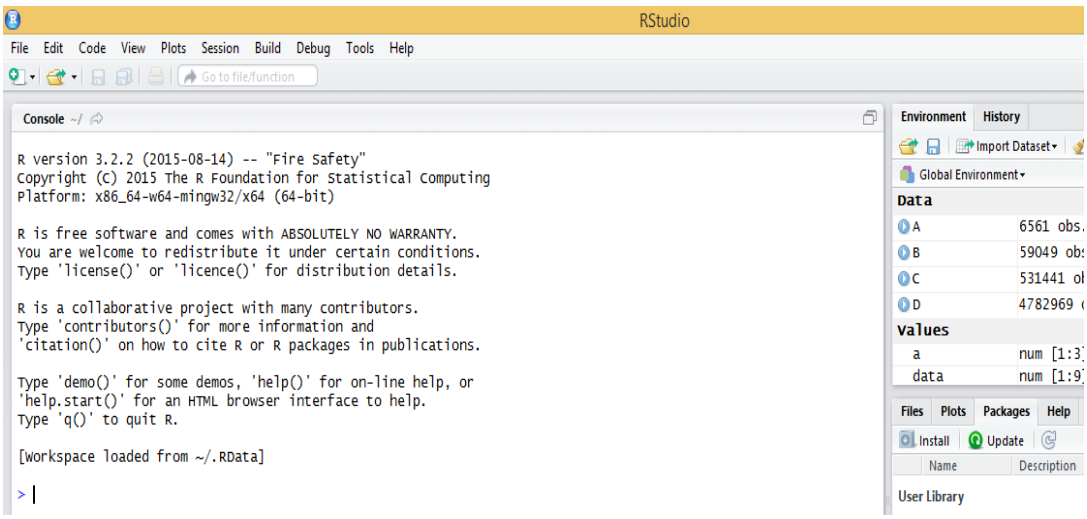
BAB 2

FUNGSI DASAR DALAM R

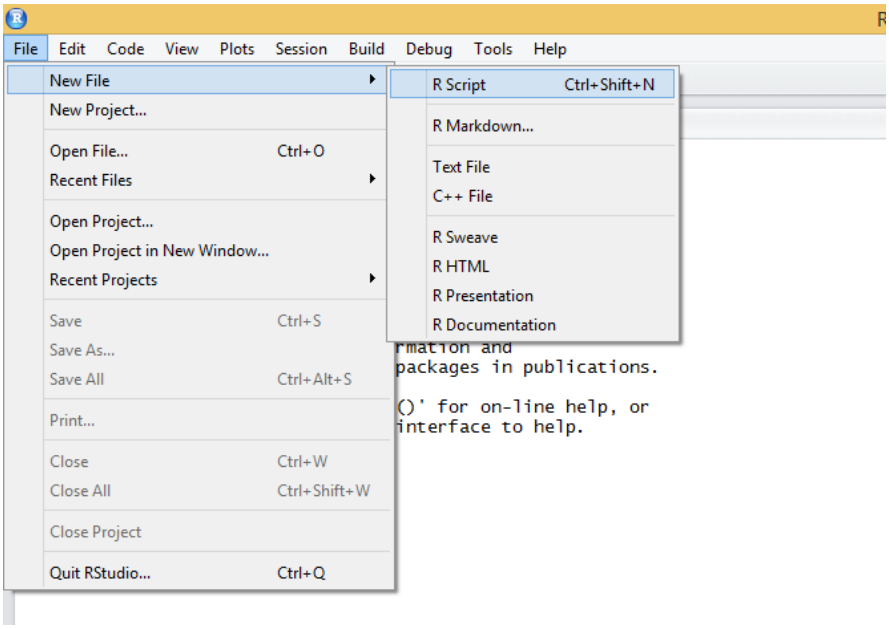
Memulai R

Berikut dipaparkan langkah-langkah untuk masuk ke dalam area kerja R. Aktifkan RStudio terlebih dahulu (Gambar 2.1), sehingga akan muncul tampilan seperti pada Gambar 2.2. Pada Gambar 2.2, pilih *File => New File => R Script* (lihat Gambar 2.3), sehingga muncul tampilan seperti pada Gambar 2.4. Gambar 2.4 merupakan area kerja R, di mana pada pembahasan selanjutnya, kode R akan diinput pada area tersebut. Setelah kode R diinput, selanjutnya kode R tersebut dieksekusi, sehingga muncul *output* berdasarkan eksekusi kode R tersebut.

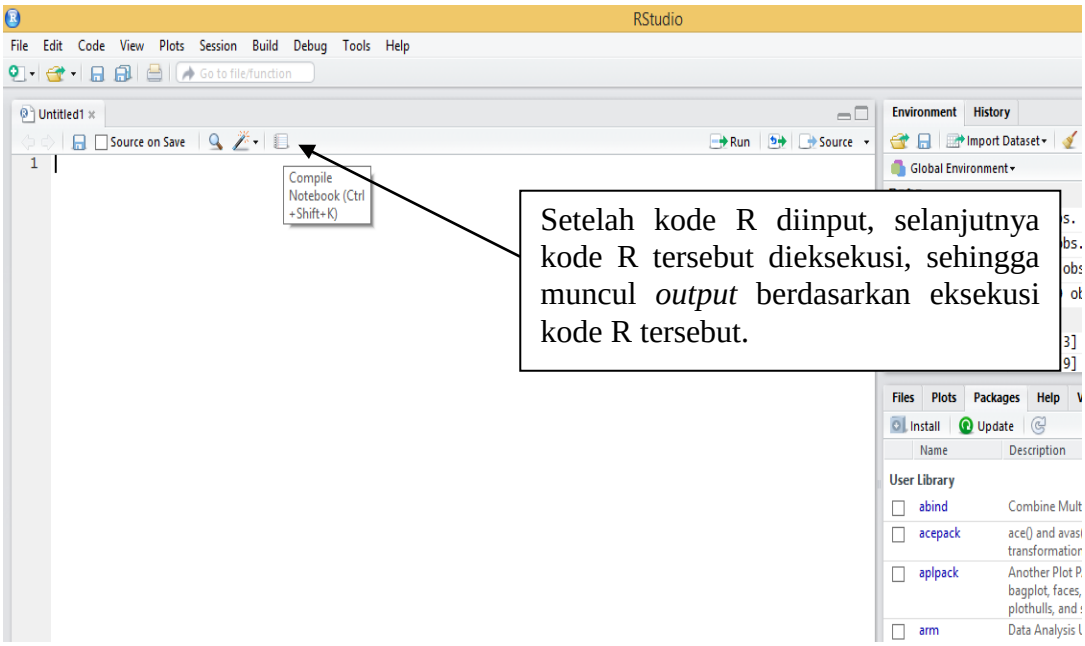
Gambar 2.1



Gambar 2.2



Gambar 2.3

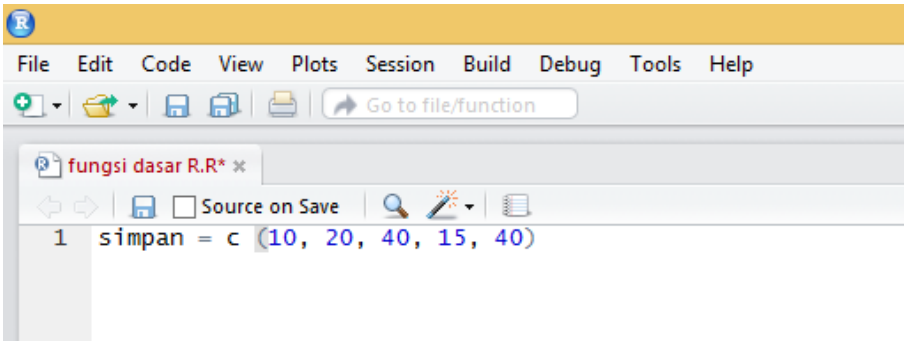


Gambar 2.4

Menyimpan Data dalam Variabel (Fungsi c)

Andaikan suatu data terdiri dari bilangan 10, 20, 40,15, 40. Misalkan data tersebut akan disimpan dalam variabel yang diberi nama **simpan**. Dalam R, fungsi **c** digunakan untuk menggabungkan satu nilai data, dengan nilai data lainnya. Perhatikan kode R berikut (lihat juga Gambar 2.5).

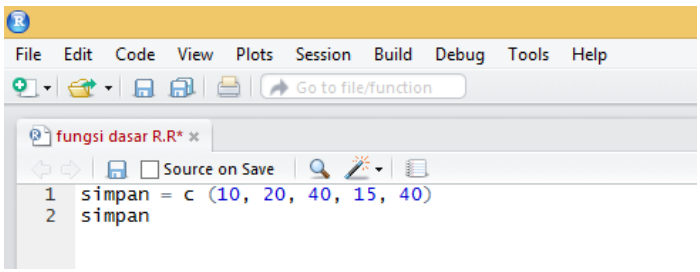
simpan = c (10, 20, 40, 15, 40)



Gambar 2.5

Kode R **simpan = c (10, 20, 40, 15, 40)** atau pada Gambar 2.5, dapat diartikan variabel **simpan** ditugaskan untuk menyimpan data dengan nilai 10, 20, 40, 15, 40. Data-data tersebut diapit oleh tanda buka-tutup kurung biasa, dan masing-masing nilai data dipisahkan oleh tanda koma. Perhatikan kode R berikut (lihat juga Gambar 2.6).

simpan

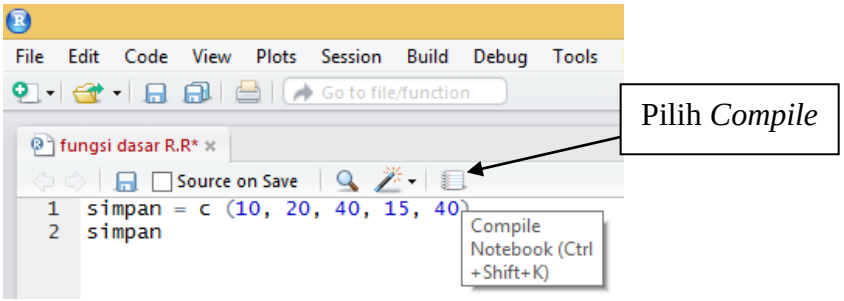


Gambar 2.6

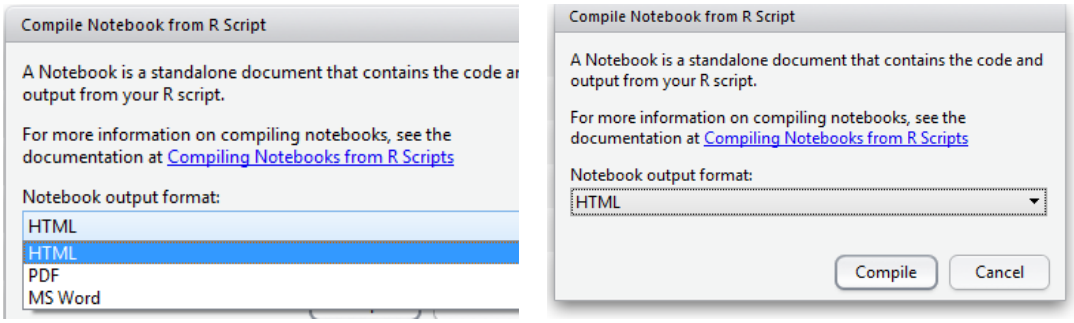
Kode R pada Gambar 2.6, baris ke-2, dapat diartikan menyajikan atau menampilkan nilai data yang disimpan dalam variabel **simpan**.

Mengeksekusi Kode R

Sekarang, kode R pada Gambar 2.6 akan dieksekusi. Pilih *Compile* (perhatikan Gambar 2.7), sehingga muncul tampilan seperti pada Gambar 2.8. Pada Gambar 2.8, *output* dari hasil eksekusi kode R pada Gambar 2.6, dapat berformat HTML, PDF, dan Ms Word. Dalam percobaan kali ini, pilih HTML dan *Compile*. Hasilnya diperlihatkan pada Gambar 2.9.



Gambar 2.7



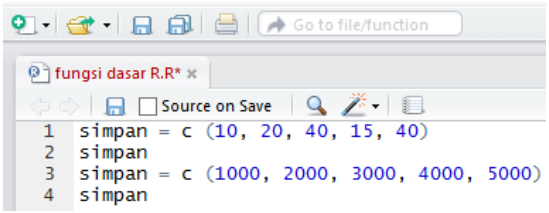
Gambar 2.8

```
simpan = c (10, 20, 40, 15, 40)
simpan
```

```
## [1] 10 20 40 15 40
```

Gambar 2.9

Pada Gambar 2.9, **## [1] 10 20 40 15 40** merupakan hasil eksekusi kode R pada baris ke-2. Perhatikan kode R pada Gambar 2.10, pada baris ke-3 dan ke-4.



Gambar 2.10

Gambar 2.11 merupakan hasil eksekusi kode R pada Gambar 2.10.

```
simpan = c (10, 20, 40, 15, 40)
simpan
```

```
## [1] 10 20 40 15 40
```

```
simpan = c (1000, 2000, 3000, 4000, 5000)
simpan
```

```
## [1] 1000 2000 3000 4000 5000
```

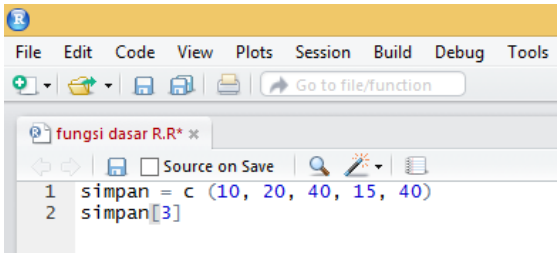
Gambar 2.11

Mengakses Nilai Data dalam Variabel secara Individual

Misalkan variabel **simpan** menyimpan nilai 10, 20, 40, 15, dan 40. Andaikan hanya ingin ditampilkan nilai dari variabel **simpan**, pada posisi ke-3, yakni nilai 40. Perhatikan kode R berikut.

simpan[3]

Kode R di atas berarti menampilkan nilai dalam variabel **simpan**, pada posisi ke-3, yakni 40. Ilustrasi dalam R diperlihatkan pada Gambar 2.12 dan Gambar 2.13.



Gambar 2.12

```
simpan = c (10, 20, 40, 15, 40)
simpan[3]
```

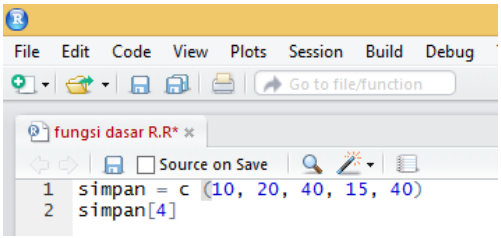
```
## [1] 40
```

Gambar 2.13

Perhatikan kode R berikut.

simpan[4]

Kode R di atas berarti menampilkan nilai dalam variabel **simpan** pada posisi ke-4, yakni 15. Ilustrasi dalam R diperlihatkan pada Gambar 2.14 dan Gambar 2.15.



Gambar 2.14

```
simpan = c (10, 20, 40, 15, 40)
simpan[4]
```

```
## [1] 15
```

Gambar 2.15

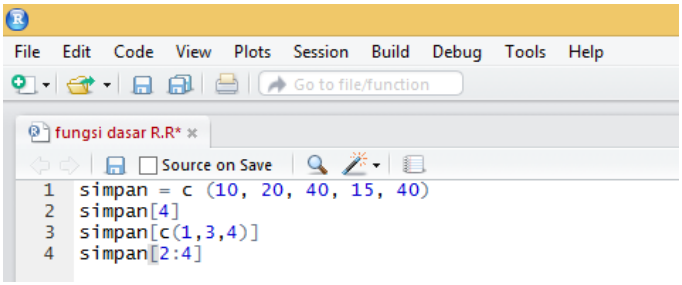
Berikut merupakan kode R untuk menampilkan 3 buah nilai dalam variabel **simpan** pada posisi ke 1,3, dan 4.

simpan[c(1,3,4)]

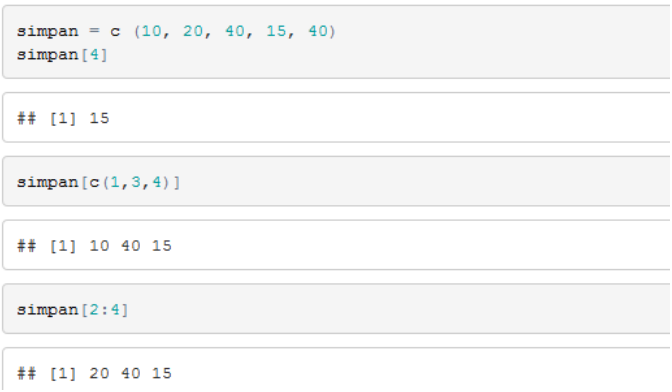
Kode **simpan[c(1,3,4)]** berarti menampilkan nilai dalam variabel **simpan** pada posisi ke 1, 3, dan 4, yakni 10, 40, dan 15. Berikut merupakan kode R untuk menampilkan 3 buah nilai dalam variabel **simpan** pada indeks ke 2,3,4.

simpan[2:4]

Kode **simpan[2:4]** berarti menampilkan nilai dalam variabel **simpan** pada posisi ke-2, sampai posisi ke-4, yakni 10, 20, dan 40. Ilustrasi dalam R diperlihatkan pada Gambar 2.16 dan Gambar 2.17.



Gambar 2.16



Gambar 2.17

Mengubah Nilai Data dalam Variabel

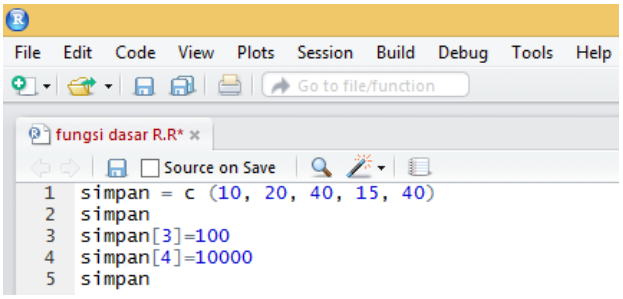
Pada pemaparan sebelumnya, diketahui variabel **simpan** menyimpan nilai 10, 20, 40, 15, dan 40. Andaikan nilai dari variabel **simpan**, pada posisi ke-3, yakni nilai 40, **akan diubah** menjadi 100. Perhatikan kode R berikut.

simpan[3]=100

Kode R tersebut, yakni **simpan[3]=100**, dapat diartikan mengubah nilai data variabel **simpan** pada posisi ke-3 dengan nilai 100. Misalkan nilai pada posisi ke-4, yakni 15, ingin diubah menjadi 10000. Berikut merupakan kode dalam R untuk mengubah nilai pada posisi ke-4, yakni 15 menjadi 10000.

simpan[4]=10000

Ilustrasi dalam R diperlihatkan pada Gambar 2.18 dan Gambar 2.19.



Gambar 2.18

```
simpan = c (10, 20, 40, 15, 40)
simpan
```

```
## [1] 10 20 40 15 40
```

```
simpan[3]=100
simpan[4]=10000
simpan
```

```
## [1]    10    20   100 10000    40
```

Gambar 2.19

Menghapus Nilai Data dalam Variabel

Misalkan suatu variabel bernama **NILAI** menyimpan 5 nilai, yakni 10, 40, 45, 30, dan 80. Berikut kode dalam R untuk menyimpan 5 nilai tersebut ke dalam variabel **NILAI**.

NILAI=c(10, 40, 45, 30, 80)

Misalkan nilai 45 pada variabel **NILAI** akan dihapus, sehingga nilai dalam variabel **NILAI** menjadi 10, 40, 30, dan 80. Diketahui nilai 45 berada pada posisi atau indeks ke-3. Berikut kode R untuk menghapus nilai 45 dalam variabel **NILAI**.

NILAI=NILAI[-3]

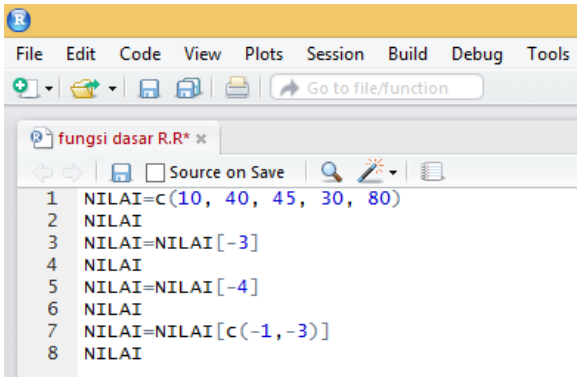
Setelah kode R **NILAI=NILAI[-3]** dieksekusi, maka nilai yang tersimpan pada variabel **NILAI** saat ini adalah 10, 20, 30, dan 80. Misalkan nilai 80 pada variabel **NILAI** akan dihapus, sehingga nilai dalam variabel **NILAI** menjadi 10, 20, dan 30. Perhatikan kode R berikut.

NILAI=NILAI[-4]

Diketahui nilai yang tersimpan pada variabel **NILAI** saat ini adalah 10, 20, dan 30. Misalkan nilai 10 dan 30 pada variabel **NILAI** akan dihapus, sehingga nilai dalam variabel **NILAI** adalah 20. Perhatikan kode R berikut.

NILAI=NILAI[c(-1,-3)]

Nilai dalam variabel **NILAI** saat ini adalah 20. Ilustrasi dalam R diperlihatkan pada Gambar 2.20 dan Gambar 2.21.



Gambar 2.20

```
NILAI=c(10, 40, 45, 30, 80)
NILAI
```

```
## [1] 10 40 45 30 80
```

```
NILAI=NILAI[-3]
NILAI
```

```
## [1] 10 40 30 80
```

```
NILAI=NILAI[-4]
NILAI
```

```
## [1] 10 40 30
```

```
NILAI=NILAI[c(-1,-3)]
NILAI
```

```
## [1] 40
```

Gambar 2.21

Penggunaan Operator > (Lebih Besar Dari)

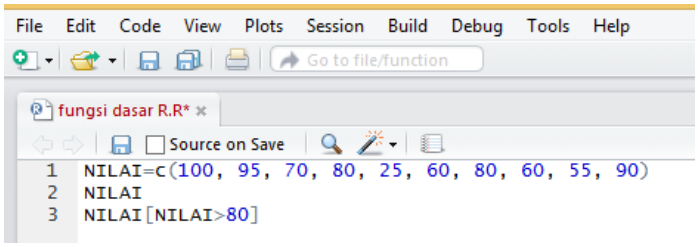
Misalkan suatu variabel bernama **NILAI** menyimpan 10 nilai, yakni 100, 95, 70, 80, 25, 60, 80, 60, 55, 90. Berikut kode R untuk menugaskan variabel **NILAI** menyimpan kesepuluh nilai tersebut.

NILAI=c(100, 95, 70, 80, 25, 60, 80, 60, 55, 90)

Misalkan akan ditampilkan nilai dari variabel **NILAI** dengan syarat lebih besar dari 80, yakni 100, 95, 90. Berikut kode dalam R.

NILAI[NILAI>80]

Ilustrasi dalam R diperlihatkan pada Gambar 2.22 dan Gambar 2.23.



Gambar 2.22

```
NILAI=c(100, 95, 70, 80, 25, 60, 80, 60, 55, 90)
NILAI
```

```
##  [1] 100  95  70  80  25  60  80  60  55  90
```

```
NILAI[NILAI>80]
```

```
## [1] 100  95  90
```

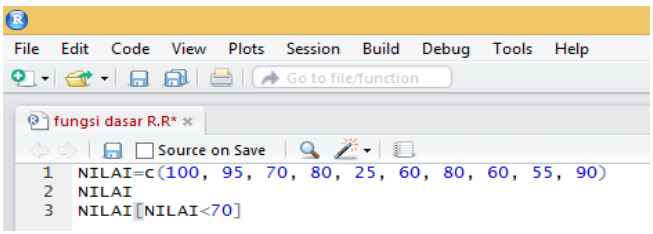
Gambar 2.23

Penggunaan Operator < (Lebih Kecil Dari)

Diketahui sebelumnya bahwa variabel bernama **NILAI** menyimpan 10 nilai, yakni 100, 95, 70, 80, 25, 60, 80, 60, 55, 90. Misalkan akan ditampilkan nilai dari variabel **NILAI** dengan syarat lebih kecil dari 70, yakni 25, 60, 60, 55. Berikut kode dalam R.

NILAI[NILAI<70]

Ilustrasi dalam R diperlihatkan pada Gambar 2.24 dan Gambar 2.25.



Gambar 2.24

```
NILAI=c(100, 95, 70, 80, 25, 60, 80, 60, 55, 90)
NILAI
```

```
##  [1] 100  95  70  80  25  60  80  60  55  90
```

```
NILAI[NILAI<70]
```

```
## [1] 25 60 60 55
```

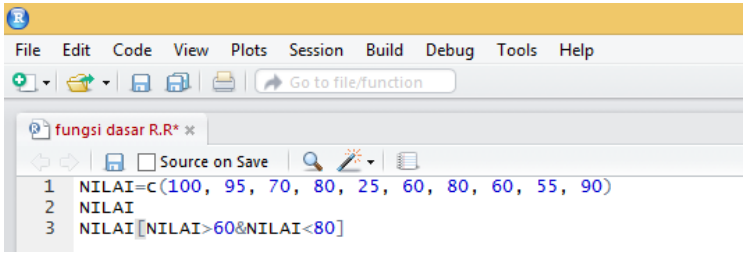
Gambar 2.25

Penggunaan Operator & (Dan)

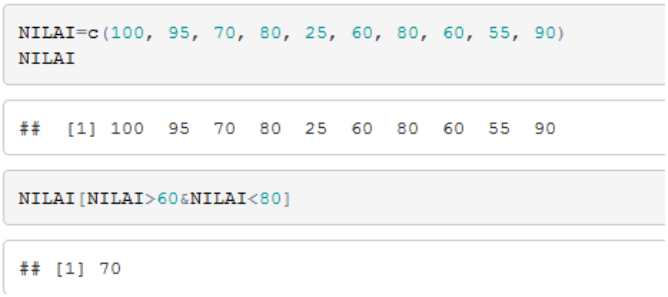
Diketahui sebelumnya bahwa variabel bernama **NILAI** menyimpan 10 buah bilangan, yakni 100, 95, 70, 80, 25, 60, 80, 60, 55, 90. Misalkan akan ditampilkan nilai dari variabel **NILAI** dengan syarat lebih besar 60 dan lebih kecil 80, yakni 70. Berikut disajikan kode R.

NILAI[NILAI>60&NILAI<80]

Ilustrasi dalam R diperlihatkan pada Gambar 2.26 dan Gambar 2.27.



Gambar 2.26



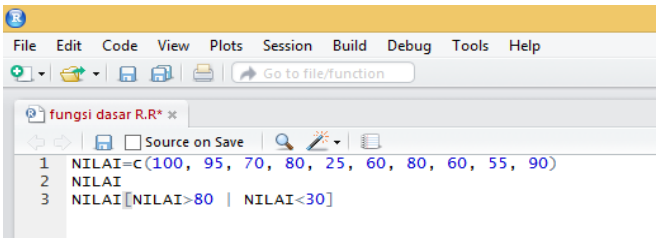
Gambar 2.27

Penggunaan Operator | (Atau)

Diketahui sebelumnya variabel bernama **NILAI** menyimpan 10 nilai, yakni 100, 95, 70, 80, 25, 60, 80, 60, 55, 90. Misalkan akan ditampilkan nilai dari variabel **NILAI** dengan syarat lebih besar 80 atau lebih kecil 30, yakni 100, 95, 25, 90. Berikut disajikan kode R.

NILAI[NILAI>80 | NILAI<30]

Ilustrasi dalam R diperlihatkan pada Gambar 2.28 dan Gambar 2.29.



Gambar 2.28

```
NILAI=c(100, 95, 70, 80, 25, 60, 80, 60, 55, 90)
NILAI
```

```
##  [1] 100  95  70  80  25  60  80  60  55  90
```

```
NILAI[NILAI>80 | NILAI<30]
```

```
## [1] 100  95  25  90
```

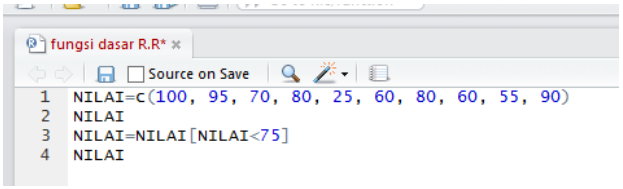
Gambar 2.29

Lebih Lanjut Penggunaan Operator < (Lebih Kecil Dari)

Diketahui variabel bernama **NILAI** menyimpan 10 nilai, yakni 100, 95, 70, 80, 25, 60, 80, 60, 55, 90. Misalkan nilai-nilai yang lebih besar atau sama dengan 75 akan dihapus dari variabel **NILAI**, sehingga nilai-nilai yang tersimpan dalam variabel **NILAI** adalah 70, 25, 60, 60, 55. Berikut disajikan kode R.

NILAI=NILAI[NILAI<75]

Ilustrasi dalam R diperlihatkan pada Gambar 2.30 dan Gambar 2.31.



Gambar 2.30

```
NILAI=c(100, 95, 70, 80, 25, 60, 80, 60, 55, 90)
NILAI
```

```
##  [1] 100  95  70  80  25  60  80  60  55  90
```

```
NILAI=NILAI[NILAI<75]
NILAI
```

```
## [1] 70 25 60 60 55
```

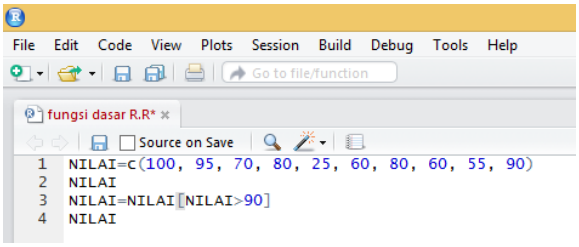
Gambar 2.31

Lebih Lanjut Penggunaan Operator > (Lebih Besar Dari)

Diketahui variabel bernama **NILAI** menyimpan 10 nilai, yakni 100, 95, 70, 80, 25, 60, 80, 60, 55, 90. Misalkan nilai-nilai yang lebih kecil atau sama dengan 90 akan dihapus dari variabel **NILAI**, sehingga nilai-nilai yang tersimpan dalam variabel **NILAI** adalah 100 dan 95. Berikut kode dalam R.

NILAI=NILAI[NILAI>90]

Ilustrasi dalam R diperlihatkan pada Gambar 2.32 dan Gambar 2.33.



Gambar 2.32

```
NILAI=c(100, 95, 70, 80, 25, 60, 80, 60, 55, 90)
NILAI
```

```
##  [1] 100  95  70  80  25  60  80  60  55  90
```

```
NILAI=NILAI[NILAI>90]
NILAI
```

```
## [1] 100  95
```

Gambar 2.33

Contoh Sederhana Penggunaan dari Bahasa Pemrograman R

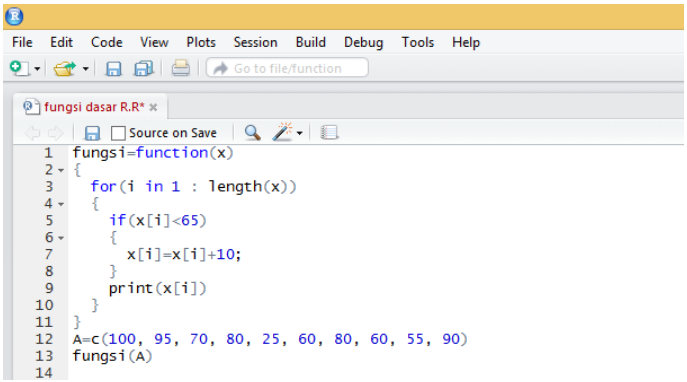
Misalkan suatu variabel bernama **NILAI** menyimpan 10 nilai, yakni 100, 95, 70, 80, 25, 60, 80, 60, 55, 90. Misalkan setiap nilai yang ada dalam variabel **NILAI**, yang **lebih kecil dari 65**, **ditambah dengan 10**. Nilai-nilai yang lebih kecil dari 65 adalah 25, 60, 60, 55. Nilai-nilai tersebut ditambah dengan 10.

25+10=35
60+10=70
60+10=70
55+10=65

Maka nilai-nilai yang akan ditampilkan adalah 100, 95, 70, 80, 35, 70, 80, 70, 65, 90. Berikut merupakan contoh kode program dalam R.

```
fungsi=function(x)
{
        for(i in 1 : length(x))
          {
                if(x[i]<65)
                  {
                        x[i]=x[i]+10;
                  }
                        print(x[i])
        }
}
A=c(100, 95, 70, 80, 25, 60, 80, 60, 55, 90)
fungsi(A)
```

Ilustrasi dalam R diperlihatkan pada Gambar 2.34 dan Gambar 2.35.



Gambar 2.34

```
fungsi=function(x)
{
  for(i in 1 : length(x))
  {
    if(x[i]<65)
    {
      x[i]=x[i]+10;
    }
    print(x[i])
  }
}
A=c(100, 95, 70, 80, 25, 60, 80, 60, 55, 90)
fungsi(A)
```

```
## [1] 100
## [1] 95
## [1] 70
## [1] 80
## [1] 35
## [1] 70
## [1] 80
## [1] 70
## [1] 65
## [1] 90
```

Gambar 2.35

Jenis Data R

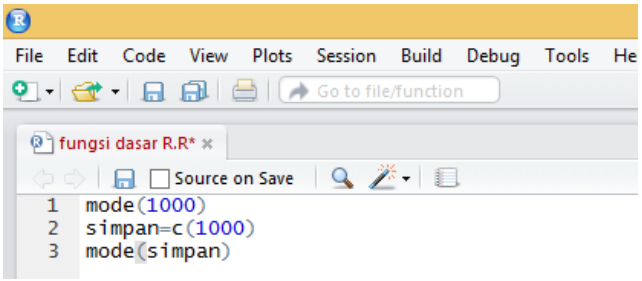
Berikut merupakan berbagai jenis dari jenis data dalam R.

- *Numeric* atau angka
- *Character* atau karakter
- *Logical* atau logika
- *Function* atau fungsi

Data yang berupa angka merupakan data *numeric*. Sebagai contoh 1, 100, 1500, 3000, dan seterusnya, merupakan data berjenis numerik. Untuk mengetahui jenis dari suatu data dalam R, digunakan perintah **mode**. Berikut merupakan kode R untuk mengetahui bahwa data 1000 termasuk ke dalam data berjenis *numeric*.

mode(1000)

Ilustrasi dalam R diperlihatkan pada Gambar 2.36 dan Gambar 2.37.



Gambar 2.36

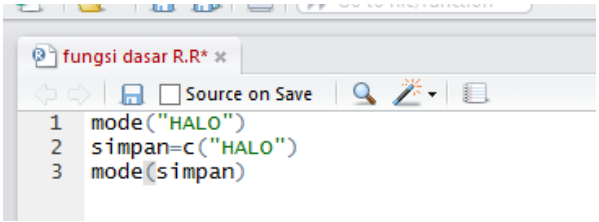
```
mode(1000)

## [1] "numeric"

simpan=c(1000)
mode(simpan)

## [1] "numeric"
```

Gambar 2.37

Data yang berupa karakter diapit dengan tanda petik ganda " ". Sebagai contoh "Halo", "A", "1", "500", "+", dan seterusnya merupakan data berjenis karakter. Berikut merupakan perintah dalam R untuk mengetahui bahwa data "HALO" termasuk ke dalam tipe data karakter.

mode("HALO")

Ilustrasi dalam R diperlihatkan pada Gambar 2.38 dan Gambar 2.39.



Gambar 2.38

```
mode("HALO")

##  [1] "character"

simpan=c("HALO")
mode(simpan)

##  [1] "character"
```

Gambar 2.39

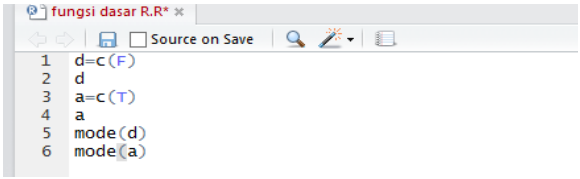
Nilai dari data yang berjenis logika terdiri dari 2 nilai, yakni T atau F. T merupakan singkatan dari *True*, yang berarti benar, sementara F merupakan singkatan dari *False*, yang berarti salah. Misalkan sebuah variabel bernama **d** menyimpan sebuah data berjenis logika, yakni F.

d=(F)

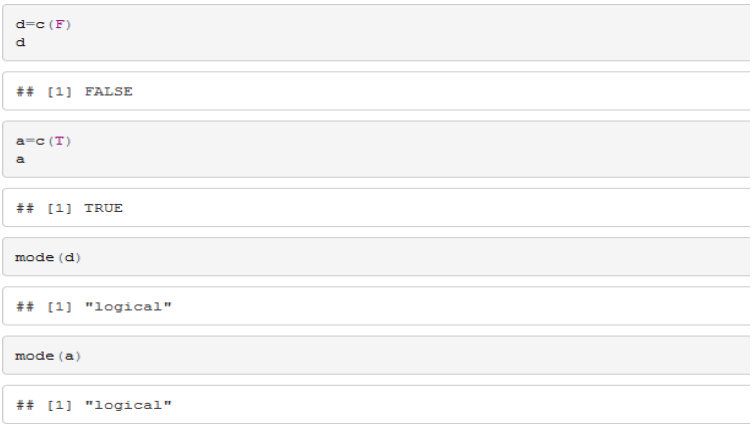
Berikut merupakan kode dalam R untuk mengetahui bahwa nilai yang tersimpan dalam variabel **d** berjenis logika.

mode(d)

Ilustrasi dalam R diperlihatkan pada Gambar 2.40 dan Gambar 2.41.



Gambar 2.40

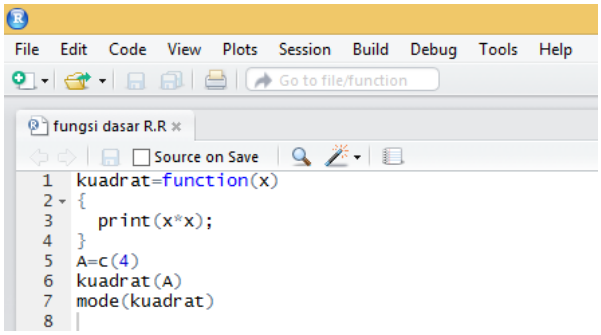


Gambar 2.41

Data yang berjenis fungsi mempunyai ciri menggunakan kata **function**. Berikut merupakan contoh kode program R pembuatan fungsi kuadrat.

```
kuadrat=function(x)
{
  print(x*x);
}
A=c(4)
kuadrat(A)
mode(kuadrat)
```

Perhatikan bahwa **kuadrat** merupakan nama variabel yang menyimpan data berjenis fungsi. Ilustrasi dalam R diperlihatkan pada Gambar 2.42 dan Gambar 2.43.



Gambar 2.42

```
kuadrat=function(x)
{
  print(x*x);
}
A=c(4)
kuadrat(A)
```

```
## [1] 16
```

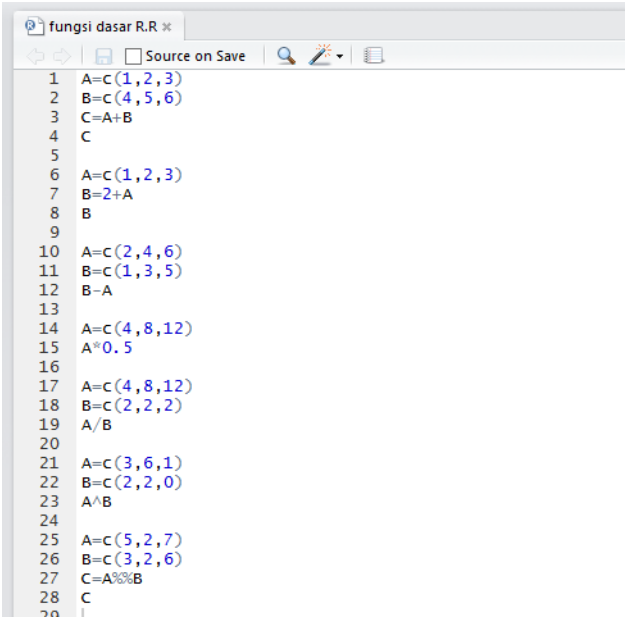
```
mode(kuadrat)
```

```
## [1] "function"
```

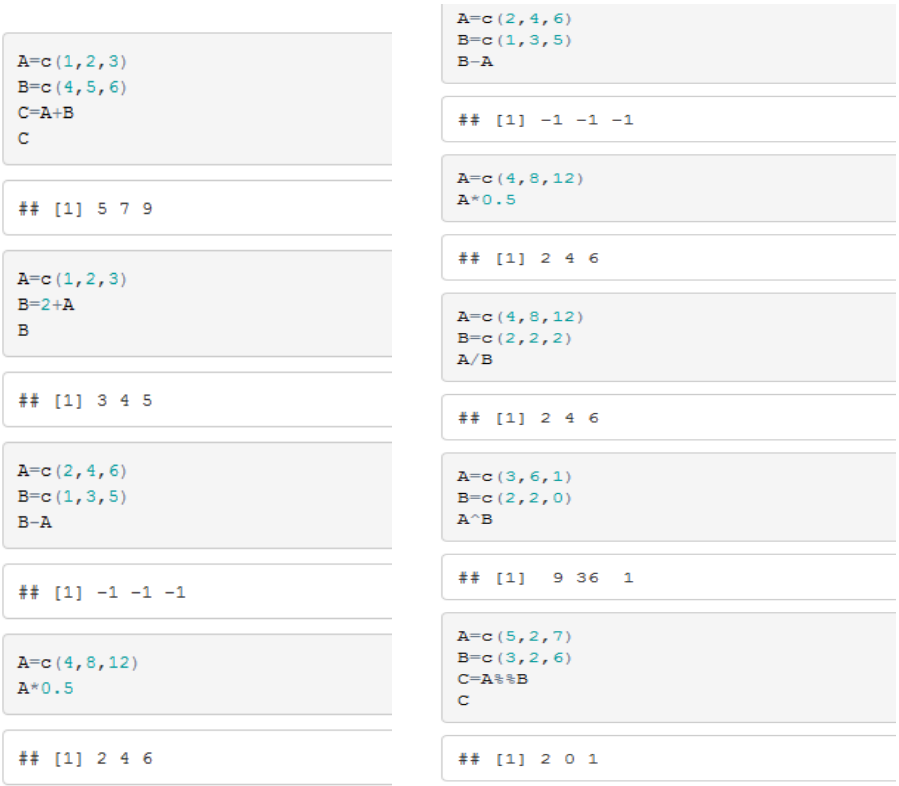
Gambar 2.43

*Operator Penjumlahan +, Pengurangan -, Perkalian *, Pembagian /, Pangkat ^, Sisa %%*

Gambar 2.44 dan Gambar 2.45 merupakan berbagai contoh kode R yang melibatkan penggunaan operator matematika.



Gambar 2.44



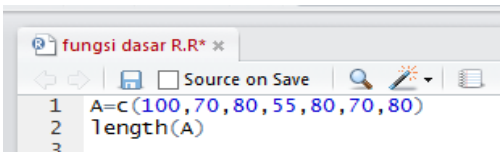
Gambar 2.45

Fungsi length

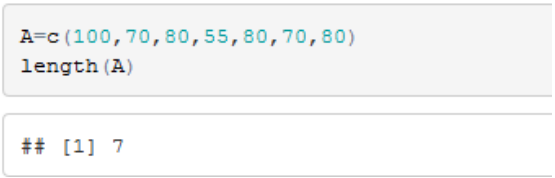
Fungsi **length** dalam R berfungsi untuk mengetahui jumlah elemen yang tersimpan atau terkandung dalam variabel. Misalkan suatu variabel bernama A menyimpan nilai 100, 70, 80, 55, 80, 70, 80. Maka banyaknya elemen dalam variabel A adalah 7. Berikut merupakan kode R untuk menentukan banyaknya elemen yang terkandung dalam variabel A.

A=c(100,70,80,55,80,70,80)
length(A)

Ilustrasi dalam R diperlihatkan pada Gambar 2.46 dan Gambar 2.47.



Gambar 2.46



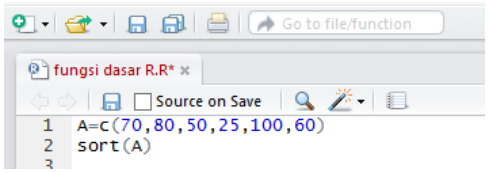
Gambar 2.47

Fungsi sort

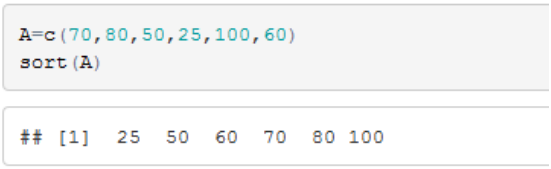
Fungsi **sort** dalam R berfungsi untuk mengurutkan data. Misalkan suatu variabel bernama A menyimpan nilai 70, 80, 50, 25, 100, 60. Berikut merupakan kode R untuk mengurutkan elemen-elemen atau nilai-nilai yang terkandung dalam variabel A.

A=c(70,80,50,25,100,60)
sort(A)

Ilustrasi dalam R diperlihatkan pada Gambar 2.48 dan Gambar 2.49.



Gambar 2.48



Gambar 2.49

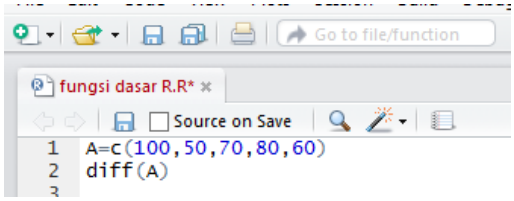
Fungsi diff

Misalkan diberikan data sebagai berikut. 100, 50, 70, 80, 60. Misalkan dilakukan perhitungan sebagai berikut.

➔$50 - 100 = -50$
➔$70 - 50 = 20$
➔$80 - 70 = 10$
➔$60 - 80 = -20$

Sehingga hasil akhirnya adalah $-50, 20, 10, -20$. Gambar 2.50 dan Gambar 2.51 merupakan penggunaan fungsi **diff** untuk mengilustrasikan contoh tersebut.

A=c(100,50,70,80,60)
diff(A)



Gambar 2.50

```
A=c(100,50,70,80,60)
diff(A)
```

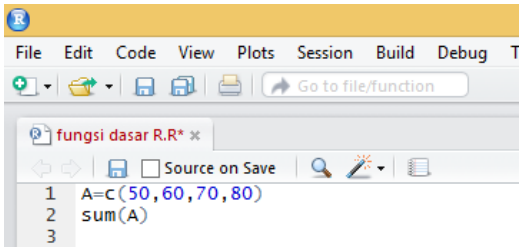
```
## [1] -50  20  10 -20
```

Gambar 2.51

Fungsi sum

Fungsi **sum** dalam R berfungsi untuk menjumlahkan seluruh nilai data. Misalkan variabel A menyimpan nilai 50, 60, 70, 80. Maka jumlah dari seluruh nilai dalam variabel A adalah 260.

A=c(50,60,70,80)
sum(A)



Gambar 2.52

```
A=c(50,60,70,80)
sum(A)
```

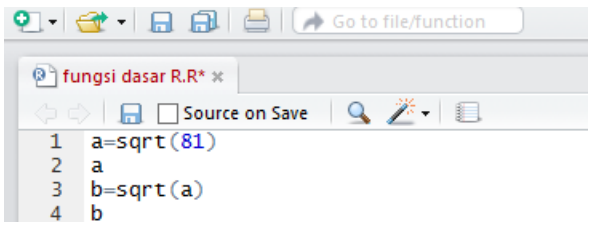
```
## [1] 260
```

Gambar 2.53

Fungsi sqrt

Fungsi **sqrt** dalam R berfungsi untuk menghitung nilai akar pangkat dua dari suatu bilangan. Sebagai contoh akar pangkat 2 dari 81 adalah 9, yakni $\sqrt[2]{81} = \sqrt{81} = 9$. Berikut merupakan kode R untuk menghitung nilai akar pangkat dua dari 81.

sqrt(81)



Gambar 2.54

```
a=sqrt(81)
a
```

```
## [1] 9
```

```
b=sqrt(a)
b
```

```
## [1] 3
```

Gambar 2.55

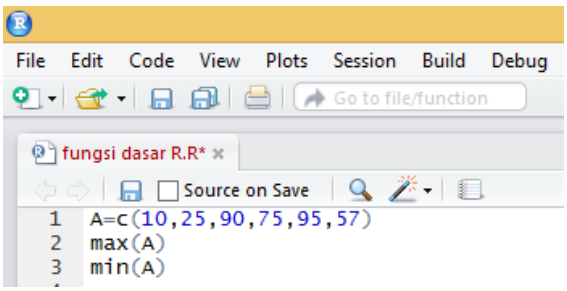
Fungsi max dan min

Fungsi **max** dalam R berfungsi untuk menentukan nilai maksimum dalam data. Misalkan diberikan data 10,25,90,75, 95, 57. Maka nilai maksimum dari data tersebut adalah 95. Berikut merupakan kode dalam R untuk menentukan nilai maksimum dari data tersebut.

A=c(10,25,90,75,95,57)
max(A)

Fungsi **min** dalam R berfungsi untuk menentukan nilai minimum dalam data. Misalkan diberikan data 10,25,90,75, 95, 57. Maka nilai minimum dari data tersebut adalah 10. Berikut merupakan kode dalam R untuk menentukan nilai minimum dari data tersebut.

A=c(10,25,90,75,95,57)
min(A)



Gambar 2.56

```
A=c(10,25,90,75,95,57)
max(A)
```

```
## [1] 95
```

```
min(A)
```

```
## [1] 10
```

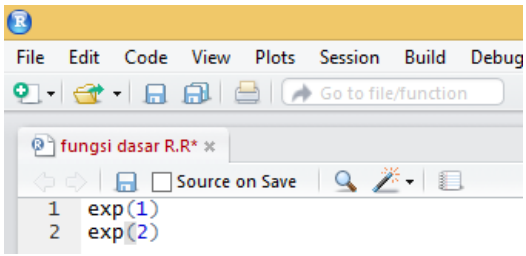
Gambar 2.57

Fungsi exp

Exp merupakan singkatan dari *exponential* atau eksponensial. Nilai dari eksponensial adalah 2,71828182845…

$$exp = 2{,}71828182845$$
$$exp^1 = 2{,}71828182845^1 = 2{,}71828182845$$
$$\exp^2 = 2{,}71828182845^2 = 7{,}389056096$$

Berikut diberikan contoh penggunaan fungsi **exp** dalam R (Gambar 2.58 dan Gambar 2.59).



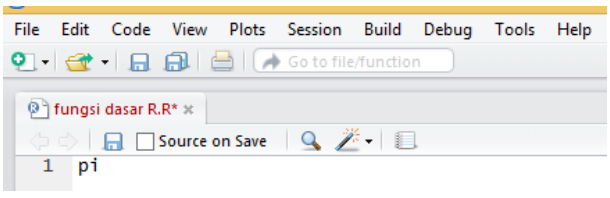
Gambar 2.58

```
exp(1)

## [1] 2.718282

exp(2)

## [1] 7.389056
```

Gambar 2.59

Fungsi pi atau π

Pi atau π bernilai 3,141593 ... Berikut diberikan contoh penggunaan fungsi **pi** dalam R (Gambar 2.60 dan Gambar 2.61).



Gambar 2.60

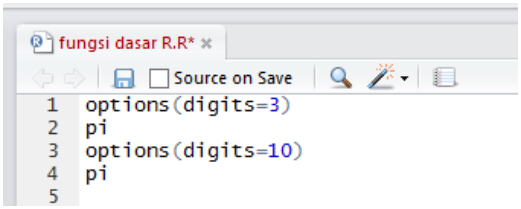
```
pi

## [1] 3.141593
```

Gambar 2.61

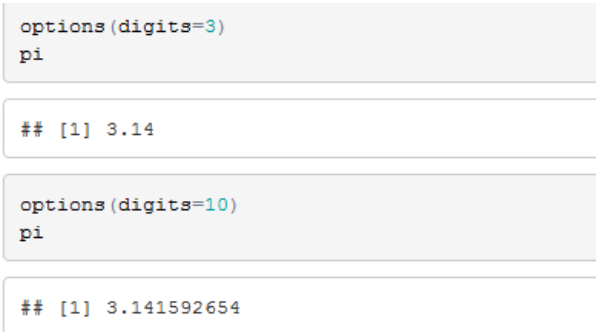
Fungsi options

Diketahui nilai pi adalah 3.141593… Misalkan hanya ingin ditampilkan 3 digit angka dari nilai pi, yakni 3.14. Berikut perintah dalam R untuk menampilkan hanya 3 digit angka dari bilangan pi.

options(digits=3)
pi



Gambar 2.62



Gambar 2.63

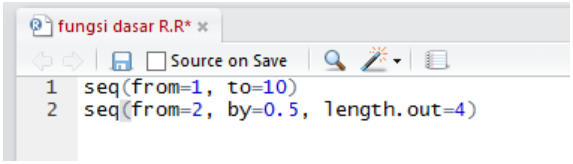
Fungsi seq

Misalkan ingin ditampilkan nilai data dari 1 sampai 10. Berikut perintah dalam R untuk menampilkan nilai data dari 1 sampai 10.

seq(from=1, to=10)

Misalkan ingin ditampilkan 4 buah nilai, dimulai dari 2 kemudian 2.5, 3, dan 3.5, dimana jaraknya adalah 0.5. Berikut merupakan perintah dalam R.

seq(from=2, by=0.5, length.out=4)



Gambar 2.64

```
seq(from=1, to=10)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

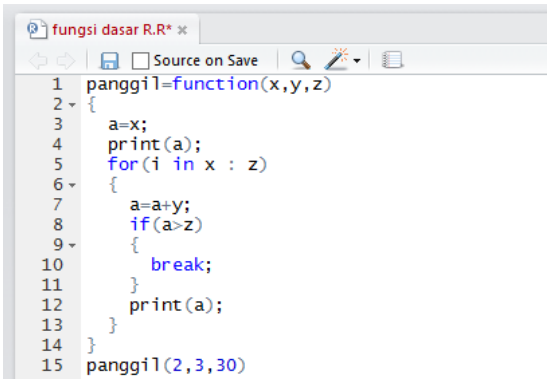
```
seq(from=2, by=0.5, length.out=4)
```

```
## [1] 2.0 2.5 3.0 3.5
```

Gambar 2.65

Misalkan ingin ditampilkan nilai-nilai kelipatan 3, dimulai dari angka 2 sampai 30. Adapun nilai-nilai tersebut adalah 2, 5, 8, 11, 14, 17, 20, 23, 26, 29. Berikut diberikan contoh kode program R untuk menyelesaikan permasalahan tersebut.

```
panggil=function(x,y,z)
{
  a=x;
  print(a);
  for(i in x : z)
    {
      a=a+y;
      if(a>z)
        {
          break;
        }
  print(a);
    }
}
panggil(2,3,30)
```



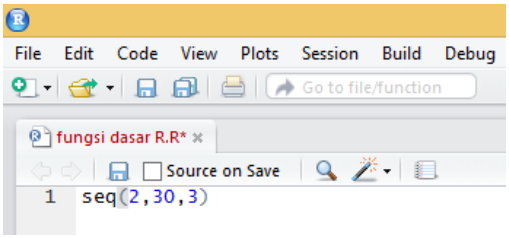
Gambar 2.66

```
panggil=function(x,y,z)
{
  a=x;
  print(a);
  for(i in x : z)
  {
    a=a+y;
    if(a>z)
    {
      break;
    }
    print(a);
  }
}
panggil(2,3,30)
```

```
## [1] 2
## [1] 5
## [1] 8
## [1] 11
## [1] 14
## [1] 17
## [1] 20
## [1] 23
## [1] 26
## [1] 29
```

Gambar 2.67

Cara lain adalah sebagai berikut.



Gambar 2.68

```
seq(2,30,3)
```

```
##  [1]  2  5  8 11 14 17 20 23 26 29
```

Gambar 2.69

Fungsi table

Fungsi **table** dalam R berfungsi untuk menyajikan data dalam bentuk tampilan tabel. Misalkan suatu variabel bernama A menyimpan data 10, 10, 30, 10, 30, 10, 10, 40, 40,70, 90, 70, 80, 60, 60, 90. Berikut merupakan perintah atau kode R untuk menyajikan data pada variabel A dalam tabel.

A=c(10, 10, 30, 10, 30, 10, 10, 40, 40,70, 90, 70, 80, 60, 60, 90)
table(A)

Penyajian secara tabel juga dapat disajikan dengan menampilkan informasi proporsi. Berikut merupakan perintah atau kode R untuk menyajikan tabel dengan informasi proporsi.

A=c(10, 10, 30, 10, 30, 10, 10, 40, 40,70, 90, 70, 80, 60, 60, 90)
table(A)/length(A)

Contoh lain misalkan suatu survey yang dilakukan terhadap 10 orang sebagai sampel untuk melihat apakah seseorang tersebut terkena insomnia (ya) atau tidak (tidak). Hasil atau data disajikan sebagai berikut.

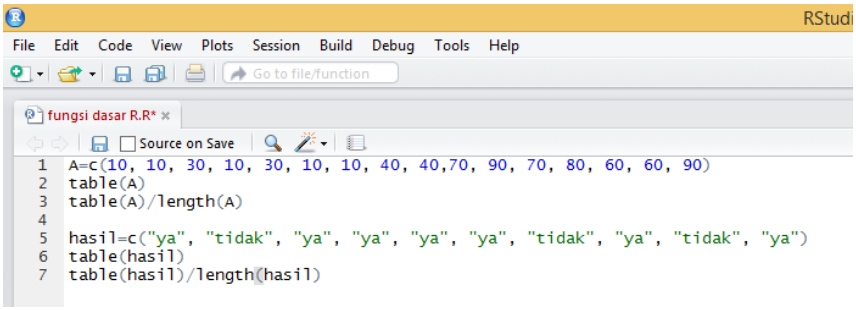
ya, tidak, ya, ya, ya, ya, tidak, ya, tidak, ya

Berikut merupakan perintah atau kode R untuk menyajikan data di atas dalam tabel.

hasil=c(“ya”, “tidak”, “ya”, “ya”, “ya”, “ya”, “tidak”, “ya”, “tidak”, “ya”)
table(hasil)

Penyajian secara tabel juga dapat disajikan dengan menampilkan informasi proporsi. Berikut merupakan perintah atau kode R untuk menyajikan tabel dengan informasi proporsi.

hasil=c(“ya”, “tidak”, “ya”, “ya”, “ya”, “ya”, “tidak”, “ya”, “tidak”, “ya”)
table(hasil)/length(hasil)



Gambar 2.70

```
A=c(10, 10, 30, 10, 30, 10, 10, 40, 40,70, 90, 70, 80, 60, 60, 90)
table(A)
```

```
## A
## 10 30 40 60 70 80 90
##  5  2  2  2  2  1  2
```

```
table(A)/length(A)
```

```
## A
##     10     30     40     60     70     80     90
## 0.3125 0.1250 0.1250 0.1250 0.1250 0.0625 0.1250
```

```
hasil=c("ya", "tidak", "ya", "ya", "ya", "ya", "tidak", "ya", "tidak", "ya")
table(hasil)
```

```
## hasil
## tidak    ya
##     3     7
```

```
table(hasil)/length(hasil)
```

```
## hasil
## tidak    ya
##   0.3   0.7
```

Gambar 2.71

Fungsi factor

Fungsi **factor** dalam R berfungsi untuk mengetahui keragaman level atau faktor dalam suatu data. Misalkan diberikan data sebagai berikut.

ikan, ikan, udang, ikan, udang, ikan, ikan, udang

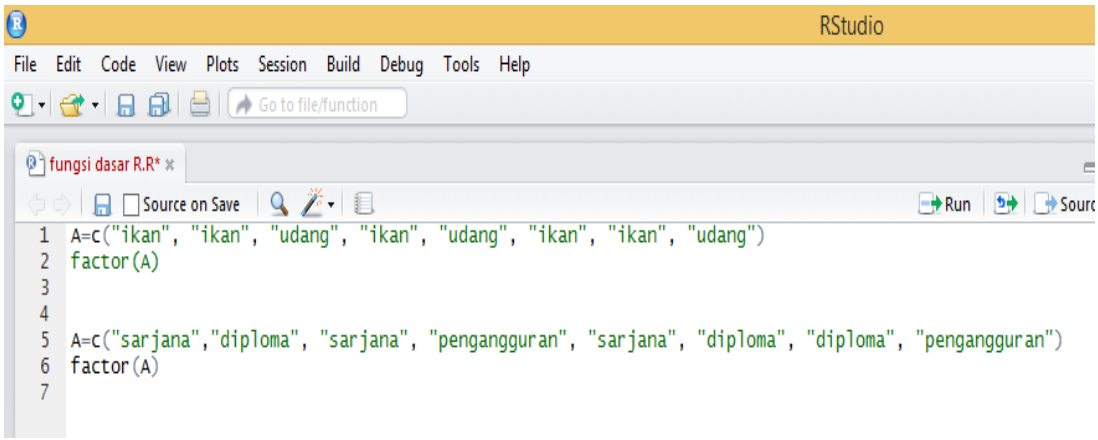
Berdasarkan data tersebut, terdapat dua faktor, yakni ikan dan udang. Misalkan diberikan data sebagai berikut.

sarjana,diploma, sarjana, pengangguran, sarjana, diploma, diploma, pengangguran

Berdasarkan data tersebut, terdapat tiga faktor, yakni sarjana, diploma, dan pengangguran. Berikut merupakan perintah atau kode R dalam penggunaan fungsi **factor( )**.

A=c(“ikan”, “ikan”, “udang”,” ikan”, “udang”, “ikan”, “ikan”,” udang”)
factor(A)

A=c(“sarjana”,”diploma”, “sarjana”, “pengangguran”, “sarjana”, “diploma”, “diploma”, “pengangguran”)
factor(A)



Gambar 2.72

```
A=c("ikan", "ikan", "udang", "ikan", "udang", "ikan", "ikan", "udang")
factor(A)
```

```
## [1] ikan  ikan  udang ikan  udang ikan  ikan  udang
## Levels: ikan udang
```

```
A=c("sarjana","diploma", "sarjana", "pengangguran", "sarjana", "diploma", "diploma", "pengangguran")
factor(A)
```

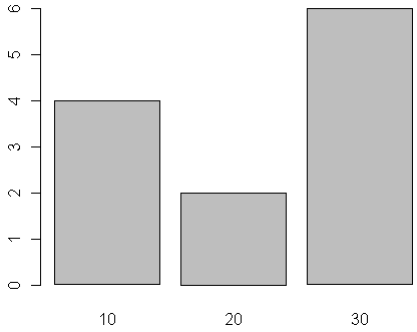
```
## [1] sarjana     diploma     sarjana     pengangguran sarjana
## [6] diploma     diploma     pengangguran
## Levels: diploma pengangguran sarjana
```

Gambar 2.73

Fungsi barplot

Fungsi **barplot** dalam R berfungsi untuk menyajikan data dalam bentuk diagram batang. Misalkan variabel A menyimpan data 10, 10, 10, 10, 20, 20, 30, 30, 30, 30, 30, 30. Berikut akan disajikan data pada variabel A dalam bentuk diagram batang.

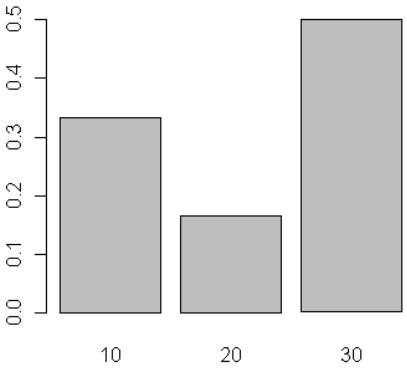
A=c(10, 10, 10, 10, 20, 20, 30, 30, 30, 30, 30, 30)
barplot(table(A))



Gambar 2.74

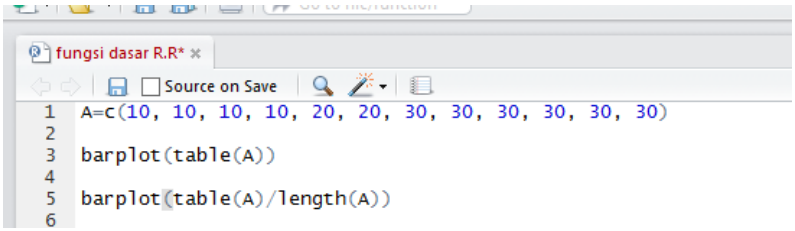
Perhatikan bahwa untuk data dengan nilai 10 mempunyai frekuensi sebanyak 4, data dengan nilai 20 mempunyai frekuensi sebanyak 2, dan data dengan nilai 30 mempunyai frekuensi sebanyak 6. Grafik batang di atas dapat diatur agar disajikan secara proporsi.

A=c(10, 10, 10, 10, 20, 20, 30, 30, 30, 30, 30, 30)
barplot(table(A)/length(A))



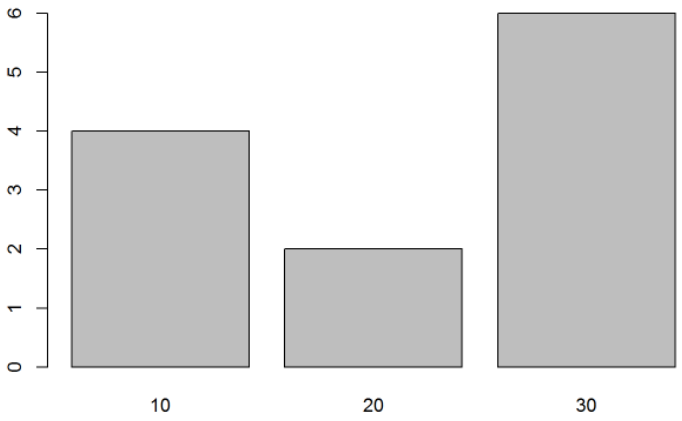
Gambar 2.75

Perhatikan bahwa nilai 0,3, 0,2, dan 0,5 masing-masing merupakan proporsi dari nilai 10, 20, dan 30.



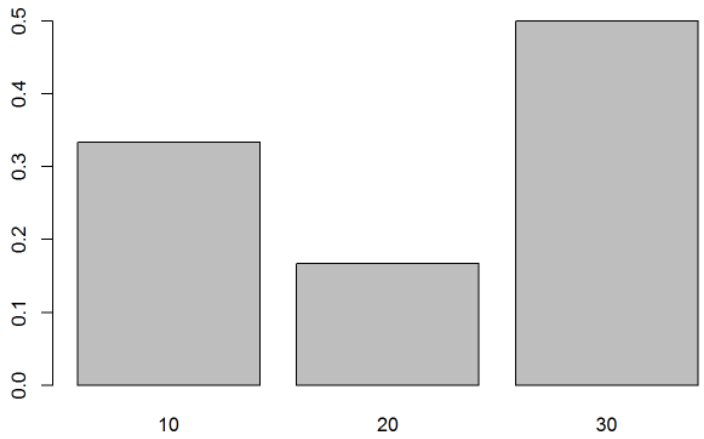
Gambar 2.76

```
A=c(10, 10, 10, 10, 20, 20, 30, 30, 30, 30, 30, 30)

barplot(table(A))
```



Gambar 2.77

```
barplot(table(A)/length(A))
```



Gambar 2.78

Fungsi plot

Misalkan variabel bernama A menyimpan data 10,10,10,10,10,20,20,20,30,30,40. Berikut akan digunakan fungsi **table** untuk mengetahui frekuensi dari masing-masing nilai data.

A=c(10,10,10,10,10,20,20,20,30,30,40)
table(A)

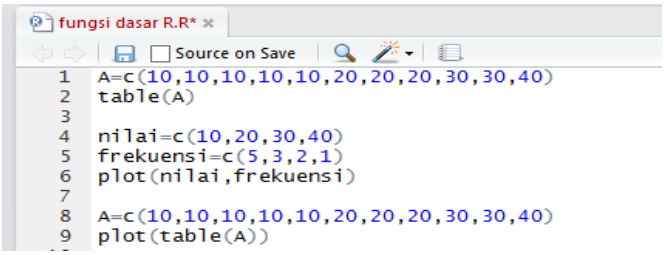
Diketahui nilai 10 muncul sebanyak 5, nilai 20 sebanyak 3, nilai 30 sebanyak 2, dan nilai 40 sebanyak 1. Berikut akan digunakan fungsi **plot( )** untuk memplot data yang tersimpan dalam variabel A.

```
nilai=c(10,20,30,40)
frekuensi=c(5,3,2,1)
plot(nilai,frekuensi)
```

Alternatif lain untuk menyajikan data.

```
A=c(10,10,10,10,10,20,20,20,30,30,40)
plot(table(A))
```

Ilustrasi dalam R diperlihatkan pada Gambar 2.79 dan Gambar 2.82.



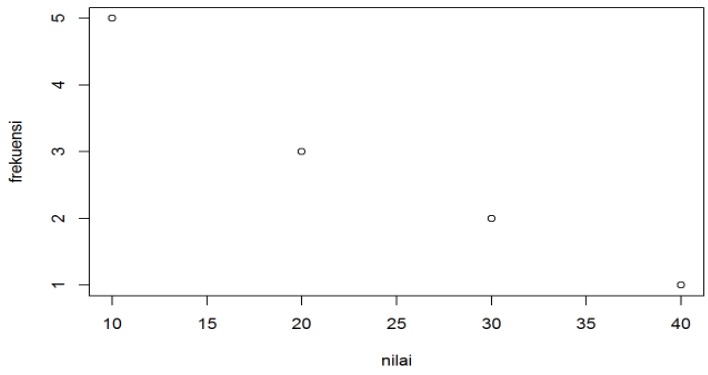
Gambar 2.79

```
A=c(10,10,10,10,10,20,20,20,30,30,40)
table(A)
```

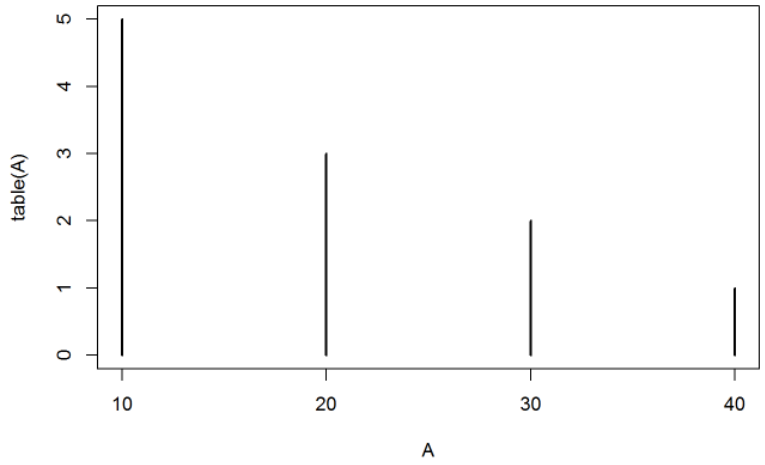
```
## A
## 10 20 30 40
##  5  3  2  1
```

```
nilai=c(10,20,30,40)
frekuensi=c(5,3,2,1)
plot(nilai,frekuensi)
```

Gambar 2.80



Gambar 2.81



Gambar 2.82

Referensi

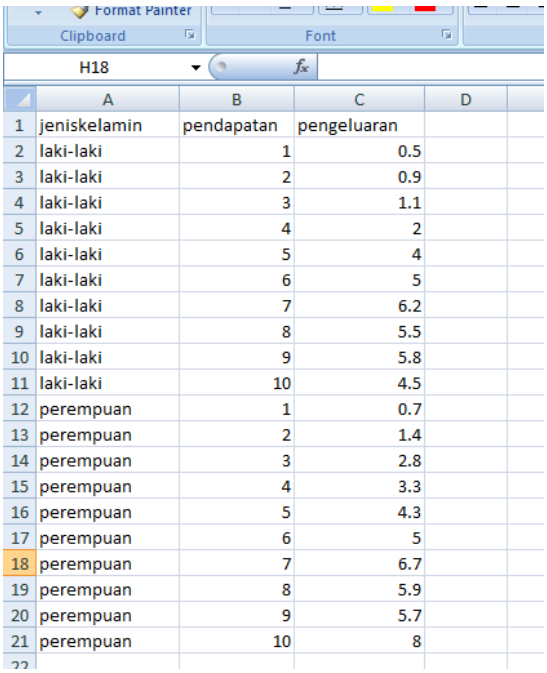
1. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.
2. http://www.statmethods.net/graphs/bar.html
3. http://www.r-tutor.com/elementary-statistics/qualitative-data/bar-graph
4. http://www.r-bloggers.com/using-r-barplot-with-ggplot2/
5. http://www.statmethods.net/graphs/line.html
6. http://www.statmethods.net/management/functions.html
7. http://www.r-bloggers.com/basic-mathematical-functions/
8. http://ww2.coastal.edu/kingw/statistics/R-tutorials/arithmetic.html

BAB 3

MENYAJIKAN DATA DALAM GRAFIK

Memplot Data dalam R (Scatter Plot)

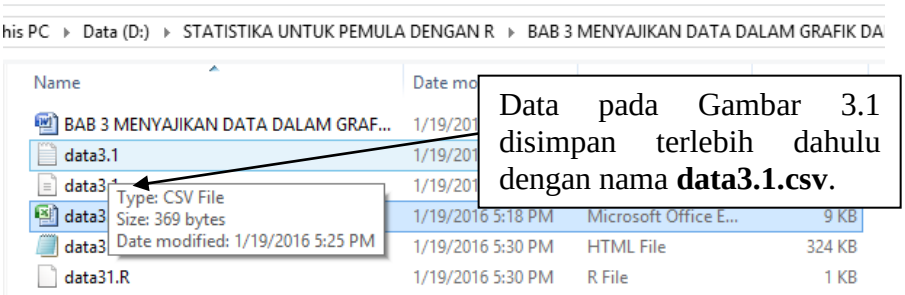
Misalkan diberikan data seperti pada Gambar 3.1. Berdasarkan Gambar 3.1, diketahui terdapat 10 responden laki-laki dan 10 responden perempuan. Masing-masing responden disajikan informasi mengenai pendapatan dan pengeluaran per-bulan, dalam jutaan. Sebagai contoh, responden ke-1 adalah laki-laki, dengan pendapatan Rp. 1.000.000, dan pengeluaran Rp. 500.000. Responden ke-20 adalah perempuan, dengan pendapatan Rp. 10.000.000, dan pengeluaran Rp. 8.000.000.



	A	B	C	D
1	jeniskelamin	pendapatan	pengeluaran	
2	laki-laki	1	0.5	
3	laki-laki	2	0.9	
4	laki-laki	3	1.1	
5	laki-laki	4	2	
6	laki-laki	5	4	
7	laki-laki	6	5	
8	laki-laki	7	6.2	
9	laki-laki	8	5.5	
10	laki-laki	9	5.8	
11	laki-laki	10	4.5	
12	perempuan	1	0.7	
13	perempuan	2	1.4	
14	perempuan	3	2.8	
15	perempuan	4	3.3	
16	perempuan	5	4.3	
17	perempuan	6	5	
18	perempuan	7	6.7	
19	perempuan	8	5.9	
20	perempuan	9	5.7	
21	perempuan	10	8	

Gambar 3.1

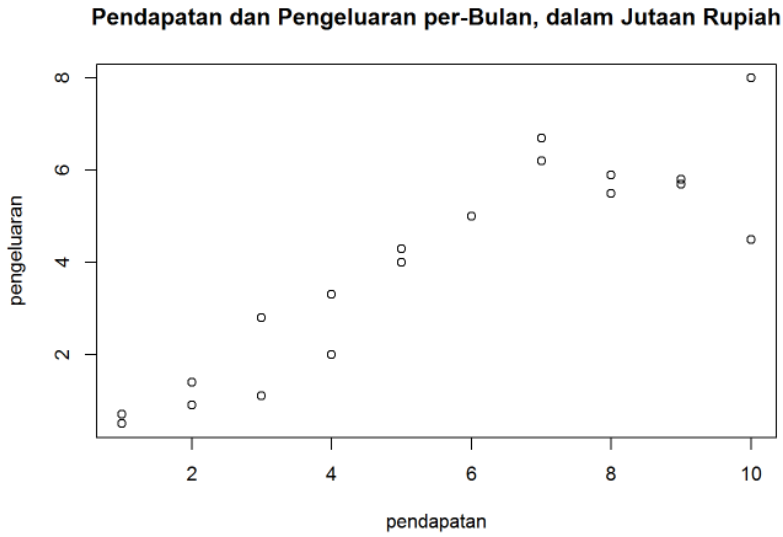
Data pada Gambar 3.1 disimpan terlebih dahulu dengan nama **data3.1.csv** (perhatikan Gambar 3.2).



Gambar 3.2

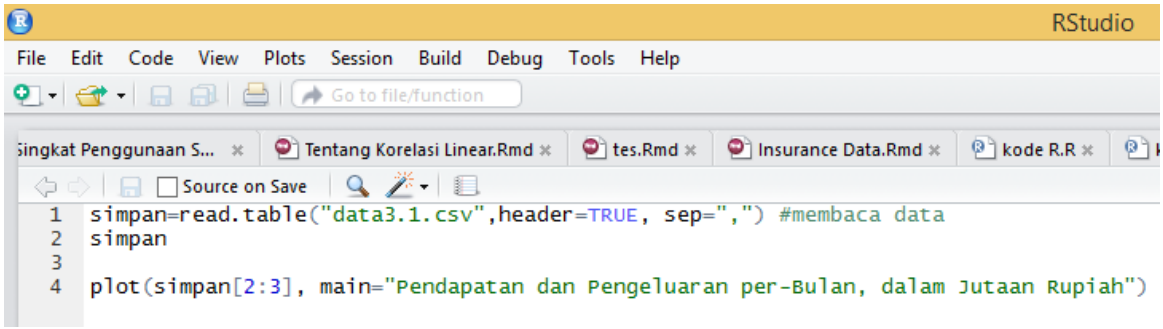
Data berdasarkan Gambar 3.1 disajikan ke dalam grafik seperti pada Gambar 3.3.

```
plot(simpan[2:3], main="Pendapatan dan Pengeluaran per-Bulan, dalam Jutaan Rupiah")
```



Gambar 3.3

Kode R untuk menyajikan data pada Gambar 3.1, seperti pada Gambar 3.3, adalah sebagai berikut (Gambar 3.4).



Gambar 3.4

Berdasarkan Gambar 3.4, perhatikan kode R berikut (kode R baris pertama).

simpan=read.table("data3.1.csv",header=TRUE, sep=",") #membaca data

Kode R tersebut (kode R baris pertama) dapat diartikan variabel **simpan** ditugaskan untuk menyimpan data pada variabel **jeniskelamin**, **pendapatan**, dan **pengeluaran** dalam *file* **data3.1.csv**. Perhatikan kode R berikut (kode R baris kedua).

simpan

Kode R baris kedua berarti menampilkan nilai yang disimpan dalam variabel **simpan**. Hasilnya seperti pada Gambar 3.5.

```
simpan=read.table("data3.1.csv",header=TRUE, sep=",") #membaca data
simpan
```

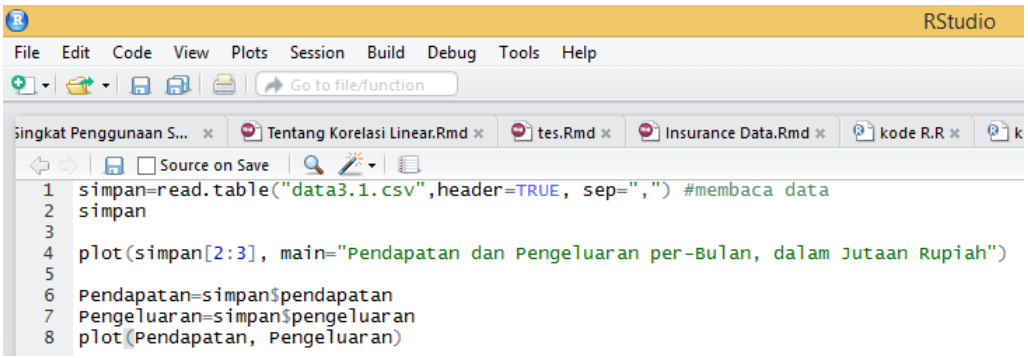
```
##    jeniskelamin pendapatan pengeluaran
## 1     laki-laki          1         0.5
## 2     laki-laki          2         0.9
## 3     laki-laki          3         1.1
## 4     laki-laki          4         2.0
## 5     laki-laki          5         4.0
## 6     laki-laki          6         5.0
## 7     laki-laki          7         6.2
## 8     laki-laki          8         5.5
## 9     laki-laki          9         5.8
## 10    laki-laki         10         4.5
## 11    perempuan          1         0.7
## 12    perempuan          2         1.4
## 13    perempuan          3         2.8
## 14    perempuan          4         3.3
## 15    perempuan          5         4.3
## 16    perempuan          6         5.0
## 17    perempuan          7         6.7
## 18    perempuan          8         5.9
## 19    perempuan          9         5.7
## 20    perempuan         10         8.0
```

Gambar 3.5

Kode R pada baris keempat (Gambar 3.6), yakni

plot(simpan[2:3], main="Pendapatan dan Pengeluaran per-Bulan, dalam Jutaan Rupiah")

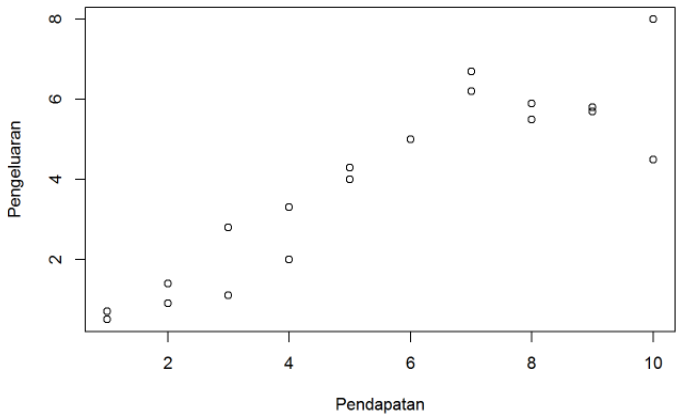
dapat diartikan data pada variabel **pendapatan** (pada kolom 2) dan data pada variabel **pengeluaran** (pada kolom 3), disajikan ke dalam grafik, seperti pada Gambar 3.3. Kode R pada baris keempat mencantumkan **main="Pendapatan dan Pengeluaran per-Bulan, dalam Jutaan Rupiah"**, yang berguna untuk memberikan judul grafik. Pada Gambar 3.6, kode R pada baris 6 sampai baris 8, apabila dieksekusi, hasilnya seperti pada Gambar 3.7.



Gambar 3.6

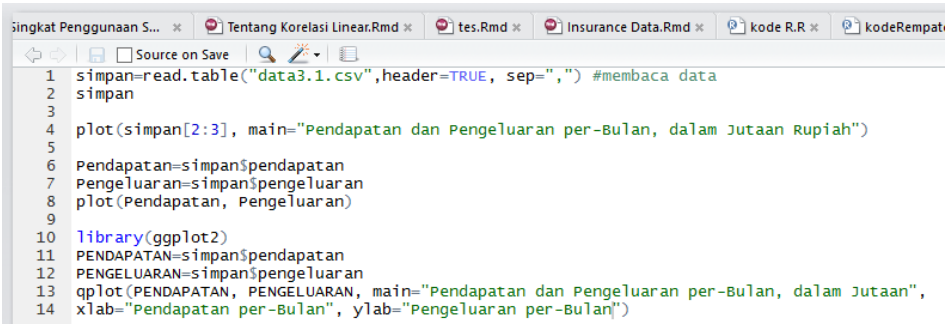
Pada Gambar 3.6, kode R pada baris keenam, yakni **Pendapatan=simpan$pendapatan**, berarti variabel **Pendapatan** ditugaskan untuk menyimpan data pada variabel **pendapatan**, dalam variabel **simpan**. Kode R pada baris ketujuh, yakni **Pengeluaran= simpan$pengeluaran**, berarti variabel **Pengeluaran** ditugaskan untuk menyimpan data pada variabel **pengeluaran**, dalam variabel **simpan**. Kode R pada baris kedelapan, yakni **plot(Pendapatan, Pengeluaran)**, berarti memplot data ke dalam grafik, dengan variabel

Pendapatan sebagai sumbu horizontal, dan variabel **Pengeluaran** sebagai sumbu vertikal. Hasilnya seperti pada Gambar 3.7.



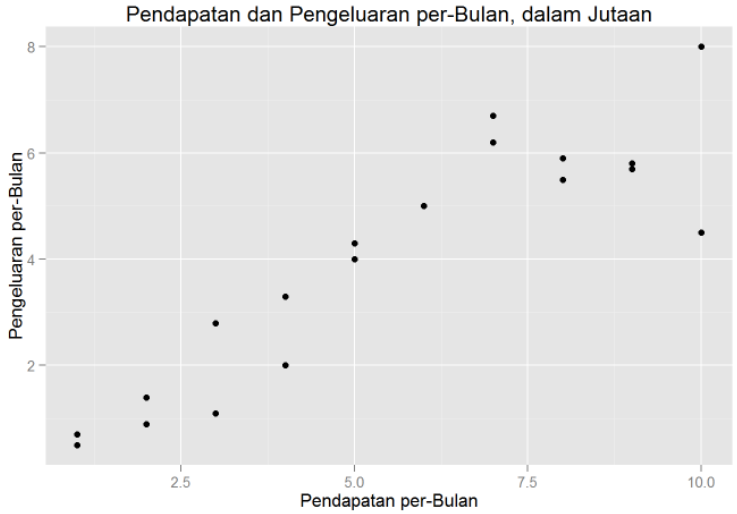
Gambar 3.7

Pada Gambar 3.8, kode R pada baris 10 sampai baris 14, apabila dieksekusi, hasilnya seperti pada Gambar 3.9.



Gambar 3.8

```
PENDAPATAN=simpan$pendapatan
PENGELUARAN=simpan$pengeluaran
qplot(PENDAPATAN, PENGELUARAN, main="Pendapatan dan Pengeluaran per-Bulan, dalam Jutaan",
xlab="Pendapatan per-Bulan", ylab="Pengeluaran per-Bulan")
```



Gambar 3.9

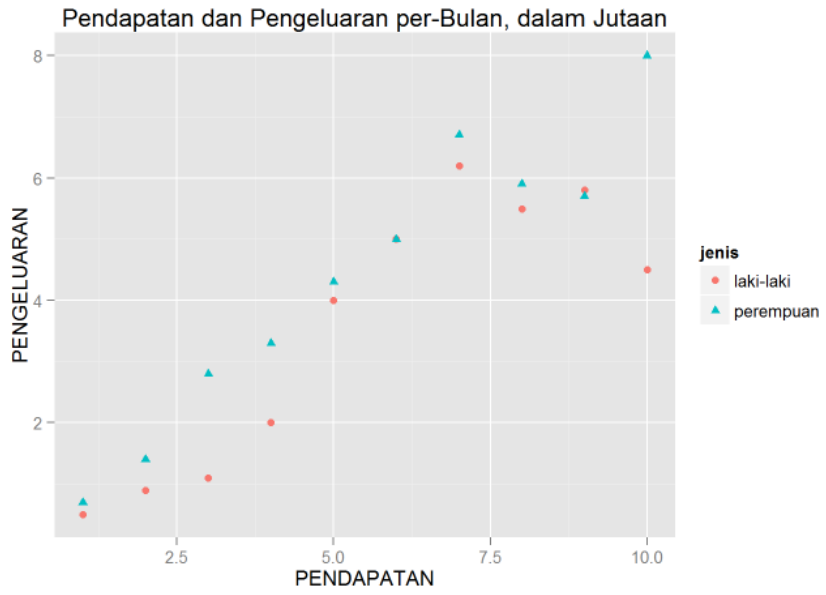
Pada Gambar 3.8, kode R pada baris kesepuluh, yakni **library(ggplot2)**, berarti mengaktifkan *package* **ggplot2**. Pengaktifkan *package* **ggplot2** bertujuan untuk menggunakan fungsi **qplot( )**. Kode R pada baris kesebelas, yakni **PENDAPATAN=simpan$pendapatan**, berarti variabel **PENDAPATAN** ditugaskan untuk menyimpan data pada variabel **pendapatan**, dalam variabel **simpan**. Kode R pada baris keduabelas, yakni **PENGELUARAN=simpan$ pengeluaran**, berarti variabel **PENGELUARAN** ditugaskan untuk menyimpan data pada variabel **pendapatan**, dalam variabel **simpan**. Kode R pada baris ketigabelas dan keempatbelas, yakni **qplot(PENDAPATAN, PENGELUARAN, main="Pendapatan dan Pengeluaran per-Bulan, dalam Jutaan", xlab="Pendapatan per-Bulan", ylab="Pengeluaran per-Bulan")**, berarti memplot data ke dalam grafik. Hasilnya seperti pada Gambar 3.9.

Pada Gambar 3.10, kode R pada baris 21 sampai baris 22, apabila dieksekusi, hasilnya seperti pada Gambar 3.11.

```
16 library(ggplot2)
17 jenis=simpan$jeniskelamin
18 qplot(PENDAPATAN, PENGELUARAN, main="Pendapatan dan Pengeluaran per-Bulan, dalam Jutaan",
19 xlab="Pendapatan", ylab="Pengeluaran", color=jenis)
20
21 qplot(PENDAPATAN, PENGELUARAN, main="Pendapatan dan Pengeluaran per-Bulan, dalam Jutaan",
22 xlab="PENDAPATAN", ylab="PENGELUARAN", color=jenis, shape=jenis)
```

Gambar 3.10

```
qplot(PENDAPATAN, PENGELUARAN, main="Pendapatan dan Pengeluaran per-Bulan, dalam Jutaan",
xlab="PENDAPATAN", ylab="PENGELUARAN", color=jenis, shape=jenis)
```



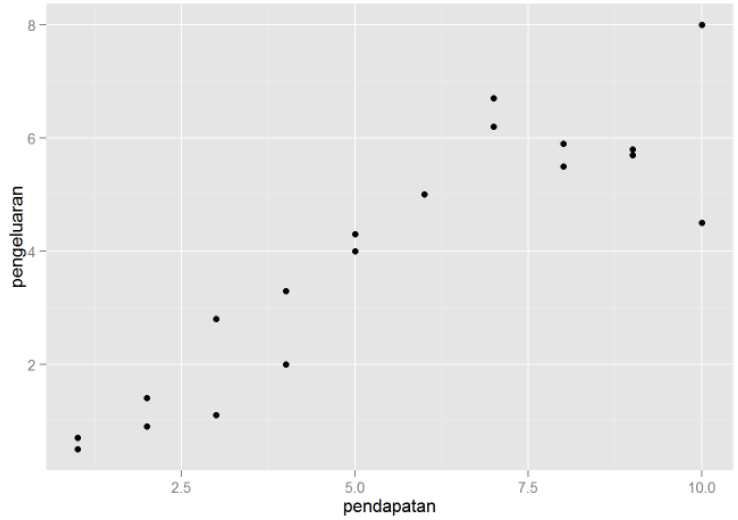
Gambar 3.11

Pada Gambar 3.12, kode R pada baris 24 sampai baris 25, apabila dieksekusi, hasilnya seperti pada Gambar 3.13.

```
19  xlab="Pendapatan", ylab="Pengeluaran", color=jenis)
20
21  qplot(PENDAPATAN, PENGELUARAN, main="Pendapatan dan Pengeluaran per-Bulan, dalam Jutaan",
22  xlab="PENDAPATAN", ylab="PENGELUARAN", color=jenis, shape=jenis)
23
24  library(ggplot2)
25  ggplot(simpan, aes(pendapatan, pengeluaran)) + geom_point()
26
```

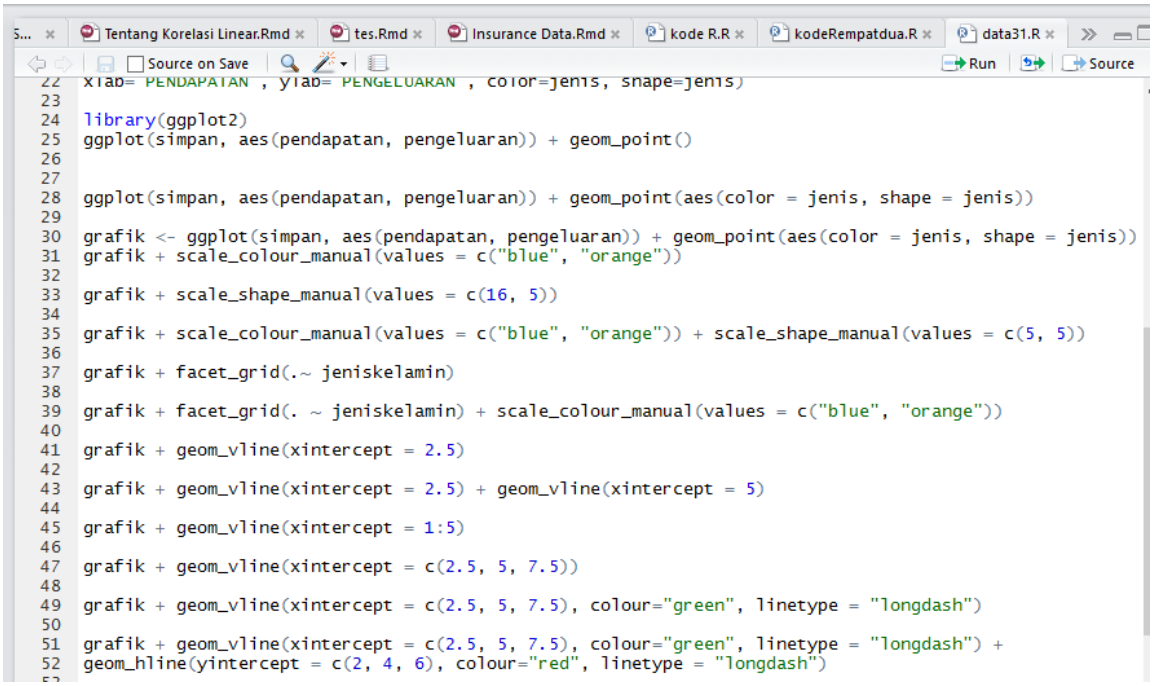
Gambar 3.12

```
library(ggplot2)
ggplot(simpan, aes(pendapatan, pengeluaran)) + geom_point()
```



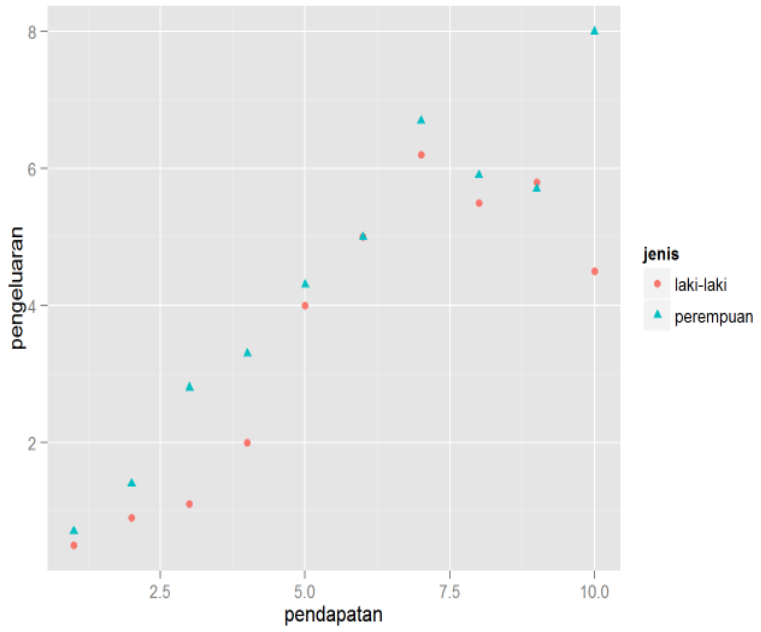
Gambar 3.13

Ketik kode R seperti pada Gambar 3.14, dan amati hasil eksekusi dari kode R tersebut.



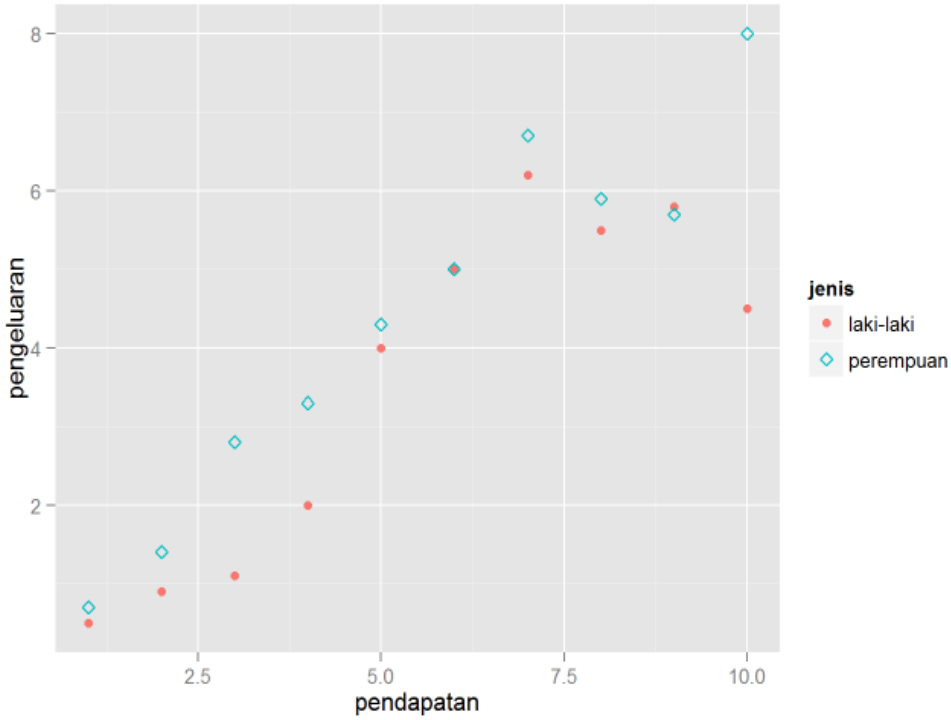
Gambar 3.14

```
ggplot(simpan, aes(pendapatan, pengeluaran)) + geom_point(aes(color = jenis, shape = jenis))
```



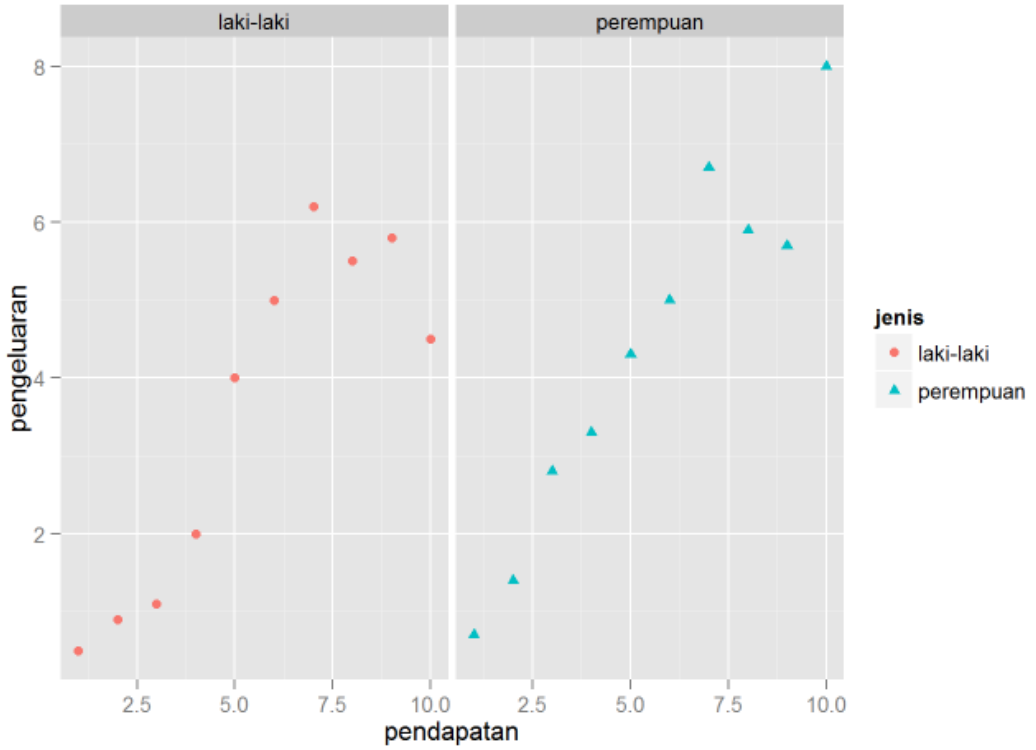
Gambar 3.15

```
grafik + scale_shape_manual(values = c(16, 5))
```



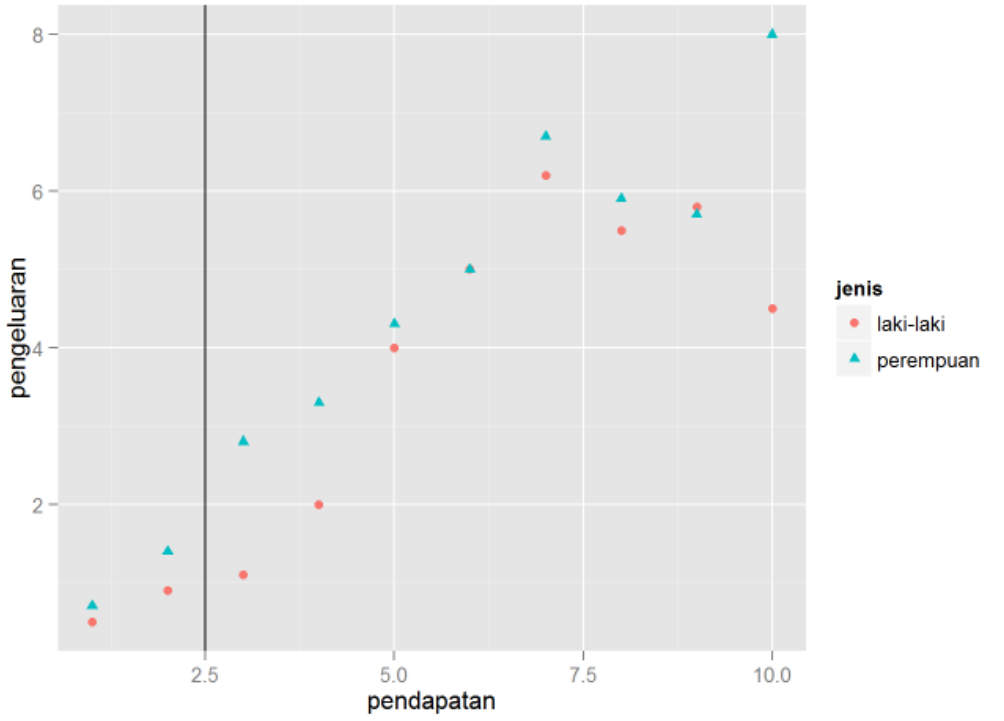
Gambar 3.16

```
grafik + facet_grid(.~ jeniskelamin)
```



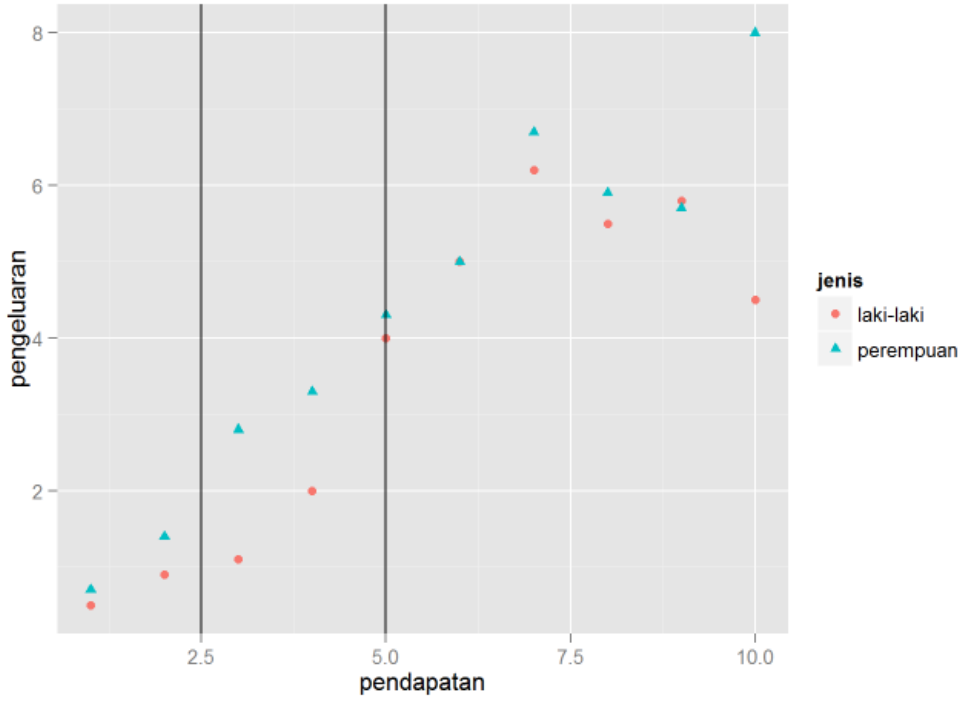
Gambar 3.17

```
grafik + geom_vline(xintercept = 2.5)
```



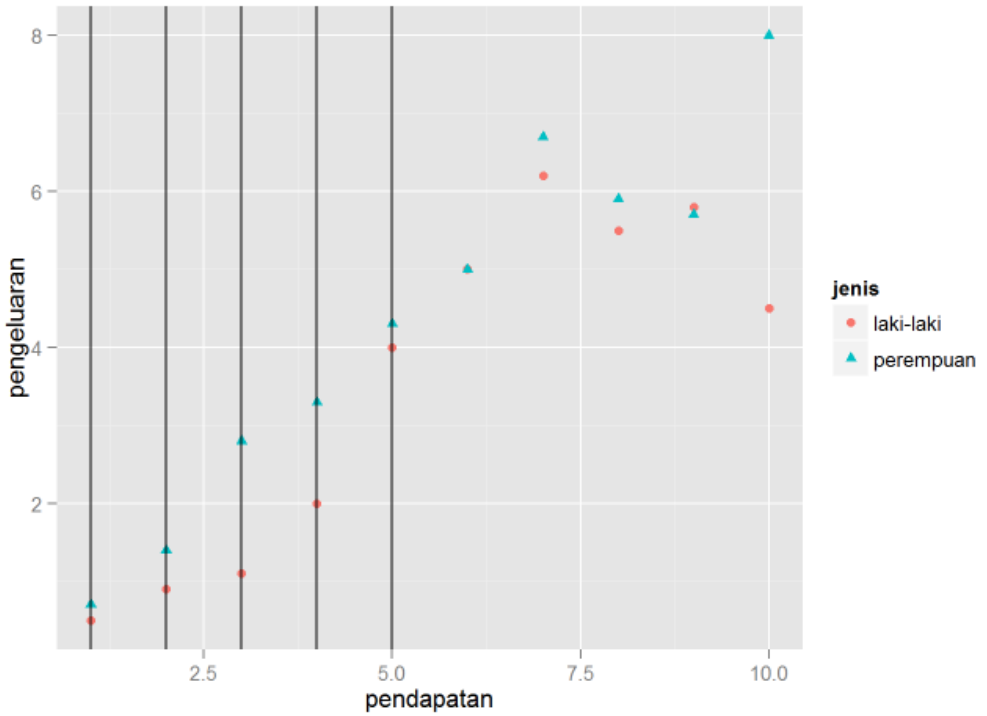
Gambar 3.18

```
grafik + geom_vline(xintercept = 2.5) + geom_vline(xintercept = 5)
```



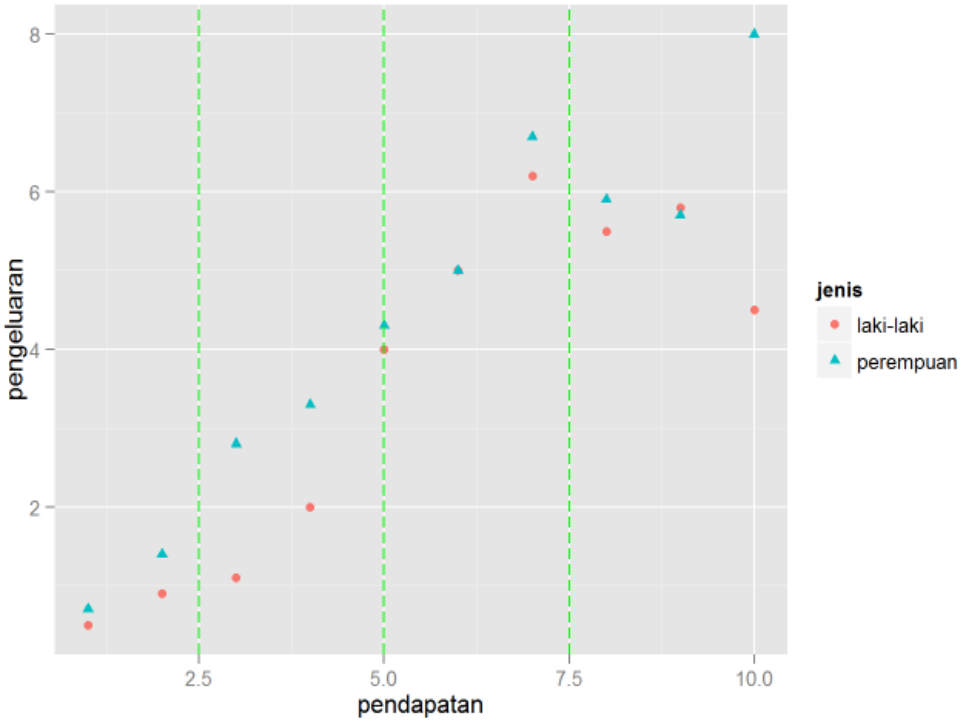
Gambar 3.19

```
grafik + geom_vline(xintercept = 1:5)
```



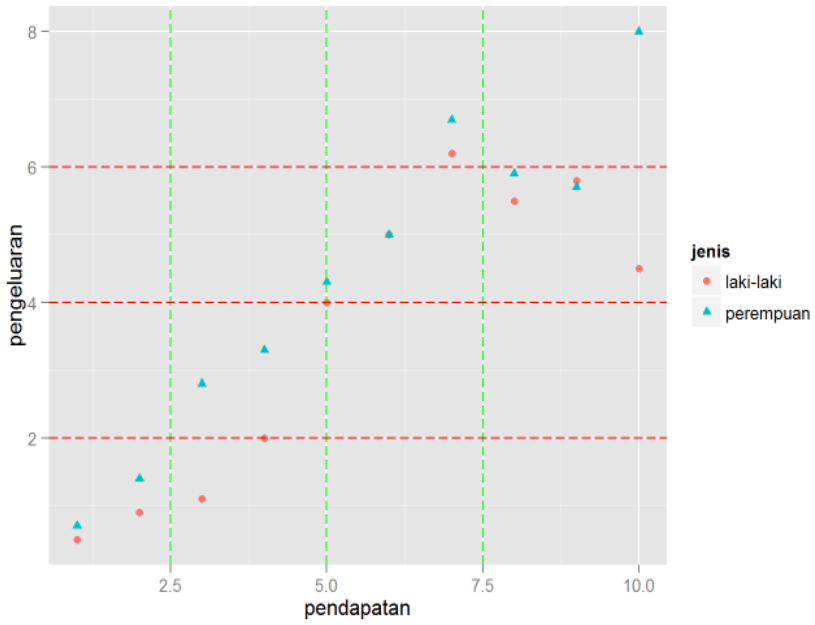
Gambar 3.20

```
grafik + geom_vline(xintercept = c(2.5, 5, 7.5), colour="green", linetype = "longdash")
```



Gambar 3.21

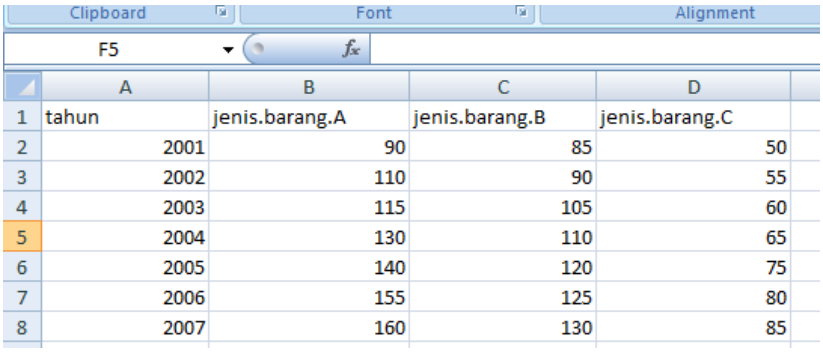
```
grafik + geom_vline(xintercept = c(2.5, 5, 7.5), colour="green", linetype = "longdash") +
geom_hline(yintercept = c(2, 4, 6), colour="red", linetype = "longdash")
```



Gambar 3.22

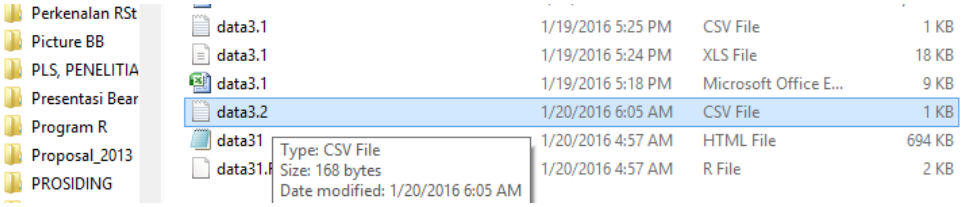
Menyajikan Data dengan Grafik Garis

Misalkan diberikan data seperti pada Gambar 3.23. Gambar 3.23 menyajikan hasil penjualan barang A, B, dan C, selama kurun waktu 2001-2007. Data pada Gambar 3.23 disimpan terlebih dahulu dengan nama **data3.2.csv** (perhatikan Gambar 3.24).



	A	B	C	D
1	tahun	jenis.barang.A	jenis.barang.B	jenis.barang.C
2	2001	90	85	50
3	2002	110	90	55
4	2003	115	105	60
5	2004	130	110	65
6	2005	140	120	75
7	2006	155	125	80
8	2007	160	130	85

Gambar 3.23



Gambar 3.24

Gambar 3.25 sampai dengan Gambar 3.29 merupakan kode R, Eksekusi kode R tersebut, dan amati hasilnya.

```
1  simpan=read.table("data3.2.csv",header=TRUE, sep=",") #membaca data
2  simpan
3
4  Tahun=simpan$tahun
5  Jumlah_A=simpan$jenis.barang.A
6  Jumlah_B=simpan$jenis.barang.B
7  Jumlah_C=simpan$jenis.barang.C
8
9  Jumlah_A
10 Jumlah_B
11 Jumlah_C
12
13 plot(Tahun,Jumlah_A)
14
15 plot(Tahun,Jumlah_A, type="o")
16
17 plot(Tahun,Jumlah_A, type="o", col="blue")
18
19 plot(Tahun,Jumlah_A, type="o", col="green")
20
21 plot(Tahun,Jumlah_A, type="o", col="red")
22 lines(Tahun, Jumlah_B, type="o", col="blue")
23
24 plot(Tahun,Jumlah_A, type="o", col="red", ylim=c(70,180))
25 lines(Tahun, Jumlah_B, type="o", col="blue")
26
27 plot(Tahun,Jumlah_A, type="o", col="red", ylim=c(40,180))
28 lines(Tahun, Jumlah_B, type="o", col="blue")
29 lines(Tahun, Jumlah_C, type="o", col="green")
30
31 plot(Tahun,Jumlah_A, type="o", col="red", ylim=c(40,180))
32 lines(Tahun, Jumlah_B, type="o", pch=22, col="blue")
33 lines(Tahun, Jumlah_C, type="o", col="green")
20:1  (Top Level)
```

Gambar. 3.25

```
34
35 plot(Tahun,Jumlah_A, type="o", col="red", ylim=c(40,180))
36 lines(Tahun, Jumlah_B, type="o", pch=22, lty=2, col="blue")
37 lines(Tahun, Jumlah_C, type="o", col="green")
38
39 plot(Tahun,Jumlah_A, type="o", pch=22, lty=2, col="red", ylim=c(40,180))
40 lines(Tahun, Jumlah_B, type="o", pch=22, lty=2, col="blue")
41 lines(Tahun, Jumlah_C, pch=22, lty=2, type="o", col="green")
42
43 plot(Tahun,Jumlah_A, type="p", pch=22, lty=2, col="red", ylim=c(40,180))
44 lines(Tahun, Jumlah_B, type="p", pch=22, lty=2, col="blue")
45 lines(Tahun, Jumlah_C, pch=22, lty=2, type="p", col="green")
46
47 plot(Tahun,Jumlah_A, type="o", pch=22, lty=2, col="red", ylim=c(40,180))
48 lines(Tahun, Jumlah_B, type="p", pch=22, lty=2, col="blue")
49 lines(Tahun, Jumlah_C, pch=22, lty=2, type="l", col="green")
50
51 plot.new()
52 plot(Tahun,Jumlah_A, type="o", col="red", ylim=c(40,180))
53 lines(Tahun, Jumlah_B, type="o", col="blue")
54 lines(Tahun, Jumlah_C, type="o", col="green")
55 title(main="Data Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
56
57 Total = Jumlah_A
58 plot.new()
59 plot(Tahun,Total, type="o", col="red", ylim=c(40,180))
60 lines(Tahun, Jumlah_B, type="o", col="blue")
61 lines(Tahun, Jumlah_C, type="o", col="green")
62 title(main="Data Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
63
```

Gambar 3.26

```
64 Total = Jumlah_A
65 plot.new()
66 plot(Tahun,Total, type="o", col="red", ylim=c(40,180))
67 lines(Tahun, Jumlah_B, type="o", col="blue")
68 lines(Tahun, Jumlah_C, type="o", col="green")
69 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), pch=21)
70 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
71
72 Total = Jumlah_A
73 plot.new()
74 plot(Tahun,Total, type="o", col="red", ylim=c(40,180))
75 lines(Tahun, Jumlah_B, type="o", col="blue")
76 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
77 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
78 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
79
80 Total = Jumlah_A
81 plot.new()
82 plot(Tahun,Total, type="o", col="red", ylim=c(40,180))
83 lines(Tahun, Jumlah_B, type="o", col="blue", lty=23)
84 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
85 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
86 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
87
88 Total = Jumlah_A
89 plot.new()
90 plot(Tahun,Total, type="o", col="red", ylim=c(40,180), lty=23)
91 lines(Tahun, Jumlah_B, type="o", col="blue", lty=23)
92 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
93 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
94 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
95
```

Gambar 3.27

```
95
96 Total = Jumlah_A
97 plot.new()
98 plot(Tahun,Total, type="o", col="red", ylim=c(40,180), lty=23)
99 lines(Tahun, Jumlah_B, type="s", col="blue", lty=23)
100 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
101 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
102 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
103
104 Total = Jumlah_A
105 plot.new()
106 plot(Tahun,Total, type="o", col="red", ylim=c(40,180), lty=23)
107 lines(Tahun, Jumlah_B, type="l", col="blue", lty=23)
108 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
109 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
110 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
111
112 Total = Jumlah_A
113 plot.new()
114 plot(Tahun,Total, type="o", col="red", ylim=c(40,180))
115 lines(Tahun, Jumlah_B, type="l", col="blue", lty=23)
116 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
117 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
118 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
119
120 Total = Jumlah_A
121 plot(Tahun,Total, type="o", col="red", ylim=c(40,180), xaxt="n")
122 Axis(at=2001:2007, side = 1, labels = c("A","B","C","D","E","F","G"))
123 lines(Tahun, Jumlah_B, type="l", col="blue", lty=23)
124 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
125 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
126 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
127
```

Gambar 3.28

```
111
112 Total = Jumlah_A
113 plot.new()
114 plot(Tahun,Total, type="o", col="red", ylim=c(40,180))
115 lines(Tahun, Jumlah_B, type="l", col="blue", lty=23)
116 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
117 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
118 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
119
120 Total = Jumlah_A
121 plot(Tahun,Total, type="o", col="red", ylim=c(40,180), xaxt="n")
122 Axis(at=2001:2007, side = 1, labels = c("A","B","C","D","E","F","G"))
123 lines(Tahun, Jumlah_B, type="l", col="blue", lty=23)
124 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
125 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
126 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
127
128 Total = Jumlah_A
129 plot(Tahun,Total, type="o", col="red", ylim=c(40,180), xaxt="n")
130 Axis(at=2001:2007, side = 1, labels = c("Tahun 1","Tahun 2","Tahun 3","Tahun 4","Tahun 5","Tahun 6","Tahun 7"))
131 lines(Tahun, Jumlah_B, type="l", col="blue", lty=23)
132 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
133 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
134 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
135
136 Total = Jumlah_A
137 plot(Tahun,Total, type="o", col="red", ylim=c(40,180), xaxt="n")
138 Axis(at=2001:2007, side = 3, labels = c("Tahun 1","Tahun 2","Tahun 3","Tahun 4","Tahun 5","Tahun 6","Tahun 7"))
139 lines(Tahun, Jumlah_B, type="l", col="blue", lty=23)
140 lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
141 legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"), lty=30)
142 title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
143
```

Gambar 3.29

```
simpan=read.table("data3.2.csv",header=TRUE, sep=",") #membaca data
simpan
```

```
##   tahun jenis.barang.A jenis.barang.B jenis.barang.C
## 1  2001             90             85             50
## 2  2002            110             90             55
## 3  2003            115            105             60
## 4  2004            130            110             65
## 5  2005            140            120             75
## 6  2006            155            125             80
## 7  2007            160            130             85
```

```
Tahun=simpan$tahun
Jumlah_A=simpan$jenis.barang.A
Jumlah_B=simpan$jenis.barang.B
Jumlah_C=simpan$jenis.barang.C

Jumlah_A
```

```
## [1]  90 110 115 130 140 155 160
```

```
Jumlah_B
```

```
## [1]  85  90 105 110 120 125 130
```

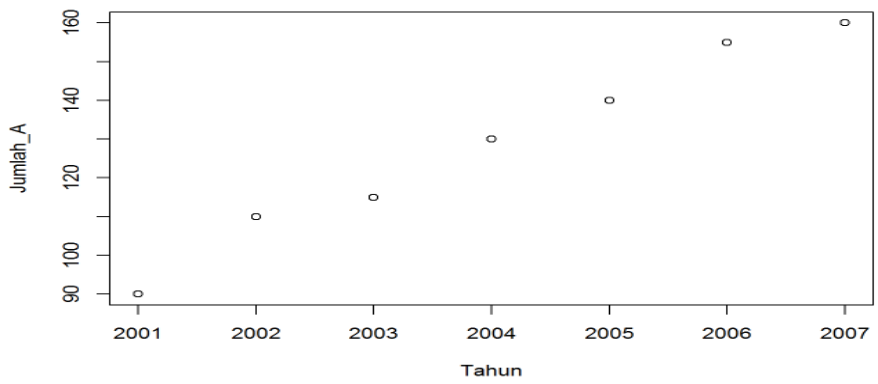
```
Jumlah_C
```

```
## [1] 50 55 60 65 75 80 85
```

Gambar 3.30 merupakan hasil eksekusi kode R pada baris 1 sampai dengan baris 11.

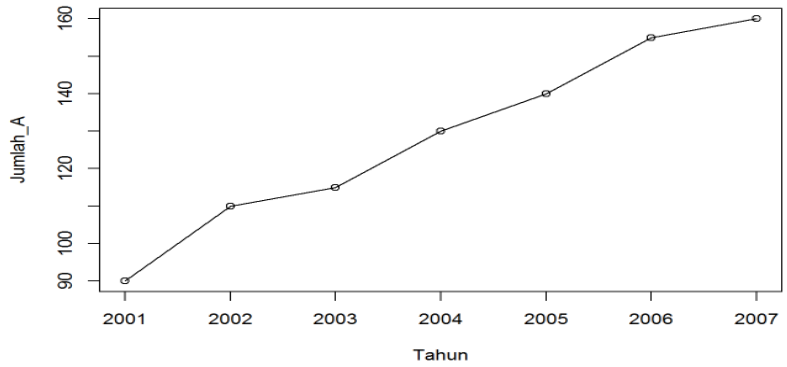
Gambar 3.30

```
plot(Tahun,Jumlah_A)
```



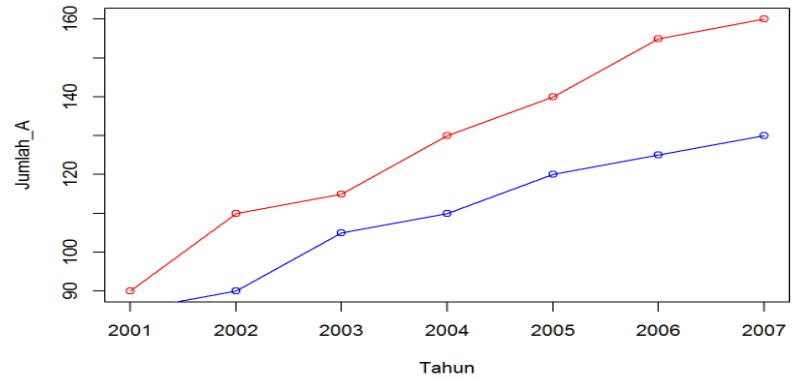
Gambar 3.31

```
plot(Tahun,Jumlah_A, type="o")
```



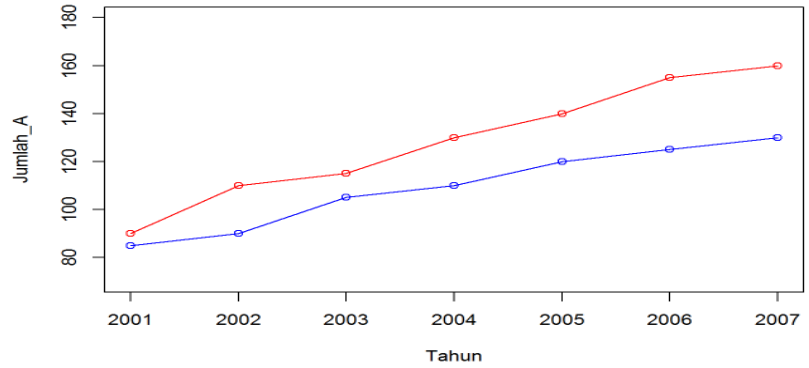
Gambar 3.32

```
plot(Tahun,Jumlah_A, type="o", col="red")
lines(Tahun, Jumlah_B, type="o", col="blue")
```



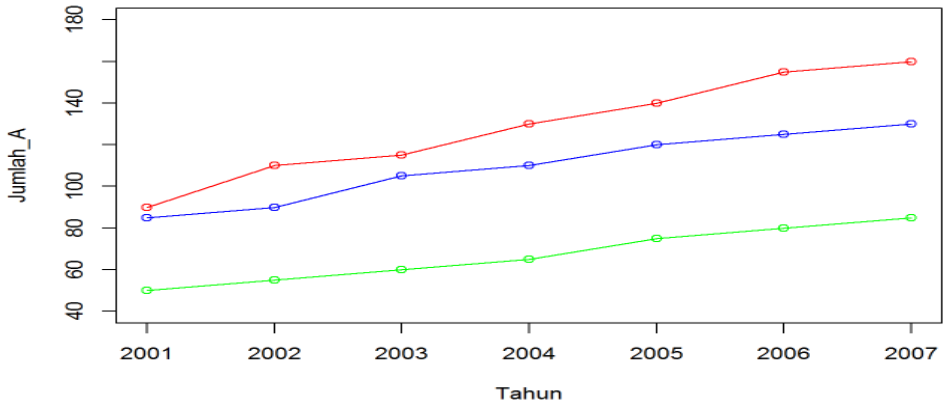
Gambar 3.33

```
plot(Tahun,Jumlah_A, type="o", col="red", ylim=c(70,180))
lines(Tahun, Jumlah_B, type="o", col="blue")
```



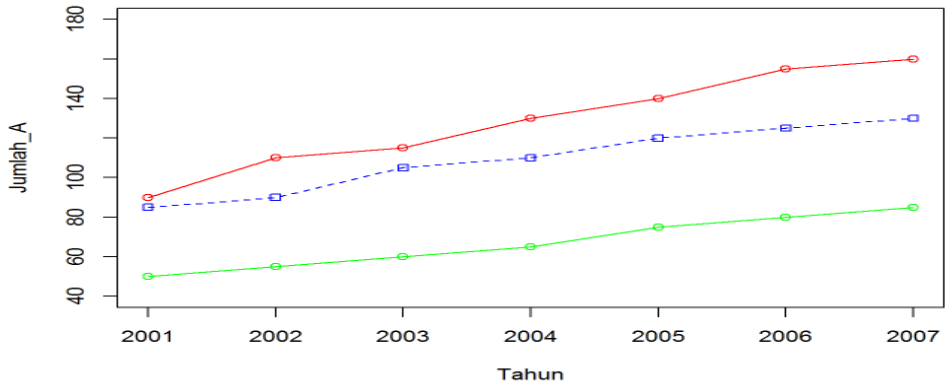
Gambar 3.34

```
plot(Tahun,Jumlah_A, type="o", col="red", ylim=c(40,180))
lines(Tahun, Jumlah_B, type="o", col="blue")
lines(Tahun, Jumlah_C, type="o", col="green")
```



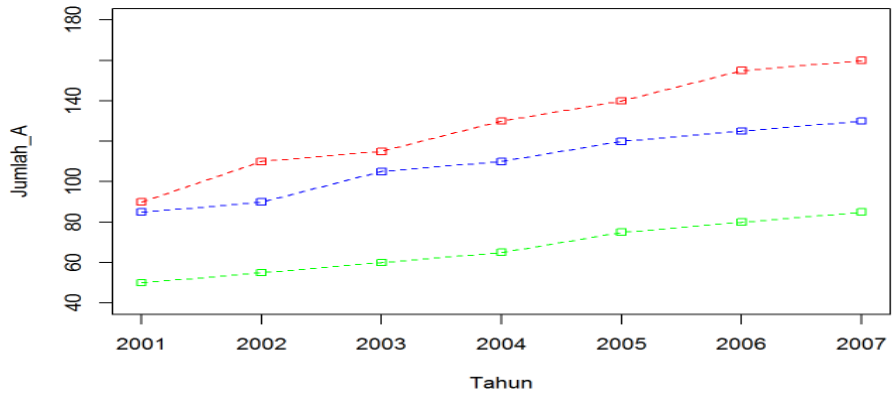
Gambar 3.35

```
plot(Tahun,Jumlah_A, type="o", col="red", ylim=c(40,180))
lines(Tahun, Jumlah_B, type="o", pch=22, lty=2, col="blue")
lines(Tahun, Jumlah_C, type="o", col="green")
```



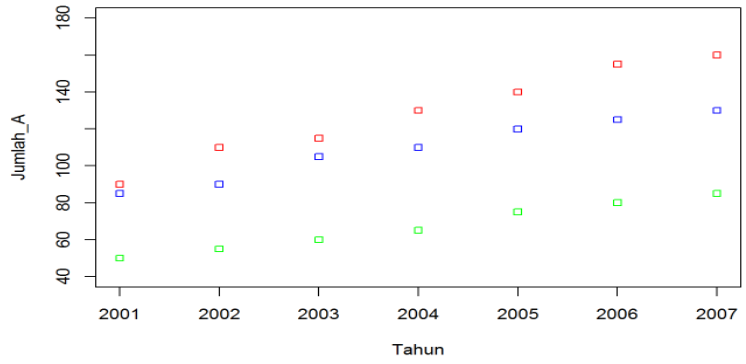
Gambar 3.36

```
plot(Tahun,Jumlah_A, type="o", pch=22, lty=2, col="red", ylim=c(40,180))
lines(Tahun, Jumlah_B, type="o", pch=22, lty=2, col="blue")
lines(Tahun, Jumlah_C, pch=22, lty=2, type="o", col="green")
```



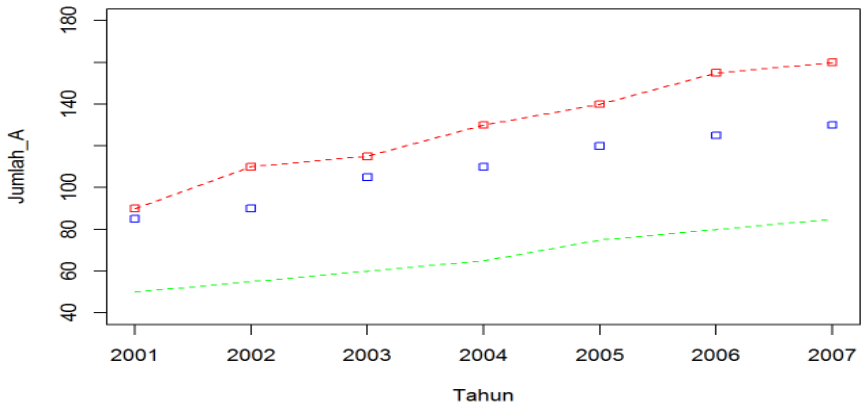
Gambar 3.37

```
plot(Tahun,Jumlah_A, type="p", pch=22, lty=2, col="red", ylim=c(40,180))
lines(Tahun, Jumlah_B, type="p", pch=22, lty=2, col="blue")
lines(Tahun, Jumlah_C, pch=22, lty=2, type="p", col="green")
```



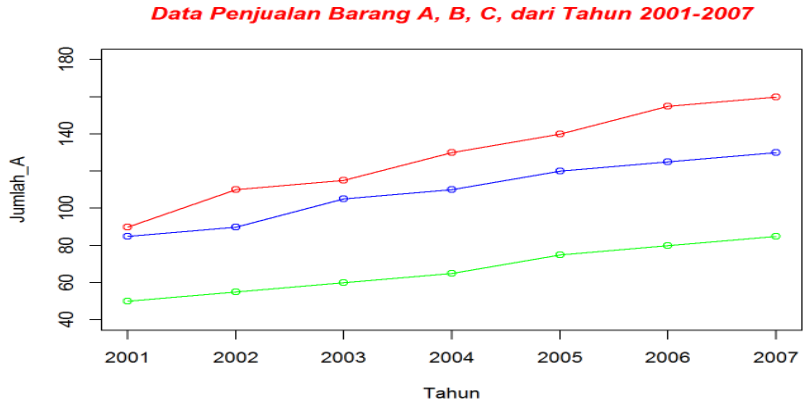
Gambar 3.38

```
plot(Tahun,Jumlah_A, type="o", pch=22, lty=2, col="red", ylim=c(40,180))
lines(Tahun, Jumlah_B, type="p", pch=22, lty=2, col="blue")
lines(Tahun, Jumlah_C, pch=22, lty=2, type="l", col="green")
```



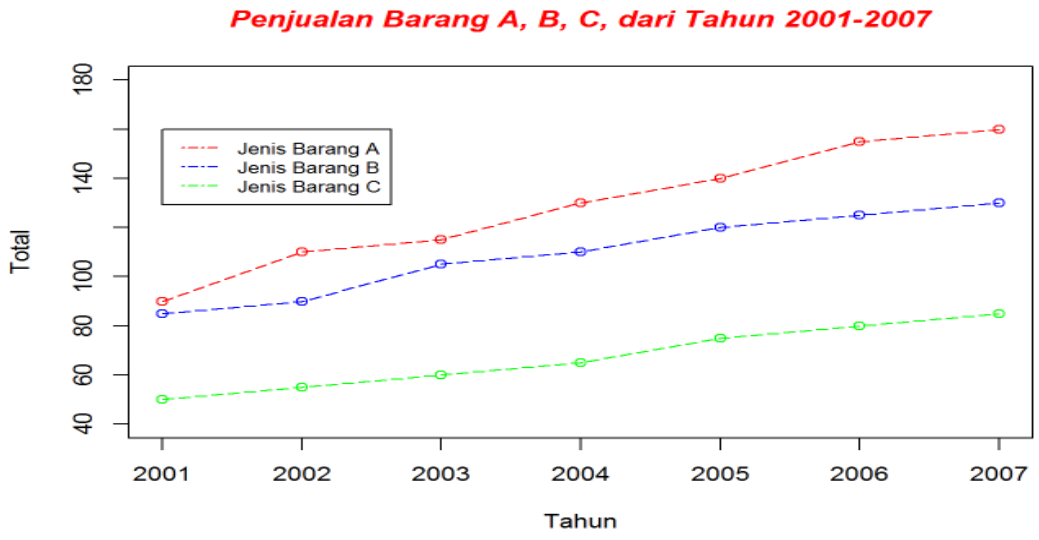
Gambar 3.39

```
plot.new()
plot(Tahun,Jumlah_A, type="o", col="red", ylim=c(40,180))
lines(Tahun, Jumlah_B, type="o", col="blue")
lines(Tahun, Jumlah_C, type="o", col="green")
title(main="Data Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
```



Gambar 3.40

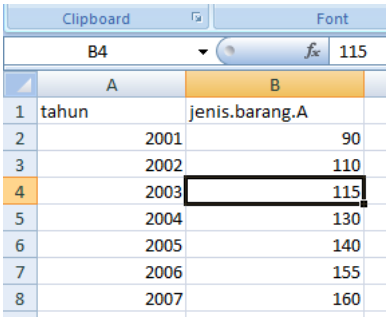
```
Total = Jumlah_A
plot.new()
plot(Tahun,Total, type="o", col="red", ylim=c(40,180), lty=23)
lines(Tahun, Jumlah_B, type="o", col="blue", lty=23)
lines(Tahun, Jumlah_C, type="o", col="green", lty=23)
legend(2001,160,c("Jenis Barang A", "Jenis Barang B", "Jenis Barang C"), cex=0.8, col=c("red","blue","green"),
lty=30)
title(main="Penjualan Barang A, B, C, dari Tahun 2001-2007", col.main="red", font.main=4)
```



Gambar 3.41

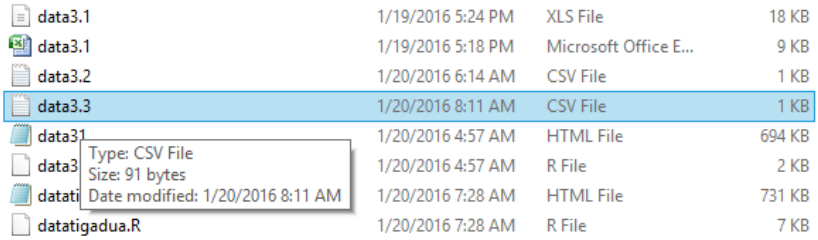
Menyajikan Data dengan Grafik Batang (Bagian Pertama)

Misalkan diberikan data seperti pada Gambar 3.42. Gambar 3.42 menyajikan hasil penjualan barang A, selama kurun waktu 2001-2007. Data pada Gambar 3.42 disimpan terlebih dahulu dengan nama **data3.3.csv** (perhatikan Gambar 3.43).



	A	B
1	tahun	jenis.barang.A
2	2001	90
3	2002	110
4	2003	115
5	2004	130
6	2005	140
7	2006	155
8	2007	160

Gambar 3.42



Gambar 3.43

Gambar 3.44 merupakan kode R. Eksekusi dan amati hasilnya.

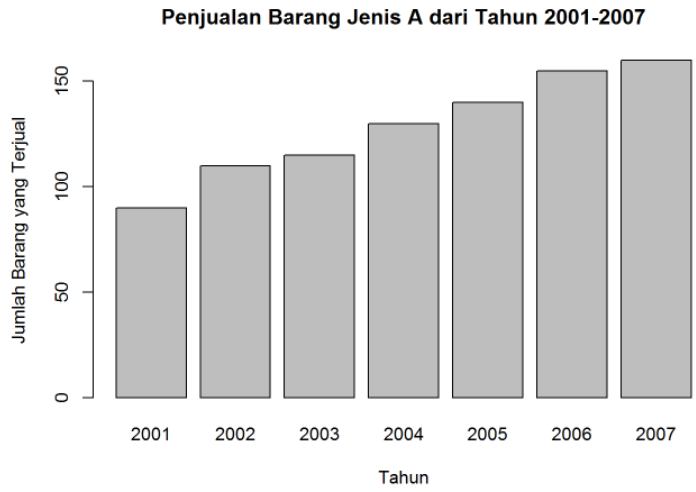
```r
simpan=read.table("data3.3.csv",header=TRUE, sep=",") #membaca data
simpan

Tahun=simpan$tahun
Jumlah_A=simpan$jenis.barang.A
barplot(Jumlah_A,Tahun)

barplot(Jumlah_A,Tahun, main="Penjualan Barang Jenis A dari Tahun 2001-2007", xlab="Tahun",
        ylab="Jumlah Barang yang Terjual", names.arg=c("2001","2002","2003","2004","2005","2006","2007"))

barplot(Jumlah_A,Tahun, main="Penjualan Barang Jenis A dari Tahun 2001-2007", xlab="Tahun",
        ylab="Jumlah Barang yang Terjual", names.arg=c("2001","2002","2003","2004","2005","2006","2007"))

barplot(Jumlah_A,Tahun, main="Penjualan Barang Jenis A dari Tahun 2001-2007", xlab="Tahun",
        ylab="Jumlah Barang yang Terjual", names.arg=c("2001","2002","2003","2004","2005","2006","2007"), border="blue")

barplot(Jumlah_A,Tahun, main="Penjualan Barang Jenis A dari Tahun 2001-2007", xlab="Tahun",
        ylab="Jumlah Barang yang Terjual", names.arg=c("2001","2002","2003","2004","2005","2006","2007"), border="red")

barplot(Jumlah_A,Tahun, main="Penjualan Barang Jenis A dari Tahun 2001-2007", xlab="Tahun",
        ylab="Jumlah Barang yang Terjual", names.arg=c("2001","2002","2003","2004","2005","2006","2007"),
        border="green",density=c(10,20,30,40,50,60,70) )


library(ggplot2)
ggplot(data=simpan, aes(x=Tahun, y=Jumlah_A)) + geom_bar(stat="identity")

ggplot(data=simpan, aes(x=Tahun, y=Jumlah_A)) + geom_bar(stat="identity", fill="darkblue")

ggplot(data=simpan, aes(x=Tahun, y=Jumlah_A)) + geom_bar(stat="identity", fill=heat.colors(7))
```

Gambar 3.44

```r
simpan=read.table("data3.3.csv",header=TRUE, sep=",") #membaca data
simpan
```

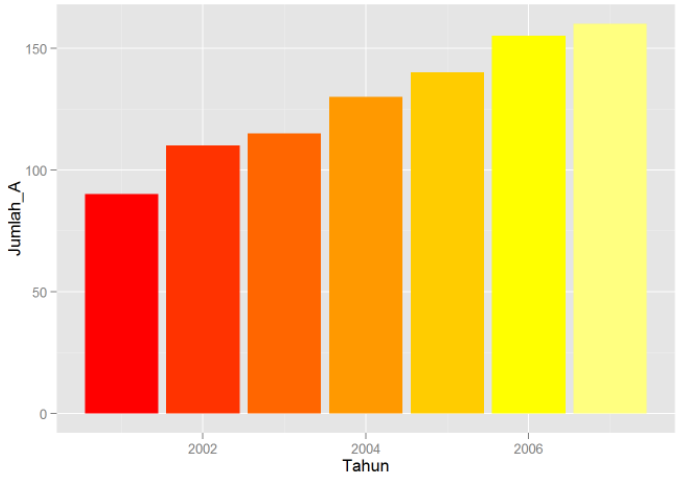
```
##   tahun jenis.barang.A
## 1  2001             90
## 2  2002            110
## 3  2003            115
## 4  2004            130
## 5  2005            140
## 6  2006            155
## 7  2007            160
```

Gambar 3.45



Gambar 3.46

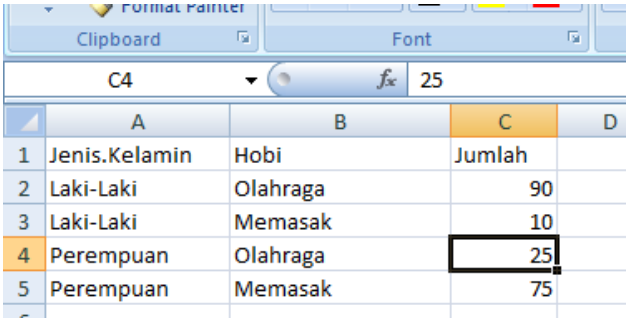
```
ggplot(data=simpan, aes(x=Tahun, y=Jumlah_A)) + geom_bar(stat="identity", fill=heat.colors(7))
```



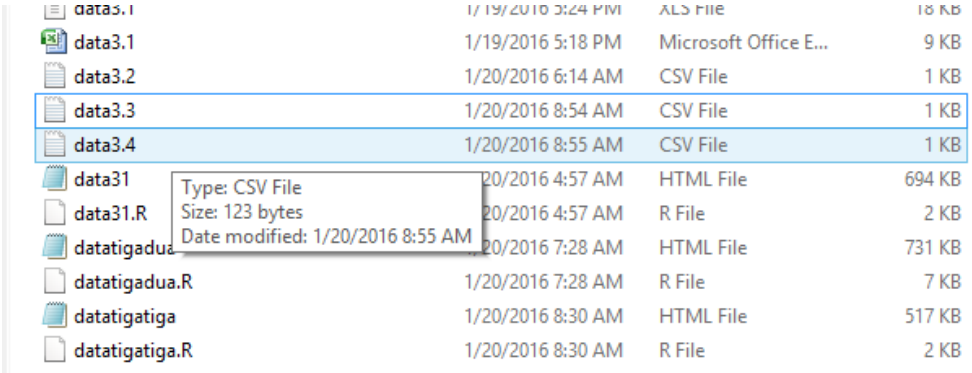
Gambar 3.47

Menyajikan Data dengan Grafik Batang (Bagian Kedua)

Misalkan diberikan data seperti pada Gambar 3.48. Berdasarkan data pada Gambar 3.48, diketahui responden laki-laki yang memiliki hobi olahraga sebanyak 90 responden, responden laki-laki yang memiliki hobi memasak sebanyak 10 responden, dan seterusnya. Data pada Gambar 3.48 disimpan terlebih dahulu dengan nama **data3.4.csv** (perhatikan Gambar 3.49).



Gambar 3.48

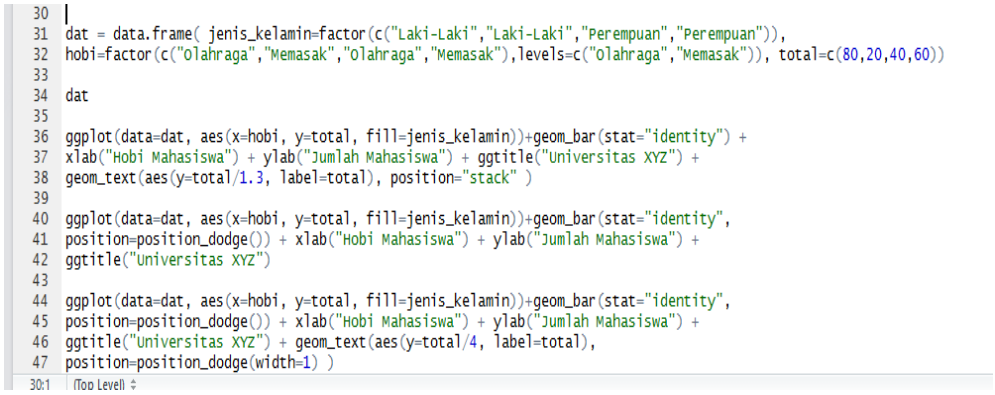


Gambar 3.49

Gambar 3.50 dan Gambar 3.51 merupakan kode R. Eksekusi kode R tersebut dan amati hasilnya.

```
simpan=read.table("data3.4.csv",header=TRUE, sep=",") #membaca data
simpan

frekuensi=c(90,10,25,75)
barplot(t(matrix(frekuensi, ncol=2, byrow=TRUE, dimnames=list(c("Laki-Laki","Perempuan"), c("Olahraga","Memasak")))),
main="Hubungan antara Jenis Kelamin dan Hobi", xlab="Jenis Kelamin",
col=c("darkblue","orange"), beside=TRUE, ylim=c(0,150), legend.text=TRUE,
args.legend=list(x="topright"))


frekuensi2=c(2,12,16,6)
barplot(frekuensi2, ylim=c(0,20), main="Jumlah Mahasiswa yang Memperoleh Nilai A, B, C, dan D, untuk
Matakuliah Matematika 1", names.arg=c("A","B","C","D"), ylab="Jumlah Mahasiswa",
xlab="Nilai Mahasiswa", cex.names=0.8, col=c("green","yellow","orange","red") )

dat = data.frame(
jenis_kelamin=factor(c("Laki-Laki","Perempuan"), levels=c("Laki-Laki","Perempuan")), total=c(20,70))

dat

library(ggplot2)
ggplot(data=dat, aes(x=jenis_kelamin, y=total))+geom_bar(stat="identity")

ggplot(data=dat, aes(x=jenis_kelamin, y=total, fill=jenis_kelamin))+geom_bar(stat="identity")

ggplot(data=dat, aes(x=jenis_kelamin, y=total, fill=jenis_kelamin))+geom_bar(stat="identity") + guides(fill=FALSE)

ggplot(data=dat, aes(x=jenis_kelamin, y=total, fill=jenis_kelamin))+geom_bar(stat="identity") +
xlab("Jenis Kelamin") + ylab("Jumlah Mahasiswa") + ggtitle("Universitas XYZ")
```

Gambar 3.50



```
dat = data.frame( jenis_kelamin=factor(c("Laki-Laki","Laki-Laki","Perempuan","Perempuan")),
hobi=factor(c("Olahraga","Memasak","Olahraga","Memasak"),levels=c("Olahraga","Memasak")), total=c(80,20,40,60))

dat

ggplot(data=dat, aes(x=hobi, y=total, fill=jenis_kelamin))+geom_bar(stat="identity") +
xlab("Hobi Mahasiswa") + ylab("Jumlah Mahasiswa") + ggtitle("Universitas XYZ") +
geom_text(aes(y=total/1.3, label=total), position="stack" )

ggplot(data=dat, aes(x=hobi, y=total, fill=jenis_kelamin))+geom_bar(stat="identity",
position=position_dodge()) + xlab("Hobi Mahasiswa") + ylab("Jumlah Mahasiswa") +
ggtitle("Universitas XYZ")

ggplot(data=dat, aes(x=hobi, y=total, fill=jenis_kelamin))+geom_bar(stat="identity",
position=position_dodge()) + xlab("Hobi Mahasiswa") + ylab("Jumlah Mahasiswa") +
ggtitle("Universitas XYZ") + geom_text(aes(y=total/4, label=total),
position=position_dodge(width=1) )
```

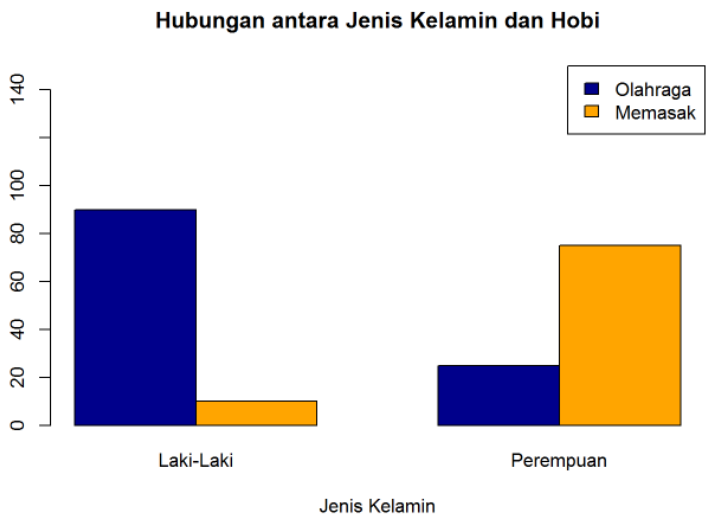
Gambar 3.51

```
simpan=read.table("data3.4.csv",header=TRUE, sep=",") #membaca data
simpan
```

```
##   Jenis.Kelamin     Hobi Jumlah
## 1     Laki-Laki Olahraga     90
## 2     Laki-Laki  Memasak     10
## 3     Perempuan Olahraga     25
## 4     Perempuan  Memasak     75
```

Gambar 3.52

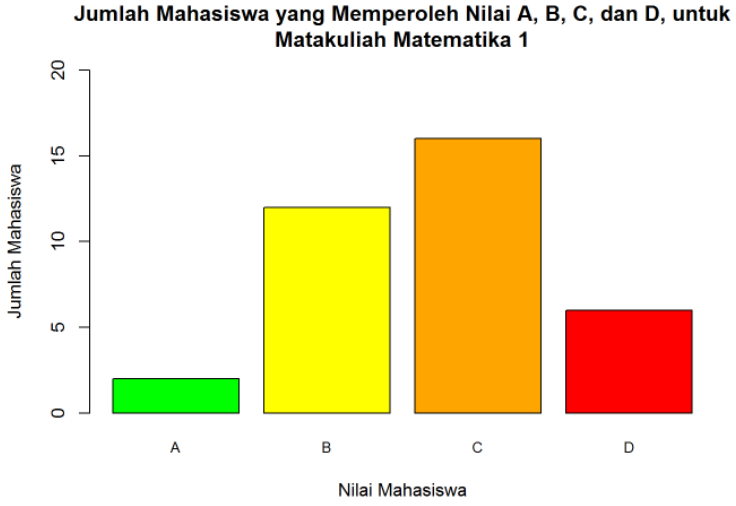
```
frekuensi=c(90,10,25,75)
barplot(t(matrix(frekuensi, ncol=2, byrow=TRUE, dimnames=list(c("Laki-Laki","Perempuan"), c("Olahraga","Memasak
"))))),
main="Hubungan antara Jenis Kelamin dan Hobi", xlab="Jenis Kelamin",
col=c("darkblue","orange"), beside=TRUE, ylim=c(0,150), legend.text=TRUE,
args.legend=list(x="topright"))
```



Gambar 3.53

```
frekuensi2=c(2,12,16,6)
barplot(frekuensi2, ylim=c(0,20), main="Jumlah Mahasiswa yang Memperoleh Nilai A, B, C, dan D, untuk
Matakuliah Matematika 1", names.arg=c("A","B","C","D"), ylab="Jumlah Mahasiswa",
xlab="Nilai Mahasiswa", cex.names=0.8, col=c("green","yellow","orange","red") )

dat = data.frame(
jenis_kelamin=factor(c("Laki-Laki","Perempuan"), levels=c("Laki-Laki","Perempuan")), total=c(20,70))

dat
```

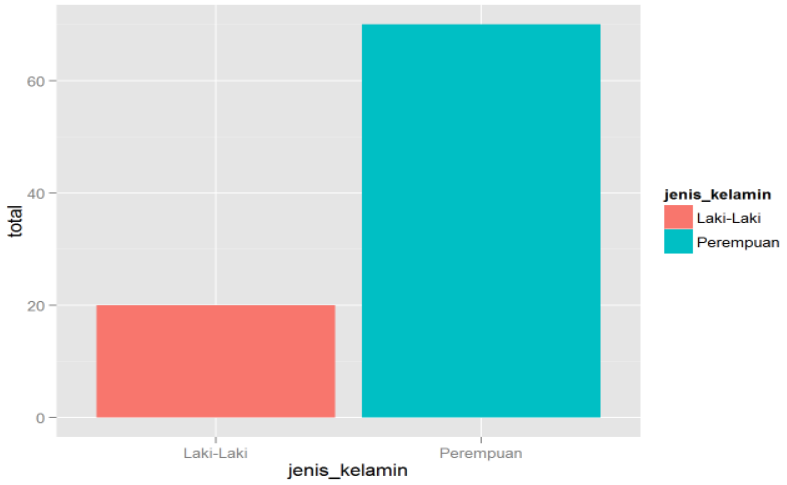
```
##   jenis_kelamin total
## 1     Laki-Laki    20
## 2     Perempuan    70
```



Gambar 3.54

```
library(ggplot2)
```

```
ggplot(data=dat, aes(x=jenis_kelamin, y=total, fill=jenis_kelamin))+geom_bar(stat="identity")
```



Gambar 3.55

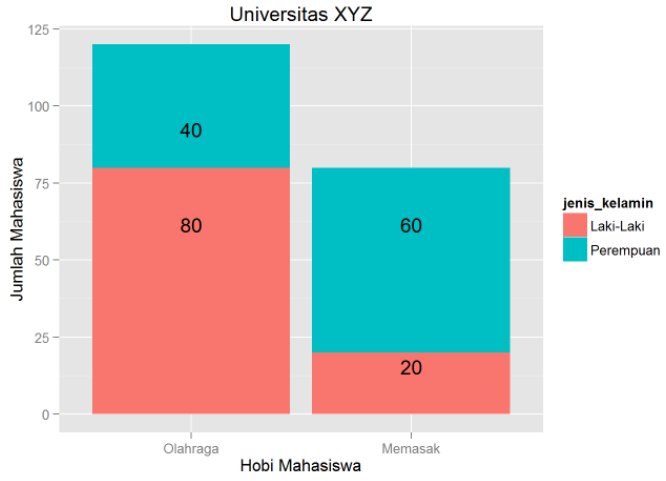
```
dat = data.frame( jenis_kelamin=factor(c("Laki-Laki","Laki-Laki","Perempuan","Perempuan")),
hobi=factor(c("Olahraga","Memasak","Olahraga","Memasak"),levels=c("Olahraga","Memasak")), total=c(80,20,40,60))

dat
```

```
##   jenis_kelamin     hobi total
## 1     Laki-Laki Olahraga    80
## 2     Laki-Laki  Memasak    20
## 3     Perempuan Olahraga    40
## 4     Perempuan  Memasak    60
```

Gambar 3.56

```
ggplot(data=dat, aes(x=hobi, y=total, fill=jenis_kelamin))+geom_bar(stat="identity") +
xlab("Hobi Mahasiswa") + ylab("Jumlah Mahasiswa") + ggtitle("Universitas XYZ") +
geom_text(aes(y=total/1.3, label=total), position="stack" )
```

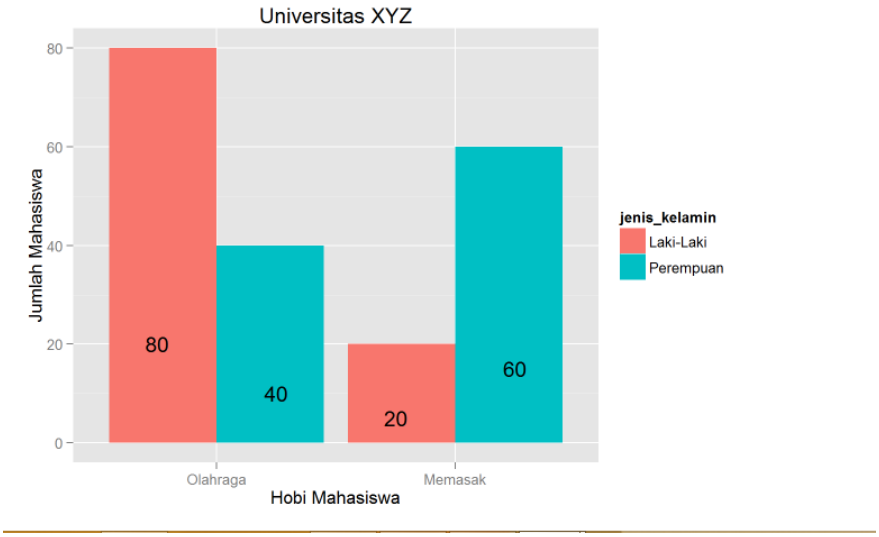
```
## ymax not defined: adjusting position using y instead
```



Gambar 3.57

```
ggplot(data=dat, aes(x=hobi, y=total, fill=jenis_kelamin))+geom_bar(stat="identity",
position=position_dodge()) + xlab("Hobi Mahasiswa") + ylab("Jumlah Mahasiswa") +
ggtitle("Universitas XYZ") + geom_text(aes(y=total/4, label=total),
position=position_dodge(width=1) )
```

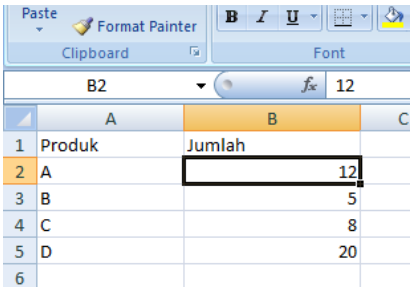
```
## ymax not defined: adjusting position using y instead
```



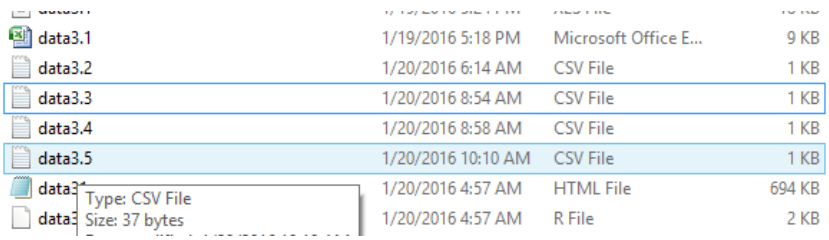
Gambar 3.58

Menyajikan Data dengan Diagram Lingkaran

Misalkan diberikan data seperti pada Gambar 3.59. Berdasarkan Gambar data pada 3.59, diketahui jumlah produk A yang terjual sebanyak 12 unit, jumlah produk B yang terjual sebanyak 5 unit, dan seterusnya. Data pada Gambar 3.59 disimpan terlebih dahulu dengan nama **data3.5.csv** (perhatikan Gambar 3.60).



Gambar 3.59



Gambar 3.60

Gambar 3.61 dan Gambar 3.62 merupakan kode R. Eksekusi kode R tersebut, dan amati hasilnya.

```r
simpan=read.table("data3.5.csv",header=TRUE, sep=",") #membaca data
simpan

pie(simpan$Jumlah,labels=simpan$Produk, main="Data Penjualan Produk A, B, C, dan D")

pie(simpan$Jumlah,labels=simpan$Produk, main="Data Penjualan Produk A, B, C, dan D", col=heat.colors(4) )

pie(simpan$Jumlah,labels=simpan$Jumlah, main="Data Penjualan Produk A, B, C, dan D", col=heat.colors(4))
colors=heat.colors(4)
legend(1,0.5, c("Produk A","Produk B","Produk C", "Produk D"), cex=0.8, fill=colors )

pie(simpan$Jumlah,labels=simpan$Jumlah, main="Data Penjualan Produk A, B, C, dan D",
col=c("darkblue","orange","yellow","red"))
colors=c("darkblue","orange","yellow","red")
legend(1,0.5, c("Produk A","Produk B","Produk C", "Produk D"), cex=0.8, fill=colors )

Persen=round(simpan$Jumlah/sum(simpan$Jumlah)*100,4)
Persen=paste(Persen,"%",sep="")
pie(simpan$Jumlah,labels=Persen, main="Data Penjualan Produk A, B, C, dan D",
col=c("darkblue","orange","yellow","red"))
colors=c("darkblue","orange","yellow","red")
legend(1,0.5, c("Produk A","Produk B","Produk C", "Produk D"), cex=0.8, fill=colors )

Persen=round(simpan$Jumlah/sum(simpan$Jumlah)*100,4)
Persen=paste(Persen,"%",sep="")
pie(simpan$Jumlah,labels=Persen, main="Data Penjualan Produk A, B, C, dan D",
col=c("darkblue","orange","yellow","red"))
colors=c("darkblue","orange","yellow","red")
legend(1,0.5, c("Produk A","Produk B","Produk C", "Produk D"), cex=0.8, fill=colors )
```

Gambar 3.61

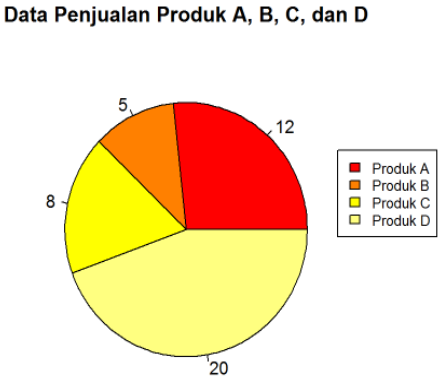
```r
Jumlah=simpan$Jumlah
Produk=simpan$Produk
library(ggplot2)
pie = ggplot(simpan, aes(x="", y=Jumlah, fill=Produk))+geom_bar(width=1,stat="identity")+coord_polar("y",start=0)

pie

library(ggplot2)
library(grid)
library(gridExtra)

blank_theme = theme(
axis.title.x=element_blank(),
axis.title.y=element_blank(),
axis.text.x = element_blank(),
axis.text.y = element_blank(),
panel.border = element_blank(),
panel.grid=element_blank(),
axis.ticks= element_blank(),
plot.title=element_text(size=14, face="bold")
)

library(scales)
pie + blank_theme + geom_text(aes(y=Jumlah/4 + c(0,cumsum(Jumlah)[-length(Jumlah)]), label=Jumlah), size=5)
pie + blank_theme + geom_text(aes(y=Jumlah/4 + c(0,cumsum(Jumlah)[-length(Jumlah)]), label=Jumlah), size=5) +
scale_fill_manual(values=c(heat.colors(4)))

Persen=round(simpan$Jumlah/sum(simpan$Jumlah)*100,2)
Persen=paste(Persen,"%",sep="")
pie + blank_theme + geom_text(aes(y=Jumlah/4 + c(0,cumsum(Jumlah)[-length(Jumlah)]),
label=Persen ), size=5) +
scale_fill_manual(values=c(heat.colors(4)))
```

Gambar 3.62

```r
simpan=read.table("data3.5.csv",header=TRUE, sep=",") #membaca data
simpan
```

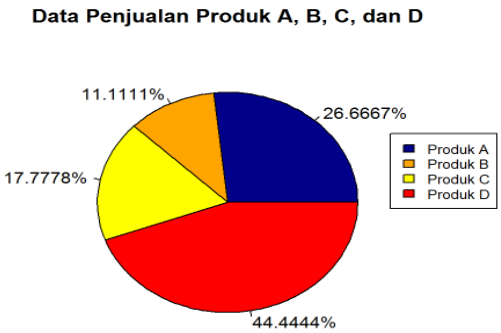
```
##   Produk Jumlah
## 1      A     12
## 2      B      5
## 3      C      8
## 4      D     20
```

```
pie(simpan$Jumlah,labels=simpan$Jumlah, main="Data Penjualan Produk A, B, C, dan D", col=heat.colors(4))
colors=heat.colors(4)
legend(1,0.5, c("Produk A","Produk B","Produk C", "Produk D"), cex=0.8, fill=colors )
```

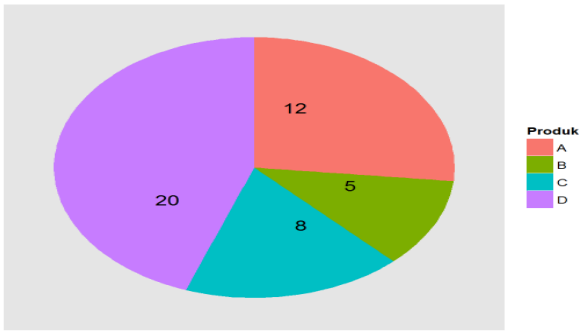


Gambar 3.63

```
Persen=round(simpan$Jumlah/sum(simpan$Jumlah)*100,4)
Persen=paste(Persen,"%",sep="")
pie(simpan$Jumlah,labels=Persen, main="Data Penjualan Produk A, B, C, dan D",
col=c("darkblue","orange","yellow","red"))
colors=c("darkblue","orange","yellow","red")
legend(1,0.5, c("Produk A","Produk B","Produk C", "Produk D"), cex=0.8, fill=colors )
```

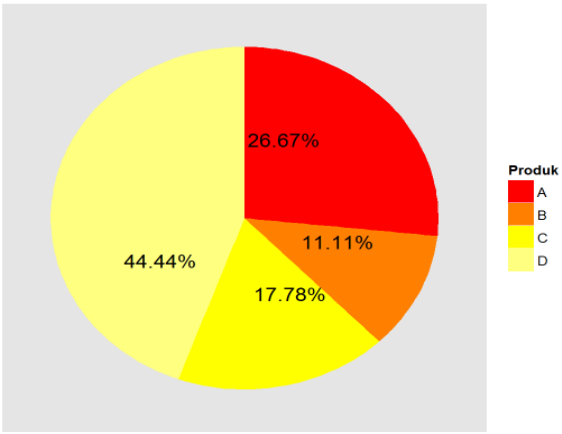


Gambar 3.64



Gambar 3.65

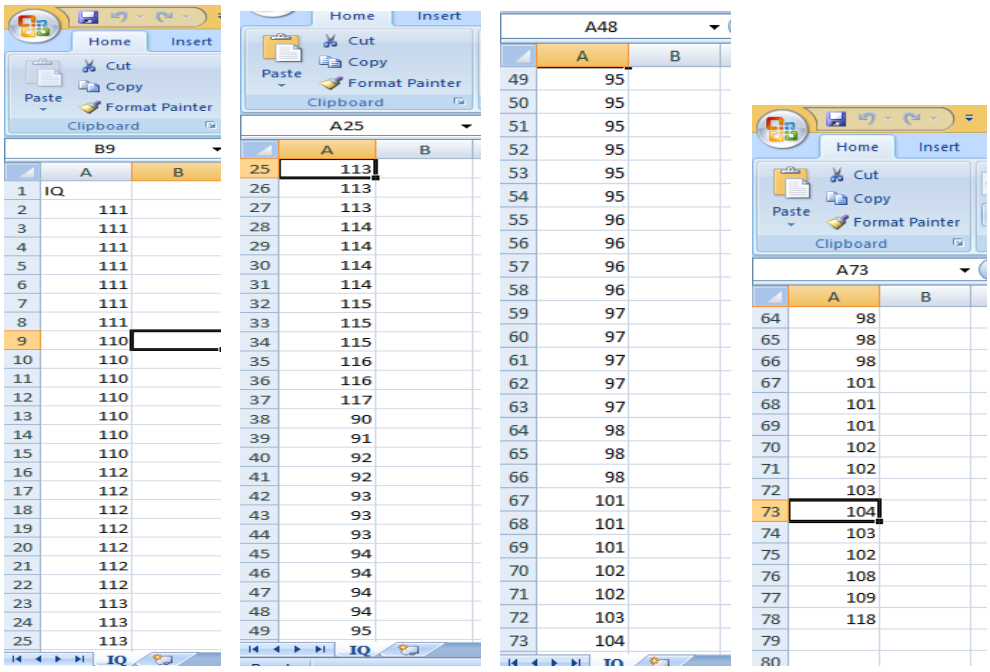
```
Persen=round(simpan$Jumlah/sum(simpan$Jumlah)*100,2)
Persen=paste(Persen,"%",sep="")
pie + blank_theme + geom_text(aes(y=Jumlah/4 + c(0,cumsum(Jumlah)[-length(Jumlah)]),
label=Persen ), size=5) +
scale_fill_manual(values=c(heat.colors(4)))
```



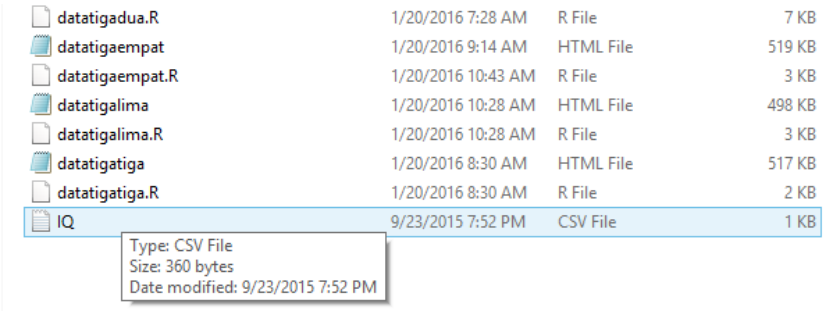
Gambar 3.66

Menyajikan Data dengan Histogram

Misalkan diberikan data mengenai skor IQ seperti pada Gambar 3.67. Berdasarkan data pada Gambar 3.67, jumlah pengamatan sebanyak 77. Data pada Gambar 3.67 disimpan terlebih dahulu dengan nama **IQ.csv** (perhatikan Gambar 3.68). Gambar 3.69 dan Gambar 3.70 disajikan kode R. Eksekusi kode R tersebut, dan amati hasilnya.



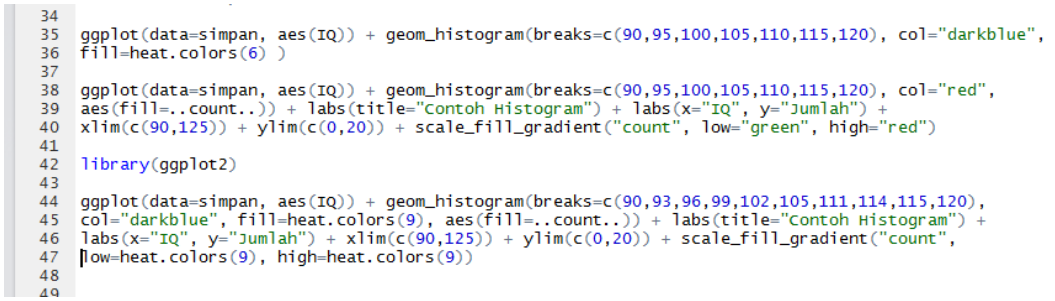
Gambar 3.67



Gambar 3.68

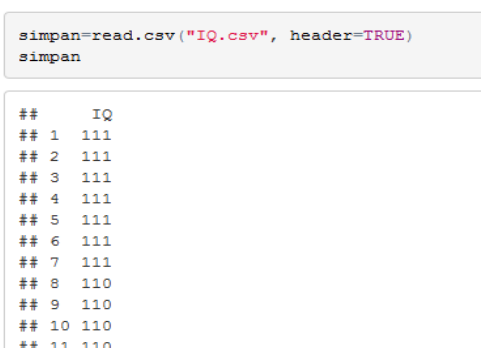
```
simpan=read.csv("IQ.csv", header=TRUE)
simpan

simpan_skor_IQ=simpan$IQ
hist(simpan_skor_IQ)

hist(simpan_skor_IQ, col="lightblue")

hist(simpan_skor_IQ, col="darkblue", ylim=c(0,40), main="Contoh Histogram", ylab="Frekuensi")

hist(simpan_skor_IQ, col="orange", ylim=c(0,40), main="Contoh Histogram", ylab="Frekuensi", breaks=c(90,100,110,120) )

hist(simpan_skor_IQ, col=heat.colors(6), ylim=c(0,30), main="Contoh Histogram", ylab="Frekuensi",
breaks=c(90,95,100,105,110,115,120), xlim=c(90,125) )

hist(simpan_skor_IQ, col=heat.colors(6), ylim=c(0,30), main="Contoh Histogram", ylab="Frekuensi",
breaks=c(90,93,96,99,102,105,108,111,114,117,120), xlim=c(90,125) )

hist(simpan_skor_IQ, breaks=6 , col=heat.colors(6), ylim=c(0,30), main="Contoh Histogram",
ylab="Frekuensi", xlim=c(90,125) )

hist(simpan_skor_IQ, breaks=c(90,117,120), ylim=c(0,50), xlim=c(90,125), main="Contoh Histogram",
col=heat.colors(2) )

hist(simpan_skor_IQ, breaks=c(90,117,120), ylim=c(0,80), xlim=c(90,125), main="Contoh Histogram",
col=heat.colors(2), freq=TRUE )

hist(simpan_skor_IQ, breaks=c(90,92,97,117,120), ylim=c(0,80), xlim=c(90,125), main="Contoh Histogram",
 col=heat.colors(4), freq=TRUE )
```

Gambar 3.69



```

ggplot(data=simpan, aes(IQ)) + geom_histogram(breaks=c(90,95,100,105,110,115,120), col="darkblue",
fill=heat.colors(6) )

ggplot(data=simpan, aes(IQ)) + geom_histogram(breaks=c(90,95,100,105,110,115,120), col="red",
aes(fill=..count..)) + labs(title="Contoh Histogram") + labs(x="IQ", y="Jumlah") +
xlim(c(90,125)) + ylim(c(0,20)) + scale_fill_gradient("count", low="green", high="red")

library(ggplot2)

ggplot(data=simpan, aes(IQ)) + geom_histogram(breaks=c(90,93,96,99,102,105,111,114,115,120),
col="darkblue", fill=heat.colors(9), aes(fill=..count..)) + labs(title="Contoh Histogram") +
labs(x="IQ", y="Jumlah") + xlim(c(90,125)) + ylim(c(0,20)) + scale_fill_gradient("count",
low=heat.colors(9), high=heat.colors(9))

```

Gambar 3.70

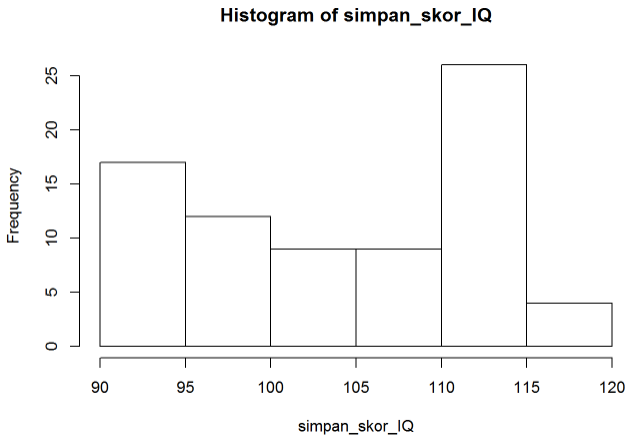


```
simpan=read.csv("IQ.csv", header=TRUE)
simpan
```

```
##     IQ
## 1  111
## 2  111
## 3  111
## 4  111
## 5  111
## 6  111
## 7  111
## 8  110
## 9  110
## 10 110
```

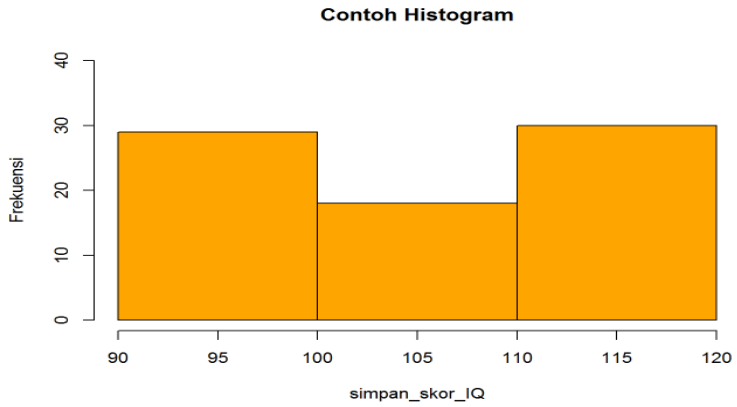
Gambar 3.71

```
simpan_skor_IQ=simpan$IQ
hist(simpan_skor_IQ)
```



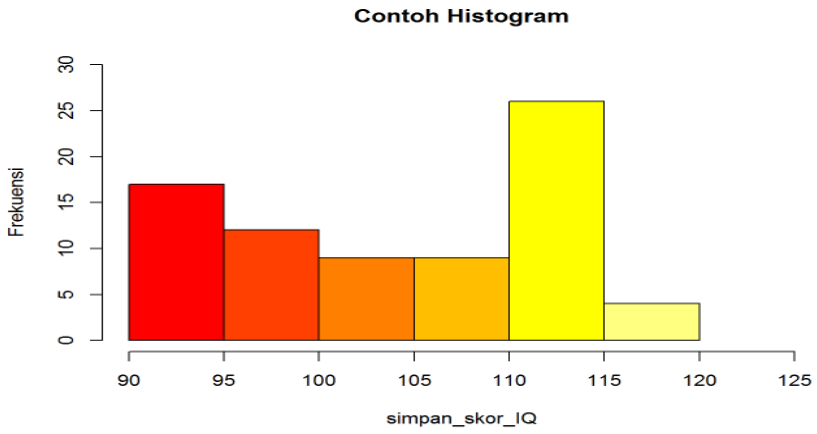
Gambar 3.72

```
hist(simpan_skor_IQ, col="orange", ylim=c(0,40), main="Contoh Histogram", ylab="Frekuensi", breaks=c(90,100,110
,120) )
```



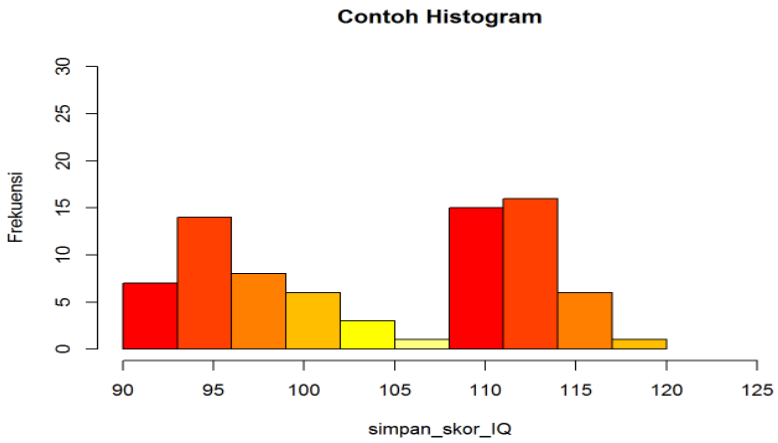
Gambar 3.73

```
hist(simpan_skor_IQ, col=heat.colors(6), ylim=c(0,30), main="Contoh Histogram", ylab="Frekuensi",
breaks=c(90,95,100,105,110,115,120), xlim=c(90,125) )
```



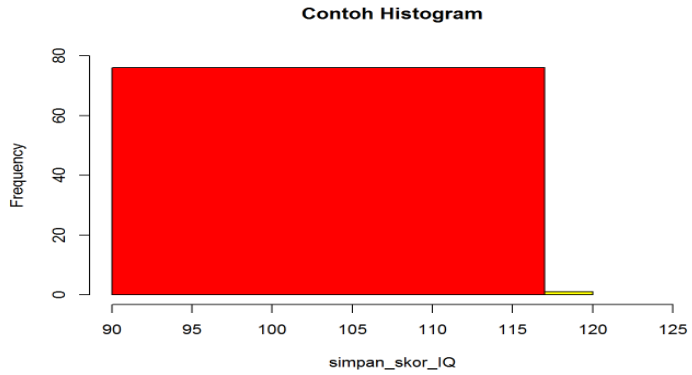
Gambar 3.74

```
hist(simpan_skor_IQ, col=heat.colors(6), ylim=c(0,30), main="Contoh Histogram", ylab="Frekuensi",
breaks=c(90,93,96,99,102,105,108,111,114,117,120), xlim=c(90,125) )
```



Gambar 3.75

```
hist(simpan_skor_IQ, breaks=c(90,117,120), ylim=c(0,80), xlim=c(90,125), main="Contoh Histogram",
col=heat.colors(2), freq=TRUE )
```

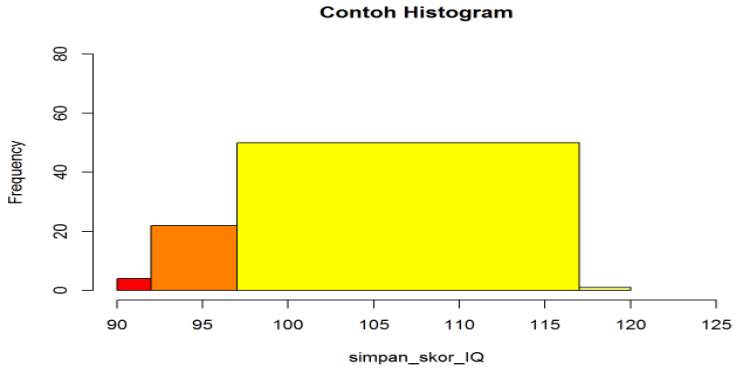


Gambar 3.76

```
hist(simpan_skor_IQ, breaks=c(90,92,97,117,120), ylim=c(0,80), xlim=c(90,125), main="Contoh Histogram",
 col=heat.colors(4), freq=TRUE )
```

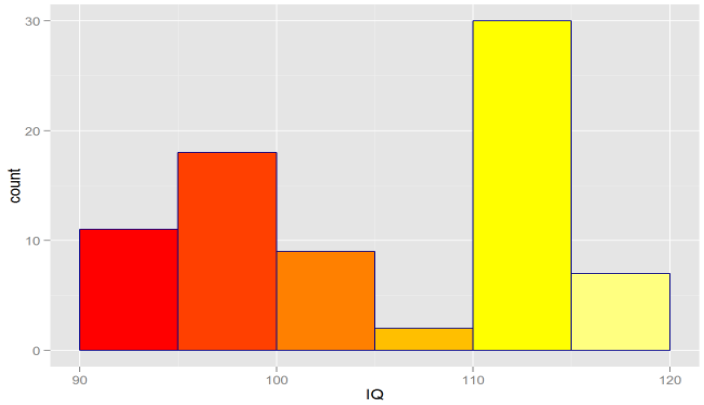
```
## Warning in plot.histogram(r, freq = freq1, col = col, border = border,
## angle = angle, : the AREAS in the plot are wrong -- rather use 'freq =
## FALSE'
```

```
library(ggplot2)
```



Gambar 3.77

```
ggplot(data=simpan, aes(IQ)) + geom_histogram(breaks=c(90,95,100,105,110,115,120), col="darkblue",
fill=heat.colors(6) )
```



Gambar 3.78

Referensi

1. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.
2. http://www.statmethods.net/advgraphs/ggplot2.html
3. https://cran.r-project.org/web/packages/ggplot2/index.html
4. http://www.r-bloggers.com/installing-r-packages/
5. http://www.r-bloggers.com/how-to-make-a-histogram-with-ggplot2/
6. http://docs.ggplot2.org/current/geom_histogram.html
7. http://www.r-bloggers.com/how-to-make-a-histogram-with-ggplot2/
8. http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)/
9. http://docs.ggplot2.org/0.9.3.1/geom_bar.html
10. http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2)/
11. http://www.r-bloggers.com/using-r-barplot-with-ggplot2/

BAB 4

UKURAN GEJALA PUSAT, LETAK, PENCARAN, KEMIRINGAN DAN KERUNCINGAN

Ukuran Gejala Pusat (Measure of Central Tendency)

Ukuran gejala pusat merupakan suatu ukuran atau nilai yang letaknya cenderung terletak dipusat data. Berikut beberapa penjelasan mengenai ukuran gejala pusat. Smidth dan Sanders (2000:73) menyatakan sebagai berikut.

"*You know from Chapter 2 that there are several measures of central tendency. The purpose of these measures is to summarize in a single value the typical size, middle property, or central location of a set of values. The most familiar measure of central tendency is, of course, the* ***arithmetic mean****, which is simply the sum of the values of a group of items divided by the number of such items. But you also saw in Chapter 2 that the* ***median*** *and* ***mode*** *are other measures of central tendency that are commonly used.*"

Spiegel dan Stephens (2008:62) menyatakan sebagai berikut.

"***An average*** *is a value that is typical, or* ***representative, of a set of data****. Since such typical values* ***tend to lie centrally within a set of data arranged according to magnitude****, averages are also called measures of central tendency.*

Several types of averages can be defined, the ***most common being the arithmetic mean****, the median, the mode, the geometric mean, and the harmonic mean.* ***Each has advantages and disadvantages, depending on the data and the intended purpose***."

Berdasarkan uraian di atas, nilai rata-rata dapat diartikan sebagai nilai tipikal atau representatif atau perwakilan dari suatu set data. Beberapa contoh dari ukuran gejala pusat atau rata-rata adalah rata-rata aritmatik (*arithmetic mean*), median, modus, rata-rata geometrik, dan rata-rata harmonik. Di antara berbagai ukuran gejala pusat tersebut memiliki kelebihan dan kekurangan, bergantung pada data dan tujuan yang dimaksud.

Smidth dan Sanders (2000:73) menyatakan sebagai berikut.

"*Data often have a tendency to congregate about some central value, and this central value may then be used as a summary measure* ***to describe the general data pattern***."

Misalkan diberikan data (sampel) seperti pada Tabel 4.1.

Tabel 4.1

Nilai	Nilai	Nilai	Nilai	Nilai
1	5	9	12	16
2	6	10	13	17
3	7	11	14	18
4	8	11	15	

Berdasarkan data pada Tabel 4.1, berikut akan dihitung jumlah keseluruhan nilai (*sum*), rata-rata aritmatik, modus, dan median.

Jumlah Keseluruhan Nilai (*Sum*)

Andaikan terdapat n buah nilai, yakni $X_1, X_2, X_3, \dots, X_n$. Jumlah dari keseluruhan nilai tersebut dihitung dengan rumus sebagai berikut.

$$jumlah\ keseluruhan\ nilai = \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \cdots + X_n$$

Jumlah keseluruhan nilai untuk data pada Tabel 4.1 adalah $1 + 2 + 3 + \cdots + 18 = 172$.

Rata-Rata Aritmatik atau Rata-Rata Hitung

Rata-rata aritmatik atau sering disebut juga dengan nama rata-rata hitung, merupakan jumlah seluruh nilai dari data, dibagi dengan banyaknya data. Berikut rumus untuk menghitung nilai rata-rata aritmatik (sampel).

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$= \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

Berikut akan dihitung nilai rata-rata aritmatik berdasarkan data (sampel) pada Tabel 4.1.

$$\bar{X} = \frac{1 + 2 + 3 + \cdots + 18}{19}$$

$$\bar{X} = 9{,}578947$$

Nilai rata-rata aritmatik berdasarkan data pada Tabel 4.1 adalah $9{,}578947$.

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 11, 12, 13, 14, 15, 16, 17, 18$

Perhatikan bahwa rata-rata hitung 9,57 cenderung terletak di pusat data.

Modus (*Mode*)

Modus merupakan nilai data dengan frekuensi atau jumlah kemunculan paling banyak. Berdasarkan data pada Tabel 4.1, nilai dengan frekuensi kemunculan paling banyak adalah nilai 11, yakni muncul sebanyak dua kali.

Median

Spiegel dan Stephens (2008:64) menyatakan sebagai berikut.

"*The median of a set of numbers arranged in order of magnitude (i.e., in an array) is either* ***the middle value*** *or* ***the arithmetic mean of the two middle values****.*

"Geometrically the median is the value of X (abscissa) corresponding to the vertical line which divides a histogram ***into two parts having equal areas****. This value of X is sometimes denoted by* $\tilde{X}$*".*

Berdasarkan uraian tersebut, median juga disebut juga dengan **nilai tengah** (*middle value*) atau **rata-rata aritmatik** dari dua nilai tengah. Nilai dari median membagi data menjadi dua bagian yang sama. Notasi atau simbol untuk rata-rata aritmatik sampel adalah $\bar{X}$, sementara notasi atau simbol median sampel adalah $\tilde{X}$. Sebelum menghitung nilai median, terlebih dahulu data diurutkan dari yang terkecil hingga terbesar. Berikut rumus menghitung median untuk data dengan jumlah genap.

$$Median = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

Berikut rumus menghitung median untuk data dengan jumlah ganjil.

$$Median = X_{\frac{n+1}{2}}$$

Perhatikan bahwa $X_{\frac{n}{2}}$ merupakan nilai X yang terletak pada urutan ke$-\frac{n}{2}$. Sebelum menghitung nilai median, data terlebih dahulu diurutkan dari yang terkecil hingga yang terbesar. Berikut disajikan kembali data pada Tabel 4.1 setelah diurutkan dari yang terkecil hingga terbesar.

1,2,3,4,5,6,7,8,9,10,11,11,12,13,14,15,16,17,18.

Diketahui banyaknya nilai $n = 19$, sehingga banyaknya data adalah ganjil.

$$Median = X_{\frac{n+1}{2}}$$

$$Median = X_{\frac{19+1}{2}}$$

$$Median = X_{10}$$

Perhatikan bahwa X_{10} berarti nilai median terletak pada data dengan urutan ke-10, yakni 10.

$1, 2, 3, 4, 5, 6, 7, 8, 9, \underline{10}, 11, 11, 12, 13, 14, 15, 16, 17, 18$

Nilai median 10 cenderung terletak di pusat data serta nilai median tersebut membagi data menjadi dua bagian yang sama.

Perhatikan bahwa nilai median membagi menjadi dua bagian yang sama. Bagian pertama adalah {1,2,3,4,5,6,7,8,9}, dan bagian kedua adalah {11,11,12,13,14,15,16,17,18}. Perhatikan bahwa masing-masing bagian terdiri dari 9 nilai.

Mann dan Lacke (2011:85) menyatakan sebagai berikut.

"*The median gives the center of a histogram, with half of the data values to the left of the median and half to the right of the median.* ***The advantage of using the median as a measure of central tendency*** *is that it is not influenced by outliers. Consequently, the median is preferred over the mean as a measure of central tendency for data sets that contain outliers. For example, when a data set has outliers, instead of using the mean, we can use either the* ***trimmed mean*** *or* ***median*** *as a measure of central tendency.*"

Berdasarkan uraian tersebut, keuntungan menggunakan median sebagai ukuran gejala pusat adalah median tidak terpengaruh oleh *outlier* (data pencilan). Oleh karena itu, median lebih disukai dibandingkan rata-rata atau *mean* (rata-rata aritmatik) sebagai ukuran gejala pusat, untuk data yang mengandung *outlier*.

Ukuran Letak (Measure of Position)

Kuartil dan desil merupakan jenis-jenis dari ukuran letak. Ukuran tersebut membagi data menjadi beberapa bagian yang sama. Sebagai contoh pada ukuran kuartil terdapat tiga buah nilai. Letak dari nilai-nilai kuartil tersebut membagi data menjadi empat bagian yang sama.

Kuartil (K)

Ukuran kuartil terdiri dari tiga buah nilai yang membagi data menjadi empat bagian yang sama.

$$1,2,\textcircled{3},4,5,\textcircled{6},7,8,\textcircled{9},10,11.$$

Nilai kuartil dikelompokkan atas tiga, yakni kuartil pertama (K_1), kuartil kedua (K_2), dan kuartil ketiga (K_3). Angka $3, 6,$ dan 9 masing-masing merupakan K_1, K_2, dan K_3. Berikut rumus untuk menghitung nilai kuartil.

$$K_i = \frac{i(n+1)}{4} \; ; i = 1,2,3$$

Perhatikan bahwa K_i merupakan nilai dari kuartil ke$-i$ dengan $i = 1, 2,$ dan 3. Berikut disajikan kembali data pada Tabel 4.1.

$$1,2,3,4,\textcircled{5},6,7,8,9,\textcircled{10},11,11,12,13,\textcircled{14},15,16,17,18$$

Diketahui banyaknya nilai data $n = 19$. Berikut akan dihitung nilai dari K_1, K_2, dan K_3.

$$K_1 = \frac{1(19+1)}{4}$$

$$K_1 = 5$$

$K_1 = 5$ berarti nilai K_1 terletak pada data dengan urutan ke-5, yakni 5.

$$K_2 = \frac{2(19+1)}{4}$$

$$K_2 = 10$$

$K_2 = 10$ berarti nilai K_2 terletak pada data dengan urutan ke-10, yakni 10.

$$K_3 = \frac{3(19+1)}{4}$$

$$K_3 = 15$$

$K_3 = 15$ berarti nilai K_3 terletak pada data dengan urutan ke-15, yakni 14. Ketiga nilai kuartil tersebut membagi data menjadi empat bagian yang sama. Bagian pertama adalah {1,2,3,4}, bagian kedua adalah {6,7,8,9}, bagian ketiga adalah {11,11,12,13}, dan bagian keempat adalah {15,16,17,18}. Perhatikan bahwa banyaknya nilai untuk masing-masing bagian adalah 4.

Desil $(\boldsymbol{D})$

Ukuran desil terdiri dari sembilan nilai yang membagi data menjadi sepuluh bagian yang sama.

1,(2),3,(4),5,(6),7,(8),9,(10),11,(11),12,(13),14,(15),16,(17),18.

Perhatikan bahwa nilai-nilai yang dilingkar merupakan nilai-nilai desil. Nilai-nilai tersebut membagi data menjadi 10 bagian yang sama. Masing-masing bagian terdiri dari 1 nilai. Terdapat sembilan nilai desil, yakni desil pertama (D_1), desil kedua (D_2), dan sampai dengan desil kesembilan (D_9). Berikut rumus untuk menghitung nilai desil.

$$D_i = \frac{i(n+1)}{10} \; ; i = 1,2,3,\ldots,9$$

Berikut akan dihitung nilai desil pertama, kedelapan, dan kesembilan berdasarkan data pada Tabel 4.1.

$$D_1 = \frac{1(19+1)}{10} = 2$$

Nilai desil ke-1 terletak pada data dengan urutan ke-2, yakni 2.

$$D_8 = \frac{8(19+1)}{10} = 16$$

Nilai desil ke-8 terletak pada data dengan urutan ke-16, yakni 15.

$$D_9 = \frac{9(19+1)}{10} = 18$$

Nilai desil ke−9 terletak pada data dengan urutan ke-18, yakni 17. Sembilan nilai desil tersebut membagi data menjadi sepuluh bagian yang sama dengan banyaknya nilai untuk masing-masing bagian adalah 1.

Ukuran Pencaran atau Dispersi atau Sebaran

Misalkan diberikan 4 data, beserta nilainya (Tabel 4.2).

Tabel 4.2

Data 1	70	70	70	70	70	$\bar{X} = 70$
Data 2	50	60	70	80	90	$\bar{X} = 70$
Data 3	20	60	70	100	100	$\bar{X} = 70$
Data 4	20	20	10	100	200	$\bar{X} = 70$

Berdasarkan Tabel 4.2, nilai rata-rata untuk data 1 adalah 70, nilai rata-rata untuk data 2 juga 70, begitu juga untuk data 3 dan data 4. Namun **nilai rata-rata untuk data manakah yang dapat mewakili data dengan baik?** Berdasarkan pengamatan, nilai rata-rata dari data 1 dapat mewakili data 1 dengan baik (secara sempurna), nilai rata-rata dari data 2 cukup baik dalam mewakili data 2, namun nilai rata-rata dari data 3 dan data 4 kurang baik dalam mewakili data 3 dan data 4.

Ukuran pencaran atau dispersi merupakan suatu nilai yang mengukur tingkat **pencaran atau sebaran nilai-nilai data terhadap nilai rata-ratanya**. **Nilai pencaran yang tinggi** menunjukkan **nilai-nilai data cenderung terletak cukup jauh** terhadap **nilai rata-rata** dari data tersebut. Dengan kata lain, data semakin bervariasi atau heterogen. Sebagaimana Mann dan Lacke (2011:92) menyatakan sebagai berikut.

"*Two data sets with the same mean may have different spreads. The variation among the values of observations for one data set may be larger or smaller than for the other data set. (Note that the words dispersion, and variation have the same meaning).*

Thus, mean, median, or mode by itself is not a sufficient measure ***to reveal shape of the distribution of a data set****. We also need a measure that can provide some information about* ***the variation among data values****. The measures that help us learn about the spread of data set are called* ***the measure of dispersion****. The measures of central tendency and dispersion* ***taken together give a better picture of a data set than the measures of central tendency alone****. This section discusses three measures of dispersion: range, variance, and standard deviation.*"

Ukuran pencaran yang akan dipaparkan dalam tulisan ini adalah *range, variance*, dan standar deviasi. Misalkan diberikan data seperti pada Tabel 4.3.

Tabel 4.3

Nilai	Nilai	Nilai	Nilai	Nilai
10	20	30	40	50
10	30	30	40	50
10	30	30	40	50
20	30	30	50	

Nilai Maksimum

Nilai maksimum merupakan nilai yang paling tinggi dari suatu data. Berdasarkan data pada Tabel 4.3, nilai maksimum adalah nilai 50.

Nilai Minimum

Nilai minimum merupakan nilai yang paling rendah dari suatu data. Berdasarkan data pada Tabel 4.3, nilai minimum adalah nilai 10.

Range

Range merupakan selisih antara nilai maksimum dengan nilai minimum. Diketahui nilai maksimum adalah 50 dan nilai minimum adalah 10, sehingga nilai *range* adalah $50 - 10 = 40$. Ukuran *range* sama seperti rata-rata aritmatik, yakni memiliki kelemahan ketika dalam suatu data mengandung *outlier*. Sebagaimana Mann dan Lacke (2011:93) menyatakan sebagai berikut.

"*The range, like the mean, has the disadvantage of being influenced by outliers. Consequently, the range is not good measure of dispersion to use for a data set that contains outliers.*

Another disadvantage of using the range as a measure of dispersion is that is calculation is based on two values only: the largest and smallest. All other values in a data set are ignored when calculation the range. Thus, ***the range is not very satisfactory measure of dispersion***."

Sebagai contoh misalkan diberikan data dengan nilai 1, 2, 3, 4, 5, 100. Nilai *range* berdasarkan data tersebut adalah $100 - 1 = 99$. Seandainya data dengan nilai 100 tidak diikutsertakan dalam penghitungan nilai *range*, maka diperoleh nilai *range* $5 - 1 = 4$. Perhatikan bahwa nilai *range* menurun, dari 100 menjadi 4. Nilai data 100 merupakan *outlier* (data pencilan).

Variance

Variance (dalam hal ini *variance* untuk sampel) dilambangkan dengan s^2. Berikut rumus untuk menghitung nilai *variance*.

$$s^2 = \frac{|X - \bar{X}|^2}{n - 1}.$$

Nilai *variance* sampel (s^2) berdasarkan data pada Tabel 4.3 adalah

$$s^2 = \frac{3 \times |10 - 31{,}6|^2 + 2 \times |20 - 31{,}6|^2 + \cdots + 4 \times |50 - 31{,}6|^2}{19 - 1}$$

$$s^2 = 180{,}7018$$

Standar Deviasi

Standar deviasi merupakan akar kuadrat positif *variance* $\left(\sqrt{s^2} = s\right)$. Nilai dari standar deviasi dapat diinterpretasi sebagai nilai yang menunjukkan seberapa dekat nilai-nilai data menyebar atau berkumpul di sekitar rata-ratanya. Standar deviasi merupakan salah satu dari ukuran pencaran yang paling sering digunakan. Mann dan Lacke (2011:93) menyatakan sebagai berikut.

"*The standard deviation is the most-used measure of dispersion. The value of standard deviation tells how closely the values of a data set are clustered around the mean. In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean. In contrast, a larger value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean.*"

Diketahui nilai *variance* adalah 180,7018, sehingga nilai standar deviasi adalah $\sqrt{180{,}7018} = 13{,}4425$. Tabel 4.4 menyajikan hasil perhitungan untuk nilai minimum, maksimum, *range*, *variance*, dan standar deviasi, berdasarkan data pada Tabel 4.2.

Berdasarkan data pada Tabel 4.4, diketahui nilai standar deviasi untuk data 1 bernilai 0, data 2 bernilai 15,811, data 3 bernilai 33,166, dan data 4 bernilai 81,240. Perhatikan bahwa pada data 1, **seluruh nilai data sama**, yakni seluruhnya 70, sehingga nilai standar deviasinya 0 (begitu juga dengan nilai *range* dan *variance*). Dapat dilihat bahwa semakin besar nilai standar deviasi dari suatu data, maka sebaran data cenderung jauh terhadap rata-ratanya (walaupun ada beberapa data yang dekat dengan rata-ratanya). Perhatikan juga Tabel 4.5.

Tabel 4.4

Data						Rata-Rata	*Range*	*Variance*	Standar Deviasi
Data 1	70	70	70	70	70	70	0	0	0
Data 2	50	60	70	80	90	70	40	250	15,811
Data 3	20	60	70	100	100	70	80	1100	33,166
Data 4	20	20	10	100	200	70	180	6600	81,240

Tabel 4.5

				Rata-Rata	Range	Variance	Standar Deviasi
Data 5	13	14	15	14	2	1	1
Data 6	12	14	16	14	4	4	2
Data 7	8	14	20	14	12	36	6
Data 8	1	14	27	14	26	169	13

Pada Tabel 4.5, nilai rata-rata untuk data 5 sampai data 8 adalah 14. Untuk data 5, jarak 13 ke 14 adalah 1, yakni $|14 - 13| = 1$, begitu juga jarak dari 15 ke 14, yakni $|15 - 14| = 1$. Nilai standar deviasinya adalah 1. Untuk data 5, jarak dari 12 ke 14 adalah 2, yakni $|14 - 12| = 2$, begitu juga jarak dari 16 ke 14, yakni $|16 - 14| = 2$. Nilai standar deviasinya adalah 2. Semakin besar nilai standar deviasi dari suatu data, maka sebaran data cenderung jauh terhadap rata-ratanya. Perhatikan juga pada Tabel 4.6.

Tabel 4.6

						Rata-Rata	Standar Deviasi
Data 9	14	15	16	17	18	16	1,58113883
Data 10	12	14	16	18	20	16	3,16227766
Data 11	10	13	16	19	22	16	4,74341649
Data 12	8	12	16	20	24	16	6,32455532
Data 13	6	11	16	21	26	16	7,90569415

Koefisien Variasi (*Coefficient of Variation*)

Misalkan diberikan data berat badan dan IQ dari 5 siswa (Tabel 4.7).

Tabel 4.7

Siswa	Berat Badan	Uang Jajan
1	54,33	20000
2	58,89	20000
3	64,33	19000
4	54,21	20000
5	53,45	19000
Rata-Rata	57,042	19600
Standar Deviasi	4,604554	547,722558
Koefisien Variasi	0,080722	0,02794503

Andaikan akan dibandingkan, data mana yang lebih bervariasi atau heterogen, apakah data berat badan atau data uang jajan? Perhatikan bahwa **satuan data untuk berat badan (puluhan) dan uang jajan (puluhan ribu) berbeda**. Berdasarkan Tabel 4.7 diketahui nilai standar deviasi dari uang jajan, yakni 547,722, lebih besar dari pada nilai standar deviasi dari berat badan, yakni 4,604. Namun belum tentu berarti bahwa data uang jajan lebih bervariasi atau heterogen dibandingkan data berat badan. Hal ini dikarenakan satuan data berbeda.

Untuk itu dapat digunakan **koefisien variasi** untuk membandingkan tingkat variasi atau heterogen di antara dua atau lebih kelompok, ketika satuan data berbeda-beda. Spiegel dan Stephens (2008:100) menyatakan sebagai berikut.

"*Note that the coefficient of variation is independent of the units used. For this reason, it is useful* ***in comparing distributions*** *where* ***the units may be different****. A disadvantage of the coefficient of variation is that it fails to be useful when* $\bar{X}$ *is close to zero."*

Nilai dari koefisien variasi dihitung sebagai berikut.

$$Koefisien\ Variasi\ (KV) = \frac{s}{\bar{X}}$$

Berdasarkan Tabel 4.7, diketahui koefisien variasi untuk data berat badan adalah 0,080722, sementara koefisien variasi untuk data uang jajan adalah 0,02794503. Sehingga data berat badan lebih bervariasi atau heterogen dibandingkan data uang jajan.

Data yang Dibakukan (*Standardized Data*)

Suatu variabel yang mengukur **deviasi dari rata-rata**, dalam **unit atau satuan standar deviasi**, disebut variabel yang dibakukan (*standardized variable*). Sebagaimana Spiegel dan Stephens (2008:101) menyatakan sebagai berikut.

"*The variable that measures the deviation from the mean in units of the standard deviation is called a standardized variable, is a dimensionless quantity (i.e., is independent of the units used), and is given by*

$$z = \frac{X - \bar{X}}{s}$$

*If the **deviations from the mean** are given **in units of the standard deviation,** they are said to be expressed in standard units, or standard scores. These are of great value in the comparison of distributions."*

Berdasarkan uraian tersebut, data dalam bentuk standar atau baku sangat berguna untuk tujuan perbandingan distribusi dari beberapa kelompok data. Suatu data dari variabel asli X, dapat ditransformasi dalam bentuk standar dengan rumus sebagai berikut.

$$Z = \frac{X - \bar{X}}{s}$$

Tabel 4.8

Siswa	Berat	Uang Jajan	Z_Baku	Z_Uang Jajan
1	54,33	20000	-0,588982091	0,730296743
2	58,89	20000	0,401341779	0,730296743
3	64,33	19000	1,582780781	-1,095445115
4	54,21	20000	-0,615043245	0,730296743
5	53,45	19000	-0,780097224	-1,095445115
Rata-Rata	57,042	19600	0	0
Standar Deviasi	4,604554	547,722558	1	1
Koefisien Variasi	0,080722	0,02794503		

Berdasarkan Tabel 4.8, nilai standar atau baku untuk uang jajan 20000 adalah 0,730296743. Nilai tersebut diperoleh sebagai berikut.

$$Z = \frac{X - \bar{X}}{s} = \frac{20000 - 19600}{547{,}722558} = 0{,}730296743$$

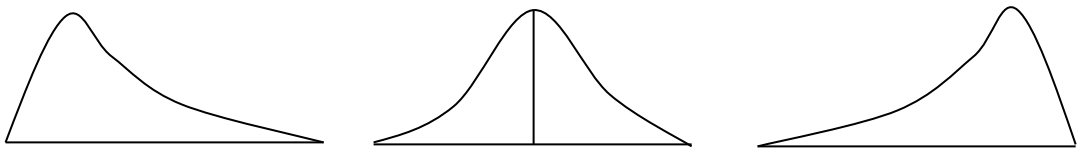
Nilai standar atau baku untuk berat badan 54,33 adalah -0,588982091. Nilai tersebut diperoleh sebagai berikut.

$$Z = \frac{X - \bar{X}}{s} = \frac{54{,}33 - 57{,}042}{4{,}604554} = -0{,}588982091$$

Data dalam bentuk standar atau baku **memiliki nilai rata-rata 0 dan standar deviasi 1.**

Ukuran Kemiringan (Skewness)

Ukuran kemiringan atau *skewness* merupakan suatu nilai yang mengukur ketidaksimetrisan distribusi data. Suatu data dikatakan berdistribusi simetris sempurna bila nilai rata-rata, median, dan modus dalam data adalah sama.



(1) (2) (3)

Gambar 4.1

Pada Gambar 4.1 (1) kurva cenderung condong ke kanan atau disebut kurva positif, sementara Gambar 4.1 (2) kurva bersifat simetris. Pada Gambar 4.1 (3) kurva cenderung condong ke kiri atau disebut kurva negatif. Berikut rumus untuk menghitung nilai kemiringan suatu data.

$$Kemiringan = \frac{n}{(n-1)(n-2)}\left(\frac{\sum(X-\bar{X})^3}{s^3}\right)$$

Bila nilai kemiringan < 0 atau negatif, maka kurva cenderung condong ke kiri (kurva negatif). Jika nilai kemiringan > 0 atau positif, maka kurva cenderung condong ke kanan (kurva positif). Jika nilai kemiringan mendekati 0 atau 0, maka kurva cenderung simetris. Spiegel dan Stephens (2008:125) menyatakan sebagai berikut.

"*Skewness is the degree of asymmetry, or departure from symmetry, of a distribution. If the frequency curve (smoothed frequency polygon) of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right, or to have positive skewness. If the reverse is true, it is said to be skewed to the left, or to have negative skewness."*

Misalkan diberikan data seperti pada Tabel 4.9. Berdasarkan data pada Tabel 4.9, berikut akan dihitung nilai kemiringan. Dari Tabel 4.10, diketahui $\bar{X} = 3{,}6$ dan $s = 1{,}454058$, sehingga nilai kemiringan dapat dihitung sebagai berikut.

$$Kemiringan = \frac{n}{(n-1)(n-2)}\left(\frac{\sum(X-\bar{X})^3}{s^3}\right)$$

$$Kemiringan = \frac{15}{(15-1)(15-2)}\left(\frac{6{,}48}{1{,}454058^3}\right)$$

$$Kemiringan = 0{,}17372$$

Tabel 4.9

Nilai (X)	Nilai (X)	Nilai (X)	Nilai (X)
1	3	4	5
2	3	4	6
2	3	4	6
3	3	5	

Tabel 4.10

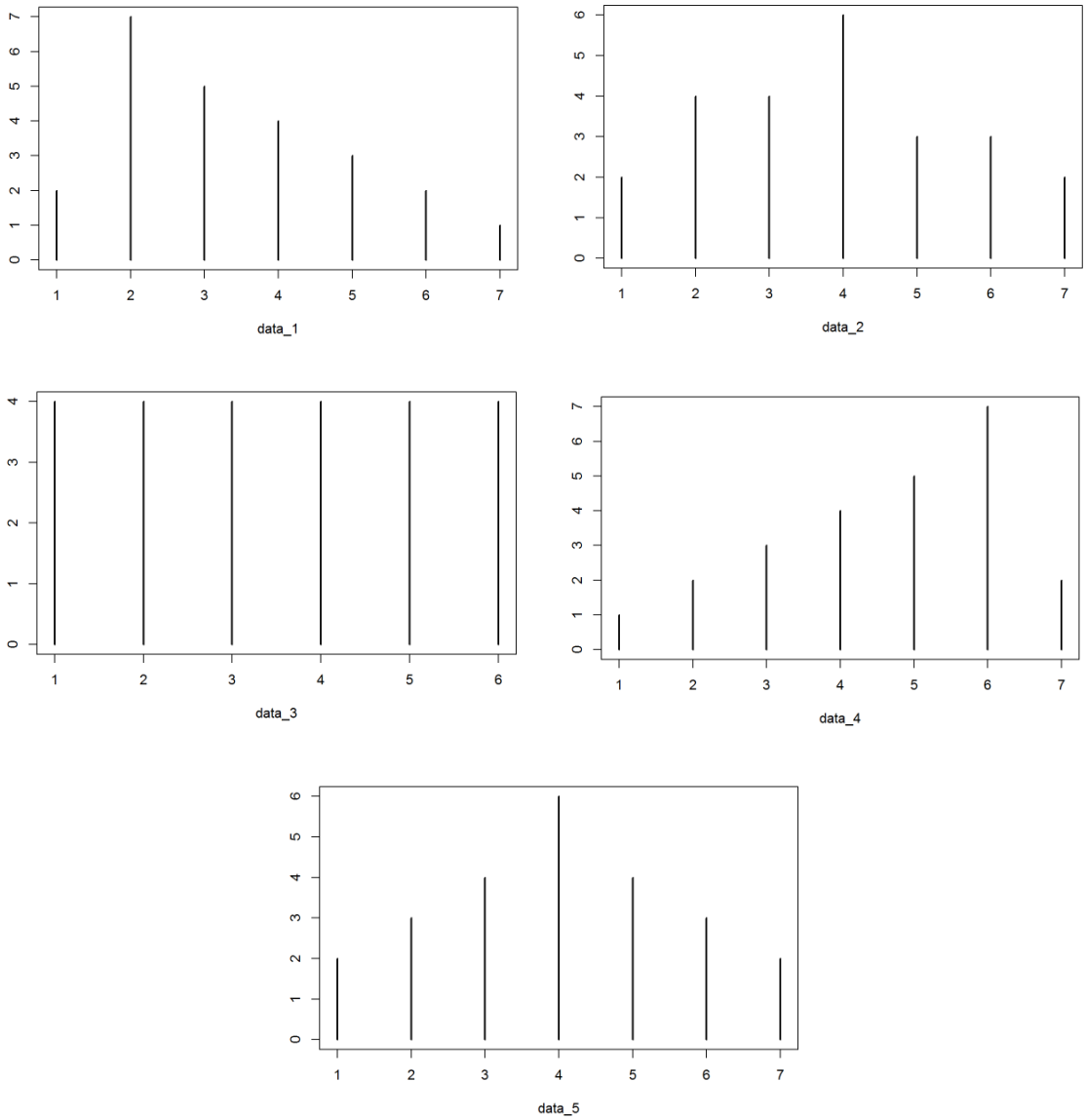
	X	f	fX	$f\sum(X-\bar{X})^3$
	1	1	1	-17,576
	2	2	4	-8,192
	3	5	15	-1,08
	4	3	12	0,192
	5	2	10	5,488
	6	2	12	27,648
Jumlah		15	54	6,48
Rata-rata ($\bar{X}$)	3,6			
Standar deviasi (s)	1,454058			

Tabel 4.11

No	Data 1	Data 2	Data 3	Data 4	Data 5
1	1	1	1	1	1
2	1	1	1	2	1
3	2	2	1	2	2
4	2	2	1	3	2
5	2	2	2	3	2
6	2	2	2	3	3
7	2	3	2	4	3
8	2	3	2	4	3
9	2	3	3	4	3
10	3	3	3	4	4
11	3	4	3	5	4
12	3	4	3	5	4
13	3	4	4	5	4
14	3	4	4	5	4
15	4	4	4	5	4
16	4	4	4	6	5
17	4	5	5	6	5
18	4	5	5	6	5
19	5	5	5	6	5
20	5	6	5	6	6
21	5	6	6	6	6
22	6	6	6	6	6
23	6	7	6	7	7
24	7	7	6	7	7
Kemiringan	0,5668	0,1545	0,0000	-0,5668	0,0000
Rata-Rata	3,375	3,875	3,5	4,625	4
Median	3	4	3,5	5	4
Modus	2	4	-	6	4

Tabel 4.12

Nilai	Frekuensi				
	Data 1	Data 2	Data 3	Data 4	Data 5
1	2	2	4	1	2
2	7	4	4	2	3
3	5	4	4	3	4
4	4	6	4	4	6
5	3	3	4	5	4
6	2	3	4	7	3
7	1	2	-	2	2



Gambar 4.2

Tabel 4.11 menyajikan 5 data, yakni data 1 sampai data 5. Kelima data tersebut masing-masing memiliki nilai data sebanyak 24. Nilai kemiringan untuk data 1 adalah 0,5668, data 2 bernilai 0,1545, data 3 adalah 0, data 4 adalah -0,5668, dan data 5 adalah 0. Perhatikan bahwa nilai kemiringan untuk data 3 dan data 5 bernilai 0 (**simetri terhadap rata-rata**). Pada Tabel

4.12 menyajikan tabel distribusi frekuensi untuk data 1 sampai dengan data 5, berdasarkan Tabel 4.11. Berdasarkan Tabel 4.12, untuk data 1, nilai 1 sebanyak 2, nilai 2 sebanyak 7, nilai 3 sebanyak 5, dan seterusnya. Pada data 1, diketahui rata-rata > median > modus (miring ke kanan). Sementara pada data 4, diketahui rata-rata < median < modus (miring ke kiri). Pada data 5, diketahui rata-rata = median = modus (simetri, kasus **unimodus** atau *unimodal*). Pada data 3, tidak bersifat unimodus (unimodus atau *unimodal* berarti jumlah modus dalam data sebanyak 1). Gambar 4.2 merupakan grafik untuk distribusi frekuensi pada Tabel 4.12.

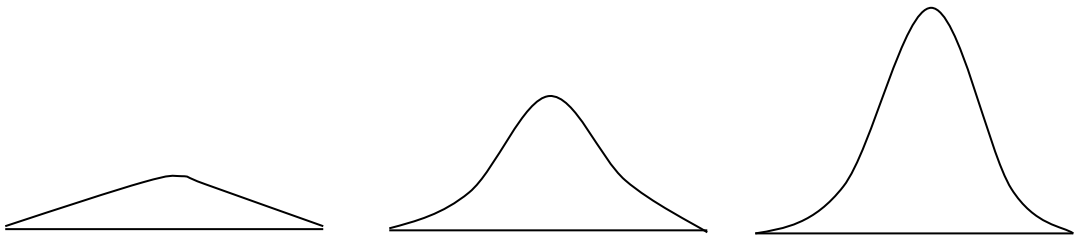
Ukuran Keruncingan (Kurtosis)

Ukuran keruncingan atau *kurtosis* merupakan suatu nilai yang mengukur tingkat keruncingan atau ketinggian puncak dari distribusi data. Berikut rumus untuk menghitung kurtosis.

$$Kurtosis = \left\{\frac{(n)(n+1)\sum(X-\bar{X})^4}{(n-1)(n-2)(n-3)s^4}\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Pada Gambar 4.3 (1), (2), dan (3), ketinggian atau keruncingan puncak kurva berbeda-beda. Gambar 4.3 (2) merupakan kurva normal atau mesokurtis (kurva tidak terlalu tajam dan datar). Pada Gambar 4.3 (1), kurva cenderung datar dan puncak tidak terlalu tinggi. Kurva ini dinamakan kurva platikurtis. Pada Gambar 4.3 (3), puncak kurva terlihat lancip dan tinggi. Kurva ini dinamakan kurva leptokurtis. Spiegel dan Stephens (2008:125) menyatakan sebagai berikut.

"*Kurtosis is the degree of peakedness of a distribution, usually taken relative to a normal distribution. A distribution having a relatively high peak is called leptokurtic, while one which is flat-topped is called platykurtic. A normal distribution, which is not very peaked or very flat-topped, is called mesokurtic.*"



(1) (2) (3)

Gambar 4.3

Berikut akan dihitung nilai kurtosis berdasarkan data pada Tabel 4.9. Berdasarkan data pada Tabel 4.10, diketahui nilai $\bar{X} = 3{,}6$ dan $s = 1{,}454058$, sehingga

$$Kurtosis = \left\{\frac{(n)(n+1)\sum(X-\bar{X})^4}{(n-1)(n-2)(n-3)s^4}\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

$$Kurtosis = \left\{\frac{(15)(15+1)(133{,}568)}{(15-1)(15-2)(15-3)(1{,}454^3)}\right\} - \frac{3(15-1)^2}{(15-2)(15-3)}$$

$$Kurtosis = -0{,}485756$$

Nilai kurtosis berdasarkan perhitungan adalah $-0{,}485756$. Tabel untuk perhitungan disajikan pada Tabel 4.13.

Tabel 4.13

	X	f	fX	$f\sum(X-\bar{X})^4$
	1	1	1	45,6976
	2	2	4	13,1072
	3	5	15	0,648
	4	3	12	0,0768
	5	2	10	7,6832
	6	2	12	66,3552
Jumlah		15	54	133,568

Tabel 4.14 menyajikan 3 data, yakni data 1 sampai data 3. Ketiga data tersebut masing-masing memiliki nilai data sebanyak 12. Nilai kurtosis untuk data 1 adalah -1,65, data 2 bernilai -0,85556, dan data 3 adalah 0,73333. Perhatikan bahwa semakin tinggi nilai kurtosis, maka puncak kurva semakin tinggi dan lancip (lihat Gambar 4.4).

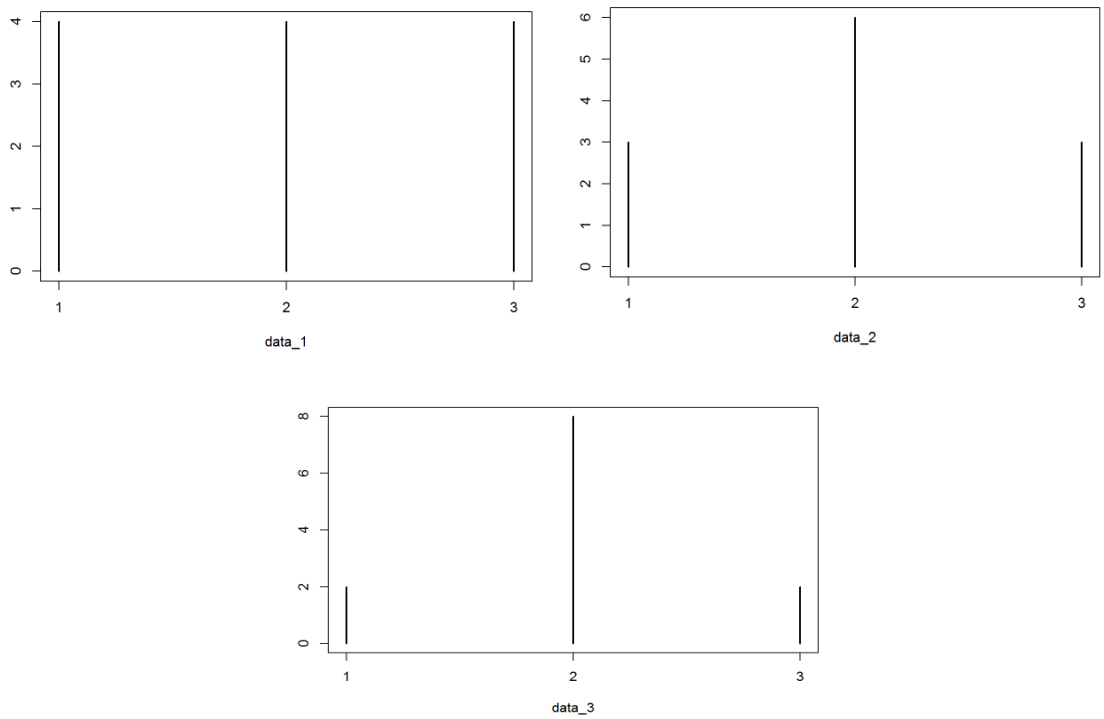
Pada Tabel 4.15 menyajikan tabel distribusi frekuensi untuk data 1 sampai dengan data 3 berdasarkan Tabel 4.14. Berdasarkan Tabel 4.15, untuk data 1, nilai 1 sebanyak 4, nilai 2 sebanyak 4, dan nilai 3 sebanyak 4. Untuk data 2, nilai 1 sebanyak 3, nilai 2 sebanyak 6, dan nilai 3 sebanyak 3. Gambar 4.4 merupakan grafik untuk distribusi frekuensi pada Tabel 4.15.

Tabel 4.14

No	data1	data2	data3
1	1	1	1
2	1	1	1
3	1	1	2
4	1	2	2
5	2	2	2
6	2	2	2
7	2	2	2
8	2	2	2
9	3	2	2
10	3	3	2
11	3	3	3
12	3	3	3
Kurtosis	-1,65	-0,85556	0,733333

Tabel 4.15

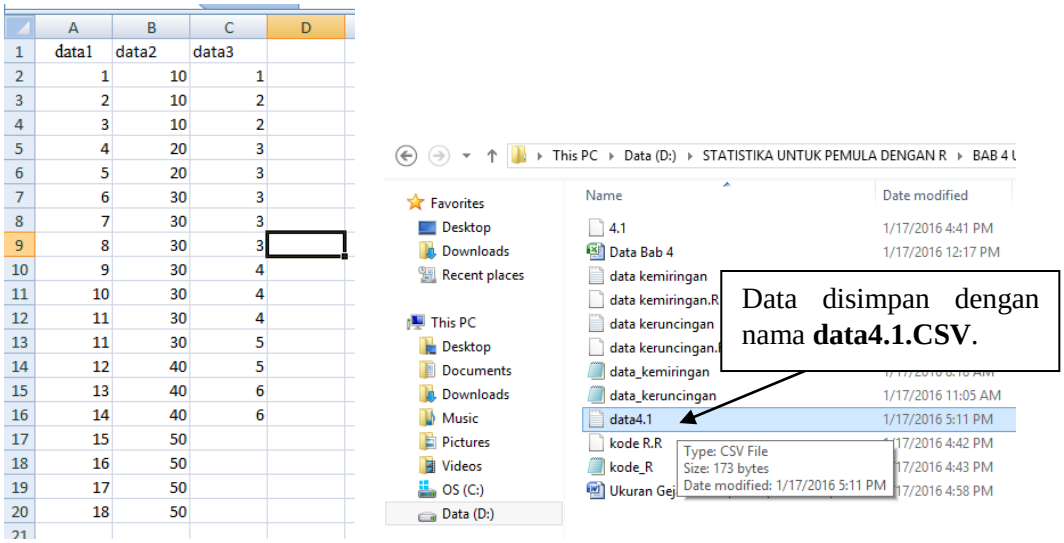
Nilai	Frekuensi		
	Data 1	Data 2	Data 3
1	4	4	4
2	3	6	3
3	2	8	2



Gambar 4.4

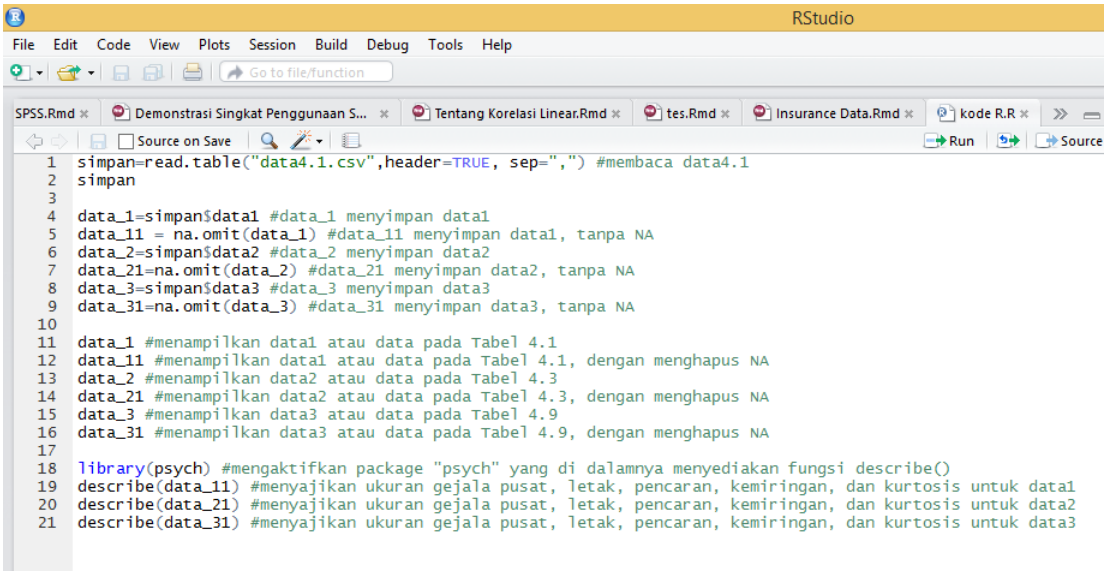
Aplikasi dalam R

Data pada Tabel 4.1, Tabel 4.3, dan Tabel 4.9 disimpan terlebih dahulu dalam *Microsoft Excel*, dan disimpan dengan format **.CSV** (perhatikan Gambar 4.5).



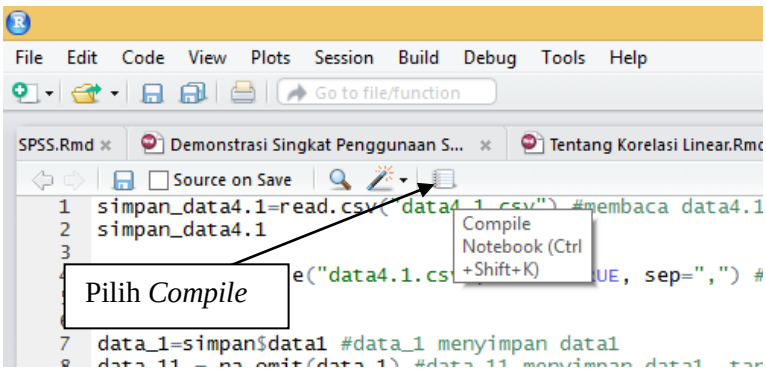
Gambar 4.5

Berikut disajikan kode R (Gambar 4.6) untuk memperoleh hasil perhitungan ukuran gejala pusat, letak, pencaran, kemiringan, dan keruncingan, berdasarkan data pada Tabel 4.1 (data1, lihat Gambar 4.5), Tabel 4.3 (data2, lihat Gambar 4.5), dan Tabel 4.9 (data3, lihat Gambar 4.5).

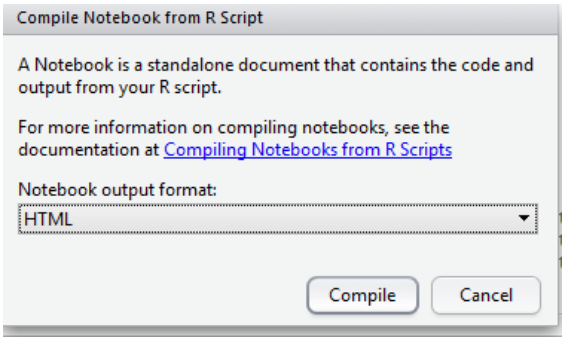


Gambar 4.6

Untuk mengeksekusi kode R pada Gambar 4.6, pilih *Compile* (Gambar 4.7). Pada Gambar 4.8, pilih format *output* HTML.



Gambar 4.7



Gambar 4.8

Interpretasi Kode R

Berikut disajikan kode R, beserta *output* dari kode R tersebut.

[1] simpan=read.table("data4.1.csv",header=TRUE, sep=",") #membaca data4.1
simpan

[2] data_1=simpan$data1 #data_1 menyimpan data1
data_11 = na.omit(data_1) #data_11 menyimpan data1, tanpa NA
data_2=simpan$data2 #data_2 menyimpan data2
data_21=na.omit(data_2) #data_21 menyimpan data2, tanpa NA
data_3=simpan$data3 #data_3 menyimpan data3
data_31=na.omit(data_3) #data_31 menyimpan data3, tanpa NA
data_1 #menampilkan data1 atau data pada Tabel 4.1
data_11 #menampilkan data1 atau data pada Tabel 4.1, dengan menghapus NA
data_2 #menampilkan data2 atau data pada Tabel 4.3
data_21 #menampilkan data2 atau data pada Tabel 4.3, dengan menghapus NA
data_3 #menampilkan data3 atau data pada Tabel 4.9
data_31 #menampilkan data3 atau data pada Tabel 4.9, dengan menghapus NA

[3] library(psych) #mengaktifkan package "psych" yang di dalamnya menyediakan fungsi describe()
describe(data_11) #menyajikan ukuran gejala pusat, letak, pencaran, kemiringan, dan kurtosis untuk data1
describe(data_21) #menyajikan ukuran gejala pusat, letak, pencaran, kemiringan, dan kurtosis untuk data2
describe(data_31) #menyajikan ukuran gejala pusat, letak, pencaran, kemiringan, dan kurtosis untuk data3

Gambar 4.9 merupakan hasil dari kode R [1]. Kode R [1] dapat diartikan variabel **simpan** ditugaskan untuk menyimpan data pada variabel **data1**, **data2**, dan **data3** dalam *file* **data4.1.csv**. Kemudian menampilkan data pada variabel **data1**, **data2**, dan **data3**. Perhatikan bahwa pada variabel **data1**, jumlah data sebanyak 19. Begitu juga pada variabel **data2**. Untuk variabel **data3**, jumlah data sebanyak 15, selebihnya adalah NA.

```
simpan=read.table("data4.1.csv",header=TRUE, sep=",") #membaca data4.1
simpan
```

```
##    data1 data2 data3
## 1      1    10     1
## 2      2    10     2
## 3      3    10     2
## 4      4    20     3
## 5      5    20     3
## 6      6    30     3
## 7      7    30     3
## 8      8    30     3
## 9      9    30     4
## 10    10    30     4
## 11    11    30     4
## 12    11    30     5
## 13    12    40     5
## 14    13    40     6
## 15    14    40     6
## 16    15    50    NA
## 17    16    50    NA
## 18    17    50    NA
## 19    18    50    NA
```

Gambar 4.9

Gambar 4.10 merupakan hasil dari kode R [2]. Sebagai contoh pada kode R **data_1=simpan$data1 #data_1 menyimpan data1**, dapat diartikan variabel **data_1** ditugaskan untuk menyimpan **data1** di dalam variabel **simpan**. Sehingga nilai dari **data_1** adalah 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 11, 12, 13, 14, 15, 16, 17, 18. Pada kode R **data_11 = na.omit(data_1) #data_11 menyimpan data1, tanpa NA** dapat diartikan variabel **data_11** ditugaskan untuk menyimpan **data_1**, dengan mengabaikan **NA**. Jadi nilai dari **data_11** adalah 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 11, 12, 13, 14, 15, 16, 17, 18.

Kode R **data_3=simpan$data3 #data_3 menyimpan data3**, dapat diartikan variabel **data_3** ditugaskan untuk menyimpan **data3** di dalam variabel **simpan**. Sehingga nilai dari **data_3** adalah 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, NA, NA, NA, NA. Pada kode R **data_31 = na.omit(data_3) #data_31 menyimpan data3, tanpa NA** dapat diartikan variabel **data_31** ditugaskan untuk menyimpan **data_3**, dengan mengabaikan **NA**. Jadi nilai dari **data_31** adalah 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6. Kode R **data_1** berarti menampilkan nilai pada variabel **data_1**, kode R **data_11** berarti menampilkan nilai pada variabel **data_11**, dan seterusnya.

```
data_1 #menampilkan data1 atau data pada Tabel 4.1
```

```
##  [1]  5  6  7  8  9 10 11 11 12 13 14 15 16 17 18 NA NA NA NA
```

```
data_11 #menampilkan data1 atau data pada Tabel 4.1, dengan menghapus NA
```

```
##  [1]  5  6  7  8  9 10 11 11 12 13 14 15 16 17 18
## attr(,"na.action")
## [1] 16 17 18 19
## attr(,"class")
## [1] "omit"
```

```
data_2 #menampilkan data2 atau data pada Tabel 4.3
```

```
##  [1] 10 10 10 20 20 30 30 30 30 30 30 30 40 40 40 50 50 50 50
```

```
data_21 #menampilkan data2 atau data pada Tabel 4.3, dengan menghapus NA
```

```
##  [1] 10 10 10 20 20 30 30 30 30 30 30 30 40 40 40 50 50 50 50
```

```
data_3 #menampilkan data3 atau data pada Tabel 4.9
```

```
##  [1]  1  2  2  3  3  3  3  3  4  4  4  5  5  6  6 NA NA NA NA
```

```
data_31 #menampilkan data3 atau data pada Tabel 4.9, dengan menghapus NA
```

Gambar 4.10

Gambar 4.11 merupakan hasil dari kode R [3]. Perhatikan kode R berikut.

library(psych)
describe(data_11)

Pada kode R tersebut, akan digunakan fungsi **describe**. Fungsi tersebut terdapat dalam *package* **psych**. Oleh karena itu, kode R **library(psych)** dapat diartikan mengaktifkan *package* **psych**. Setelah *package* **psych** diaktifkan, barulah fungsi **describe** dapat digunakan. Fungsi **describe** dalam hal ini digunakan untuk menentukan banyaknya data (n), rata-rata aritmatik (*mean*), standar deviasi (sd), median, minimum (min), maksimum (max), *range*, kemiringan (*skew*), dan kurtosis.

```
library(psych) #mengaktifkan package "psych" yang di dalamnya menyediakan fungsi describe()
describe(data_11) #menyajikan ukuran gejala pusat, letak, pencaran, kemiringan, dan kurtosis untuk d
```

```
##   vars  n mean  sd median trimmed  mad min max range  skew kurtosis   se
## 1    1 19 9.58 5.2     10    9.59 5.93   1  18    17 -0.04    -1.32 1.19
```

```
describe(data_21) #menyajikan ukuran gejala pusat, letak, pencaran, kemiringan, dan kurtosis untuk d
```

```
##   vars  n  mean    sd median trimmed   mad min max range  skew kurtosis
## 1    1 19 31.58 13.44     30   31.76 14.83  10  50    40 -0.14    -1.13
##     se
## 1 3.08
```

```
describe(data_31) #menyajikan ukuran gejala pusat, letak, pencaran, kemiringan, dan kurtosis untuk d
```

```
##   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## 1    1 15  3.6 1.45      3    3.62 1.48   1   6     5 0.14    -1.01 0.38
```

Gambar 4.11

Pada hasil R Gambar 4.11, nilai kemiringan dihitung dengan rumus sebagai berikut.

$$Kemiringan = \frac{\sum(X - \bar{X})^3}{ns^3} - 3$$

Sementara, dalam *Microsoft excel*, nilai kemiringan dihitung dengan rumus sebagai berikut.

$$Kemiringan = \frac{n}{(n-1)(n-2)}\left(\frac{\sum(X - \bar{X})^3}{s^3}\right)$$

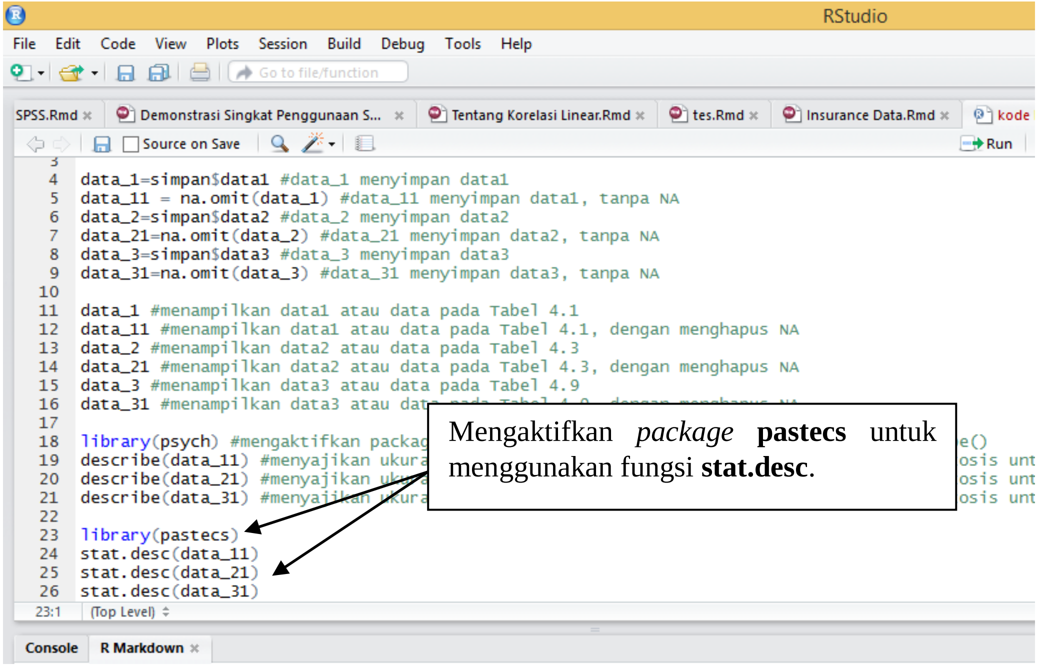
Pada hasil R Gambar 4.11, nilai kurtosis dihitung dengan rumus sebagai berikut.

$$Kurtosis = \frac{\sum(X - \bar{X})^4}{ns^4} - 3$$

Sementara, dalam *Microsoft excel*, nilai kurtosis dihitung dengan rumus sebagai berikut.

$$Kurtosis = \left\{\frac{(n)(n+1)\sum(X - \bar{X})^4}{(n-1)(n-2)(n-3)s^4}\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Pada Gambar 4.12, mengaktifkan *package* **pastecs** untuk menggunakan fungsi **stat.desc**. Hasilnya diperlihatkan pada Gambar 4.13. Pada penggunaan fungsi **stat.desc**, menyajikan beberapa nilai statistik, seperti *variance*, jumlah keseluruhan (*sum*), *standard error mean* (*SE.mean*), dan koefisien variasi, yang **sebelumnya tidak tersaji** pada penggunaan fungsi **describe**.

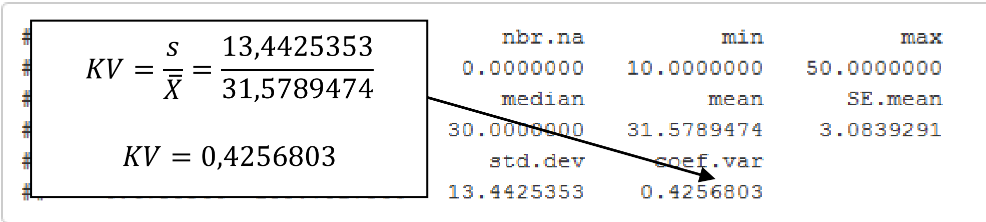


Gambar 4.12

```
stat.desc(data_11)
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##    19.0000000     0.0000000     0.0000000     1.0000000    18.0000000
##         range           sum        median          mean       SE.mean
##    17.0000000   182.0000000    10.0000000     9.5789474     1.1928535
## CI.mean.0.95           var       std.dev      coef.var
##     2.5060921    27.0350877     5.1995276     0.5428078
```

```
stat.desc(data_21)
```



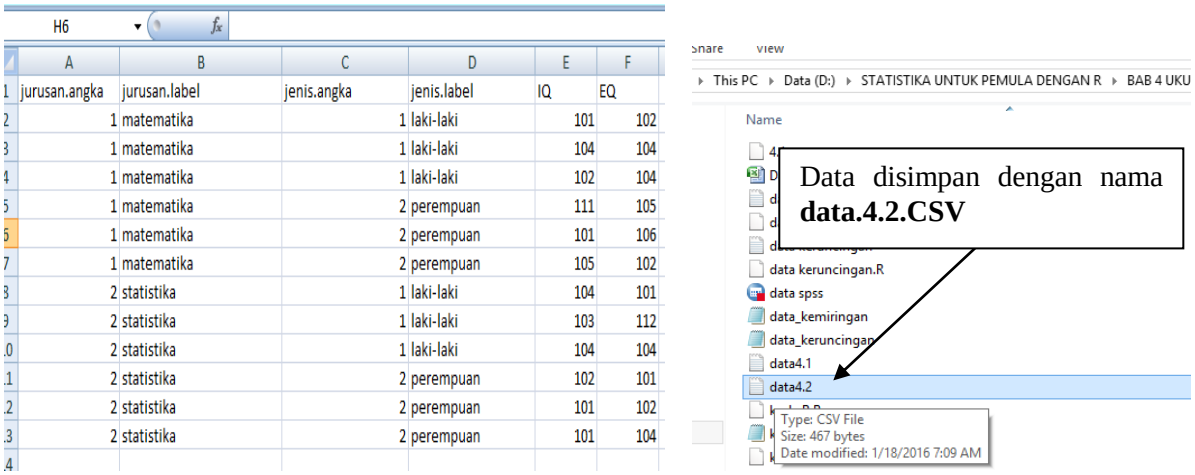

```
stat.desc(data_31)
```

```
##                        x
## nbr.val       15.0000000
## nbr.null       0.0000000
## nbr.na         0.0000000
## min            1.0000000
## max            6.0000000
## range          5.0000000
## sum           54.0000000
## median         3.0000000
## mean           3.6000000
## SE.mean        0.3754363
## CI.mean.0.95   0.8052307
## var            2.1142857
## std.dev        1.4540584
## coef.var       0.4039051
```

Gambar 4.13

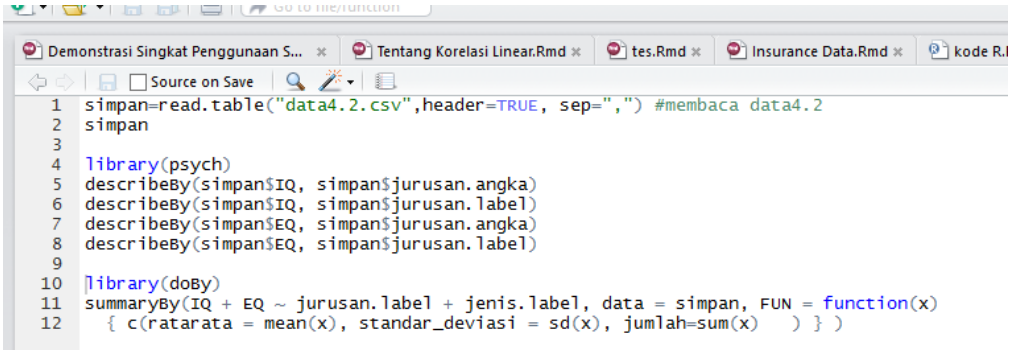
Aplikasi dalam R (Data Berkelompok)

Andaikan diberikan data, seperti pada Gambar 4.14. Data tersebut disimpan dengan nama **data4.2.CSV** (perhatikan Gambar 4.14).



Gambar 4.14

Berikut disajikan kode R (Gambar 4.15).



Gambar 4.15

Untuk mengeksekusi kode R pada Gambar 4.15, pilih *Compile* dan pilih format *output* HTML.

Interpretasi Kode R

Berikut disajikan kode R, beserta *output* dari kode R tersebut.

[1] simpan=read.table("data4.2.csv",header=TRUE, sep=",") #membaca data4.2 Simpan

[2] library(psych)
describeBy(simpan$IQ, simpan$jurusan.angka)
describeBy(simpan$IQ, simpan$jurusan.label)
describeBy(simpan$EQ, simpan$jurusan.angka)
describeBy(simpan$EQ, simpan$jurusan.label)

[3] library(doBy)
summaryBy(IQ + EQ ~ jurusan.label + jenis.label, data = simpan, FUN = function(x) { c(ratarata = mean(x), standar_deviasi = sd(x), jumlah=sum(x)) })

Gambar 4.16 merupakan hasil dari kode R [1]. Kode R [1] dapat diartikan variabel **simpan** ditugaskan untuk menyimpan data pada variabel **jurusan.angka**, **jurusan.label**, **jenis.angka**, **jenis.label**, **IQ**, dan **EQ** dalam *file* **data4.2.csv**. Kemudian menampilkan data yang tersimpan pada variabel **simpan**.

```
simpan=read.table("data4.2.csv",header=TRUE, sep=",") #membaca data4.2
simpan
```

```
##    jurusan.angka jurusan.label jenis.angka jenis.label  IQ  EQ
## 1              1    matematika           1   laki-laki 101 102
## 2              1    matematika           1   laki-laki 104 104
## 3              1    matematika           1   laki-laki 102 104
## 4              1    matematika           2   perempuan 111 105
## 5              1    matematika           2   perempuan 101 106
## 6              1    matematika           2   perempuan 105 102
## 7              2     statistika          1   laki-laki 104 101
## 8              2     statistika          1   laki-laki 103 112
## 9              2     statistika          1   laki-laki 104 104
## 10             2     statistika          2   perempuan 102 101
## 11             2     statistika          2   perempuan 101 102
## 12             2     statistika          2   perempuan 101 104
```

Gambar 4.16

```
library(psych)
describeBy(simpan$IQ, simpan$jurusan.angka)
```

```
## group: 1
##   vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## 1    1 6  104 3.79    103     104 2.97 101 111    10 0.86    -0.93 1.55
## ------------------------------------------------------------
## group: 2
##   vars n  mean   sd median trimmed  mad min max range skew kurtosis   se
## 1    1 6 102.5 1.38  102.5   102.5 2.22 101 104     3    0    -2.06 0.56
```

```
describeBy(simpan$IQ, simpan$jurusan.label)
```

```
## group: matematika
##   vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## 1    1 6  104 3.79    103     104 2.97 101 111    10 0.86    -0.93 1.55
## ------------------------------------------------------------
## group: statistika
##   vars n  mean   sd median trimmed  mad min max range skew kurtosis   se
## 1    1 6 102.5 1.38  102.5   102.5 2.22 101 104     3    0    -2.06 0.56
```

```
describeBy(simpan$EQ, simpan$jurusan.angka)
```

```
## group: 1
##   vars n   mean  sd median trimmed  mad min max range  skew kurtosis   se
## 1    1 6 103.83 1.6    104  103.83 2.22 102 106     4 -0.02    -1.82 0.65
## ------------------------------------------------------------
## group: 2
##   vars n mean   sd median trimmed  mad min max range skew kurtosis   se
## 1    1 6  104 4.15    103     104 2.22 101 112    11 1.05    -0.59 1.69
```

Gambar 4.17

Gambar 4.17 merupakan hasil dari kode R [2]. Kode R **library(psych)** dapat diartikan untuk mengaktifkan *package* **psych**. Pengaktifan *package* **psych** dimaksudkan untuk penggunaan

fungsi **describeBy**. Kode R **describeBy(simpan$IQ, simpan$jurusan.angka)** dapat diartikan akan disajikan nilai-nilai statistik, seperti rata-rata (*mean*), median, *range*, dan seterusnya, berdasarkan variabel **IQ** untuk kelompok pada variabel **jurusan.angka**. Kode R **describeBy(simpan$IQ, simpan$jurusan.label)** dapat diartikan akan disajikan nilai-nilai statistik, seperti rata-rata (*mean*), median, *range*, dan seterusnya, berdasarkan variabel **IQ** untuk kelompok pada variabel **jurusan.label**. Kode R **describeBy(simpan$EQ, simpan$jurusan.label)** dapat diartikan akan disajikan nilai-nilai statistik, seperti rata-rata (*mean*), median, *range*, dan seterusnya, berdasarkan variabel **EQ** untuk kelompok pada variabel **jurusan.label**.

Gambar 4.18 merupakan hasil dari kode R [3]. Kode R **library(doBy)** dapat diartikan untuk mengaktifkan *package* **doBy**. Pengaktifan *package* **doBy** dimaksudkan untuk penggunaan fungsi **summaryBy**. Kode R **summaryBy(IQ + EQ ~ jurusan.label + jenis.label, data = simpan, FUN = function(x) { c(ratarata = mean(x), standar_deviasi = sd(x), jumlah=sum(x)) })** dapat diartikan akan disajikan nilai-nilai statistik, seperti rata-rata (*mean*), median, *range*, dan seterusnya, berdasarkan variabel **IQ** dan **EQ**, untuk kombinasi kategori dari variabel **jurusan.label** dan **jenis.label**.

```
library(doBy)
```

```
## Loading required package: survival
```

```
summaryBy(IQ + EQ ~ jurusan.label + jenis.label, data = simpan, FUN = function(x)
  { c(ratarata = mean(x), standar_deviasi = sd(x), jumlah=sum(x)    ) } )
```

```
##   jurusan.label jenis.label IQ.ratarata IQ.standar_deviasi IQ.jumlah
## 1    matematika   laki-laki    102.3333          1.5275252       307
## 2    matematika   perempuan    105.6667          5.0332230       317
## 3    statistika   laki-laki    103.6667          0.5773503       311
## 4    statistika   perempuan    101.3333          0.5773503       304
##   EQ.ratarata EQ.standar_deviasi EQ.jumlah
## 1    103.3333           1.154701       310
## 2    104.3333           2.081666       313
## 3    105.6667           5.686241       317
## 4    102.3333           1.527525       307
```

Gambar 4.18

Referensi

1. Agresti, A. dan B. Finlay. 2009. *Statistical Methods for the Social Sciences, 4th Edition*. United States of America: Prentice Hall.

2. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

3. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

4. Johnson, R.A. dan G.K. Bhattacharyya. 2011. *Statistics, Principles and Methods, 6th Edition*. John Wiley and Sons, Inc.

5. Mann, P. S. dan C.J. Lacke. 2011. *Introductory Statistics, International Student Version, 7th Edition*. Asia: John Wiley & Sons, Inc.

6. Montgomery, D. C. dan G. C. Runger. 2011. *Applied Statistics and Probability for Engineers, 5th Edition*. United States of America: John Wiley & Sons, Inc.

7. Ott, R.L. dan M. Longnecker. 2001. *An Introduction to Statistical Methods and Data Analysis, 5th Edition*. United States of America: Duxbury.

8. Smidth, R. K. dan D. H. Sanders. 2000. *Statistics a First Course, 6th Edition*. United States of America: McGraw-Hill Companies.

9. Spiegel, M.R. dan L.J. Stephens. 2008. *Statistics, 4th Edition*. McGraw-Hill.

10. https://cran.r-project.org/web/packages/psych/psych.pdf

11. https://cran.r-project.org/web/packages/pastecs/pastecs.pdf

12. https://cran.r-project.org/web/packages/doBy/doBy.pdf

BAB 5

DISTRIBUSI SAMPLING

Distribusi Populasi (Population Distribution)

Distribusi populasi dapat diartikan sebagai distribusi probabilitas dari data populasi. Andaikan dalam suatu kelas hanya terdiri lima mahasiswa jurusan matematika. Berikut disajikan nilai ujian matakuliah kalkulus dari lima mahasiswa tersebut.

70, 75, 80, 80, 90

Andaikan X menyatakan nilai ujian matakuliah kalkulus dan $P(X = x)$ atau $f(x)$ menyatakan probabilitas dari suatu nilai ujian matakuliah kalkulus. Berikut disajikan distribusi probabilitas dari data populasi nilai ujian matakuliah kalkulus (Tabel 5.1).

Tabel 5.1 Distribusi Probabilitas dari Data Populasi Nilai Ujian Kalkulus

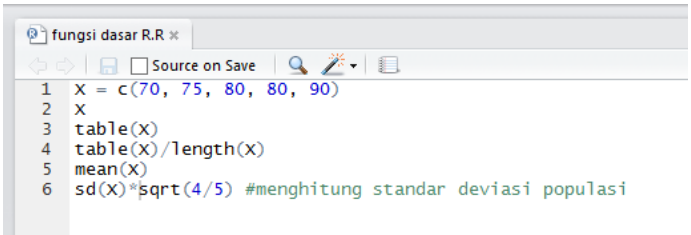
X	$P(X = x)$
70	0.2
75	0.2
80	0.4
90	0.2
$\sum P(X = x) = \sum f(x) = 1$	

Nilai rata-rata dan standar deviasi berdasarkan data pada Tabel 5.1 dihitung sebagai berikut.

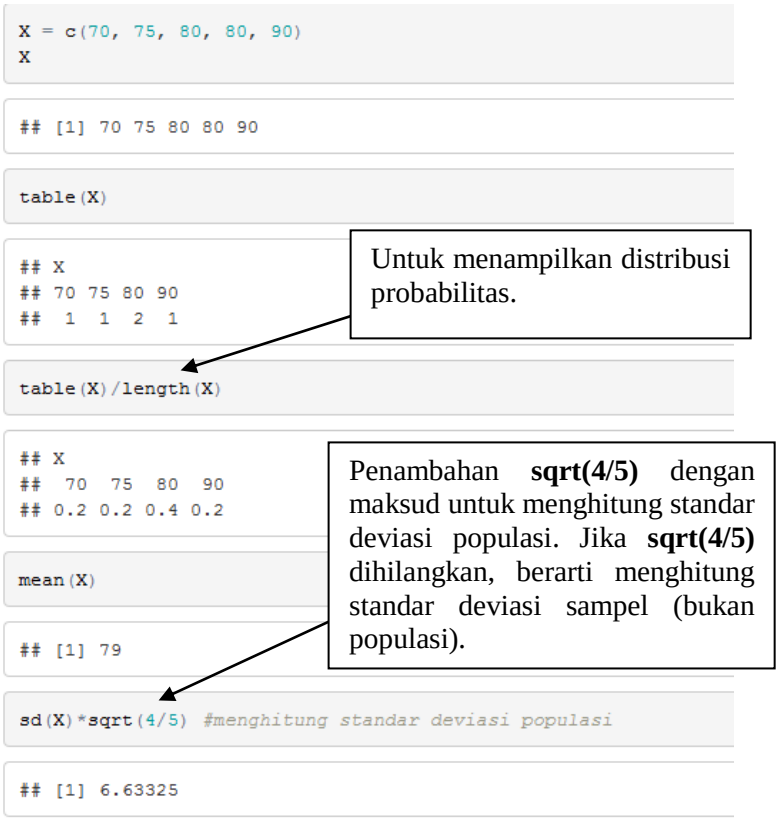
$$\mu = \frac{70 + 75 + 80 + 80 + 90}{5} = 79$$

$$\sigma = \sqrt{\frac{(70 - 79)^2 + (75 - 79)^2 + \cdots + (90 - 79)^2}{5}} = 6{,}633$$

Perhatikan bahwa μ dan σ merupakan nilai-nilai parameter populasi. Parameter dapat diartikan sebagai suatu nilai atau ukuran yang dihitung berdasarkan populasi. Gambar 5.1 dan Gambar 5.2 merupakan ilustrasi dalam R.



Gambar 5.1



Gambar 5.2

Distribusi Sampling Rata-Rata Sampel $\bar{X}$ (Sampling Distribution of $\bar{X}$)

Berbeda dengan statistika deskriptif yang rangkaian pengerjaannya meliputi mengorganisasi (*organizing*), menampilkan (*displaying*), dan menjelaskan data dengan menggunakan tabel, grafik, serta ukuran-ukuran seperti rata-rata, median, serta modus, pada statistika inferensi sampai pada tahap pengambilan keputusan atau prediksi mengenai populasi berdasarkan sampel yang diteliti. Konsep mengenai **distribusi sampling** memberikan teori yang penting untuk membuat prosedur-prosedur statistik inferensi. Daniel (2005:129) menyatakan sebagai berikut.

"*Sampling distributions serve two purposes: (1) they allow us to answer probability questions about sample statistics, and (2) they provide the necessary theory for making statistical inference procedures valid*".

Nilai dari parameter suatu populasi bersifat konstan. Dalam hal ini, untuk setiap data populasi hanya memiliki satu nilai rata-rata populasi μ. Namun hal ini belum tentu berlaku untuk rata-rata sampel $\bar{X}$. Sampel-sampel yang ditarik dari populasi yang sama dan dengan ukuran yang sama dapat menghasilkan nilai rata-rata sampel yang berbeda-beda. Jadi, nilai rata-rata sampel bergantung pada nilai-nilai yang berada dalam sampel tersebut. **Oleh karena itu, rata-rata sampel $\bar{X}$ merupakan variabel acak (*random variable*)**. Sebagaimana pada variabel acak, **maka rata-rata sampel $\bar{X}$ memiliki distribusi probabilitas**. Distribusi probabilitas $\bar{X}$ sering disebut dengan istilah **distribusi sampling dari $\bar{X}$**. Ukuran-ukuran statistik lainnya seperti median, modus, dan standar deviasi juga memiliki distribusi sampling (Mann dan Lacke, 2011:302).

Pada pembahasan sebelumnya mengenai “Distribusi Probabilitas”, diketahui data populasi sebagai berikut.

70, 75, 80, 80, 90

Andaikan masing-masing nilai diberi kode huruf sebagai berikut.

V = 70, W = 75, X = 80, Y = 80, dan Z = 90

Maka, V, W, X, Y, dan Z merupakan kode-kode huruf yang menyatakan kelima nilai ujian matakuliah kalkulus. Kemudian misalkan akan diambil sampel yang terdiri tiga nilai tanpa pengembalian (*without replacement*). Maka banyaknya kemungkinan sampel yang terambil sebagai berikut.

$$C_3^5 = \frac{5!}{(5-3)!\,3!} = \frac{5.4.3.2.1}{(2.1)(3.2.1)} = 10\ kemungkinan\ sampel$$

VWX, VWY, VWZ, VXY, VXZ, VYZ, WXY, WXZ, WYZ, XYZ

Tabel 5.3 Sampel-Sampel yang Mungkin Terambil beserta Nilai Rata-Rata

Sampel	Nilai-Nilai dalam Sampel			$\bar{X}$
VWX	70	75	80	75
VWY	70	75	80	75
VWZ	70	75	90	78.33
VXY	70	80	80	76.67
VXZ	70	80	90	80
VYZ	70	80	90	80
WXY	75	80	80	78.33
WXZ	75	80	90	81.67
WYZ	75	80	90	81.67
XYZ	80	80	90	83.33

Perhatikan bahwa terdapat 10 kemungkinan sampel. Sampel VWX berarti mengandung nilai 70, 75, dan 80, sampel WYZ berarti mengandung nilai 75, 80, dan 90, dan seterusnya. Tabel 5.3 menyajikan sampel-sampel yang mungkin terambil beserta penghitungan nilai rata-rata. Berdasarkan Tabel 5.3, selanjutnya dibentuk tabel distribusi frekuensi dan frekuensi relatif berdasarkan nilai rata-rata sampel (Tabel 5.4). Tabel 5.5 menyajikan distribusi sampling dari rata-rata sampel $\bar{X}$ berdasarkan data pada Tabel 5.3.

Tabel 5.5 menyajikan distribusi probabilitas dari rata-rata sampel $\bar{X}$. Sebagai contoh probabilitas untuk memperoleh sampel yang memiliki nilai rata-rata 76,67 sebesar 0,2. Atau dapat dinyatakan sebagai berikut.

$$P(\bar{X} = 81.67) = 0.20$$

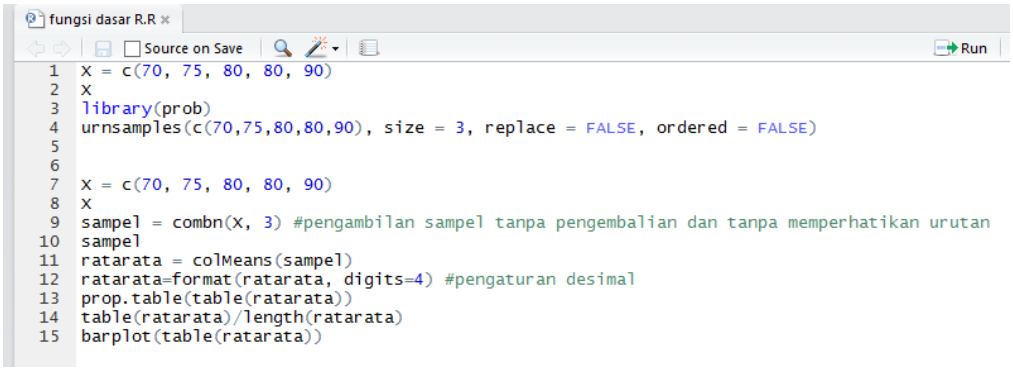
Tabel 5.4 Distribusi Frekuensi dan Frekuensi Relatif Berdasarkan Nilai Rata-Rata Sampel

$\bar{X}$	Frekuensi	Frekuensi Relatif
75	2	0.2
76.67	1	0.1
78.33	2	0.2
80	2	0.2
81.67	2	0.2
83.33	1	0.1
Jumlah	10	1

Tabel 5.5 Distribusi Sampling dari $\bar{X}$ dengan Ukuran Sampel sebanyak 3

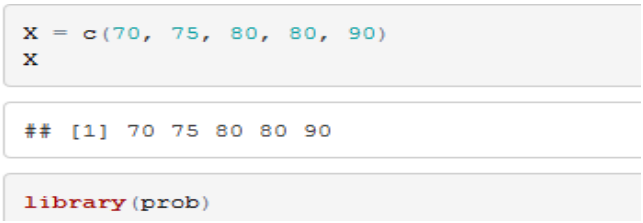
$\bar{X}$	$P(\bar{X} = \bar{x}) = f(\bar{x})$
75	0.2
76.67	0.1
78.33	0.2
80	0.2
81.67	0.2
83.33	0.1
$\sum P(\bar{X} = \bar{x}) = 1$	

Berikut diberikan ilustrasi dalam R.



Gambar 5.3

Pada Gambar 5.3 mengaktifkan *package* **prob** (kode R baris 3) dengan maksud untuk menggunakan fungsi **urnsamples**.



Gambar 5.4

```
urnsamples(c(70,75,80,80,90), size = 3, replace = FALSE, ordered = FALSE)
```

```
##    X1 X2 X3
## 1  70 75 80
## 2  70 75 80
## 3  70 75 90
## 4  70 80 80
## 5  70 80 90
## 6  70 80 90
## 7  75 80 80
## 8  75 80 90
## 9  75 80 90
## 10 80 80 90
```

Gambar 5.5

Pada Gambar 5.5, penggalan kode R **replace = FALSE** berarti pengambilan sampel tanpa pengembalian, serta pada penggalan kode R **ordered = FALSE** berarti tanpa memperhatikan urutan.

```
sampel = combn(X, 3) #pengambilan sampel tanpa pengembalian dan tanpa memperhatikan urutan
sampel
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]   70   70   70   70   70   70   75   75   75    80
## [2,]   75   75   75   80   80   80   80   80   80    80
## [3,]   80   80   90   80   90   90   80   90   90    90
```

Gambar 5.6

Pada Gambar 5.6 menyajikan alternatif kode R (dari yang sebelumnya) untuk menampilkan seluruh kemungkinan sampel yang mungkin terambil. Pada Gambar 5.6 menggunakan fungsi **combn** (*combination*).

```
ratarata = colMeans(sampel)
ratarata=format(ratarata, digits=4) #pengaturan desimal
prop.table(table(ratarata))
```

```
## ratarata
## 75.00 76.67 78.33 80.00 81.67 83.33
##   0.2   0.1   0.2   0.2   0.2   0.1
```

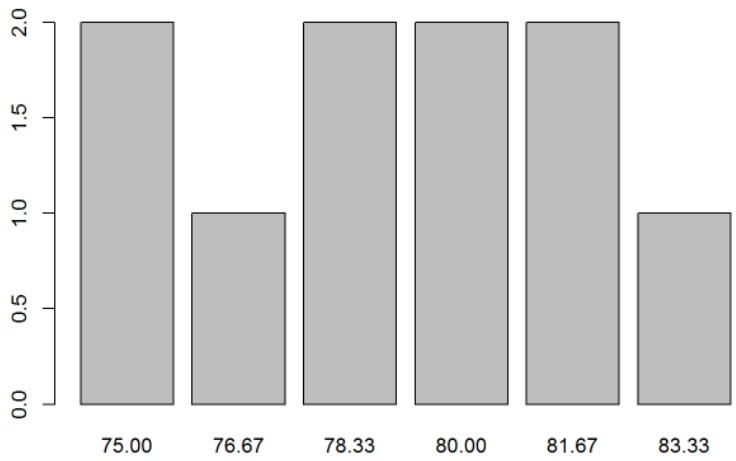
```
table(ratarata)/length(ratarata)
```

```
## ratarata
## 75.00 76.67 78.33 80.00 81.67 83.33
##   0.2   0.1   0.2   0.2   0.2   0.1
```

Gambar 5.7

Pada Gambar 5.7 menyajikan distribusi probabilitas dari rata-rata sampel $\bar{X}$. Pada Gambar 5.8 menyajikan grafik batang yang menyajikan frekuensi dari setiap nilai rata-rata sampel $\bar{X}$. Berdasarkan Gambar 5.8, nilai rata-rata 75 sebanyak 2, nilai rata-rata 76,67 sebanyak 1, dan seterusnya.

```
barplot(table(ratarata))
```



Gambar 5.8

Rata-Rata dari Distribusi Sampling Rata-Rata Sampel $\bar{X}$

Rata-rata dari distribusi sampling $\bar{X}$ (*mean of the sampling distribution of* $\bar{X}$) atau rata-rata dari $\bar{X}$ dilambangkan dengan $\mu_{\bar{X}}$. Berdasarkan Tabel 5.3, berikut akan dihitung rata-rata dari distribusi sampling $\bar{X}$ serta rata-rata populasinya.

$$\mu_{\bar{X}} = \frac{75 + 75 + 78{,}33 + \cdots + 83{,}33}{10} = 79$$

$$\mu = \frac{70 + 75 + 80 + 80 + 90}{5} = 79$$

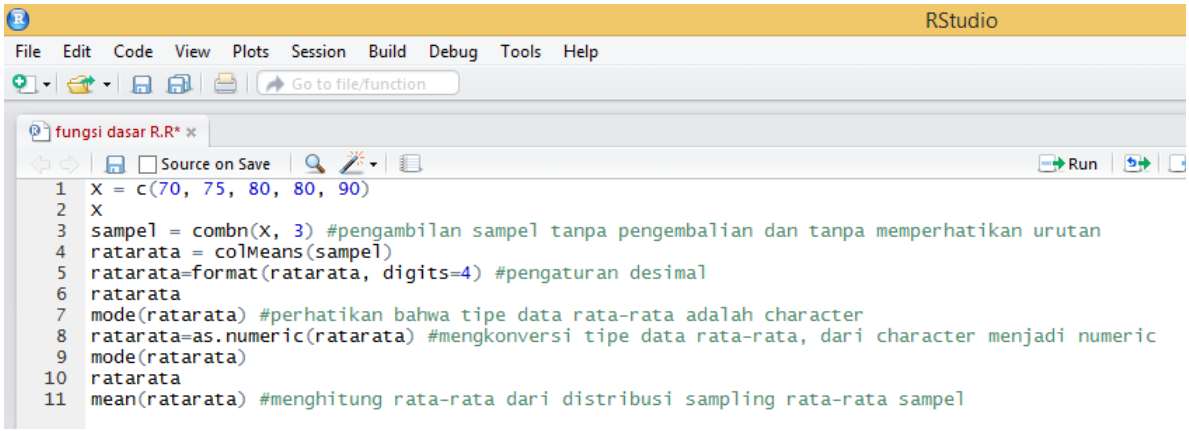
Perhatikan bahwa berdasarkan perhitungan diperoleh $\mu_{\bar{X}} = 79$ dan $\mu = 79$. Mann dan Lacke (2011:307) menyatakan sebagai berikut.

"*The mean of the sampling distribution of* $\bar{X}$ ***is always equal*** *to the mean of the population. Thus,* $\mu_{\bar{X}} = \mu$".

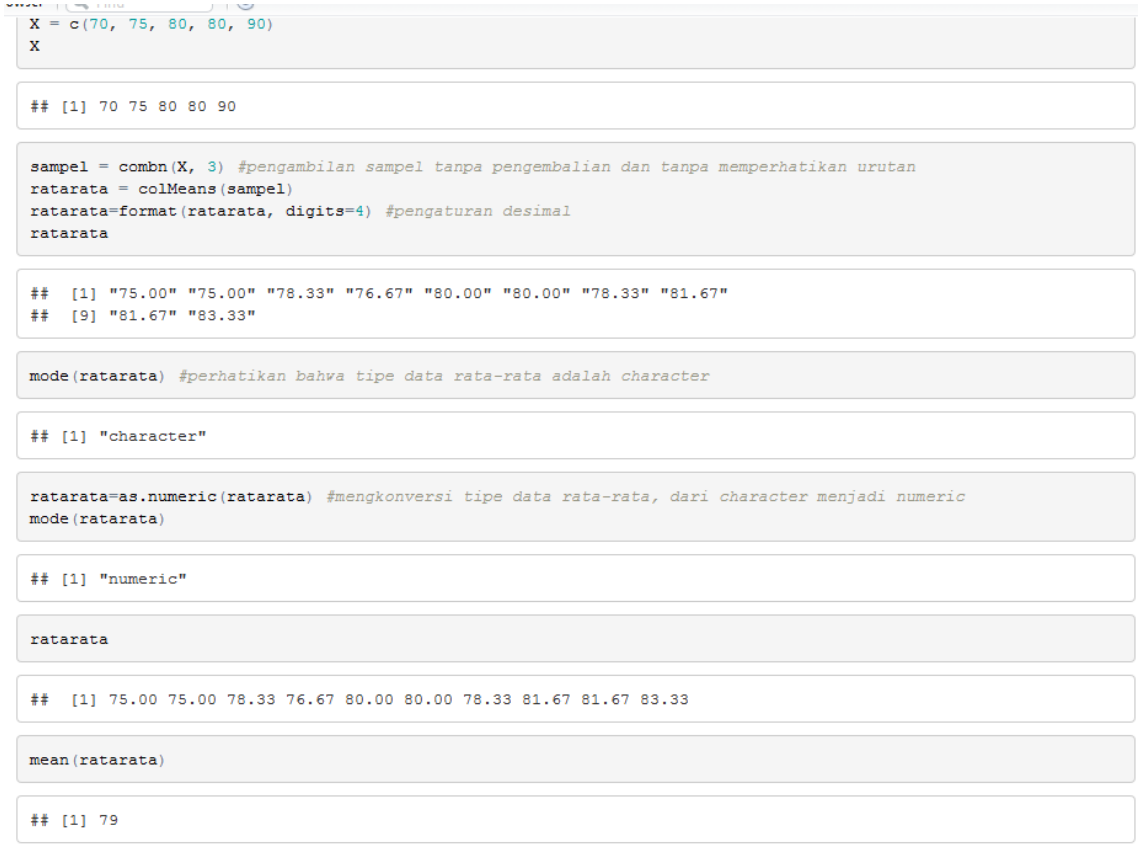
Rata-rata sampel $\bar{X}$ disebut juga sebagai *estimator* atau penduga terhadap rata-rata populasi μ. Suatu statistik dikatakan sebagai estimator tak-bias atau *unbiased estimator* jika nilai rata-rata dari distribusi sampling statistik tersebut sama dengan nilai parameter tertentu. Perhatikan bahwa statistik rata-rata sampel $\bar{X}$ merupakan estimator tak-bias dari parameter rata-rata populasi (μ), karena nilai rata-rata dari distribusi sampling rata-rata $\bar{X}$ selalu sama dengan rata-rata populasi, yakni

$$\mu_{\bar{X}} = \mu.$$

Berikut diberikan ilustrasi dalam R.

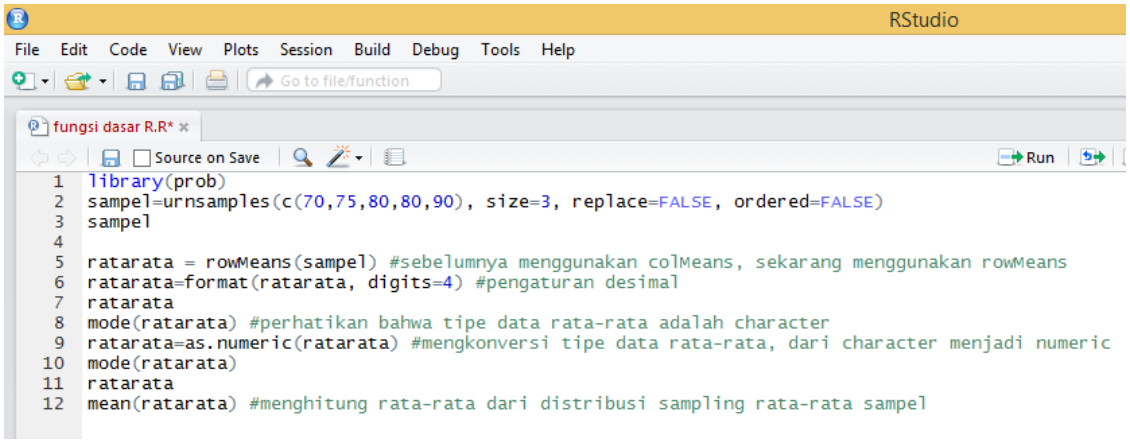


Gambar 5.9



Gambar 5.10

Berdasarkan Gambar 5.9, kode R pada baris 7 bertujuan untuk mengetahui tipe atau jenis data dari variabel **ratarata**. Sementara kode R pada baris 8 bertujuan untuk mengkonversi jenis data variabel **ratarata**, dari *character* menjadi *numeric*. Setelah dikonversi menjadi *numeric*, barulah bisa dihitung nilai rata-rata dari distribusi sampling rata-rata sampel (kode R pada baris 11). Berikut alternatif kode R untuk memperoleh rata-rata dari distribusi sampling rata-rata sampel (perhatikan Gambar 5.11 sampai dengan Gambar 5.13).



Gambar 5.11

```
sampel=urnsamples(c(70,75,80,80,90), size=3, replace=FALSE, ordered=FALSE)
sampel
```

```
##    X1 X2 X3
## 1  70 75 80
## 2  70 75 80
## 3  70 75 90
## 4  70 80 80
## 5  70 80 90
## 6  70 80 90
## 7  75 80 80
## 8  75 80 90
## 9  75 80 90
## 10 80 80 90
```

```
ratarata = rowMeans(sampel)
ratarata=format(ratarata, digits=4) #pengaturan desimal
ratarata
```

```
##  [1] "75.00" "75.00" "78.33" "76.67" "80.00" "80.00" "78.33" "81.67"
##  [9] "81.67" "83.33"
```

```
mode(ratarata) #perhatikan bahwa tipe data rata-rata adalah character
```

```
## [1] "character"
```

```
ratarata=as.numeric(ratarata) #mengkonversi tipe data rata-rata, dari character menjadi numeric
mode(ratarata)
```

```
## [1] "numeric"
```

```
ratarata
```

```
##  [1] 75.00 75.00 78.33 76.67 80.00 80.00 78.33 81.67 81.67 83.33
```

```
mean(ratarata) #menghitung rata-rata dari distribusi sampling rata-rata sampel
```

```
## [1] 79
```

Gambar 5.12

Standar Deviasi dari Distribusi Sampling Rata-Rata Sampel $\bar{X}$

Diketahui pada pembahasan sebelumnya bahwa rata-rata dari distribusi sampling rata-rata $\bar{X}$ dilambangkan dengan simbol $\mu_{\bar{X}}$, sedangkan rata-rata populasi dilambangkan dengan simbol μ. Standar deviasi dari distribusi sampling rata-rata $\bar{X}$ dilambangkan dengan simbol $\sigma_{\bar{X}}$, sedangkan standar deviasi populasi dilambangkan dengan simbol σ. Pada pembahasan sebelumnya diketahui bahwa rata-rata dari distribusi sampling rata-rata $\bar{X}$ sama dengan rata-rata populasi μ, yakni

$$\mu_{\bar{X}} = \mu.$$

Namun pada standar deviasi dari distribusi sampling rata-rata $\bar{X}$ tidak sama dengan standar deviasi populasi (kecuali jika $n = 1$). Sebagai contoh untuk kasus $n = 1$, misalkan suatu populasi terdiri dari tiga angka, yakni 1, 2, 3. Misalkan dari populasi yang terdiri dari tiga angka tersebut, akan diambil sampel yang terdiri atas satu angka. Maka sampel-sampel yang mungkin adalah

1 2 3.

Diketahui rata-rata dari setiap sampel tersebut adalah

1 2 3.

Maka rata-rata dari distribusi sampling rata-rata $\bar{X}$ tersebut adalah

$$\mu_{\bar{X}} = \frac{1+2+3}{3} = 2.$$

Sedangkan standar deviasi dari distribusi sampling rata-rata $\bar{X}$ tersebut adalah

$$\sigma_{\bar{X}} = \sqrt{\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3}} = 0{,}8165,$$

yang mana

$$\sigma_{\bar{X}} = \sigma \quad (ketika\ n = 1).$$

Mann dan Lacke (2011:307) menyatakan rumus

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

berlaku ketika paling tidak memenuhi salah satu dari kriteria sebagai berikut.

- ⇨ Jumlah elemen dalam populasi berhingga (*finite*) dan pengambilan elemen untuk sampel dari suatu populasi dengan pengembalian (*with replacement*).

- ⇨ Jumlah elemen dalam populasi tak berhingga (*infinite*) dan pengambilan elemen untuk sampel dari suatu populasi tanpa pengembalian (*without replacement*).

Namun kriteria-kriteria tersebut dapat diganti ketika ukuran sampel kecil (*sample size is small*) dalam perbandingannya terhadap ukuran populasi (*in comparison to the population size*). Ukuran sampel dapat dipandang (*is considered*) kecil dalam perbandingannya terhadap ukuran populasi ketika ukuran sampel lebih kecil atau sama dengan 5% dari ukuran populasi, yakni

$$\frac{n}{N} \leq 0{,}05,$$

dengan n merupakan ukuran sampel dan N ukuran populasi. Namun ketika tidak terpenuhi, maka penghitungan $\sigma_{\bar{X}}$ dihitung dengan rumus

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

di mana

$$\sqrt{\frac{N-n}{N-1}}$$

merupakan faktor koreksi populasi berhingga (Mann dan Lacke, 2011:307).

Berikut diberikan contoh kasus untuk perhitungan standar deviasi dari distribusi sampling $\bar{X}$ dengan rumus $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. Misalkan suatu populasi terdiri dari tiga angka, yakni 1, 2, 3. Misalkan dari populasi yang terdiri dari tiga angka tersebut, akan diambil sampel yang terdiri atas dua angka dengan pengembalian (*with replacement*). Maka sampel-sampel yang mungkin adalah sebagai berikut.

(1,1)	(1,2)	(1,3)
(2,1)	(2,2)	(2,3)
(3,1)	(3,2)	(3,3)

Perhatikan bahwa karena jumlah elemen dalam populasi berhingga, yakni tiga, dan pengambilan elemen sampel dengan pengembalian, maka standar deviasi dari distribusi sampling rata-rata $\bar{X}$ dihitung dengan rumus sebagai berikut.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Hasil perhitungan rata-rata untuk setiap sampel sebagai berikut.

1	1,5	2
1,5	2	2,5
2	2,5	3

Maka rata-rata dari distribusi sampling rata-rata $\bar{X}$ tersebut adalah

$$\mu_{\bar{X}} = \frac{1 + 1{,}5 + 2 + 1{,}5 + 2 + 2{,}5 + 2 + 2{,}5 + 3}{9} = \frac{18}{9} = 2.$$

Berikut perhitungan standar deviasi dari distribusi sampling rata-rata $\bar{X}$.

$$\begin{matrix} (1-2)^2 & (1{,}5-2)^2 & (2-2)^2 \\ (1{,}5-2)^2 & (2-2)^2 & (2{,}5-2)^2 \\ (2-2)^2 & (2{,}5-2)^2 & (3-2)^2 \end{matrix}$$

Maka diperoleh hasil sebagai berikut.

$$\begin{matrix} 1 & 0{,}25 & 0 \\ 0{,}25 & 0 & 0{,}25 \\ 0 & 0{,}25 & 1 \end{matrix}$$

Sehingga

$$\sigma_{\bar{X}} = \sqrt{\frac{(1-2)^2 + (1{,}5-2)^2 + (2-2)^2 + \cdots + (3-2)^2}{9}}$$

$$\sigma_{\bar{X}} = \sqrt{\frac{1 + 0{,}25 + 0 + 0{,}25 + 0 + 0{,}25 + 0 + 0{,}25 + 1}{9}}$$

$$\sigma_{\bar{X}} = \sqrt{\frac{3}{9}} = \sqrt{0{,}3333333} = 0{,}57735$$

Perhatikan bahwa berdasarkan perhitungan sebelumnya diperoleh

$$\mu_{\bar{X}} = 2$$

$$\sigma_{\bar{X}} = 0{,}57735.$$

Diketahui

$$\mu = \frac{1+2+3}{3} = 2$$

$$\sigma = \sqrt{\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3}} = \sqrt{0{,}6666666} = 0{,}81649658.$$

Perhatikan bahwa

$$\sigma_{\bar{X}} \neq \sigma,$$

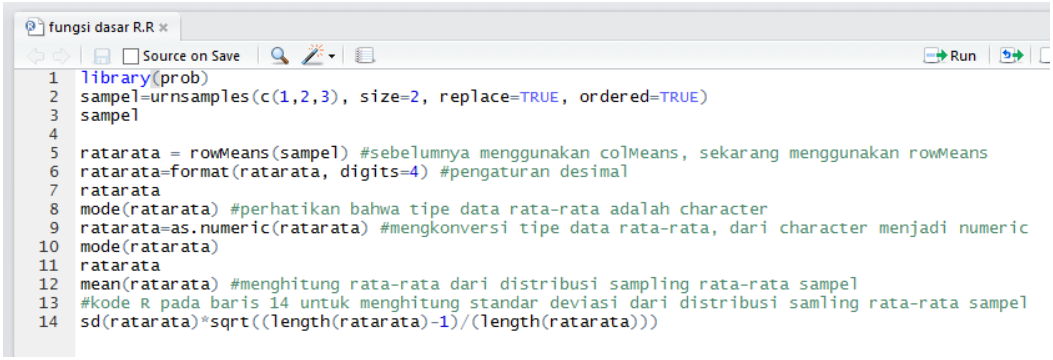
namun

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

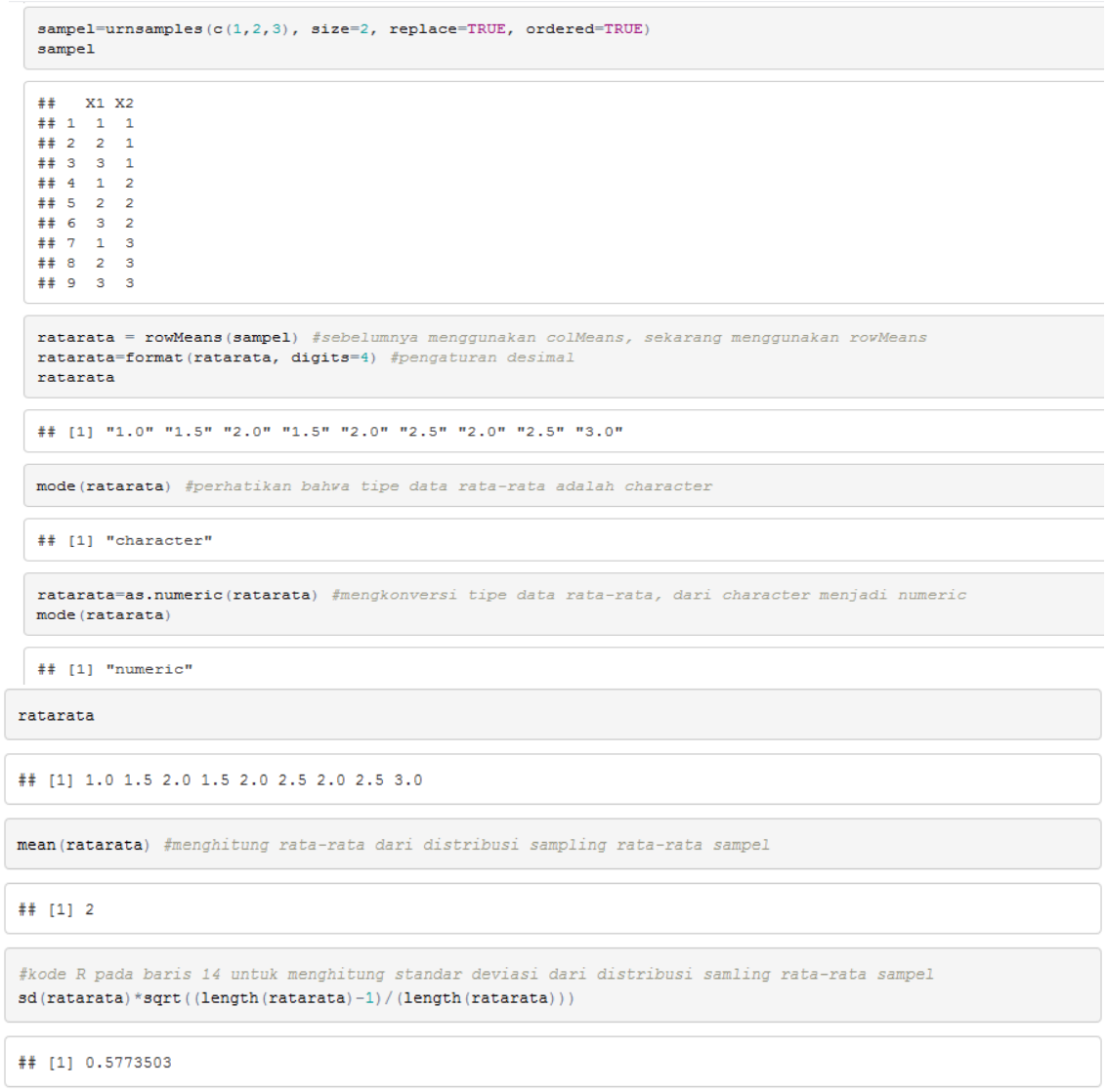
$$0{,}57735 = \frac{0{,}81649658}{\sqrt{2}}$$

$$0{,}57735 = 0{,}57735.$$

Berikut diberikan ilustrasi dalam R.



Gambar 5.13



Gambar 5.14

Berikut diberikan contoh kasus untuk perhitungan standar deviasi dari distribusi sampling $\bar{X}$ dengan rumus $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$. Misalkan suatu populasi terdiri dari tiga angka, yakni 1, 2, 3. Misalkan dari populasi yang terdiri dari tiga angka tersebut, akan diambil sampel yang terdiri atas dua angka tanpa pengembalian (*without replacement*). Maka sampel-sampel yang mungkin adalah

(1,2) (1,3) (2,3)

Perhatikan bahwa karena jumlah elemen dalam populasi berhingga, yakni tiga, namun pengambilan elemen sampel tanpa pengembalian, maka standar deviasi dari distribusi sampling rata-rata $\bar{X}$ dihitung dengan rumus sebagai berikut.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Diketahui rata-rata dari setiap sampel tersebut adalah

1,5 2 2,5,

sehingga rata-rata dari distribusi sampling rata-rata $(\bar{X})$ tersebut adalah

$$\mu_{\bar{X}} = \frac{1{,}5 + 2 + 2{,}5}{3} = \frac{6}{3} = 2.$$

Standar deviasi dari distribusi sampling rata-rata $\bar{X}$ tersebut adalah

$$\sigma_{\bar{X}} = \sqrt{\frac{(1{,}5-2)^2 + (2-2)^2 + (2{,}5-2)^2}{3}}$$

$$\sigma_{\bar{X}} = \sqrt{\frac{0{,}25 + 0 + 0{,}25}{3}}$$

$$\sigma_{\bar{X}} = \sqrt{\frac{0{,}5}{3}} = \sqrt{0{,}16666667} = 0{,}408248.$$

Perhatikan bahwa berdasarkan perhitungan sebelumnya diperoleh

$$\mu_{\bar{X}} = 2$$

$$\sigma_{\bar{X}} = 0{,}408248.$$

Diketahui

$$\mu = \frac{1 + 2 + 3}{3} = 2$$

$$\sigma = \sqrt{\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3}} = \sqrt{0{,}6666666} = 0{,}81649658.$$

Perhatikan bahwa

$$\sigma_{\bar{X}} \neq \sigma.$$

Namun

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$$0{,}408248 = \frac{0{,}81649658}{\sqrt{2}} \sqrt{\frac{3-2}{3-1}}$$

$$0{,}408248 = \frac{0{,}81649658}{\sqrt{2}} \sqrt{\frac{1}{2}}$$

$$0{,}408248 = \frac{0{,}81649658}{2}$$

$$0{,}408248 = 0{,}408248$$

Beberapa hal penting mengenai distribusi sampling rata-rata $\bar{X}$, yakni:

- ⇨ Nilai standar deviasi dari distribusi sampling rata-rata $\bar{X}$ lebih kecil dibandingkan nilai standar deviasi populasi, yakni $\sigma_{\bar{X}} < \sigma$ ketika n lebih besar dari 1. Hal ini terlihat jelas dari rumus

  $$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

  Sebagai contoh misalkan $\sigma = 20$ dan $n = 4$, maka

  $$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{4}} = 10.$$

  Perhatikan bahwa

  $$\sigma_{\bar{X}} < \sigma$$

  $$10 < 20.$$

- ⇨ Nilai dari standar deviasi dari distribusi sampling rata-rata $\bar{X}$ akan semakin mengecil ketika ukuran sampel n semakin besar.

  ketika n ↑ maka $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$↓

Sebagai contoh misalkan $\sigma = 20$ dan $n = 4$, maka

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{4}} = 10.$$

Untuk $n = 20$ maka

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{20}} = 4{,}4721.$$

Untuk $n = 50$ maka

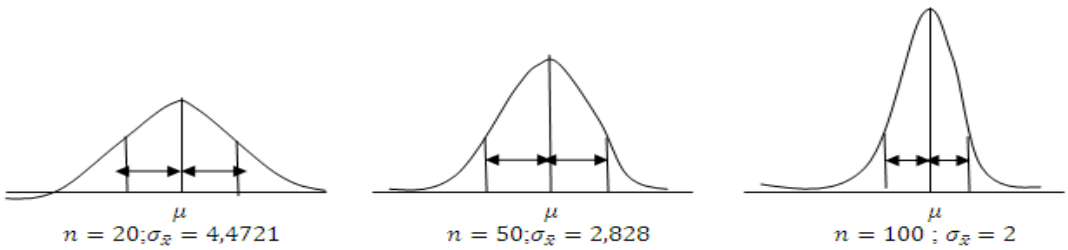
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{50}} = 2{,}828.$$

Untuk $n = 100$ maka

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{100}} = 2\,.$$

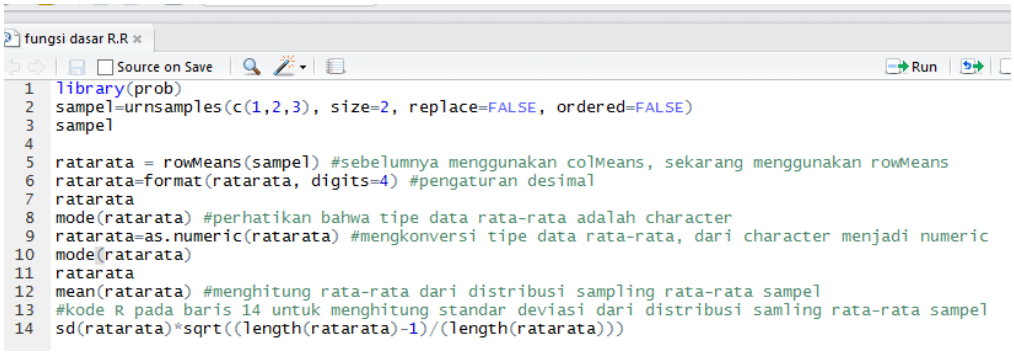
Perhatikan bahwa nilai $\sigma_{\bar{X}}$ semakin mengecil ketika ukuran sampel n semakin besar. Suatu statistik dikatakan estimator konsisten jika nilai standar deviasi dari distribusi sampling statistik tersebut semakin mengecil ketika ukuran sampel n semakin besar, sehingga statistik rata-rata $\bar{X}$ merupakan estimator konsisten dari parameter rata-rata μ (Mann dan Lacke, 2011:307)

Standar deviasi dari distribusi sampling rata-rata $\bar{X}$ merupakan suatu nilai yang mengukur pencaran atau sebaran dari rata-rata sampel dari distribusi sampling rata-rata $\bar{X}$ terhadap rata-rata populasinya μ. Semakin kecil nilai standar deviasi dari distribusi sampling rata-rata $\bar{X}$, maka rata-rata sampel dari distribusi sampling rata-rata $\bar{X}$ semakin mengumpul atau lebih dekat terhadap rata-rata populasinya μ. Pada pembahasan sebelumnya, diketahui untuk untuk $n = 20$ diperoleh $\sigma_{\bar{X}} = 4{,}4721$, untuk $n = 50$ diperoleh $\sigma_{\bar{X}} = 2{,}828$, dan untuk $n = 100$ diperoleh $\sigma_{\bar{X}} = 2$. Perhatikan ilustrasi gambar berikut ini (Gambar 5.15).



Gambar 5.15

Berikut diberikan ilustrasi dalam R (perhatikan Gambar 5.16 dan Gambar 5.17).



Gambar 5.16

```
sampel=urnsamples(c(1,2,3), size=2, replace=FALSE, ordered=FALSE)
sampel
```

```
##   X1 X2
## 1  1  2
## 2  1  3
## 3  2  3
```

```
ratarata = rowMeans(sampel) #sebelumnya menggunakan colMeans, sekarang menggunakan rowMeans
ratarata=format(ratarata, digits=4) #pengaturan desimal
ratarata
```

```
## [1] "1.5" "2.0" "2.5"
```

```
mode(ratarata) #perhatikan bahwa tipe data rata-rata adalah character
```

```
## [1] "character"
```

```
ratarata=as.numeric(ratarata) #mengkonversi tipe data rata-rata, dari character menjadi numeric
mode(ratarata)
```

```
## [1] "numeric"
```

```
ratarata
```

```
## [1] 1.5 2.0 2.5
```

```
mean(ratarata) #menghitung rata-rata dari distribusi sampling rata-rata sampel
```

```
## [1] 2
```

```
#kode R pada baris 14 untuk menghitung standar deviasi dari distribusi samling rata-rata sampel
sd(ratarata)*sqrt((length(ratarata)-1)/(length(ratarata)))
```

```
## [1] 0.4082483
```

Gambar 5.17

Bentuk Distribusi Sampling dari Rata-Rata Sampel $\bar{X}$

Mann dan Lacke (2011:310) menyatakan bentuk distribusi sampling dari rata-rata $\bar{X}$ berkenaan (*relates*) atas dua hal, yakni:

⇨ Sampel yang ditarik dari populasi yang berdistribusi normal.

⇨ Sampel yang ditarik dari populasi yang tidak berdistribusi normal.

Jika sampel-sampel yang ditarik berasal dari populasi yang berdistribusi normal dengan rata-rata dan standar deviasi masing-masing μ dan σ, maka:

⇨ Rata-rata distribusi sampling rata-rata $\bar{X}$ sama dengan rata-rata populasi, yakni

$$\mu_{\bar{X}} = \mu.$$

⇨ Standar deviasi distribusi sampling rata-rata $\bar{X}$ sama dengan $\frac{\sigma}{\sqrt{n}}$, dengan asumsi (*assuming*) $n/N \leq 0{,}05$.

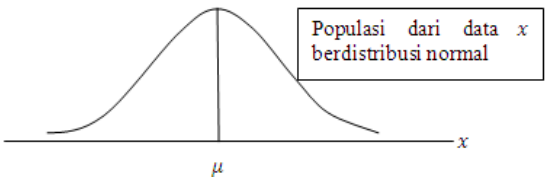
⇨ Bentuk dari distribusi sampling rata-rata $\bar{X}$ berbentuk normal, untuk berapapun ukuran sampel n.

Jadi, jika sampel-sampel yang ditarik berasal dari populasi yang berdistribusi normal dengan rata-rata adalah μ dan standar deviasi adalah σ, maka distribusi sampling dari rata-rata $\bar{X}$ juga terdistribusi secara normal, dengan rata-rata dan standar deviasi

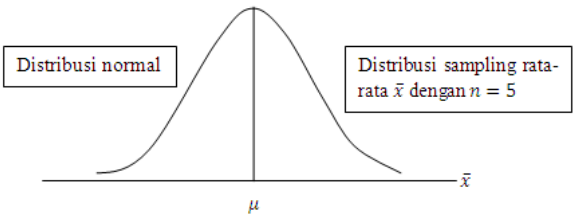
$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad ; \frac{n}{N} \leq 0{,}05.$$

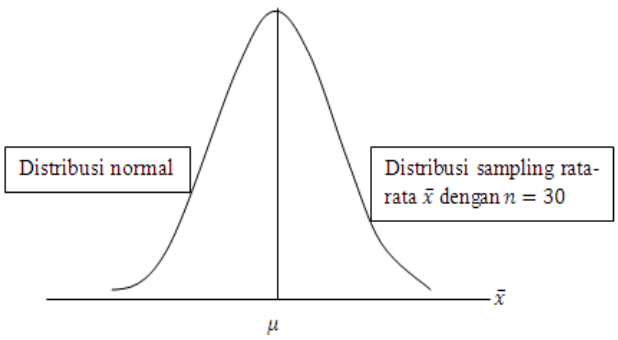
Perhatikan Gambar 5.18 hingga Gambar 5.21.



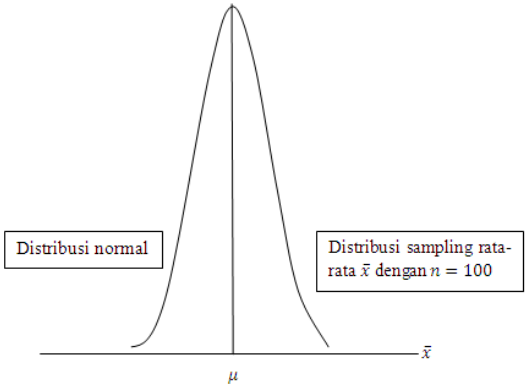
Gambar 5.18



Gambar 5.19



Gambar 5.20



Gambar 5.21

Perhatikan bahwa pada Gambar 5.18 menjelaskan data X berasal dari populasi berdistribusi normal. Pada Gambar 5.19 merupakan kurva dari distribusi sampling rata-rata $\bar{X}$ dengan $n = 5$. Pada Gambar 5.20 merupakan kurva dari distribusi sampling rata-rata $\bar{X}$ dengan $n = 30$.

Pada Gambar 5.21 merupakan kurva dari distribusi sampling rata-rata $\bar{X}$ dengan $n = 100$. Perhatikan bahwa karena sampel-sampel ditarik dari populasi yang berdistribusi normal, maka kurva dari distribusi sampling rata-rata $\bar{X}$ membentuk kurva normal (Gambar 5.19 sampai Gambar 5.21). Perhatikan bahwa standar deviasi dari distribusi sampling rata-rata $\bar{X}$ pada Gambar 5.20 lebih kecil daripada Gambar 5.19, standar deviasi dari distribusi sampling rata-rata $\bar{X}$ pada Gambar 5.21 lebih kecil daripada Gambar 5.20. Perhatikan bahwa semakin besar ukuran sampel, maka akan semakin kecil nilai standar deviasi dari distribusi sampling rata-rata $\bar{X}$. Dalam prakteknya, seringkali populasi yang diteliti tidak berdistribusi normal. Teorema yang sangat penting untuk menyimpulkan bentuk dari distribusi sampling rata-rata $\bar{X}$ adalah **Teorema Limit Sentral** (***Central Limit Theorem***).

Teorema limit sentral menyatakan bahwa untuk sampel berukuran besar, distribusi sampling rata-rata $\bar{X}$ akan mendekati normal, tidak peduli apakah sampel-sampel tersebut ditarik dari populasi yang berdistribusi normal atau tidak, dengan rata-rata dan standar deviasi dari distribusi sampling rata-rata $\bar{X}$ sebagai berikut.

$$\mu_{\bar{X}} = \mu \quad \text{dan} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

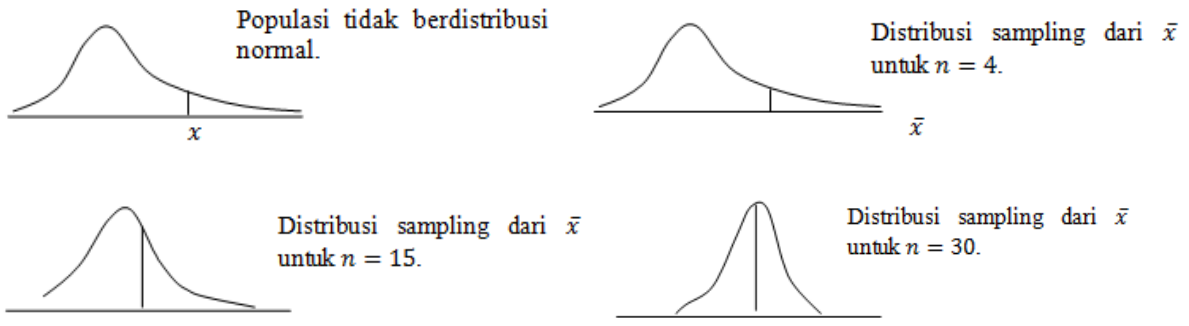
Ukuran sampel n dipertimbangkan cukup besar, yakni $n \geq 30$. Berdasarkan teorema limit sentral, perlu diperhatikan bahwa, jika populasi tidak berdistribusi normal, bentuk dari distribusi sampling rata-rata $\bar{X}$ **tidak secara tepat normal**, **namun mendekati normal**, ketika sampel berukuran besar. Semakin besar ukuran sampel, maka bentuk dari distribusi sampling rata-rata $(\bar{X})$ akan semakin mendekati normal. Berdasarkan teori limit sentral (Mann dan Lacke, 2011:313),

- ⇨ Ketika ukuran sampel $n \geq 30$, maka bentuk dari distribusi sampling rata-rata $(\bar{X})$ mendekati normal, tidak peduli apakah sampel-sampel tersebut ditarik dari populasi berdistribusi normal atau tidak.

- ⇨ Rata-rata dari distribusi sampling rata-rata $(\bar{X})$, yakni $\mu_{\bar{X}}$ sama dengan rata-rata populasi, yakni μ.

- ⇨ Standar deviasi dari distribusi sampling rata-rata $(\bar{X})$, yakni $\sigma_{\bar{X}}$ sama dengan $\sigma/\sqrt{n}$ dengan syarat $n/N \leq 0{,}05$.

Perhatikan ilustrasi gambar berikut.



Gambar 5.22

Berdasarkan Gambar 5.22, populasi tidak berdistribusi normal. Semakin meningkat ukuran sampel, maka distribusi sampling rata-rata $\bar{X}$ semakin berbentuk distribusi normal. Semakin

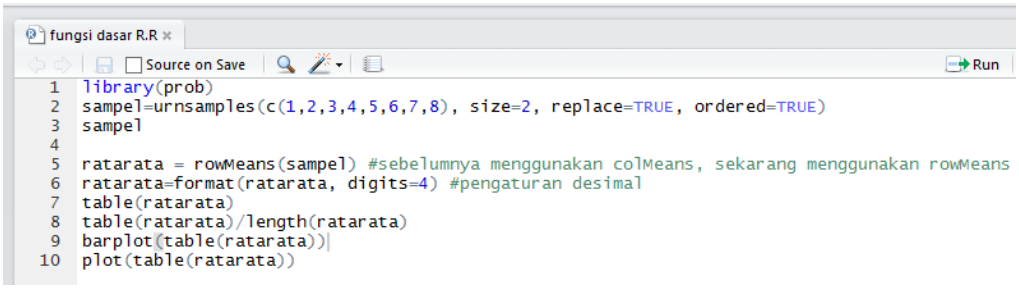
meningkat ukuran sampel, semakin kecil nilai standar deviasi dari distribusi sampling rata-rata $\bar{X}$.

Simulasi Distribusi Sampling dalam R (Bagian 1)

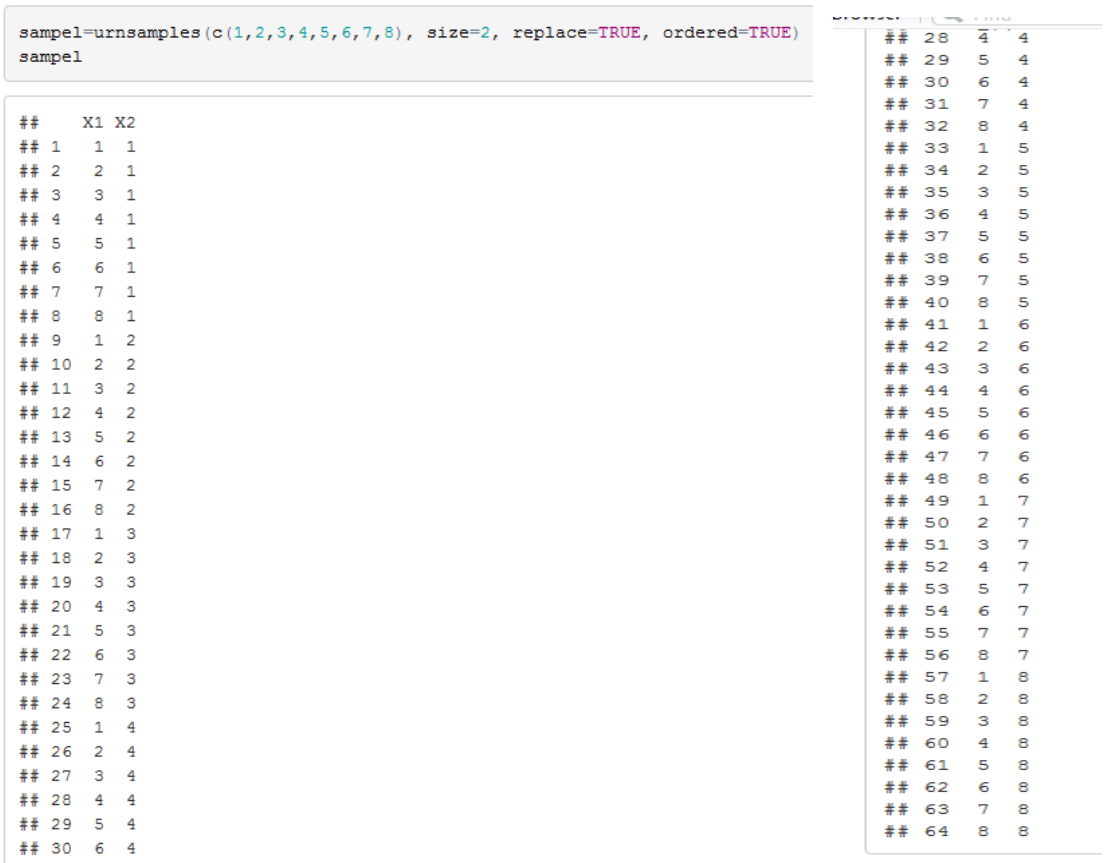
Andaikan diberikan data populasi sebagai berikut.

1,2,3,4,5,6,7,8

Dari data populasi tersebut, **akan diambil sampel yang terdiri dari 2 angka**. Pengambilan sampel **dengan pengembalian** dan **memperhatikan urutan**. Dengan menggunakan R, berikut akan ditentukan seluruh kemungkinan sampel yang mungkin terambil, distribusi frekuensi dari rata-rata sampel, distribusi probabilitas dari rata-rata sampel atau distribusi sampling dari rata-rata sampel, dan disajikan secara visual.



Gambar 5.23



Gambar 5.24

```
ratarata = rowMeans(sampel) #sebelumnya menggunakan colMeans, sekarang menggunakan rowMeans
ratarata=format(ratarata, digits=4) #pengaturan desimal
table(ratarata)
```

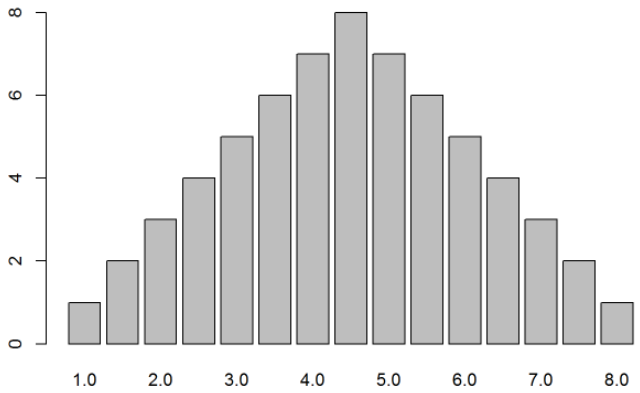
```
## ratarata
## 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
##   1   2   3   4   5   6   7   8   7   6   5   4   3   2   1
```

```
table(ratarata)/length(ratarata)
```

```
## ratarata
##       1.0       1.5       2.0       2.5       3.0       3.5       4.0       4.5
## 0.015625 0.031250 0.046875 0.062500 0.078125 0.093750 0.109375 0.125000
##       5.0       5.5       6.0       6.5       7.0       7.5       8.0
## 0.109375 0.093750 0.078125 0.062500 0.046875 0.031250 0.015625
```

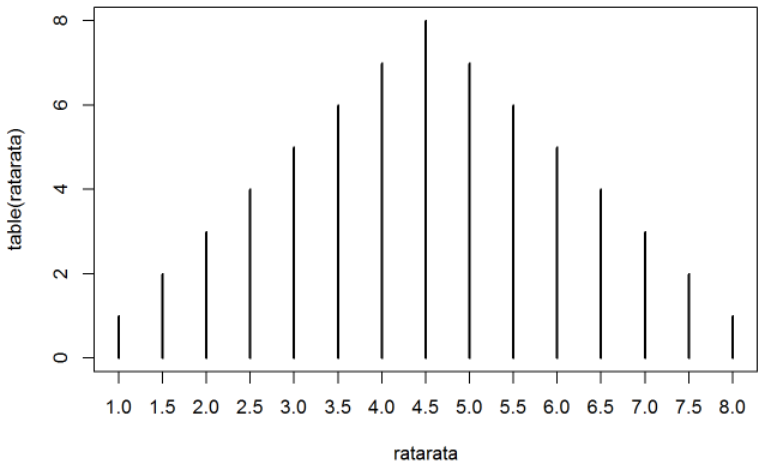
Gambar 5.25

```
barplot(table(ratarata))
```



Gambar 5.26

```
plot(table(ratarata))
```



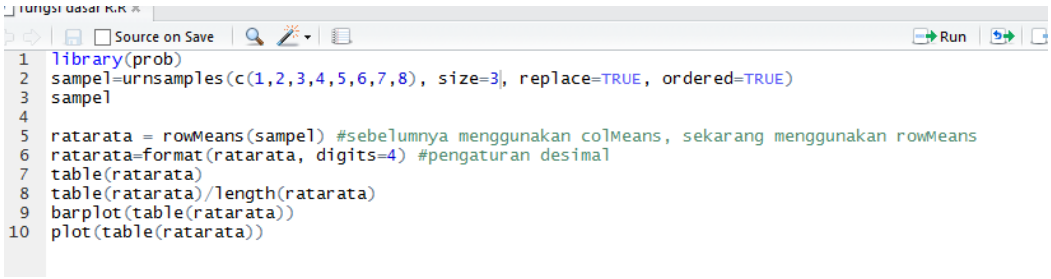
Gambar 5.27

Simulasi Distribusi Sampling dalam R (Bagian 2)

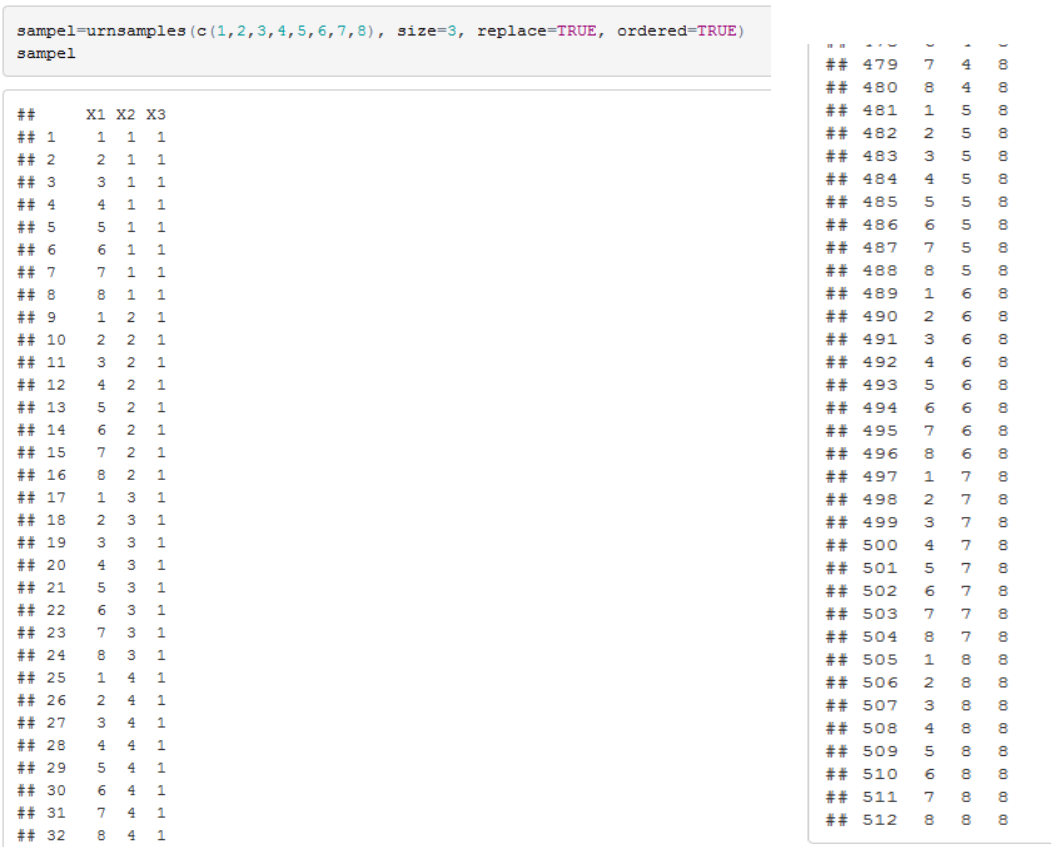
Andaikan diberikan data populasi sebagai berikut.

1,2,3,4,5,6,7,8

Dari data populasi tersebut, **akan diambil sampel yang terdiri dari 3 angka**. Pengambilan sampel **dengan pengembalian** dan **memperhatikan urutan**. Dengan menggunakan R, berikut akan ditentukan seluruh kemungkinan sampel yang mungkin terambil, distribusi frekuensi dari rata-rata sampel, distribusi probabilitas dari rata-rata sampel atau distribusi sampling dari rata-rata sampel, dan disajikan secara visual.



Gambar 5.28



```
sampel=urnsamples(c(1,2,3,4,5,6,7,8), size=3, replace=TRUE, ordered=TRUE)
sampel
```

```
##      X1 X2 X3
## 1     1  1  1
## 2     2  1  1
## 3     3  1  1
## 4     4  1  1
## 5     5  1  1
## 6     6  1  1
## 7     7  1  1
## 8     8  1  1
## 9     1  2  1
## 10    2  2  1
## 11    3  2  1
## 12    4  2  1
## 13    5  2  1
## 14    6  2  1
## 15    7  2  1
## 16    8  2  1
## 17    1  3  1
## 18    2  3  1
## 19    3  3  1
## 20    4  3  1
## 21    5  3  1
## 22    6  3  1
## 23    7  3  1
## 24    8  3  1
## 25    1  4  1
## 26    2  4  1
## 27    3  4  1
## 28    4  4  1
## 29    5  4  1
## 30    6  4  1
## 31    7  4  1
## 32    8  4  1
```

```
## 479   7  4  8
## 480   8  4  8
## 481   1  5  8
## 482   2  5  8
## 483   3  5  8
## 484   4  5  8
## 485   5  5  8
## 486   6  5  8
## 487   7  5  8
## 488   8  5  8
## 489   1  6  8
## 490   2  6  8
## 491   3  6  8
## 492   4  6  8
## 493   5  6  8
## 494   6  6  8
## 495   7  6  8
## 496   8  6  8
## 497   1  7  8
## 498   2  7  8
## 499   3  7  8
## 500   4  7  8
## 501   5  7  8
## 502   6  7  8
## 503   7  7  8
## 504   8  7  8
## 505   1  8  8
## 506   2  8  8
## 507   3  8  8
## 508   4  8  8
## 509   5  8  8
## 510   6  8  8
## 511   7  8  8
## 512   8  8  8
```

Gambar 5.29

```
ratarata = rowMeans(sampel) #sebelumnya menggunakan colMeans, sekarang menggunakan rowMeans
ratarata=format(ratarata, digits=4) #pengaturan desimal
table(ratarata)
```

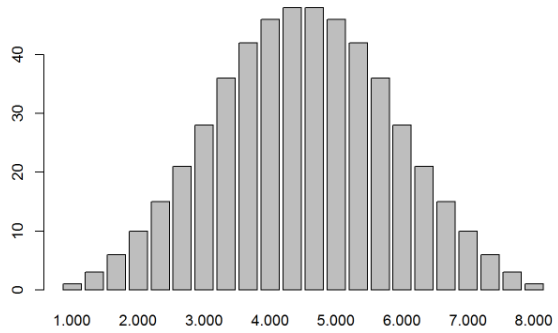
```
## ratarata
## 1.000 1.333 1.667 2.000 2.333 2.667 3.000 3.333 3.667 4.000 4.333 4.667
##     1     3     6    10    15    21    28    36    42    46    48    48
## 5.000 5.333 5.667 6.000 6.333 6.667 7.000 7.333 7.667 8.000
##    46    42    36    28    21    15    10     6     3     1
```

```
table(ratarata)/length(ratarata)
```

```
## ratarata
##       1.000       1.333       1.667       2.000       2.333       2.667
## 0.001953125 0.005859375 0.011718750 0.019531250 0.029296875 0.041015625
##       3.000       3.333       3.667       4.000       4.333       4.667
## 0.054687500 0.070312500 0.082031250 0.089843750 0.093750000 0.093750000
##       5.000       5.333       5.667       6.000       6.333       6.667
## 0.089843750 0.082031250 0.070312500 0.054687500 0.041015625 0.029296875
##       7.000       7.333       7.667       8.000
## 0.019531250 0.011718750 0.005859375 0.001953125
```

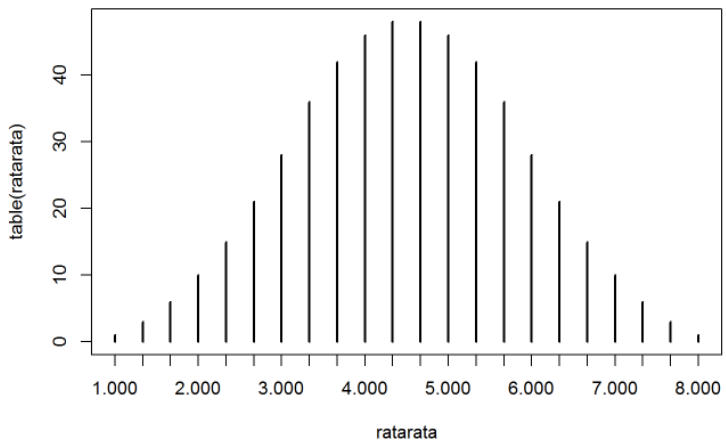
Gambar 5.30

```
barplot(table(ratarata))
```



Gambar 5.31

```
plot(table(ratarata))
```



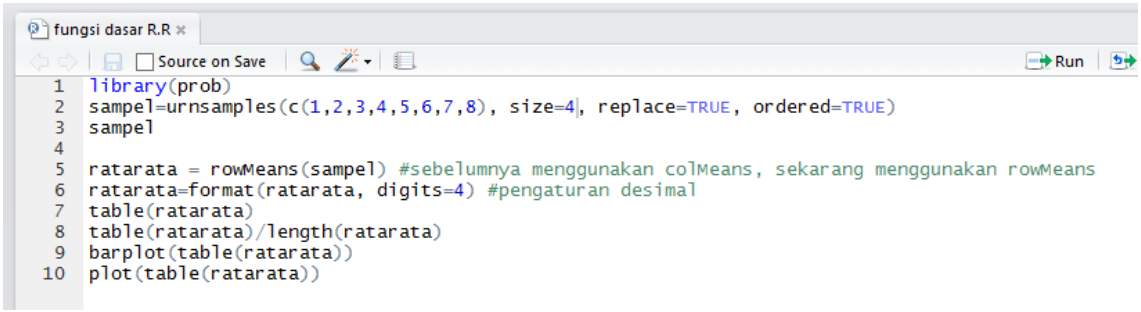
Gambar 5.32

Simulasi Distribusi Sampling dalam R (Bagian 3)

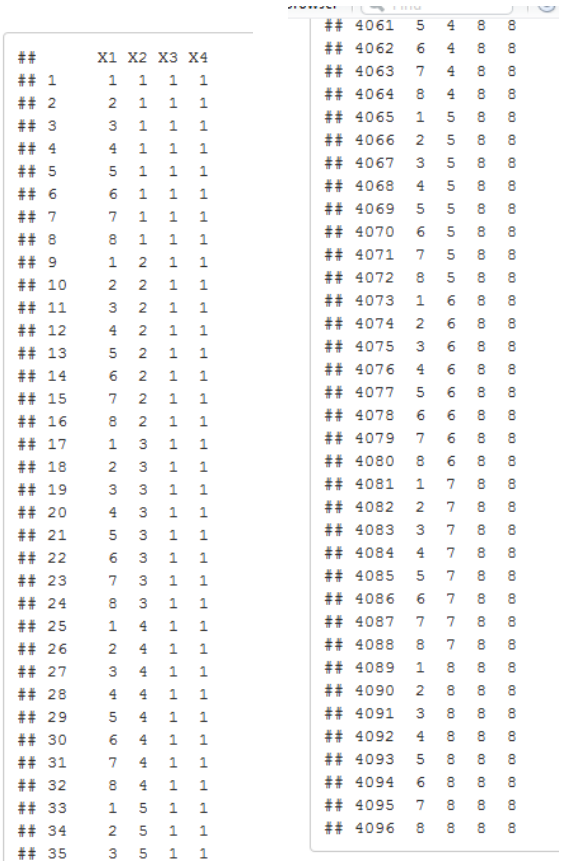
Andaikan diberikan data populasi sebagai berikut.

1,2,3,4,5,6,7,8

Dari data populasi tersebut, **akan diambil sampel yang terdiri dari 4 angka**. Pengambilan sampel **dengan pengembalian** dan **memperhatikan urutan**. Dengan menggunakan R, berikut akan ditentukan seluruh kemungkinan sampel yang mungkin terambil, distribusi frekuensi dari rata-rata sampel, distribusi probabilitas dari rata-rata sampel atau distribusi sampling dari rata-rata sampel, dan disajikan secara visual.



Gambar 5.33



```
##       X1 X2 X3 X4
## 1      1  1  1  1
## 2      2  1  1  1
## 3      3  1  1  1
## 4      4  1  1  1
## 5      5  1  1  1
## 6      6  1  1  1
## 7      7  1  1  1
## 8      8  1  1  1
## 9      1  2  1  1
## 10     2  2  1  1
## 11     3  2  1  1
## 12     4  2  1  1
## 13     5  2  1  1
## 14     6  2  1  1
## 15     7  2  1  1
## 16     8  2  1  1
## 17     1  3  1  1
## 18     2  3  1  1
## 19     3  3  1  1
## 20     4  3  1  1
## 21     5  3  1  1
## 22     6  3  1  1
## 23     7  3  1  1
## 24     8  3  1  1
## 25     1  4  1  1
## 26     2  4  1  1
## 27     3  4  1  1
## 28     4  4  1  1
## 29     5  4  1  1
## 30     6  4  1  1
## 31     7  4  1  1
## 32     8  4  1  1
## 33     1  5  1  1
## 34     2  5  1  1
## 35     3  5  1  1
```

```
## 4061  5  4  8  8
## 4062  6  4  8  8
## 4063  7  4  8  8
## 4064  8  4  8  8
## 4065  1  5  8  8
## 4066  2  5  8  8
## 4067  3  5  8  8
## 4068  4  5  8  8
## 4069  5  5  8  8
## 4070  6  5  8  8
## 4071  7  5  8  8
## 4072  8  5  8  8
## 4073  1  6  8  8
## 4074  2  6  8  8
## 4075  3  6  8  8
## 4076  4  6  8  8
## 4077  5  6  8  8
## 4078  6  6  8  8
## 4079  7  6  8  8
## 4080  8  6  8  8
## 4081  1  7  8  8
## 4082  2  7  8  8
## 4083  3  7  8  8
## 4084  4  7  8  8
## 4085  5  7  8  8
## 4086  6  7  8  8
## 4087  7  7  8  8
## 4088  8  7  8  8
## 4089  1  8  8  8
## 4090  2  8  8  8
## 4091  3  8  8  8
## 4092  4  8  8  8
## 4093  5  8  8  8
## 4094  6  8  8  8
## 4095  7  8  8  8
## 4096  8  8  8  8
```

Gambar 5.34

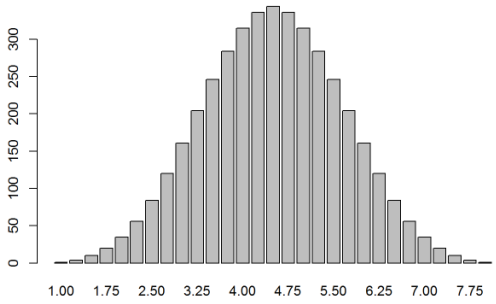
```
ratarata = rowMeans(sampel) #sebelumnya menggunakan colMeans, sekarang menggunakan rowMeans
ratarata=format(ratarata, digits=4) #pengaturan desimal
table(ratarata)
```

```
## ratarata
## 1.00 1.25 1.50 1.75 2.00 2.25 2.50 2.75 3.00 3.25 3.50 3.75 4.00 4.25 4.50
##    1    4   10   20   35   56   84  120  161  204  246  284  315  336  344
## 4.75 5.00 5.25 5.50 5.75 6.00 6.25 6.50 6.75 7.00 7.25 7.50 7.75 8.00
##  336  315  284  246  204  161  120   84   56   35   20   10    4    1
```

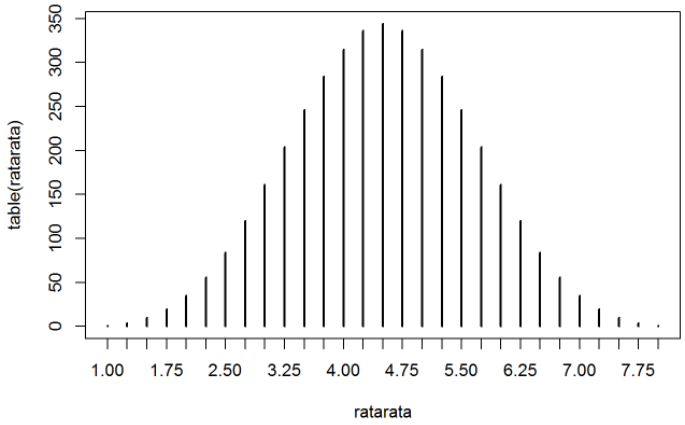
```
table(ratarata)/length(ratarata)
```

```
## ratarata
##          1.00          1.25          1.50          1.75          2.00
## 0.0002441406 0.0009765625 0.0024414062 0.0048828125 0.0085449219
##          2.25          2.50          2.75          3.00          3.25
## 0.0136718750 0.0205078125 0.0292968750 0.0393066406 0.0498046875
##          3.50          3.75          4.00          4.25          4.50
## 0.0600585938 0.0693359375 0.0769042969 0.0820312500 0.0839843750
##          4.75          5.00          5.25          5.50          5.75
## 0.0820312500 0.0769042969 0.0693359375 0.0600585938 0.0498046875
##          6.00          6.25          6.50          6.75          7.00
## 0.0393066406 0.0292968750 0.0205078125 0.0136718750 0.0085449219
##          7.25          7.50          7.75          8.00
## 0.0048828125 0.0024414062 0.0009765625 0.0002441406
```

Gambar 5.35



Gambar 5.36



Gambar 5.37

Simulasi Distribusi Sampling dalam R (Bagian 4)

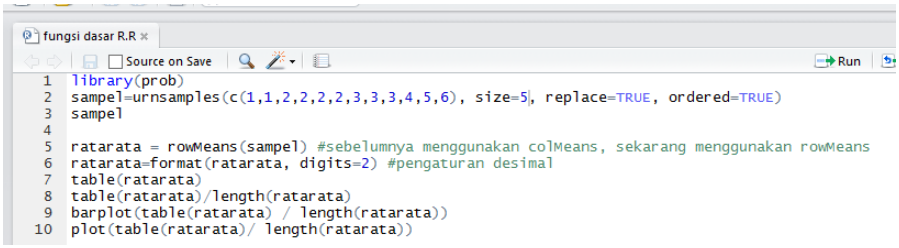
Andaikan diberikan data populasi sebagai berikut.

1,1,2,2,2,2,3,3,3,4,5,6

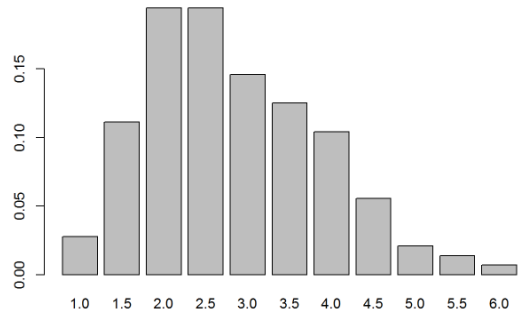
Dari data populasi tersebut, misalkan:

➔**akan diambil sampel yang terdiri dari 2 angka**.
➔**akan diambil sampel yang terdiri dari 3 angka**.
➔**akan diambil sampel yang terdiri dari 4 angka**.
➔**akan diambil sampel yang terdiri dari 5 angka**.

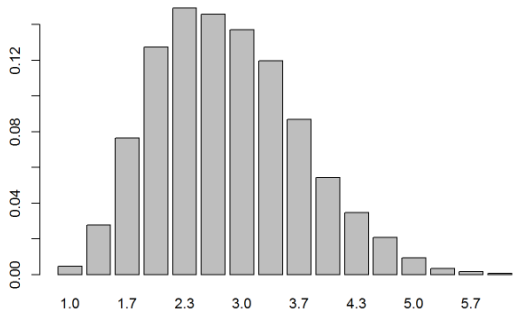
Pengambilan sampel **dengan pengembalian** dan **memperhatikan urutan**. Dengan menggunakan R, berikut akan disajikan secara visual distribusi sampling dari rata-rata sampel.



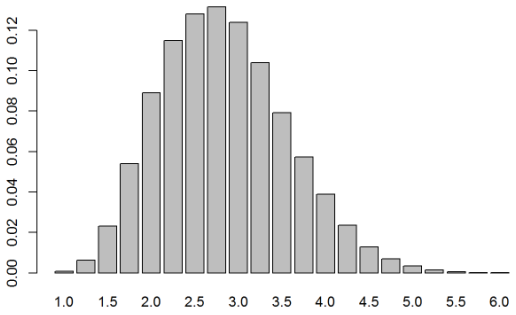
Gambar 5.38



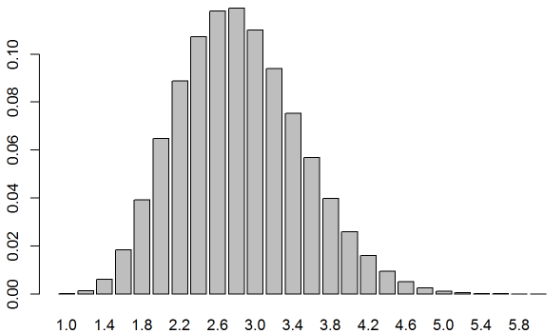
$n = 2$



$n = 3$



$n = 3$



$n = 4$

Gambar 5.39

Referensi

1. Agresti, A. dan B. Finlay. 2009. *Statistical Methods for the Social Sciences, 4th Edition*. United States of America: Prentice Hall.

2. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

3. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

4. Johnson, R.A. dan G.K. Bhattacharyya. 2011. *Statistics, Principles and Methods, 6th Edition*. John Wiley and Sons, Inc.

5. Mann, P. S. dan C.J. Lacke. 2011. *Introductory Statistics, International Student Version, 7th Edition*. Asia: John Wiley & Sons, Inc.

6. Montgomery, D. C. dan G. C. Runger. 2011. *Applied Statistics and Probability for Engineers, 5th Edition*. United States of America: John Wiley & Sons, Inc.

7. Ott, R.L. dan M. Longnecker. 2001. *An Introduction to Statistical Methods and Data Analysis, 5th Edition*. United States of America: Duxbury.

8. Smidth, R. K. dan D. H. Sanders. 2000. *Statistics a First Course, 6th Edition*. United States of America: McGraw-Hill Companies.

9. http://www.dummies.com/how-to/content/how-to-format-numbers-in-r.html

10. http://stackoverflow.com/questions/13033914/sampling-distribution-of-the-sample-mean

11. https://cran.r-project.org/web/packages/prob/prob.pdf

BAB 6

UJI NORMALITAS POPULASI

Uji Normalitas dengan Uji Kolmogorov-Smirnov

Uji Kolmogorov-Smirnov dapat digunakan untuk menguji suatu asumsi apakah suatu data sampel berasal dari populasi yang berdistribusi normal atau tidak. Pada pembahasan Bab 5 telah dibahas mengenai distribusi sampling dari rata-rata $\bar{X}$. Apabila data sampel berasal dari populasi yang berdistribusi normal, maka distribusi sampling dari rata-rata $\bar{X}$ juga mengikuti distribusi normal. Asumsi normalitas memiliki peranan penting dalam uji-uji parametrik, seperti uji beda rata-rata dari dua populasi dengan uji t dan analisis varians. Hal ini karena uji-uji parametrik akan bekerja dengan baik ketika asumsi normalitas dipenuhi. Conover (1999:115) menyatakan sebagai berikut.

"Most parametric methods are based on the normality assumption because the theory behind the test can be worked out with the normal population distribution. The resulting procedures are efficient and powerful procedures for normally distributed data. Other parametric procedures have been developed assuming the population has other distributions, such as the exponential, Weibull, and soon".

Pada uji Kolmogorov-Smirnov, hipotesis nol menyatakan data yang diteliti berasal dari populasi yang berdistribusi normal, sedangkan hipotesis alternatif menyatakan data yang diteliti tidak berasal dari populasi yang berdistribusi normal. Andaikan $X_1, X_2, X_3, \dots, X_k$ merupakan nilai-nilai pada sampel acak (*random sample*). Misalkan $f(X_i)$ menyatakan probabilitas dari nilai X_i, sedangkan $F(X_i) = f(X \leq X_i)$ menyatakan probabilitas kumulatif dari nilai X_i, di mana $i = 1,2,3, \dots, k$. Selanjutnya andaikan Z_i merupakan nilai normal (sampel) terstandarisasi dari hasil transformasi nilai X_i dan $F(Z_i) = f(Z \leq Z_i)$ menyatakan probabilitas kumulatif dari nilai normal Z_i terstandarisasi. Nilai normal Z_i terstandarisasi merupakan hasil transformasi dari nilai X_i yang dihitung dengan rumus sebagai berikut.

$$Z_i = \frac{X_i - \bar{X}}{s}, i = 1,2,3, \dots, k$$

Perhatikan bahwa $\bar{X}$ merupakan rata-rata sampel sebagai estimasi dari rata-rata populasi μ, sedangkan s merupakan standar deviasi sampel sebagai estimasi dari standar deviasi populasi σ. Misalkan D_i menyatakan nilai mutlak dari selisih antara $F(Z_i)$ dan $F(X_i)$, yakni

$$\mathrm{D_i} = |F(Z_i) - F(X_i)|, i = 1,2,3, \dots, k.$$

Nilai D_i paling besar (*maximum*) atau D_{max} merupakan nilai statistik dari uji Kolmogorov-Smirnov. Nilai statistik dari uji Kolmogorov-Smirnov (D_{max}) kemudian dibandingkan dengan nilai kritis berdasarkan tabel distribusi Kolmogorov-Smirnov untuk pengambilan keputusan terhadap hipotesis. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji Kolmogorov-Smirnov.

$$Jika\ D_{max} \leq nilai\ kritis, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ D_{max} > nilai\ kritis, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Tabel 6.1 merupakan tabel distribusi Kolmogorov-Smirnov. Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan membandingkan nilai probabilitas (*p-value*) dari uji Kolmogorov-Smirnov terhadap tingkat signifikansi α (*significance level*). Berikut aturan pengambilan keputusan berdasarkan pendekatan nilai probabilitas.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Tabel 6.1

n	$\alpha = 0{,}2$	$\alpha = 0{,}1$	$\alpha = 0{,}05$	$\alpha = 0{,}02$	$\alpha = 0{,}01$
1	0,900	0,950	0,975	0,990	0,995
2	0,684	0,776	0,842	0,900	0,929
3	0,565	0,636	0,708	0,785	0,829
4	0,493	0,565	0,624	0,689	0,734
5	0,447	0,509	0,563	0,627	0,669
6	0,410	0,468	0,519	0,577	0,617
7	0,381	0,436	0,483	0,538	0,576
8	0,359	0,410	0,454	0,507	0,542
9	0,339	0,387	0,430	0,480	0,513
10	0,323	0,369	0,409	0,457	0,486
11	0,308	0,352	0,391	0,437	0,468
12	0,296	0,338	0,375	0,419	0,449
13	0,285	0,325	0,361	0,404	0,432
14	0,275	0,314	0,349	0,390	0,418
15	0,266	0,304	0,338	0,377	0,404
16	0,258	0,295	0,327	0,366	0,392
17	0,250	0,286	0,318	0,355	0,381
18	0,244	0,279	0,309	0,346	0,371

Contoh Kasus Uji Normalitas Populasi dengan Uji Kolmogorov-Smirnov (Contoh Perhitungan)

Misalkan seorang mahasiswa semester 8 sedang menyusun tugas akhir dan baru saja mengumpulkan data sampel mengenai nilai ujian matematika kelas 6 SD sebanyak 16 siswa. Berikut data yang telah dikumpulkan oleh mahasiswa tersebut.

Tabel 6.2 (Data Fiktif)

Nomor	Nama	Nilai	Nomor	Nama	Nilai	Nomor	Nama	Nilai
1	A	40	7	H	70	13	N	80
2	B	50	8	I	70	14	O	90
3	C	50	9	J	70	15	P	90
4	D	60	10	K	70	16	Q	100
5	F	60	11	L	80			
6	G	60	12	M	80			

Berikut akan digunakan pendekatan uji Kolmogorov-Smirnov untuk menguji hipotesis apakah data tersebut ditarik dari populasi yang berdistribusi normal atau tidak (misalkan tingkat signifikansi yang digunakan $\alpha = 5\%$). Perhitungan akan dilakukan secara manual.

➔Menghitung nilai rata-rata $(\bar{X})$ dan standar deviasi (s).

Tabel 6.3

No.	X	Frekuensi	$f(X)$	$F(X)$	Z	$F(Z)$	$D=\|F(Z)-F(X)\|$
1	40	1	0,0625	0,0625	-1,83712	0,033096276	0,029403724
2	50	2	0,125	0,1875	-1,22474	0,110335658	0,077164342
3	60	3	0,1875	0,375	-0,61237	0,270145667	0,104854333
4	70	4	0,25	0,625	0	0,5	**0,125**
5	80	3	0,1875	0,8125	0,612372	0,729854333	0,082645667
6	90	2	0,125	0,9375	1,224745	0,889664342	0,047835658
7	100	1	0,0625	1	1,837117	0,966903724	0,033096276

Berdasarkan Tabel 6.3, berikut akan dihitung nilai rata-rata hitung $(\bar{X})$ dan standar deviasi (s).

$$\bar{X} = \frac{\sum X}{n}$$

$$\bar{X} = \frac{(40 \times 1) + (50 \times 2) + (60 \times 3) + (70 \times 4) + (80 \times 3) + (90 \times 2) + (100 \times 1)}{16}$$

$$\bar{X} = 70$$

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

$$s = \sqrt{\frac{4000}{15}}$$

$$s = 16{,}330.$$

➔Menghitung probabilitas dari X_i atau $f(X_i)$.

Setelah diperoleh $\bar{X} = 70$ dan $s = 16{,}330$, selanjutnya akan dihitung probabilitas dari X_i atau $f(X_i)$. Probabilitas untuk nilai $X = 40$ atau $f(40)$ adalah $\frac{1}{16} = 0{,}0625$, probabilitas untuk nilai $X = 50$ atau $f(50)$ adalah $\frac{2}{16} = 0{,}125$, probabilitas untuk nilai $X = 70$ atau $f(70)$ adalah $\frac{4}{16} = 0{,}25$, dan seterusnya.

➔Menghitung probabilitas kumulatif dari X_i atau $F(X_i) = f(X \leq X_i)$.

Nilai dari $F(40) = 0{,}0625$, nilai dari $F(50) = f(X \leq 50) = f(40) + f(50) = 0{,}0625 + 0{,}125 = 0{,}1875$, nilai dari $F(60) = f(X \leq 60) = f(40) + f(50) + f(60) = 0{,}375$, dan seterusnya.

➔Mentransformasi nilai X_i menjadi nilai normal Z_i terstandarisasi.

Selanjutnya mentransformasi nilai X_i ke dalam nilai normal Z_i terstandarisasi yang dihitung dengan rumus

$$Z_i = \frac{X_i - \bar{X}}{s}.$$

Untuk $X = 40$, maka

$$Z(X = 40) = \frac{40 - 70}{16{,}330} = -1{,}837.$$

Untuk $X = 50$, maka

$$Z(X = 50) = \frac{50 - 70}{16{,}330} = -1{,}2247,$$

dan seterusnya.

➔ Menghitung probabilitas kumulatif dari Z_i atau $F(Z_i) = f(Z \leq Z_i)$.

Setelah diperoleh nilai-nilai normal terstandarisasi, maka akan dihitung probabilitas kumulatif dari nilai-nilai normal terstandarisasi tersebut. Probabilitas kumulatif dari $Z = -1{,}837$ atau $f(Z \leq -1{,}837)$ berdasarkan tabel distribusi normal kumulatif adalah $0{,}033$, probabilitas kumulatif dari $Z = 0{,}61$ atau $f(Z \leq 0{,}61)$ berdasarkan tabel distribusi normal kumulatif adalah $0{,}729$, dan seterusnya.

➔ Menghitung nilai mutlak dari selisih antara $F(Z_i)$ dan $F(X_i)$.

Selanjutnya menghitung nilai mutlak dari selisih antara $F(Z_i)$ dan $F(X_i)$.

$$\mathrm{D_i} = |F(Z_i) - F(X_i)|.$$

Nilai D untuk $X = 40$ adalah $|0{,}033 - 0{,}0625| = 0{,}0295$, nilai D untuk $X = 50$ adalah $|0{,}110 - 0{,}1875| = 0{,}077$, dan seterusnya.

➔ Menghitung nilai statistik dari uji Kolmogorov-Smirnov (D_{max}).

Nilai statistik dari uji Kolmogorov-Smirnov merupakan nilai D yang paling besar atau maksimum. Berdasarkan Tabel 6.3, nilai D terbesar adalah $0{,}125$, sehingga nilai statistik dari uji Kolmogorov-Smirnov adalah $0{,}125$ atau $D_{max} = 0{,}125$.

➔ Menghitung nilai kritis Kolmogorov-Smirnov.

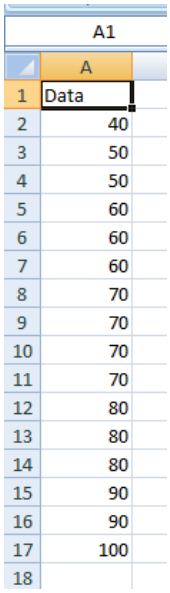
Nilai kritis Kolmogorov-Smirnov pada tingkat signifikansi 5% dan jumlah elemen sampel 16 berdasarkan tabel distribusi Kolmogorov-Smirnov adalah 0,327.

➔ Pengambilan keputusan terhadap hipotesis.

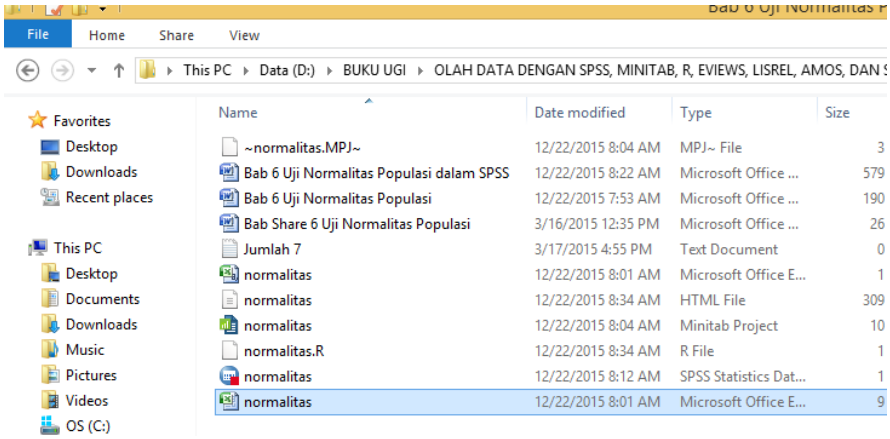
Perhatikan bahwa karena nilai statistik dari uji Kolmogorov-Smirnov (0,125) lebih kecil dibandingkan nilai kritis Kolmogorov-Smirnov (0,327), maka hipotesis nol diterima dan hipotesis alternatif ditolak, sehingga asumsi mengenai data nilai ujian matematika kelas 6 SD ditarik dari populasi yang berdistribusi normal dapat diterima pada tingkat signifikansi 5%.

Penyelesaian dalam R untuk Uji Normalitas Populasi dengan Uji Kolmogorov-Smirnov

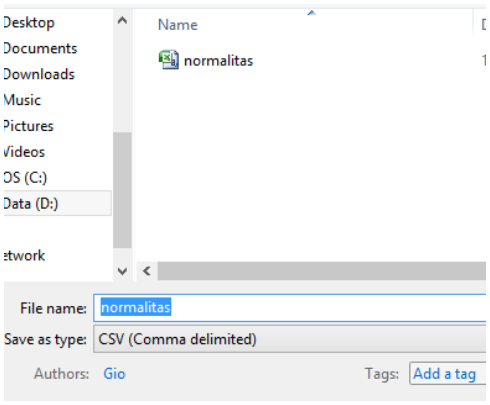
Data terlebih dahulu dibuat dalam *Microsoft Excel* (Gambar 6.1) dan disimpan dengan format tipe **.csv** (Gambar 6.2 dan Gambar 6.3). Ketik kode R seperti pada Gambar 6.4. Kemudian *Compile* dan pilih HTML (Gambar 6.5). Hasilnya seperti pada Gambar 6.6.



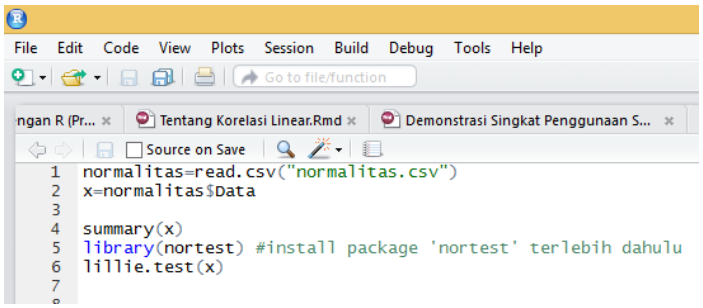
Gambar 6.1



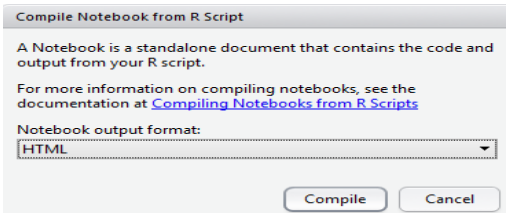
Gambar 6.2



Gambar 6.3



Gambar 6.4



Gambar 6.5

```
normalitas=read.csv("normalitas.csv")
x=normalitas$Data

summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      40      60      70      70      80     100
```

```
lillie.test(x)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.125, p-value = 0.7235
```

Gambar 6.6

Pada Gambar 6.6, terlihat bahwa nilai statistik dari uji Kolmogorov-Smirnov (D) 0,125, lebih kecil dibandingkan nilai kritis Kolmogorov-Smirnov 0,327, maka hipotesis nol diterima dan hipotesis alternatif ditolak, sehingga asumsi mengenai data nilai ujian matematika kelas 6 SD ditarik dari populasi yang berdistribusi normal dapat diterima pada tingkat signifikansi 5%.

Perhatikan juga bahwa nilai probabilitas atau *p-value* adalah 0,7235. Karena nilai probabilitas, yakni 0,7235, lebih besar dibandingkan tingkat signifikansi, yakni 0,05, maka hipotesis nol diterima, dan hipotesis alternatif ditolak. Hal ini berarti asumsi mengenai data nilai ujian matematika kelas 6 SD ditarik dari populasi yang berdistribusi normal dapat diterima pada tingkat signifikansi 5%.

Pada Gambar 6.4, *package* **nortest** diaktifkan dengan maksud untuk menggunakan fungsi **lillie.test**. Fungsi **lillie.test** digunakan untuk menghitung nilai statistik dari uji Kolmogorov-Smirnov, dan probabilitasnya.

Uji Normalitas Populasi dengan Uji Jarque-Bera (Contoh Perhitungan dan Penyelesaian dalam R)

Berdasarkan data pada Tabel 6.2, berikut akan digunakan pendekatan uji Jarque-Bera (JB) untuk menguji hipotesis apakah data tersebut ditarik dari populasi yang berdistribusi normal atau tidak (misalkan tingkat signifikansi yang digunakan $\alpha = 5\%$). Perhitungan akan dilakukan secara manual. Nilai statistik dari uji JB dihitung dengan rumus sebagai berikut (Gujarati, 2003:148).

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right]$$

Perhatikan bahwa n menyatakan banyaknya elemen dalam sampel, S menyatakan kemiringan atau *skewness*, dan K menyatakan kurtosis. Untuk variabel yang terdistribusi secara normal, $S = 0$ dan $K = 3$. Oleh karena itu, uji normalitas JB merupakan suatu uji dari hipotesis gabungan (*joint hypothesis*), yakni S dan K masing-masing bernilai 0 dan 3. Dalam hal ini, nilai statistik dari uji JB diharapkan 0 (Gujarati, 2003:148).
Untuk kemiringan dan kurtosis dihitung dengan rumus sebagai berikut.

$$Kemiringan = \frac{\frac{1}{n}\sum(X-\bar{X})^3}{\left(\frac{1}{n}\sum(X-\bar{X})^2\right)^{3/2}}$$

$$Kurtosis = \frac{\frac{1}{n}\sum(X-\bar{X})^4}{\left(\frac{1}{n}\sum(X-\bar{X})^2\right)^{2}}$$

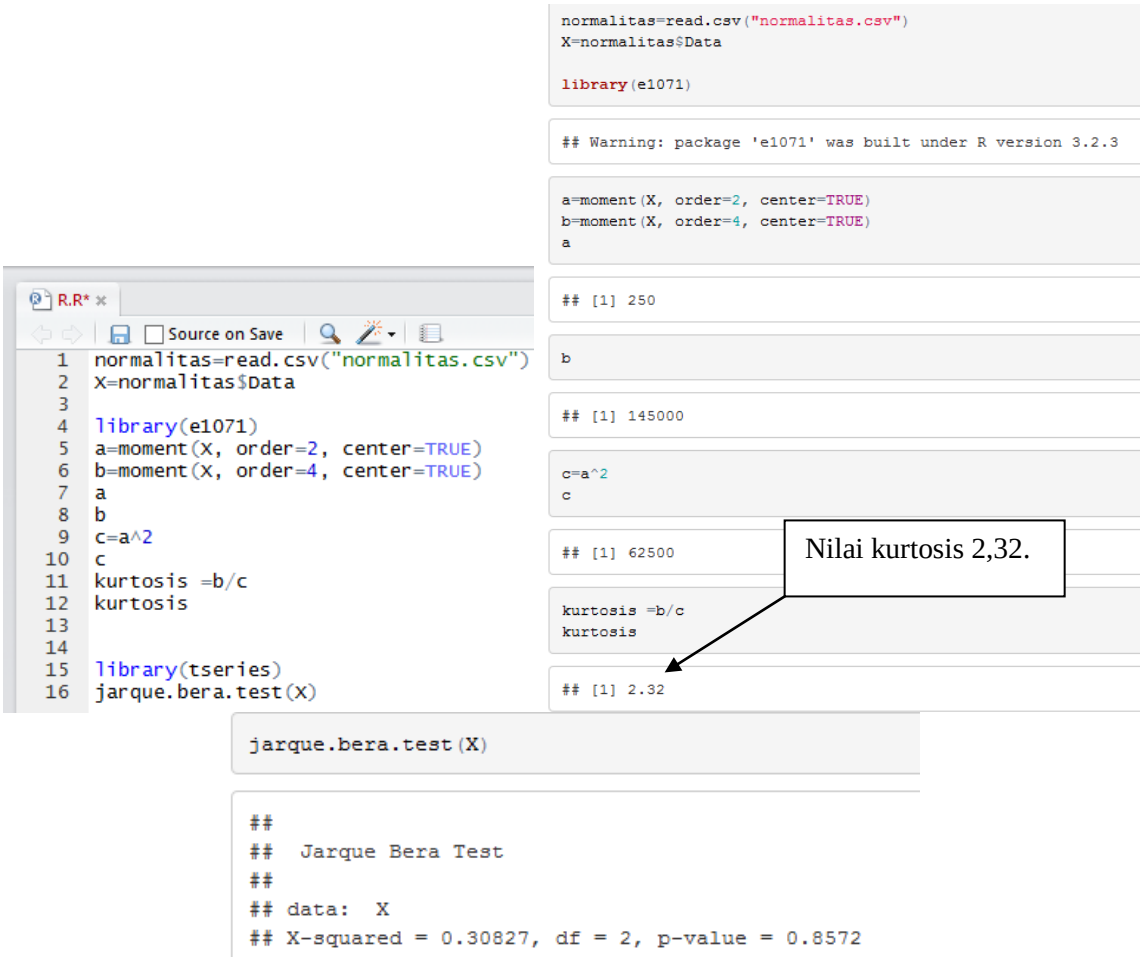
Tabel 6.4

	X	$(X-\bar{X})^2$	$(X-\bar{X})^3$	$(X-\bar{X})^4$
	40	900	-27000	810000
	50	400	-8000	160000
	50	400	-8000	160000
	60	100	-1000	10000
	60	100	-1000	10000
	60	100	-1000	10000
	70	0	0	0
	70	0	0	0
	70	0	0	0
	70	0	0	0
	80	100	1000	10000
	80	100	1000	10000
	80	100	1000	10000
	90	400	8000	160000
	90	400	8000	160000
	100	900	27000	810000
Jumlah	1120	4000	0	2320000
Rata-Rata	70	250	0	145000
Standar Deviasi	16.32993	296.6479395	10708.25227	268179.0447

$$Kemiringan = \frac{\frac{1}{n}\sum(X-\bar{X})^3}{\left(\frac{1}{n}\sum(X-\bar{X})^2\right)^{3/2}} = \frac{0}{\left(\frac{1}{n}\sum(X-\bar{X})^2\right)^{3/2}} = 0$$

$$Kurtosis = \frac{\frac{1}{n}\sum(X-\bar{X})^4}{\left(\frac{1}{n}\sum(X-\bar{X})^2\right)^{2}} = \frac{\frac{1}{16}(2320000)}{\left(\frac{1}{16}(4000)\right)^{2}} = \frac{145000}{62500} = 2{,}32$$

Gambar 6.7 menyajikan hasil perhitungan kurtosis. Berdasarkan Gambar 6.7, nilai dari kurtosis adalah 2,32.



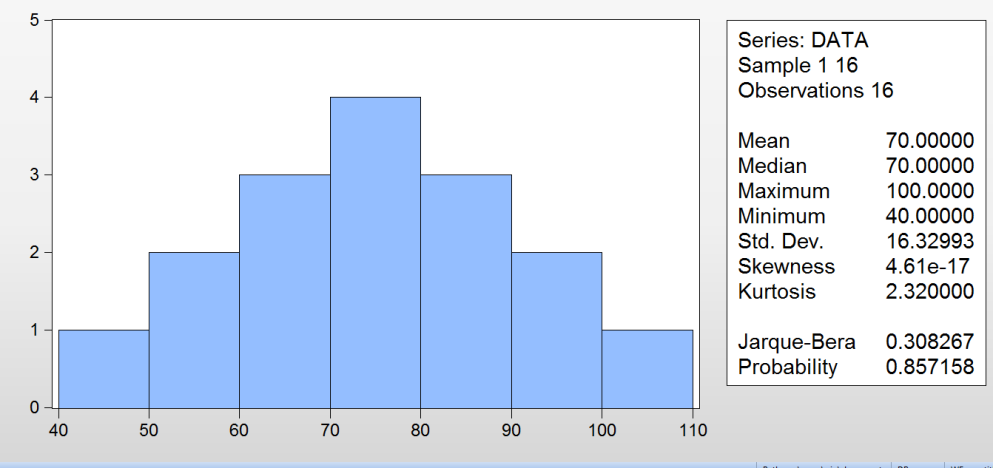
Gambar 6.7

Diketahui nilai kemiringan adalah 0 dan nilai kurtosis adalah 2,32. Sehingga nilai statistik dari uji JB dihitung sebagai berikut.

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right] = 16\left[\frac{0^2}{6} + \frac{(2{,}32-3)^2}{24}\right]$$

$$JB = 0{,}308267$$

Gambar 6.8 ditampilkan hasil perhitungan nilai statistik dari uji JB berdasarkan *software* EViews. Untuk hasil perhitungan nilai statistik dari uji JB berdasarkan R, disajikan pada Gambar 6.7 (*X-squared* = 0,30827).



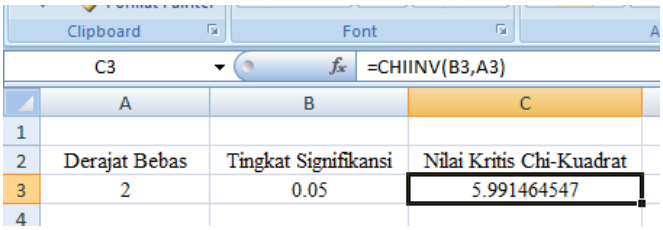
Gambar 6.8

Pengambilan keputusan terhadap hipotesis, dapat dilakukan dengan membandingkan nilai statisik dari uji Jarque-Bera terhadap nilai kritis chi-kuadrat χ^2_{kritis}. Statistik dari uji Jarque-Bera berdistribusi sampling chi-kuadrat dengan derajat bebas 2 untuk ukuran sampel yang besar. Gujarati (2003:148) menyatakan sebagai berikut.

"*Under the null hypothesis that the residuals are normally distributed, Jarque and Bera showed that asymptotically (i.e., in large samples) the JB statistic given in (5.12.1)* ***follows the chi-square distribution with 2 df***."

Berikut aturan pengambilan keputusan terhadap hipotesis.

$$Jika\ nilai\ statistik\ JB \leq \chi^2_{kritis}, H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ statistik\ JB > \chi^2_{kritis}, H_0\ ditolak\ dan\ H_1\ diterima.$$



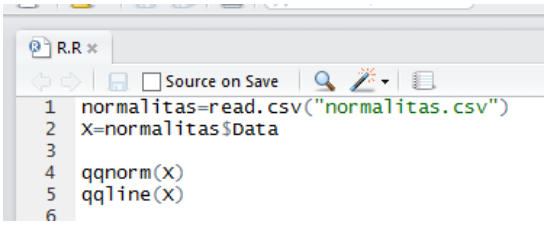
Gambar 6.9 Menghitung Nilai Kritis Chi-kuadrat dengan *Microsoft Excel*

Berdasarkan Gambar 6.9, diketahui nilai kritis chi-kuadrat bernilai 5,991. Karena nilai statisik dari uji Jarque-Bera, yakni 0,308, lebih kecil dibandingkan nilai kritis chi-kuadrat, yakni 5,991, maka hipotesis nol diterima dan hipotesis alternatif ditolak, sehingga asumsi mengenai data nilai ujian matematika kelas 6 SD ditarik dari populasi yang berdistribusi normal dapat diterima pada tingkat signifikansi 5%.

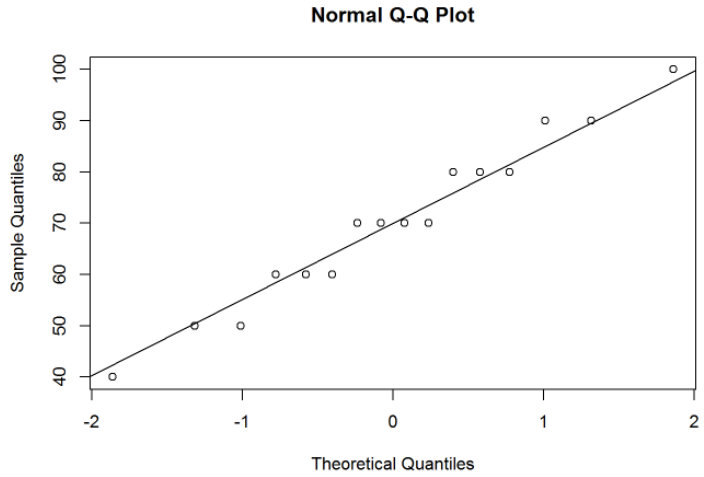
Perhatikan juga bahwa nilai probabilitas atau *p-value* adalah 0,8572 (lihat Gambar 6.7). Karena nilai probabilitas, yakni 0,8572, lebih besar dibandingkan tingkat signifikansi, yakni 0,05, maka hipotesis nol diterima, dan hipotesis alternatif ditolak. Hal ini berarti asumsi mengenai data nilai ujian matematika kelas 6 SD ditarik dari populasi yang berdistribusi normal dapat diterima pada tingkat signifikansi 5%.

Uji Normalitas Populasi dengan Quantile-Quantile Plot (Q-Q Plot)

Untuk menguji asumsi normalitas juga dapat digunakan pendekatan analisis grafik, yakni *Q-Q (quantile-quantile) plot*. Pada pendekatan *Q-Q plot*, jika titik-titik (*dots*) menyebar jauh (menyebar berliku-liku pada garis diagonal seperti ular) dari garis diagonal, maka diindikasi asumsi normalitas tidak dipenuhi. Jika titik-titik menyebar sangat dekat pada garis diagonal, maka asumsi normalitas dipenuhi. Ilustrasi dalam R diperlihatkan pada Gambar 6.10 dan Gambar 6.11.



Gambar 6.10



Gambar 6.11

Berdasarkan Gambar 6.11, titik-titik (*dots*) menyebar cukup dekat dari garis diagonal, maka maka asumsi normalitas dipenuhi.

Referensi

1. Conover, W.J. 1999. *Practical Nonparametric Statistics*, 3rd *Edition*. New York: John Wiley & Sons, Inc.

2. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

3. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

4. Mann, P. S. dan C.J. Lacke. 2011. *Introductory Statistics, International Student Version, 7th Edition*, Asia: John Wiley & Sons, Inc.

5. Montgomery, D.C. dan G.C. Runger. 2011. *Applied Statistics and Probability for Engineers, 5th Edition*. United States of America: John Wiley & Sons, Inc.

6. http://www.r-tutor.com/elementary-statistics/numerical-measures/skewness

7. http://www.r-tutor.com/elementary-statistics/numerical-measures/moment

8. http://stats.stackexchange.com/questions/130368/why-do-i-get-this-p-value-doing-the-jarque-bera-test-in-r

9. http://www.inside-r.org/packages/cran/tseries/docs/jarque.bera.test

10. https://cran.r-project.org/web/packages/nortest/nortest.pdf

11. https://cran.r-project.org/web/packages/e1071/e1071.pdf

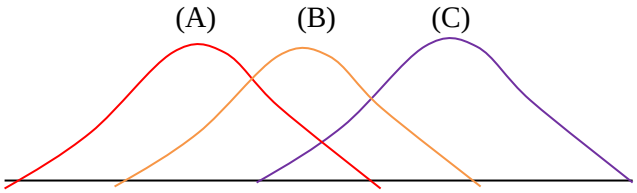
12. https://cran.r-project.org/web/packages/tseries/tseries.pdf

BAB 7

UJI KESAMAAN VARIANS POPULASI

Uji Kesaman Varians Populasi dengan Uji Levene

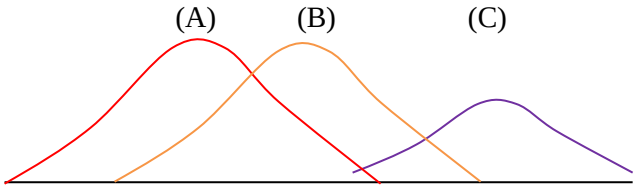
Uji Levene merupakan salah satu uji dalam statistika yang dapat digunakan untuk menguji kesamaan varians dari dua atau lebih populasi. Selain uji Levene, dapat juga digunakan uji F, uji Hartley, dan uji Bartlett untuk menguji kesamaan varians populasi. Varians populasi dilambangkan dengan σ^2, sedangkan varians sampel dilambangkan dengan s^2.



$$\sigma_A^2 = \sigma_B^2 = \sigma_C^2$$

Gambar 7.1

Pada Gambar 7.1, varians dari populasi A, B, dan C adalah sama, namun rata-ratanya berbeda. Pada Gambar 7.2, varians dari populasi A dan B sama, namun berbeda dengan C.



Gambar 7.2

Pada uji Levene, hipotesis nol menyatakan tidak terdapat perbedaan varians di antara populasi, sedangkan hipotesis alternatif menyatakan terdapat paling tidak sepasang varians populasi yang berbeda. Field (2009:150) menyatakan sebagai berikut.

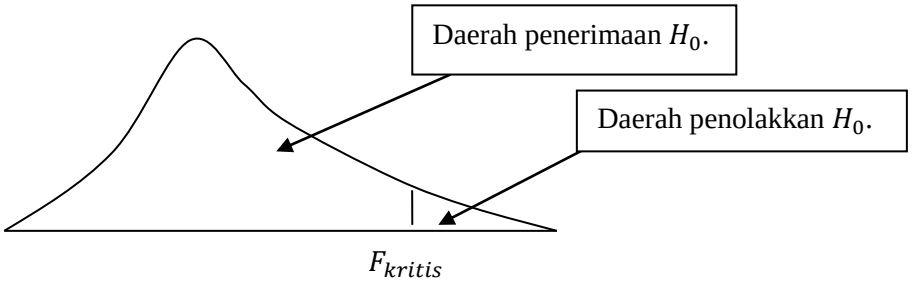
"*Levene's test tests null hypothesis that the variances in different groups are equal (i.e. the difference between the variances is zero).*"

Untuk pengambilan keputusan terhadap hipotesis dapat dilakukan dengan membandingkan nilai statistik dari uji Levene (L) terhadap nilai kritis dari tabel distribusi F (F_{kritis}). Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji Levene.

$$Jika\ L \leq nilai\ kritis\ F, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ L > nilai\ kritis\ F, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan membandingkan nilai probabilitas dari uji Levene terhadap tingkat signifikansi α (*significance level*).

$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$
$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$



Contoh Kasus Uji Kesamaan Varians Populasi dengan Uji Levene (Contoh Perhitungan)

Misalkan diberikan data mengenai nilai ujian matematika kelas 1,2, dan 3 SMA (Tabel 7.1). Berdasarkan data pada Tabel 7.1, X menyatakan nilai ujian matematika siswa kelas 1 SMA, Y menyatakan nilai ujian matematika siswa kelas 2 SMA, dan Z menyatakan nilai ujian siswa kelas 3 SMA. Berikut akan digunakan pendekatan uji Levene untuk menguji apakah asumsi populasi X, Y, dan Z memiliki varians yang sama (secara statistika), dapat diterima atau tidak, pada tingkat signifikansi 5%.

Tabel 7.1 (Data Fiktif)

Nilai Ujian Matematika		
X	Y	Z
70	80	70
80	85	87
87	70	90
77	77	77
80	85	76
	60	87
	80	

Tabel 7.2 menyajikan proses perhitungan untuk memperoleh nilai statistik dari uji Levene (L).

Tabel 7.2

	X	Y	Z	$a = \lvert X - \bar{X}\rvert$	$b = \lvert Y - \bar{Y}\rvert$	$c = \lvert Z - \bar{Z}\rvert$
	70	80	70	8,8	3,28571429	11,16666667
	80	85	87	1,2	8,28571429	5,833333333
	87	70	90	8,2	6,71428571	8,833333333
	77	77	77	1,8	0,28571429	4,166666667
	80	85	76	1,2	8,28571429	5,166666667
		60	87		16,7142857	5,833333333
		80			3,28571429	
Jumlah	394	537	487	21,2	46,8571429	41
Rata-rata	78,8	76,71429	81,16667	4,24	6,69387755	6,833333333

	$d=(a-\bar{a})^2$	$e=\left(b-\bar{b}\right)^2$	$f=(c-\bar{c})^2$
	20,7936	11,61557684	18,77777778
	9,2416	2,53394419	1
	15,6816	0,000416493	4
	5,9536	41,06455643	7,111111111
	9,2416	2,53394419	2,777777778
		100,4085798	1
		11,61557684	
Jumlah	60,912	169,7725948	34,66666667
Rata-rata			

➔Menghitung rata-rata gabungan dari data a, b, dan c.

$$\bar{X}_{a,b,c}=\frac{\sum a+\sum b+\sum c}{n_a+n_b+n_c}$$

$$\bar{X}_{a,b,c}=\frac{21{,}2+46{,}8571429+41}{5+7+6}$$

$$\bar{X}_{a,b,c}=6{,}05873.$$

➔Menghitung nilai statistik dari uji Levene (L).

$$L=\frac{\frac{n_a\left(\bar{X}_a-\bar{X}_{a,b,c}\right)^2+n_b\left(\bar{X}_b-\bar{X}_{a,b,c}\right)^2+n_c\left(\bar{X}_c-\bar{X}_{a,b,c}\right)^2}{(k-1)}}{\frac{(\sum d+\sum e+\sum f)}{(N-k)}}.$$

$$n_a\left(\bar{X}_a-\bar{X}_{a,b,c}\right)^2=(5)(4{,}24-6{,}05873)^2=16{,}5389$$

$$n_b\left(\bar{X}_b-\bar{X}_{a,b,c}\right)^2=(7)(6.69387755-6{,}05873)^2=2{,}823885$$

$$n_c\left(\bar{X}_c-\bar{X}_{a,b,c}\right)^2=(6)(6.833333333-6{,}05873)^2=3{,}60006$$

$$L=\frac{\frac{16{,}5389+2{,}823885+3{,}60006}{3-1}}{\frac{60{,}912+169{,}7725948+34{,}66667}{18-3}}$$

$$L=\frac{\frac{22{,}96284}{2}}{\frac{265{,}3513}{15}}$$

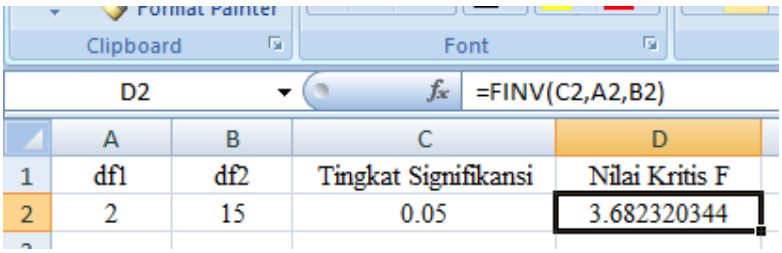
$$L=0{,}64903148.$$

➔Menghitung nilai kritis F.

Berikut rumus untuk menghitung nilai derajat bebas pembilang dan derajat bebas penyebut.

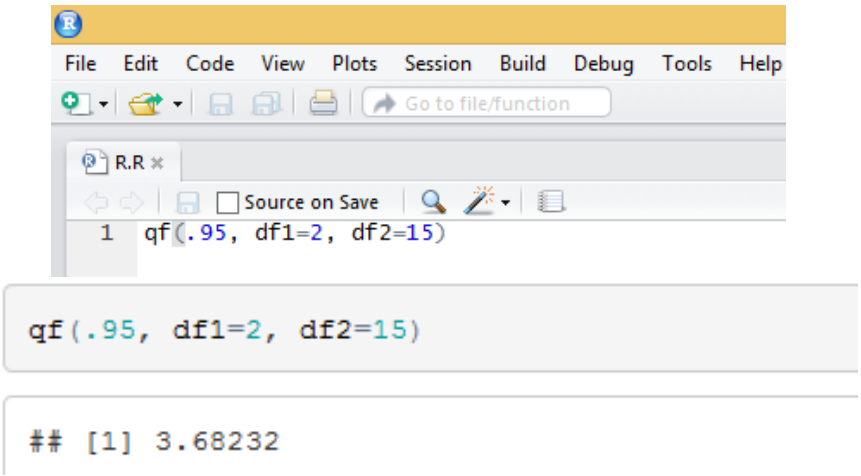
$$Derajat\ bebas\ pembilang = k - 1.$$
$$Derajat\ bebas\ penyebut = N - k.$$

Perhatikan bahwa k menyatakan banyaknya sampel, sedangkan N merupakan jumlah elemen atau pengamatan dari seluruh sampel. Diketahui nilai k adalah 3, sedangkan nilai N adalah 18 ($n_1 + n_2 + n_3 = 5 + 7 + 6 = 18$). Diketahui tingkat signifikansi yang digunakan adalah 5%, sehingga nilai kritis F dengan derajat bebas pembilang $3 - 1 = 2$, derajat bebas penyebut $18 - 3 = 15$, dan tingkat signifikansi 5% adalah 3,68.



	A	B	C	D
1	df1	df2	Tingkat Signifikansi	Nilai Kritis F
2	2	15	0.05	3.682320344

Gambar 7.3 Menentukan Nilai Kritis F dengan *Microsoft Excel*



Gambar 7.4 Menentukan Nilai Kritis F dengan R

➔Pengambilan keputusan terhadap hipotesis.

Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji Levene.

$$Jika\ L \leq nilai\ kritis\ F, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ L > nilai\ kritis\ F, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

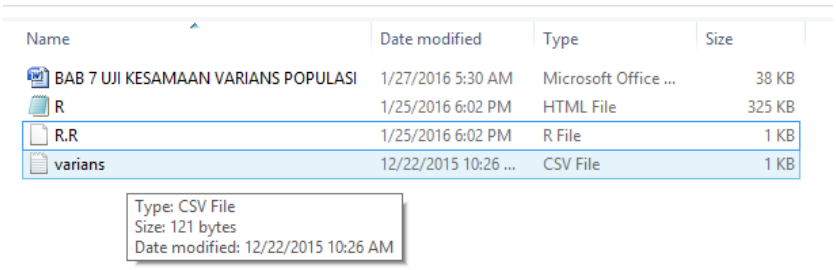
Perhatikan bahwa karena nilai statistik dari uji Levene, yakni 0,649, lebih kecil dibandingkan nilai kritis F, yakni 3,68, maka hipotesis nol diterima dan hipotesis alternatif ditolak, sehingga asumsi bahwa sampel X, Y, dan Z berasal dari populasi-populasi yang memiliki varians populasi yang sama, dapat diterima pada tingkat signifikansi 5%.

Penyelesaian dalam R untuk Uji Kesamaan Varians Populasi dengan Uji Levene

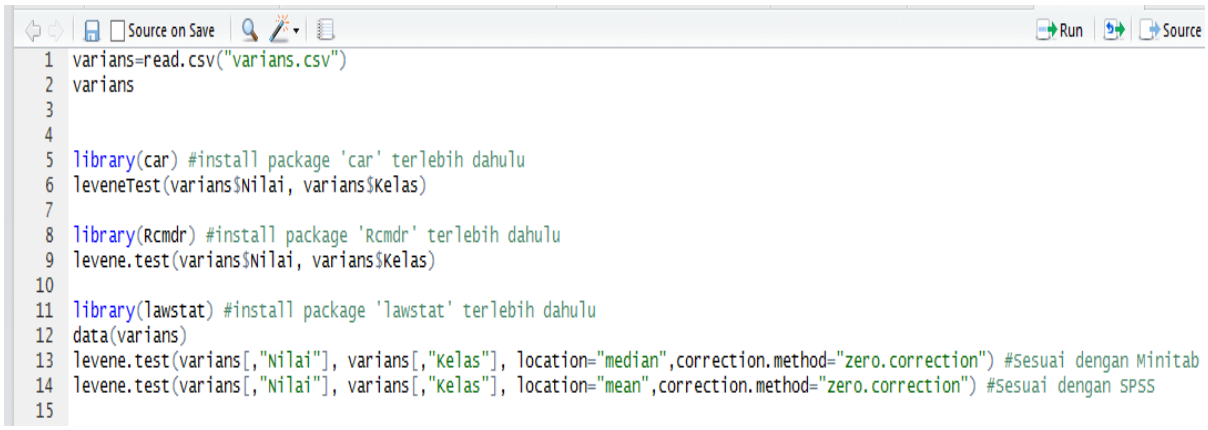
Data terlebih dahulu dibuat dalam *Microsoft Excel* (Gambar 7.5) dan disimpan dengan format tipe **.csv** (Gambar 7.6). Ketik kode R seperti pada Gambar 7.7. Kemudian *Compile* dan pilih HTML (Gambar 7.8). Hasilnya seperti pada Gambar 7.9 dan Gambar 7.10.

	A	B
1	Nilai	Kelas
2	70	1
3	80	1
4	87	1
5	77	1
6	80	1
7	80	2
8	85	2
9	70	2
10	77	2
11	85	2
12	60	2
13	80	2
14	70	3
15	87	3
16	90	3
17	77	3
18	76	3
19	87	3
20		

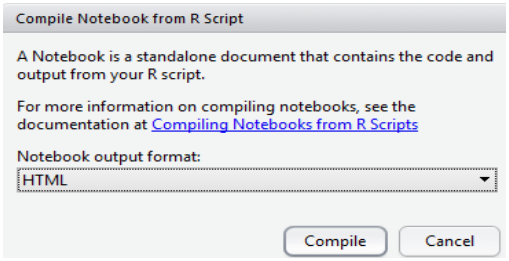
Gambar 7.5



Gambar 7.6



Gambar 7.7



Gambar 7.8

```
varians=read.csv("varians.csv")
varians
```

```
##    Nilai Kelas
## 1     70     1
## 2     80     1
## 3     87     1
## 4     77     1
## 5     80     1
## 6     80     2
## 7     85     2
## 8     70     2
## 9     77     2
## 10    85     2
## 11    60     2
## 12    80     2
## 13    70     3
## 14    87     3
## 15    90     3
## 16    77     3
## 17    76     3
## 18    87     3
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.4267 0.6604
##       15
```

Gambar 7.9

```
levene.test(varians[,"Nilai"], varians[,"Kelas"], location="median",correction.method="zero.correction") #Sesua
i dengan Minitab
```

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median with modified structural zero
##  removal method and correction factor
##
## data:  varians[, "Nilai"]
## Test Statistic = 0.4372, p-value = 0.6557
```

```
levene.test(varians[,"Nilai"], varians[,"Kelas"], location="mean",correction.method="zero.correction") #Sesuai
dengan SPSS
```

```
##
##  classical Levene's test based on the absolute deviations from the
##  mean ( zero.correction not applied because the location is not
##  set to median )
##
## data:  varians[, "Nilai"]
## Test Statistic = 0.64903, p-value = 0.5366
```

Gambar 7.10

Perhatikan Gambar 7.10. Nilai statistik dari uji Levene dengan pendekatan *Location = "median"* adalah 0,4372, yang mana hasil ini sama dengan hasil Minitab. Namun nilai statistik dari uji Levene dengan pendekatan *Location = "mean"* adalah 0,649, yang mana hasil ini sama dengan hasil SPSS.

Diketahui juga berdasarkan Gambar 7.10 nilai probabilitas (*p-value*) adalah 0,5366, yakni lebih besar dibandingkan tingkat signifikansi 0,05, maka hipotesis nol diterima dan hipotesis alternatif ditolak, sehingga asumsi bahwa sampel *X*, *Y*, dan *Z* berasal dari populasi-populasi yang memiliki varians populasi yang sama, dapat diterima pada tingkat signifikansi 5%.

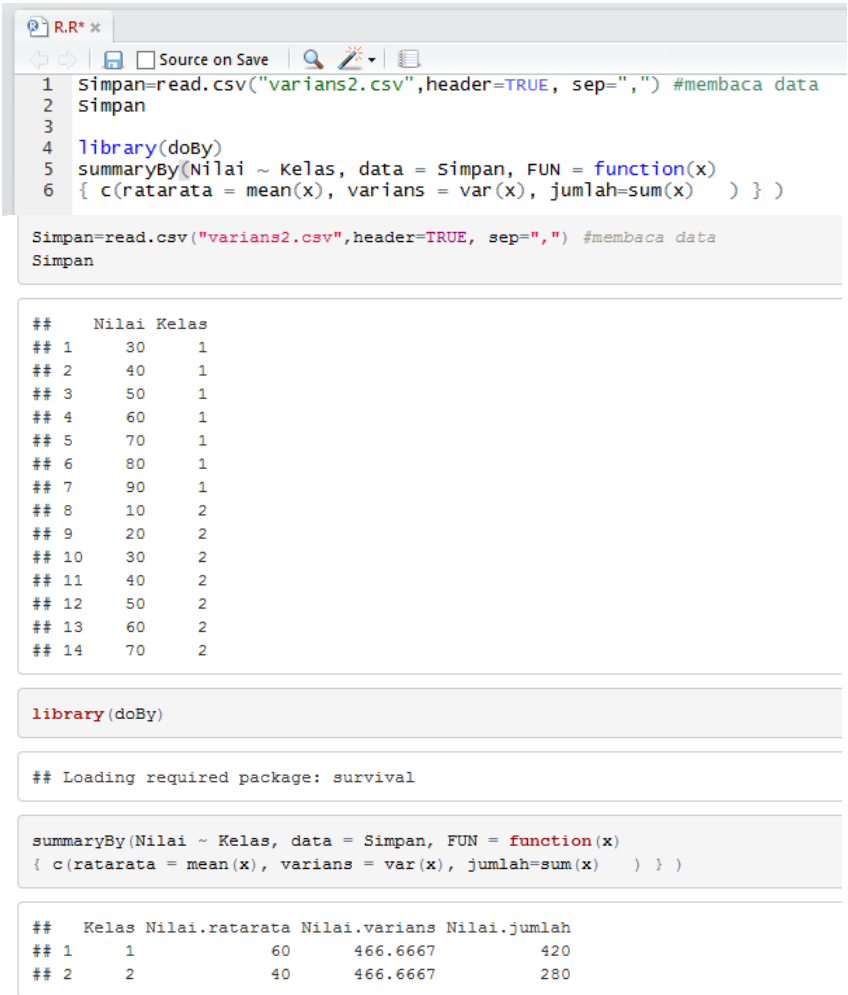
Contoh Kasus 2, Uji Kesamaan Varians Populasi dengan Uji Levene (Contoh Perhitungan dan Penyelesaian dengan R)

Misalkan diberikan data mengenai nilai ujian matematika kelas 1 dan 2 SMA (Tabel 7.3). Berdasarkan data pada Tabel 7.3, X menyatakan nilai ujian matematika siswa kelas 1 SMA, dan Y menyatakan nilai ujian matematika siswa kelas 2 SMA. Berikut akan digunakan pendekatan uji Levene untuk menguji apakah asumsi populasi X, Y, dan Z memiliki varians yang sama, dapat diterima atau tidak, pada tingkat signifikansi 5%.

Tabel 7.3 (Data Fiktif)

X	Y
30	10
40	20
50	30
60	40
70	50
80	60
90	70

Perhatikan bahwa sudah bisa diduga atau ditebak bahwa hipotesis nol diterima, yakni sampel X dan sampel Y ditarik dari populasi-populasi yang memiliki varians (*variance*) yang sama. Hal ini dikarenakan nilai nilai varians dari X dan Y **bernilai sama,** yakni 466,67 (lihat Gambar 7.11).



Gambar 7.11

Berdasarkan Gambar 7.11, diketahui nilai varians (*variance*) dari sampel X dan sampel Y, masing-masing adalah 466,6667. Tabel 7.4 menyajikan proses perhitungan untuk memperoleh nilai statistik dari uji Levene (L).

Tabel 7.4

	X	Y	$a = \lvert X - \bar{X} \rvert$	$b = \lvert Y - \bar{Y} \rvert$	$c = (a - \bar{a})^2$	$d = \left(b - \bar{b}\right)^2$
	30	10	30	30	165,3061	165,3061
	40	20	20	20	8,163265	8,163265
	50	30	10	10	51,02041	51,02041
	60	40	0	0	293,8776	293,8776
	70	50	10	10	51,02041	51,02041
	80	60	20	20	8,163265	8,163265
	90	70	30	30	165,3061	165,3061
Rata-Rata	60	40	17,14285714	17,14286	106,1224	106,1224
Jumlah	420	280	120	120	742,8571	742,8571

➔Menghitung rata-rata gabungan dari data a dan b.

$$\bar{X}_{a,b} = \frac{\sum a + \sum b}{n_a + n_b} = \frac{120 + 120}{7 + 7} = 17{,}14285714.$$

➔Menghitung nilai statistik dari uji Levene (L).

$$L = \frac{\dfrac{n_a\left(\bar{X}_a - \bar{X}_{a,b}\right)^2 + n_b\left(\bar{X}_b - \bar{X}_{a,b}\right)^2}{(k-1)}}{\dfrac{(\sum c + \sum d)}{(N-k)}}.$$

$$n_a\left(\bar{X}_a - \bar{X}_{a,b}\right)^2 = (7)(17{,}1428 - 17{,}1428)^2 = 0$$

$$n_b\left(\bar{X}_b - \bar{X}_{a,b}\right) = (7)(17{,}1428 - 17{,}1428)^2 = 0$$

$$L = \frac{\dfrac{0 + 0}{2 - 1}}{\dfrac{742{,}8571 + 742{,}8571}{14 - 2}}$$

$$L = 0.$$

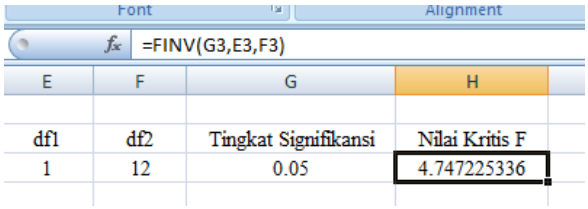
➔Menghitung nilai kritis F.

Berikut rumus untuk menghitung nilai derajat bebas pembilang dan derajat bebas penyebut.

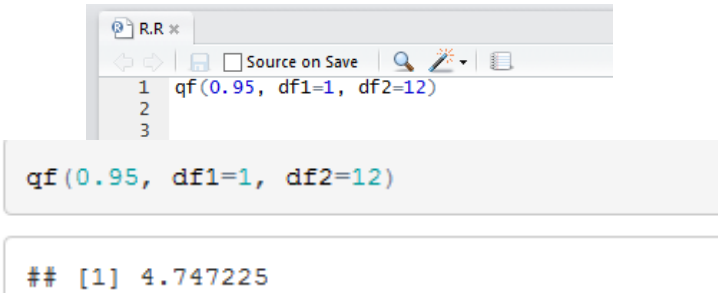
$$Derajat\ bebas\ pembilang = k - 1.$$
$$Derajat\ bebas\ penyebut = N - k.$$

Perhatikan bahwa k menyatakan banyaknya sampel, sedangkan N merupakan jumlah elemen atau pengamatan dari seluruh sampel. Diketahui nilai k adalah 2, sedangkan nilai N adalah 14 ($n_1 + n_2 = 7 + 7 = 14$). Diketahui tingkat signifikansi yang digunakan adalah 5%,

sehingga nilai kritis F dengan derajat bebas pembilang $2 - 1 = 1$, derajat bebas penyebut $14 - 2 = 12$, dan tingkat signifikansi 5% adalah 4,747.



Gambar 7.12 Menentukan Nilai Kritis F dengan *Microsoft Excel*



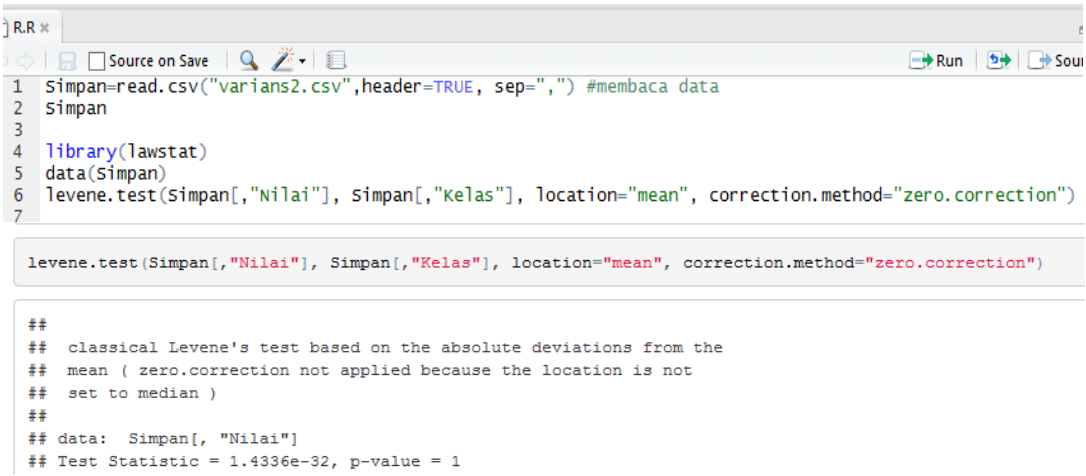
Gambar 7.13 Menentukan Nilai Kritis F dengan R

➔Pengambilan keputusan terhadap hipotesis.

Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji Levene.

$$Jika\ L \leq nilai\ kritis\ F, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ L > nilai\ kritis\ F, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Perhatikan bahwa karena nilai statistik dari uji Levene, yakni 0, lebih kecil dibandingkan nilai kritis F, yakni 4,747, maka hipotesis nol diterima dan hipotesis alternatif ditolak, sehingga asumsi bahwa sampel X dan sampel Y berasal dari populasi-populasi yang memiliki varians populasi yang sama, dapat diterima pada tingkat signifikansi 5%. Gambar 7.14 menyajikan hasil penyelesaian dengan R.



Gambar 7.14

Perhatikan Gambar 7.14. Nilai statistik dari uji Levene dengan pendekatan *Location = "mean"* adalah $1{,}4336 \times 10^{-32} = 0{,}0000000$ ..... Diketahui juga berdasarkan Gambar 7.14 nilai probabilitas (*p-value*) adalah 1, yakni lebih besar dibandingkan tingkat signifikansi 0,05, maka hipotesis nol diterima dan hipotesis alternatif ditolak, sehingga asumsi bahwa sampel *X* dan sampel *Y* berasal dari populasi-populasi yang memiliki varians populasi yang sama, dapat diterima pada tingkat signifikansi 5%.

.

Referensi

1. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

2. Gamst, G., L.S. Meyers dan A.J. Guarino. 2008. *Analysis of Variance Designs*. New York: Cambridge University Press.

3. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

4. Ott, R.L. dan M. Longnecker. 2001. *An Introduction to Statistical Methods and Data Analysis, 5th Edition*. United States of America: Duxbury.

5. https://cran.r-project.org/web/packages/lawstat/lawstat.pdf

6. https://cran.r-project.org/web/packages/doBy/doBy.pdf

7. https://cran.r-project.org/web/packages/car/car.pdf

8. https://cran.r-project.org/web/packages/Rcmdr/index.html

BAB 8

UJI KESAMAAN RATA-RATA DARI DUA POPULASI UNTUK DATA BERPASANGAN DAN SALING BERHUBUNGAN (UJI t)

Uji Kesamaan Rata-Rata dari Dua Populasi untuk Data Berpasangan dan Saling Berhubungan dengan Uji t *(Paired t Test for Dependent Populations)*

Dalam uji kesamaan rata-rata dari dua populasi untuk data berpasangan dan saling berhubungan dengan uji t, pengamatan-pengamatan dari dua populasi dinyatakan dalam berpasangan. Sebagai contoh misalkan $(X_1, Y_1), (X_2, Y_2), \dots, (X_k, Y_k)$ merupakan pengamatan-pengamatan dari dua populasi, yakni populasi X dan Y yang dinyatakan dalam berpasangan.

Berikut beberapa contoh kasus yang dapat diselesaikan dengan pendekatan uji kesamaan rata-rata dari dua populasi untuk data berpasangan dan saling berhubungan dengan uji t.

- ⇨ Menguji ada tidaknya pengaruh yang signifikan secara statistika penggunaan suplemen X terhadap berat badan, sebelum dan sesudah mengkonsumsi suplemen X selama satu minggu.

- ⇨ Menguji ada tidaknya pengaruh yang signifikan secara statistika penggunaan suplemen Y terhadap tinggi badan, sebelum dan sesudah mengkonsumsi suplemen Y selama satu bulan.

- ⇨ Menguji ada tidaknya pengaruh yang signifikan secara statistika pada program kursus matematika terhadap nilai ujian matematika siswa, sebelum dan sesudah mengikuti kursus matematika.

Misalkan D_i menyatakan selisih dari pasangan pengamatan ke-i dari dua populasi, yakni X dan Y, maka $D_1 = Y_1 - X_1, D_2 = Y_2 - X_2, \dots, D_k = Y_k - X_k$. Dalam uji kesamaan rata-rata dari dua populasi untuk data berpasangan dan saling berhubungan dengan uji t, data dari selisih pasangan pengamatan (D) diasumsikan berdistribusi normal, dengan rata-rata μ_D.

Dalam uji kesamaan rata-rata dari dua populasi untuk data berpasangan dan saling berhubungan dengan uji t, hipotesis nol menyatakan tidak terdapat pengaruh yang signifikan secara statistika, sesudah dan sebelum perlakuan. Dengan kata lain, selisih rata-rata antara kelompok sesudah dan sebelum perlakuan sama dengan nol $(\mu_2 - \mu_1 = 0)$. Hipotesis alternatif menyatakan terdapat pengaruh yang signifikan secara statistika, sesudah dan sebelum perlakuan. Dengan kata lain, selisih rata-rata antara kelompok sesudah dan sebelum perlakuan berbeda dari nol $(\mu_2 - \mu_1 \neq 0)$. Nilai statistik dari uji t $\left(t_{hitung}\right)$ dihitung dengan rumus sebagai berikut.

$$t = \frac{\bar{d} - \mu_D}{s_d/\sqrt{n}}.$$

Perhatikan bahwa $\bar{d}$ merupakan rata-rata dari selisih pasangan pengamatan dari dua sampel, μ_D merupakan rata-rata dari selisih pasangan pengamatan dari dua populasi, serta s_d merupakan nilai standar deviasi dari selisih pasangan pengamatan dari dua sampel. Berikut rumus untuk menghitung nilai s_d.

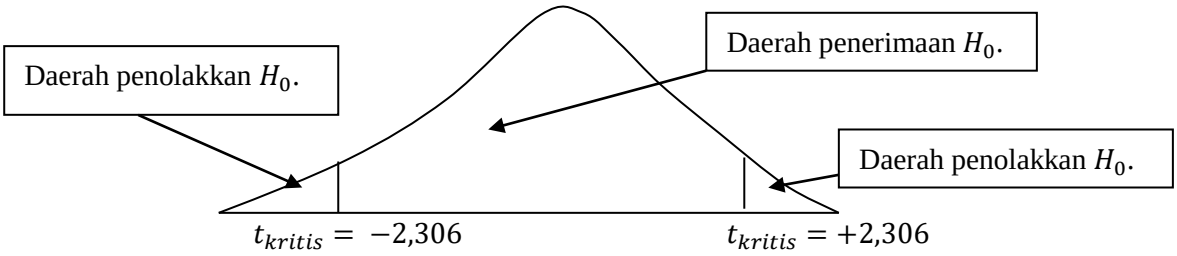
$$s_d = \sqrt{\frac{\sum(d-\bar{d})^2}{n-1}}.$$

Untuk pengambilan keputusan terhadap hipotesis, dapat dilakukan dengan membandingkan nilai statistik dari uji t terhadap nilai kritis berdasarkan tabel distribusi t (t_{kritis}). Sebelum menghitung nilai kritis t, terlebih dahulu menghitung nilai derajat bebas. Berikut rumus untuk menghitung nilai derajat bebas.

$$Derajat\ bebas = n - 1.$$

Perhatikan bahwa n menyatakan banyaknya pasangan pengamatan. Andaikan banyaknya pasangan pengamatan sebanyak 9, tingkat signifikansi yang digunakan adalah 5%, sehingga nilai kritis t dengan derajat bebas $9 - 1 = 8$ dan tingkat signifikansi 5% adalah $\pm 2{,}306$. Diketahui nilai kritis $t = \pm 2{,}306$. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji t (pengujian dua arah).

$$Jika\ |t_{hitung}| \leq |t_{kritis}|, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ |t_{hitung}| > |t_{kritis}|, maka\ H_0\ ditolak\ dan\ H_1\ ditolak.$$



Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan menggunakan nilai probabilitas dari uji t. Nilai probabilitas dari uji t dibandingkan dengan tingkat signifikansi yang digunakan. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan pendekatan nilai probabilitas.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Uji Asumsi Normalitas

Dalam uji kesamaan rata-rata dari dua populasi untuk data berpasangan dan saling berhubungan dengan uji t, data dari **selisih pasangan pengamatan** $(\boldsymbol{D})$ **diasumsikan berdistribusi normal, dengan rata-rata** $\boldsymbol{\mu_D}$. Field (2009:326) menyatakan sebagai berikut.

"*The sampling distribution is normally distributed. In the dependent t-test this means that the sampling distribution of the* ***differences between scores should be normal****, not the scores themselves (see section 9.4.3)*".

Sejalan dengan Field, Mann dan Lacke (2011:465) menyatakan sebagai berikut.

"*If the sample size is small, then the population of paired differences is normally distributed*".

Lebih lanjut, Mann dan Lacke (2011:465) menyatakan sebagai berikut.

"*However, usually* σ_d *is* ***never known****. Then ,if the standard deviation* σ_d *of the population paired differences is unknown and* ***either the sample size is large (i.e.,*** $\boldsymbol{n \geq 30}$***) or the population of paired differences is normally distributed (with*** $\boldsymbol{n < 30}$***), then the t distribution is used to make a confidence interval and test hypothesis about*** $\boldsymbol{\mu_d}$."

Namun ketika ukuran sampel cukup besar, yakni $\geq$ **30**, maka populasi tidak harus berdistribusi normal (Mann dan Lacke, 2011:465). Hal ini karena berdasarkan sifat teorema limit sentral (*central limit theorem*). Untuk menguji asumsi normalitas tersebut, dapat digunakan pendekatan grafik, yakni *Q-Q plot*. Pada pendekatan *Q-Q plot*, jika titik-titik (*dots*) menyebar jauh (menyebar jauh berliku-liku pada garis diagonal seperti ular) dari garis diagonal, maka diindikasi asumsi normalitas tidak dipenuhi. Jika titik-titik menyebar sangat dekat pada garis diagonal, maka asumsi normalitas dipenuhi. Di samping itu, dapat juga digunakan pendekatan uji Kolmogorov-Smirnov atau uji Jarque-Bera, untuk menguji asumsi normalitas.

Dalam pendekatan uji Kolmogorov-Smirnov atau uji Jarque-Bera, data dari selisih pasangan pengamatan diuji normalitasnya. Hipotesis nol menyatakan data dari selisih pasangan pengamatan (D) berdistribusi normal, sedangkan hipotesis alternatif menyatakan data dari selisih pasangan pengamatan (D) tidak berdistribusi normal.

Untuk pengambilan keputusan terhadap hipotesis, dapat dibandingkan antara nilai probabilitas dari uji Kolmogorov-Smirnov atau uji Jarque-Bera, dengan tingkat signifikansi yang digunakan (α). Berikut aturan pengambilan keputusan terhadap hipotesis.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, H_0\ ditolak\ dan\ H_1\ diterima.$$

Contoh Kasus Uji Kesamaan Rata-Rata dari Dua Populasi untuk Data Berpasangan dan Saling Berhubungan dengan Uji t *(Contoh Perhitungan)*

Misalkan seorang peneliti ingin meneliti mengenai pengaruh penggunaan obat A terhadap jumlah denyut jantung per-menit pada manusia. Peneliti tersebut mengambil sampel sebanyak 9 responden. Pertama, sebelum pemberian obat A, peneliti mencatat jumlah denyut jantung yang terjadi dalam satu menit dari 9 responden tersebut. Kemudian, 9 responden tersebut mengkonsumsi obat A. Setelah 15 menit, peneliti tersebut mencatat kembali jumlah denyut jantung yang terjadi dalam satu menit. Berikut data dari 9 responden mengenai jumlah denyut jantung yang terjadi dalam satu menit sebelum dan sesudah mengkonsumsi obat A (Tabel 8.1).

Tabel 8.1 (Data Fiktif)

Responden	X	Y
1	78	100
2	75	95
3	67	70
4	77	90
5	70	90
6	72	90
7	78	89
8	74	90
9	77	100

Berdasarkan data pada Tabel 8.1, diketahui jumlah denyut jantung dalam satu menit dari responden ke-3 ketika belum mengkonsumsi obat A sebanyak 67, dan setelah mengkonsumsi obat A sebanyak 70. Peneliti akan menguji apakah terdapat pengaruh yang signifikan secara statistika dalam hal jumlah denyut jantung yang terjadi dalam satu menit, sebelum dan sesudah mengkonsumsi obat A pada tingkat signifikansi $\alpha = 5\%$. Berikut akan dihitung standar deviasi dari data selisih pasangan pengamatan s_d.

Tabel 8.2

	X	Y	$d = Y - X$	$d - \bar{d}$	$(d - \bar{d})^2$
	78	100	22	5,777778	33,38272
	75	95	20	3,777778	14,2716
	67	70	3	-13,2222	174,8272
	77	90	13	-3,22222	10,38272
	70	90	20	3,777778	14,2716
	72	90	18	1,777778	3,160494
	78	89	11	-5,22222	27,2716
	74	90	16	-0,22222	0,049383
	77	100	23	6,777778	45,93827
Jumlah	668	814	146		323,5556
Rata-Rata	74,22222	90,44444	16,22222		35,95062

$$s_d = \sqrt{\frac{\sum(d - \bar{d})^2}{n - 1}}$$

$$s_d = \sqrt{\frac{323{,}555556}{9 - 1}}$$

$$s_d = 6{,}35959468.$$

Berdasarkan perhitungan diperoleh nilai standar deviasi dari data selisih pasangan pengamatan, yakni $s_d = 6{,}360$. Selanjutnya akan dihitung nilai statistik dari uji t.

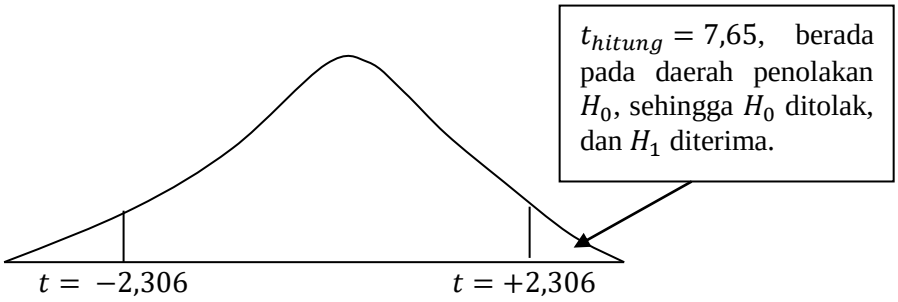
$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

$$t = \frac{16{,}2222 - 0}{6{,}35959468/\sqrt{9}}$$

$$t = 7{,}652468821.$$

Berdasarkan perhitungan, nilai statistik dari uji t adalah 7,652468821. Diketahui derajat bebas (*df*) bernilai $9 - 1 = 8$. Nilai kritis t dengan derajat bebas 8 dan tingkat signifikansi 5% adalah $\pm 2{,}306$. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji t.

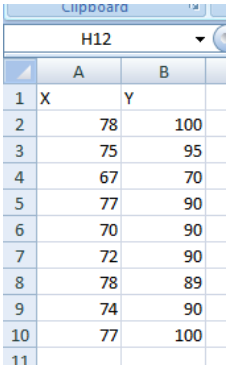
$$Jika\ |t_{hitung}| \leq |t_{kritis}|, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ |t_{hitung}| > |t_{kritis}|, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$



Perhatikan bahwa karena $|t_{hitung}| > |t_{kritis}|$, yakni 7,652 > 2,306, maka disimpulkan bahwa hipotesis nol ditolak dan hipotesis alternatif diterima. Hal ini berarti terdapat pengaruh yang signifikan secara statistika dalam hal jumlah denyut jantung, sebelum dan sesudah mengkonsumsi obat *A* pada tingkat signifikansi 5%.

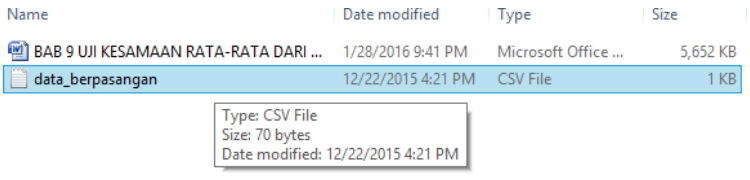
Penyelesaian dalam R untuk Uji Kesamaan Rata-Rata dari Dua Populasi untuk Data Berpasangan dan Saling Berhubungan dengan Uji t

Data terlebih dahulu dibuat dalam *Microsoft Excel* (Gambar 8.1) dan disimpan dengan format tipe **.csv** (Gambar 8.2). Ketik kode R seperti pada Gambar 8.3. Kemudian *Compile* dan pilih HTML. Hasilnya seperti pada Gambar 8.4 dan Gambar 8.5.

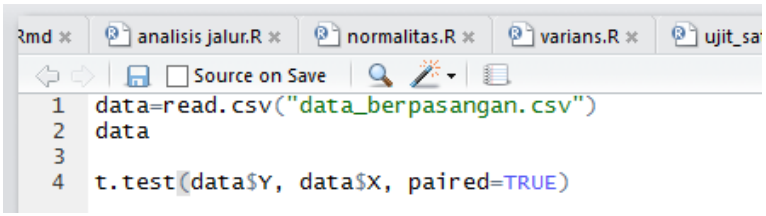


	A	B
1	X	Y
2	78	100
3	75	95
4	67	70
5	77	90
6	70	90
7	72	90
8	78	89
9	74	90
10	77	100
11		

Gambar 8.1



Gambar 8.2



Gambar 8.3

```
data=read.csv("data_berpasanga
data
```

```
##    X   Y
## 1 78 100
## 2 75  95
## 3 67  70
## 4 77  90
## 5 70  90
## 6 72  90
## 7 78  89
## 8 74  90
## 9 77 100
```

Gambar 8.4

```
t.test(data$Y, data$X, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  data$Y and data$X
## t = 7.6525, df = 8, p-value = 6.003e-05
## alternative hypothesis: true difference in means is n
## 95 percent confidence interval:
##  11.33381 21.11064
## sample estimates:
## mean of the differences
##                16.22222
```

Gambar 8.5

Berdasarkan Gambar 8.5, diketahui nilai statistik dari uji t (t) adalah 7,6525, sementara nilai probabilitas (*p-value*) adalah 0,00006003 (atau 6.003e-05). Berdasarkan Gambar 8.5, diketahui nilai derajat bebas (df) adalah 8. Perhatikan bahwa karena $|t_{hitung}| > |t_{kritis}|$, yakni 7,652 > 2,306, maka disimpulkan bahwa hipotesis nol ditolak dan hipotesis alternatif diterima. Hal ini berarti terdapat pengaruh yang signifikan secara statistika dalam hal jumlah denyut jantung, sebelum dan sesudah mengkonsumsi obat *A* pada tingkat signifikansi 5%.

Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan menggunakan nilai probabilitas dari uji t (*p-value*). Nilai probabilitas dari uji t dibandingkan dengan tingkat signifikansi yang digunakan. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan pendekatan nilai probabilitas.

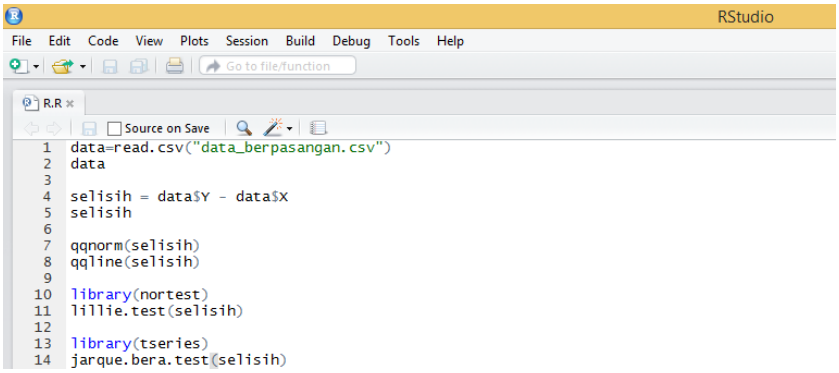
$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Berdasarkan Gambar 8.5, diketahui nilai probabilitas dari uji t (*p-value*) adalah 0,00006003. Karena nilai probabilitas tersebut lebih kecil dibandingkan tingkat signifikansi $\alpha = 0{,}05$, maka hipotesis nol ditolak dan hipotesis alternatif diterima. Hal ini berarti terdapat pengaruh yang signifikan secara statistika dalam hal jumlah denyut jantung, sebelum dan sesudah mengkonsumsi obat *A* pada tingkat signifikansi 5%.

Uji Asumsi Normalitas dalam R

Dalam uji kesamaan rata-rata dari dua populasi untuk data berpasangan dan saling berhubungan dengan uji t, data dari **selisih pasangan pengamatan** $(\boldsymbol{D})$ **diasumsikan berdistribusi normal, dengan rata-rata** $\boldsymbol{\mu_D}$. Gambar 8.6, disajikan kode R untuk uji

normalitas untuk data selisih pasangan pengamatan. Hasil eksekusi kode R pada Gambar 8.6, disajikan pada Gambar 8.7 hingga Gambar Gambar 8.10.



Gambar 8.6

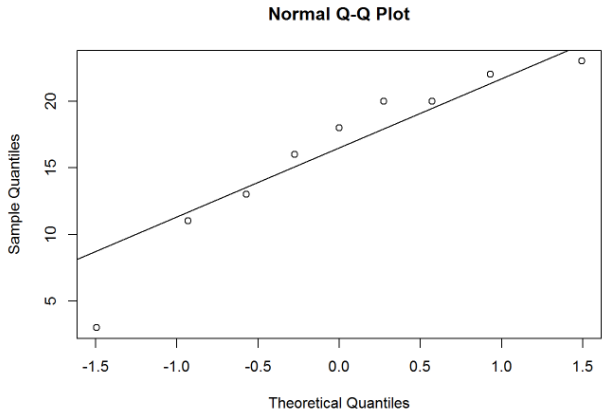
```
data=read.csv("data_berpasangan.csv")
data
```

```
##    X   Y
## 1 78 100
## 2 75  95
## 3 67  70
## 4 77  90
## 5 70  90
## 6 72  90
## 7 78  89
## 8 74  90
## 9 77 100
```

```
selisih = data$Y - data$X
selisih
```

```
## [1] 22 20  3 13 20 18 11 16 23
```

Gambar 8.7



Gambar 8.8

Berdasarkan Gambar 8.7, diperoleh nilai selisih untuk setiap pasangan nilai data. Pasangan nilai data pertama adalah $(X = 76, Y = 100)$, maka nilai selisihnya adalah $100 - 78 = 23$. Pasangan nilai data kedua adalah $(X = 75, Y = 95)$, maka nilai selisihnya adalah $95 - 75 = 20$.

```
lillie.test(selisih)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  selisih
## D = 0.1682, p-value = 0.6544
```

Gambar 8.9

```
jarque.bera.test(selisih)
```

```
##
##  Jarque Bera Test
##
## data:  selisih
## X-squared = 1.3935, df = 2, p-value = 0.4982
```

Gambar 8.10

Berdasarkan Gambar 8.8, perhatikan bahwa titik-titik menyebar sangat dekat pada garis diagonal, maka disimpulkan bahwa asumsi normalitas data dari selisih pasangan pengamatan dipenuhi. Perhatikan juga bahwa berdasarkan Gambar 8.9, nilai probabilitas dari uji Kolmogorov-Smirnov (*p-value*) adalah 0,6544, sementara berdasarkan Gambar 8.10, nilai probabilitas dari uji Jarque-Bera (*p-value*) adalah 0,4982. Karena masing-masing nilai probabilitas lebih besar dibandingkan tingkat signifikansi, yakni 0,05, maka hipotesis nol diterima, dan hipotesis alternatif ditolak. Hal ini berarti asumsi normalitas data dari selisih pasangan pengamatan dipenuhi.

Referensi

1. Agresti, A. dan B. Finlay. 2009. *Statistical Methods for the Social Sciences, 4th Edition.* United States of America: Prentice Hall.

2. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

3. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

4. Mann, P. S. dan C.J. Lacke. 2011. *Introductory Statistics, International Student Version, 7th Edition*, Asia: John Wiley & Sons, Inc.

5. Montgomery, D. C. dan G. C. Runger. 2011. *Applied Statistics and Probability for Engineers, 5th Edition*. United States of America: John Wiley & Sons, Inc.

6. Smidth, R. K. dan D. H. Sanders. 2000. *Statistics a First Course, 6th Edition*, United States of America: McGraw-Hill Companies.

7. http://www.r-bloggers.com/paired-students-t-test/

BAB 9

UJI KESAMAAN RATA-RATA DARI DUA POPULASI TIDAK BERHUBUNGAN, DENGAN ASUMSI VARIANS POPULASI SAMA (UJI t)

Uji Kesamaan Rata-Rata dari Dua Populasi yang Tidak Berhubungan (Independen) dengan Asumsi Varians yang Sama (t Test for Independent Populations with Assumption $\sigma_1^2 = \sigma_2^2$)

Dalam uji kesamaan rata-rata dari dua populasi yang tidak berhubungan dengan asumsi varians yang sama, menguji ada tidaknya perbedaan rata-rata antara populasi pertama dan populasi kedua. Dengan kata lain, menguji apakah selisih rata-rata antara kelompok kedua dan pertama berbeda atau sama dengan nol. Dalam uji ini, pengamatan-pengamatan pada populasi pertama saling bebas atau independen dengan pengamatan-pengamatan pada populasi kedua (*independent populations*). **Uji ini didasarkan pada ketidaktahuan (*unknown*) mengenai nilai varians dari dua populasi**, **namun diasumsikan varians dari dua populasi tersebut sama.**

Berikut beberapa contoh kasus yang dapat diselesaikan dengan pendekatan uji kesamaan rata-rata dari dua populasi independen dengan asumsi varians yang sama dengan uji t.

- ⇨ Menguji ada tidaknya perbedaan (perbedaan yang signifikan secara statistika) nilai indeks prestasi (secara rata-rata) antara mahasiswa laki-laki dan perempuan.
- ⇨ Menguji ada tidaknya perbedaan harga saham antara perusahaan manufaktur dan *real estate*.
- ⇨ Menguji ada tidaknya perbedaan uang jajan antara mahasiswa kedokteran dan mahasiswa matematika.
- ⇨ Menguji ada tidaknya perbedaan indeks prestasi antara mahasiswa dominan otak kanan dan dominan kotak kiri.

Dalam uji kesamaan rata-rata dari dua populasi yang tidak berhubungan dengan asumsi varians yang sama, hipotesis nol menyatakan tidak terdapat perbedaan rata-rata antara populasi pertama dan populasi kedua. Dengan kata lain, selisih rata-rata antara populasi kedua dan pertama sama dengan nol $(\mu_2 - \mu_1 = 0)$. Hipotesis alternatif menyatakan terdapat perbedaan rata-rata antara populasi pertama dan populasi kedua. Dengan kata lain, selisih rata-rata antara populasi kedua dan pertama berbeda dari nol $(\mu_2 - \mu_1 \neq 0)$. Nilai statistik dari uji t (t_{hitung}) dihitung dengan rumus sebagai berikut.

$$t = \frac{\bar{X}_2 - \bar{X}_1}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

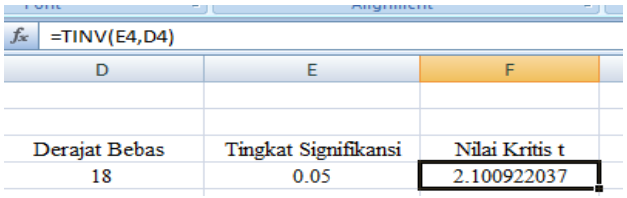
Perhatikan bahwa t merupakan nilai statistik dari uji t, $\bar{X}_1$ merupakan nilai rata-rata dari sampel pertama, $\bar{X}_2$ merupakan nilai rata-rata dari sampel kedua, n_1 merupakan jumlah pengamatan dalam sampel pertama, dan n_2 merupakan jumlah pengamatan dalam sampel kedua. Berikut rumus untuk menghitung s_p.

$$s_p = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}.$$

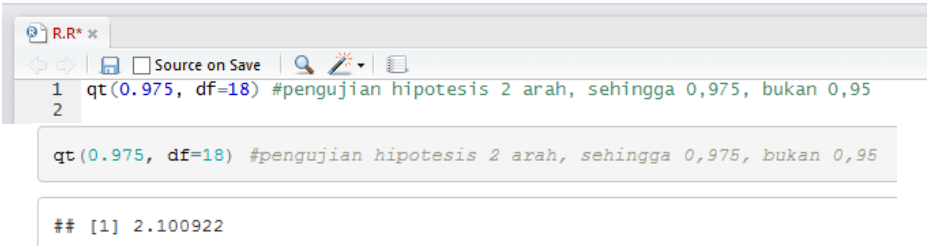
Perhatikan bahwa s_p disebut *pooled estimator standard deviation for two samples*, yang mana merupakan estimator dari σ. Untuk pengambilan keputusan terhadap hipotesis, dapat dilakukan dengan membandingkan nilai statistik dari uji t terhadap nilai kritis t (t_{kritis}). Sebelum menghitung nilai kritis t, terlebih dahulu menghitung nilai derajat bebas. Berikut rumus untuk menghitung nilai derajat bebas.

$$Derajat\ bebas = n_1 + n_2 - 2.$$

Perhatikan bahwa n_1 menyatakan banyaknya pengamatan/elemen pada sampel pertama, n_2 menyatakan banyaknya pengamatan/elemen pada sampel kedua. Andaikan $n_1 = n_2 = 10$ dan tingkat signifikansi yang digunakan $\alpha = 5\%$, maka nilai kritis t adalah $\pm 2{,}101$.



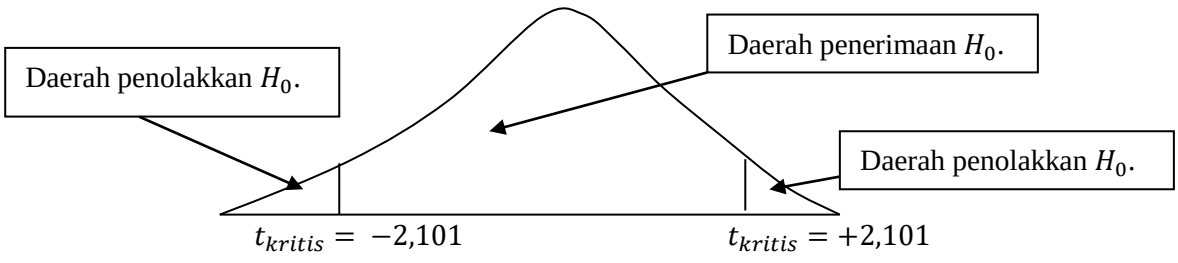
Gambar 9.1 Menentukan Nilai Kritis *t* dengan *Microsoft Excel*



Gambar 9.2 Menentukan Nilai Kritis *t* dengan R

Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji t.

$$Jika\ |t_{hitung}| \leq |t_{kritis}|, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ |t_{hitung}| > |t_{kritis}|, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$



Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan menggunakan pendekatan nilai probabilitas dari uji t. Nilai probabilitas dari uji t dibandingkan dengan tingkat signifikansi yang digunakan. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan nilai probabilitas.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Uji Asumsi Normalitas

Dalam uji kesamaan rata-rata dari dua populasi yang tidak berhubungan dengan asumsi varians yang sama, populasi pertama dan populasi kedua diasumsikan berdistribusi normal (Mann dan Lacke, 2011:448). Namun ketika ukuran sampel cukup besar, yakni masing-masing sampel berukuran ≥ 30, maka populasi tidak harus berdistribusi normal (Mann dan Lacke, 2011:465). Untuk menguji asumsi normalitas tersebut, dapat digunakan pendekatan grafik, yakni *Q-Q plot*. Pada pendekatan *Q-Q plot*, jika titik-titik (*dots*) menyebar jauh (menyebar jauh berliku-liku pada garis diagonal seperti ular) dari garis diagonal, maka diindikasi asumsi normalitas tidak dipenuhi. Jika titik-titik menyebar sangat dekat pada garis diagonal, maka asumsi normalitas dipenuhi. Di samping itu, dapat juga digunakan pendekatan uji Kolmogorov-Smirnov atau uji Jarque-Bera, untuk menguji asumsi normalitas. Hipotesis nol menyatakan data sampel ditarik dari populasi yang berdistribusi normal, sedangkan hipotesis alternatif menyatakan data sampel ditarik dari populasi yang tidak berdistribusi normal.

Untuk pengambilan keputusan terhadap hipotesis, dapat dibandingkan antara nilai probabilitas dari uji Kolmogorov-Smirnov atau uji Jarque-Bera, dengan tingkat signifikansi yang digunakan (α). Berikut aturan pengambilan keputusan terhadap hipotesis.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, H_0\ ditolak\ dan\ H_1\ diterima.$$

Uji Asumsi Kesamaan Varians

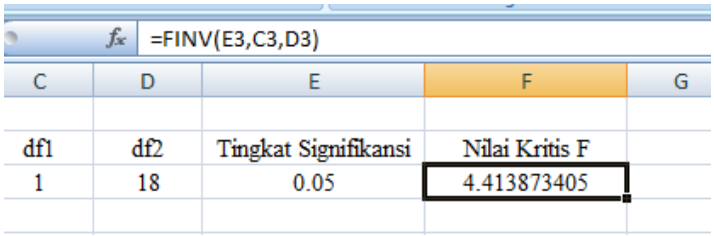
Selain asumsi normalitas, asumsi lain yang dikenakan adalah asumsi kesamaan varians, yakni sampel-sampel yang diteliti berasal dari populasi-populasi yang memiliki varians yang sama. Untuk menguji apakah sampel-sampel yang diteliti berasal dari populasi-populasi yang memiliki varians yang sama, dapat digunakan uji Levene. Pada uji Levene, hipotesis nol menyatakan sampel-sampel yang diambil berasal dari populasi-populasi yang memiliki varians yang sama, sedangkan hipotesis alternatif menyatakan paling tidak terdapat sepasang populasi yang memiliki varians yang berbeda.

Pengambilan keputusan terhadap hipotesis dilakukan dengan membandingkan nilai statistik dari uji Levene (L) dengan nilai kritis F (F_{kritis}). Sebelum menghitung nilai kritis F, terlebih dahulu menghitung nilai dari derajat bebas pembilang dan derajat bebas penyebut. Berikut rumus untuk menghitung nilai dari derajat bebas pembilang dan derajat bebas penyebut.

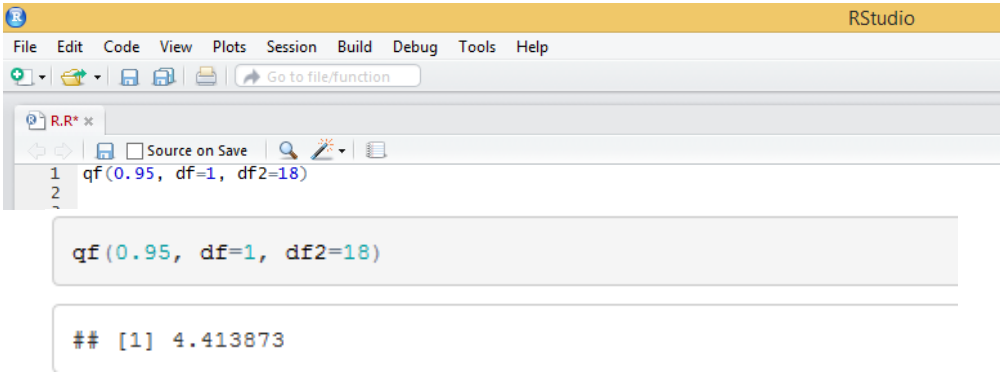
$$Derajat\ bebas\ pembilang = k - 1.$$
$$Derajat\ bebas\ penyebut = N - k.$$

Perhatikan bahwa k menyatakan banyaknya sampel/populasi yang diteliti, sedangkan N merupakan jumlah pengamatan/elemen dari seluruh sampel. Diketahui misalkan nilai k

adalah 2, sedangkan nilai N adalah 20 ($n_1 + n_2 = 10 + 10 = 20$). Misalkan tingkat signifikansi yang digunakan adalah 5%, sehingga nilai kritis F dengan derajat bebas pembilang $2 - 1 = 1$, derajat bebas penyebut $20 - 2 = 18$, dan tingkat signifikansi 5% adalah 4,41.



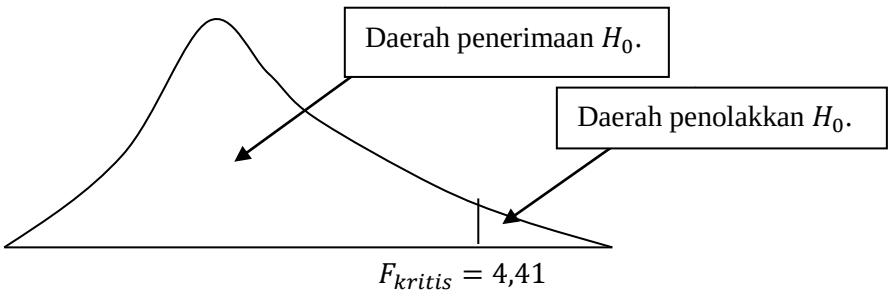
Gambar 9.3 Menentukan Nilai Kritis *F* dengan *Microsoft Excel*



Gambar 9.4 Menentukan Nilai Kritis F dengan R

Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji Levene (aturan distribusi F).

$$Jika\ L \leq F_{kritis}, maka\ \ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$jika\ L > F_{kritis}, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$



Pengambilan keputusan terhadap hipotesis dapat juga digunakan pendekatan nilai probabilitas dari uji Levene. Nilai probabilitas tersebut dibandingkan dengan tingkat signifikansi (α). Berikut aturan pengambilan keputusan terhadap hipotesis.

$$Jika\ nilai\ probabilitas \geq \alpha, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < \alpha, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Contoh Kasus Uji Kesamaan Rata-Rata dari Dua Populasi yang Tidak Berhubungan (Independen) dengan Asumsi Varians yang Sama (Contoh Perhitungan)

Misalkan seorang peneliti akan meneliti mengenai ada tidaknya perbedaan (secara rata-rata) nilai ujian matematika dasar antara mahasiswa laki-laki dan perempuan. Untuk keperluan penelitian, peneliti tersebut mengambil sampel sebanyak 20 nilai ujian matakuliah matematika dasar yang terdiri dari 10 nilai ujian mahasiswa laki-laki dan 10 nilai ujian mahasiswa perempuan. Data yang telah dikumpulkan disajikan dalam Tabel 9.1. Peneliti akan menguji apakah terdapat perbedaan (secara rata-rata) yang signifikan secara statistika dari nilai ujian matematika dasar antara mahasiswa laki-laki dan perempuan dengan tingkat signifikansi 5%.

Tabel 9.1 (Data Fiktif)

Nama Mahasiswa Laki-Laki	X	Nama Mahasiswa Perempuan	Y
Ugi	65	Ulan	85
Mifdhal	68	Fitri	75
Iqbal	70	Evelin	75
Alan	80	Melda	80
John	75	Dina	75
Andre	72	Suci	75
Ridho	65	Febri	75
Hanafi	60	Oshin	80
Romi	88	Wilya	90
Udin	70	Windy	85

Tabel 9.2

	X	Y
	65	85
	68	75
	70	75
	80	80
	75	75
	72	75
	65	75
	60	80
	88	90
	70	85
Rata-Rata	71,3	79,5
Standar Deviasi	8,097325	5,502525

Berdasarkan data pada Tabel 9.2, diketahui $\bar{X} = 71{,}3$; $\bar{Y} = 79{,}5$; $s_X = 8{,}097325$; $s_Y = 5{,}502525$, sehingga

$$s_p = \sqrt{\frac{s_X^2(n_X - 1) + s_Y^2(n_Y - 1)}{n_X + n_Y - 2}}$$

$$s_p = \sqrt{\frac{(8{,}097325)^2(10-1) + (5{,}502525)^2(10-1)}{10+10-2}}$$

$$s_p = 6{,}922588.$$

Nilai statistik dari uji t (t_{hitung}) dihitung sebagai berikut.

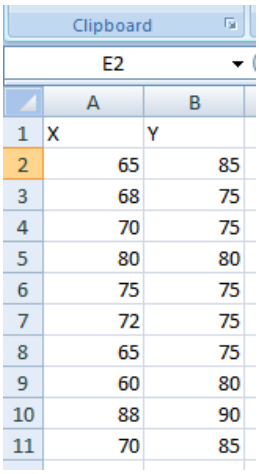
$$t = \frac{\bar{Y} - \bar{X}}{s_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

$$t = \frac{79{,}5 - 71{,}3}{6{,}922588\sqrt{\frac{1}{10} + \frac{1}{10}}} = 2{,}648685349.$$

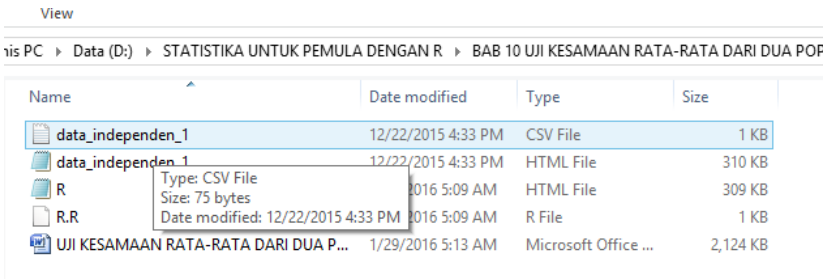
Nilai statistik dari uji t berdasarkan perhitungan adalah 2,648685394. Perhatikan bahwa karena $|t_{hitung}| > |t_{kritis}|$, yakni 2,6487 > 2,101, maka disimpulkan bahwa hipotesis nol ditolak dan hipotesis alternatif diterima. Hal ini berarti terdapat perbedaan (secara rata-rata) yang signifikan secara statistika dari nilai ujian matematika dasar antara mahasiswa laki-laki dan perempuan dengan tingkat signifikansi 5%.

Penyelesaian dalam R untuk Uji Kesamaan Rata-Rata dari Dua Populasi yang Tidak Berhubungan (Independen) dengan Asumsi Varians yang Sama

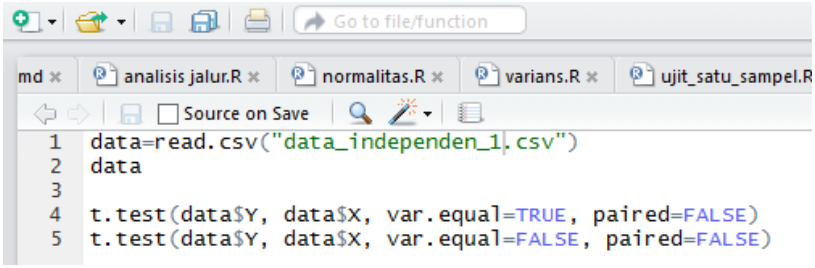
Data terlebih dahulu dibuat dalam *Microsoft Excel* (Gambar 9.5) dan disimpan dengan format tipe **.csv** (Gambar 9.6). Ketik kode R seperti pada Gambar 9.7. Kemudian *Compile* dan pilih HTML. Hasilnya seperti pada Gambar 9.8 dan Gambar 9.9.



Gambar 9.5



Gambar 9.6



Gambar 9.7

```
data=read.csv("data_independen_1.csv")
data
```

```
##      X  Y
## 1   65 85
## 2   68 75
## 3   70 75
## 4   80 80
## 5   75 75
## 6   72 75
## 7   65 75
## 8   60 80
## 9   88 90
## 10 70 85
```

Gambar 9.8

```
t.test(data$Y, data$X, var.equal=TRUE, paired=FALSE)
```

```
##
##  Two Sample t-test
##
## data:  data$Y and data$X
## t = 2.6487, df = 18, p-value = 0.01633
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.695807 14.704193
## sample estimates:
## mean of x mean of y
##      79.5      71.3
```

Ketika asumsi kesamaan varians populasi dipenuhi.

```
t.test(data$Y, data$X, var.equal=FALSE, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data$Y and data$X
## t = 2.6487, df = 15.851, p-value = 0.01762
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.632025 14.767975
## sample estimates:
## mean of x mean of y
##      79.5      71.3
```

Ketika asumsi kesamaan varians populasi tidak dipenuhi.

Gambar 9.9

Berdasarkan Gambar 9.9, perhatikan *output* "**ketika asumsi varians populasi dipenuhi**". Diketahui nilai statistik dari uji *t* (*t*) adalah 2,6487, sementara nilai probabilitas (*p-value*)

adalah 0,01633. Perhatikan bahwa karena $|t_{hitung}| > |t_{kritis}|$, yakni 2,6487 > 2,101, maka disimpulkan bahwa hipotesis nol ditolak dan hipotesis alternatif diterima. Hal ini berarti terdapat perbedaan (secara rata-rata) yang signifikan secara statistika dari nilai ujian matematika dasar antara mahasiswa laki-laki dan perempuan dengan tingkat signifikansi 5%.

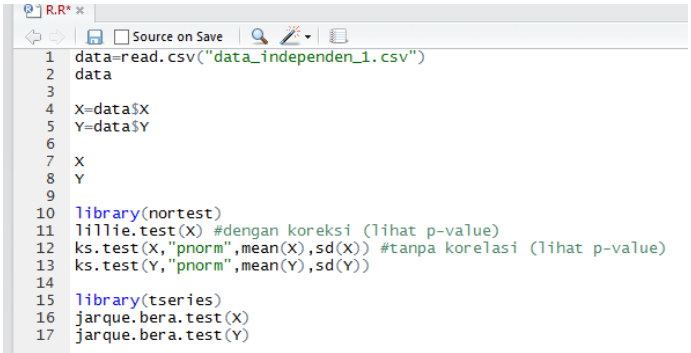
Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan menggunakan nilai probabilitas dari uji t (*p-value*). Nilai probabilitas dari uji t dibandingkan dengan tingkat signifikansi yang digunakan. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan nilai probabilitas.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

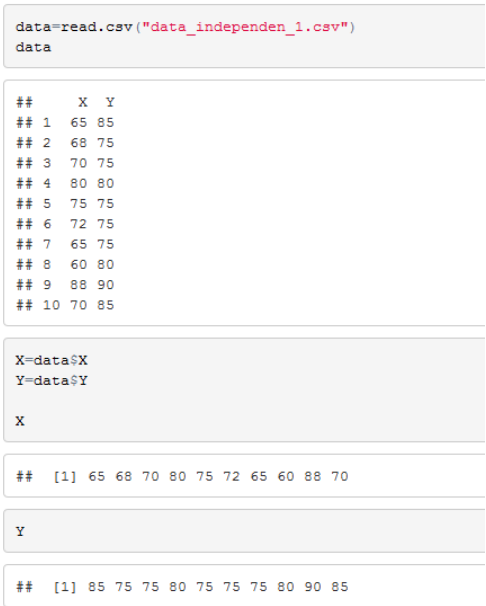
Diketahui nilai probabilitas dari uji t (*p-value*) adalah 0,01633. Karena nilai probabilitas tersebut lebih kecil dibandingkan tingkat signifikansi $\alpha = 0{,}05$, maka hipotesis nol ditolak dan hipotesis alternatif diterima. Hal ini berarti terdapat perbedaan (secara rata-rata) yang signifikan secara statistika dari nilai ujian matematika dasar antara mahasiswa laki-laki dan perempuan dengan tingkat signifikansi 5%.

Uji Asumsi Normalitas dalam R

Dalam uji kesamaan rata-rata dari dua populasi yang tidak berhubungan dengan asumsi varians yang sama, populasi pertama dan populasi kedua diasumsikan berdistribusi normal (Mann dan Lacke, 2011:448). Pada Gambar 9.10, disajikan kode R untuk uji asumsi bahwa sampel X dan sampel Y ditarik dari populasi-populasi yang berdistribusi normal. Hasil eksekusi kode R pada Gambar 9.10, disajikan pada Gambar 9.11 hingga Gambar 9.14.



Gambar 9.10



Gambar 9.11

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  X
## D = 0.16556, p-value = 0.6103
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  X
## D = 0.16556, p-value = 0.9469
## alternative hypothesis: two-sided
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  Y
## D = 0.29327, p-value = 0.356
## alternative hypothesis: two-sided
```

Gambar 9.12

One-Sample Kolmogorov-Smirnov Test

		VAR00001	VAR00002
N		10	10
Normal Parameters[a,,b]	Mean	71.3000	79.5000
	Std. Deviation	8.09732	5.50252
Most Extreme Differences	Absolute	.166	.293
	Positive	.166	.293
	Negative	-.118	-.207
Kolmogorov-Smirnov Z		.524	.927
Asymp. Sig. (2-tailed)		.947	.356

a. Test distribution is Normal.

b. Calculated from data.

Gambar 9.13 Hasil Berdasarkan SPSS

```
jarque.bera.test(X)
```

```
##
##  Jarque Bera Test
##
## data:  X
## X-squared = 0.92911, df = 2, p-value = 0.6284
```

```
jarque.bera.test(Y)
```

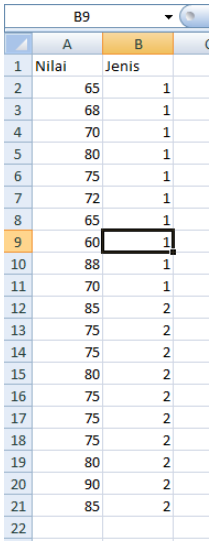
```
##
##  Jarque Bera Test
##
## data:  Y
## X-squared = 1.1767, df = 2, p-value = 0.5552
```

Gambar 9.14

Perhatikan bahwa berdasarkan Gambar 9.12, nilai probabilitas dari uji Kolmogorov-Smirnov untuk sampel *X* (*p-value*) adalah 0,9469, sementara untuk sampel *Y* adalah 0,356. Berdasarkan Gambar 9.14, nilai probabilitas dari uji Jarque-Bera untuk sampel *X* (*p-value*) adalah 0,6284, sementara untuk sampel *Y* adalah 0,5552. Karena masing-masing nilai probabilitas lebih besar dibandingkan tingkat signifikansi, yakni 0,05, maka hipotesis nol diterima, dan hipotesis alternatif ditolak. Hal ini berarti asumsi normalitas dipenuhi.

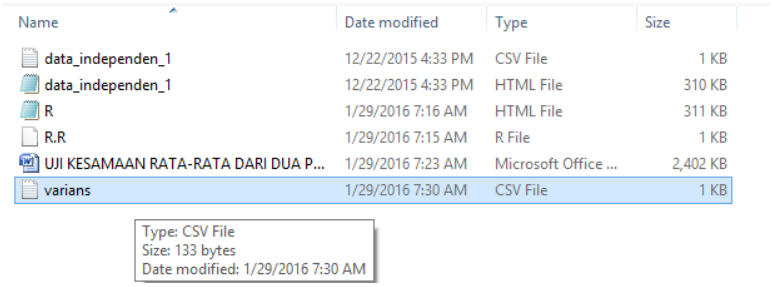
Uji Asumsi Kesamaan Varians dalam R

Selain asumsi normalitas, asumsi lain yang dikenakan adalah asumsi kesamaan varians, yakni sampel-sampel yang diteliti berasal dari populasi-populasi yang memiliki varians yang sama. Untuk menguji apakah sampel-sampel yang diteliti berasal dari populasi-populasi yang memiliki varians yang sama, dapat digunakan uji Levene. Data terlebih dahulu dibuat dalam *Microsoft Excel* (Gambar 9.15) dan disimpan dengan format tipe **.csv** (Gambar 9.16). Ketik kode R seperti pada Gambar 9.17. Kemudian *Compile* dan pilih HTML (Gambar 9.18). Hasilnya seperti pada Gambar 9.19 dan Gambar 9.20.

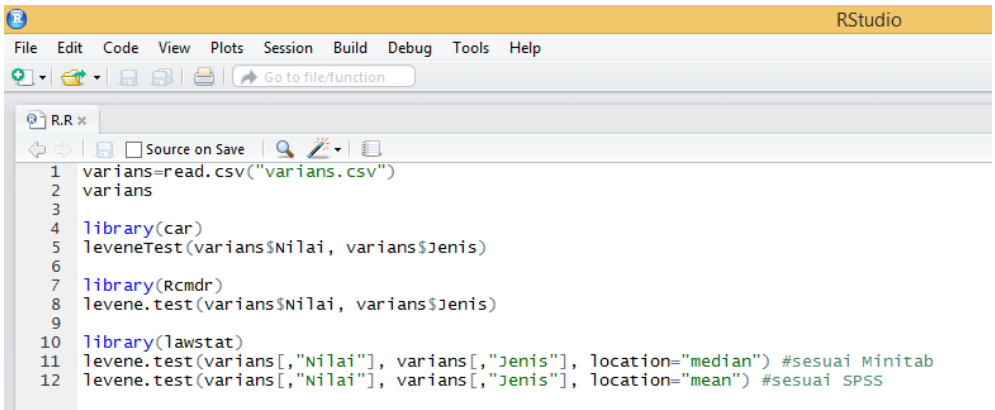


	A	B
1	Nilai	Jenis
2	65	1
3	68	1
4	70	1
5	80	1
6	75	1
7	72	1
8	65	1
9	60	1
10	88	1
11	70	1
12	85	2
13	75	2
14	75	2
15	80	2
16	75	2
17	75	2
18	75	2
19	80	2
20	90	2
21	85	2
22		

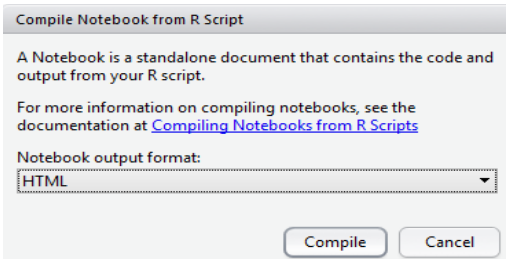
Gambar 9.15



Gambar 9.16



Gambar 9.17



Gambar 9.18

```
varians=read.csv("varians.csv")
varians
```

```
##    Nilai Jenis
## 1     65     1
## 2     68     1
## 3     70     1
## 4     80     1
## 5     75     1
## 6     72     1
## 7     65     1
## 8     60     1
## 9     88     1
## 10    70     1
## 11    85     2
## 12    75     2
## 13    75     2
## 14    80     2
## 15    75     2
## 16    75     2
## 17    75     2
## 18    80     2
## 19    90     2
## 20    85     2
```

```
library(car)
leveneTest(varians$Nilai, varians$Jenis)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.3305 0.5725
##       18
```

```
library(Rcmdr)
```

```
levene.test(varians$Nilai, varians$Jenis)
```

```
## Warning: 'levene.test' is deprecated.
## Use 'leveneTest' instead.
## See help("Deprecated") and help("car-deprecated").
```

```
## Warning in leveneTest.default(...): varians$Jenis coerced to factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.3305 0.5725
##       18
```

Gambar 9.19

```
levene.test(varians[,"Nilai"], varians[,"Jenis"], location="median") #sesuai Minitab
```

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  varians[, "Nilai"]
## Test Statistic = 0.33053, p-value = 0.5725
```

```
levene.test(varians[,"Nilai"], varians[,"Jenis"], location="mean") #sesuai SPSS
```

```
##
##  classical Levene's test based on the absolute deviations from the
##  mean ( none not applied because the location is not set to median
##  )
##
## data:  varians[, "Nilai"]
## Test Statistic = 0.62924, p-value = 0.438
```

Gambar 9.20

Perhatikan Gambar 9.20. Nilai statistik dari uji Levene dengan pendekatan *Location = "median"* adalah 0,33053, yang mana hasil ini sama dengan hasil Minitab. Namun nilai statistik dari uji Levene dengan pendekatan *Location = "mean"* adalah 0,62924, yang mana hasil ini sama dengan hasil SPSS.

Test of Homogeneity of Variances

Dependent Variables: Nilai

Levene Statistic	df1	df2	Sig.
0.6292360373125517	18		.438

Gambar 9.21 Hasil berdasarkan SPSS

Diketahui juga berdasarkan Gambar 9.20 nilai probabilitas (*p-value*) adalah 0,438, yakni lebih besar dibandingkan tingkat signifikansi 0,05, maka hipotesis nol diterima dan hipotesis alternatif ditolak, sehingga asumsi bahwa populasi *X* dan populasi *Y* memiliki varians yang sama dapat diterima pada tingkat signifikansi 5%.

Referensi

1. Agresti, A. dan B. Finlay. 2009. *Statistical Methods for the Social Sciences, 4th Edition.* United States of America: Prentice Hall.

2. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

3. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

4. Mann, P. S. dan C.J. Lacke. 2011. *Introductory Statistics, International Student Version, 7th Edition*, Asia: John Wiley & Sons, Inc.

5. Montgomery, D. C. dan G. C. Runger. 2011. *Applied Statistics and Probability for Engineers, 5th Edition*. United States of America: John Wiley & Sons, Inc.

6. Smidth, R. K. dan D. H. Sanders. 2000. *Statistics a First Course, 6th Edition*, United States of America: McGraw-Hill Companies.

7. http://www.statmethods.net/stats/ttest.html

BAB 10

UJI KESAMAAN RATA-RATA DARI DUA POPULASI TIDAK BERHUBUNGAN, DENGAN ASUMSI VARIANS POPULASI BERBEDA (UJI t)

Uji Kesamaan Rata-Rata dari Dua Populasi yang Tidak Berhubungan (Independen) dengan Asumsi Varians Berbeda (t Test for Independent Populations with Assumption $\sigma_1^2 \neq \sigma_2^2$)

Dalam uji kesamaan rata-rata dari dua populasi yang tidak berhubungan dengan asumsi varians yang berbeda (tidak sama), menguji ada tidaknya perbedaan rata-rata antara populasi pertama dan populasi kedua. Dengan kata lain, menguji apakah selisih rata-rata antara kelompok kedua dan pertama berbeda atau sama dengan nol. Dalam uji ini, pengamatan-pengamatan pada populasi pertama saling bebas/independen (*independent*) dengan pengamatan-pengamatan pada populasi kedua (*independent populations*). **Uji ini didasarkan pada ketidaktahuan (*unknown*) mengenai nilai varians dari dua populasi, namun diasumsikan varians dari dua populasi tersebut tidak sama**.

Berikut beberapa contoh kasus yang dapat diselesaikan dengan pendekatan uji kesamaan rata-rata dari dua populasi independen dengan asumsi varians yang sama dengan uji t.

- ⇨ Menguji ada tidaknya perbedaan (perbedaan yang signifikan secara statistika) nilai indeks prestasi (secara rata-rata) antara mahasiswa laki-laki dan perempuan.
- ⇨ Menguji ada tidaknya perbedaan harga saham antara perusahaan manufaktur dan *real estate*.
- ⇨ Menguji ada tidaknya perbedaan uang jajan antara mahasiswa kedokteran dan mahasiswa matematika.
- ⇨ Menguji ada tidaknya perbedaan indeks prestasi antara mahasiswa dominan otak kanan dan dominan kotak kiri.

Dalam uji kesamaan rata-rata dari dua populasi yang tidak berhubungan dengan asumsi varians yang berbeda, hipotesis nol menyatakan tidak terdapat perbedaan rata-rata antara populasi pertama dan populasi kedua. Dengan kata lain, selisih rata-rata antara populasi kedua dan pertama sama dengan nol $(\mu_2 - \mu_1 = 0)$. Hipotesis alternatif menyatakan terdapat perbedaan rata-rata antara populasi pertama dan populasi kedua. Dengan kata lain, selisih rata-rata antara populasi kedua dan pertama berbeda dari nol $(\mu_2 - \mu_1 \neq 0)$. Nilai statistik dari uji t (t_{hitung}) dihitung dengan rumus sebagai berikut.

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

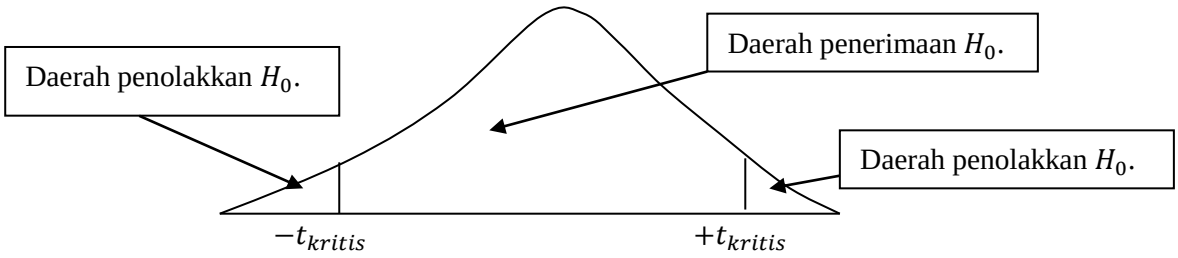
Perhatikan bahwa t merupakan nilai statistik dari uji t, $\bar{X}_1$ merupakan nilai rata-rata dari sampel pertama, $\bar{X}_2$ merupakan nilai rata-rata dari sampel kedua, s_1 merupakan nilai standar deviasi dari sampel pertama, s_2 merupakan nilai standar deviasi dari sampel kedua, n_1 merupakan jumlah pengamatan dalam sampel pertama, dan n_2 merupakan jumlah pengamatan dalam sampel kedua.

Untuk pengambilan keputusan terhadap hipotesis, dapat dilakukan dengan membandingkan nilai statistik dari uji t terhadap nilai kritis t (t_{kritis}). Sebelum menghitung nilai kritis t, terlebih dahulu menghitung nilai derajat bebas. Berikut rumus untuk menghitung nilai derajat bebas.

$$Derajat\ bebas = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 1)}}.$$

Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji t.

$$Jika\ |t_{hitung}| \leq |t_{kritis}|, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ |t_{hitung}| > |t_{kritis}|, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$



Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan menggunakan pendekatan nilai probabilitas dari uji t. Nilai probabilitas dari uji t dibandingkan dengan tingkat signifikansi yang digunakan. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan pendekatan nilai probabilitas.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Uji Asumsi Normalitas

Dalam uji kesamaan rata-rata dari dua populasi yang tidak berhubungan dengan asumsi varians yang berbeda, populasi pertama dan populasi kedua diasumsikan berdistribusi normal (Mann dan Lacke, 2011:458). Namun ketika ukuran sampel cukup besar, yakni masing-masing sampel berukuran ≥ 30, maka populasi tidak harus berdistribusi normal (Mann dan Lacke, 2011:465). Untuk menguji asumsi normalitas tersebut, dapat digunakan pendekatan grafik, yakni *Q-Q plot*. Pada pendekatan *Q-Q plot*, jika titik-titik (*dots*) menyebar jauh (menyebar jauh berliku-liku pada garis diagonal seperti ular) dari garis diagonal, maka diindikasi asumsi normalitas tidak dipenuhi. Jika titik-titik menyebar sangat dekat pada garis diagonal, maka asumsi normalitas dipenuhi. Di samping itu, dapat juga digunakan pendekatan

uji Kolmogorov-Smirnov atau uji Jarque-Bera, untuk menguji asumsi normalitas. Hipotesis nol menyatakan data sampel ditarik dari populasi yang berdistribusi normal, sedangkan hipotesis alternatif menyatakan data sampel ditarik dari populasi yang tidak berdistribusi normal.

Untuk pengambilan keputusan terhadap hipotesis, dapat dibandingkan antara nilai probabilitas dari uji Kolmogorov-Smirnov atau uji Jarque-Bera, dengan tingkat signifikansi yang digunakan (α). Berikut aturan pengambilan keputusan terhadap hipotesis.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, H_0\ ditolak\ dan\ H_1\ diterima.$$

Uji Asumsi Ketidaksamaan Varians

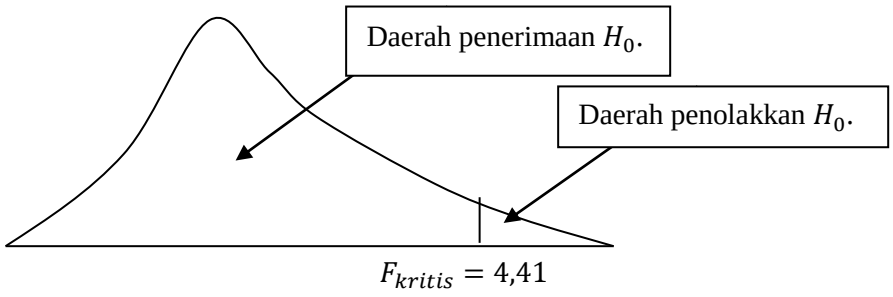
Selain asumsi normalitas, asumsi lain yang dikenakan adalah asumsi ketidaksamaan varians, yakni sampel-sampel yang diteliti berasal dari populasi-populasi yang memiliki varians yang berbeda. Untuk menguji apakah sampel-sampel yang diteliti berasal dari populasi-populasi yang memiliki varians berbeda, dapat digunakan uji Levene. Pada uji Levene, hipotesis nol menyatakan sampel-sampel yang diambil berasal dari populasi-populasi yang memiliki varians yang sama, sedangkan hipotesis alternatif menyatakan paling tidak terdapat sepasang populasi yang memiliki varians yang berbeda.

Pengambilan keputusan terhadap hipotesis dilakukan dengan membandingkan nilai statistik dari uji Levene (L) dengan nilai kritis F (F_{kritis}). Sebelum menghitung nilai kritis F, terlebih dahulu menghitung nilai dari derajat bebas pembilang dan derajat bebas penyebut. Berikut rumus untuk menghitung nilai dari derajat bebas pembilang dan derajat bebas penyebut.

$$Derajat\ bebas\ pembilang = k - 1.$$
$$Derajat\ bebas\ penyebut = N - k.$$

Perhatikan bahwa k menyatakan banyaknya elemen sampel, sedangkan N merupakan jumlah elemen/pengamatan dari seluruh sampel. Diketahui misalkan nilai k adalah 2, sedangkan nilai N adalah $20\ (n_1 + n_2 = 10 + 10 = 20)$. Diketahui misalkan tingkat signifikansi yang digunakan adalah 5%, sehingga nilai kritis F dengan derajat bebas pembilang $2 - 1 = 1$, derajat bebas penyebut $20 - 2 = 18$, dan tingkat signifikansi 5% adalah 4,41. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji Levene.

$$Jika\ nilai\ statistik\ dari\ uji\ Levene \leq F_{kritis}, maka\ H_0\ diterima\ dan\ H_1 \text{ ditolak.}$$
$$jika\ nilai\ statistik\ dari\ uji\ Levene > F_{kritis}, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$



Pengambilan keputusan terhadap hipotesis juga dapat digunakan pendekatan nilai probabilitas dari uji Levene. Nilai probabilitas tersebut dibandingkan dengan tingkat signifikansi (α). Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan pendekatan nilai probabilitas.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, H_0\ ditolak\ dan\ H_1\ diterima.$$

Contoh Kasus Uji Kesamaan Rata-Rata dari Dua Populasi yang Tidak Berhubungan (Independen) dengan Asumsi Varians yang Berbeda (Contoh Perhitungan)

Misalkan seorang peneliti akan meneliti mengenai ada tidaknya perbedaan nilai ujian matakuliah matematika dasar antara mahasiswa laki-laki dan mahasiswa perempuan. Untuk keperluan penelitian, peneliti tersebut mengambil sampel sebanyak 20 nilai ujian matakuliah matematika dasar yang terdiri dari 10 nilai ujian mahasiswa laki-laki dan 10 nilai ujian mahasiswa perempuan. Data yang telah dikumpulkan disajikan dalam Tabel 10.1. Peneliti akan menguji apakah terdapat perbedaan (secara rata-rata) yang signifikan secara statistika dari nilai ujian matematika dasar antara mahasiswa laki-laki dan perempuan dengan tingkat signifikansi 5%.

Tabel 10.1 (Data Fiktif)

Nama Mahasiswa Laki-laki	Nilai (X)	Nama Mahasiswa Perempuan	Nilai Ujian (Y)
Ugi	70	Ulan	90
Mifdhal	71	Fitri	91
Iqbal	72	Evelin	92
Alan	70	Melda	93
John	71	Dina	94
Andre	72	Suci	95
Ridho	70	Febri	86
Hanafi	70	Oshin	97
Romi	71	Wilya	98
Hasoloan	72	Windy	100

Berikut akan dihitung nilai derajat bebas (*degree of freedom*).

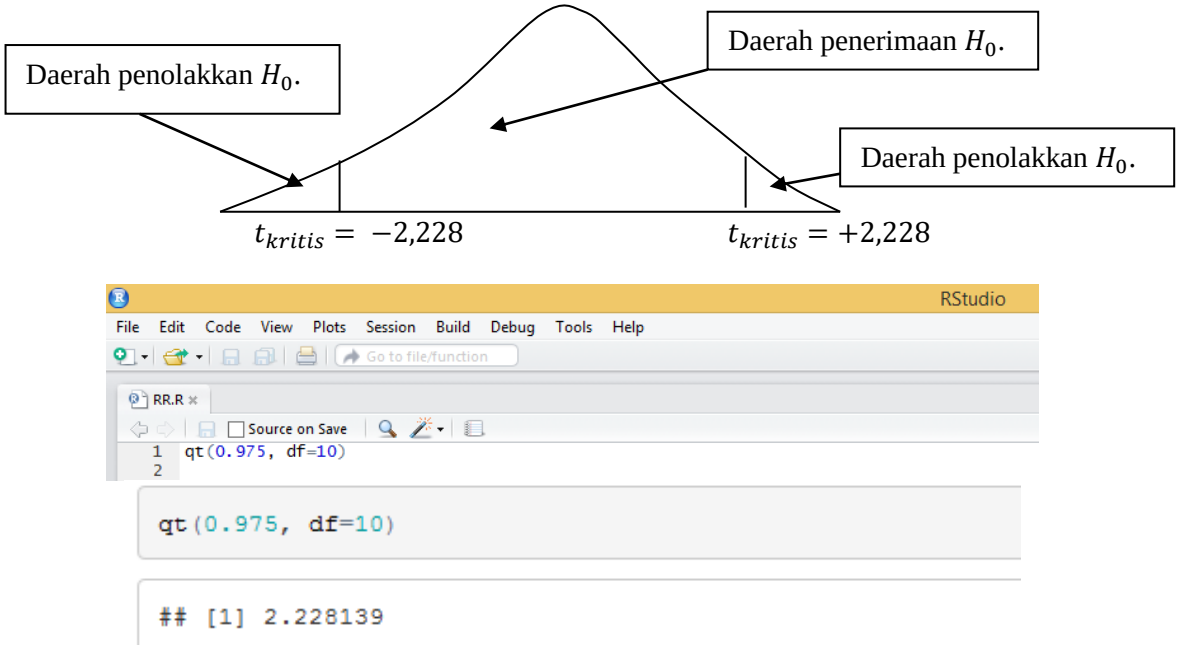
$$derajat\ bebas = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{s_X^2}{n_X}\right)^2}{(n_X - 1)} + \frac{\left(\frac{s_Y^2}{n_Y}\right)^2}{(n_Y - 1)}}$$

$$derajat\ bebas = \frac{\left(\frac{0{,}875595^2}{10} + \frac{4{,}141927^2}{10}\right)^2}{\frac{\left(\frac{0{,}875595^2}{10}\right)^2}{(10 - 1)} + \frac{\left(\frac{4{,}141927^2}{10}\right)^2}{(10 - 1)}}$$

$$derajat\ bebas = 9{,}802 \cong 10.$$

Diketahui derajat bebas (*df*) bernilai 9,802 ≅ 10. Nilai kritis t dengan derajat bebas 10 dan tingkat signifikansi 5% adalah $\pm 2{,}228$. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji t.

$$Jika\ |t_{hitung}| \leq |t_{kritis}|, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ |t_{hitung}| > |t_{kritis}|, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$



Gambar 10.1 Menentukan Nilai Kritis *t* dengan R

Tabel 10.2

	X	Y
	70	90
	71	91
	72	92
	70	93
	71	94
	72	95
	70	86
	70	97
	71	98
	72	100
$rata - rata$	70,9	93,6
standar deviasi	0,875595	4,141927

Selanjutnya akan dihitung nilai statistik dari uji t (t_{hitung}).

$$t = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

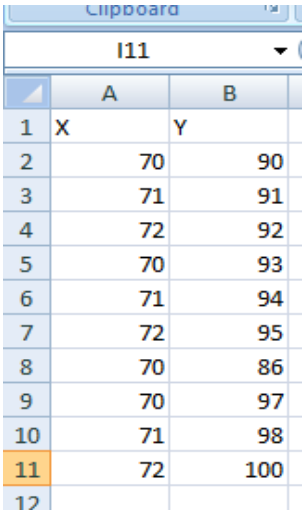
$$t = \frac{93{,}6 - 70{,}9}{\sqrt{\frac{(0{,}8755)^2}{10} + \frac{(4{,}141927)^2}{10}}}$$

$$t = 16{,}9563.$$

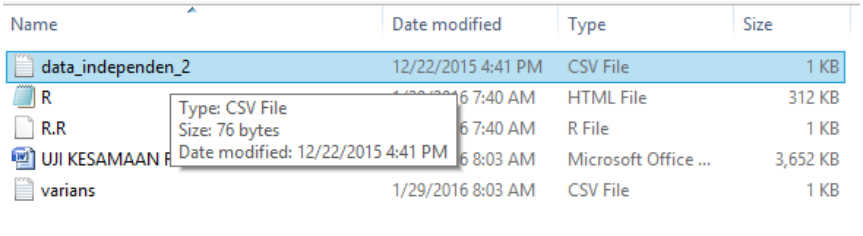
Nilai statistik dari uji t berdasarkan perhitungan adalah 16,9563. Perhatikan bahwa karena $|t_{hitung}| > |t_{kritis}|$, yakni 16,956 > 2,228, maka disimpulkan bahwa hipotesis nol ditolak dan hipotesis alternatif diterima. Hal ini berarti terdapat perbedaan (secara rata-rata) yang signifikan secara statistika dari nilai ujian matematika dasar antara mahasiswa laki-laki dan perempuan dengan tingkat signifikansi 5%.

Penyelesaian dalam R untuk Uji Kesamaan Rata-Rata dari Dua Populasi yang Tidak Berhubungan (Independen) dengan Asumsi Varians yang Berbeda

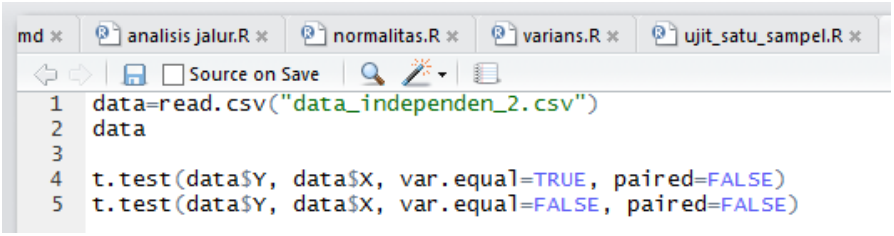
Data terlebih dahulu dibuat dalam *Microsoft Excel* (Gambar 10.2) dan disimpan dengan format tipe **.csv** (Gambar 10.3). Ketik kode R seperti pada Gambar 10.4. Kemudian *Compile* dan pilih HTML. Hasilnya seperti pada Gambar 10.5 hingga Gambar 10.6. Berdasarkan Gambar 10.6, perhatikan *output* pada bagian "**ketika asumsi kesamaan varians populasi tidak dipenuhi**". Diketahui nilai statistik dari uji t (t) adalah 16,956, sementara nilai probabilitas (*p-value*) adalah 0,0000001374. Diketahui nilai derajat bebas (*df*) adalah 9,8028≅ 10.



Gambar 10.2



Gambar 10.3



Gambar 10.4

```
data=read.csv("data_independen_2.csv")
data
```

```
##     X   Y
## 1  70  90
## 2  71  91
## 3  72  92
## 4  70  93
## 5  71  94
## 6  72  95
## 7  70  86
## 8  70  97
## 9  71  98
## 10 72 100
```

Gambar 10.5

```
t.test(data$Y, data$X, var.equal=TRUE, paired=FALSE)
```

```
##
##  Two Sample t-test
##
## data:  data$Y and data$X
## t = 16.956, df = 18, p-value = 1.63e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  19.88741 25.51259
## sample estimates:
## mean of x mean of y
##      93.6      70.9
```

Ketika asumsi kesamaan varians populasi dipenuhi.

```
t.test(data$Y, data$X, var.equal=FALSE, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  data$Y and data$X
## t = 16.956, df = 9.8028, p-value = 1.374e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  19.70895 25.69105
## sample estimates:
## mean of x mean of y
##      93.6      70.9
```

Ketika asumsi kesamaan varians populasi tidak dipenuhi.

Gambar 10.6

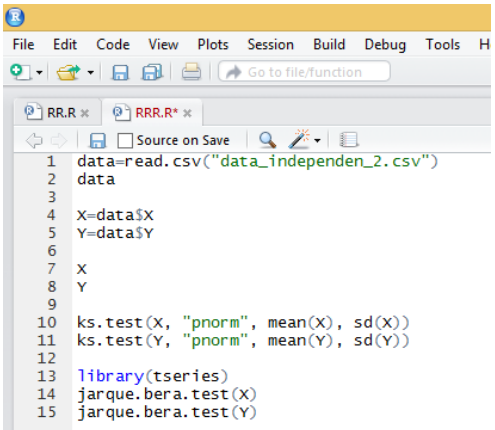
Perhatikan bahwa karena $|t_{hitung}| > |t_{kritis}|$, yakni 16,956 > 2,228, maka disimpulkan bahwa hipotesis nol ditolak dan hipotesis alternatif diterima. Hal ini berarti terdapat perbedaan (secara rata-rata) yang signifikan secara statistika dari nilai ujian matematika dasar antara mahasiswa laki-laki dan perempuan dengan tingkat signifikansi 5%. Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan menggunakan nilai probabilitas dari uji t (*p-value*). Nilai probabilitas dari uji t dibandingkan dengan tingkat signifikansi yang digunakan. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan nilai probabilitas.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Diketahui nilai probabilitas (*p-value*) dari uji t adalah 0,0000001374. Karena nilai probabilitas tersebut (*p-value*) lebih kecil dibandingkan tingkat signifikansi $\alpha = 0{,}05$, maka hipotesis nol ditolak dan hipotesis alternatif diterima. Hal ini berarti terdapat perbedaan (secara rata-rata) yang signifikan secara statistika dari nilai ujian matematika dasar antara mahasiswa laki-laki dan perempuan dengan tingkat signifikansi 5%.

Uji Asumsi Normalitas dalam R

Dalam uji kesamaan rata-rata dari dua populasi yang tidak berhubungan dengan asumsi varians yang berbeda, populasi pertama dan populasi kedua diasumsikan berdistribusi normal (Mann dan Lacke, 2011:448). Pada Gambar 10.7, disajikan kode R untuk uji asumsi bahwa sampel X dan sampel Y ditarik dari populasi-populasi yang berdistribusi normal. Hasil eksekusi kode R pada Gambar 10.7, disajikan pada Gambar 10.8 hingga Gambar 10.11.



Gambar 10.7

```
data=read.csv("data_independen_2.csv")
data
```

```
##     X   Y
## 1  70  90
## 2  71  91
## 3  72  92
## 4  70  93
## 5  71  94
## 6  72  95
## 7  70  86
## 8  70  97
## 9  71  98
## 10 72 100
```

```
X=data$X
Y=data$Y

X
```

```
##  [1] 70 71 72 70 71 72 70 70 71 72
```

```
Y
```

```
##  [1]  90  91  92  93  94  95  86  97  98 100
```

```
ks.test(X, "pnorm", mean(X), sd(X))
```

Gambar 10.8

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  X
## D = 0.248, p-value = 0.57
## alternative hypothesis: two-sided
```

```
##  One-sample Kolmogorov-Smirnov test
##
## data:  Y
## D = 0.094141, p-value = 0.9999
## alternative hypothesis: two-sided
```

Gambar 10.9

One-Sample Kolmogorov-Smirnov Test

		VAR00001	VAR00002
N		10	10
Normal Parameters[a,,b]	Mean	70.9000	93.6000
	Std. Deviation	.87560	4.14193
Most Extreme Differences	Absolute	.248	.094
	Positive	.248	.068
	Negative	-.195	-.094
Kolmogorov-Smirnov Z		.784	.298
Asymp. Sig. (2-tailed)		.570	.999

a. Test distribution is Normal.

b. Calculated from data.

Gambar 10.10 Hasil Berdasarkan SPSS

```
jarque.bera.test(X)
```

```
##
##  Jarque Bera Test
##
## data:  X
## X-squared = 1.0296, df = 2, p-value = 0.5976
```

```
jarque.bera.test(Y)
```

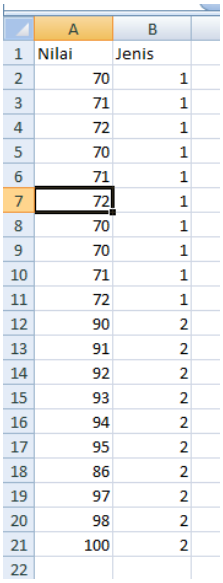
```
##
##  Jarque Bera Test
##
## data:  Y
## X-squared = 0.20791, df = 2, p-value = 0.9013
```

Gambar 10.11

Perhatikan bahwa berdasarkan Gambar 10.9, nilai probabilitas dari uji Kolmogorov-Smirnov untuk sampel *X* (*p-value*) adalah 0,57, sementara untuk sampel *Y* adalah 0,999. Berdasarkan Gambar 10.11, nilai probabilitas dari uji Jarque-Bera untuk sampel *X* (*p-value*) adalah 0,5976, sementara untuk sampel *Y* adalah 0,9013. Karena masing-masing nilai probabilitas lebih besar dibandingkan tingkat signifikansi, yakni 0,05, maka hipotesis nol diterima, dan hipotesis alternatif ditolak. Hal ini berarti asumsi normalitas dipenuhi.

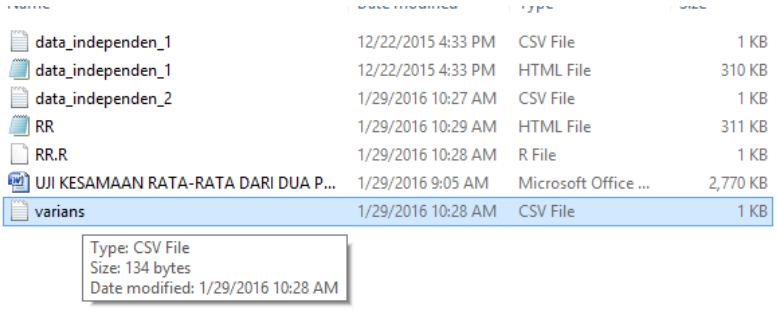
Uji Asumsi Ketidaksamaan Varians dalam R

Selain asumsi normalitas, asumsi lain yang dikenakan adalah asumsi ketidaksamaan varians, yakni sampel-sampel yang diteliti berasal dari populasi-populasi yang memiliki varians yang berbeda. Untuk menguji apakah sampel-sampel yang diteliti berasal dari populasi-populasi yang memiliki varians yang berbeda, dapat digunakan uji Levene. Data terlebih dahulu dibuat dalam *Microsoft Excel* (Gambar 10.12) dan disimpan dengan format tipe **.csv** (Gambar 10.13). Ketik kode R seperti pada Gambar 10.14. Kemudian *Compile* dan pilih HTML (Gambar 10.15). Hasilnya seperti pada Gambar 10.16 dan Gambar 10.17.

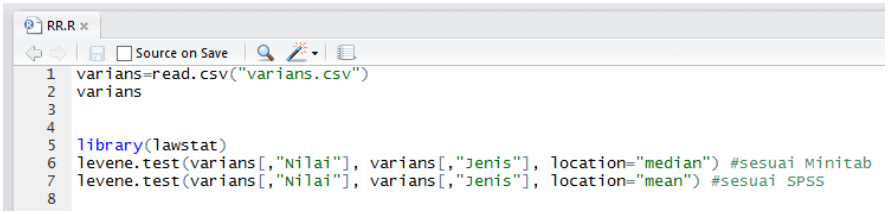


	A	B
1	Nilai	Jenis
2	70	1
3	71	1
4	72	1
5	70	1
6	71	1
7	72	1
8	70	1
9	70	1
10	71	1
11	72	1
12	90	2
13	91	2
14	92	2
15	93	2
16	94	2
17	95	2
18	86	2
19	97	2
20	98	2
21	100	2
22		

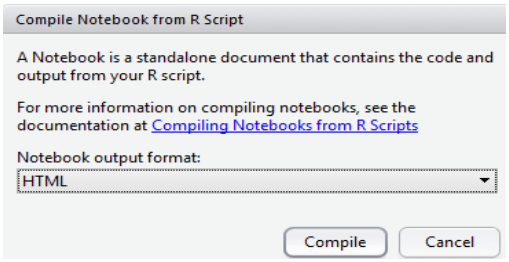
Gambar 10.12



Gambar 10.13



Gambar 10.14



Gambar 10.15

```
varians=read.csv("varians.csv")
varians
```

```
##    Nilai Jenis
## 1     70     1
## 2     71     1
## 3     72     1
## 4     70     1
## 5     71     1
## 6     72     1
## 7     70     1
## 8     70     1
## 9     71     1
## 10    72     1
## 11    90     2
## 12    91     2
## 13    92     2
## 14    93     2
## 15    94     2
## 16    95     2
## 17    86     2
## 18    97     2
## 19    98     2
## 20   100     2
```

Gambar 10.16

```
levene.test(varians[,"Nilai"], varians[,"Jenis"], location="median") #sesuai Minitab
```

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  varians[, "Nilai"]
## Test Statistic = 10.378, p-value = 0.004733
```

```
levene.test(varians[,"Nilai"], varians[,"Jenis"], location="mean") #sesuai SPSS
```

```
##
##  classical Levene's test based on the absolute deviations from the
##  mean ( none not applied because the location is not set to median
##  )
##
## data:  varians[, "Nilai"]
## Test Statistic = 10.305, p-value = 0.004853
```

Gambar 10.17

Diketahui juga berdasarkan Gambar 10.17 nilai probabilitas (*p-value*) adalah 0,004853 (*location=”mean”*), yakni lebih kecil dibandingkan tingkat signifikansi 0,05, maka hipotesis nol diterima dan hipotesis alternatif ditolak, sehingga asumsi bahwa populasi *X* dan populasi *Y* memiliki varians yang berbeda (ketidaksamaan varians) dapat diterima pada tingkat signifikansi 5%.

Referensi

1. Agresti, A. dan B. Finlay. 2009. *Statistical Methods for the Social Sciences, 4th Edition*. United States of America: Prentice Hall.

2. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

3. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

4. Mann, P. S. dan C.J. Lacke. 2011. *Introductory Statistics, International Student Version, 7th Edition*, Asia: John Wiley & Sons, Inc.

5. Montgomery, D. C. dan G. C. Runger. 2011. *Applied Statistics and Probability for Engineers, 5th Edition*. United States of America: John Wiley & Sons, Inc.

6. Smidth, R. K. dan D. H. Sanders. 2000. *Statistics a First Course, 6th Edition*, United States of America: McGraw-Hill Companies.

7. http://www.statmethods.net/stats/ttest.html

8. http://www.r-bloggers.com/two-sample-students-t-test-1/

9. http://stats.stackexchange.com/questions/110225/two-sample-t-test-for-equal-means-with-unequal-variances-for-large-samples

10. http://www.r-bloggers.com/paired-students-t-test/

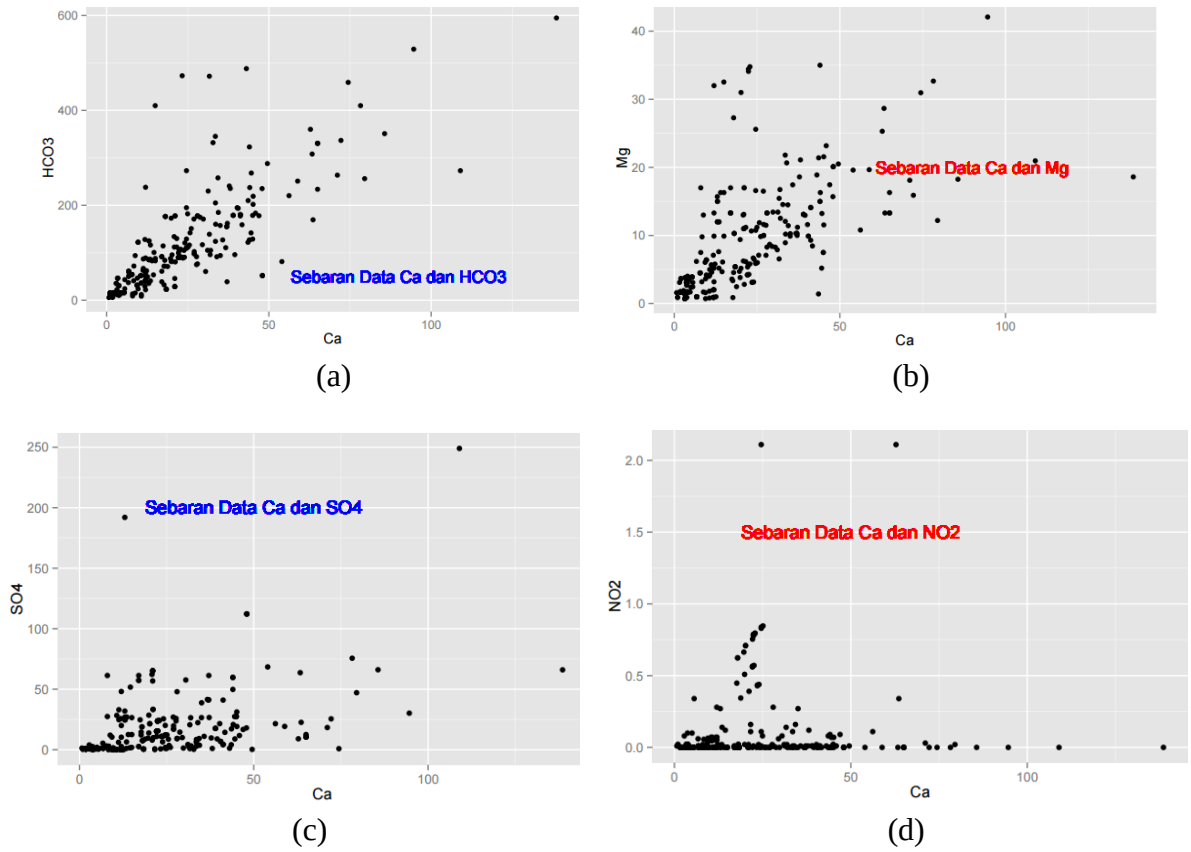
11. http://www.r-tutor.com/elementary-statistics/inference-about-two-populations/population-mean-between-two-matched-samples

BAB 11

KORELASI LINEAR PEARSON

Analisis Korelasi (Hubungan) Linear dengan Grafik

Berikut disajikan grafik dari sebaran data antara Ca v/s HCO3, Ca v/s Mg, Ca v/s SO4, dan Ca v/s NO2.



(a) (b)

(c) (d)

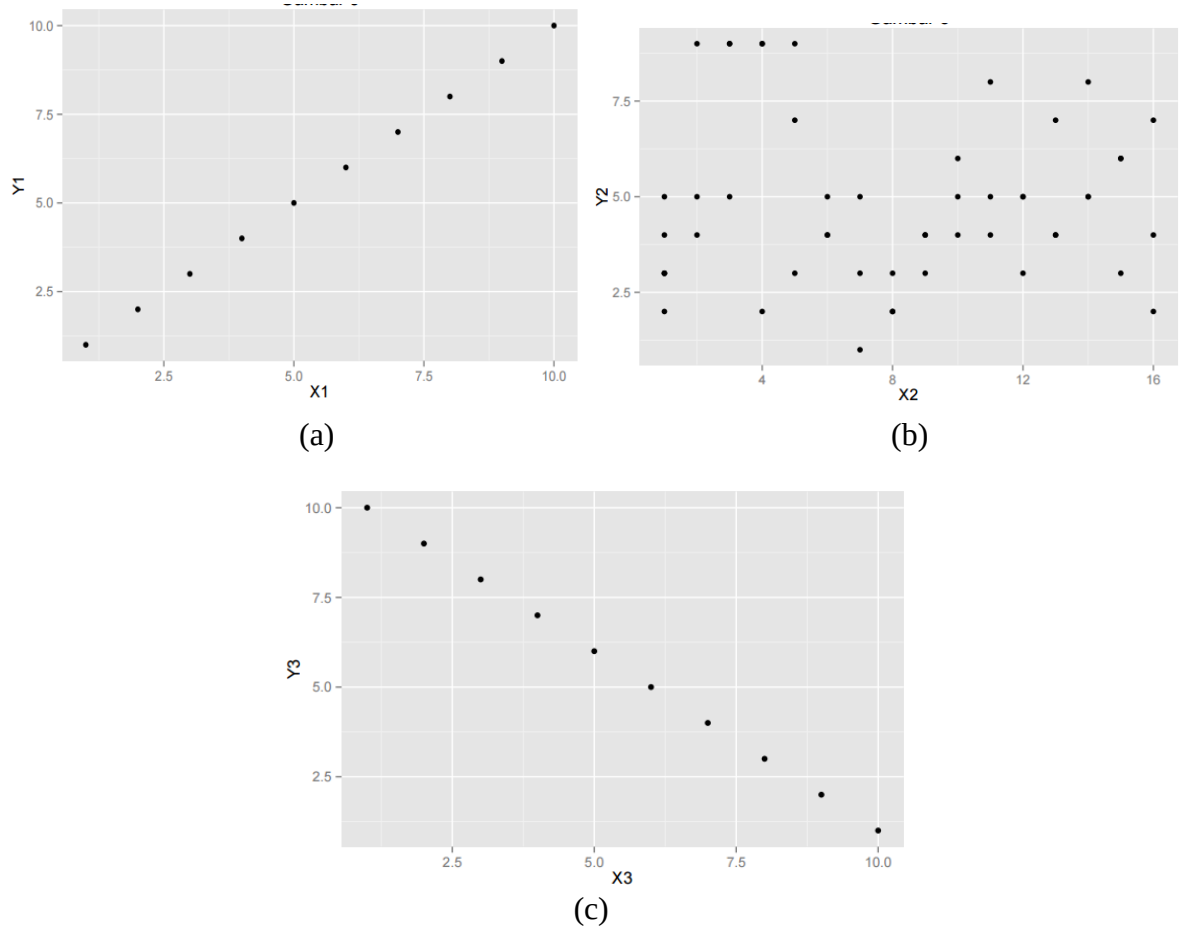
Gambar 11.1

Berdasarkan sebaran data dari Gambar 11.1(a) hingga Gambar 11.1(d), gambar manakah yang kira-kira memiliki sebaran data paling linear? Gambar manakah yang kira-kira memiliki sebaran data paling tidak linear? Pada pembahasan selanjutnya akan diperkenalkan suatu nilai yang dapat mengukur seberapa linear sebaran data untuk dua variabel.

Koefisien Korelasi Linear Pearson

Koefisien korelasi Pearson (dalam hal ini korelasi linear) merupakan suatu nilai yang dapat mengukur seberapa erat hubungan linear yang terjadi di antara dua variabel. Nilai dari koefisien korelasi Pearson berkisar dari -1 sampai 1. Nilai koefisien korelasi Pearson yang semakin mendekati 1 atau -1 menandakan terjadi hubungan linear yang kuat antara dua variabel, sementara jika mendekati 0 menandakan terjadi hubungan linear yang lemah antara dua variabel (mungkin bisa didekati dengan hubungan non-linear, alternatif dari hubungan

linear). Hubungan linear yang terjadi dapat bersifat positif, yakni ditandai dengan nilai koefisien korelasi Pearson yang bernilai positif, atau dapat bersifat negatif, ditandai dengan nilai koefisien korelasi Pearson yang bernilai negatif. Perhatikan Gambar 11.2(a) hingga Gambar 11.2(c).



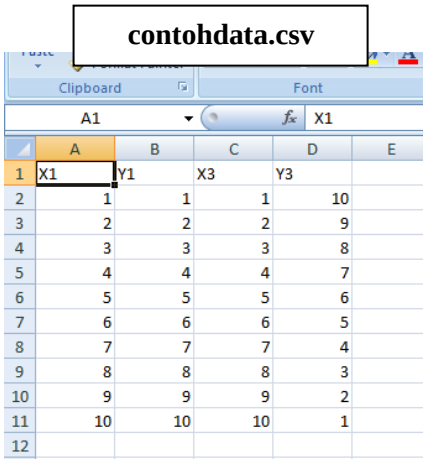
Gambar 11.2

Gambar 11.2(a) menunjukkan terjadinya hubungan linear positif yang **sempurna** antara X1 dan Y1 (apabila nilai koefisien korelasi Pearson dihitung, maka akan bernilai 1). **Hubungan positif berarti sebaran data cenderung menyebar dari kiri bawah ke kanan atas**. Sebaran data pada Gambar 11.2(b) cenderung acak (tidak beraturan), sehingga hubungan linear yang terjadi antara X2 dan Y2 lemah. Apabila nilai koefisien korelasi Pearson dihitung, maka akan bernilai mendekati 0. Gambar 11.2(c) menunjukkan terjadinya hubungan linear negatif yang **sempurna** antara X3 dan Y3 (apabila nilai koefisien korelasi Pearson dihitung, maka akan bernilai -1). **Hubungan negatif berarti sebaran data cenderung menyebar dari kiri atas ke kanan bawah**.

Menyajikan Grafik Sebaran Data dan Menghitung Koefisien Korelasi Linear Pearson dengan R

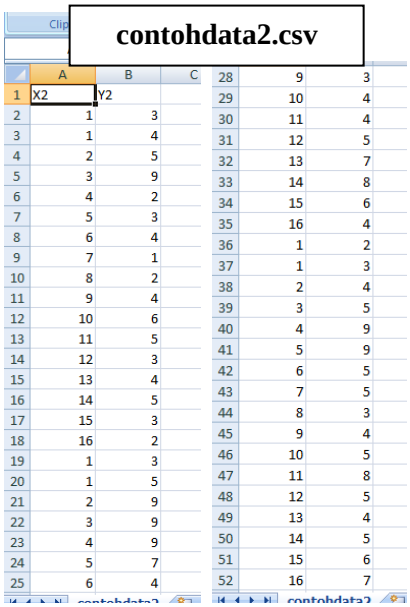
Misalkan diberikan data seperti pada Gambar 11.3 dengan nama *file* **contohdata.csv**, dan Gambar 11.4 dengan nama *file* **contohdata2.csv**. Gambar 11.6 disajikan kode R. Apabila kode R pada Gambar 11.6 dieksekusi, hasilnya seperti pada Gambar 11.7 dan Gambar 11.8. Berdasarkan Gambar 11.8, diketahui nilai koefisien korelasi linear Pearson antara X1 dan Y1

adalah 1. Hal ini berarti sebaran data bersifat positif dan linear sempurna (positif berarti sebaran data cenderung bergerak dari kiri bawah ke kanan atas). Perhatikan Gambar 11.2(a). Berdasarkan Gambar 11.8, diketahui nilai koefisien korelasi linear Pearson antara X2 dan Y2 adalah -0,0255. Perhatikan bahwa nilai koefisien korelasi Pearson -0,0255 mendekati 0. Hal ini berarti hubungan linear yang terjadi antara X2 dan Y2 lemah (perhatikan bahwa data menyebar cenderung acak, Gambar 11.2(b)). Berdasarkan Gambar 11.8, diketahui nilai koefisien korelasi Pearson antara X3 dan Y3 adalah -1. Hal ini berarti sebaran data bersifat negatif dan linear sempurna (negatif berarti sebaran data cenderung bergerak dari kiri atas ke kanan bawah). Perhatikan Gambar 11.2(c).



	A	B	C	D	E
1	X1	Y1	X3	Y3	
2	1	1	1	10	
3	2	2	2	9	
4	3	3	3	8	
5	4	4	4	7	
6	5	5	5	6	
7	6	6	6	5	
8	7	7	7	4	
9	8	8	8	3	
10	9	9	9	2	
11	10	10	10	1	
12					

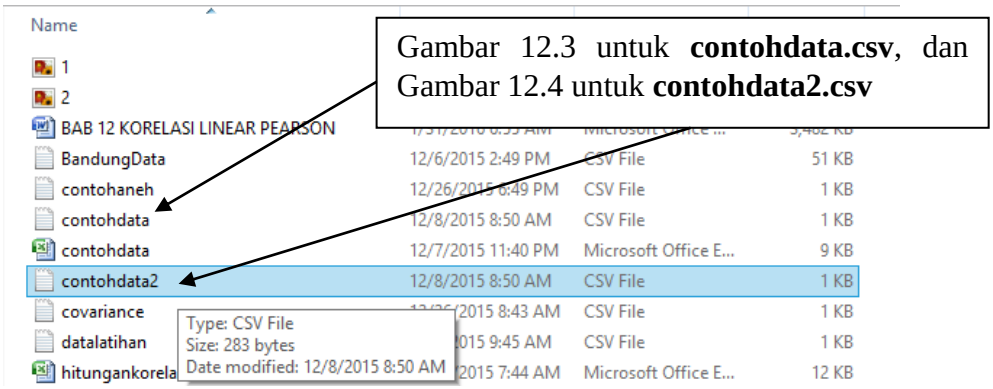
Gambar 11.3



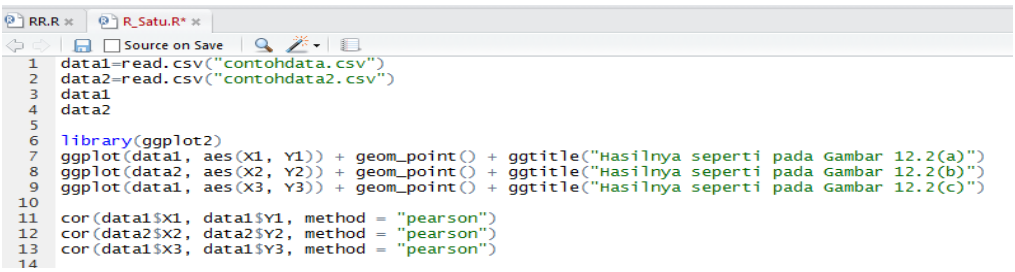
	A	B	C
1	X2	Y2	
2	1	3	
3	1	4	
4	2	5	
5	3	9	
6	4	2	
7	5	3	
8	6	4	
9	7	1	
10	8	2	
11	9	4	
12	10	6	
13	11	5	
14	12	3	
15	13	4	
16	14	5	
17	15	3	
18	16	2	
19	1	3	
20	1	5	
21	2	9	
22	3	9	
23	4	9	
24	5	7	
25	6	4	

28	9	3	
29	10	4	
30	11	4	
31	12	5	
32	13	7	
33	14	8	
34	15	6	
35	16	4	
36	1	2	
37	1	3	
38	2	4	
39	3	5	
40	4	9	
41	5	9	
42	6	5	
43	7	5	
44	8	3	
45	9	4	
46	10	5	
47	11	8	
48	12	5	
49	13	4	
50	14	5	
51	15	6	
52	16	7	

Gambar 11.4



Gambar 11.5




```
1  data1=read.csv("contohdata.csv")
2  data2=read.csv("contohdata2.csv")
3  data1
4  data2
5
6  library(ggplot2)
7  ggplot(data1, aes(X1, Y1)) + geom_point() + ggtitle("Hasilnya seperti pada Gambar 12.2(a)")
8  ggplot(data2, aes(X2, Y2)) + geom_point() + ggtitle("Hasilnya seperti pada Gambar 12.2(b)")
9  ggplot(data1, aes(X3, Y3)) + geom_point() + ggtitle("Hasilnya seperti pada Gambar 12.2(c)")
10
11 cor(data1$X1, data1$Y1, method = "pearson")
12 cor(data2$X2, data2$Y2, method = "pearson")
13 cor(data1$X3, data1$Y3, method = "pearson")
14
```

Gambar 11.6

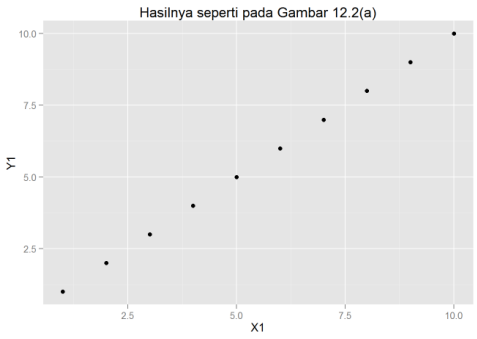
```
data1=read.csv("contohdata.csv")
data2=read.csv("contohdata2.csv")
data1
```

```
##    X1 Y1 X3 Y3
## 1   1  1  1 10
## 2   2  2  2  9
## 3   3  3  3  8
## 4   4  4  4  7
## 5   5  5  5  6
## 6   6  6  6  5
## 7   7  7  7  4
## 8   8  8  8  3
## 9   9  9  9  2
## 10 10 10 10  1
```

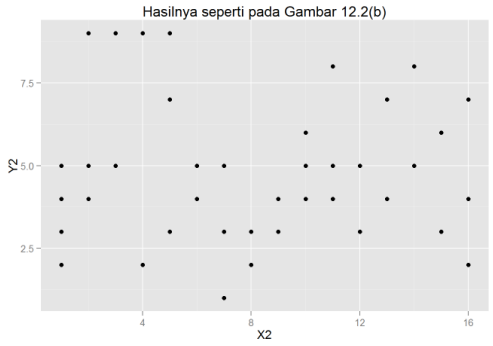
```
data2
```

```
##    X2 Y2
## 1   1  3
## 2   1  4
## 3   2  5
## 4   3  9
## 5   4  2
## 6   5  3
## 7   6  4
## 8   7  1
## 9   8  2
## 10  9  4
## 11 10  6
## 12 11  5
## 13 12  3
## 14 13  4
## 15 14  5
## 16 15  3
## 17 16  2
## 18  1  3
## 19  1  5
## 20  2  9
## 21  3  9
## 22  4  9
## 23  5  7
## 24  6  4
## 25  7  3
## 26  8  2
## 27  9  3
## 28 10  4
## 29 11  4
## 30 12  5
## 31 13  7
## 32 14  8
## 33 15  6
## 34 16  4
## 35  1  2
## 36  1  3
## 37  2  4
## 38  3  5
## 39  4  9
## 40  5  9
## 41  6  5
## 42  7  5
## 43  8  3
## 44  9  4
## 45 10  5
## 46 11  8
## 47 12  5
## 48 13  4
## 49 14  5
## 50 15  6
## 51 16  7
```

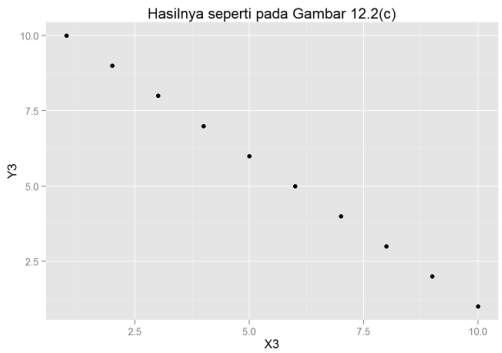
```
library(ggplot2)
ggplot(data1, aes(X1, Y1)) + geom_point() + ggtitle("Hasilnya seperti pada Gambar 12.2(a)")
```




```
ggplot(data2, aes(X2, Y2)) + geom_point() + ggtitle("Hasilnya seperti pada Gambar 12.2(b)")
```




```
ggplot(data1, aes(X3, Y3)) + geom_point() + ggtitle("Hasilnya seperti pada Gambar 12.2(c)")
```



Gambar 11.7

```
cor(data1$X1, data1$Y1, method = "pearson")
```

```
## [1] 1
```

```
cor(data2$X2, data2$Y2, method = "pearson")
```

```
## [1] -0.02557102
```

```
cor(data1$X3, data1$Y3, method = "pearson")
```

```
## [1] -1
```

Gambar 11.8

Berdasarkan Gambar 11.6, secara umum, perintah untuk menghitung koefisien korelasi linear Pearson dalam R sebagai berikut.

cor(variabel1, variabel2, method = ”pearson”)

Berdasarkan Gambar 11.6, *package* **ggplot2** diaktifkan (kode R pada baris 6), dengan maksud untuk menggunakan fungsi **ggplot**. Fungsi **ggplot** bertujuan untuk menyajikan grafik sebaran data.

Menyajikan Grafik Sebaran Data dalam R (Bagian 2)

Grafik dari sebaran data antara Ca v/s HCO3, Ca v/s Mg, Ca v/s SO4, dan Ca v/s NO2, seperti pada Gambar 11.1, akan disajikan kembali, seperti pada Gambar 11.10. Kode R disajikan pada Gambar 11.9.

```
1  data=read.csv("BandungData.csv")
2
3  Ca=data$Ca
4  HCO3=data$HCO3
5  Mg=data$Mg
6  SO4=data$SO4
7  NO2=data$NO2
8
9  library(ggplot2)
10 df <- data.frame(Ca, HCO3, Mg, SO4, NO2)
11 library(reshape)
12 df.melted <- melt(df, id = "Ca")
13 ggplot(data = df.melted, aes(Ca, y = value)) + geom_point(color="firebrick") + facet_grid(variable ~ ., scales='free_y')
14
```

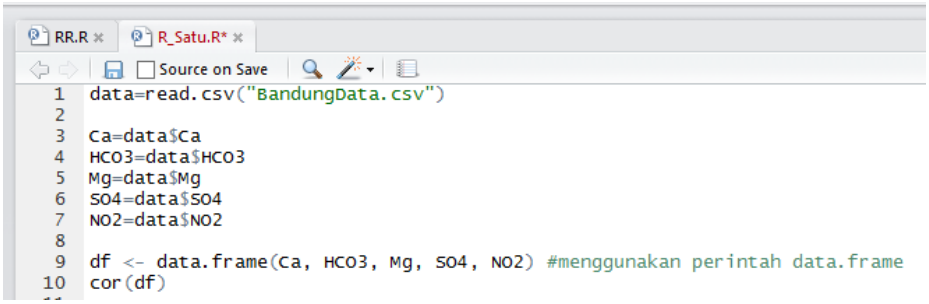
Gambar 11.9

```
data=read.csv("BandungData.csv")

Ca=data$Ca
HCO3=data$HCO3
Mg=data$Mg
SO4=data$SO4
NO2=data$NO2

library(ggplot2)
df <- data.frame(Ca, HCO3, Mg, SO4, NO2)
library(reshape)
df.melted <- melt(df, id = "Ca")
ggplot(data = df.melted, aes(Ca, y = value)) + geom_point(color="firebrick") + facet_grid(variable ~ ., scales=
'free_y')
```

Gambar 11.10

Menghitung Koefisien Korelasi Linear Pearson secara Sekaligus dengan R

Pada pemaparan sebelumnya, penghitungan nilai koefisien korelasi linear Pearson dilakukan secara satu persatu (Gambar 11.8). Dalam R, perhitungan nilai koefisien korelasi linear Pearson dapat dilakukan secara sekaligus dengan menggunakan perintah **data.frame** terlebih dahulu. Perhatikan ilustrasi berikut.



Gambar 11.11

```
data=read.csv("BandungData.csv")

Ca=data$Ca
HCO3=data$HCO3
Mg=data$Mg
SO4=data$SO4
NO2=data$NO2

df <- data.frame(Ca, HCO3, Mg, SO4, NO2) #menggunakan perintah data.frame
cor(df)
```

```
##              Ca       HCO3        Mg         SO4         NO2
## Ca    1.00000000 0.80131620 0.6469660  0.48130799  0.07567947
## HCO3  0.80131620 1.00000000 0.6848155  0.20305583  0.08185679
## Mg    0.64696597 0.68481548 1.0000000  0.44023223  0.16807301
## SO4   0.48130799 0.20305583 0.4402322  1.00000000 -0.01023144
## NO2   0.07567947 0.08185679 0.1680730 -0.01023144  1.00000000
```

Gambar 11.12

Berdasarkan Gambar 11.12, nilai koefisien korelasi linear Pearson antara Ca dan HCO3 adalah 0,80131620, nilai koefisien korelasi linear Pearson antara Ca dan Mg adalah 0,6469660, nilai koefisien korelasi linear Pearson antara Ca dan SO4 adalah 0,48130799, dan seterusnya. Di antara variabel HCO3, Mg, SO4, dan NO2, variabel HCO3 yang memiliki keeratan linear yang paling tinggi terhadap variabel Ca, yakni bernilai 0,80131620.

Contoh Perhitungan Koefisien Korelasi Linear Pearson dan Penyelesaian dalam R

Misalkan diberikan data seperti pada Tabel 11.1. Berdasarkan data pada Tabel 11.1, berikut rumus untuk menghitung nilai koefisien korelasi linear Pearson (r).

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}}$$

Tabel 11.1

X	Y
10	3,01
12	3,15
9	2,9
10	3,1
8	2,7
11	3,25
15	3,6
17	3,7
16	3,65
10	3,15

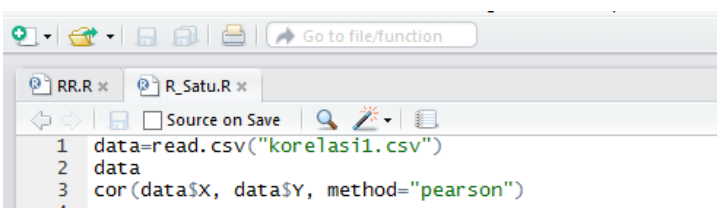
Tabel 11.2

	X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
	10	3,01	-1,8	-0,211	3,24	0,044521	0,3798
	12	3,15	0,2	-0,071	0,04	0,005041	-0,0142
	9	2,9	-2,8	-0,321	7,84	0,103041	0,8988
	10	3,1	-1,8	-0,121	3,24	0,014641	0,2178
	8	2,7	-3,8	-0,521	14,44	0,271441	1,9798
	11	3,25	-0,8	0,029	0,64	0,000841	-0,0232
	15	3,6	3,2	0,379	10,24	0,143641	1,2128
	17	3,7	5,2	0,479	27,04	0,229441	2,4908
	16	3,65	4,2	0,429	17,64	0,184041	1,8018
	10	3,15	-1,8	-0,071	3,24	0,005041	0,1278
Jumlah	118	32,21	0	0	87,6	1,00169	9,072
Rata-Rata	11,8	3,221					

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}} = \frac{9{,}072}{\sqrt{87{,}6}\sqrt{1{,}00169}} = \frac{9{,}072}{(9{,}359487165)(1{,}000844643)}$$

$$r = 0{,}968465868$$

Berdasarkan perhitungan secara manual, diperolah nilai koefisien korelasi linear Pearson $r = 0{,}968465868$. Berikut hasil perhitungan nilai koefisien korelasi linear Pearson berdasarkan R.



Gambar 11.13

```
data=read.csv("korelasi1.csv")
data
```

```
##     X    Y
## 1  10 3.01
## 2  12 3.15
## 3   9 2.90
## 4  10 3.10
## 5   8 2.70
## 6  11 3.25
## 7  15 3.60
## 8  17 3.70
## 9  16 3.65
## 10 10 3.15
```

```
cor(data$X, data$Y, method="pearson")
```

```
## [1] 0.9684659
```

Gambar 11.14

Contoh Perhitungan Covariance dan Penyelesaian dalam R

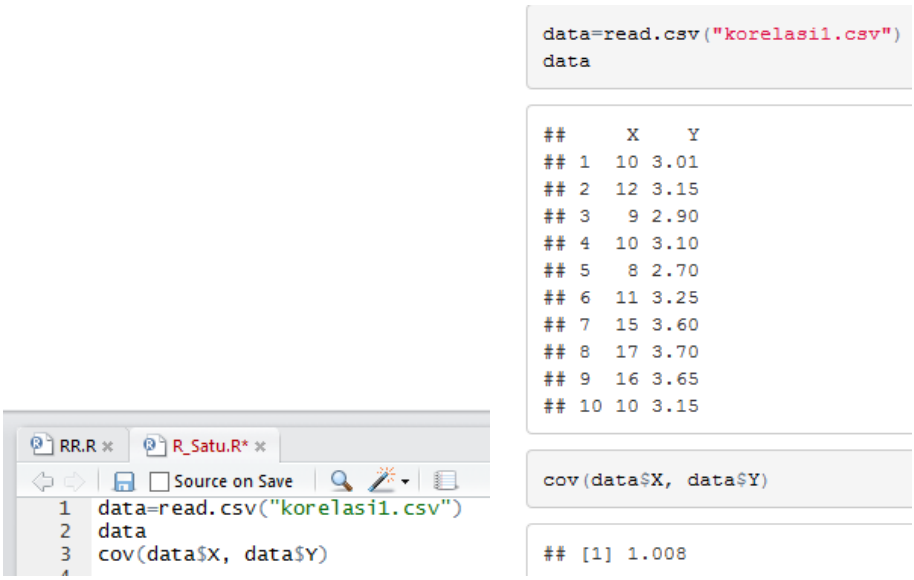
Berdasarkan data pada Tabel 11.1, berikut rumus untuk menghitung *covariance* antara variabel X dan variabel Y $(cov(X,Y))$.

$$cov(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1}$$

Perhatikan bahwa n menyatakan banyaknya data, yakni $n = 10$.

$$cov(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1} = \frac{9{,}072}{10-1} = 1{,}008$$

Berdasarkan perhitungan secara manual, diperolah nilai $cov(X,Y) = 1{,}008$. Berikut hasil perhitungan nilai $cov(X,Y)$ berdasarkan R.



Gambar 11.15

Referensi

1. Agresti, A. dan B. Finlay. 2009. *Statistical Methods for the Social Sciences, 4th Edition*. United States of America: Prentice Hall.

2. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

3. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

4. Mann, P. S. dan C.J. Lacke. 2011. *Introductory Statistics, International Student Version, 7th Edition*. Asia: John Wiley & Sons, Inc.

5. Montgomery, D. C. dan G. C. Runger. 2011. *Applied Statistics and Probability for Engineers, 5th Edition*. United States of America: John Wiley & Sons, Inc.

6. Ott, R.L. dan M. Longnecker. 2001. *An Introduction to Statistical Methods and Data Analysis, 5th Edition*. United States of America: Duxbury.

7. Smidth, R. K. dan D. H. Sanders. 2000. *Statistics a First Course, 6th Edition*. United States of America: McGraw-Hill Companies.

8. http://www.statmethods.net/stats/correlations.html

9. http://www.r-bloggers.com/correlation-and-linear-regression/

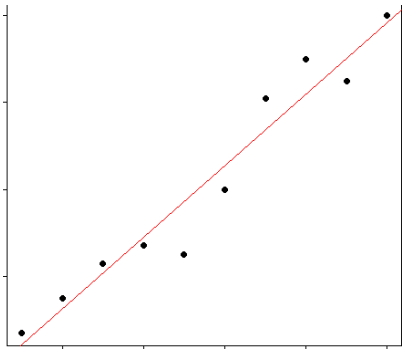
10. http://www.r-bloggers.com/pairwise-complete-correlation-considered-dangerous/

BAB 12

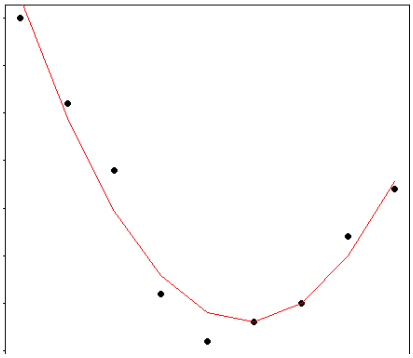
REGRESI LINEAR BERGANDA

Sekilas Regresi Linear Berganda

Regresi linear berganda (*multiple linear regression*) merupakan suatu teknik statistika yang menghasilkan suatu persamaan linear. Persamaan linear tersebut menerangkan atau menjelaskan hubungan antara variabel-variabel bebas terhadap variabel tak bebas. Dari persamaan linear tersebut juga dapat diketahui variabel bebas manakah yang memiliki kontribusi terbesar dalam pengaruhnya terhadap variabel tak bebas. Di samping itu, persamaan linear tersebut dapat digunakan untuk keperluan prediksi suatu nilai dari variabel tak bebas berdasarkan masukkan dari nilai-nilai variabel tak bebas. Gambar 12.1 dan Gambar 12.2 menyajikan kurva dari persamaan regresi linear dan persamaan persamaan regresi nonlinear.



Kurva Persamaan Regresi Linear
Gambar 12.1



Kurva Persamaan Regresi Nonlinear
Gambar 12.2

Variabel bebas dan tak bebas yang digunakan untuk membuat persamaan regresi linear bersifat metrik (interval atau rasio). Hair dkk. (2010:151) menyatakan sebagai berikut.

Multiple regression analysis is a general statistical technique used to analyze the relationship between a single dependent variable and independent variables. Its basic formulation is

$$\underset{(metric)}{Y_1} = \underset{(metric)}{X_1 + \cdots + X_n}$$

Sebagai contoh dari variabel yang bersifat metrik adalah pendapatan per bulan, penghasilan per bulan, produksi beras per tahun, tinggi badan, berat badan, dan sebagainya. **Jika variabel tak bebas bersifat non-metrik atau kategori** (nominal atau ordinal), maka alternatif teknik statistika yang dapat digunakan adalah regresi logistik, analisis diskriminan, atau pohon klasifikasi (*classification tree*). Namun **jika variabel tak bebas dan variabel bebas bersifat kategori**, maka alternatif teknik statistika yang dapat digunakan adalah regresi logistik atau pohon klasifikasi.

Beberapa Contoh Aplikasi dari Regresi Linear Berganda

Berikut diberikan beberapa contoh aplikasi dari regresi linear berganda.

- ⇨ Membuat suatu persamaan linear untuk memprediksi indeks harga saham gabungan (IHSG) berdasarkan informasi dari tingkat inflasi, harga emas dunia, dan harga minyak mentah dunia. Kemudian dari ketiga faktor tersebut, dapat ditentukan, faktor mana yang memberikan kontribusi terbesar dalam pengaruhnya terhadap indeks harga saham gabungan.

- ⇨ Membuat suatu persamaan linear untuk memprediksi atau mengestimasi laba perusahaan berdasarkan umur perusahaan, tingkat penjualan, dan besarnya ukuran perusahaan. Di samping itu, dapat diketahui seberapa besar kontribusi yang diberikan dari faktor umur perusahaan terhadap naik/turunnya laba perusahaan, dengan mengontrol pengaruh tingkat penjualan dan besarnya perusahaan. Dapat juga diketahui seberapa besar kontribusi yang diberikan dari faktor tingkat penjualan perusahaan terhadap naik/turunnya laba perusahaan, dengan mengontrol pengaruh umur perusahaan dan besarnya perusahaan.

- ⇨ Membuat suatu persamaan linear untuk memprediksi pengeluaran per bulan dari suatu rumah tangga berdasarkan informasi penghasilan per bulan, jumlah anggota keluarga, dan jumlah kendaraan yang dimiliki. Kemudian dari ketiga faktor tersebut, dapat ditentukan, faktor mana yang memberikan kontribusi terbesar dalam pengaruhnya terhadap pengeluaran per bulan.

Koefisien Korelasi Linear Pearson (Mengukur Keeratan Hubungan Linear antar Variabel)

Misalkan seorang peneliti ingin membuat model regresi linear berganda dengan menggunakan variabel indeks prestasi (Y) sebagai variabel tak bebas, variabel jumlah jam belajar dalam sehari (X_1) dan uang jajan dalam sehari (X_2) sebagai variabel bebas. Data yang telah dikumpulkan oleh peneliti disajikan dalam Tabel 12.1.

Tabel 12.1 (Data Fiktif)

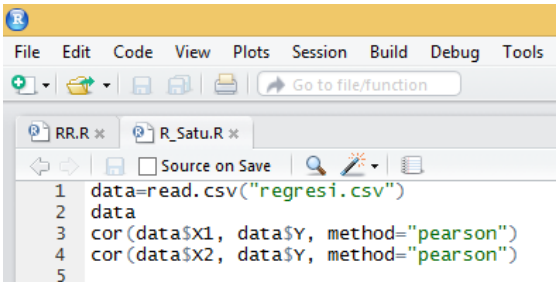
Nama	X_1	X_2	Y	Nama	X_1	X_2	Y
Ugi	10	7	3,01	Iqbal	10	7	3,02
Niar	10	7	3,15	Edi	12	7,2	3,16
Alvi	9	11	2,9	Budi	9	6	2,95
Fitri	10	8	3,1	Indah	10	8	3,12
Ridho	8	7,5	2,7	Tari	8	12	2,8
Mifdhal	11	8	3,25	Maura	11	11	3,3
Romi	13	7	3,6	Nina	15	10	3,57
Wilya	13	12	3,7	Suci	17	8	3,64
Windi	15	9,5	3,65	Febri	16	9,5	3,6
Evelin	10	10	3,15	Iman	10	10	3,15

Berdasarkan data pada Tabel 12.1, jumlah responden yang diteliti sebanyak $n = 20$ responden. Misalkan responden yang diteliti adalah mahasiswa matematika. Diketahui

responden ke-1 bernama Ugi menghabiskan waktu untuk belajar dalam sehari selama 10 jam, uang jajan satu hari Rp. 7000, dan meraih IP 3,01. Responden ke-3 bernama Alvi menghabiskan waktu untuk belajar dalam sehari selama 9 jam, uang jajan dalam satu hari Rp. 11000, dan meraih IP 2,9, dan seterusnya. Misalkan akan ditentukan:

- ⇨ Nilai koefisien korelasi linear Pearson antara X_1 dan Y
- ⇨ Nilai koefisien korelasi linear Pearson antara X_2 dan Y

Berikut hasil perhitungan nilai koefisien korelasi linear Pearson berdasarkan R.



Gambar 12.3

```
data=read.csv("regresi.csv")
data
```

```
##    X1   X2    Y
## 1  10  7.0 3.01
## 2  10  7.0 3.15
## 3   9 11.0 2.90
## 4  10  8.0 3.10
## 5   8  7.5 2.70
## 6  11  8.0 3.25
## 7  13  7.0 3.60
## 8  13 12.0 3.70
## 9  15  9.5 3.65
## 10 10 10.0 3.15
## 11 10  7.0 3.02
## 12 12  7.2 3.16
## 13  9  6.0 2.95
## 14 10  8.0 3.12
## 15  8 12.0 2.80
## 16 11 11.0 3.30
## 17 15 10.0 3.57
## 18 17  8.0 3.64
## 19 16  9.5 3.60
## 20 10 10.0 3.15
```

```
cor(data$X1, data$Y, method="pearson")
```

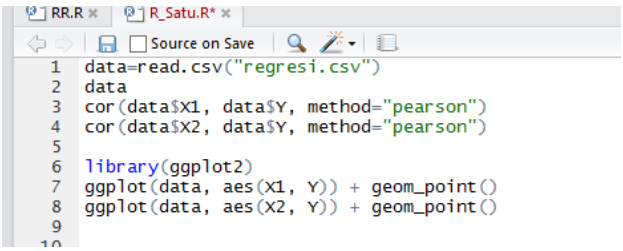
```
## [1] 0.9168692
```

```
cor(data$X2, data$Y, method="pearson")
```

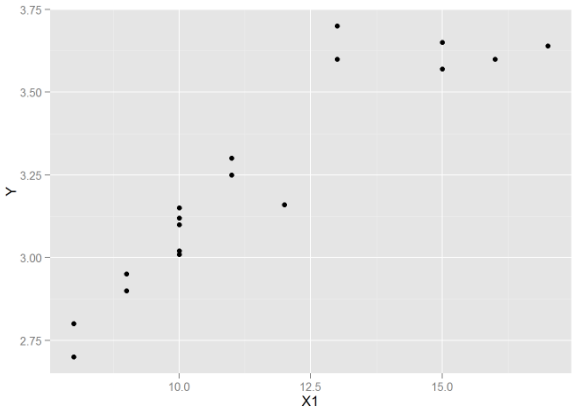
```
## [1] 0.1907533
```

Gambar 12.4

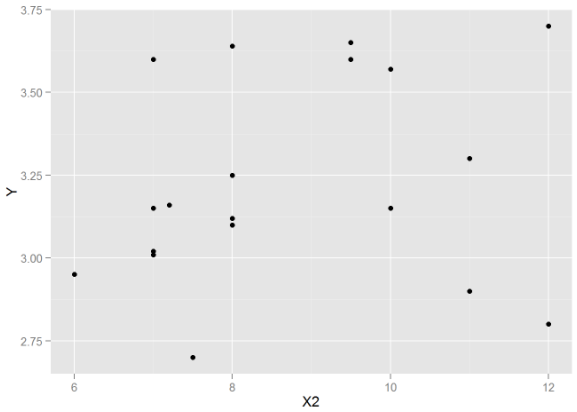
Berdasarkan Gambar 12.4, diketahui nilai koefisien korelasi linear Pearson antara X_1 dan Y sebesar 0,92 (dibulatkan 2 angka di belakang koma), dan nilai koefisien korelasi linear Pearson antara X_2 dan Y sebesar 0,19. Berdasarkan hasil tersebut, **diketahui variabel X_1 memiliki keeratan linear lebih erat terhadap Y, dibandingkan X_2**. Grafik sebaran data antara X_1 dan Y, serta X_2 dan Y, disajikan sebagai berikut (Gambar 12.6 dan Gambar 12.7).



Gambar 12.5



Gambar 12.6



Gambar 12.7

Perhatikan bahwa Gambar 12.6 merupakan grafik sebaran data antara X_1 dan Y, sementara Gambar 12.7 merupakan grafik sebaran data antara X_2 dan Y. **Dapat dilihat bahwa sebaran data pada Gambar 12.6 lebih linear dibandingkan sebaran data pada Gambar 12.7. Hal dapat diartikan bahwa variabel jam (X_1) memiliki keeratan linear lebih tinggi terhadap variabel IP (Y), dibandingkan variabel uang jajan (X_2).**

Mengestimasi Persamaan Regresi Linear Berganda

Pada pembahasan sebelumnya, diketahui bahwa variabel jam (X_1) memiliki keeratan linear lebih tinggi terhadap variabel IP (Y), dibandingkan variabel uang jajan (X_2). Selanjutnya akan diestimasi persamaan regresi linear berganda. Persamaan regresi linear berganda untuk kasus ini memiliki bentuk sebagai berikut.

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Berikut akan dihitung $\hat{\alpha}$, $\hat{\beta}_1$, dan $\hat{\beta}_2$.

$$p = n\sum X_1 Y - \sum X_1 \sum Y = (20)(746{,}3) - (227)(64{,}52) = 279{,}96$$

$$q = n\sum X_2^2 - \left(\sum X_2\right)^2 = (20)(1606{,}59) - (175{,}7)^2 = 1261{,}31$$

$$r = n\sum X_1 X_2 - \sum X_1 \sum X_2 = (20)(2001{,}9) - (227)(175{,}7) = 154{,}1$$

$$s = n\sum X_2 Y - \sum X_2 \sum Y = (20)(568{,}817) - (175{,}7)(64{,}52) = 40{,}176$$

$$t = n\sum X_1^2 - \left(\sum X_1\right)^2 = (20)(2709) - (227)^2 = 2651$$

$$u = tq - r^2 = (2651)(1261{,}31) - (154{,}1)^2 = 3319986$$

Tabel 12.2

	X_1	X_2	Y	X_1Y	X_2Y	X_1^2	X_2^2	Y^2	X_1X_2
	10	7	3.01	30.1	21.07	100	49	9.0601	70
	10	7	3.15	31.5	22.05	100	49	9.9225	70
	9	11	2.9	26.1	31.9	81	121	8.41	99
	10	8	3.1	31	24.8	100	64	9.61	80
	8	7.5	2.7	21.6	20.25	64	56.25	7.29	60
	11	8	3.25	35.75	26	121	64	10.5625	88
	13	7	3.6	46.8	25.2	169	49	12.96	91
	13	12	3.7	48.1	44.4	169	144	13.69	156
	15	9.5	3.65	54.75	34.675	225	90.25	13.3225	142.5
	10	10	3.15	31.5	31.5	100	100	9.9225	100
	10	7	3.02	30.2	21.14	100	49	9.1204	70
	12	7.2	3.16	37.92	22.752	144	51.84	9.9856	86.4
	9	6	2.95	26.55	17.7	81	36	8.7025	54
	10	8	3.12	31.2	24.96	100	64	9.7344	80
	8	12	2.8	22.4	33.6	64	144	7.84	96
	11	11	3.3	36.3	36.3	121	121	10.89	121
	15	10	3.57	53.55	35.7	225	100	12.7449	150
	17	8	3.64	61.88	29.12	289	64	13.2496	136
	16	9.5	3.6	57.6	34.2	256	90.25	12.96	152
	10	10	3.15	31.5	31.5	100	100	9.9225	100
Jumlah	227	175.7	64.52	746.3	568.817	2709	1606.59	209.9	2001.9

$$\hat{\beta}_1 = \frac{pq - rs}{u}$$

$$\hat{\beta}_1 = \frac{(279{,}96)(1261{,}31) - (154{,}1)(40{,}176)}{3319986} = 0{,}104496$$

$$\hat{\beta}_2 = \frac{st - pr}{u}$$

$$\hat{\beta}_2 = \frac{(40{,}176)(2651\ ) - (279{,}96)(154{,}1)}{3319986} = 0{,}019086$$

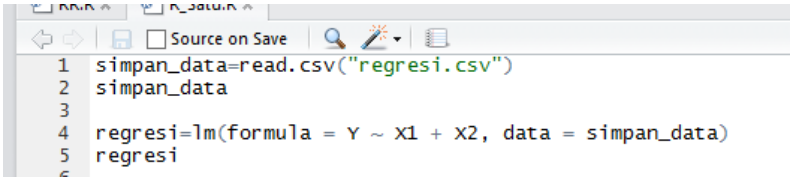
$$\hat{\alpha} = \frac{\sum Y - \hat{\beta}_1 \sum X_1 - \hat{\beta}_2 \sum X_2}{n}$$

$$\hat{\alpha} = \frac{64{,}52 - (0{,}104496)(227) - (0{,}019086)(175{,}7)}{20} = 1{,}872301$$

Maka diperoleh persamaan regresi linear berganda

$$\hat{Y} = 1{,}872301 \ + \ 0{,}104496X_1 + 0{,}019086X_2.$$

Berikut disajikan hasil perhitungan berdasarkan R.



Gambar 12.8

```
## Call:
## lm(formula = Y ~ X1 + X2, data = simpan_data)
##
## Coefficients:
## (Intercept)           X1           X2
##     1.87230      0.10450      0.01909
```

Gambar 12.9

Memprediksi Nilai Variabel Tak Bebas

Persamaan regresi linear berganda yang telah dihasilkan sebelumnya, dapat digunakan untuk memprediksi atau mengestimasi nilai dari variabel tak bebas, berdasarkan masukan nilai-nilai dari variabel bebas. Diketahui persamaan regresi linear berganda berdasarkan perhitungan sebelumnya sebagai berikut.

$$\hat{Y} = 1{,}872301 + 0{,}104496X_1 + 0{,}019086X_2$$

Misalkan akan diprediksi nilai IP, ketika jumlah jam belajar dalam sehari $X_1 = 10$ dan uang jajan dalam sehari $X_2 = 7$ (dalam ribuan).

$$\hat{Y} = 1{,}872301 + 0{,}104496(10) + 0{,}019086(7) = 3{,}050862$$

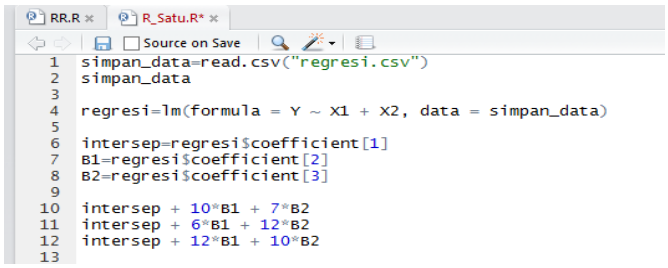
Misalkan akan diprediksi nilai IP, ketika jumlah jam belajar dalam sehari $X_1 = 6$ dan uang jajan dalam sehari $X_2 = 12$ (dalam ribuan).

$$\hat{Y} = 1{,}872301 + 0{,}104496(6) + 0{,}019086(12) = 2{,}728307$$

Misalkan akan diprediksi nilai IP, ketika jumlah jam belajar dalam sehari $X_1 = 12$ dan uang jajan dalam sehari $X_2 = 10$ (dalam ribuan).

$$\hat{Y} = 1{,}872301 + 0{,}104496(12) + 0{,}019086(10) = 3{,}317112$$

Berikut disajikan ilustrasi dalam R.



Gambar 12.10

```
regresi=lm(formula = Y ~ X1 + X2, data = simpan_data)

intersep=regresi$coefficient[1]
B1=regresi$coefficient[2]
B2=regresi$coefficient[3]

intersep + 10*B1 + 7*B2
```

```
## (Intercept)
##    3.050862
```

```
intersep + 6*B1 + 12*B2
```

```
## (Intercept)
##    2.728307
```

```
intersep + 12*B1 + 10*B2
```

```
## (Intercept)
##    3.317112
```

Gambar 12.11

Menghitung Nilai Residual untuk Setiap Pengamatan

Residual (dilambangkan dengan $\hat{e}$) merupakan selisih antara nilai variabel tak bebas (Y) dan nilai estimasi dari variabel tak bebas $(\hat{Y})$.

Tabel 12.3

No	X_1	X_2	Y	$\widehat{Y}$	$\hat{e} = Y - \hat{Y}$
1	10	7	3.01	3.050862	-0.04086
2	10	7	3.15	3.050862	0.099138
3	9	11	2.9	3.02271	-0.12271
4	10	8	3.1	3.069948	0.030052
5	8	7.5	2.7	2.851413	-0.15141
6	11	8	3.25	3.174444	0.075556
7	13	7	3.6	3.36435	0.23565
8	13	12	3.7	3.459779	0.240221
9	15	9.5	3.65	3.621057	0.028943
10	10	10	3.15	3.10812	0.04188
11	10	7	3.02	3.050862	-0.03086
12	12	7.2	3.16	3.263671	-0.10367
13	9	6	2.95	2.92728	0.02272
14	10	8	3.12	3.069948	0.050052
15	8	12	2.8	2.937299	-0.1373
16	11	11	3.3	3.231702	0.068298
17	15	10	3.57	3.6306	-0.0606
18	17	8	3.64	3.80142	-0.16142
19	16	9.5	3.6	3.725553	-0.12555
20	10	10	3.15	3.10812	0.04188

Berdasarkan Tabel 12.3, nilai estimasi Y untuk responden ke-1 adalah 3,050862. Nilai tersebut diperoleh berdasarkan hasil perhitungan berikut.

$$\hat{Y} = 1{,}872301 + 0{,}104496(10) + 0{,}019086(7) = 3{,}050862$$

Nilai residual untuk responden ke-1 dihitung sebagai berikut.

$$e = \hat{Y} - Y = 3{,}01 - 3{,}050862 = -0{,}04086$$

Nilai estimasi Y untuk responden ke-20 adalah 3,15. Nilai tersebut diperoleh berdasarkan hasil perhitungan berikut.

$$\hat{Y} = 1{,}872301 + 0{,}104496(10) + 0{,}019086(10) = 3{,}10812$$

Nilai residual untuk responden ke-20 dihitung sebagai berikut.

$$e = \hat{Y} - Y = 3{,}15 - 3{,}1082 = -0{,}04086$$

Berikut hasil perhitungan dengan R untuk memperoleh nilai estimasi IP dari tiap-tiap responden, beserta residualnya.

```
simpan_data=read.csv("regresi.csv")
simpan_data

X1=simpan_data$X1
X2=simpan_data$X2
Y=simpan_data$Y

regresi=lm(formula = Y ~ X1 + X2, data = simpan_data)

intersep=regresi$coefficient[1]
B1=regresi$coefficient[2]
B2=regresi$coefficient[3]

Y_estimasi = intersep + B1*X1 + B2*X2

Y_estimasi

Residual = Y-Y_estimasi

Residual
```

Gambar 12.12

```
X2=simpan_data$X2
Y=simpan_data$Y

regresi=lm(formula = Y ~ X1 + X2, data = simpan_data)

intersep=regresi$coefficient[1]
B1=regresi$coefficient[2]
B2=regresi$coefficient[3]

Y_estimasi = intersep + B1*X1 + B2*X2

Y_estimasi
```

```
##  [1] 3.050862 3.050862 3.022710 3.069948 2.851413 3.174444 3.364350
##  [8] 3.459779 3.621057 3.108120 3.050862 3.263671 2.927280 3.069948
## [15] 2.937299 3.231702 3.630600 3.801420 3.725553 3.108120
```

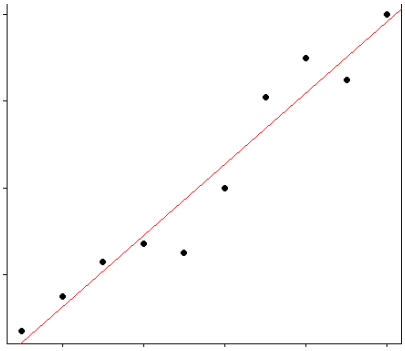
```
Residual = Y-Y_estimasi

Residual
```

```
##  [1] -0.04086218  0.09913782 -0.12270957  0.03005198 -0.15141312
##  [6]  0.07555599  0.23564985  0.24022062  0.02894325  0.04188029
## [11] -0.03086218 -0.10367133  0.02271966  0.05005198 -0.13729942
## [16]  0.06829845 -0.06059967 -0.16141996 -0.12555274  0.04188029
```

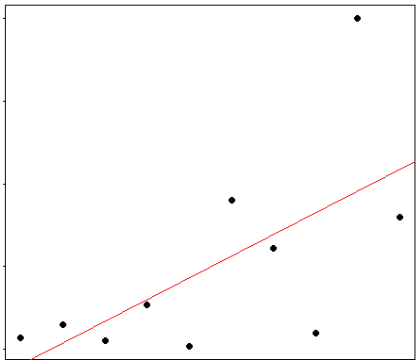
Gambar 12.13

Mengukur Kecocokkan Model Regresi Linear Berganda terhadap Data dengan Koefisien Determinasi (r^2)

Dalam regresi linear, baik sederhana (melibatkan satu variabel bebas) maupun berganda (melibatkan lebih dari satu variabel bebas), nilai dari koefisien determinasi (r^2) digunakan untuk mengukur kemampuan persamaan regresi linear dalam mencocokkan atau menyesuiakan (*fits*) data. Sebagai ilustrasi perhatikan Gambar 12.14 dan Gambar 12.15. Pada Gambar 12.14 dan Gambar 12.15 menyajikan garis persamaan regresi linear. Pada Gambar 12.14, garis persamaan regresi linear lebih baik dalam hal mencocokkan data dibandingkan garis persamaan regresi linear pada Gambar 12.15. Pada Gambar 12.14, titik-titik cenderung menyebar lebih dekat pada garis persamaan regresi linear, dibandingkan pada Gambar 12.15.

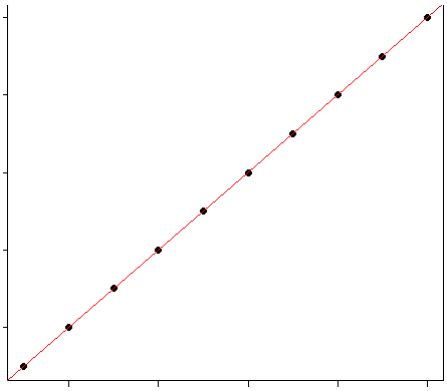


Gambar 12.14



Gambar 12.15

Nilai koefisien determinasi berkisar di antara 0 dan 1. Nilai koefisien determinasi yang bernilai 1 berarti persamaan regresi linear secara sempurna dalam mencocokkan data (Gambar 12.16). Nilai koefisien determinasi yang semakin mendekati 0, berarti kemampuan persamaan regresi linear semakin tidak baik dalam mencocokkan data. Dengan kata lain, kemampuan variabel-variabel bebas yang digunakan dalam persamaan regresi linear secara bersamaan atau simultan kurang mampu dalam hal menjelaskan *variation* variabel tak bebas (Gambar 12.15).



Gambar 12.16

Semakin tinggi nilai koefisien determinasi (mendekati 1), maka akan semakin baik suatu persamaan regresi linear dalam mencocokkan data. Dengan kata lain, kemampuan variabel-variabel bebas yang digunakan dalam persamaan regresi linear secara bersamaan atau simultan semakin baik dalam hal menjelaskan *variation* variabel tak bebas (Gujarati, 2993:87).

Pada pembahasan sebelumnya, telah diperoleh persamaan regresi linear berganda sebagai berikut.

$$\hat{Y} = 1{,}872301 + 0{,}104496X_1 + 0{,}019086X_2$$

Berikut akan dihitung nilai koefisien determinasi dari persamaan regresi linear berganda tersebut.

$$r^2 = \frac{\sum\left(\hat{Y} - \bar{Y}\right)^2}{\sum(Y - \bar{Y})^2}$$

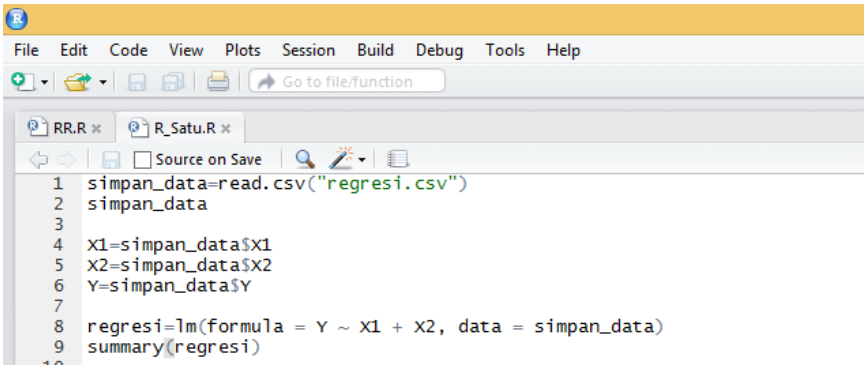
$$r^2 = \frac{1{,}501074522}{1{,}75848}$$

$$r^2 = 0{,}85362$$

Tabel 12.4

	Y	$\hat{Y}$	$\left(\hat{Y} - \bar{Y}\right)^2$	$(Y - \bar{Y})^2$
	3.01	3.050862	0.030673257	0.046656
	3.15	3.050862	0.030673257	0.005776
	2.9	3.02271	0.041326999	0.106276
	3.1	3.069948	0.024352219	0.015876
	2.7	2.851413	0.140315331	0.276676
	3.25	3.174444	0.00265802	0.000576
	3.6	3.36435	0.019140764	0.139876
	3.7	3.459779	0.054652798	0.224676
	3.65	3.621057	0.156069832	0.179776
	3.15	3.10812	0.013895762	0.005776
	3.02	3.050862	0.030673257	0.042436
	3.16	3.263671	0.001419129	0.004356
	2.95	2.92728	0.089233434	0.076176
	3.12	3.069948	0.024352219	0.011236
	2.8	2.937299	0.083348022	0.181476
	3.3	3.231702	3.25077E-05	0.005476
	3.57	3.6306	0.163700891	0.118336
	3.64	3.80142	0.331108128	0.171396
	3.6	3.725553	0.249552936	0.139876
	3.15	3.10812	0.013895762	0.005776
Jumlah	64.52	64.52	1.501074522	1.75848
Rata-Rata	3.226	3.226	0.075053726	0.087924

Gambar 12.18 merupakan hasil perhitungan dengan R. Pada Gambar 12.18, nilai koefisien determinasi (***R-squared***) bernilai 0,8536. Nilai tersebut dapat diinterpretasikan variabel jumlah jam belajar dan uang jajan mampu menjelaskan atau menerangkan *variation* dari variabel IP sebesar 85,36%, sisanya sebesar 14,64% dijelaskan oleh variabel atau faktor lain.



Gambar 12.17

```
X1=simpan_data$X1
X2=simpan_data$X2
Y=simpan_data$Y

regresi=lm(formula = Y ~ X1 + X2, data = simpan_data)
summary(regresi)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = simpan_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16142 -0.10843  0.02583  0.05461  0.24022
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.87230    0.17730  10.560 6.93e-09 ***
## X1           0.10450    0.01073   9.742 2.27e-08 ***
## X2           0.01909    0.01555   1.227    0.236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1231 on 17 degrees of freedom
## Multiple R-squared:  0.8536, Adjusted R-squared:  0.8364
## F-statistic: 49.57 on 2 and 17 DF,  p-value: 8.065e-08
```

Gambar 12.18

Menguji Kecocokan Persamaan Regresi Linear terhadap Data dengan Uji F

Uji *F* digunakan untuk menguji apakah persamaan regresi linear yang telah diperoleh benar-benar bermakna atau signifikan secara statistika (*statistically significant*) mampu, dalam hal mencocokkan data. Hipotesis nol menyatakan bahwa kemampuan persamaan regresi linear dalam mencocokkan data tidak signifikan. Dengan kata lain, kemampuan variabel-variabel bebas secara simultan atau bersamaan dalam menjelaskan *variation* variabel tak bebas tidak signifikan. Secara matematis, untuk hipotesis nol dapat dinyatakan dalam persamaan sebagai berikut.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_m = 0.$$

Pada persamaan di atas berarti seluruh koefisien regresi populasi dari variabel bebas bernilai 0. Perhatikan bahwa$m$ menyatakan jumlah variabel bebas yang digunakan dalam persamaan regresi linear. Hipotesis alternatif menyatakan bahwa kemampuan persamaan regresi linear

dalam mencocokkan data signifikan secara statiska mampu menjelaskan *variation* dari variabel bebas.

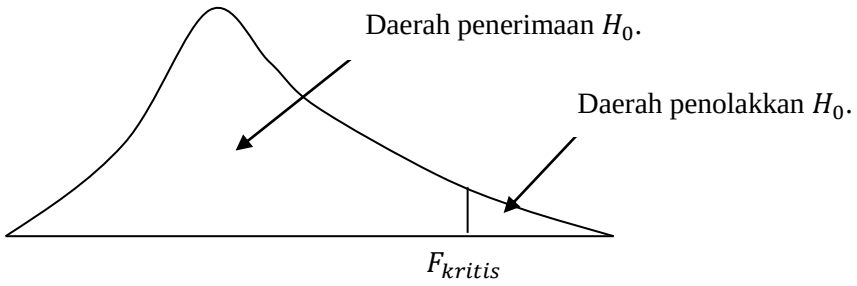
Nilai statistik dari uji F (F_{hitung}) dibandingkan dengan nilai kritis F (F_{kritis}) untuk pengambilan keputusan terhadap hipotesis. Untuk menentukan nilai kritis F, terlebih dahulu menghitung nilai derajat bebas pembilang (*numerator*) dan derajat bebas penyebut (*denominator*). Derajat bebas pembilang dan derajat bebas penyebut dihitung dengan rumus sebagai berikut.

$$Derajat\ bebas\ pembilang = k - 1.$$
$$Derajat\ bebas\ penyebut = n - k.$$

Perhatikan bahwa k menyatakan jumlah variabel, sedangkan n menyatakan jumlah pengamatan atau elemen dalam sampel. Berikut aturan pengambilan keputusan berdasarkan uji F.

$$Jika\ F_{hitung} \leq F_{kritis}, maka\ \ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ F_{hitung} > F_{kritis}, maka\ \ H_0\ ditolak\ dan\ H_1\ diterima.$$

Gambar 12.19 menyajikan daerah keputusan untuk uji F.

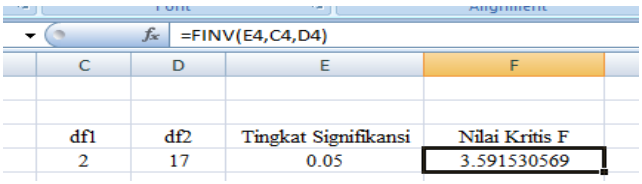


Gambar 12.19

Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan menggunakan pendekatan nilai probabilitas dari uji F. Nilai probabilitas dari uji F dibandingkan dengan tingkat signifikansi yang digunakan. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan pendekatan nilai probabilitas.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

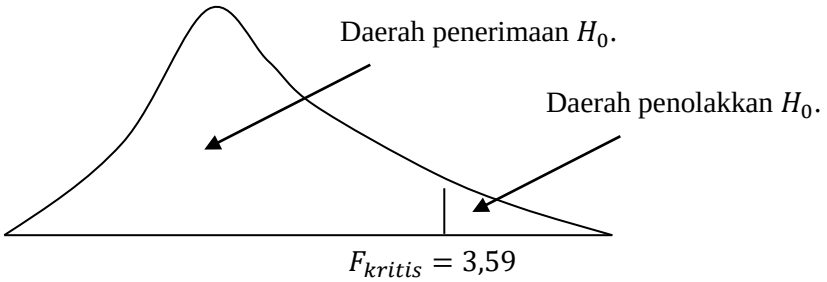
Diketahui nilai statistik dari uji F (*F-statistic*) adalah 49,57 (perhatikan Gambar 3.28). Diketahui jumlah pengamatan atau elemen dalam sampel adalah $n = 20$ dan jumlah variabel adalah $k = 3$. Maka nilai derajat bebas pembilang adalah $k - 1 = 3 - 1 = 2$ dan nilai derajat bebas penyebut adalah $n - k = 20 - 3 = 17$. Nilai kritis F dengan derajat bebas pembilang 2, derajat bebas penyebut 17, dan tingkat signifikansi 5% adalah 3,59.



Gambar 12.20 Menentukan Nilai Kritis F dengan *Microsoft Excel*

Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji F.

$$Jika\ F_{hitung} \leq F_{kritis}, maka\ \ H_0\ diterima\ dan\ H_1 \text{ ditolak.}$$
$$jika\ F_{hitung} > F_{kritis}, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$



Gambar 12.21

Perhatikan bahwa karena nilai statistik dari uji F, yakni 49,57 lebih besar dibandingkan nilai kritis F, maka hipotesis nol ditolak dan hipotesis alternatif diterima. Hal ini berarti persamaan regresi linear yang dihasilkan signifikan secara statistika mampu dalam hal mencocokkan data.

Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan menggunakan pendekatan nilai probabilitas dari uji F. Nilai probabilitas dari uji F dibandingkan dengan tingkat signifikansi yang digunakan. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan nilai probabilitas.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Berdasarkan Gambar 12.18, diketahui nilai probabilitas (*p-value* atau *probability-value*) adalah

$$8.065e - 08 = \frac{8.065}{10^8} = 0{,}0000000865.$$

Karena nilai probabilitas tersebut lebih kecil dibandingkan $\alpha = 0{,}05$, maka hipotesis nol ditolak dan hipotesis alternatif diterima. Berikut rumus untuk menghitung nilai statistik dari uji *F*.

$$F_{hitung} = \frac{\frac{r^2}{k-1}}{\frac{1-r^2}{n-k}}$$

Sehingga nilai statistik dari uji F diperoleh sebagai berikut.

$$F = \frac{\frac{0{,}85362}{3-1}}{\frac{1-0{,}85362}{20-3}}$$

$$F = 49{,}56804208$$

Uji Signifikansi Koefisien Regresi Secara Individu dengan Uji t

Dalam regresi linear berganda, uji t digunakan untuk menguji signifikansi dari masing-masing koefisien regresi populasi. Signifikansi koefisien regresi populasi diuji berdasarkan koefisien regresi sampel. Berikut perumusan hipotesis untuk uji signifikansi koefisien regresi secara individu.

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$

Perhatikan bahwa hipotesis nol menyatakan koefisien regresi populasi ke-i (β_i) bernilai nol. Dengan kata lain, variabel bebas ke-i memiliki pengaruh yang tidak signifikan secara statistika terhadap variabel tak bebas, dengan mengontrol pengaruh dari variabel bebas lain. Hipotesis alternatif menyatakan koefisien regresi populasi ke-i (β_i) tidak bernilai nol. Dengan kata lain, variabel bebas ke-i memiliki pengaruh yang signifikan secara statistika terhadap variabel tak bebas, dengan mengontrol pengaruh dari variabel bebas lain.

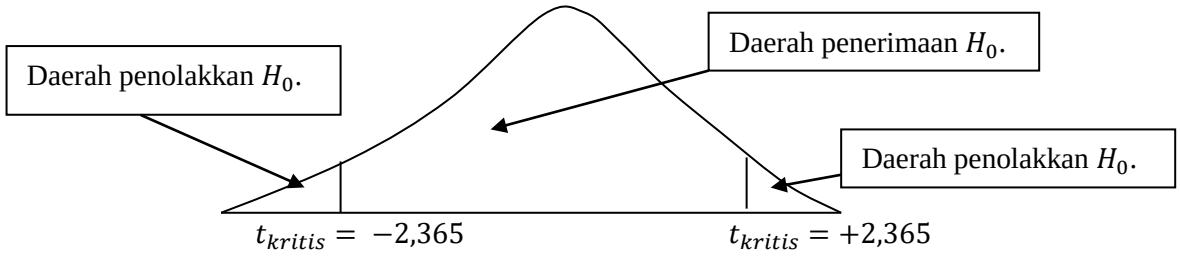
Untuk pengambilan keputusan terhadap hipotesis, dapat dilakukan dengan membandingkan nilai statistik dari uji t (t_{hitung}) terhadap nilai kritis t (t_{kritis}). Sebelum menghitung nilai kritis t, terlebih dahulu menghitung nilai derajat. Berikut rumus untuk menghitung nilai derajat bebas.

$$Derajat\ bebas = n - k.$$

Perhatikan bahwa n menyatakan jumlah pengamatan atau elemen dalam sampel, sedangkan k merupakan jumlah variabel. Andaikan jumlah pengamatan atau elemen dalam sampel sebanyak 10 dan jumlah variabel adalah 3 (jumlah variabel bebas adalah 2 dan variabel tak bebas adalah 1), sehingga derajat bebas adalah $10 - 3 = 7$. Misalkan tingkat signifikansi yang digunakan adalah 5%, sehingga nilai kritis t dengan derajat bebas 7 dan tingkat signifikansi 5% adalah $\pm 2{,}365$. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji t.

$$Jika\ |t_{hitung}| \leq |t_{kritis}|, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ |t_{hitung}| > |t_{kritis}|, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan menggunakan pendekatan nilai probabilitas dari uji t. Nilai probabilitas dari uji t dibandingkan dengan tingkat signifikansi yang digunakan. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan pendekatan nilai probabilitas.



Gambar 12.22

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

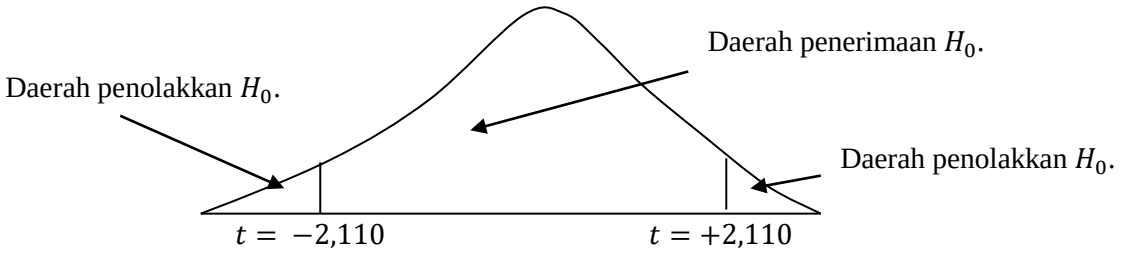
Berikut akan ditentukan apakah faktor jumlah jam belajar dalam sehari mempengaruhi IP secara signifikan (signifikan secara statistika), dengan mengontrol pengaruh uang jajan dalam sehari. *Output* R pada Gambar 12.18 disajikan kembali pada Gambar 12.23.

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.87230    0.17730  10.560 6.93e-09 ***
## X1           0.10450    0.01073   9.742 2.27e-08 ***
## X2           0.01909    0.01555   1.227    0.236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1231 on 17 degrees of freedom
## Multiple R-squared:  0.8536, Adjusted R-squared:  0.8364
## F-statistic: 49.57 on 2 and 17 DF,  p-value: 8.065e-08
```

Gambar 12.23

Diketahui nilai statistik dari uji t untuk variabel jumlah jam belajar dalam sehari (X_2) adalah 9,742. Nilai kritis t dengan derajat bebas $n - k = 20 - 3 = 17$ dan tingkat signifikansi 5% adalah $\pm 2{,}110$. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan uji t.

$$Jika\ |t_{hitung}| \leq |t_{kritis}|, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ |t_{hitung}| > |t_{kritis}|, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$



Gambar 12.24

Perhatikan bahwa karena $|t_{hitung}| > |t_{kritis}|$, yakni $9{,}742 > 2{,}110$, maka disimpulkan bahwa faktor jumlah jam belajar memiliki pengaruh yang signifikan (signifikan secara statistika) terhadap IP, dengan mengontrol pengaruh uang jajan dalam sehari.

Pengambilan keputusan terhadap hipotesis juga dapat dilakukan dengan menggunakan pendekatan nilai probabilitas dari uji t. Nilai probabilitas dari uji t dibandingkan dengan tingkat signifikansi yang digunakan. Berikut aturan pengambilan keputusan terhadap hipotesis berdasarkan nilai probabilitas.

$$Jika\ nilai\ probabilitas \geq tingkat\ signifikansi, maka\ H_0\ diterima\ dan\ H_1\ ditolak.$$
$$Jika\ nilai\ probabilitas < tingkat\ signifikansi, maka\ H_0\ ditolak\ dan\ H_1\ diterima.$$

Nilai probabilitas dari uji t berdasarkan variabel jumlah jam belajar dalam sehari (lihat kolom Pr(> |t|)) adalah

$$2.27e - 08 = \frac{2.27}{10^8} = 0.0000000227.$$

Karena nilai probabilitas tersebut lebih kecil dibandingkan tingkat signifikansi $\alpha = 5\%$, maka disimpulkan bahwa faktor jumlah jam belajar dalam sehari mempengaruhi IP secara signifikan (signifikan secara statistika), dengan mengontrol pengaruh uang jajan dalam sehari. Diketahui nilai koefisien regresi untuk variabel jumlah jam belajar dalam sehari adalah 0,10450. Nilai tersebut dapat diinterpretasikan ketika jumlah jam belajar dalam sehari ditambah satu jam, maka diharapkan (*expected*) nilai indeks prestasi meningkat sebesar 0,1045, ketika pengaruh dari uang jajan dalam sehari dipertahankan konstan.

Selanjutnya akan ditentukan apakah faktor uang jajan dalam sehari (X_2) mempengaruhi IP secara signifikan (signifikan secara statistika), dengan mengontrol pengaruh jumlah jam belajar dalam sehari. Perhatikan bahwa karena $|t_{hitung}| \leq |t_{kritis}|$, yakni $1{,}227 < 2{,}110$, maka disimpulkan bahwa faktor uang jajan dalam sehari tidak mempengaruhi IP secara signifikan (signifikan secara statistika), dengan mengontrol pengaruh uang jajan dalam sehari. Dengan kata lain, pengaruh yang diberikan oleh faktor uang jajan dalam sehari terhadap IP sangat lemah, dengan mengontrol pengaruh jumlah jam belajar dalam sehari. Nilai probabilitas dari uji t berdasarkan variabel uang jajan dalam sehari adalah 0,236. Karena nilai probabilitas tersebut lebih besar dibandingkan tingkat signifikansi $\alpha = 5\%$, maka disimpulkan bahwa faktor uang jajan dalam sehari tidak mempengaruhi IP secara signifikan (signifikan secara statistika), dengan mengontrol pengaruh jumlah jam belajar dalam sehari.

Referensi

1. Agresti, A. dan B. Finlay. 2009. *Statistical Methods for the Social Sciences, 4th Edition*. United States of America: Prentice Hall.

2. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

3. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

4. Gujarati, D.N. 2003. *Basic Econometrics, 4th Edition*. New York: McGraw-Hill.

5. Hair, J.F Jr., R.E. Anderson, B.J. Babin, dan W.C. Black. 2010. *Multivariate Data Analysis, 7th Edition*. Pearson Prentice Hall.

6. Johnson, R.A. dan D.W. Wichern. 2007. *Applied Multivariate Statistical Analysis, 6th Edition*. United States of America: Prentice Hall.

7. Malhotra, N.K. dan D.F. Birks. 2006. *Marketing Research, An Applied Approach, 2nd European Edition*. London: Prentice Hall.

8. Montgomery, D.C. dan G.C. Runger. 2011. *Applied Statistics and Probability for Engineers, 5th Edition*. United States of America: John Wiley & Sons, Inc.

9. Stevens, J.P. 2009. *Applied Multivariate Statistics For The Social Science, 5th Edition*. New York: Routledge.

10. Supranto, J. 2004. Ekonometri, Buku Kedua. Jakarta: Ghalia Indonesia.

11. Supranto, J. 2005. Ekonometri, Buku Kesatu. Jakarta: Ghalia Indonesia.

BAB 13

REGRESI LOGISTIK

Sekilas Regresi Logistik

Dalam regresi linear, baik sederhana maupun berganda, variabel tak bebas bersifat metrik (interval atau rasio), sedangkan dalam regresi logistik, **variabel tak bebas bersifat non-metrik** (memiliki kategori). Pada regresi linear, variabel bebas bersifat metrik (interval atau rasio), sedangkan dalam regresi logistik, **variabel bebas dapat bersifat metrik atau non-metrik atau kombinasi dari keduanya**. Hair dkk. (2010:314) menyatakan sebagai berikut.

"*Logistic regression may be described as estimating the relationship between a single non-metric (binary) dependent variable and set of metric or non-metric independent variables, in this general form:*

$$\underset{(\textit{binary non-metric})}{Y_1} = \underset{(\textit{non-metric and metric})}{X_1 + X_2 + X_3 + \cdots + X_n}$$"

Sejalan dengan Hair, Field (2009:265) menyatakan sebagai berikut.

"Logistic regression is multiple regression but with an outcome variable that is a categorical variable and predictors variables that are continuous or categorical".

Pada regresi logistik, jika variabel tak bebas memiliki dua kategori, maka disebut regresi logistik biner (*binary regression logistic*). Namun, jika variabel tak bebas memiliki lebih dari dua kategori, maka disebut regresi logistik multinomial (*multinomial/polychotomous logistic regression*). Secara umum, persamaan regresi logistik sederhana (melibatkan satu variabel bebas) memiliki bentuk sebagai berikut.

$$\ln\left[\frac{P(y=1)}{1-P(y=1)}\right] = \alpha + \beta x$$

Perhatikan bahwa $P(y=1)$ menyatakan probabilitas terjadinya kejadian sukses (*success*), sedangkan $1-P(y=1)$ menyatakan probabilitas terjadinya kejadian gagal (*failure*). Rasio dari $\frac{P(y=1)}{1-P(y=1)}$ disebut dengan *odds*. Sebagai contoh misalkan $P(y=1) = 0{,}8$, maka

$$\frac{P(y=1)}{1-P(y=1)} = \frac{0{,}8}{1-0{,}8} = 4.$$

Nilai 4 tersebut dapat diartikan kejadian untuk terjadinya sukses 4 kali lebih mungkin (*as likely as*) dibandingkan untuk terjadinya gagal. Misalkan diberikan data seperti pada Tabel 13.1. Berdasarkan data pada Tabel 13.1, pada variabel kelulusan, misalkan nilai 1 menyatakan lulus, sedangkan nilai 0 menyatakan tidak lulus. Probabilitas untuk lulus dengan menggunakan metode A adalah $\frac{1}{4}$, maka probabilitas untuk tidak lulus dengan menggunakan metode A adalah $1-\frac{1}{4}=\frac{3}{4}$. Nilai *odds* pada metode A adalah

$$\frac{P(y=1)}{1-P(y=1)} = \frac{1/4}{1-1/4} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Nilai $\frac{1}{3}$ tersebut dapat diartikan kejadian untuk lulus dengan menggunakan metode A $\frac{1}{3}$ kali lebih mungkin dibandingkan untuk tidak lulus. Dengan kata lain, kejadian untuk tidak lulus dengan menggunakan metode A 3 kali lebih mungkin dibandingkan untuk lulus. Probabilitas untuk lulus dengan menggunakan metode B adalah $\frac{3}{4}$, maka probabilitas untuk tidak lulus dengan menggunakan metode B adalah $1 - \frac{3}{4} = \frac{1}{4}$. Maka nilai *odds* pada metode B adalah

$$\frac{P(y=1)}{1-P(y=1)} = \frac{3/4}{1-3/4} = \frac{3/4}{1/4} = 3.$$

Nilai 3 tersebut menyatakan kejadian untuk lulus dengan menggunakan metode B 3 kali lebih mungkin dibandingkan untuk tidak lulus. Jika nilai *odds* pada metode B dibagi dengan nilai *odds* pada metode A, maka diperoleh

$$\frac{odds\ metode\ B}{odds\ metode\ A} = \frac{3}{\frac{1}{3}} = 9.$$

Nilai 9 dapat diinterpretasikan mahasiswa dengan menggunakan metode B untuk lulus 9 kali lebih mungkin dibandingkan dengan mahasiswa dengan menggunakan metode A. Nilai 9 tersebut disebut *odds ratio*.

Tabel 13.1

Responden	Kelulusan	Metode
1	1	A
2	0	A
3	0	A
4	0	A
5	1	B
6	1	B
7	1	B
8	0	B

Persamaan regresi logistik sederhana untuk probabilitas terjadinya sukses memiliki bentuk sebagai berikut.

$$P(y=1) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Persamaan regresi logistik untuk probabilitas dapat digunakan untuk mengestimasi probabilitas atau kemungkinan terjadinya suatu variabel tak bebas. Persamaan regresi logistik biner berganda memiliki bentuk umum

$$\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

dan persamaan regresi logistik biner berganda untuk probabilitas terjadinya sukses memiliki bentuk umum

$$P(y=1)=\frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}.$$

Contoh Kasus Regresi Logistik

Andaikan diberikan data dari 100 responden mengenai usia, serta ada tidaknya penyakit gula.

Tabel 13.2

Responden	Penyakit Gula	Usia	Responden	Penyakit Gula	Usia
1	Tidak	20	51	Ya	44
2	Tidak	21	52	Ya	44
3	Tidak	23	53	Tidak	45
4	Tidak	25	54	Ya	45
5	Tidak	25	55	Tidak	46
6	Tidak	27	56	Ya	46
7	Ya	26	57	Tidak	47
8	Tidak	29	58	Tidak	47
9	Tidak	28	59	Ya	47
10	Tidak	29	60	Tidak	48
11	Tidak	30	61	Ya	48
12	Tidak	30	62	Ya	48
13	Tidak	30	63	Tidak	49
14	Tidak	30	64	Tidak	49
15	Tidak	31	65	Ya	49
16	Ya	31	66	Tidak	50
17	Tidak	32	67	Ya	50
18	Tidak	32	68	Tidak	51
19	Tidak	33	69	Tidak	52
20	Tidak	34	70	Ya	52
21	Tidak	34	71	Ya	53
22	Tidak	34	72	Ya	53
23	Ya	34	73	Ya	54
24	Tidak	34	74	Tidak	55
25	Tidak	34	75	Ya	55
26	Tidak	35	76	Ya	55
27	Tidak	35	77	Ya	56
28	Tidak	36	78	Ya	56
29	Ya	36	79	Ya	56
30	Tidak	36	80	Tidak	57
31	Tidak	37	81	Tidak	57
32	Ya	37	82	Ya	57
33	Tidak	37	83	Ya	57

34	Tidak	38	84	Ya	57
35	Tidak	38	85	Ya	57
36	Tidak	39	86	Tidak	58
37	Ya	39	87	Ya	58
38	Tidak	40	88	Ya	58
39	Ya	40	89	Ya	59
40	Tidak	41	90	Ya	59
41	Tidak	41	91	Tidak	60
42	Tidak	41	92	Ya	60
43	Tidak	42	93	Ya	61
44	Tidak	42	94	Ya	62
45	Ya	43	95	Ya	62
46	Tidak	43	96	Ya	63
47	Tidak	43	97	Tidak	64
48	Ya	43	98	Ya	64
49	Tidak	44	99	Ya	65
50	Tidak	44	100	Ya	69

Berdasarkan data pada Tabel 13.2, diketahui responden ke-1 berusia 20 tahun dan tidak terkena penyakit gula, responden ke-2 berusia 21 tahun dan tidak terkena penyakit gula, responden ke-100 berusia 69 tahun dan terkena penyakit gula. Data pada Tabel 13.2 disajikan dalam tabel, seperti pada Tabel 13.3.

Tabel 13.3

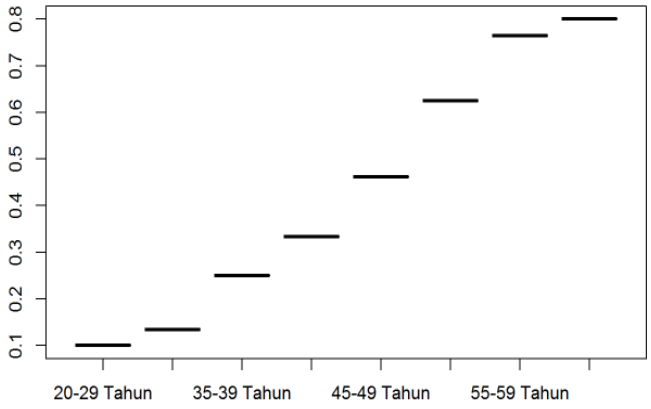
Kelompok Usia	Frekuensi	Penyakit Gula		Rata-Rata (Proporsi)
		Tidak	Ya	
20-29 Tahun	10	9	1	0.1
30-34 Tahun	15	13	2	0.133333333
35-39 Tahun	12	9	3	0.25
40-44 Tahun	15	10	5	0.333333333
45-49 Tahun	13	7	6	0.461538462
50-54 Tahun	8	3	5	0.625
55-59 Tahun	17	4	13	0.764705882
60-69 Tahun	10	2	8	0.8
Jumlah	100	57	43	0.43

Berdasarkan Tabel 13.3, diketahui dari 10 responden pada kelompok usia 20-29 tahun, sebanyak 1 (10% responden dari kelompok usia 20-29 tahun mengalami penyakit gula) responden yang mengalami penyakit gula. Diketahui dari 15 responden pada kelompok usia 30-34 tahun, sebanyak 2 (13,3% responden dari kelompok usia 30-34 tahun mengalami penyakit gula) responden yang mengalami penyakit gula. Data pada Tabel 13.3 disajikan secara visual, seperti pada Gambar 13.1. Pada Gambar 13.1, sumbu horizontal menyatakan kelompok usia, sementara sumbu vertikal menyatakan persentase. Berdasarkan gambar 13.1, **semakin tinggi kelompok usia responden**, **maka resiko untuk terkena penyakit gula juga semakin tinggi**.

```
simpan_data=read.csv("proporsi.csv")
simpan_data
```

```
##          Usia Persentase
## 1 20-29 Tahun  0.1000000
## 2 30-34 Tahun  0.1333333
## 3 35-39 Tahun  0.2500000
## 4 40-44 Tahun  0.3333333
## 5 45-49 Tahun  0.4615385
## 6 50-54 Tahun  0.6250000
## 7 55-59 Tahun  0.7647059
## 8 60-69 Tahun  0.8000000
```

```
Usia=simpan_data$Usia
Persentase=simpan_data$Persentase

plot(Usia, Persentase)
```



Gambar 13.1

Mengestimasi Persamaan Regresi Logistik

Berdasarkan data pada Tabel 13.2, diketahui variabel tak bebas (*dependent*) **penyakit gula** bersifat non-metrik, yakni berupa kategori. Kategori "Ya" diberi kode angka 1, sementara kategori "Tidak" diberi kode angka 0. Pada variabel bebas (*independent*) **usia** bersifat metrik. **Salah satu syarat penggunaan metode regresi logistik ialah data pada variabel tak bebas bersifat non-metrik (kategori)**. Gambar 13.2 menyajikan kode R, yang apabila dieksekusi kode tersebut, akan diperoleh persamaan regresi logistik (Gambar 13.4).

```
simpan_data=read.csv("data1.csv")
simpan_data
Penyakit=simpan_data$Penyakitgula
Usia=simpan_data$Usia
regresi_logistik=glm(formula = Penyakit ~ Usia , family = "binomial")
regresi_logistik
summary(regresi_logistik)

```

Gambar 13.2

Data disimpan dengan nama **data1.csv**.

	A	B
1	Penyakitgula	Usia
2	Tidak	20
3	Tidak	21
4	Tidak	23
5	Tidak	25
6	Tidak	25
7	Tidak	27
8	Ya	26
9	Tidak	29
10	Tidak	28
11	Tidak	29
12	Tidak	30
13	Tidak	30
14	Tidak	30
15	Tidak	30
16	Tidak	31
17	Ya	31
18	Tidak	32
19	Tidak	32
20	Tidak	33
21	Tidak	34
22	Tidak	34
23	Tidak	34

	A	B
22	Tidak	34
23	Tidak	34
24	Ya	34
25	Tidak	34
26	Tidak	34
27	Tidak	35
28	Tidak	35
29	Tidak	36
30	Ya	36
31	Tidak	36
32	Tidak	37
33	Ya	37
34	Tidak	37
35	Tidak	38
36	Tidak	38
37	Tidak	39
38	Ya	39
39	Tidak	40
40	Ya	40
41	Tidak	41
42	Tidak	41
43	Tidak	41
44	Tidak	42
45	Tidak	42
46	Ya	43

	A	B
46	Ya	43
47	Tidak	43
48	Tidak	43
49	Ya	43
50	Tidak	44
51	Tidak	44
52	Ya	44
53	Ya	44
54	Tidak	45
55	Ya	45
56	Tidak	46
57	Ya	46
58	Tidak	47
59	Tidak	47
60	Ya	47
61	Tidak	48
62	Ya	48
63	Ya	48
64	Tidak	49
65	Tidak	49
66	Ya	49
67	Tidak	50
68	Ya	50
69	Tidak	51
70	Tidak	52

	A	B
78	Ya	56
79	Ya	56
80	Ya	56
81	Tidak	57
82	Tidak	57
83	Ya	57
84	Ya	57
85	Ya	57
86	Ya	57
87	Tidak	58
88	Ya	58
89	Ya	58
90	Ya	59
91	Ya	59
92	Tidak	60
93	Ya	60
94	Ya	61
95	Ya	62
96	Ya	62
97	Ya	63
98	Tidak	64
99	Ya	64
100	Ya	65

Gambar 13.3

```
summary(regresi_logistik)
```

```
##
## Call:
## glm(formula = Penyakit ~ Usia, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9845  -0.8411  -0.4626   0.8200   2.2562
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.39351    1.14585  -4.707 2.51e-06 ***
## Usia         0.11269    0.02431   4.635 3.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 136.66  on 99  degrees of freedom
## Residual deviance: 106.79  on 98  degrees of freedom
## AIC: 110.79
##
## Number of Fisher Scoring iterations: 4
```

Gambar 13.4

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Usia	.113	.024	21.488	1	.000	1.119
	Constant	-5.394	1.146	22.156	1	.000	.005

a. Variable(s) entered on step 1: Usia.

Gambar 13.5 Hasil berdasarkan SPSS

Berdasarkan Gambar 13.4 atau Gambar 13.5, diperoleh persamaan regresi logistik untuk memprediksi probabilitas terjadinya penyakit gula sebagai berikut.

$$\hat{P}(\boldsymbol{y}=\mathbf{1}) = \frac{e^{-5{,}39351+0{,}11269Usia}}{1+e^{-5{,}39351+0{,}11269Usia}}$$

Mengestimasi atau Memprediksi Nilai Peluang atau Probabilitas Responden (Predicted Probability)

Persamaan regresi logistik untuk probabilitas, seperti yang telah dihasilkan sebelumnya, dapat digunakan untuk memprediksi atau mengestimasi peluang terjadinya penyakit gula, berdasarkan usia responden. Sebagai contoh, misalkan ingin diketahui perkiraan atau prediksi peluang seseorang terkena penyakit gula, ketika berusia 20 tahun. Perhitungannya sebagai berikut.

$$\hat{P}(\boldsymbol{y}=\mathbf{1}) = \frac{e^{-5{,}39351+0{,}11269Usia}}{1+e^{-5{,}39351+0{,}11269Usia}} = \frac{e^{-5{,}39351+0{,}11269(20)}}{1+e^{-5{,}39351+0{,}11269(20)}} = 0.041498653$$

Misalkan ingin diketahui prediksi peluang seseorang terkena penyakit gula, ketika berusia 21 tahun. Perhitungannya sebagai berikut.

$$\hat{P}(\boldsymbol{y}=\mathbf{1}) = \frac{e^{-5{,}39351+0{,}11269Usia}}{1+e^{-5{,}39351+0{,}11269Usia}} = \frac{e^{-5{,}39351+0{,}11269(21)}}{1+e^{-5{,}39351+0{,}11269(21)}} = 0{,}046220019$$

Misalkan ingin diketahui prediksi peluang seseorang terkena penyakit gula, ketika berusia 45 tahun. Perhitungannya sebagai berikut.

$$\hat{P}(\boldsymbol{y}=\mathbf{1}) = \frac{e^{-5,39351+0,11269Usia}}{1+e^{-5,39351+0,11269Usia}} = \frac{e^{-5,39351+0,11269(45)}}{1+e^{-5,39351+0,11269(25)}} = 0,420076344$$

Misalkan ingin diketahui prediksi peluang seseorang terkena penyakit gula, ketika berusia 60 tahun. Perhitungannya sebagai berikut.

$$\hat{P}(\boldsymbol{y}=\mathbf{1}) = \frac{e^{-5,39351+0,11269Usia}}{1+e^{-5,39351+0,11269Usia}} = \frac{e^{-5,39351+0,11269(60)}}{1+e^{-5,39351+0,11269(60)}} = 0,797039037$$

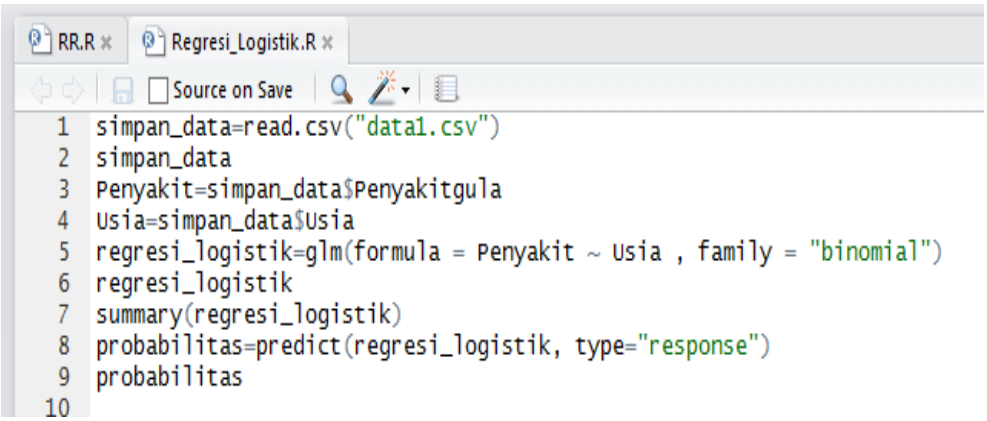
Berdasarkan perhitungan yang telah diperoleh, dapat dilihat bahwa semakin tinggi usia seseorang, maka peluang untuk terkena penyakit gula juga semakin meningkat. Selanjutnya, prediksi peluang seseorang untuk tidak terkena penyakit gula, ketika berusia 60 tahun adalah

$$1 - 0,797039037 = 0,202960963.$$

Perhatikan bahwa

$$\frac{0,797039037}{1-0,797039037} = \frac{0,797039037}{0,202960963} = 3,92 \cong 4.$$

Nilai tersebut dapat diartikan, ketika seseorang berusia 60 tahun, diprediksi terjadinya penyakit gula 4 kali lebih mungkin, dibandingkan tidak terkena penyakit gula. Gambar 13.6 dan Gambar 13.7 menyajikan hasil perhitungan prediksi peluang terjadinya penyakit gula, dari 100 responden, dengan menggunakan R. Berdasarkan Gambar 13.7, diketahui prediksi peluang responden ke-1 terkena penyakit gula sebesar 0,04150107, prediksi peluang responden ke-2 terkena penyakit gula sebesar 0,04622284, prediksi peluang responden ke-3 terkena penyakit gula sebesar 0,05723942, dan seterusnya. Gambar 13.8 menyajikan hasil perhitungan prediksi peluang terjadinya penyakit gula, dari 100 responden, dengan menggunakan SPSS.



Gambar 13.6

```
probabilitas=predict(regresi_logistik, type="response")
probabilitas
```

```
##          1          2          3          4          5          6
## 0.04150107 0.04622284 0.05723942 0.07068705 0.07068705 0.08700246
##          7          8          9         10         11         12
## 0.07845763 0.10665149 0.09638058 0.10665149 0.11787416 0.11787416
##         13         14         15         16         17         18
## 0.11787416 0.11787416 0.13010577 0.13010577 0.14340030 0.14340030
##         19         20         21         22         23         24
## 0.15780686 0.17336784 0.17336784 0.17336784 0.17336784 0.17336784
##         25         26         27         28         29         30
## 0.17336784 0.19011686 0.19011686 0.20807669 0.20807669 0.20807669
##         31         32         33         34         35         36
## 0.22725707 0.22725707 0.22725707 0.24765257 0.24765257 0.26924073
##         37         38         39         40         41         42
## 0.26924073 0.29198043 0.29198043 0.31581068 0.31581068 0.31581068
##         43         44         45         46         47         48
## 0.34065009 0.34065009 0.36639696 0.36639696 0.36639696 0.36639696
##         49         50         51         52         53         54
## 0.39293012 0.39293012 0.39293012 0.39293012 0.42011072 0.42011072
##         55         56         57         58         59         60
## 0.44778465 0.44778465 0.47578584 0.47578584 0.47578584 0.50394011
##         61         62         63         64         65         66
## 0.50394011 0.50394011 0.53206941 0.53206941 0.53206941 0.55999634
##         67         68         69         70         71         72
## 0.55999634 0.58754856 0.61456298 0.61456298 0.64088948 0.64088948
##         73         74         75         76         77         78
## 0.66639391 0.69096044 0.69096044 0.69096044 0.71449296 0.71449296
```

Gambar 13.7

	Penyakitgula	Usia	PRE_1
1	Tidak	20	0.04150
2	Tidak	21	0.04622
3	Tidak	23	0.05724
4	Tidak	25	0.07069
5	Tidak	25	0.07069
6	Tidak	27	0.08700
7	Ya	26	0.07846
8	Tidak	29	0.10665
9	Tidak	28	0.09638
10	Tidak	29	0.10665
11	Tidak	30	0.11787
12	Tidak	30	0.11787
13	Tidak	30	0.11787
14	Tidak	30	0.11787
15	Tidak	31	0.13011
16	Ya	31	0.13011
17	Tidak	32	0.14340
18	Tidak	32	0.14340
19	Tidak	33	0.15781

	Penyakitgula	Usia	PRE_1
28	Tidak	36	0.20808
29	Ya	36	0.20808
30	Tidak	36	0.20808
31	Tidak	37	0.22726
32	Ya	37	0.22726
33	Tidak	37	0.22726
34	Tidak	38	0.24765
35	Tidak	38	0.24765
36	Tidak	39	0.26924
37	Ya	39	0.26924
38	Tidak	40	0.29198
39	Ya	40	0.29198
40	Tidak	41	0.31581
41	Tidak	41	0.31581
42	Tidak	41	0.31581
43	Tidak	42	0.34065
44	Tidak	42	0.34065
45	Ya	43	0.36640
46	Tidak	43	0.36640
47	Tidak	43	0.36640
48	Ya	43	0.36640
49	Tidak	44	0.39293
50	Tidak	44	0.39293
51	Ya	44	0.39293

82	Ya	57	0.73692
83	Ya	57	0.73692
84	Ya	57	0.73692
85	Ya	57	0.73692
86	Tidak	58	0.75817
87	Ya	58	0.75817
88	Ya	58	0.75817
89	Ya	59	0.77823
90	Ya	59	0.77823
91	Tidak	60	0.79707
92	Ya	60	0.79707
93	Ya	61	0.81469
94	Ya	62	0.83110
95	Ya	62	0.83110
96	Ya	63	0.84634
97	Tidak	64	0.86043
98	Ya	64	0.86043
99	Ya	65	0.87342
100	Ya	69	0.91547
101			

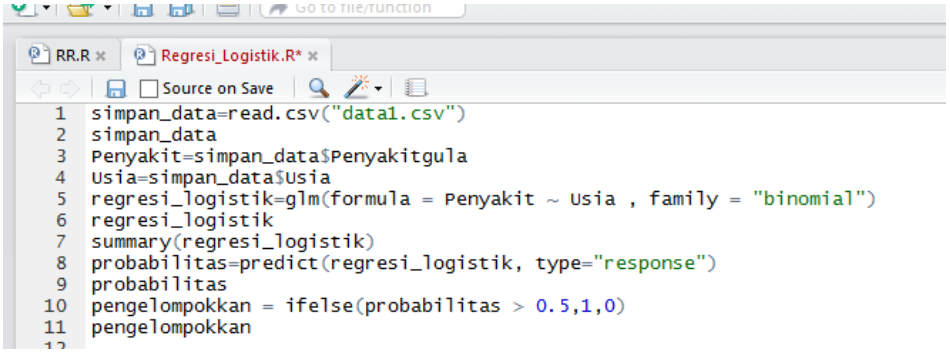
Gambar 13.8

Mengestimasi atau Memprediksi Keanggotaan Responden dalam Kelompok (Predicted Group)

Pada pembahasan sebelumnya, telah dihitung nilai prediksi peluang terjadinya penyakit gula untuk tiap-tiap responden. Berdasarkan nilai prediksi peluang tersebut, dapat diprediksi apakah responden tersebut masuk ke dalam kelompok terkena penyakit gula "Ya" atau tidak terkena penyakit gula "Tidak". Apabila nilai prediksi peluang responden > 0,5, maka responden tersebut diprediksi masuk ke dalam kelompok terkena penyakit gula "Ya". Sementara apabila nilai prediksi peluang responden < 0,5, maka responden tersebut diprediksi masuk ke dalam kelompok tidak terkena penyakit gula "Tidak".

Berdasarkan Gambar 13.7, diketahui prediksi peluang responden ke-1 terkena penyakit gula sebesar 0,04150107, yakni < 0,5, maka responden ke-1 diprediksi masuk ke dalam kelompok tidak terkena penyakit gula "Tidak". Diketahui **pada keadaan sebenarnya, responden ke-1 memang tidak terkena penyakit gula** (tidak terjadi kesalahan klasifikasi atau pengelompokkan). Diketahui prediksi peluang responden ke-2 terkena penyakit gula sebesar 0,04622284, yakni < 0,5, maka responden ke-2 diprediksi masuk ke dalam kelompok tidak terkena penyakit gula "Tidak". Diketahui **pada keadaan sebenarnya, responden ke-1 memang tidak terkena penyakit gula** (tidak terjadi kesalahan klasifikasi). Diketahui prediksi peluang responden ke-7 terkena penyakit gula sebesar 0,07845763, yakni < 0,5, maka responden ke-7 diprediksi masuk ke dalam kelompok tidak terkena penyakit gula "Tidak". Diketahui **pada keadaan sebenarnya, responden ke-7 terkena penyakit gula** (**terjadi kesalahan klasifikasi**).

Gambar 13.9 dan Gambar 13.10 menyajikan hasil prediksi pengelompokkan responden dengan R. Sementara pada Gambar 13.11 menyajikan hasil prediksi pengelompokkan responden dengan SPSS. Berdasarkan Gambar 13.10, responden ke-1 diprediksi masuk ke dalam kelompok tidak terkena penyakit gula (diberi angka 0), responden ke-60 diprediksi masuk ke dalam kelompok terkena penyakit gula (diberi angka 1), dan seterusnya.

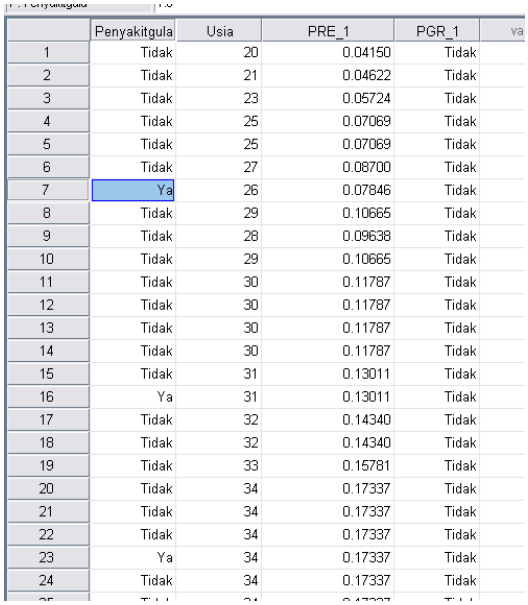


Gambar 13.9

```
pengelompokkan = ifelse(probabilitas > 0.5,1,0)
pengelompokkan
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##   0   0   0   0   0   1   1   1   1   1   1   1   1   1   1   1   1   1
##  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  91  92  93  94  95  96  97  98  99 100
##   1   1   1   1   1   1   1   1   1   1
```

Gambar 13.10



	Penyakitgula	Usia	PRE_1	PGR_1	ya
1	Tidak	20	0.04150	Tidak	
2	Tidak	21	0.04622	Tidak	
3	Tidak	23	0.05724	Tidak	
4	Tidak	25	0.07069	Tidak	
5	Tidak	25	0.07069	Tidak	
6	Tidak	27	0.08700	Tidak	
7	Ya	26	0.07846	Tidak	
8	Tidak	29	0.10665	Tidak	
9	Tidak	28	0.09638	Tidak	
10	Tidak	29	0.10665	Tidak	
11	Tidak	30	0.11787	Tidak	
12	Tidak	30	0.11787	Tidak	
13	Tidak	30	0.11787	Tidak	
14	Tidak	30	0.11787	Tidak	
15	Tidak	31	0.13011	Tidak	
16	Ya	31	0.13011	Tidak	
17	Tidak	32	0.14340	Tidak	
18	Tidak	32	0.14340	Tidak	
19	Tidak	33	0.15781	Tidak	
20	Tidak	34	0.17337	Tidak	
21	Tidak	34	0.17337	Tidak	
22	Tidak	34	0.17337	Tidak	
23	Ya	34	0.17337	Tidak	
24	Tidak	34	0.17337	Tidak	

Gambar 13.11

Menghitung Tingkat Keakuratan Model Regresi Logistik dalam Memprediksi Pengelompokkan

Pada pembahasan sebelumnya, berdasarkan nilai prediksi peluang dari responden, dapat diprediksi responden tersebut masuk ke dalam kelompok tidak terkena penyakit gula "Tidak" atau terkena penyakit gula "Ya". Dalam proses pengelompokkan tersebut, bisa saja terjadi kesalahan pengelompokkan. Sebagai contoh, responden ke-7 diprediksi masuk ke dalam kelompok tidak terkena penyakit gula "Tidak". Diketahui **pada keadaan sebenarnya, responden ke-7 terkena penyakit gula** (**terjadi kesalahan klasifikasi**).

Gambar 13.12 dan Gambar 13.13 menyajikan hasil prediksi pengelompokkan responden dengan R. Berdasarkan Gambar 4.13, terdapat 67 responden **yang tidak terkena penyakit gula**. Kemudian dari 67 responden tersebut, **diprediksi 45 responden masuk ke dalam kelompok tidak terkena penyakit gula "Tidak", dan 12 responden masuk ke dalam kelompok terkena penyakit gula "Ya"**. Dalam hal ini terjadi 12 **kesalahan pengelompokkan**. Kemudian berdasarkan Gambar 4.13, terdapat 33 responden **yang terkena penyakit gula**. Kemudian dari 33 responden tersebut, **diprediksi 29 responden masuk ke dalam kelompok terkena penyakit gula "Ya", dan 14 responden masuk ke dalam kelompok tidak terkena penyakit gula "Tidak"**. Dalam hal ini terjadi 14 **kesalahan pengelompokkan**.

Sehingga persentase ketepatan model dapat memprediksi dengan benar (berdasarkan data 100 responden)

$$\frac{45+29}{45+12+14+29}=\frac{74}{100}=74\%.$$

Gambar 4.14 menyajikan hasil prediksi pengelompokkan responden dengan SPSS.

```
RR.R  Regresi_Logistik.R*
Source on Save
1  simpan_data=read.csv("data1.csv")
2  simpan_data
3  Penyakit=simpan_data$Penyakitgula
4  Usia=simpan_data$Usia
5  regresi_logistik=glm(formula = Penyakit ~ Usia , family = "binomial")
6  regresi_logistik
7  summary(regresi_logistik)
8  probabilitas=predict(regresi_logistik, type="response")
9  probabilitas
10 pengelompokkan = ifelse(probabilitas > 0.5,"Ya","Tidak")
11 pengelompokkan
12 table(Penyakit, pengelompokkan)
```

Gambar 13.12

```
table(Penyakit, pengelompokkan)
```

```
##          pengelompokkan
## Penyakit Tidak Ya
##    Tidak    45 12
##    Ya       14 29
```

Gambar 13.13

Classification Table[a]

Observed			Predicted		
			Penyakitgula		
			Tidak	Ya	Percentage Correct
Step 1	Penyakitgula	Tidak	45	12	78.9
		Ya	14	29	67.4
	Overall Percentage				74.0

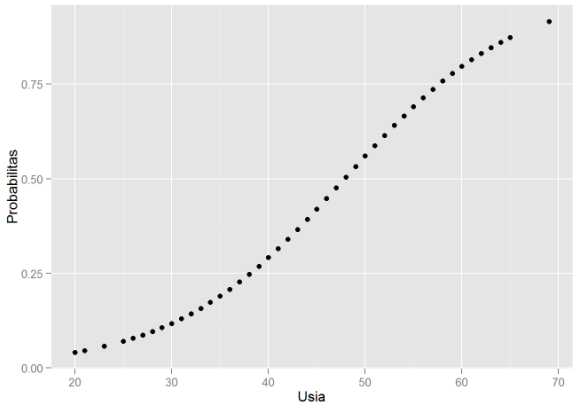
a. The cut value is .500

Gambar 13.14

Grafik Usia v/s Nilai Prediksi Probabilitas

Berikut disajikan grafik antara usia (sumbu horizontal) dan nilai prediksi probabilitas (sumbu vertikal) (nonlinear).

```
library(ggplot2)
ggplot(simpan_data, aes(Usia, Probabilitas)) + geom_point()
```



Gambar 13.15

Referensi

1. Agresti, A. dan B. Finlay. 2009. *Statistical Methods for the Social Sciences, 4th Edition*. United States of America: Prentice Hall.

2. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

3. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

4. Gujarati, D.N. 2003. *Basic Econometrics, 4th Edition*. New York: McGraw-Hill.

5. Hosmer, D.W. dan S. Lemeshow. 2000. *Applied Logistic Regression, 2nd Edition*. United States of America: John Wiley & Sons, Inc.

6. Hair, J. F Jr., R.E. Anderson, B.J. Babin, dan W.C. Black. 2010. *Multivariate Data Analysis, 7th Edition*. Pearson Prentice Hall.

7. Kleinbaum, D.G. dan M. Klein. 2010. *Logistic Regression, 3rd Edition*. New York: Springer.

8. Meyers, L.S., G. Gamst, dan A.J. Guarino. 2005. *Applied Multivariate Research, Design and Interpretation*. Sage.

9. Stevens, J.P. 2009. *Applied Multivariate Statistics For The Social Science, 5th Edition*. New York: Routledge.

10. Supranto, J. 2004. Ekonometri, Buku Kedua. Jakarta: Ghalia Indonesia.

11. http://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

12. http://ww2.coastal.edu/kingw/statistics/R-tutorials/logistic.html

13. http://www.statmethods.net/advstats/glm.html

BAB 14

ANALISIS KLASTER

Sekilas Analisis Klaster

Berikut pemaparan singkat mengenai analisis klaster menurut Hair dkk. (2010:477).

"*Cluster analysis groups individuals or objects into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters. The attempt is to maximize the homogeneity of objects within the clusters while also maximizing the heterogeneity between clusters.*"

Malhotra dan Birks (2006:597) menyatakan sebagai berikut.

"*Cluster analysis is a class of techniques used to classify objects or cases into relatively homogeneous groups called clusters. Objects in each cluster tend to be similar to each other and dissimilar to objects in the other clusters. Cluster analysis is also called classification analysis or numerical taxonomy[3]. Both cluster analysis and discriminant analysis are concerned with classification. Discriminant analysis, however, requires prior knowledge of the cluster or group membership for each object or case included, to develop the classification rule. In contrast, in cluster analysis there is no a priori information about the group or cluster membership for any of the objects. Groups or clusters are suggested by the data, not defined a priori[5].*

Janssens dkk. (2008:317) menyatakan sebagai berikut.

"*The objective of cluster analysis is to take a sample of n individuals or objects, each of which is measured for p variables, and group it into g classes, where g is less than n. In other words, the goal is to sort cases (individuals, products, brands, stimuli) into groups so that a high degree of similarity exists between cases in the same group, and a low degree of similarity between cases belonging to different groups. This similarity is evaluated on the basis of the value of each case (individual, product, etc.) for the variables (characteristics, attributes) upon which the cluster analysis is performed.*"

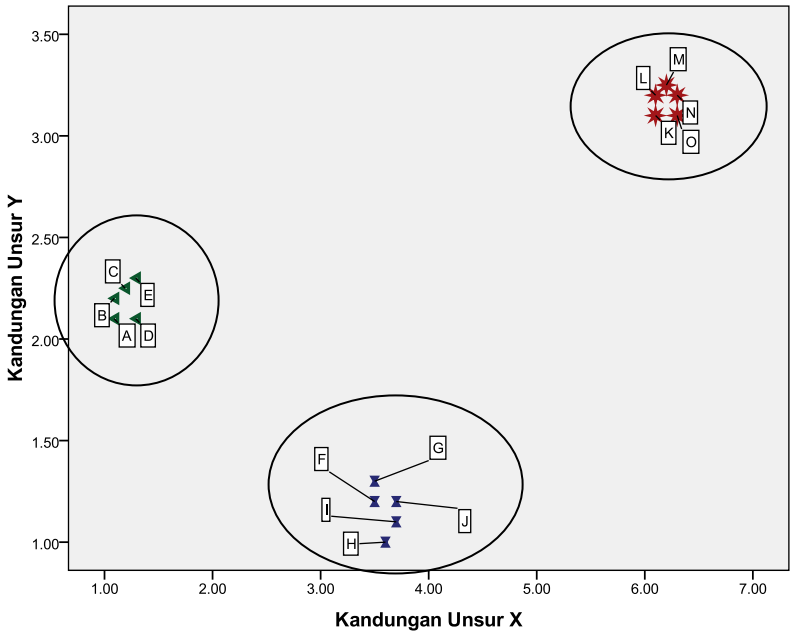
Berdasarkan uraian di atas, analisis klaster (*cluster analysis*) merupakan suatu teknik statistika yang digunakan untuk mengelompokkan (*group*) sekumpulan objek (manusia, produk, tanaman, dan sebagainya) ke dalam beberapa klaster. Perhatikan bahwa suatu objek hanya bisa masuk atau tergabung dalam satu klaster. Beberapa objek yang berada dalam satu klaster cenderung saling mirip, namun cenderung berbeda terhadap objek-objek yang berada dalam klaster lainnya. Sebagai contoh perhatikan data pada Tabel 14.1. Berdasarkan data pada Tabel 14.1, objek yang diteliti adalah batu, sebanyak 15 batu. Masing-masing batu memiliki kadar X dan kadar Y. Gambar 14.1 memberikan gambaran yang cukup jelas untuk pengelompokkan (*cluster*). Berdasarkan Gambar 14.2, jika dibentuk klaster sebanyak 3, maka:

- ⇨ Batu A, B, C, D, dan E berada dalam satu klaster, misalkan klaster pertama.
- ⇨ Batu F, G, H, I, dan J berada dalam satu klaster, misalkan klaster kedua.
- ⇨ Batu K, L, M, N, dan O berada dalam satu klaster, misalkan klaster ketiga.

Perhatikan bahwa batu A, B, C, D, dan E cenderung mirip, karena berada di dalam satu klaster, yakni klaster pertama, namun cenderung berbeda terhadap batu-batu yang berada dalam klaster yang berbeda. Tiga klaster yang tersaji dalam Gambar 14.1 melibatkan dua variabel klaster, yakni variabel **kadar X** (sumbu horizontal) dan **kadar Y** (sumbu vertikal).

Tabel 14.1

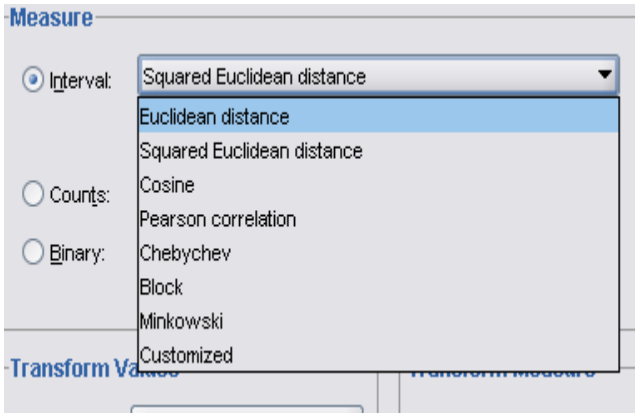
Batu	kadar X	kadar Y
A	1.1	2.1
B	1.1	2.2
C	1.2	2.25
D	1.3	2.1
E	1.3	2.3
F	3.5	1.2
G	3.5	1.3
H	3.6	1
I	3.7	1.1
J	3.7	1.2
K	6.1	3.1
L	6.1	3.2
M	6.2	3.25
N	6.3	3.2
O	6.3	3.1



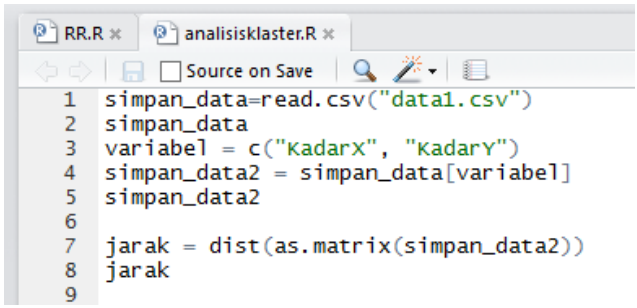
Gambar 14.1

Ukuran Kemiripan (Measure of Similarity)

Gambar 14.2 menyajikan beberapa ukuran kemiripan, yakni di antaranya adalah *Euclidean distance* dan *Squared Euclidean distance*. Gambar 14.5 menyajikan *Euclidean distance* (jarak *Euclidean*) untuk tiap-tiap pasang objek (batu). Berdasarkan Gambar 14.1, suatu objek akan semakin mirip dengan objek yang lain, jika posisinya semakin berdekatan. Dengan kata lain, jarak di antara objek tersebut semakin kecil (nilai *Euclidean distance* semakin kecil).



Gambar 14.2 Berbagai Ukuran Kemiripan



Gambar 14.3

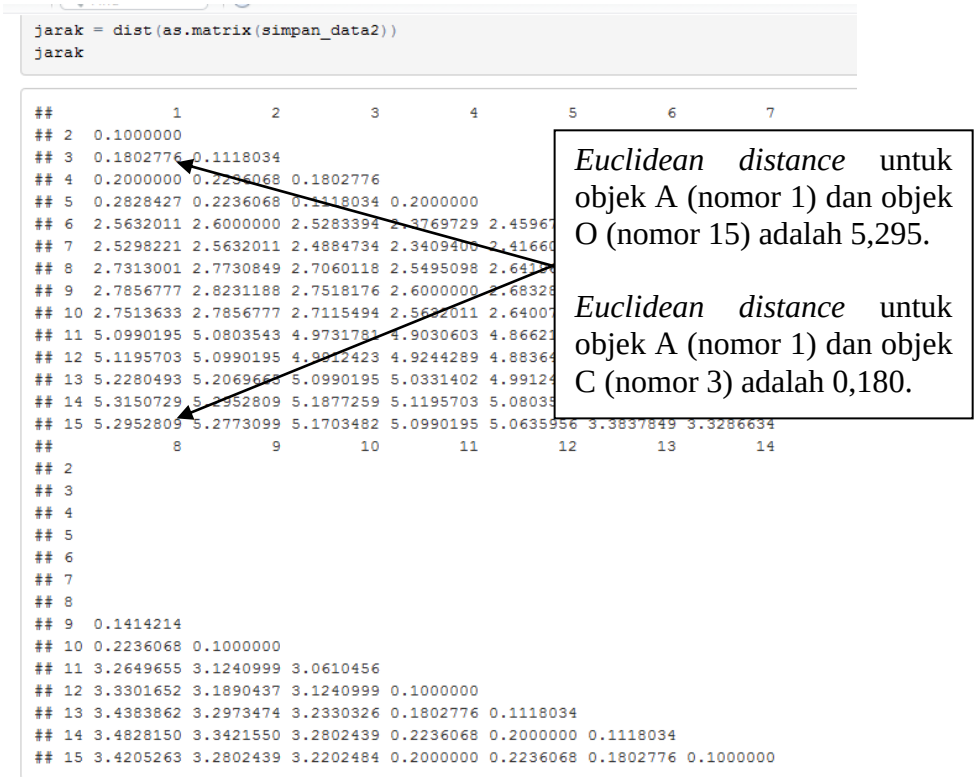
```
simpan_data=read.csv("data1.csv")
simpan_data
```

```
##    Batu KadarX KadarY
## 1     A    1.1   2.10
## 2     B    1.1   2.20
## 3     C    1.2   2.25
## 4     D    1.3   2.10
## 5     E    1.3   2.30
## 6     F    3.5   1.20
## 7     G    3.5   1.30
## 8     H    3.6   1.00
## 9     I    3.7   1.10
## 10    J    3.7   1.20
## 11    K    6.1   3.10
## 12    L    6.1   3.20
## 13    M    6.2   3.25
## 14    N    6.3   3.20
## 15    O    6.3   3.10
```

```
variabel = c("KadarX", "KadarY")
simpan_data2 = simpan_data[variabel]
simpan_data2
```

```
##    KadarX KadarY
## 1     1.1   2.10
## 2     1.1   2.20
## 3     1.2   2.25
## 4     1.3   2.10
## 5     1.3   2.30
## 6     3.5   1.20
## 7     3.5   1.30
## 8     3.6   1.00
## 9     3.7   1.10
## 10    3.7   1.20
## 11    6.1   3.10
## 12    6.1   3.20
## 13    6.2   3.25
## 14    6.3   3.20
## 15    6.3   3.10
```

Gambar 14.4



Gambar 14.5 *Euclidean Distance* untuk Tiap-Tiap Pasang Objek (Batu)

Berdasarkan Gambar 14.5, diketahui *Euclidean distance* untuk objek A (nomor 1) dan objek C (nomor 3) adalah 0,180. Nilai tersebut dihitung sebagai berikut.

$$\sqrt{(1{,}2-1{,}1)^2+(2{,}25-2{,}1)^2} = 0{,}180277\ atau\ dibulatkan\ 0{,}180.$$

Diketahui *Euclidean distance* untuk objek A (nomor 1) dan objek O (nomor 15) adalah 5,295. Nilai tersebut dihitung sebagai berikut.

$$\sqrt{(6{,}3-1{,}1)^2+(3{,}1-2{,}1)^2} = 5{,}295280\ atau\ dibulatkan\ 5{,}295.$$

Gambar 14.7 menyajikan *Squared Euclidean distance* (jarak *Euclidean* yang dikuadratkan) untuk tiap-tiap pasang objek (batu). Berdasarkan Gambar 14.7, diketahui *Squared Euclidean distance* untuk objek A dan objek C adalah 0,032. Nilai tersebut dihitung sebagai berikut.

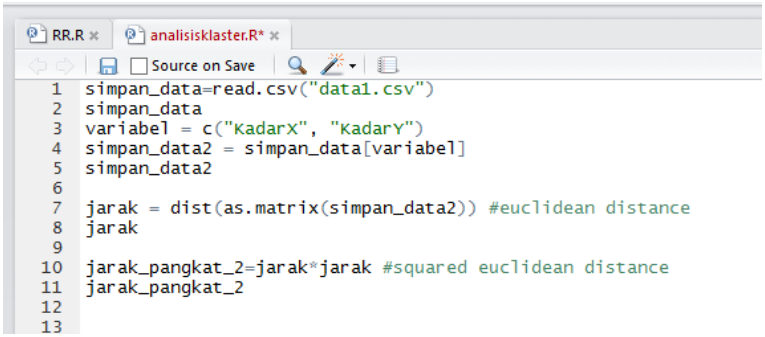
$$(1{,}2-1{,}1)^2+(2{,}25-2{,}1)^2 = 0{,}0325.$$

Diketahui *Squared Euclidean distance* untuk objek A dan objek O adalah 28,040. Nilai tersebut dihitung sebagai berikut.

$$(6{,}3-1{,}1)^2+(3{,}1-2{,}1)^2 = 28{,}04.$$

Diketahui *Squared Euclidean distance* untuk objek C dan objek D adalah 0,0325. Nilai tersebut dihitung sebagai berikut.

$$(1{,}3-1{,}2)^2+(2{,}1-2{,}25)^2 = 0{,}0325.$$



Gambar 14.6

```
jarak_pangkat_2=jarak*jarak
jarak_pangkat_2
```

```
##          1       2       3       4       5       6       7       8       9
## 2   0.0100
## 3   0.0325  0.0125
## 4   0.0400  0.0500  0.0325
## 5   0.0800  0.0500  0.0125  0.0400
## 6   6.5700  6.7600  6.3925  5.6500  6.0500
## 7   6.4000  6.5700  6.1925  5.4800  5.8400  0.0100
## 8   7.4600  7.6900  7.3225  6.5000  6.9800  0.0500  0.1000
## 9   7.7600  7.9700  7.5725  6.7600  7.2000  0.0500  0.0800  0.0200
## 10  7.5700  7.7600  7.3525  6.5700  6.9700  0.0400  0.0500  0.0500  0.0100
## 11 26.0000 25.8100 24.7325 24.0400 23.6800 10.3700 10.0000 10.6600  9.7600
## 12 26.2100 26.0000 24.9125 24.2500 23.8500 10.7600 10.3700 11.0900 10.1700
## 13 27.3325 27.1125 26.0000 25.3325 24.9125 11.4925 11.0925 11.8225 10.8725
## 14 28.2500 28.0400 26.9125 26.2100 25.8100 11.8400 11.4500 12.1300 11.1700
## 15 28.0400 27.8500 26.7325 26.0000 25.6400 11.4500 11.0800 11.7000 10.7600
##         10      11      12      13      14
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11  9.3700
## 12  9.7600  0.0100
## 13 10.4525  0.0325  0.0125
## 14 10.7600  0.0500  0.0400  0.0125
## 15 10.3700  0.0400  0.0500  0.0325  0.0100
```

Gambar 14.7 *Squared Euclidean Distance* untuk Tiap-Tiap Pasang Objek (Batu)

Malhotra dan Birks (2006:600) menyatakan sebagai berikut

"*Because the objective of clustering is to group similar objects together, some measure is needed to assess how similar or different the objects are. The most common approach is to measure similarity in terms of distance between pairs of objects. Objects with smaller distances between them are more similar to each other than are those at larger distances. There are several ways to compute the distance between two objects*[9]. ***The most commonly used measure of similarity is the euclidean distance or its square***[10]*. The euclidean distance is the square root of the sum of the squared differences in values for each variable. Other distance measures are also available.* ***The city-block or Manhattan distance between two objects is the sum of the absolute differences in values for each variable***. ***The Chebychev distance between two objects is the maximum absolute difference in values for any variable***. *For our example, we use the squared euclidean distance.*"

Berdasarkan uraian tersebut, secara umum, ukuran kemiripan yang umum digunakan adalah *Euclidean distance* atau *Squared Euclidean distance*. Lebih lanjut Malhotra dan Birks (2006:600) dan Hair dkk. (2010:496-497) menganjurkan untuk melakukan standarisasi data (data ditransormasi ke dalam bentuk normal, dengan rata-rata 0, dan standar deviasi 1) untuk tiap-tiap variabel klaster, apabila data pada variabel-variabel kluster memiliki satuan yang berbeda-beda. Di sisi lain, data yang termasuk *outlier* juga dianjurkan untuk dihapus (Malhotra dan Birks, 2006:601).

Selanjutnya Malhotra dan Birks (2006:601) menyatakan penggunaan ukuran kemiripan (*measure of similarity*) yang berbeda-beda, dapat mempengaruhi hasil klaster, sehingga disarankan untuk menggunakan berbagai ukuran kemiripan dan hasil klaster tersebut diperbandingkan.

Sejalan dengan Malhotra dan Birks, Hair dkk. (2010:496) menyatakan sebagai berikut.

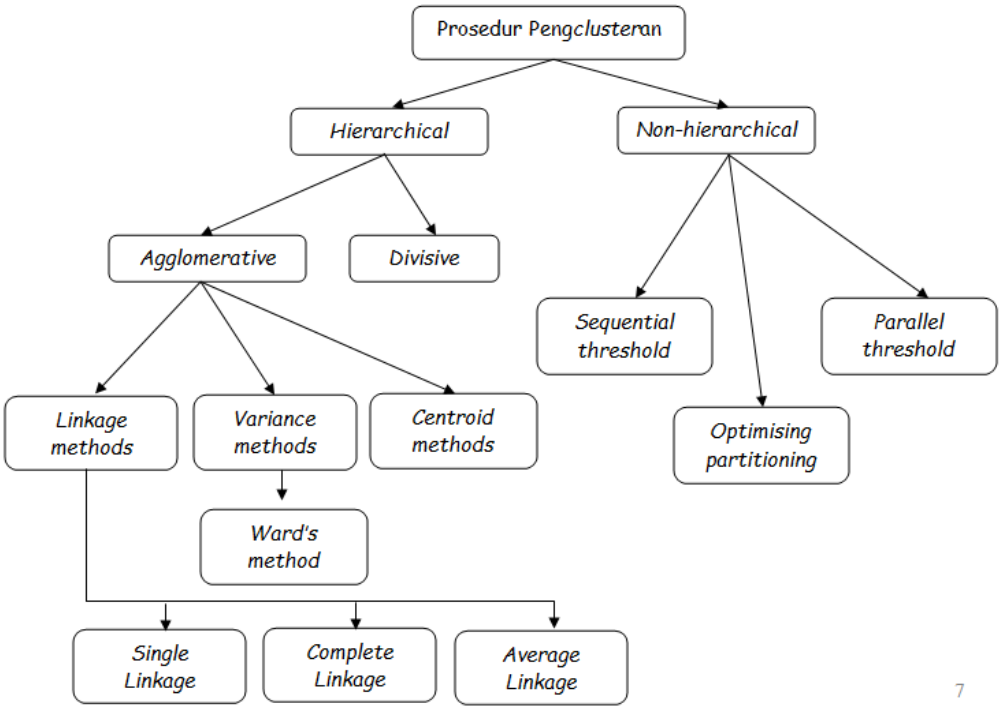
"***Which Distance Measures is Best?*** *In attempting to select a particular distance measure, the researcher should remember the following caveats: Difference distance measures or a change in the scales of the variables may lead to different cluster solutions. Thus, it is advisable to use several measures and compare the results with theoretical or know patterns. When the variables are correlated (either positively or negatively) the Mahalanobis distance measure is likely to be the most appropriate because it adjusts for correlations and weights all variable equally. Alternatively, the researcher may wish to avoid using highly redundant variables as input to cluster analysis.*"

Prosedur Pengklasteran

Gambar 14.8 menyajikan prosedur pengklasteran dalam analisis klaster (Malhotra dan Birks, 2006:601). Berdasarkan Gambar 14.8, prosedur pengklasteran dapat menggunakan metode *hierarchical* atau metode *non-hierarchical*. Pada metode *hierarchical*, jumlah klaster belum atau tidak diketahui sebelumnya, sementara pada metode *non-hierarchical* jumlah klaster ditetapkan terlebih dahulu, sebelum melakukan pengklasteran objek. Dengan kata lain, pada metode *non-hierarchical*, tahap awal ialah menentukan jumlah klaster yang diinginkan, kemudian tiap-tiap objek pengamatan digabungkan ke dalam salah satu klaster yang telah ditetapkan.

Selanjutnya, dalam metode *hierarchical* terdiri dari dua metode, yakni metode *agglomerative* dan metode *divisive*. Metode *agglomerative* dimulai dengan menganggap tiap-tiap objek sebagai klaster-klaster yang berbeda atau terpisah. Kemudian dua klaster atau objek paling dekat digabung menjadi satu klaster. Proses ini terus berlanjut, sampai seluruh objek bergabung menjadi satu klaster. Sementara pada metode *divisive* merupakan kebalikan dari metode *agglomerative*, yakni dimulai dengan menganggap tiap-tiap objek berasal dalam satu klaster, kemudian dipecah atau dipisahkan sampai setiap objek berada dalam klaster-klaster yang terpisah (Malhotra dan Birks, 2006:601).

Metode *agglomerative* terdiri dari 3 metode, yakni metode *linkage, variance,* dan *centroid*. Metode *linkage* terdiri dari metode *single linkage, complete linkage*, dan *average linkage*, sementara pada metode *variance* terdiri dari metode *ward*. Pada metode *non-hierarchical* terdiri dari metode *sequential threshold, optimising partitioning,* dan *parallel threshold*. Metode *non-hierarchical* sering disebut dengan istilah *k-means clustering*.



Gambar 14.8 Prosedur Pengklasteran (Malhotra dan Birks, 2006:601)

Analisis Klaster dengan Metode Average Linkage

Berikut diberikan contoh penggunaan analisis klaster metode *average linkage*. Diberikan data seperti pada Gambar 14.9. Data pada Gambar 14.9 disajikan dalam grafik seperti pada Gambar 14.10. Berikut akan digunakan analisis klaster metode *average linkage* untuk pengklasteran. Gambar 14.11 menyajikan *Squared Euclidean distance* (matriks jarak/*distance matrix*).

	Batu	A	B
1	A	1.10	1.10
2	B	1.20	.85
3	C	1.30	.97
4	D	1.40	.90
5	E	1.35	1.00
6	F	1.20	1.00

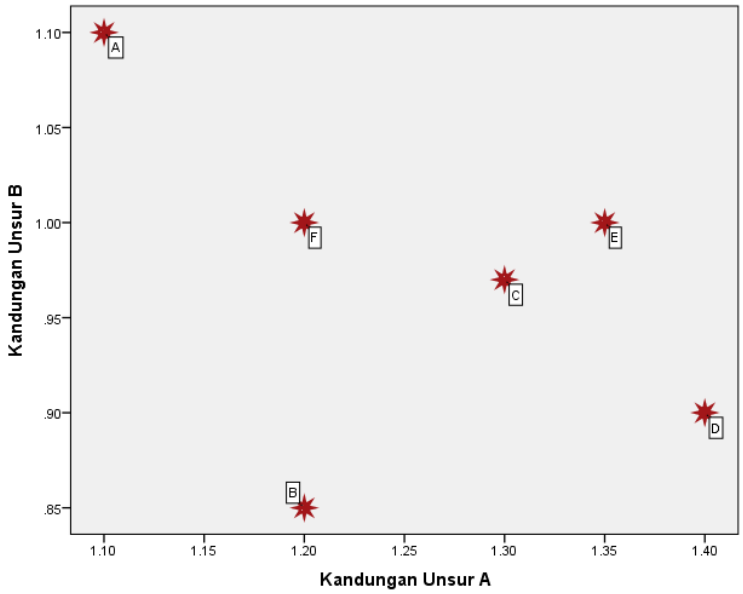
Gambar 14.9

Berdasarkan Gambar 14.11, diketahui *Squared Euclidean distance* untuk objek A dan objek C adalah 0,057. Nilai tersebut dihitung sebagai berikut.

$$(1{,}1 - 0{,}97)^2 + (1{,}1 - 1{,}3)^2 = 0{,}0569 \; atau \; dibulatkan \; 0{,}057.$$

Diketahui *Squared Euclidean distance* untuk objek B dan objek F adalah 0,023. Nilai tersebut dihitung sebagai berikut.

$$(0{,}85 - 1)^2 + (1{,}2 - 1{,}2)^2 = 0{,}0225 \; atau \; dibulatkan \; 0{,}023.$$



Gambar 14.10

Proximity Matrix

Case	Squared Euclidean Distance					
	1:A	2:B	3:C	4:D	5:E	6:F
1:A	.000	.073	.057	.130	.073	.020
2:B	.073	.000	.024	.042	.045	.023
3:C	.057	.024	.000	.015	.003	.011
4:D	.130	.042	.015	.000	.012	.050
5:E	.073	.045	.003	.012	.000	.023
6:F	.020	.023	.011	.050	.023	.000

This is a dissimilarity matrix

Gambar 14.11 *Squared Euclidean Distance* (Matriks Jarak)

Berdasarkan Gambar 14.11, diketahui nilai *Squared Euclidean distance* **paling kecil** berada pada pasangan objek C dan objek E (pasangan objek yang berbeda), yakni bernilai 0,003 (Perhatikan Gambar 14.11). **Maka objek C dan objek E bergabung menjadi *cluster* (C,E)**. Pada Gambar 14.12, terlihat bahwa pada *Stage 1*, objek C (3) dan objek E (5) bergabung menjadi *cluster* (C,E). Perhatikan juga bahwa **nilai *coefficient* 0,003**, yang merupakan jarak antara objek C dan objek E.

Average Linkage (Between Groups)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	5	.003	0	0	2
2	3	4	.014	1	0	4
3	1	6	.020	0	0	5
4	2	3	.037	0	2	5
5	1	2	.055	3	4	0

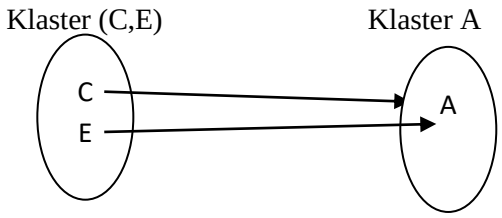
Gambar 14.12 *Output* SPSS untuk Analisis Klaster Metode *Average Linkage*

Selanjutnya menghitung jarak antara *cluster* (C,E) terhadap objek lainnya.

⇨ Menghitung jarak antara *cluster* (C,E) terhadap objek A.

$$d_{(C,E)A} = \frac{d_{(C,A)} + d_{(E,A)}}{N_{(C,E)} \times N_{(A)}} = \frac{0{,}0569 + 0{,}0725}{2 \times 1} = 0{,}0647.$$

Perhatikan bahwa $N_{(C,E)}$ dan N_A masing-masing menyatakan jumlah objek dalam klaster (C,E) dan A.



Menentukan jarak antara klaster (C,E) dan klaster A adalah hitung jarak dar C ke A, dan jarak dari E ke A. Kemudian jumlahkan dan bagi 2. 2 dalam hal ini $2 \times 1 = 2$. 2 menyatakan jumlah anggota klaster (C,E) dan 1 menyatakan jumlah anggota klaster A.

⇨ Menghitung jarak antara *cluster* (C,E) terhadap objek B.

$$d_{(C,E)B} = \frac{d_{(C,B)} + d_{(E,B)}}{N_{(C,)} \times N_{(B)}} = \frac{0{,}0244 + 0{,}045}{2 \times 1} = 0{,}0347.$$

Gambar 14.13 menyajikan jarak antara *cluster* (C,E) terhadap masing-masing objek.

	A	B	C	D	E	F
A	0.000000	0.072500	0.056900	0.130000	0.072500	0.020000
B	0.072500	0.000000	0.024400	0.042500	0.045000	0.022500
C	0.056900	0.024400	0.000000	0.014900	0.003400	0.010900
D	0.130000	0.042500	0.014900	0.000000	0.012500	0.050000
E	0.072500	0.045000	0.003400	0.012500	0.000000	0.022500
F	0.020000	0.022500	0.010900	0.050000	0.022500	0.000000
Jarak	0.0647	0.0347		0.0137		0.0167

Gambar 14.13

Sehingga diperoleh matriks jarak yang baru seperti pada Gambar 14.14.

	C,E	A	B	D	F
C,E	0	0.0647	0.0347	0.0137	0.0167
A	0.0647	0	0.072500	0.130000	0.020000
B	0.0347	0.072500	0	0.042500	0.022500
D	0.0137	0.130000	0.042500	0	0.050000
F	0.0167	0.020000	0.022500	0.050000	0

Gambar 14.14 Matriks Jarak

Berdasarkan Gambar 14.14, diketahui nilai **jarak paling kecil** berada pada pasangan (C,E) dan D (pasangan objek yang berbeda), yakni bernilai 0,0137, **maka (C,E) dan D bergabung menjadi *cluster* (C,E,D)**. Pada Gambar 14.12, terlihat bahwa pada *Stage 2*, objek C (3) dan

objek D (4) bergabung. Perhatikan juga bahwa **nilai *coefficient* 0,014 (pembulatan dari 0,0137)**.

Selanjutnya menghitung jarak antara *cluster* (C,E,D) terhadap objek lainnya.

- ⇨ Menghitung jarak antara *cluster* (C,E,D) terhadap objek A.

$$d_{(C,E,D)A} = \frac{d_{(C,A)} + d_{(E,A)} + d_{(D,A)}}{N_{(C,E,D)} \times N_{(A)}} = \frac{\mathbf{0,0647} + \mathbf{0,0647} + 0{,}13}{3 \times 1} = 0{,}086467.$$

- ⇨ Menghitung jarak antara *cluster* (C,E,D) terhadap objek B.

$$d_{(C,E,D)B} = \frac{d_{(C,B)} + d_{(E,B)} + d_{(D,B)}}{N_{(C,E,D)} \times N_{(B)}} = \frac{\mathbf{0,0347} + \mathbf{0,0347} + 0{,}0425}{3 \times 1} = 0{,}0373.$$

- ⇨ Menghitung jarak antara *cluster* (C,E,D) terhadap objek F.

$$d_{(C,E,D)F} = \frac{d_{(C,F)} + d_{(E,F)} + d_{(D,F)}}{N_{(C,E,D)} \times N_{(B)}} = \frac{\mathbf{0,0167} + \mathbf{0,0167} + 0{,}05}{3 \times 1} = 0{,}0278.$$

Gambar 14.15 menyajikan jarak antara *cluster* (C,E,D) terhadap masing-masing objek.

	C,E	A	B	D	F
C,E	0	0.0647	0.0347	0.0137	0.0167
A	0.0647	0	0.0725	0.13	0.02
B	0.0347	0.0725	0	0.0425	0.0225
D	0.0137	0.13	0.0425	0	0.05
F	0.0167	0.02	0.0225	0.05	0
Jarak		0.086467	0.0373		0.0278

Gambar 14.15

Sehingga diperoleh matriks jarak yang baru seperti pada Gambar 14.16.

	C,E,D	A	B	F
C,E,D	0	0.086467	0.0373	0.0278
A	0.086467	0	0.0725	0.02
B	0.0373	0.0725	0	0.0225
F	0.0278	0.02	0.0225	0

Gambar 14.16

Berdasarkan Gambar 14.16, diketahui nilai **jarak paling kecil** berada pada pasangan objek A dan objek F, yakni bernilai 0,02, **maka objek A dan objek F bergabung menjadi *cluster* (A,F)**. Pada Gambar 14.12, terlihat bahwa pada *Stage 3*, objek 1 (A) dan objek 6 (F). Perhatikan juga bahwa **nilai *coefficient* 0,02**.

Selanjutnya menghitung jarak antara *cluster* (A,F) terhadap objek lainnya.

- ⇨ Menghitung jarak antara *cluster* (A,F) terhadap objek B.

$$d_{(A,F)B} = \frac{d_{(A,B)} + d_{(F,B)}}{N_{(A,F)} \times N_{(B)}} = \frac{0{,}0725 + 0{,}0225}{2 \times 1} = 0{,}0475.$$

⇨ Menghitung jarak antara *cluster* (A,F) terhadap *cluster* (C,E,D).

$$d_{(A,F)(C,E,D)} = \frac{d_{(A,C)} + d_{(A,E)} + d_{(A,D)} + d_{(F,C)} + d_{(F,E)} + d_{(F,D)}}{6}$$

$$d_{(A,F)(C,E,D)} = \frac{(\mathbf{3 \times 0{,}086467}) + (\mathbf{3 \times 0{,}0278})}{6} = 0{,}057133.$$

Gambar 14.17 menyajikan jarak antara *cluster* (A,F) terhadap masing-masing objek.

	C,E,D	A	B	F
C,E,D	0	0.086467	0.0373	0.0278
A	0.086467	0	0.0725	0.02
B	0.0373	0.0725	0	0.0225
F	0.0278	0.02	0.0225	0
Jarak	0.057133		0.0475	

Gambar 14.17

Sehingga diperoleh matriks jarak yang baru seperti pada Gambar 14.18.

	A,F	C,E,D	B
A,F	0	0.057133	0.0475
C,E,D	0.057133	0	0.0373
B	0.0475	0.0373	0

Gambar 14.18

Berdasarkan Gambar 14.18, diketahui nilai **jarak paling kecil** berada pada pasangan (C,E,D) dan B, yakni bernilai 0,0373, **maka (C,E,D) dan B bergabung menjadi *cluster* (C,E,D,B)**. Pada Gambar 14.12, terlihat bahwa pada *Stage 4*, objek 2 dan objek 3. Perhatikan juga bahwa **nilai *coefficient* 0,037**.

Selanjutnya menghitung jarak antara *cluster* (C,E,D,B) terhadap objek lainnya.

⇨ Menghitung jarak antara *cluster* (C,E,D,B) terhadap *cluster* (A,F).

$$d_{(C,E,D,B)(A,F)} = \frac{d_{(A,C)} + d_{(A,E)} + d_{(A,D)} + d_{(A,B)} + d_{(F,C)} + d_{(F,E)} + d_{(F,D)} + d_{(F,B)}}{N_{(C,E,D,B)} \times N_{(A,F)}}$$

$$d_{(C,E,D,B)(A,F)} = \frac{(6 \times 0{,}057133) + (2 \times 0{,}0475)}{8} = 0{,}054725.$$

Gambar 14.19 menyajikan jarak antara *cluster* (A,F) terhadap *cluster* (A,F).

	A,F	C,E,D	B
A,F	0	0.057133	0.0475
C,E,D	0.057133	0	0.0373
B	0.0475	0.0373	0
Jarak	0.054725		

Gambar 14.19

Sehingga diperoleh matriks jarak yang baru seperti pada Gambar 14.20.

	A,F	C,E,D,B
A,F	0	0.054725
C,E,D,B	0.054725	0

Gambar 14.20

Gambar 14.20 menyajikan jarak antara *cluster* (C,E,F,D,B) terhadap *cluster* (A,F). Diketahui jarak antara *cluster* (C,E,F,D,B) dan *cluster* (A,F)adalah 0,054725. **Pada Gambar 14.12**, yakni *Stage 5* (**objek 1 dan objek 2 bergabung**). Diketahui **nilai *coefficient*** adalah 0,054725. Berdasarkan hasil perhitungan diketahui:

- ⇨ Berdasarkan Gambar 14.20, jika dibentuk dua klaster, maka klaster-klaster tersebut adalah {A,F} dan {C,E,D,B}.

- ⇨ Berdasarkan Gambar 14.18, jika dibentuk tiga klaster, maka klaster-klaster tersebut adalah {A,F}, {C,E,D}, dan {B}.

Hasil tersebut sesuai dengan hasil SPSS seperti pada Gambar 14.21. Gambar 14.21 merupakan hasil berdasarkan SPSS untuk analisis klaster metode *average linkage*.

Cluster Membership

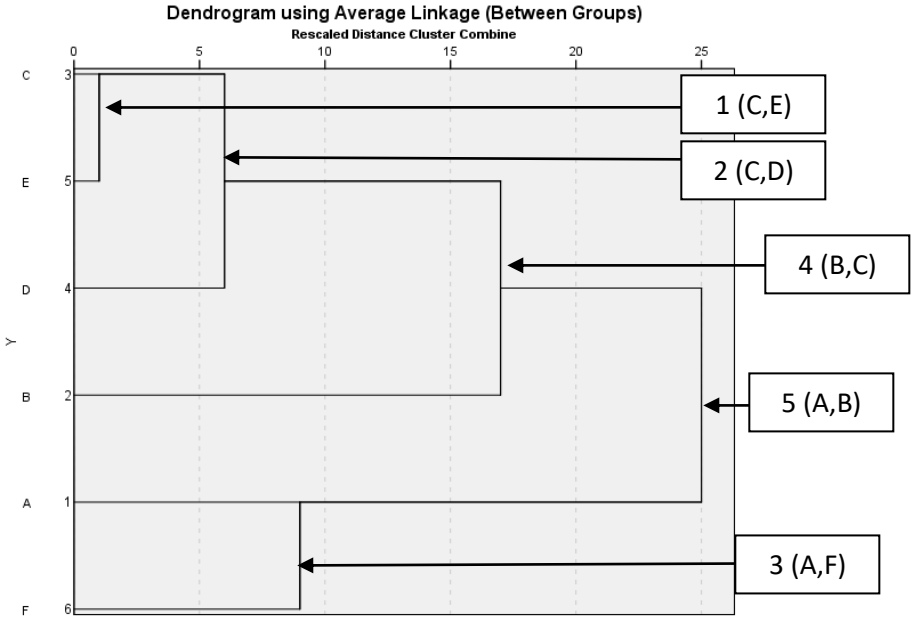
Case	3 Clusters	2 Clusters
1:A	1	1
2:B	2	2
3:C	3	2
4:D	3	2
5:E	3	2
6:F	1	1

Gambar 14.21

Jadi, pada metode *average linkage* memperlakukan jarak di antara dua klaster sebagai jarak rata-rata antara seluruh objek dalam klaster pertama terhadap seluruh objek dalam klaster kedua. Gambar 14.22 menyajikan dendogram. Dendogram menyajikan proses pengklasteran mulai dari *Stage 1* hingga *Stage 6*. Gambar 14.23 disajikan kode R. Sedangkan Gambar 14.24 hingga Gambar 14.26 merupakan hasil ekesekusi dari kode R pada Gambar 14.23.

- ⇨ Berdasarkan Gambar 14.25, jika dibentuk dua klaster, maka klaster-klaster tersebut adalah {1A,6F} dan {3C,5E,4D,2B}.

- ⇨ Berdasarkan Gambar 14.18, jika dibentuk tiga klaster, maka klaster-klaster tersebut adalah {A,F}, {C,E,D}, dan {B}.



Gambar 14.22

RR.R | analisisklaster.R

Source on Save

```
1  simpan_data=read.csv("data2.csv")
2  simpan_data
3  variabel = c("A", "B")
4  simpan_data2 = simpan_data[variabel]
5  simpan_data2
6
7  jarak = dist(as.matrix(simpan_data2)) #euclidean distance
8  jarak
9
10 jarak_pangkat_2=jarak*jarak #squared euclidean distance
11 jarak_pangkat_2
12
13 klaster1 = hclust(jarak_pangkat_2, method="average")
14 plot(klaster1) # display dendogram
15 dua_klaster = cutree(klaster1, k=2) # cut tree into 5 clusters
16 rect.hclust(klaster1, k=2, border="red")
17
18 klaster2 = hclust(jarak_pangkat_2, method="average")
19 plot(klaster2) # display dendogram
20 tiga_klaster = cutree(klaster2, k=3) # cut tree into 5 clusters
21 rect.hclust(klaster2, k=3, border="red")
```

Gambar 14.23

```
simpan_data=read.csv("data2.csv")
simpan_data
```

```
##   Batu    A    B
## 1    A 1.10 1.10
## 2    B 1.20 0.85
## 3    C 1.30 0.97
## 4    D 1.40 0.90
## 5    E 1.35 1.00
## 6    F 1.20 1.00
```

```
variabel = c("A", "B")
simpan_data2 = simpan_data[variabel]
simpan_data2
```

```
##      A    B
## 1 1.10 1.10
## 2 1.20 0.85
## 3 1.30 0.97
## 4 1.40 0.90
## 5 1.35 1.00
## 6 1.20 1.00
```

```
jarak = dist(as.matrix(simpan_data2)) #euclidean distance
jarak
```

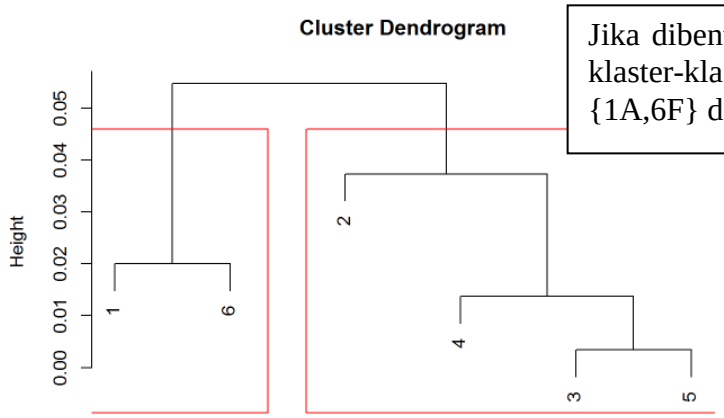
```
##            1          2          3          4          5
## 2 0.26925824
## 3 0.23853721 0.15620499
## 4 0.36055513 0.20615528 0.12206556
## 5 0.26925824 0.21213203 0.05830952 0.11180340
## 6 0.14142136 0.15000000 0.10440307 0.22360680 0.15000000
```

```
jarak_pangkat_2=jarak*jarak #squared euclidean distance
jarak_pangkat_2
```

```
##        1      2      3      4      5
## 2 0.0725
## 3 0.0569 0.0244
## 4 0.1300 0.0425 0.0149
## 5 0.0725 0.0450 0.0034 0.0125
## 6 0.0200 0.0225 0.0109 0.0500 0.0225
```

Gambar 14.24

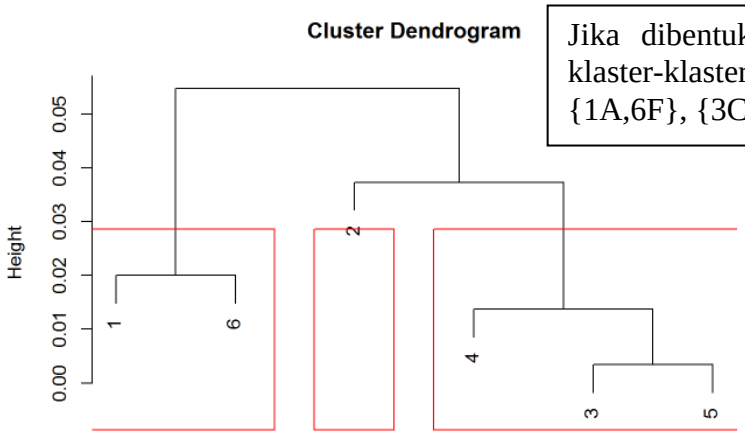
```
klaster1 = hclust(jarak_pangkat_2, method="average")
plot(klaster1) # display dendogram
dua_klaster = cutree(klaster1, k=2) # cut tree into 5 clusters
rect.hclust(klaster1, k=2, border="red")
```



Jika dibentuk dua klaster, maka klaster-klaster tersebut adalah {1A,6F} dan {3C,5E,4D,2B}.

Gambar 14.25

```
klaster2 = hclust(jarak_pangkat_2, method="average")
plot(klaster2) # display dendogram
tiga_klaster = cutree(klaster2, k=3) # cut tree into 5 clusters
rect.hclust(klaster2, k=3, border="red")
```



Jika dibentuk tiga klaster, maka klaster-klaster tersebut adalah {1A,6F}, {3C,5E,4D}, dan {2B}.

Gambar 14.26

Analisis Klaster dengan Metode Single Linkage

Berdasarkan Gambar 14.11, diketahui nilai *Squared Euclidean distance* **paling kecil** berada pada pasangan objek C dan objek E (pasangan objek yang berbeda), yakni bernilai 0,003 (Perhatikan Gambar 14.27). **Maka objek C dan objek E bergabung menjadi *cluster* (C,E)**. Pada Gambar 14.28, terlihat bahwa pada *Stage 1*, objek C dan objek E bergabung menjadi *cluster* (C,E). Perhatikan juga bahwa **nilai *coefficient* 0,003**, yang merupakan jarak antara objek C dan objek E.

	A	B	C	D	E	F
A	0	0.073	0.057	0.13	0.073	0.02
B	0.073	0	0.024	0.042	0.045	0.023
C	0.057	0.024	0	0.015	0.003	0.011
D	0.13	0.042	0.015	0	0.012	0.05
E	0.073	0.045	0.003	0.012	0	0.023
F	0.02	0.023	0.011	0.05	0.023	0

Gambar 14.27 *Squared Euclidean Distance* (Matriks Jarak)

Single Linkage

3 dalam hal ini adalah objek C, dan 5 dalam hal ini adalah objek E.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	5	.003	0	0	2
2	3	6	.011	1	0	3
3	3	4	.012	2	0	4
4	1	3	.020	0	3	5
5	1	2	.023	4	0	0

Gambar 14.28

Selanjutnya menghitung jarak antara *cluster* (C,E) terhadap objek lainnya.

⇨ Menghitung jarak antara *cluster* (C,E) terhadap objek A.

$$d_{(C,E)A} = \min\{d_{(C,A)}; d_{(E,A)}\} = \min\{0{,}057; 0{,}073\} = 0{,}057.$$

Dapat diartikan bahwa jarak antara objek C ke objek A lebih dekat, dibandingkan jarak antara objek E ke objek A.

⇨ Menghitung jarak antara *cluster* (C,E) terhadap objek B.

$$d_{(C,E)B} = \min\{d_{(C,B)}; d_{(E,B)}\} = \min\{0{,}024; 0{,}045\} = 0{,}024.$$

Dapat diartikan bahwa jarak antara objek C ke objek B lebih dekat, dibandingkan jarak antara objek E ke objek B.

⇨ Menghitung jarak antara *cluster* (C,E) terhadap objek D.

$$d_{(C,E)D} = \min\{d_{(C,D)}; d_{(E,D)}\} = \min\{0{,}015; 0{,}012\} = 0{,}012.$$

⇨ Menghitung jarak antara *cluster* (C,E) terhadap objek F.

$$d_{(C,E)F} = \min\{d_{(C,F)}; d_{(E,F)}\} = \min\{0{,}011; 0{,}023\} = 0{,}011.$$

Gambar 14.29 menyajikan jarak antara *cluster* (C,E) terhadap masing-masing objek.

	A	B	C	D	E	F
A	0	0.073	0.057	0.13	0.073	0.02
B	0.073	0	0.024	0.042	0.045	0.023
C	0.057	0.024	0	0.015	0.003	0.011
D	0.13	0.042	0.015	0	0.012	0.05
E	0.073	0.045	0.003	0.012	0	0.023
F	0.02	0.023	0.011	0.05	0.023	0
Minimum	0.057	0.024		0.012		0.011

Gambar 14.29

Sehingga diperoleh matriks jarak yang baru seperti pada Gambar 14.30.

	C,E	A	B	D	F
C,E	0	0.057	0.024	0.012	0.011
A	0.057	0	0.073	0.13	0.02
B	0.024	0.073	0	0.042	0.023
D	0.012	0.13	0.042	0	0.05
F	0.011	0.02	0.023	0.05	0

Gambar 14.30 Matriks Jarak

Sampai pada tahap ini, telah terbentuk 5 klaster, yakni {C,E}, {A}, {B}, {D}, dan {F}. Gambar 14.31 disajikan *output* SPSS.

Jika dibentuk 5 klaster, maka diperoleh klaster {C,E}, {A}, {B}, {D}, dan {F}.

Cluster Membership

Case	5 Clusters	4 Clusters	3 Clusters	2 Clusters
1:A	1	1	1	1
2:B	2	2	2	2
3:C	3	3	3	1
4:D	4	4	3	1
5:E	3	3	3	1
6:F	5	3	3	1

Gambar 14.31 *Output* SPSS untuk Analisis Klaster Metode *Single Linkage*

Berdasarkan Gambar 14.30, diketahui nilai **jarak paling kecil** berada pada pasangan (C,E) dan F, yakni bernilai 0,011. **Maka (C,E) dan F bergabung menjadi *cluster* (C,E,F)**. **Pada Gambar 14.28**, yakni *Stage 2* (**objek 3 dan objek 6 bergabung**). Diketahui **nilai *coefficient*** adalah 0,011 (lihat juga nilai *coefficient* pada Gambar 14.28, *Stage 2*).

Selanjutnya menghitung jarak antara *cluster* (C,E,F) terhadap objek lainnya.

⇨ Menghitung jarak antara *cluster* (C,E,F) terhadap objek A.

$$d_{(C,E,F)A} = \min\{d_{(C,E)A}; d_{(F,A)}\} = \min\{0{,}057; 0{,}02\} = 0{,}02.$$

⇨ Menghitung jarak antara *cluster* (C,E,F) terhadap objek B.

$$d_{(C,E,F)B} = \min\{d_{(C,E)B}; d_{(F,B)}\} = \min\{0{,}024; 0{,}023\} = 0{,}023.$$

⇨ Menghitung jarak antara *cluster* (C,E,F) terhadap objek D.

$$d_{(C,E,F)D} = \min\{d_{(C,E)D}; d_{(F,D)}\} = \min\{0{,}012; 0{,}05\} = 0{,}012.$$

Gambar 14.32 menyajikan jarak antara *cluster* (C,E,F) terhadap masing-masing objek.

	C,E	A	B	D	F
C,E	0	0.057	0.024	0.012	0.011
A	0.057	0	0.073	0.13	0.02
B	0.024	0.073	0	0.042	0.023
D	0.012	0.13	0.042	0	0.05
F	0.011	0.02	0.023	0.05	0
Minimum		0.02	0.023	0.012	

Gambar 14.32

Sehingga diperoleh matriks matriks jarak yang baru seperti pada Gambar 14.33.

	C,E,F	A	B	D
C,E,F	0	0.02	0.023	0.012
A	0.02	0	0.073	0.13
B	0.023	0.073	0	0.042
D	0.012	0.13	0.042	0

Gambar 14.33

Berdasarkan Gambar 14.33, diketahui nilai **jarak paling kecil** berada pada pasangan (C,E,F) dan D, yakni bernilai 0,012. **Maka (C,E,F) dan D bergabung menjadi *cluster* (C,E,F,D)**. **Pada Gambar 14.28**, yakni *Stage 3* (**objek 3 dan objek 4 bergabung**). Diketahui **nilai *coefficient*** adalah 0,012 (lihat juga nilai *coefficient* pada Gambar 14.28, *Stage 3*).

Selanjutnya menghitung jarak antara *cluster* (C,E,F,D) terhadap objek lainnya.

⇨ Menghitung jarak antara *cluster* (C,E,F,D) terhadap objek A.

$$d_{(C,E,F,D)A} = \min\{d_{(C,E,F)A}; d_{(D,A)}\} = \min\{0{,}02; 0{,}13\} = 0{,}02.$$

⇨ Menghitung jarak antara *cluster* (C,E,F,D) terhadap objek B.

$$d_{(C,E,F,D)B} = \min\{d_{(C,E,F)B}; d_{(D,B)}\} = \min\{0{,}023; 0{,}042\} = 0{,}023.$$

Gambar 14.34 menyajikan jarak antara *cluster* (C,E,F,D) terhadap masing-masing objek.

	C,E,F	A	B	D
C,E,F	0	0.02	0.023	0.012
A	0.02	0	0.073	0.13
B	0.023	0.073	0	0.042
D	0.012	0.13	0.042	0
Minimum		0.02	0.023	

Gambar 14.34

Sehingga diperoleh matriks jarak yang baru seperti pada Gambar 14.35.

	C,E,F,D	A	B
C,E,F,D	0	0.02	0.023
A	0.02	0	0.073
B	0.023	0.073	0

Gambar 14.35

Berdasarkan Gambar 14.35, diketahui **nilai jarak paling kecil** berada pada pasangan (C,E,F,D) dan A, yakni bernilai 0,002. **Maka (C,E,F,D) dan A bergabung menjadi *cluster* (C,E,F,D,A)**. **Pada Gambar 14.28**, yakni *Stage 4* (**objek 3 dan objek 1 bergabung**). Diketahui **nilai *coefficient*** adalah 0,02 (lihat juga nilai *coefficient* pada Gambar 14.28, *Stage 4*).

Selanjutnya menghitung jarak antara *cluster* (C,E,F,D,A) terhadap objek lainnya.

- ⇨ Menghitung jarak antara *cluster* (C,E,F,D,A) terhadap objek B.

$$d_{(C,E,F,D,A)B} = \min\{d_{(C,E,F,D)B}; d_{(A,B)}\} = \min\{0{,}023; 0{,}073\} = 0{,}023.$$

Gambar 14.36 menyajikan jarak antara *cluster* (C,E,F,D,A) terhadap objek B. diketahui jarak antara *cluster* (C,E,F,D,A) dan B adalah 0,023. **Pada Gambar 14.28**, yakni *Stage 5* (**objek 1 dan objek 2 bergabung**). Diketahui **nilai *coefficient*** adalah 0,023 (lihat juga nilai *coefficient* pada Gambar 14.28, *Stage 5*).

	C,D,E,F,A	B
C,D,E,F,A	0	0.023
B	0.023	0

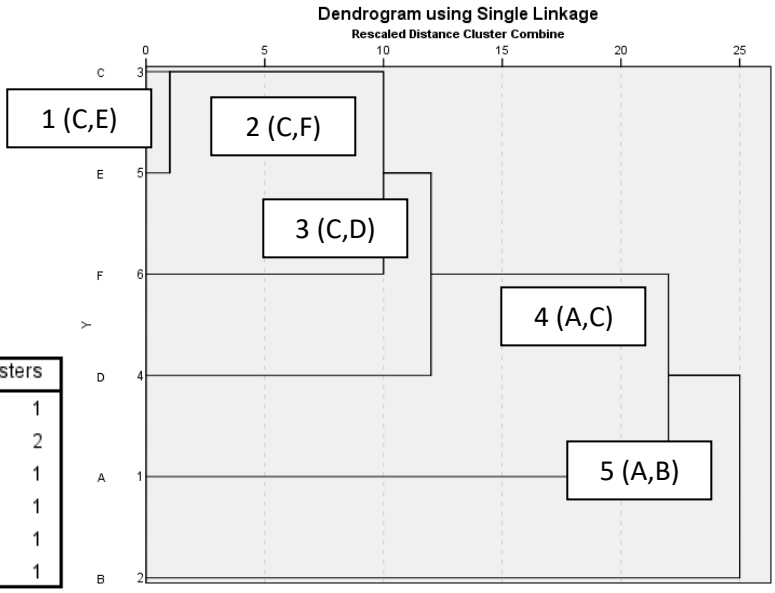
Gambar 14.36

Dari hasil yang telah diperoleh, dapat disimpulkan bahwa:

- ⇨ Jika dibentuk 3 klaster, maka {C,E,F,D} , {A}, dan {B} (lihat Gambar 14.35).
- ⇨ Jika dibentuk 2 klaster, maka {C,E,F,D,A} dan {B} (lihat Gambar 14.36).

Hasil pengklasteran tersebut sesuai dengan hasil yang diperoleh dengan SPSS, seperti pada Gambar 14.37. Berdasarkan Gambar 14.37, jika dibentuk 3 klaster, maka diperoleh klaster {A}, {B}, dan {C,D,E,F}, dan jika dibentuk dua klaster, maka diperoleh klaster {A,C,D,E,F} dan {B}. Gambar 14.34 menyajikan *dendogram* dengan menggunakan metode *single linkage*. Berdasarkan dendogram tersebut, dapat ditarik informasi:

- ⇨ Pertama, objek 3 (C) dan objek 5 (E) bergabung menjadi *cluster* (3,5).
- ⇨ Selanjutnya, *cluster* (3,5) bergabung dengan objek 6 (F) membentuk *cluster* (3,5,6).
- ⇨ Kemudian *cluster* (3,5,6) bergabung dengan objek 4 (D) membentuk *cluster* (3,5,6,4).
- ⇨ *Cluster* (3,5,6,4) bergabung dengan objek 1 (A) membentuk *cluster* (3,5,6,4,1).
- ⇨ Dan terakhir *cluster* (3,5,6,4,1) bergabung dengan objek 2 (B) membentuk *cluster* (3,5,6,4,1,2).



Cluster Membership

Case	3 Clusters	2 Clusters
1:A	1	1
2:B	2	2
3:C	3	1
4:D	3	1
5:E	3	1
6:F	3	1

Gambar 14.37 **Gambar 14.38**

Jadi, pada metode *average linkage*, pertama menentukan jarak paling minimum antara dua objek. Misalkan objek i dan objek k memiliki jarak yang paling minimum, maka objek i dan objek k bergabung menjadi suatu klaster (i, k). Langkah selanjutnya menghitung jarak antara klaster (i, k) terhadap klaster/objek lainnya (misalkan klaster l), dengan rumus sebagai berikut.

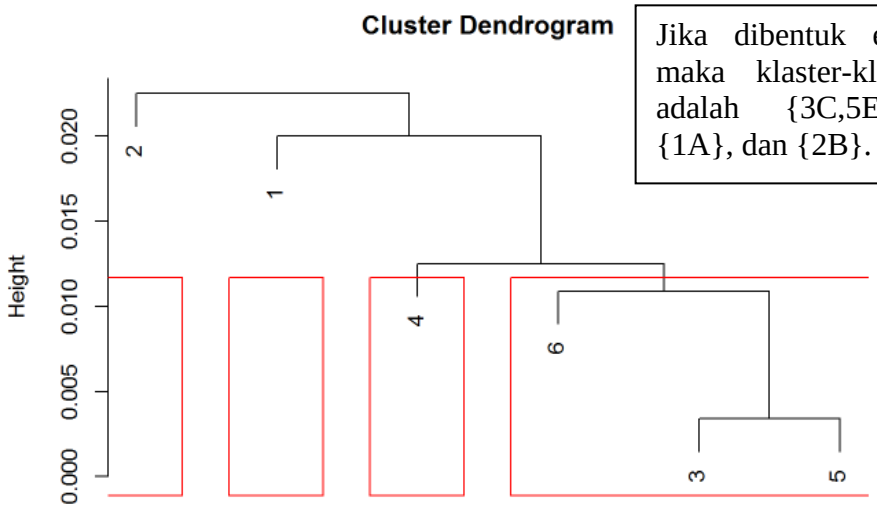
$$d_{(i,k),l} = \min\left(d_{(i,k)}; d_{(i,l)}\right).$$

Gambar 14.39 disajikan kode R. Sedangkan Gambar 14.40 hingga Gambar 14.42 merupakan hasil ekesekusi dari kode R pada Gambar 14.39.

```
Source on Save
simpan_data=read.csv("data2.csv")
simpan_data
variabel = c("A", "B")
simpan_data2 = simpan_data[variabel]
simpan_data2

jarak = dist(as.matrix(simpan_data2)) #euclidean distance
jarak

jarak_pangkat_2=jarak*jarak #squared euclidean distance
jarak_pangkat_2

klaster1 = hclust(jarak_pangkat_2, method="single")
plot(klaster1) # display dendogram
dua_klaster = cutree(klaster1, k=2) # cut tree into 5 clusters
rect.hclust(klaster1, k=2, border="red")

klaster2 = hclust(jarak_pangkat_2, method="single")
plot(klaster2) # display dendogram
tiga_klaster = cutree(klaster2, k=3) # cut tree into 5 clusters
rect.hclust(klaster2, k=3, border="red")

klaster3 = hclust(jarak_pangkat_2, method="single")
plot(klaster3) # display dendogram
empat_klaster = cutree(klaster3, k=4) # cut tree into 5 clusters
rect.hclust(klaster3, k=4, border="red")
```

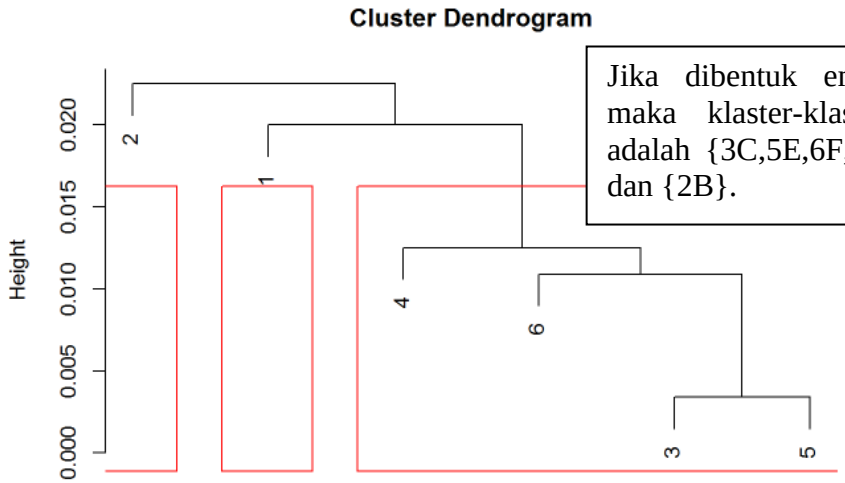
Gambar 14.39

```
klaster3 = hclust(jarak_pangkat_2, method="single")
plot(klaster3) # display dendogram
empat_klaster = cutree(klaster3, k=4) # cut tree into 5 clusters
rect.hclust(klaster3, k=4, border="red")
```



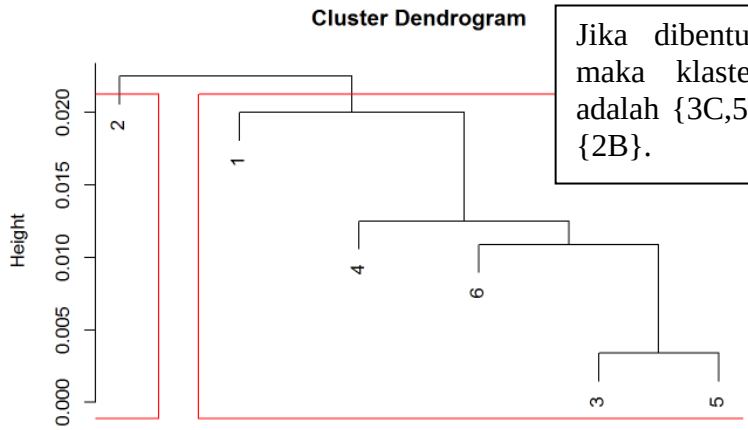
Jika dibentuk empat klaster, maka klaster-klaster tersebut adalah {3C,5E,6F}, {4D}, {1A}, dan {2B}.

Gambar 14.40



Jika dibentuk empat klaster, maka klaster-klaster tersebut adalah {3C,5E,6F, 4D}, {1A}, dan {2B}.

Gambar 14.41



Jika dibentuk empat klaster, maka klaster-klaster tersebut adalah {3C,5E,6F, 4D,1A} dan {2B}.

Gambar 14.42

Referensi

1. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

2. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

3. Hair, J.F Jr., R.E. Anderson, B.J. Babin, dan W.C. Black. 2010. *Multivariate Data Analysis, 7th Edition*. Pearson Prentice Hall.

4. Janssens, W., K. Wijnen, P.D. Pelsmacker, dan P.V. Kenhove. 2008. *Marketing Research with SPSS*. Pearson Prenticel Hall.

5. Johnson, R.A. dan D.W. Wichern. 2007. *Applied Multivariate Statistical Analysis, 6th Edition.* Pearson Prentice Hall.

6. Malhotra, N.K. dan D.F. Birks. 2006. *Marketing Research, An Applied Approach, 2nd European Edition*. London: Prentice Hall.

7. Stevens, J.P. 2009. *Applied Multivariate Statistics For The Social Science, 5th Edition*. New York: Routledge.

8. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html

9. http://www.r-tutor.com/gpu-computing/clustering/hierarchical-cluster-analysis

10. http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio3.pdf

BAB 15

PRINCIPAL COMPONENT ANALYSIS

Sekilas Principal Component Analysis (PCA) dan Factor Analysis (FA)

Principal component analysis (PCA) biasa disebut juga dengan analisis komponen utama, sementara *factor analysis* (FA) biasa disebut juga dengan analisis faktor. Kedua metode ini, yakni PCA dan FA, sama-sama **mereduksi** sekumpulan variabel-variabel asli (*original variables*) menjadi **beberapa variabel baru**, yang disebut dengan **faktor** atau **dimensi** atau **komponen**, namun pada dasarnya berbeda. PCA dan FA berusaha menghasilkan faktor dengan jumlah seminimal mungkin, yang mana faktor-faktor tersebut mampu menjelaskan jumlah maksimal dari *variance* (*explaining the maximum amount of common variance in a correlation matrix*) dalam matriks korelasi atau matriks R (keseluruhan variabel). Seringkali PCA dan FA memberikan hasil yang sama atau mirip. Supranto (2010:253) menyatakan sebagai berikut.

"*Untuk menyatakan dimensi yang mendasari evaluasi kepuasan pelanggan, kita menggunakan teknik seperti analisis faktor (AF) dan analisis komponen utama (AKU). Banyak sekali peneliti secara salah menganggap kedua analisis tersebut sebagai famili analisis faktor. Perlu disebutkan di sini, bahwa analisis faktor dan analisis komponen utama,* ***keduanya merupakan teknik mereduksi dimensi akan tetapi sebetulnya tak sama (not interchangeable). Namun demikian, keduanya sering memberikan hasil yang sama/mirip (similar result)***."

Terkait PCA dan FA, Field (2009:638) juga menyatakan sebagai berikut.

"*However, we should consider whether the techniques provide different solutions to the same problem. Based on an extensive literature review, Guadagnoli and Velicer (1988) concluded that the* ***solutions generated from principal component analysis differ little*** *from those derived from factor analytic techniques. In reality, there are some circumstances for which this statement is untrue. Stevens (2002) summarizes the evidence and concludes that with 30 or more variables and communalities greater than 0.7 for all variables, different solutions are unlikely; however, with fewer than 20 variables and any low communalities (< 0.4) differences can occur."*

Supranto (2010:262) menyatakan dalam PCA, faktor atau komponen ke-i, yakni F_i, merupakan kombinasi linear dari variabel asli, yakni

$$F_i = w_{i1}X_1 + w_{i2}X_2 + \cdots + w_{ij}X_j + \cdots + w_{ip}X_p.$$

Sementara pada FA, suatu variabel merupakan kombinasi linear dari faktor. Di samping itu, pada FA, **diasumsikan variabel (asli) dipengaruhi oleh variabel laten yang tak teramati** (*unobservable latent constructs*) (Supranto, 2010:261).

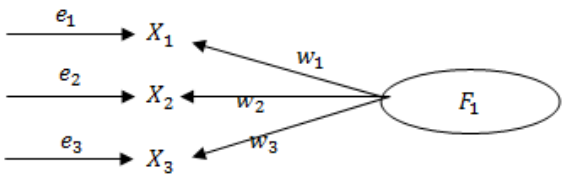
Meyers dkk (2005:488-489) juga menyatakan sebagai berikut.

"*We have indicated that the **component is analogous to the dependent variable in a multiple regression analysis**. This is because principal components are latent or composites descriptive of the information contained in the measured variables (the variables in the analysis). In some sense, the components "arise from" the measured variables. From a causal modeling perpective, the causal flow is **from the measured variables to the latent components**. Because we think of independent variables as causes of dependent variables, the measured variables are analogous to independent variables and the components are analogous to dependent variables. **Factor analysis shifts this conception around**. The measured variables are taken as "indicators" of the factors. Here, **the causal flow is from the factor (still a latent variable) to the measured indicator variables**. Thus, the factors are analogous to the independent variables, and the measured variables are analogous to the dependent variables.*

Jollife dalam Supranto (2010:261) menyatakan sebagai berikut.

"*A **final difference** between principal components and common factors is that the former can be calculated exactly from x, whereas the latter typically cannot. The PCS (Principal Component) **are exact linear function of x**. **The factors, however, are not exact linear function of x,** instead **x is defined as a linear function of f (the factors) apart from an error term…** .*"

Berikut diberikan ilustrasi untuk FA (Gambar 15.1).



Gambar 15.1

Berdasarkan Gambar 15.1, diperoleh persamaan sebagai berikut.

$$X_1 = w_1F_1 + e_1$$

$$X_2 = w_2F_1 + e_2$$

$$X_3 = w_2F_1 + e_3$$

Gambar 15.1 merupakan ilustrasi dari FA, di mana tanda panah bergerak dari faktor atau komponen menuju variabel, yang mana ini merupakan asumsi awal dari FA, **variabel (asli) dipengaruhi oleh variabel laten yang tak teramati** (*unobservable latent constructs*).

Misalkan diberikan data seperti pada Tabel 15.1. Berdasarkan data pada Tabel 15.1, terdapat tujuh variabel, yakni X1, X2, X3, X4, X5, X6, dan X7. Tabel 15.2 menyajikan matriks korelasi (matriks R), yakni menyajikan nilai korelasi (korelasi Pearson) antar dua variabel. Berdasarkan Tabel 15.2, nilai korelasi (korelasi Pearson) antara X1 dan X2 adalah -0,271, korelasi antara X1 dan X5 adalah -0,301, dan seterusnya. Perhatikan bahwa berdasarkan Tabel 15.2:

- ⇨ Terdapat korelasi yang tinggi antara X1 dan X6 (nilai korelasi 0,8992).
- ⇨ Terdapat korelasi yang tinggi antara X2 dan X5.
- ⇨ Terdapat korelasi yang tinggi di antara X3, X4, dan X7.

Sehingga **diduga** akan terbentuk tiga komponen, yakni komponen pertama meliputi X1 dan X6, komponen kedua meliputi X2 dan X5, dan komponen ketiga meliputi X3, X4, dan X7.

Tabel 15.1

No	X1	X2	X3	X4	X5	X6	X7
1	1	5	1	1	5	4	1
2	2	6	4	4	10	2	4
3	3	7	2	2	7	3	2
4	4	8	3	3	8	1	3
5	5	9	3	3	9	5	3
6	6	4	2	2	6	6	2
7	7	1	3	3	1	7	3
8	8	2	3	3	2	8	3
9	9	3	1	1	3	9	1
10	8	4	2	2	4	10	2
11	1	5	1	1	5	1	1
12	2	1	1	1	1	2	1
13	3	2	2	4	2	3	3
14	4	3	3	2	3	4	3
15	5	4	4	3	4	5	1

Tabel 15.2

Korelasi	X1	X2	X3	X4	X5	X6	X7
X1	1	-0.271	0.1653	0.087	-0.301	0.8992	0.0916
X2		1	0.215	0.1309	0.9214	-0.347	0.1794
X3			1	0.8043	0.3278	-0.008	0.673
X4				1	0.2559	-0.08	0.8076
X5					1	-0.368	0.3218
X6						1	-0.074
X7							1

Mereduksi Variabel dan Eigenvalues

Selanjutnya mereduksi variabel-variabel atau indikator-indikator (dalam contoh kasus ini terdapat 7 variabel) menjadi beberapa komponen (yang jumlahnya lebih sedikit). ***Eigenvalues*** (nilai-nilai eigen) merupakan salah satu pendekatan yang dapat digunakan untuk menentukan **jumlah komponen yang akan dipertahankan dalam analisis**. (selain pendekatan *eigenvalues*, terdapat pendekatan *scree plot*). Pada Gambar 15.2 terdapat 7 komponen yang terbentuk (diketahui jumlah variabel juga 7), **namun tidak semua komponen akan dipertahankan dalam analsis selanjutnya**. Berdasarkan Gambar 15.2, dari 7 komponen yang terbentuk, hanya 3 komponen yang dipertahankan dalam analisis selanjutnya, yakni komponen 1, 2, dan 3. Sebagaimana Field (2009:639) menyatakan sebagai berikut.

*"**Not all factors are retained in an analysis**, and there is debate over the criterion used to decide whether a factor is statistically important. I mentioned above that eigenvalues associated with a variate indicate the substantive importance of that factor. Therefore, it seems logical that we should retain only factors with large eigenvalues… Typically there will be a few factors with quite high eigenvalues, and many factors with relatively low eigenvalues, …"*

Lebih lanjut, Field (2009:640) menyatakan sebagai berikut.

*"Although scree plots are very useful, factor selection should not be based on this criterion alone. **Kaiser (1960) recommended retaining all factors with eigenvalues greater than 1**. This criterion is based on the idea that the eigenvalues represent the amount of variation explained by a factor and that an eigenvalue of 1 represents a substantial amount of variation."*

Berdasarkan uraian di atas, Kaiser (1960) memberi rekomendasi bahwa *eigenvalue* dari suatu faktor atau komponen yang lebih besar dari 1, agar dipertahankan dalam proses analisis. Perhatikan bahwa berdasarkan Gambar 15.2, *eigenvalues* untuk komponen 1, 2, dan 3 adalah 2,988, 2,277, dan 1,126, di mana lebih besar dari 1, sehingga komponen 1, 2, dan 3 dipertahankan untuk analisis selanjutnya (terbentuk tiga komponen).

Berdasarkan Gambar 15.2, diketahui komponen pertama mampu menjelaskan 42,679% dari *total variance*, komponen kedua mampu menjelaskan 32,531% dari *total variance* dan komponen ketiga mampu menjelaskan 16,082% dari *total variance*. Jadi, ketiga komponen tersebut mampu menjelaskan 91,293% dari *total variance.*

```
##        Eigenvalue Proportion_Variance Cummulative_Proportion_Variance
## Comp.1 2.98751893          42.6788419                          42.67884
## Comp.2 2.27718852          32.5312646                          75.21011
## Comp.3 1.12577018          16.0824311                          91.29254
## Comp.4 0.32242662           4.6060946                          95.89863
## Comp.5 0.13928171           1.9897387                          97.88837
## Comp.6 0.09359966           1.3371380                          99.22551
## Comp.7 0.05421438           0.7744911                         100.00000
```

Gambar 15.2

Principal Component Analysis: X1, X2, X3, X4, X5, X6, X7

Eigenanalysis of the Correlation Matrix

Eigenvalue	2.9875	2.2772	1.1258	0.3224	0.1393	0.0936	0.0542
Proportion	0.427	0.325	0.161	0.046	0.020	0.013	0.008
Cumulative	0.427	0.752	0.913	0.959	0.979	0.992	1.000

Gambar 15.3 Hasil berdasarkan Minitab

Analisis Nilai Loading

Berdasarkan pemaparan sebelumnya, diketahui dipertahankan tiga komponen. Gambar 15.4 menyajikan nilai *loading* antara variabel dan komponen. Diketahui nilai *loading* antara variabel X1 dan *Comp.1* adalah 0,165, nilai *loading* antara variabel X1 dan *Comp.2* adalah 0,546, dan seterusnya.

```
PCA$loading
```

```
##
## Loadings:
##    Comp.1 Comp.2 Comp.3
## X1  0.165  0.546 -0.421
## X2 -0.389 -0.278 -0.548
## X3 -0.424  0.343
## X4 -0.433  0.346  0.270
## X5 -0.449 -0.233 -0.460
## X6  0.261  0.485 -0.440
## X7 -0.429  0.313  0.192
```

Gambar 16.4 merupakan *output* R yang menyajikan *loading* antara komponen dan variabel.

***Loading* yang kosong antara Comp.3 dan X3.**

Gambar 15.4

Variable	PC1	PC2	PC3
X1	-0.165	0.546	-0.421
X2	0.389	-0.278	-0.548
X3	0.424	0.343	0.087
X4	0.433	0.346	0.270
X5	0.449	-0.233	-0.460
X6	-0.261	0.485	-0.440
X7	0.429	0.313	0.192

Gambar 16.5 merupakan *output* Minitab yang menyajikan *loadng* antara komponen dan variabel. **Dalam Minitab**, **seluruh *loading* ditampilkan.**

Dalam R, |*loading*| < 0,1 tidak ditampilkan. Dalam hal ini menggunakan fungsi princomp().

Gambar 15.5 Hasil berdasarkan Minitab

Berdasarkan nilai *loading* tersebut, dapat **digunakan untuk menentukan apakah suatu variabel masuk ke dalam komponen pertama, kedua, atau ketiga**. Berdasarkan Gambar 15.4, dapat ditarik informasi sebagai berikut berikut.

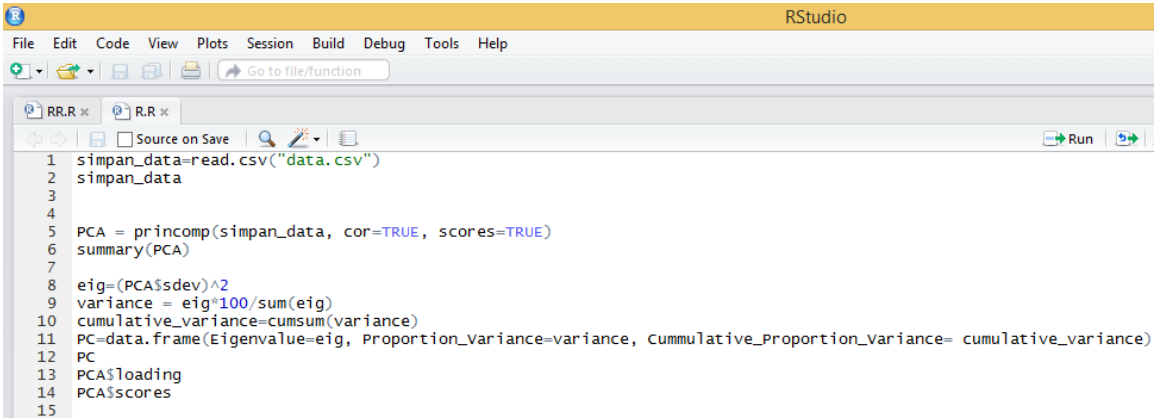
- ⇨ Variabel X1 masuk ke dalam komponen 2
- ⇨ Variabel X2 masuk ke dalam komponen 3
- ⇨ Variabel X3 masuk ke dalam komponen 1
- ⇨ Variabel X4 masuk ke dalam komponen 1
- ⇨ Variabel X5 masuk ke dalam komponen 3
- ⇨ Variabel X6 masuk ke dalam komponen 2
- ⇨ Variabel X7 masuk ke dalam komponen 1

Pada pembahasan sebelumnya, **telah diduga** melalui analisis korelasi, yakni:

- ⇨ Terdapat korelasi yang tinggi antara X1 dan X6 (nilai korelasi 0,8992).
- ⇨ Terdapat korelasi yang tinggi antara X2 dan X5.
- ⇨ Terdapat korelasi yang tinggi di antara X3, X4, dan X7.

Sehingga **diduga** akan terbentuk tiga komponen, yakni komponen pertama meliputi X1 dan X6, komponen kedua meliputi X2 dan X5, dan komponen ketiga meliputi X3, X4, dan X7.

Berikut disajikan kode R, yang apabila dieksekusi, akan menghasilkan *output* R sebelumnya.



Gambar 15.6

Referensi

1. Field, A. 2009. *Discovering Statistics Using SPSS, 3rd Edition*. London: Sage.

2. Gio, P.U. dan E. Rosmaini, 2015. Belajar Olah Data dengan SPSS, Minitab, R, Microsoft Excel, EViews, LISREL, AMOS, dan SmartPLS. USUpress.

3. Hair, J.F Jr., R.E. Anderson, B.J. Babin, dan W.C. Black. 2010. *Multivariate Data Analysis, 7th Edition*. Pearson Prentice Hall.

4. Malhotra, N.K. dan D.F. Birks. 2006. *Marketing Research, An Applied Approach, 2nd European Edition*. London: Prentice Hall.

5. Stevens, J.P. 2009. *Applied Multivariate Statistics For The Social Science, 5th Edition*. New York: Routledge.

6. Supranto, J. 2010. Analisis Multivariat, Arti & Interpretasi. Jakarta: Rineka Cipta.

7. http://www.r-bloggers.com/computing-and-visualizing-pca-in-r/

8. http://www.statmethods.net/advstats/factor.html

9. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/princomp.html

10. http://bioconductor.org/packages/release/bioc/html/pcaMethods.html

BAB 16

POHON KEPUTUSAN (*DECISION TREE*)

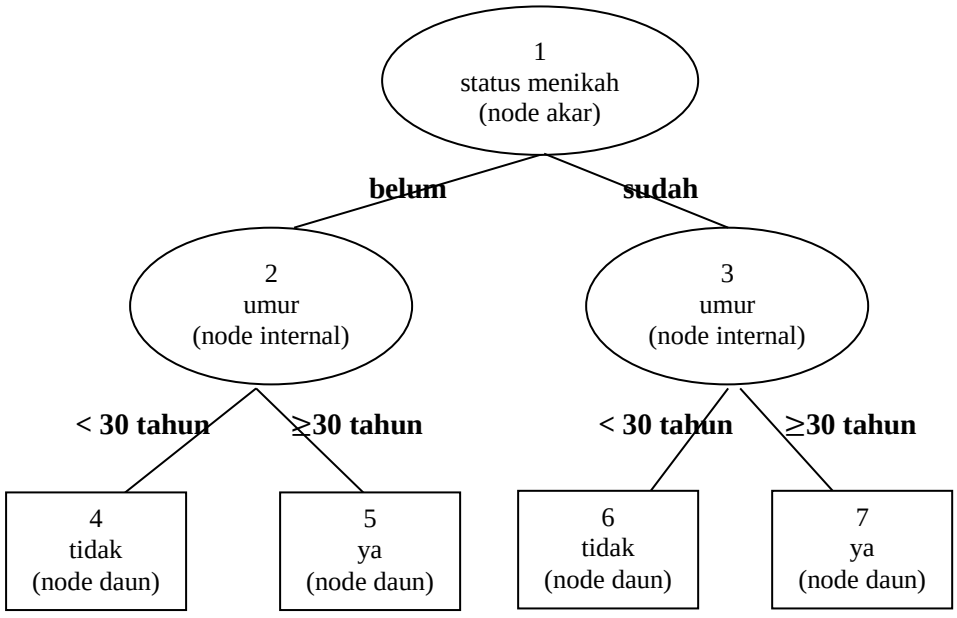
Sekilas Pohon Keputusan

Berikut diberikan data mengenai catatan kepemilikan mobil (Tabel 16.1).

Tabel 16.1 Data mengenai Catatan Kepemilikan Mobil

Nama	Umur	Status Menikah	Kepemilikan
A	25	sudah	ya
B	26	belum	tidak
C	28	belum	tidak
D	19	sudah	tidak
E	28	sudah	ya
F	40	sudah	ya
G	35	sudah	ya
H	32	belum	ya
I	33	sudah	tidak
J	55	sudah	ya

Berdasarkan data pada Tabel 16.1, diketahui responden bernama A, berusia 25 tahun, dengan status sudah menikah, memiliki mobil. Responden bernama H, berusia 32 tahun, dengan status belum menikah, memiliki mobil, dan seterusnya. Berdasarkan data pada Tabel 16.1, dibentuk pohon keputusan (*decision tree*) sebagai berikut (Gambar 16.1).



Gambar 16.1 Pohon Keputusan untuk Klasifikasi Kepemilikan Mobil berdasarkan Umur dan Status

Berdasarkan Gambar 16.1, pohon keputusan terdiri dari:

- ⇨ Akar pohon atau node akar (*root node*). Variabel **status menikah** berkedudukan sebagai node akar. Pada node akar tidak memiliki cabang (*branch*) masukan, namun dapat tidak memiliki atau memiliki cabang keluaran. Pada Gambar 16.1, node akar dari variabel **status menikah** memiliki dua cabang keluaran.

- ⇨ Node internal (*internal node*). Variabel **umur** bertindak sebagai node internal. Pada node internal **umur** memiliki jumlah cabang keluaran sebanyak dua. Pada node internal, cabang keluaran dapat berjumlah dua atau lebih, namun jumlah cabang masukkan tepat satu.

- ⇨ Node daun (*leaf node*/*terminal node*). Pada Gambar 16.1, node daun direpresentasikan dengan bentuk persegi. Pada node daun memiliki tepat satu cabang masukkan, dan tidak memiliki cabang keluaran.

Pohon keputusan pada Gambar 16.1 dibangun berdasarkan data pada Tabel 16.1. Maka data pada Tabel 16.1 disebut juga dengan istilah **data latih** (*training data*). Pembuatan pohon keputusan berdasarkan data latih pada Tabel 16.1 disebut juga dengan istilah **induksi**. Andaikan diketahui seseorang bernama Andi, berusia 35 tahun dan sudah menikah. Maka dengan menggunakan pohon keputusan, Andi dapat diprediksi, apakah termasuk ke dalam kelompok orang yang memiliki mobil atau tidak. Berdasarkan pohon keputusan pada Gambar 16.1, diketahui Andi diprediksi termasuk ke dalam kelompok yang memiliki mobil. Prediksi pengelompokkan yang baru saja dilakukan disebut juga dengan istilah **deduksi**.

Gorunescu (2011:161) menyatakan pohon keputusan memiliki tiga pendekatan klasik, yakni sebagai berikut.

1. Pohon klasifikasi (*classification trees*), digunakan ketika hasil prediksi merupakan keanggotaan dari salah satu kelompok yang ada. Pada pohon klasifikasi, variabel tak bebas (*dependent variable*) bersifat kategori. Pohon keputusan pada Gambar 16.1 termasuk ke dalam pendekatan pohon klasifikasi. Diketahui variabel dependen **kepemilikan** memiliki dua kategori, yakni "ya" (memiliki mobil) dan "tidak" (tidak memiliki mobil).

2. Pohon regresi (*regression trees*), digunakan ketika hasil prediksi berupa nilai atau angka real. Contoh variabel dependen untuk pendekatan pohon regresi adalah harga minyak, harga rumah, harga beras, dan sebagainya.

3. *Classification and Regression Tree* yang merupakan kombinasi antara (1) dan (2).

Untuk membuat pohon keputusan, terdapat beberapa algoritma yang dapat digunakan, yakni di antaranya sebagai berikut (Gorunescu, 2011:164).

1. ID3, C4.5, dan C5.0 – *Machine learning*;

2. CART (C&RT) – *Statistics*;

3. CHAID – *Pattern recognition*.

Gorunescu (2011:165) menyatakan salah satu kriteria yang dapat digunakan untuk menentukan titik pemecah terbaik (*optimal splitting point*) adalah GINI *index*, yang biasanya

digunakan dalam algoritma CART (C&RT) dan SPRINT. Lebih lanjut Gorunescu (2011:166-167) menyatakan dalam penerapan GINI *index* untuk data berskala *continuous*, terdapat beberapa metode yang dapat digunakan untuk menentukan titik pemecah terbaik, yakni metode *brute-force* dan *metode midpoints* (Gorunescu, 2011:166-167).

Membuat Pohon Klasifikasi dengan Satu Variabel Bebas Continuous, Kriteria Pemecah GINI, dengan Metode Brute-Force dan Metode Midpoints (Contoh Perhitungan dan Penyelesaian R)

Misalkan diberikan data seperti pada Tabel 16.2. Berdasarkan data pada Tabel 16.2, terdapat satu variabel tak bebas (Y) dan satu variabel bebas (X_1). Diketahui **terdapat dua kategori pada variabel tak bebas**, yakni A dan B.

Tabel 16.2

Y	X_1
A	3
A	1
A	5
B	9
B	12
B	7

Berikut akan dibentuk pohon klasifikasi berdasarkan kriteria pemecah GINI, dengan metode *brute-force* dan metode *midpoints*. Berikut akan dihitung nilai GINI *index* dan GINI *splitting index* dengan metode ***brute-force***.

$$I_{GINI}(X_1 \leq 1) = 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 1) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right) = 1 - 0{,}52 = 0{,}48$$

$$GINI_{split} = \left(\frac{1}{6}\right)(0) + \left(\frac{5}{6}\right)(0{,}48) = 0{,}4$$

$$I_{GINI}(X_1 \leq 3) = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 3) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) = 1 - 0{,}625 = 0{,}375$$

$$GINI_{split} = \left(\frac{2}{6}\right)(0) + \left(\frac{4}{6}\right)(0{,}375) = 0{,}25$$

$$I_{GINI}(X_1 \leq 5) = 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 5) = 1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right) = 1 - 1 = 0$$

$$GINI_{split} = \left(\frac{3}{6}\right)(0) + \left(\frac{3}{6}\right)(0) = 0$$

$$I_{GINI}(X_1 \leq 7) = 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) = 1 - 0{,}625 = 0{,}375$$

$$I_{GINI}(X_1 > 7) = 1 - \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right) = 1 - 1 = 0$$

$$GINI_{split} = \left(\frac{4}{6}\right)(0{,}375) + \left(\frac{2}{6}\right)(0) = 0{,}25$$

$$I_{GINI}(X_1 \leq 9) = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right) = 1 - 0{,}52 = 0{,}48$$

$$I_{GINI}(X_1 > 9) = 1 - \left(\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right) = 1 - 1 = 0$$

$$GINI_{split} = \left(\frac{5}{6}\right)(0{,}48) + \left(\frac{1}{6}\right)(0) = 0{,}4$$

$$I_{GINI}(X_1 \leq 12) = 1 - \left(\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right) = 1 - 0{,}5 = 0{,}5$$

$$I_{GINI}(X_1 > 12) = 1 - (0^2 + 0^2) = 1$$

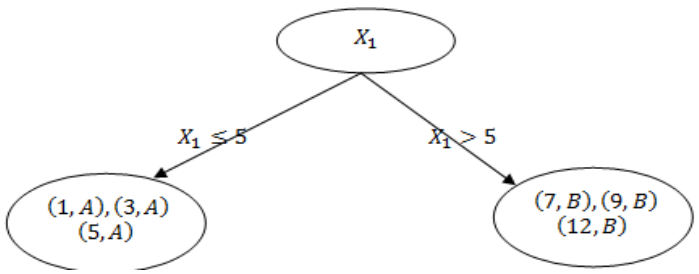
$$GINI_{split} = \left(\frac{6}{6}\right)(0{,}5) + (0)(0) = 0{,}5$$

Hasil perhitungan sebelumnya dapat diringkas seperti pada Tabel 16.3.

Tabel 16.3

X_1	*GINI Split Index*
1	0,4
3	0,25
5	0
7	0,25
9	0,4
12	0,5

Berdasarkan Tabel 16.3, diketahui nilai *GINI split index* terkecil berada pada nilai $X_1 = 5$, yang mana merupakan **titik pemecah optimal**. Perhatikan gambar pohon klasifikasi berikut.



Gambar 16.2

Berdasarkan Gambar 16.2, perhatikan bahwa tidak terjadi kesalahan klasifikasi. Berikut akan dihitung nilai GINI *index* dan GINI *splitting index* dengan **metode *midpoints***.

$$I_{GINI}(X_1 \leq 2) = 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right) = 1 - 1 = 0$$
$$I_{GINI}(X_1 > 2) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right) = 1 - 0{,}52 = 0{,}48$$
$$GINI_{split} = \left(\frac{1}{6}\right)(0) + \left(\frac{5}{6}\right)(0{,}48) = 0{,}4$$

$$I_{GINI}(X_1 \leq 4) = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right) = 1 - 1 = 0$$
$$I_{GINI}(X_1 > 4) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) = 1 - 0{,}625 = 0{,}375$$
$$GINI_{split} = \left(\frac{2}{6}\right)(0) + \left(\frac{4}{6}\right)(0{,}375) = 0{,}25$$

$$I_{GINI}(X_1 \leq 6) = 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right) = 1 - 1 = 0$$
$$I_{GINI}(X_1 > 6) = 1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right) = 1 - 1 = 0$$
$$GINI_{split} = \left(\frac{3}{6}\right)(0) + \left(\frac{3}{6}\right)(0) = 0$$

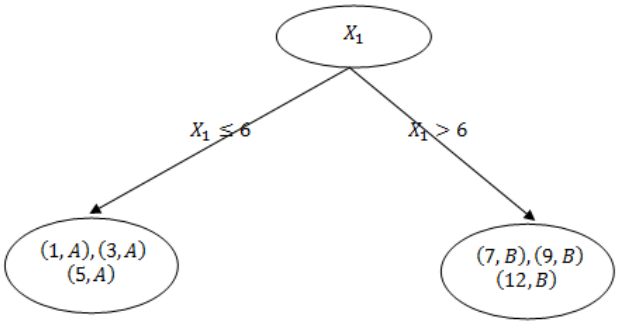
$$I_{GINI}(X_1 \leq 8) = 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) = 1 - 0{,}625 = 0{,}375$$
$$I_{GINI}(X_1 > 8) = 1 - \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right) = 1 - 1 = 0$$
$$GINI_{split} = \left(\frac{4}{6}\right)(0{,}375) + \left(\frac{2}{6}\right)(0) = 0{,}25$$

$$I_{GINI}(X_1 \leq 10{,}5) = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right) = 1 - 0{,}52 = 0{,}48$$
$$I_{GINI}(X_1 > 10{,}5) = 1 - \left(\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right) = 1 - 1 = 0$$
$$GINI_{split} = \left(\frac{5}{6}\right)(0{,}48) + \left(\frac{1}{6}\right)(0) = 0{,}4$$

Tabel 16.4

X_1	*GINI Split Index*
2	0,4
4	0,25
6	0
8	0,25
10,5	0,4

Diketahui nilai *GINI split index* terkecil berada pada nilai $X_1 = 6$, yang mana merupakan **titik pemecah optimal**. Perhatikan gambar pohon klasifikasi berikut (Gambar 16.3).



Gambar 16.3

Gambar 16.4 menyajikan kode R untuk membentuk pohon klasifikasi, seperti pada Gambar 16.7. Pada Gambar 16.4, digunakan fungsi **rpart** dan **tree** untuk membentuk pohon klasifikasi. Fungsi **rpart** tersedia dalam *package* **rpart**, dan fungsi **tree** tersedia dalam *package* **tree**. Gambar 16.5 hingga Gambar 16.9 merupakan hasil eksekusi kode R pada Gambar 16.4.

```r
simpan_data=read.csv("data1.csv")
simpan_data

library(rpart)
tree <- rpart(Y ~ X1, simpan_data, minsplit=1)
summary(tree)
print(tree)


library(rpart.plot)
prp(tree, faclen = 0, cex = 0.8, extra = 1)

treePrediction <- predict(tree, simpan_data, type = "class")
treePrediction

library(caret)
confusionMatrix(treePrediction, simpan_data$Y)

#Berikut akan digunakan R untuk membuat pohon klasifikasi dengan package Tree.

library(tree)
tree1 <- tree(Y ~ X1,simpan_data, split = c("gini"), control=tree.control(nobs=6, mincut = 1, minsize = 2))
summary(tree1)
print(tree1)
plot(tree1)
text(tree1)
```

Gambar 16.4

```r
simpan_data=read.csv("data1.csv")
simpan_data
```

```
##   Y X1
## 1 A  3
## 2 A  1
## 3 A  5
## 4 B  9
## 5 B 12
## 6 B  7
```

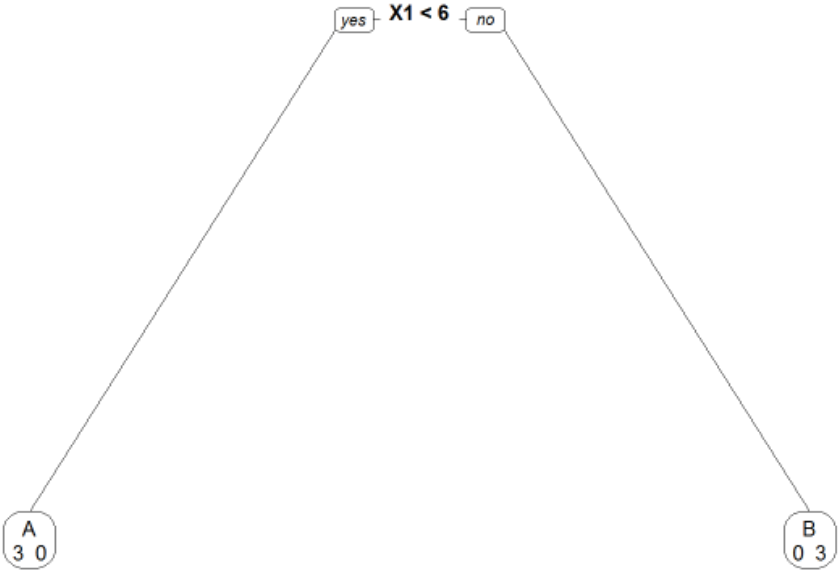
Gambar 16.5

```
## n= 6
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 6 3 A (0.5000000 0.5000000)
##   2) X1< 6 3 0 A (1.0000000 0.0000000) *
##   3) X1>=6 3 0 B (0.0000000 1.0000000) *
```

```
library(rpart.plot)
prp(tree, faclen = 0, cex = 0.8, extra = 1)

treePrediction <- predict(tree, simpan_data, type = "class")
treePrediction
```

```
## 1 2 3 4 5 6
## A A A B B B
## Levels: A B
```

Gambar 16.6



Gambar 16.7

```
confusionMatrix(treePrediction, simpan_data$Y)
```

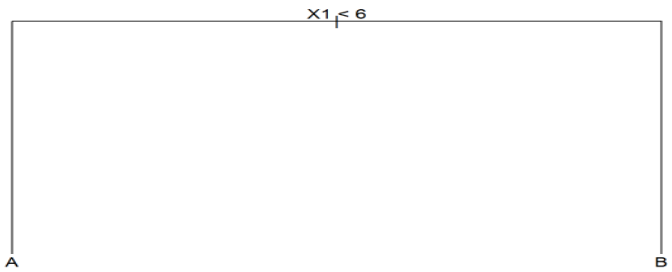
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction A B
##          A 3 0
##          B 0 3
```

Gambar 16.8

```
print(tree1)
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
## 1) root 6 8.318 A ( 0.5 0.5 )
##   2) X1 < 6 3 0.000 A ( 1.0 0.0 ) *
##   3) X1 > 6 3 0.000 B ( 0.0 1.0 ) *
```

```
plot(tree1)
text(tree1)
```



Gambar 16.9

Perhatikan hasil pengelompokkan untuk fungsi **rpart** pada Gambar 16.6, yakni

$$X_1 < 6; X_1 \geq 6.$$

Sementara hasil pengelompokkan untuk fungsi **tree** pada Gambar 16.9, yakni

$$X_1 < 6;\ X_1 > 6.$$

Pohon klasifikasi dibentuk berdasarkan kriteria pemecah GINI, dengan metode *midpoints*

Membuat Pohon Klasifikasi dengan Satu Variabel Bebas Continuous, Kriteria Pemecah GINI, dengan Metode Midpoints (Contoh Perhitungan dan Penyelesaian R)

Misalkan diberikan data seperti pada Tabel 16.5. Berdasarkan data pada Tabel 16.5, terdapat satu variabel tak bebas (Y) dan satu variabel bebas (X_1). Diketahui **terdapat dua kategori pada variabel tak bebas**, yakni A dan B.

Tabel 16.5

Y	X_1
A	3
A	1
A	5
B	9
A	12
B	7

Berikut akan dibentuk pohon klasifikasi berdasarkan kriteria pemecah GINI, dengan metode *midpoints*. Berikut akan dihitung nilai GINI *index* dan GINI *splitting index* dengan metode *midpoints*.

$$I_{GINI}(X_1 \leq 2) = 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right) = 1 - 1 = 0$$
$$I_{GINI}(X_1 > 2) = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right) = 1 - 0{,}52 = 0{,}48$$
$$GINI_{split} = \left(\frac{1}{6}\right)(0) + \left(\frac{5}{6}\right)(0{,}48) = 0{,}4$$

$$I_{GINI}(X_1 \leq 4) = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right) = 1 - 1 = 0$$
$$I_{GINI}(X_1 > 4) = 1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right) = 1 - 0{,}5 = 0{,}5$$
$$GINI_{split} = \left(\frac{2}{6}\right)(0) + \left(\frac{4}{6}\right)(0{,}5) = 0{,}33$$

$$I_{GINI}(X_1 \leq 6) = 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right) = 1 - 1 = 0$$
$$I_{GINI}(X_1 > 6) = 1 - \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) = 1 - 0{,}555 = 0{,}445$$
$$GINI_{split} = \left(\frac{3}{6}\right)(0) + \left(\frac{3}{6}\right)(0{,}445) = 0{,}2225$$

$$I_{GINI}(X_1 \leq 8) = 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) = 1 - 0{,}625 = 0{,}375$$
$$I_{GINI}(X_1 > 8) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = 1 - 0{,}5 = 0{,}5$$
$$GINI_{split} = \left(\frac{4}{6}\right)(0{,}375) + \left(\frac{2}{6}\right)(0{,}5) = 0{,}416$$

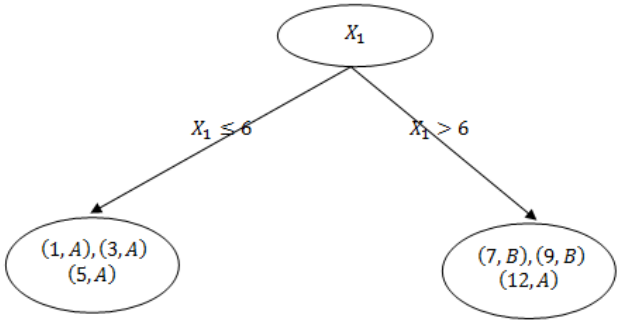
$$I_{GINI} = (X_1 \leq 10{,}5) = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right) = 1 - 0{,}52 = 0{,}48$$
$$I_{GINI} = (X_1 > 10{,}5) = 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right) = 1 - 1 = 0$$
$$GINI_{split} = \left(\frac{5}{6}\right)(0{,}48) + \left(\frac{1}{6}\right)(0) = 0{,}4$$

Berdasarkan perhitungan di atas, dapat diringkas seperti pada Tabel 16.6.

Tabel 16.6

X_1	*GINI Split Index*
2	0,4
4	0,33
6	0,2225
8	0,416
10,5	0,4

Di ketahui nilai *GINI split index* terkecil berada pada nilai $X_1 = 6$, yang mana merupakan titik pemecah optimal. Perhatikan gambar pohon klasifikasi berikut.



Gambar 16.10

Perhatikan bahwa terjadi kesalahan klasifikasi sebanyak 1.

Tabel 16.7

Y	X_1
B	7
B	9
A	12

Menghitung nilai *GINI index* dan *GINI splitting index*.

$$I_{GINI}(X_1 \leq 8) = 1 - \left(\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 8) = 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) = 1 - 0{,}5 = 0{,}5$$

$$GINI_{split} = \left(\frac{1}{3}\right)(0) + \left(\frac{2}{3}\right)(0{,}5) = 0{,}333$$

$$I_{GINI}(X_1 \leq 10{,}5) = 1 - \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 10{,}5) = 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right) = 1 - 1 = 0$$

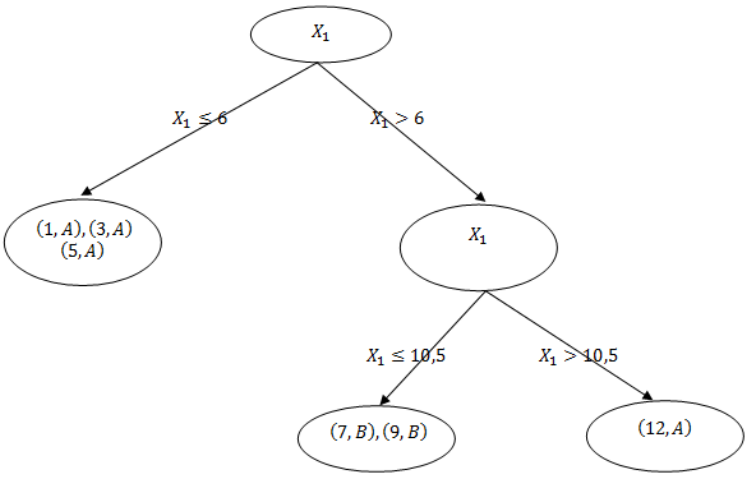
$$GINI_{split} = \left(\frac{2}{3}\right)(0) + \left(\frac{1}{3}\right)(0) = 0$$

Berdasarkan perhitungan di atas, dapat diringkas sebagai berikut (Tabel 16.8).

Tabel 16.8

X_1	*GINI Split Index*
8	0,333
10,5	0

Diketahui nilai *GINI split index* terkecil berada pada nilai $X_1 = 10{,}5$, yang mana merupakan titik pemecah optimal. Perhatikan gambar pohon klasifikasi seperti pada Gambar 16.11.



Gambar 16.11

Gambar 16.12 menyajikan kode R untuk membentuk pohon klasifikasi. Gambar 16.13 hingga Gambar 16.16 merupakan hasil eksekusi kode R pada Gambar 16.12.

RR.R | R.R | code-ipaconf2016.R | kode_R_sedikit_modifikasi.R

Source on Save | Run

```r
simpan_data=read.csv("data2.csv")
simpan_data

library(rpart)
tree <- rpart(Y ~ X1, simpan_data, minsplit=1)
summary(tree)
print(tree)


library(rpart.plot)
prp(tree, faclen = 0, cex = 0.8, extra = 1, digits = 4)

treePrediction <- predict(tree, simpan_data, type = "class")
treePrediction

library(caret)
confusionMatrix(treePrediction, simpan_data$Y)

#Berikut akan digunakan R untuk membuat pohon klasifikasi dengan package Tree.

library(tree)
tree1 <- tree(Y ~ X1,simpan_data, split = c("gini"), control=tree.control(nobs=6, mincut = 1, minsize = 2))
summary(tree1)
print(tree1)
plot(tree1)
text(tree1)
```

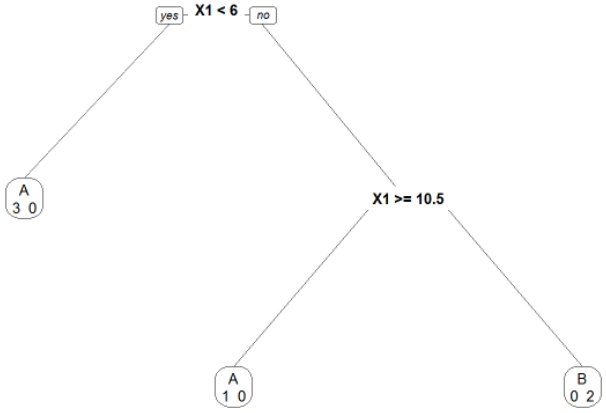
Gambar 16.12

```
## n= 6
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 6 2 A (0.6666667 0.3333333)
##   2) X1< 6 3 0 A (1.0000000 0.0000000) *
##   3) X1>=6 3 1 B (0.3333333 0.6666667)
##     6) X1>=10.5 1 0 A (1.0000000 0.0000000) *
##     7) X1< 10.5 2 0 B (0.0000000 1.0000000) *
```

```r
library(rpart.plot)
prp(tree, faclen = 0, cex = 0.8, extra = 1, digits = 4)

treePrediction <- predict(tree, simpan_data, type = "class")
treePrediction
```

```
## 1 2 3 4 5 6
## A A A B A B
## Levels: A B
```

Gambar 16.13




```
confusionMatrix(treePrediction, simpan_data$Y)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction A B
##          A 4 0
##          B 0 2
```

Gambar 16.14

```
library(tree)
tree1 <- tree(Y ~ X1,simpan_data, split = c("gini"), control=tree.control(nobs=6, mincut = 1, minsize = 2))
summary(tree1)
```

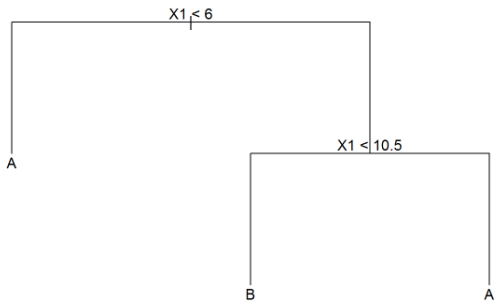
```
##
## Classification tree:
## tree(formula = Y ~ X1, data = simpan_data, control = tree.control(nobs = 6,
##     mincut = 1, minsize = 2), split = c("gini"))
## Number of terminal nodes:  3
## Residual mean deviance:  0 = 0 / 3
## Misclassification error rate: 0 = 0 / 6
```

```
print(tree1)
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
## 1) root 6 7.638 A ( 0.6667 0.3333 )
##   2) X1 < 6 3 0.000 A ( 1.0000 0.0000 ) *
##   3) X1 > 6 3 3.819 B ( 0.3333 0.6667 )
##     6) X1 < 10.5 2 0.000 B ( 0.0000 1.0000 ) *
##     7) X1 > 10.5 1 0.000 A ( 1.0000 0.0000 ) *
```

Gambar 16.15

```
plot(tree1)
text(tree1)
```



Gambar 16.16

Membuat Pohon Klasifikasi dengan Dua Variabel Bebas Continuous, Kriteria Pemecah GINI, dengan Metode Midpoints (Contoh Perhitungan dan Penyelesaian R)

Misalkan diberikan data seperti pada Tabel 16.9. Berdasarkan data pada Tabel 16.9, terdapat satu variabel tak bebas (Y) dan dua variabel bebas (X_1 dan X_2). Diketahui **terdapat dua kategori pada variabel tak bebas**, yakni A dan B.

Tabel 16.9

Y	X_1	X_2
A	3	25
A	1	27
A	2	4
A	3	1
B	9	20
B	8	24
B	10	23
B	14	21
A	13	3
A	12	13

Berikut akan dibentuk pohon klasifikasi berdasarkan kriteria pemecah GINI, dengan metode *midpoints*. Pertama, akan dihitung GINI *index* dan GINI *splitting index* dengan metode *midpoints* pada variabel X_1 di node akar. Berikut disajikan data untuk variabel Y dan X_1 (Tabel 16.10).

Tabel 16.10

Y	X_1
A	3
A	1
A	2
A	3
B	9
B	8
B	10
B	14
A	13
A	12

Data pada Tabel 16.10 diurutkan, seperti pada Tabel 16.11.

Tabel 16.11

Y	X_1
A	1
A	2
A	3
A	3
B	8
B	9
B	10
A	12
A	13
B	14

$$I_{GINI}(X_1 \leq 1{,}5) = 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 1{,}5) = 1 - \left(\left(\frac{5}{9}\right)^2 + \left(\frac{4}{9}\right)^2\right) = 1 - 0{,}506173 = 0{,}493872$$

$$GINI_{split} = \left(\frac{1}{10}\right)(0) + \left(\frac{9}{10}\right)(0{,}493872) = 0{,}4444$$

$$I_{GINI}(X_1 \leq 2{,}5) = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 2{,}5) = 1 - \left(\left(\frac{4}{8}\right)^2 + \left(\frac{4}{8}\right)^2\right) = 1 - 0{,}5 = 0{,}5$$

$$GINI_{split} = \left(\frac{2}{10}\right)(0) + \left(\frac{8}{10}\right)(0{,}5) = 0{,}4$$

$$I_{GINI}(X_1 \leq 3) = 1 - \left(\left(\frac{4}{4}\right)^2 + \left(\frac{0}{4}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 3) = 1 - \left(\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right) = 1 - 0{,}6944 = 0{,}3055$$

$$GINI_{split} = \left(\frac{4}{10}\right)(0) + \left(\frac{6}{10}\right)(0{,}3055) = 0{,}1833$$

Perhatikan bahwa seandainya aturan pengelompokannya diubah menjadi $I_{GINI}(X_1 < 3)$ dan $I_{GINI}(X_1 \geq 3)$, maka diperoleh

$$I_{GINI}(X_1 < 3) = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 \geq 3) = 1 - \left(\left(\frac{4}{8}\right)^2 + \left(\frac{4}{8}\right)^2\right) = 1 - 0{,}5 = 0{,}5$$

$$GINI_{split} = \left(\frac{2}{10}\right)(0) + \left(\frac{8}{10}\right)(0{,}5) = 0{,}4$$

Dalam penggunaan *software* R untuk fungsi **rpart**, aturan pengelompokannya adalah $I_{GINI}(X_1 < 3)$ dan $I_{GINI}(X_1 \geq 3)$.

$$I_{GINI}(X_1 < 3) = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 3) = 1 - \left(\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right) = 1 - 0{,}55 = 0{,}44$$

$$GINI_{split} = \left(\frac{2}{8}\right)(0) + \left(\frac{6}{8}\right)(0{,}44) = 0{,}33$$

Dalam penggunaan *software* R untuk fungsi **tree**, aturan pengelompokannya adalah $I_{GINI}(X_1 < 3)$ dan $I_{GINI}(X_1 > 3)$.

$$I_{GINI}(X_1 \leq 5{,}5) = 1 - \left(\left(\frac{4}{4}\right)^2 + \left(\frac{0}{4}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 5{,}5) = 1 - \left(\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right) = 1 - 0{,}6944 = 0{,}3055$$

$$GINI_{split} = \left(\frac{4}{10}\right)(0) + \left(\frac{6}{10}\right)(0{,}3055) = 0{,}1833$$

$$I_{GINI}(X_1 \leq 8{,}5) = 1 - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right) = 1 - 0{,}68 = 0{,}32$$

$$I_{GINI}(X_1 > 8{,}5) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right) = 1 - 0{,}52 = 0{,}48$$

$$GINI_{split} = \left(\frac{5}{10}\right)(0{,}32) + \left(\frac{5}{10}\right)(0{,}48) = 0{,}4$$

$$I_{GINI}(X_1 \leq 9{,}5) = \cdots (silahkan\ hitung)$$
$$I_{GINI}(X_1 > 9{,}5) = \cdots$$
$$GINI_{split} = \cdots$$

$$I_{GINI}(X_1 \leq 11) = \cdots$$
$$I_{GINI}(X_1 > 11) = \cdots$$
$$GINI_{split} = \cdots$$

$$I_{GINI}(X_1 \leq 12{,}5) = \cdots$$
$$I_{GINI}(X_1 > 12{,}5) = \cdots$$
$$GINI_{split} = \cdots$$

$$I_{GINI}(X_1 \leq 13{,}5) = \cdots$$
$$I_{GINI}(X_1 > 13{,}5) = \cdots$$
$$GINI_{split} = \cdots$$

Berdasarkan perhitungan di atas, diketahui nilai GINI *split index* terkecil berada pada nilai $X_1 = 5{,}5$, yakni dengan nilai GINI *split index* 0,1833. Selanjutnya, menghitung GINI *index* dan GINI *splitting index* dengan metode *midpoints* pada variabel X_2 di node akar.

Berikut disajikan data untuk variabel Y dan X_2 (Tabel 16.12).

Tabel 16.12

Y	X_2
A	25
A	27
A	4
A	1
B	20
B	24
B	23
B	21
A	3
A	13

Data pada Tabel 16.12 diurutkan, seperti pada Tabel 16.13.

Tabel 16.13

Y	X_2
A	1
A	3
A	4
A	13
B	20
B	21
B	23
B	24
A	25
A	27

$$I_{GINI}(X_2 \leq 2) = 1 - \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_2 > 2) = 1 - \left(\left(\frac{5}{9}\right)^2 + \left(\frac{4}{9}\right)^2\right) = 1 - 0{,}506173 = 0{,}493872$$

$$GINI_{split} = \left(\frac{1}{10}\right)(0) + \left(\frac{9}{10}\right)(0{,}493872) = 0{,}4444$$

$$I_{GINI}(X_2 \leq 3{,}5) = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_2 > 3{,}5) = 1 - \left(\left(\frac{4}{8}\right)^2 + \left(\frac{4}{8}\right)^2\right) = 1 - 0{,}5 = 0{,}5$$

$$GINI_{split} = \left(\frac{2}{10}\right)(0) + \left(\frac{8}{10}\right)(0{,}5) = 0{,}4$$

$$I_{GINI}(X_2 \leq 8{,}5) = 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_2 > 8{,}5) = 1 - \left(\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right) = 1 - 0{,}5102 = 0{,}4898$$
$$GINI_{split} = \left(\frac{3}{10}\right)(0) + \left(\frac{7}{10}\right)(0{,}4898) = 0{,}34286$$

$$I_{GINI}(X_2 \leq 16{,}5) = 1 - \left(\left(\frac{4}{4}\right)^2 + \left(\frac{0}{4}\right)^2\right) = 1 - 1 = 0$$
$$I_{GINI}(X_2 > 16{,}5) = 1 - \left(\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2\right) = 1 - 0{,}55555 = 0{,}4444$$
$$GINI_{split} = \left(\frac{4}{10}\right)(0) + \left(\frac{6}{10}\right)(0{,}44444) = 0{,}2667$$

$$I_{GINI}(X_2 \leq 20{,}5) = 1 - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right) = 1 - 0{,}68 = 0{,}32$$
$$I_{GINI}(X_2 > 20{,}5) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right) = 1 - 0{,}52 = 0{,}48$$
$$GINI_{split} = \left(\frac{5}{10}\right)(0{,}32) + \left(\frac{5}{10}\right)(0{,}48) = 0{,}4$$

$$I_{GINI}(X_2 \leq 22) = \cdots (silahkan\ hitung)$$
$$I_{GINI}(X_2 > 22) = \cdots$$
$$GINI_{split} = \cdots$$

$$I_{GINI}(X_2 \leq 23{,}5) = \cdots$$
$$I_{GINI}(X_2 > 23{,}5) = \cdots$$
$$GINI_{split} = \cdots$$

$$I_{GINI}(X_2 \leq 24{,}5) = \cdots$$
$$I_{GINI}(X_2 > 24{,}5) = \cdots$$
$$GINI_{split} = \cdots$$

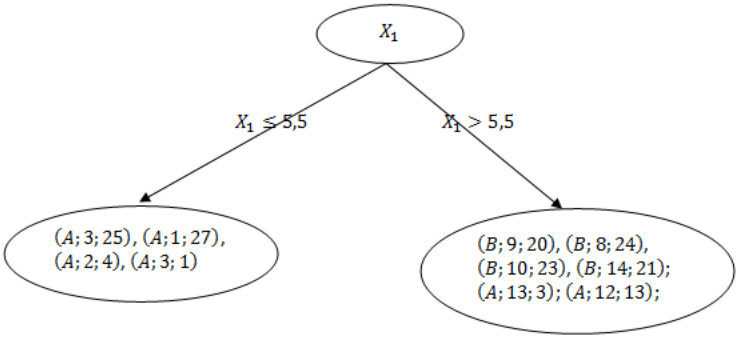
$$I_{GINI}(X_2 \leq 26) = \cdots$$
$$I_{GINI}(X_2 > 26) = \cdots$$
$$GINI_{split} = \cdots$$

Berdasarkan perhitungan di atas, diketahui nilai GINI *split index* terkecil berada pada nilai $X_2 = 16{,}5$, yakni dengan nilai GINI *split index* 0,2667.

Tabel 16.14

Pengelompokkan	$X_1 = 5{,}5$	$X_2 = 16{,}5$
Gini split index	0,1833 $(minimum)$	0,2667

Berdasarkan Tabel 16.13, maka variabel X_1 bertindak sebagai **node akar**. Perhatikan pohon klasifikasi berikut (Gambar 16.17).



Gambar 16.17

Tabel 16.15

Y	X_1	X_2
B	9	20
B	8	24
B	10	23
B	14	21
A	13	3
A	12	13

Selanjutnya, menghitung GINI *index* dan GINI *splitting index* dengan metode *midpoints* pada variabel X_1 di node internal (berdasarkan data pada Tabel 16.15). Berikut disajikan data untuk variabel Y dan X_1, setelah diurutkan.

Tabel 16.16

Y	X_1
B	8
B	9
B	10
A	12
A	13
B	14

$$I_{GINI}(X_1 \leq 8{,}5) = 1 - \left(\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 8{,}5) = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right) = 1 - 0{,}52 = 0{,}48$$

$$GINI_{split} = \left(\frac{1}{6}\right)(0) + \left(\frac{5}{6}\right)(0{,}48) = 0{,}15275$$

$$I_{GINI}(X_1 \leq 9{,}5) = 1 - \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right) = 1 - 1 = 0$$

$$I_{GINI}(X_1 > 9{,}5) = 1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right) = 1 - 0{,}5 = 0{,}5$$

$$GINI_{split} = \left(\frac{2}{6}\right)(0) + \left(\frac{4}{6}\right)(0{,}5) = 0{,}333$$

$$I_{GINI}(X_1 \leq 11) = 1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right) = 1 - 1 = 0$$
$$I_{GINI}(X_1 > 11) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) = 1 - 0{,}5555 = 0{,}4444$$
$$GINI_{split} = \left(\frac{3}{6}\right)(0) + \left(\frac{3}{6}\right)(0{,}4444) = 0{,}2222$$

$$I_{GINI}(X_1 \leq 12{,}5) = \cdots (silahkan\ hitung)$$
$$I_{GINI}(X_1 > 12{,}5) = \cdots$$
$$GINI_{split} = \cdots$$

$$I_{GINI}(X_1 \leq 13{,}5) = \cdots$$
$$I_{GINI}(X_1 > 13{,}5) = \cdots$$
$$GINI_{split} = \cdots$$

Berdasarkan perhitungan di atas, diketahui nilai GINI *split index* terkecil berada pada nilai $X_1 = 11$, yakni dengan nilai *GINI split index* 0,2222. Selanjutnya menghitung GINI *index* dan GINI *splitting index* dengan metode *midpoints* pada Variabel X_2 di node internal. Berikut disajikan data untuk variabel Y dan X_2 setelah diurutkan (Tabel 16.17).

Tabel 16.17

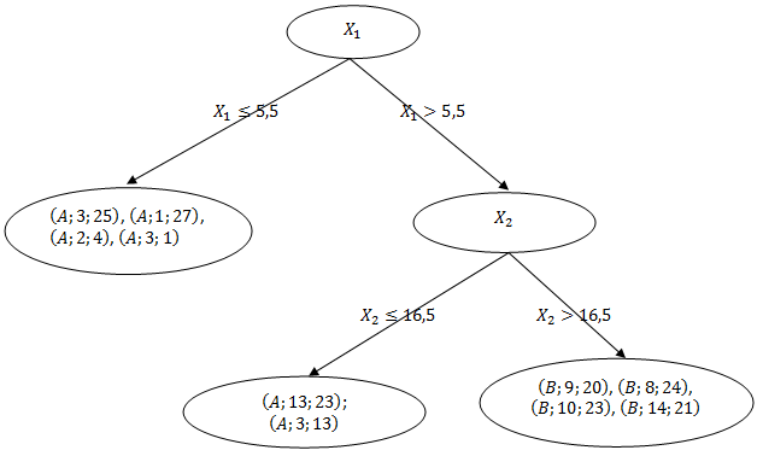
Y	X_2
A	3
A	13
B	20
B	21
B	23
B	24

Berdasarkan data pada Tabel 16.17, nilai GINI *split index* terkecil berada pada nilai $X_2 = 16{,}5$, yakni dengan nilai GINI *split index* sebagai berikut.

$$I_{GINI}(X_1 \leq 16{,}5) = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right) = 1 - 1 = 0$$
$$I_{GINI}(X_1 > 16{,}5) = 1 - \left(\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2\right) = 1 - 1 = 0$$
$$GINI_{split} = 0$$

Tabel 16.18

Pengelompokkan	$X_1 = 11$	$X_2 = 16{,}5$
Gini split index	0,222	0 $(minimum)$



Gambar 16.18

Gambar 16.19 menyajikan kode R untuk membentuk pohon klasifikasi. Gambar 16.20 hingga Gambar 16.24 merupakan hasil eksekusi kode R pada Gambar 16.19.

RR.R | R.R | code-ipaconf2016.R | kode_R_sedikit_modifikasi.R

Source on Save | Run

```
1  simpan_data=read.csv("data3.csv")
2  simpan_data
3
4  library(rpart)
5  tree <- rpart(Y ~ X1+X2, simpan_data, minsplit=1)
6  summary(tree)
7  print(tree)
8
9
10 library(rpart.plot)
11 prp(tree, faclen = 0, cex = 0.8, extra = 1, digits = 4)
12
13 treePrediction <- predict(tree, simpan_data, type = "class")
14 treePrediction
15
16 library(caret)
17 confusionMatrix(treePrediction, simpan_data$Y)
18
19 #Berikut akan digunakan R untuk membuat pohon klasifikasi dengan package Tree.
20
21 library(tree)
22 tree1 <- tree(Y ~ X1+X2,simpan_data, split = c("gini"), control=tree.control(nobs=10, mincut = 1, minsize = 2))
23 summary(tree1)
24 print(tree1)
25 plot(tree1)
26 text(tree1)
```

Gambar 16.19

```
simpan_data=read.csv("data3.csv")
simpan_data
```

```
##    Y X1 X2
## 1  A  3 25
## 2  A  1 27
## 3  A  2  4
## 4  A  3  1
## 5  B  9 20
## 6  B  8 24
## 7  B 10 23
## 8  B 14 21
## 9  A 13  3
## 10 A 12 13
```

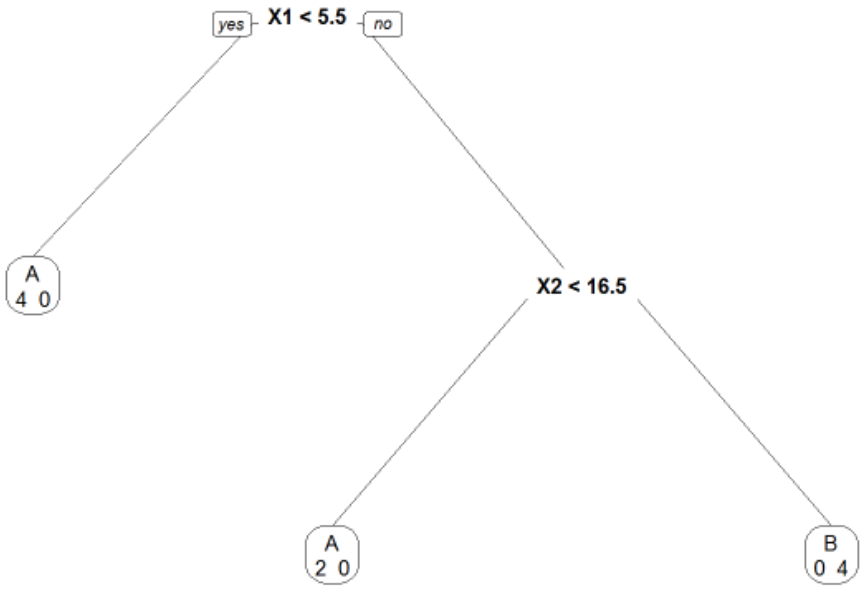
```
print(tree)
```

```
## n= 10
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 10 4 A (0.6000000 0.4000000)
##   2) X1< 5.5 4 0 A (1.0000000 0.0000000) *
##   3) X1>=5.5 6 2 B (0.3333333 0.6666667)
##     6) X2< 16.5 2 0 A (1.0000000 0.0000000) *
##     7) X2>=16.5 4 0 B (0.0000000 1.0000000) *
```

```
library(rpart.plot)
prp(tree, faclen = 0, cex = 0.8, extra = 1, digits = 4)

treePrediction <- predict(tree, simpan_data, type = "class")
treePrediction
```

```
##  1  2  3  4  5  6  7  8  9 10
##  A  A  A  A  B  B  B  B  A  A
## Levels: A B
```

Gambar 16.20



Gambar 16.21

```
confusionMatrix(treePrediction, simpan_data$Y)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction A B
##          A 6 0
##          B 0 4
```

Gambar 16.22

```
library(tree)
tree1 <- tree(Y ~ X1+X2,simpan_data, split = c("gini"), control=tree.control(nobs=10, mincut = 1, minsize = 2))
summary(tree1)
```

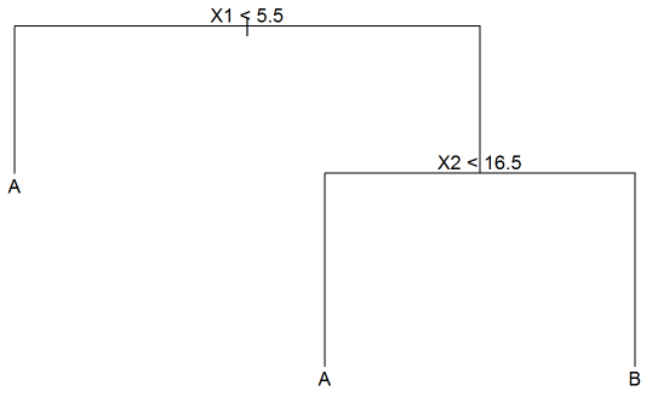
```
##
## Classification tree:
## tree(formula = Y ~ X1 + X2, data = simpan_data, control = tree.control(nobs = 10,
##     mincut = 1, minsize = 2), split = c("gini"))
## Number of terminal nodes:  3
## Residual mean deviance:  0 = 0 / 7
## Misclassification error rate: 0 = 0 / 10
```

```
print(tree1)
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
## 1) root 10 13.460 A ( 0.6000 0.4000 )
##   2) X1 < 5.5 4  0.000 A ( 1.0000 0.0000 ) *
##   3) X1 > 5.5 6  7.638 B ( 0.3333 0.6667 )
##     6) X2 < 16.5 2  0.000 A ( 1.0000 0.0000 ) *
##     7) X2 > 16.5 4  0.000 B ( 0.0000 1.0000 ) *
```

Gambar 16.23

```
plot(tree1)
text(tree1)
```



Gambar 16.24

Referensi

1. Bramer, Max. 2007. *Principles of Data Mining*. Springer.
2. Gorunescu, Florin. 2011. *Data Mining, Concepts, Models, and Techniques*. Springer.
3. Hermawati, F.A. 2013. *Data Mining*. Penerbit Andi.
4. Prasetyo, Eko. 2014. *Data Mining*, Mengolah Data Menjadi Informasi Menggunakan Matlab. Penerbit Andi.
5. https://cran.r-project.org/web/packages/rpart/rpart.pdf
6. https://cran.r-project.org/web/packages/tree/tree.pdf
7. https://cran.r-project.org/web/packages/rpart.plot/rpart.plot.pdf
8. https://cran.r-project.org/web/packages/caret/caret.pdf
9. http://www.milbo.org/rpart-plot/prp.pdf
10. http://www.r-bloggers.com/draw-nicer-classification-and-regression-trees-with-the-rpart-plot-package/
11. https://rpubs.com/minma/cart_with_rpart