

Statistika Lingkungan Menggunakan R

Moh. Rosidi

2019-05-20

Contents

Pengantar	15
Bahasa Pemrograman R	19
1 Mengenal Bahasa R	19
1.1 Sejarah R	19
1.2 Fitur dan Karakteristik R	20
1.3 Kelebihan dan Kekurangan R	20
1.4 RStudio	21
1.5 Menginstall R dan RStudio	21
1.6 Working Directory	22
1.7 Fasilitas Help	23
1.8 Referensi	28
2 Sintaks Bahasa R	29
2.1 Operator Aritmatika	29
2.2 Fungsi Aritmetik	30
2.3 Operator Relasi	31
2.4 Operator Logika	32
2.5 Memasukkan Nilai Kedalam Variabel	34
2.6 Tipe Data	35
2.7 Vektor	38
2.8 Matriks	42
2.9 Faktor	48
2.10 Data Frames	50
2.11 List	53
2.12 Loop	56
2.13 Decision Making	59
2.14 Fungsi	61
2.15 Referensi	63

3 Manajemen Data R	65
3.1 Import File	65
3.2 Ekspor File	70
3.3 Tibble Data Format	73
3.4 Merapikan Data	79
3.5 Transformasi Data	87
3.6 Referensi	101
 Visualisasi Data - R	 105
4 Visualisasi Data Menggunakan Fungsi Dasar R	105
4.1 Visualisasi Data Menggunakan Fungsi plot()	105
4.2 Matriks Scatterplot	108
4.3 Box plot	113
4.4 Bar Plot	117
4.5 Line Plot	118
4.6 Pie Chart	120
4.7 Histogram dan Density Plot	121
4.8 QQ Plot	123
4.9 Dot Chart	124
4.10 Kustomisasi Parameter Grafik	125
4.11 Alternatif Library Dasar Lain	143
4.12 Referensi	148
 5 Visualisasi Data Menggunakan GGPLOT	 149
5.1 Scatterplot	150
5.2 Box Plot dan Violin Plot	155
5.3 Bar Plot	158
5.4 Line Plot	162
5.5 Pie Chart	163
5.6 Histogram dan Desity Plot	164
5.7 QQ Plot	166
5.8 Dot Plot	168
5.9 ECDF Plot	170
5.10 Parameter Grafik	170
5.11 Referensi	208

Statistika Deskriptif - R	213
6 Ringkasan Numerik	213
6.1 Ukuran Pemusatan Data	213
6.2 Ukuran Sebaran Data	221
6.3 Ringkasan Data Menggunakan Fungsi summary() dan stat.desc()	224
6.4 Ukuran Kemencenggan Data	225
6.5 Outlier	227
6.6 Transformasi Data	228
6.7 Referensi	230
7 Ekplorasi Data Menggunakan Grafik	233
7.1 Grafik Untuk Melihat Distrobusi Data	233
7.2 Grafik Untuk Melihat Beda Distribusi Data Antar Grup	238
7.3 Grafik Untuk Memvisualisasikan Korelasi Antar Variabel	240
7.4 Grafik Yang Digunakan Untuk Memvisualisasikan Asosiasi Antar Variabel	241
7.5 Grafik Yang Digunakan Untuk Memvisualisasikan Ukuran Sampel dan Perubahan Sepanjang Waktu	242
7.6 Referensi	243
Probabilitas dan Distribusi Probabilitas	247
8 Probabilitas	247
8.1 Aturan Dasar Probabilitas	248
8.2 Teori Bayes	255
8.3 Ekspektasi Matematis	256
8.4 Referensi	257
9 Distribusi Probabilitas	259
9.1 Properti Umum dari Distribusi Probabilitas	259
9.2 Distribusi Binomial dan Multinomial	260
9.3 Distribusi Hipergeometris	264
9.4 Distribusi Binomial Negatif dan Distribusi Geometris	267
9.5 Distribusi Poisson	271
9.6 Distribusi Uniform	272
9.7 Distribusi Normal	274
9.8 Distribusi Gamma dan Eksponensial	282
9.9 Distribusi Chi-Square, Student's t, dan Snedecor's F	286
9.10 Distribusi Kontinu Lainnya	288
9.11 Referensi	294

Statistika Inferensi - R	299
10 Penaksiran Secara Statistika	299
10.1 Definisi Interval Estimasi	299
10.2 Interpretasi Interval Estimasi	300
10.3 Interval Kepercayaan Median	301
10.4 Interval Kepercayaan Mean	311
10.5 Interval Prediksi Nonparametrik	316
10.6 Interval Prediksi Parametrik	320
10.7 Interval Kepercayaan Persentil (Interval Toleransi)	324
10.8 Interval Kepercayaan Menggunakan Metode Bootstrap	331
10.9 Kegunaan Lain Dari Interval Kepercayaan	337
10.10 Referensi	339

List of Tables

2.2	Operator Relasi R	32
2.3	Operator logika R	33
2.4	Tipe Data R	36
2.5	Daftar percabangan pada R	59
5.1	20 observasi pertama dataset gapminder	151
6.1	Data Debit Sampel (m ³ /detik)	215
6.2	Kosentrasi TDS dan Uranium dalam berbagai kondisi kesadahan	231
8.1	Populasi orang yang telah menyelesaikan masa studinya di suatu kota.	253
10.1	Sepuluh interval kepercayaan 90% sekitar nilai mean sebenarnya sebesar 10 (Data berdistribusi normal dan Tanda plus menyatakan data tidak disertakan dalam nilai mean sebenarnya)	300
10.2	Sepuluh interval kepercayaan 90% sekitar nilai mean sebenarnya sebesar 1 (Data tidak berdistribusi normal dan Tanda plus menyatakan data tidak disertakan dalam nilai mean sebenarnya)	301
10.3	Konsentrasi Arsenik dalam air tanah (ppb)	304
10.4	Tranformasi logaritmik konsentrasi Arsenik dalam air tanah (ppb)	309

List of Figures

1.1	Logo R.	20
1.2	Jendela R.	22
1.3	Jendela RStudio.	23
1.4	Mengubah working directory.	24
1.5	Merubah working directory melalui Global options.	24
1.6	Jendela help dokumentasi fungsi mean().	25
1.7	Jendela general help dokumentasi fungsi mean().	27
1.8	Jendela help search dokumentasi fungsi mean().	28
2.1	Diagram umum loop (sumber: Primartha, 2018).	56
2.2	Diagram if statement (sumber: Primartha, 2018).	60
2.3	Diagram if else statement (sumber: Primartha, 2018).	61
2.4	Diagram switch statement (sumber: Primartha, 2018).	62
3.1	Visualisasi 3 rule tidy data	80
3.2	Diagram operasi Boolean	89
3.3	Jarak vs rata-rata delay	100
4.1	Plot berbagai jenis setting type	106
4.2	Scatterplot Height vs Volume	107
4.3	Matriks scatterplot dataset trees	108
4.4	Plot diagnostik regresi linier	109
4.5	Matriks scatterplot iris	110
4.6	Matriks scatterplot iris tanpa panel bawah	111
4.7	Matriks scatterplot iris tanpa panel bawah	112
4.8	Matriks scatterplot iris dengan koefisien korelasi	113
4.9	Matriks scatterplot iris dengan koefisien korelasi di panel atas	114
4.10	Boxplot variabel Sepal.Length	114
4.11	Boxplot berdasarkan variabel species	115

4.12 Boxplot dengan warna berdasarkan spesies	116
4.13 Boxplot multiple group	117
4.14 a. bar plot vertikal; b. bar plot horizontal	118
4.15 Kustomisasi bar plot	119
4.16 Stacked bar plot	119
4.17 Grouped bar plot	120
4.18 Line plot	121
4.19 Pie chart	122
4.20 Histogram	122
4.21 Density plot	123
4.22 Density plot dan histogram	124
4.23 QQ plot	125
4.24 Dot chart	126
4.25 Menambahkan Judul	127
4.26 Menambahkan Judul (2)	128
4.27 Menambahkan Judul (3)	129
4.28 Menambahkan legend	130
4.29 Menambahkan legend (2)	131
4.30 Menambahkan legend (3)	132
4.31 Kustomisasi posisi legend	133
4.32 Menambahkan teks	134
4.33 Menambahkan teks (2)	135
4.34 Menambahkan teks (3)	135
4.35 Menambahkan garis	136
4.36 Symbol plot	138
4.37 Line type	139
4.38 Menambahkan axis	140
4.39 Mengubah rentang dan skala axis	141
4.40 Kustomisasi tick mark	142
4.41 Nama warna	143
4.42 Enhanced scatterplot	144
4.43 Enhanced scatterplot matrices	145
4.44 Enhanced box plot	146
4.45 Enhanced qq plot	147
4.46 Plot group means	147
5.1 Scatterplot lifeExp vs gdpPerCap	152

5.2 Scatterplot lifeExp vs gdpPercap tiap benua (1)	152
5.3 Scatterplot lifeExp vs gdpPercap tiap benua (2)	153
5.4 Scatterplot lifeExp vs gdpPercap dan populasi tiap negara dan benua	153
5.5 Scatterplot lifeExp vs gdpPercap dengan garis penghalusan regresi linier	154
5.6 Box plot variabel lifeExp	155
5.7 Box plot variabel lifeExp pada tiap continent	156
5.8 Box plot variabel lifeExp pada tiap continent (1952 dan 2007)	157
5.9 Box plot variabel lifeExp Benua Asia	157
5.10 Violin plot variabel lifeExp pada masing-masing benua	158
5.11 Violin plot variabel lifeExp pada masing-masing benua (2)	159
5.12 Bar plot rata-rata lifeExp masing-masing benua	160
5.13 Bar plot rata-rata lifeExp masing-masing benua dengan confidence interval	161
5.14 Bar plot rata-rata lifeExp masing-masing benua (1952 dan 2007) dengan confidence interval .	161
5.15 Line plot lifeExp masing-masing benua	162
5.16 Histogram lifeExp	163
5.17 Pie chart pop	164
5.18 Histogram lifeExp	165
5.19 Histogram lifeExp berdasarkan benua	165
5.20 Density plot lifeExp	166
5.21 Density plot lifeExp berdasarkan benua	167
5.22 histogram dan density plot lifeExp	167
5.23 QQ plot variabel lifeExp	168
5.24 Dot plot variabel lifeExp masing-masing benua (1952-2007)	169
5.25 Dot plot variabel lifeExp masing-masing benua (1952-2007) (2)	169
5.26 ECDF plot variabel lifeExp	170
5.27 Mengubah judul grafik dan keterangan axis	171
5.28 Mengubah keterangan legend pada grafik	172
5.29 Kustomisasi judul grafik dan keterangan axis	173
5.30 Kustomisasi posisi legend berdasarkan karakter	174
5.31 Kustomisasi posisi legend berdasarkan vektor numerik	175
5.32 Kustomisasi tampilan legend	176
5.33 Menghilangkan seluruh legend	177
5.34 Menghilangkan sebagian legend legend	178
5.35 Merubah warna grup berdasarkan satu warna	178
5.36 Merubah warna grup secara otomatis	179
5.37 Merubah pencahayaan dan intensitas warna	180

5.38 Merubah warna secara manual	180
5.39 Palet warna RColorBrewer	181
5.40 Merubah warna menggunakan palet	182
5.41 Merubah warna menggunakan palet gray	183
5.42 Kustomisasi jenis, ukuran dan warna titik	183
5.43 Kustomisasi jenis, ukuran dan warna titik untuk multiple group secara otomatis	184
5.44 Kustomisasi jenis, ukuran dan warna titik untuk multiple group secara manual	185
5.45 Kustomisasi jenis, ukuran dan warna garis	186
5.46 Kustomisasi jenis, ukuran dan warna garis untuk multiple group secara otomatis	186
5.47 Kustomisasi jenis, ukuran dan warna garis untuk multiple group secara manual	187
5.48 Scatterplot variabel pop vs gdpPercap	188
5.49 Scatterplot variabel pop vs gdpPercap dengan label	189
5.50 Scatterplot variabel pop vs gdpPercap dengan label dan notasi	189
5.51 Scatterplot variabel pop vs gdpPercap dengan label dan notasi pada tiap panel	190
5.52 Scatterplot dengan tema black and white	191
5.53 Scatterplot dengan tema Wall Street Journal	192
5.54 Scatterplot dengan axis limits	194
5.55 Scatterplot dengan axis limits (2)	195
5.56 Scatterplot dengan transformasi axis	196
5.57 Scatterplot dengan transformasi tick mark axis	197
5.58 Mengubah tampilan dari tick mark	198
5.59 Menyembunyikan tampilan dari tick mark	199
5.60 Kustomisasi tampilan dari garis axis	200
5.61 Kustomisasi tick mark	200
5.62 Penerapan vline	201
5.63 Penerapan hline	202
5.64 Penerapan abline	203
5.65 Penerapan garis segmen	204
5.66 Rotasi axis	205
5.67 Pembalikan sumbu y	206
5.68 Facet horizontal satu variabel	206
5.69 Facet vertikal satu variabel	207
5.70 Facet dua variabel	208
5.71 Facet dua variabel dengan skala bebas pada sumbu y	209
6.1 Nilai mean (segitiga) sebagai titik kesetimbangan pada data.	214
6.2 Pergeseran nilai mean (segitiga) ke kiri setelah penghilangan outlier.	214

6.3	Visualisasi debit sungai pada sampel	216
6.4	Visualisasi konsentrasi TDS pada air tanah	217
6.5	Visualisasi konsentrasi Uranium pada air tanah	218
6.6	Jendela diagram trimmed mean.	220
6.7	a) Kemencengan negatif, b) Kemencengan positif.	226
6.8	Box plot untuk data dengan a) Kemencengan negatif, b) Kemencengan positif.	226
6.9	Ladder of power	229
6.10	Visualisasi konsentrasi Uranium hasil transformasi pada air tanah	230
7.1	Scatterplot dengan koefisien korelasi $r=0,7$	234
7.2	Histogram dengan bin.width=default debit sungai Saddle	235
7.3	Histogram dengan bin.width=500 debit sungai Saddle	236
7.4	Density plot debit sungai Saddle	237
7.5	QQ plot debit sungai Saddle	237
7.6	Box plot dan violin plot debit sungai Saddle	238
7.7	Box plot konsentrasi Atrazine pada bulan Juni dan September	240
7.8	Bar plot konsentrasi Atrazine pada bulan Juni dan September	241
7.9	Scatterplot hubungan antara konsentrasi TDS dan Uranium pada airtanah	242
7.10	Bar plot Jumlah rata-rata corbicula pada sungai Tennessee	243
7.11	Line plot perubahan jumlah rata-rata corbicula di sungai Tennessee	244
8.1	Diagram venn peristiwa mutually exclusive	250
8.2	Diagram venn peristiwa not mutually exclusive	252
9.1	Distribusi uniform dengan nilai min 1 dan max 3	273
9.2	Probabilitas distribusi uniform pada rentang nilai x 3 sampai 4	274
9.3	Distribusi normal dengan nilai mean sama dan simpangan baku berbeda.	275
9.4	Distribusi normal dengan nilai mean sama dan simpangan baku berbeda.	276
9.5	Distribusi normal dengan nilai mean berbeda dan simpangan baku berbeda.	276
9.6	Luas area di bawah kurva normal.	277
9.7	Luas area masa layan lampu antara 750 sampai 830 jam.	278
9.8	Luas area masa layan lampu lebih dari atau sama dengan 830 jam.	279
9.9	Visualisasi distribusi konsentrasi ozon Kota New York a)density plot, b)boxplot, c)ecdf, d)qq-plot	281
9.10	Luas area jumlah pompa rusak kurang dari 30.	283
9.11	Visualisasi distribusi gamma dengan variasi alpha dengan beta 1 a) density plot, b)ecdf . . .	284
9.12	Visualisasi distribusi chi-square dengan variasi derajat kebebasan a) density plot, b)ecdf . . .	287
9.13	Visualisasi distribusi t dengan variasi derajat kebebasan a) density plot, b)ecdf	288

9.14 Visualisasi distribusi F dengan variasi derajat kebebasan a) density plot, b)ecdf	289
9.15 Visualisasi distribusi beta dengan variasi derajat kebebasan a) density plot, b)ecdf	290
9.16 Visualisasi distribusi beta dengan variasi mean dan sd a) density plot, b)ecdf	291
9.17 Visualisasi distribusi t dengan variasi m dan beta a) density plot, b)ecdf	292
9.18 Visualisasi distribusi logistik dengan variasi mean dan simpangan baku, a) density plot, b)ecdf	293
9.19 Visualisasi distribusi weibull dengan variasi alpha dengan beta 1 a) density plot, b)ecdf . . .	294
 10.1 Sepuluh interval kepercayaan 90 persen data dengan nilai mean sebenarnya 10 (Helsel dan Hirsch, 2002)	301
10.2 Histogram data dengan nilai mean populasi 1 dan simpangan baku populasi 0.75 (Helsel dan Hirsch, 2002)	302
10.3 Sepuluh interval kepercayaan 90 persen data dengan nilai mean sebenarnya (Helsel dan Hirsch, 2002)	302
10.4 Probabilitas median populasi P50 pada dua sisi interval estimasi (Helsel dan Hirsch, 2002) . .	303
10.5 Distribusi konsentrasi arsenik dalam air tanah	303
10.6 Lokasi probabilitas x berdasarkan tabel distribusi binomial	305
10.7 Distribusi logaritmik konsentrasi arsenik dalam air tanah	310
10.8 Prediksi interval dua sisi (Helsel dan Hirsch, 2002)	317
10.9 Prediksi interval satu sisi (Helsel dan Hirsch, 2002)	318
10.10 Interval estimasi persentil X_p sebagai pengujii apakah $X_p = X_0$. A) X_0 didalam interval estimasi sehingga X_p tidak berbeda secara signifikan dari X_0 , B) X_0 berada diluar rentang estimasi sehingga X_p berbeda secara signifikan dari X_0 . (Helsel dan Hirsch, 2002)	328
10.11 Interval estimasi persentil X_p sebagai pengujii apakah $X_p > X_0$. A) X_0 didalam interval estimasi sehingga X_p tidak signifikan lebih besar dari X_0 , B) X_0 berada diluar rentang estimasi sehingga X_p signifikan lebih besar dari X_0 . (Helsel dan Hirsch, 2002)	328
10.12 Interval estimasi persentil X_p sebagai pengujii apakah $X_p > X_0$. A) X_0 didalam interval estimasi sehingga X_p tidak signifikan lebih kecil dari X_0 , B) X_0 berada diluar rentang estimasi sehingga X_p signifikan lebih kecil dari X_0 . (Helsel dan Hirsch, 2002)	329
10.13 Distribusi bootstrap median	332
10.14 Distribusi bootstrap mean	334
10.15 Distribusi bootstrap persentil 90	336

Pengantar

Buku ini menyajikan penerapan program R dalam **Statistika Lingkungan**. Buku ini akan disajikan secara ringkas menggunakan sejumlah contoh kasus yang relevan dalam bidang lingkungan.

Penulis berharap buku ini dapat menjadi referensi sumber terbuka bagi mahasiswa yang ingin menggunakan R untuk kegiatan analisa data. Sehingga dapat mengurangi ketergantungan pada penggunaan aplikasi yang berlisensi.

Bahasa Pemrograman R

Chapter 1

Mengenal Bahasa R

Dewasa ini tersedia banyak sekali *software* yang dapat digunakan untuk membantu kita dalam melakukan analisa data. *software* yang digunakan dapat berupa *software* berbayar atau gratis.

R merupakan salah satu *software* gratis yang sangat populer di Indonesia. Kemudahan penggunaan serta banyaknya dukungan komunitas membuat R menjadi salah satu bahasa pemrograman paling populer di dunia.

Paket yang disediakan untuk analisis statistika juga sangat lengkap dan terus bertambah setiap saat. Hal ini membuat R banyak digunakan oleh para analis data.

Pada *chapter* ini penulis akan memperkenalkan kepada pembaca mengenai bahasa pemrograman R. Mulai dari sejarah, cara instalasi sampai dengan bagaimana kita memanfaatkan fitur dasar bantuan untuk menggali lebih jauh tentang fungsi-fungsi R.

1.1 Sejarah R

R Merupakan bahasa yang digunakan dalam komputasi **statistik** yang pertama kali dikembangkan oleh **Ross Ihaka** dan **Robert Gentleman** di University of Auckland New Zealand yang merupakan akronim dari nama depan kedua pembuatnya. Sebelum R dikenal ada S yang dikembangkan oleh **John Chambers** dan rekan-rekan dari **Bell Laboratories** yang memiliki fungsi yang sama untuk komputasi statistik. Hal yang membedakan antara keduanya adalah R merupakan sistem komputasi yang bersifat gratis. Logo R dapat dilihat pada Gambar 1.1.

```
## Warning: package 'knitr' was built under R version  
## 3.5.3
```

R dapat dibilang merupakan aplikasi sistem **statistik** yang kaya. Hal ini disebabkan banyak sekali paket yang dikembangkan oleh pengembang dan komunitas untuk keperluan analisa statistik seperti *linear regression*, *clustering*, *statistical test*, dll. Selain itu, R juga dapat ditambahkan paket-paket lain yang dapat meningkatkan fiturnya.

Sebagai sebuah bahasa pemrograman yang banyak digunakan untuk keperluan analisa data, R dapat dioperasikan pada berbagai sistem operasi pada komputer. Adapun sistem operasi yang didukung antara lain: **UNIX**, **Linux**, **Windows**, dan **MacOS**.



Figure 1.1: Logo R.

1.2 Fitur dan Karakteristik R

R memiliki karakteristik yang berbeda dengan bahasa pemrograman lain seperti C++, python, dll. R memiliki aturan/sintaks yang berbeda dengan bahasa pemrograman yang lain yang membuatnya memiliki ciri khas tersendiri dibanding bahasa pemrograman yang lain.

Beberapa ciri dan fitur pada R antara lain:

- Bahasa R bersifat case sensitif.** maksudnya adalah dalam proses input R huruf besar dan kecil sangat diperhatikan. Sebagai contoh kita ingin melihat apakah objek A dan B pada sintaks berikut:

```
A <- "Andi"
B <- "andi"

# cek kedua objek A dan B
A == B

## [1] FALSE

# Kesimpulan : Kedua objek berbeda
```

- Segala sesuatu yang ada pada program R akan dianggap sebagai objek.** konsep objek ini sama dengan bahasa pemrograman berbasis objek yang lain seperti Java, C++, python, dll. Perbedaannya adalah bahasa R relatif lebih sederhana dibandingkan bahasa pemrograman berbasis objek yang lain.
- interpreted language atau script.** Bahasa R memungkinkan pengguna untuk melakukan kerja pada R tanpa perlu kompilasi kode program menjadi bahasa mesin.
- Mendukung proses **loop**, **decision making**, dan menyediakan berbagai jenis **operator** (aritmatika, logika, dll).
- Mendukung export dan import berbagai format file**, seperti:TXT, CSV, XLS, dll.
- Mudah ditingkatkan melalui penambahan fungsi atau paket.** Penambahan paket dapat dilakukan secara online melalui CRAN atau melalui sumber seperti github.
- Menyediakan berbagai fungsi untuk keperluan visualisasi data.** Visualisasi data pada R dapat menggunakan paket bawaan atau paket lain seperti ggplot2, ggviz, dll.

1.3 Kelebihan dan Kekurangan R

Selain karena R dapat digunakan secara gratis terdapat **kelebihan** lain yang ditawarkan, antara lain:

1. **Probability.** Penggunaan software dapat digunakan kapanpun tanpa terikat oleh masa berakhirnya lisensi.
2. **Multiplatform.** R bersifat *Multiplatform Operating Systems*, dimana *software* R lebih kompatibel dibanding *software* statistika lainnya. Hal ini berdampak pada kemudahan dalam penyesuaian jika pengguna harus berpindah sistem operasi karena R baik pada sistem operasi seperti **windows** akan sama pengoperasianya dengan yang ada di **Linux** (paket yang digunakan sama).
3. **General** dan **Cutting-edge**. Berbagai metode statistik baik metode klasik maupun baru telah diprogram kedalam R. Dengan demikian *software* ini dapat digunakan untuk analisis statistika dengan pendekatan klasik dan pendekatan modern.
4. **Programable.** Pengguna dapat memprogram metode baru atau mengembangkan modifikasi dari analisis statistika yang telah ada pada sistem R.
5. **Berbasis analisis matriks.** Bahasa R sangat baik digunakan untuk *programming* dengan basis matriks.
6. Fasilitas grafik yang lengkap.

Adapun kekurangan dari R antara lain:

1. **Point and Click GUI.** Interaksi utama dengan R bersifat *CLI (Command Line Interface)*, walaupun saat ini telah dikembangkan paket yang memungkinkan kita berinteraksi dengan R menggunakan *GUI (Graphical User Interface)* sederhana menggunakan paket **R-Commander** yang memiliki fungsi yang terbatas. **R- Commander** sendiri merupakan *GUI* yang diciptakan dengan tujuan untuk keperluan pengajaran sehingga analisis statistik yang disediakan adalah yang klasik. Meskipun terbatas paket ini berguna jika kita membutuhkan analisis statistik sederhana dengan cara yang simpel.
2. **Missing statistical function.** Meskipun analisis statistika dalam R sudah cukup lengkap, namun tidak semua metode statistika telah diimplementasikan ke dalam R. Namun karena R merupakan *lingua franca* untuk keperluan komputasi statistika modern saat ini, dapat dikatakan ketersediaan fungsi tambahan dalam bentuk paket hanya masalah waktu saja.

1.4 RStudio

Aplikasi R pada dasarnya berbasis teks atau *command line* sehingga pengguna harus mengetikkan perintah-perintah tertentu dan harus halal perintah-perintahnya. Setidaknya jika kita ingin melakukan kegiatan analisa data menggunakan R kita harus selalu siap dengan perintah-perintah yang hendak digunakan sehingga buku manual menjadi sesuatu yang wajib adasaat berkerja dengan R.

Kondisi ini sering kali membingungkan bagi pengguna pemula maupun pengguna mahir yang sudah terbiasa dengan aplikasi statistik lain seperti SAS, SPSS, Minitab, dll. Alasan itulah yang menyebabkan pengembang R membuat berbagai *frontend* untuk R yang berguna untuk memudahkan dalam pengoperasian R.

RStudio merupakan salah satu bentuk *frontend* R yang cukup populer dan nyaman digunakan. Selain nyaman digunakan, **RStudio** memungkinkan kita melakukan penulisan laporan menggunakan **Rmarkdown** atau **RNotebook** serta membuat berbagai bentuk project seperti **shiny**, dll. Pada **R studio** juga memungkinkan kita mengatur *working directory* tanpa perlu mengetikkan sintaks pada Commander, yang diperlukan hanya memilihnya di menu **RStudio**. Selain itu, kita juga dapat meng-import file berisikan data tanpa perlu mengetikkan pada Commander dengan cara memilih pada menu **Environment**.

1.5 Menginstall R dan RStudio

Pada tutorial ini hanya akan dijelaskan bagaimana menginstal R dan **RStudio** pada sistem operasi **windows**. Sebelum memulai menginstal sebaiknya pembaca mengunduh terlebih dahulu *installer* R dan **RStudio**.

1. Jalankan proses pemasangan dengan meng-klik *installer* aplikasi R dan **RStudio**.

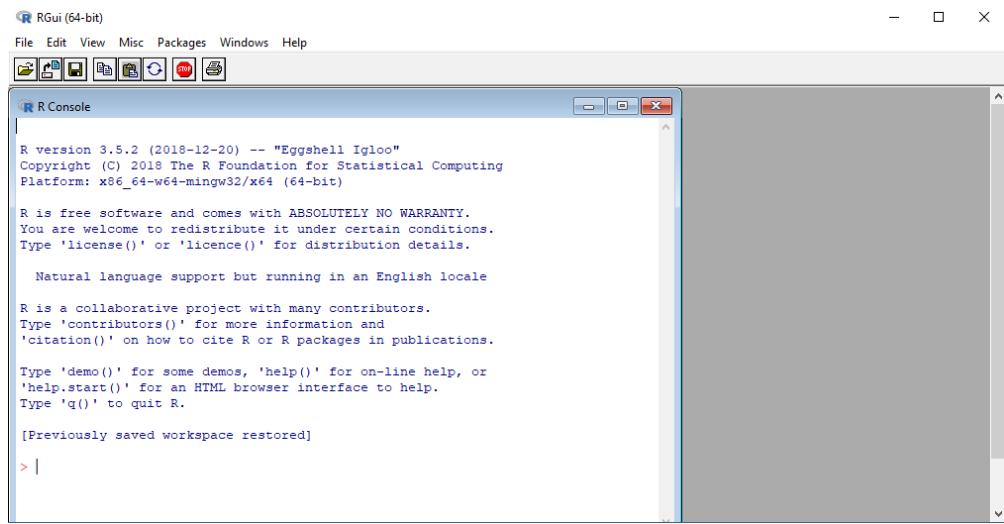


Figure 1.2: Jendela R.

2. Ikuti langkah proses pemasangan aplikasi yang ditampilkan dengan klik **OK** atau **Next**.
3. Apabila pemasangan telah dilakukan, jalankan aplikasi yang telah terpasang untuk menguji jika aplikasi telah berjalan dengan baik.

Jendela aplikasi yang telah terpasang ditampilkan pada Gambar 1.2 dan Gambar 1.3.

Note: Sebaiknya install R terlebih dahulu sebelum RStudio

1.6 Working Directory

Setiap pengguna akan bekerja pada tempat khusus yang disebut sebagai *working directory*. *working directory* merupakan sebuah folder dimana R akan membaca dan menyimpan file kerja kita. Pada pengguna windows, *working directory* secara default pada saat pertama kali menginstall R terletak pada folder c:\\Document.

1.6.1 Mengubah Lokasi Working Directory

Kita dapat mengubah lokasi *working directory* berdasarkan lokasi yang kita inginkan, misalnya letak data yang akan kita olah tidak ada pada folder default atau kita ingin pekerjaan kita terkait R dapat berlangsung pada satu folder khusus.

Berikut adalah cara mengubah *working directory* pada R.

1. Buatlah folder pada drive (kita bisa membuat folder pada selain drive c) dan namai dengan nama yang kalian inginkan. Pada tutorial ini penulis menggunakan nama folder R.
2. Jika pengguna menggunakan RStudio, pada menu RStudio pilih **Session > Set Working Directory > Chooses Directory**. Proses tersebut ditampilkan pada Gambar 1.4
3. Pilih folder yang telah dibuat pada step 1 sebagai *working directory.

Note: Data atau file yang hendak dibaca selama proses kerja pada R harus selalu diletakkan pada working directory. Jika tidak maka data atau file tidak akan terbaca.

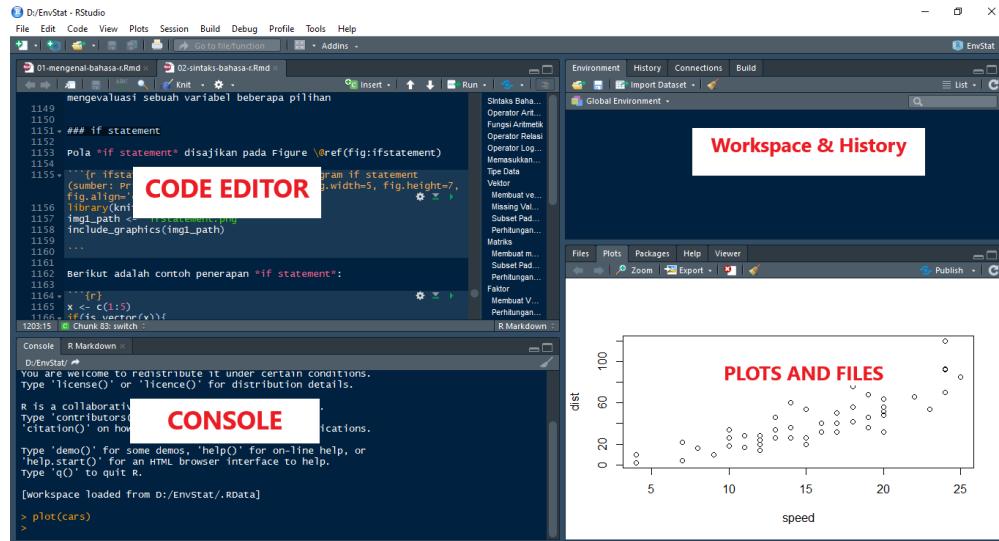


Figure 1.3: Jendela RStudio.

Untuk mengecek apakah proses perubahan telah terjadi, kita dapat mengeceknya dengan menjalankan perintah berikut untuk melihat lokasi *working directory* kita yang baru.

```
getwd()
```

Selain itu kita dapat mengubah *working directory* menggunakan perintah berikut:

```
# Ubah working directori pada folder R
setwd("/Documents/R")
```

Note: Pada proses pengisian lokasi folder pastikan pemisah pada lokasi folder menggunakan tanda “/” bukan “”

1.6.2 Mengubah Lokasi Working Directory Default

Pada proses yang telah penulis jelaskan sebelumnya. Proses perubahan *working directory* hanya berlaku pada saat pekerjaan tersebut dilakukan. Setelah pekerjaan selesai dan kita menjalankan kembali R maka *working directory* akan kembali secara default pada working directory lama.

Untuk membuat lokasi default *working directory* pindah, kita dapat melakukannya dengan memilih pada menu: **Tools > Global options > pada “General” klik pada “Browse” dan pilih lokasi working directory yang diinginkan.** Proses tersebut ditampilkan pada Gambar 1.5

1.7 Fasilitas Help

Agar dapat menggunakan R dengan secara lebih baik, pengetahuan untuk mengakses fasilitas *help* cukup penting untuk disampaikan. Adapun cara yang dapat digunakan adalah sebagai berikut.

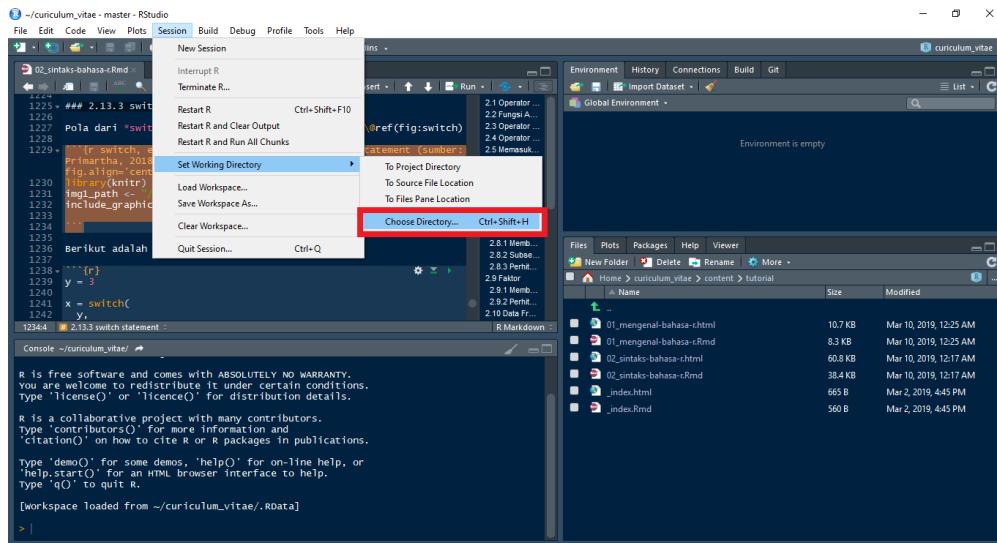


Figure 1.4: Mengubah working directory.

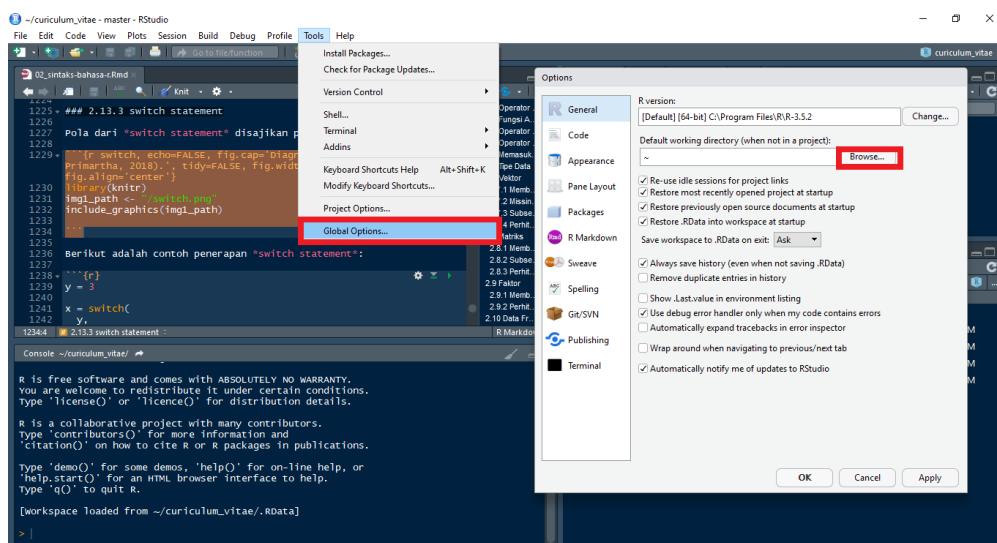


Figure 1.5: Merubah working directory melalui Global options.

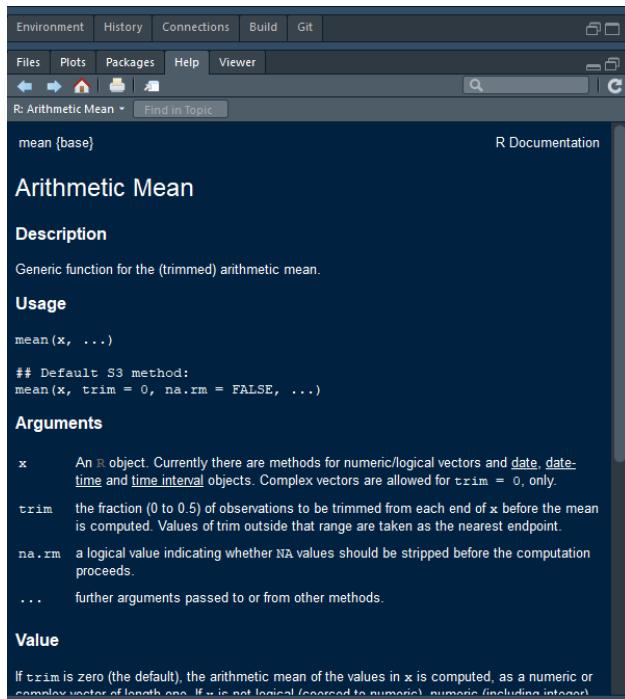


Figure 1.6: Jendela help dokumentasi fungsi mean().

1.7.1 Mencari Help dari Suatu Perintah Tertentu

Untuk memperoleh bantuan terkait suatu perintah tertentu kita dapat menggunakan fungsi `help()`. Secara umum format yang digunakan adalah sebagai berikut:

```
help(nama_perintah)
```

atau dapat juga menggunakan tanda tanya (?) pada awal `nama_perintah` seperti berikut:

```
?nama_perintah
```

Misalkan kita kebingungan terkait bagaimana cara menuliskan perintah untuk menghitung rata-rata suatu vektor. Kita dapat mengetikkan perintah berikut untuk mengakses fasilitas `help`.

```
help(mean)
#atau
?mean
```

Perintah tersebut akan memunculkan hasil berupa dokumentasi yang ditampilkan pada Gambar 1.6.

Keterangan pada jendela pada Gambar 1.6 adalah sebagai berikut:

1. Pada bagian jendela kiri atas jendela `help`, diberikan keterangan nama dari perintah yang sedang ditampilkan.
2. Selanjutnya, pada bagian atas dokumen, ditampilkan infomasi terkait nama perintah, dan nama *library* yang memuat perintah tersebut. Pada gambar diatas informasi terkait perintah dan nama *library* ditunjukkan pada teks `mean {base}` yang menunjukkan perintah `mean()` pada paket (*library*) `base` (paket bawaan R).

3. Setiap jendela *help* dari suatu perintah tertentu selanjutnya akan memuat bagian-bagian berikut:

- *Title*
- *Description* : deskripsi singkat tentang perintah.
- *Usage* : menampilkan sintaks perintah untuk penggunaan perintah tersebut.
- *Arguments* : keterangan mengenai *argument/input* yang diperlukan pada perintah tersebut.
- *Details* : keterangan lebih lengkap tentang perintah tersebut.
- *Value* : keterangan tentang *output* suatu perintah dapat diperoleh pada bagian ini.
- *Author(s)* : memberikan keterangan tentang *Author* dari perintah tersebut.
- *References* : seringkali referensi yang dapat digunakan untuk memperoleh keterangan lebih lanjut terhadap suatu perintah ditampilkan pada bagian ini.
- *See also*: bagian ini berisikan daftar perintah/fungsi yang berhubungan erat dengan perintah tersebut.
- *Example* : berisikan contoh-contoh penggunaan perintah tersebut.

Kita juga dapat melihat contoh penggunaan dari perintah tersebut. Untuk melakukannya kita dapat menggunakan fungsi `example()`. Fungsi tersebut akan menampilkan contoh kode penerapan dari fungsi yang kita inginkan. Secara sederhana fungsi tersebut dapat dituliskan sebagai berikut:

```
example(nama_perintah)
```

Untuk mengetahui contoh kode fungsi `mean()`, ketikkan sintaks berikut:

```
example(mean)
```

```
##  
## mean> x <- c(0:10, 50)  
##  
## mean> xm <- mean(x)  
##  
## mean> c(xm, mean(x, trim = 0.10))  
## [1] 8.75 5.50
```

kita juga dapat mencoba kode yang dihasilkan pada console R. Berikut adalah contoh penerapannya:

```
# Menghitung rata-rata bilangan 1 sampai 10 dan 50  
# membuat vektor  
x <- c(0:10, 50)  
  
# Print  
x
```

```
## [1] 0 1 2 3 4 5 6 7 8 9 10 50
```

```
# mean  
mean(x)
```

```
## [1] 8.75
```

Pembaca dapat mencoba melakukannya sendiri dengan mengganti nilai yang telah ada serta mencoba contoh kode yang lain.

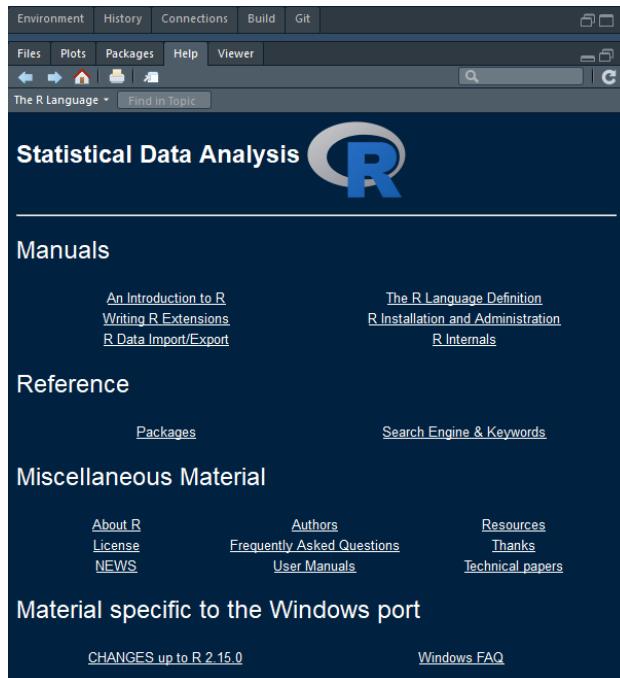


Figure 1.7: Jendela general help dokumentasi fungsi mean().

1.7.2 General Help

Kita juga dapat membaca beberapa dokumen manual yang ada pada R. Untuk melakukannya jalankan perintah berikut:

```
help.start()
```

Output yang dihasilkan berupa link pada sejumlah dokumen yang dapat kita klik. Tampilan halaman yang dihasilkan disajikan pada Gambar 1.7.

1.7.3 Fasilitas Help Lainnya

Selain yang telah penulis sebutkan sebelumnya. Kita juga dapat memanfaatkan fasilitas *help* lainnya melalui fungsi *apropos()* dan *help.search()*.

apropos (): mengembalikan daftar objek, berisi pola yang pembaca cari, dengan pencocokan sebagian. Ini berguna ketika pembaca tidak ingat persis nama fungsi yang akan digunakan. Berikut adalah contoh ketika penulis ingin mengetahui fungsi yang digunakan untuk menghitung median.

```
apropos("med")
```

```
## [1] "elNamed"          "elNamed<-"        "median"
## [4] "median.default"   "medpolish"       "runmed"
```

List yang dihasilkan berupa fungsi-fungsi yang memiliki elemen kata “med”. Berdasarkan pencarian tersebut penulis dapat mencoba menggunakan fungsi “median” untuk menghitung median.

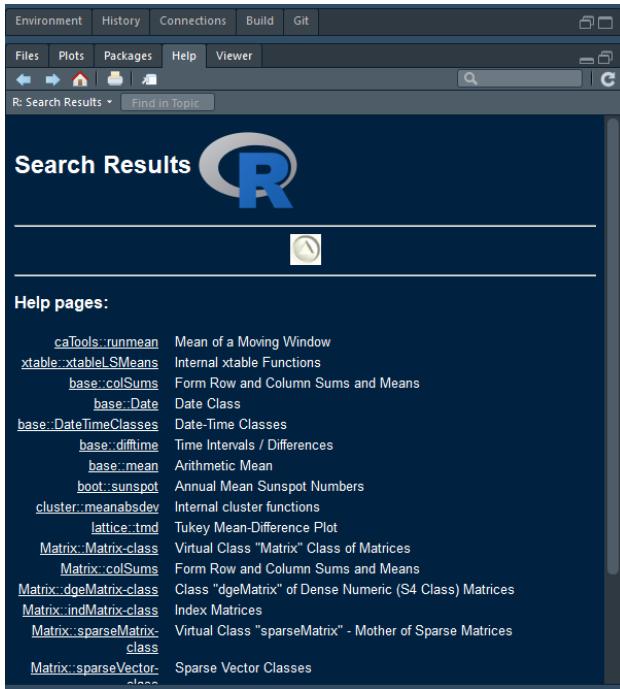


Figure 1.8: Jendela help search dokumentasi fungsi mean().

`help.search()` (sebagai alternatif `??`): mencari dokumentasi yang cocok dengan karakter yang diberikan dengan cara yang berbeda. Ini mengembalikan daftar fungsi yang mengandung istilah yang pembaca cari dengan deskripsi singkat dari fungsi.

Berikut adalah contoh penerapan dari fungsi tersebut:

```
help.search("mean")
# atau
??mean
```

Output yang dihasilkan akan tampak seperti pada Gambar 1.8.

1.8 Referensi

1. Primartha, R. 2018. **Belajar Machine Learning Teori dan Praktik.** Penerbit Informatika : Bandung
2. Rosadi,D. 2016. **Analisis Statistika dengan R.** Gadjah Mada University Press: Yogyakarta
3. STHDA. Running RStudio and Setting Up Your Working Directory - Easy R Programming .<http://www.sthda.com/english/wiki/running-rstudio-and-setting-up-your-working-directory-easy-r-programming#set-your-working-directory>
4. STDHA. **Getting Help With Functions In R Programming.** <http://www.sthda.com/english/wiki/getting-help-with-functions-in-r-programming> .
5. Venables, W.N. Smith D.M. and R Core Team. 2018. **An Introduction to R.** R Manuals.

Chapter 2

Sintaks Bahasa R

Pada *chapter* ini penulis hendak mengajak pembaca lebih familiar dengan sintaks atau perintah yang ada pada R. Pembaca akan mempelajari penggunaan operator dalam melakukan operasi pengolahan data pada R, jenis data yang ada pada R, sampai dengan bagaimana kita melakukan proses *decision making* menggunakan R.

2.1 Operator Aritmatika

Proses perhitungan akan ditangani oleh fungsi khusus. R akan memahami urutannya secara benar. Kecuali kita secara eksplisit menetapkan yang lain. Sebagai contoh jalankan sintaks berikut:

```
2+4*2
```

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

Bandingkan dengan sintaks berikut:

```
(2+4)*2
```

```
## [1] 12
```

R dapat digunakan sebagai kalkulator

Berdasarkan kedua hasil tersebut dapat disimpulkan bahwa ketika kita tidak menetapkan urutan perhitungan menggunakan tanda kurung, R akan secara otomatis akan menghitung terlebih dahulu perkalian atau pembagian.

Operator aritmatika yang disediakan R disajikan pada Tabel 2.1:

Simbol	Keterangan
+	<i>Addition</i> , untuk operasi penjumlahan
-	<i>Subtraction</i> , untuk operasi pengurangan
*	<i>Multiplication</i> , untuk operasi pembagian
/	<i>Division</i> , untuk operasi pembagian
[^]	<i>Eksponentiation</i> , untuk operasi pemangkatan
%%	<i>Modulus</i> , Untuk mencari sisa pembagian
/%	<i>Integer</i> , Untuk mencari bilangan bulat hasil pembagian saja dan tanpa sisa pembagian

*: Operator Aritmatika R.**

Untuk lebih memahaminya berikut contoh sintaks penerapan operator tersebut.

```
# Addition
5+3
```

```
## [1] 8
```

```
# Subtraction
5-3
```

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

```
# Multiplication
5*3
```

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

```
# Division
5/3
```

```
## [1] 1.667
```

```
# Eksponetiation
5^3
```

```
## [1] 125
```

```
# Modulus
5%3
```

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

```
# Integer
5%/%3
```

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

Note: Pada R tanda # berfungsi menambahkan keterangan untuk menjelaskan sebuah sintaks pada R.

2.2 Fungsi Aritmetik

Selain fungsi operator aritmetik, pada R juga telah tersedia fungsi aritmetik yang lain seperti logaritmik, eksponensial, trigonometri, dll.

1. Logaritma dan eksponensial

Untuk contoh fungsi logaritmik dan eksponensial jalankan sintaks berikut:

```

log2(8) # logaritma basis 2 untuk 8

## [1] 3

log10(8) # logaritma basis 10 untuk 8

## [1] 0.9031

exp(8) # eksponensial 8

## [1] 2981

```

2. Fungsi trigonometri

fungsi trigonometri yang ditampilkan seperti sin,cos, tan, dll.

```

cos(x) # cos x
sin(x) # Sin x
tan(x) # Tan x
acos(x) # arc-cos x
asin(x) # arc-sin x
atan(x) #arc-tan x

```

Note: x dalam fungsi trigonometri memiliki satuan radian

Berikut adalah salah satu contoh penggunaannya:

```
cos(pi)
```

```
## [1] -1
```

3. Fungsi matematik lainnya

Fungsi lainnya yang dapat digunakan adalah fungsi absolut, akar kuadrat, dll. Berikut adalah contoh sintaks penggunaan fungsi absolut dan akar kuadrat.

```

abs(-2) # nilai absolut -2

## [1] 2

sqrt(4) # akar kuadrat 4

## [1] 2

```

2.3 Operator Relasi

Operator relasi digunakan untuk membandingkan satu objek dengan objek lainnya. Operator yang disediakan R disajikan pada Tabel 2.2.

Table 2.2: Operator Relasi R.

Simbol	Keterangan
“>”	Lebih besar dari
“<”	Lebih Kecil dari
“==”	Sama dengan
“>=”	Lebih besar sama dengan
“<=”	Lebih kecil sama dengan
“!=”	Tidak sama dengan

Berikut adalah penerapan operator pada tabel tersebut:

```
x <- 34
y <- 35

# Operator >
x > y

## [1] FALSE

# Operator <
x < y

## [1] TRUE

# operator ==
x == y

## [1] FALSE

# Operator >=
x >= y

## [1] FALSE

# Operator <=
x <= y

## [1] TRUE

# Operator !=
x != y

## [1] TRUE
```

2.4 Operator Logika

Operator logika hanya berlaku pada vektor dengan tipe logical, numeric, atau complex. Semua angka bernilai 1 akan dianggap bernilai logika TRUE. Operator logika yang disediakan R dapat dilihat pada Tabel 2.3.

Table 2.3: Operator logika R.

Simbol	Keterangan
&&	Operator logika AND
!	Opearator logika NOT
&	Operator logika AND element wise
	Operator logika OR element wise

Penerapannya terdapat pada sintaks berikut:

```
v <- c(TRUE,TRUE, FALSE)
t <- c(FALSE,FALSE, FALSE)
```

```
# Operator &&
print(v&&t)
```

```
## [1] FALSE
```

```
# Operator ||
print(v||t)
```

```
## [1] TRUE
```

```
# Operator !
print(!v)
```

```
## [1] FALSE FALSE  TRUE
```

```
# operator &
print(v&t)
```

```
## [1] FALSE FALSE FALSE
```

```
# Operator /
print(v|t)
```

```
## [1]  TRUE  TRUE FALSE
```

Note:

operator & dan | akan mengecek logika tiap elemen pada vektor secara berpasangan (sesuai urutan dari kiri ke kanan).

Operator %% dan || hanya mengecek dari kiri ke kanan pada observasi pertama. Misal saat menggunakan && jika observasi pertama TRUE maka observasi pertama pada vektor lainnya akan dicek, namun jika observasi pertama FALSE maka proses akan segera dihentikan dan menghasilkan FALSE.

2.5 Memasukkan Nilai Kedalam Variabel

Variabel pada R dapat digunakan untuk menyimpan nilai. Sebagai contoh jalankan sintaks berikut:

```
# Harga sebuah lemon adalah 500 rupiah
lemon <- 500

# Atau
500 -> lemon

# dapat juga menggunakan tanda "="
lemon = 500
```

Note:

1. R memungkinkan penggunaan <-,>, atau = sebagai perintah pengisi nilai variabel
2. R bersifat *case-sensitive*. Maksudnya adalah variabel Lemon tidak sama dengan lemon (Besar kecil huruf berpengaruh)

Untuk mengetahui nilai dari objek `lemon` kita dapat menggunakan fungsi `print()` atau mengetikkan nama objeknya secara langsung.

```
# Menggunakan fungsi print()
print(lemon)
```

```
## [1] 500
```

```
# Atau
lemon
```

```
## [1] 500
```

R akan menyimpan variabel `lemon` sebagai objek pada memori. Sehingga kita dapat melakukan operasi terhadap objek tersebut seperti mengalikannya atau menjumlahkannya dengan bilangan lain. Sebagai contoh jalankan sintaks berikut:

```
# Operasi perkalian terhadap objek lemon
5*lemon
```

```
## [1] 2500
```

Kita dapat juga mengubah nilai dari objek `lemon` dengan cara menginput nilai baru terhadap objek yang sama. R secara otomatis akan menggantikan nilai sebelumnya. Untuk lebih memahaminya jalankan sintaks berikut:

```
lemon <- 1000

# Print lemon
print(lemon)
```

```
## [1] 1000
```

Untuk lebih memahaminya berikut adalah sintaks untuk menghitung volume suatu objek.

```
# Dimensi objek
panjang <- 10
lebar <- 5
tinggi <- 5

# Menghitung volume
volume <- panjang*lebar*tinggi

# Print objek volume
print(volume)
```

```
## [1] 250
```

Untuk mengetahui objek apa saja yang telah kita buat sepanjang artikel ini kita dapat menggunakan fungsi `ls()`.

```
ls()
```

```
## [1] "A"          "B"          "img1_path"  "lebar"
## [5] "lemon"      "panjang"     "t"          "tinggi"
## [9] "v"          "volume"     "x"          "xm"
## [13] "y"
```

Kumpulan objek yang telah tersimpan dalam memori disebut sebagai **workspace**

Untuk menghapus objek pada memori kita dapat menggunakan fungsi `rm()`. Pada sintaks berikut penulis hendak menghapus objek `lemon` dan `volume`.

```
# Menghapus objek lemon dan volume
rm(lemon, volume)

# Tampilkan kembali objek yang tersisa
ls()
```

```
## [1] "A"          "B"          "img1_path"  "lebar"
## [5] "panjang"    "t"          "tinggi"     "v"
## [9] "x"          "xm"         "y"
```

Note: Setiap variabel atau objek yang dibuat akan menempati sejumlah memori pada komputer sehingga jika kita bekerja dengan jumlah data yang banyak pastikan kita menghapus seluruh objek pada memori sebelum memulai kerja.

2.6 Tipe Data

Data pada R dapat dikelompokan berdasarkan beberapa tipe. Tipe data pada R disajikan pada Tabel 2.4.

Table 2.4: Tipe Data R.

Tipe Data	Contoh	Keterangan
Logical	TRUE, FALSE	Nilai Boolean
Numeric	12.3, 5, 999	Segala jenis angka
Integer	23L, 97L, 3L	Bilangan integer (bilangan bulat)
Complex	2i, 3i, 9i	Bilangan kompleks
Character	'a', "b", "123"	Karakter dan string
Raw	Identik dengan "hello"	Segala jenis data yang disimpan sebagai raw bytes

Sintaks berikut adalah contoh dari tipe data pada R. Untuk mengetahui tipa data suatu objek kita dapat menggunakan perintah `class()`

```
# Logical
apel <- TRUE
class(apel)
```

```
## [1] "logical"
```

```
# Numeric
x <- 2.3
class(x)
```

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

```
# Integer
y <- 2L
class(y)
```

```
## [1] "integer"
```

```
# Kompleks
z <- 5+2i
class(z)
```

```
## [1] "complex"
```

```
# string
w <- "saya"
class(w)
```

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

```
# Raw
xy <- charToRaw("hello world")
class(xy)
```

```
## [1] "raw"
```

Keenam jenis data tersebut disebut sebagai tipe data atomik. Hal ini disebabkan karena hanya dapat menangani satu tipe data saja. Misalnya hanya numeric atau hanya integer.

Selain menggunakan fungsi `class()`, kita dapat pula menggunakan fungsi `is.numeric()`, `is.character()`, `is.logical()`, dan sebagainya berdasarkan jenis data apa yang ingin kita cek. Berbeda dengan fungsi `class()`, output yang dihasilkan pada fungsi seperti `is.numeric()` adalah nilai Boolean sehingga fungsi ini hanya digunakan untuk mengecek apakah jenis data pada objek sama seperti yang kita pikirkan. Sebagai contoh disajikan pada sintaks berikut:

```
data <- 25

# Cek apakah objek berisi data numerik
is.numeric(data)
```

```
## [1] TRUE

# Cek apakah objek adalah karakter
is.character(data)
```

```
## [1] FALSE
```

Kita juga dapat mengubah jenis data menjadi jenis lainnya seperti integer menjadi numeric atau sebaliknya. Fungsi yang digunakan adalah `as.numeric()` jika ingin mengubah suatu jenis data menjadi numeric. Fungsi lainnya juga dapat digunakan sesuai dengan kita ingin mengubah jenis data objek menjadi jenis data lainnya.

```
# Integer
apel <- 2L

# Ubah menjadi numeric
as.numeric(apel)
```

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

```
# Cek
is.numeric(apel)
```

```
## [1] TRUE
```

```
# Logical
nangka <- TRUE

# Ubah logical menjadi numeric
as.numeric(nangka)
```

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

```
# Karakter
minum <- "minum"

# ubah karakter menjadi numeric
as.numeric(minum)
```

```
## Warning: NAs introduced by coercion
## [1] NA
```

Note: Konversi karakter menjadi numerik akan menghasilkan output NA (*not available*). R tidak mengetahui bagaimana cara merubah karakter menjadi bentuk numerik.

Berdasarkan Tabel 2, vektor karakter dapat dibuat menggunakan tanda kurung baik *double quote* ("") maupun *single quote* (''). Jika pada teks yang kita tuliskan mengandung *quote* maka kita harus menghentikannya menggunakan tanda (). Sbegai contoh kita ingin menuliskan ‘My friend’s name is “Adi”, pada sintaks akan dituliskan:

```
'My friend`\s name is "Adi"'
```

```
## [1] "My friend`s name is \"Adi\""
```

```
# Atau
```

```
"My friend's name \"Adi\""
```

```
## [1] "My friend's name \"Adi\""
```

2.7 Vektor

Vektor merupakan kombinasi berbagai nilai (numerik, karakter, logical, dan sebagainya berdasarkan jenis input data) pada objek yang sma. Pada contoh kasus berikut, pembaca akan memiliki sesuai jenis data input yaituvektor numerik, vector karakter, vektor logical, dll.

2.7.1 Membuat vektor

Vektor dibuat dengan menggunakan fungsi `c()`(concatenate) seperti yang disajikan pada sintaks berikut:

```
# membuat vektor numerik
x <- c(3,3.5,4,7)
x # print vektor

## [1] 3.0 3.5 4.0 7.0

# membuat vektor karakter
y <- c("Apel", "Jeruk", "Rambutan", "Salak")
y # print vektor

## [1] "Apel"      "Jeruk"     "Rambutan"  "Salak"

# membuat vektor logical
t <- c("TRUE", "FALSE", "TRUE")
t # print vektor
```

```
## [1] "TRUE"  "FALSE" "TRUE"
```

selain menginput nilai pada vektor, kita juga dapat memberi nama nilai setiap vektor menggunakan fungsi `names()`.

```
# Membuat vektor jumlah buah yang dibeli
Jumlah <- c(5,5,6,7)
names(Jumlah) <- c("Apel", "Jeruk", "Rambutan", "Salak")

# Atau
Jumlah <- c(Apel=5, Jeruk=5, Rambutan=6, Salak=7)
```

```
# Print
Jumlah
```

```
##      Apel    Jeruk Rambutan     Salak
##      5        5       6        7
```

Note: Vektor hanya dapat memuat satu buah jenis data. Vektor hanya dapat mengandung jenis data numerik saja, karakter saja, dll.

Untuk menentukan panjang sebuah vektor kita dapat menggunakan fungsi `length()`.

```
length(Jumlah)
```

```
## [1] 4
```

2.7.2 Missing Values

Seringkali nilai pada vektor kita tidak lengkap atau terdapat nilai yang hilang (*missing value*) pada vektor. *Missing value* pada R dilambangkan oleh NA(*not available*). Berikut adalah contoh vektor dengan *missing value*.

```
Jumlah <- c(Apel=5, Jeruk=NA, Rambutan=6, Salak=7)
```

Untuk mengecek apakah dalam objek terdapat *missing value* dapat menggunakan fungsi `is.na()`. output dari fungsi tersebut adalah nilai Boolean. Jika terdapat *Missing value*, maka output yang dihasilkan akan memberikan nilai TRUE.

```
is.na(Jumlah)
```

```
##      Apel    Jeruk Rambutan     Salak
##      FALSE   TRUE    FALSE   FALSE
```

Note:

Selain NA terdapat NaN (*not a number*) sebagai *missing value*. Nilai tersebut muncul ketika fungsi matematika yang digunakan pada proses perhitungan tidak bekerja sebagaimana mestinya. Contoh: $0/0 = \text{NaN}$

`is.na()` juga akan menghasilkan nilai TRUE pada NaN. Untuk membedakannya dengan NA dapat digunakan fungsi `is.nan()`.

2.7.3 Subset Pada Vektor

Subsetting vector terdiri atas tiga jenis, yaitu: *positive indexing*, *Negative Indexing*, dan .

- **Positive indexing:** memilih elemen vektor berdasarkan posisinya (indeks) dalam kurung siku.

```
# Subset vektor pada urutan kedua
Jumlah[2]
```

```
## Jeruk
## NA
```

```
# Subset vektor pada urutan 2 dan 4
Jumlah[c(2, 4)]
```

```
## Jeruk Salak
## NA 7
```

Selain melalui urutan (indeks), kita juga dapat melakukan subset berdasarkan nama elemen vektornya.

```
Jumlah["Jeruk"]
```

```
## Jeruk
## NA
```

Note: Indeks pada R dimulai dari 1. Sehingga kolom atau elemen pertama vektor dimulai dari [1]

- **Negative indexing:** mengecualikan (*exclude*) elemen vektor.

```
# mengecualikan elemen vektor 2 dan 4
Jumlah[-c(2,4)]
```

```
## Apel Rambutan
## 5 6
```

```
# mengecualikan elemen vektor 1 sampai 3
Jumlah[-c(1:3)]
```

```
## Salak
## 7
```

- **Subset berdasarkan vektor logical:** Hanya, elemen-elemen yang nilai yang bersesuaian dalam vektor pemilihan bernilai TRUE, akan disimpan dalam subset.

Note: panjang vektor yang digunakan untuk subset harus sama.

```

Jumlah <- c(Apel=5, Jeruk=NA, Rambutan=6, Salak=7)

# selecting vector
merah <- c(TRUE, FALSE, TRUE, FALSE)

# Subset
Jumlah[merah==TRUE]

##      Apel Rambutan
##      5         6

# Subset untuk elemen vektor bukan missing value
Jumlah[!is.na(Jumlah)]

##      Apel Rambutan     Salak
##      5         6         7

```

2.7.4 Perhitungan Menggunakan Vektor

Jika pembaca melakukan operasi dengan vektor, operasi akan diterapkan ke setiap elemen vektor. Contoh disediakan pada sintaks di bawah ini:

```

pendapatan <- c(2000, 1800, 2500, 3000)
names(pendapatan) <- c("Andi", "Joni", "Lina", "Rani")
pendapatan

## Andi Joni Lina Rani
## 2000 1800 2500 3000

# Kalikan pendapatan dengan 3
pendapatan*3

## Andi Joni Lina Rani
## 6000 5400 7500 9000

```

Seperti yang dapat dilihat, R mengalikan setiap elemen dengan bilangan pengali.

Kita juga dapat mengalikan vektor dengan vektor lainnya. Contohnya disajikan pada sintaks berikut:

```

# membuat vektor dengan panjang sama dengan dengan vektor pendapatan
coefs <- c(2, 1.5, 1, 3)

# Mengalikan pendapatan dengan vektor coefs
pendapatan*coefs

## Andi Joni Lina Rani
## 4000 2700 2500 9000

```

Berdasarkan sintaks tersebut dapat terlihat bahwa operasi matematik terhadap masing-masing vektor dapat berlangsung jika panjang vektornya sama.

Berikut adalah fungsi lain yang dapat digunakan pada operasi matematika vektor.

```
max(x) # memperoleh nilai maksimum x
min(x) # memperoleh nilai minimum x
range(x) # memperoleh range vektor x
length(x) # memperoleh jumlah elemen vektor x
sum(x) # memperoleh total penjumlahan elemen vektor x
prod(x) # memperoleh produk elemen vektor x
mean(x) # memperoleh nilai rata-rata seluruh elemen vektor x
sd(x) # standar deviasi vektor x
var(x) # varian vektor x
sort(x) # mengurutkan elemen vektor x dari yang terbesar
```

Contoh penggunaan fungsi tersebut disajikan beberapa pada sintaks berikut:

```
# Menghitung range pendapatan
range(pendapatan)

## [1] 1800 3000

# menghitung rata-rata dan standar deviasi pendapatan
mean(pendapatan)

## [1] 2325

sd(pendapatan)

## [1] 537.7
```

2.8 Matriks

Matriks seperti Excel sheet yang berisi banyak baris dan kolom (kumpulan beberapa vektor). Matriks digunakan untuk menggabungkan vektor dengan tipe yang sama, yang bisa berupa numerik, karakter, atau logis. Matriks digunakan untuk menyimpan tabel data dalam R. Baris-baris matriks pada umumnya adalah individu / pengamatan dan kolom adalah variabel.

2.8.1 Membuat matriks

Untuk membuat matriks kita dapat menggunakan fungsi `cbind()` atau `rbind()`. Berikut adalah contoh sintaks untuk membuat matriks.

```
# membuat vektor numerik
col1 <- c(5, 6, 7, 8, 9)
col2 <- c(2, 4, 5, 9, 8)
col3 <- c(7, 3, 4, 8, 7)

# menggabungkan vektor berdasarkan kolom
my_data <- cbind(col1, col2, col3)
my_data
```

```

##      col1 col2 col3
## [1,]    5    2    7
## [2,]    6    4    3
## [3,]    7    5    4
## [4,]    8    9    8
## [5,]    9    8    7

# Mengubah atau menambahkan nama baris
rownames(my_data) <- c("row1", "row2", "row3", "row4", "row5")
my_data

```

```

##      col1 col2 col3
## row1    5    2    7
## row2    6    4    3
## row3    7    5    4
## row4    8    9    8
## row5    9    8    7

```

Note:

- **cbind()**: menggabungkan objek R berdasarkan kolom
- **rbind()**: menggabungkan objek R berdasarkan baris
- **rownames()**: mengambil atau menetapkan nama-nama baris dari objek seperti-matriks
- **colnames()**: mengambil atau menetapkan nama-nama kolom dari objek seperti-matriks

Kita dapat melakukan transpose (merotasi matriks sehingga kolom menjadi baris dan sebaliknya) menggunakan fungsi **t()**. Berikut adalah contoh penerapannya:

```
t(my_data)
```

```

##      row1 row2 row3 row4 row5
## col1    5    6    7    8    9
## col2    2    4    5    9    8
## col3    7    3    4    8    7

```

Selain melalui pembentukan sejumlah objek vektor, kita juga dapat membuat matriks menggunakan fungsi **matrix()**. Secara sederhana fungsi tersebut dapat dituliskan sebagai berikut:

```
matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE,
       dimnames = NULL)
```

Note:

- **data**: vektor data opsional
- **nrow, ncol**: jumlah baris dan kolom yang diinginkan, masing-masing.
- **byrow**: nilai logis. Jika FALSE (default) matriks diisi oleh kolom, jika tidak, matriks diisi oleh baris.
- **dimnames**: Daftar dua vektor yang memberikan nama baris dan kolom masing-masing.

Dalam kode R di bawah ini, data input memiliki panjang 6. Kita ingin membuat matriks dengan dua kolom. Kita tidak perlu menentukan jumlah baris (di sini **nrow = 3**). R akan menyimpulkan ini secara otomatis. Matriks diisi kolom demi kolom saat argumen **byrow = FALSE**. Jika kita ingin mengisi matriks dengan baris, gunakan **byrow = TRUE**. Berikut adalah contoh pembuatan matriks menggunakan fungsi **matrix()**.

```

data <- matrix(
  data = c(1,2,3, 11,12,13),
  nrow = 2, byrow = TRUE,
  dimnames = list(c("row1", "row2"), c("C.1", "C.2", "C.3"))
)
data

##      C.1 C.2 C.3
## row1   1   2   3
## row2  11  12  13

```

Untuk mengetahui dimensi dari suatu matriks, kita dapat menggunakan fungsi `ncol()` untuk mengetahui jumlah kolom matriks dan `nrow()` untuk mengetahui jumlah baris pada matriks. Berikut adalah contoh penerapannya:

```

# mengetahui jumlah kolom
ncol(my_data)

```

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

```

# mengetahui jumlah baris
nrow(my_data)

```

```
## [1] 5
```

Jika ingin memperoleh ringkasan terkait dimensi matriks kita juga dapat menggunakan fungsi `dim()` untuk mengetahui jumlah baris dan kolom matriks. Berikut adalah contoh penerapannya:

```

dim(my_data) # jumlah baris dan kolom

```

```
## [1] 5 3
```

2.8.2 Subset Pada Matriks

Sama dengan vektor, subset juga dapat dilakukan pada matriks. Bedanya subset dilakukan berdasarkan baris dan kolom pada matriks.

- **Memilih baris/kolom** berdasarkan pengindeksan positif

baris atau kolom dapat diseleksi menggunakan format `data[row, col]`. Cara selesa ini sama dengan vektor, bedanya kita harus menetukan baris dan kolom dari data yang akan kita pilih. Berikut adalah contoh penerapannya:

```

# Pilih baris ke-2
my_data[2,]

```

```

## col1 col2 col3
##     6    4    3

```

```
# Pilih baris 2 sampai 4
my_data[2:4,]

##      col1 col2 col3
## row2     6     4     3
## row3     7     5     4
## row4     8     9     8

# Pilih baris 2 dan 4
my_data[c(2,4),]

##      col1 col2 col3
## row2     6     4     3
## row4     8     9     8

# Pilih baris 2 dan kolom 3
my_data[2, 3]
```

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

- **Pilih berdasarkan nama baris/kolom**

Berikut adalah contoh subset berdasarkan nama baris atau kolom.

```
# Pilih baris 1 dan kolom 3
my_data["row1","col3"]

## [1] 7

# Pilih baris 1 sampai 4 dan kolom 3
baris <- c("row1","row2","row3")
my_data[baris, "col3"]

## row1 row2 row3
##    7    3    4
```

- **Kecualikan baris/kolom** dengan pengindeksan negatif

Sama seperti vektor pengecualian data dapat dilakukan di matriks menggunakan pengindeksan negatif. Berikut cara melakukannya:

```
# Kecualikan baris 2 dan 3 serta kolom 3
my_data[-c(2,3), -3]
```

```
##      col1 col2
## row1     5     2
## row4     8     9
## row5     9     8
```

- **Pilihan dengan logik**

Dalam kode R di bawah ini, misalkan kita ingin hanya menyimpan baris di mana $\text{col3} \geq 4$:

```
col3 <- my_data[, "col3"]
my_data[col3 >= 4, ]
```

```
##      col1 col2 col3
## row1    5    2    7
## row3    7    5    4
## row4    8    9    8
## row5    9    8    7
```

2.8.3 Perhitungan Menggunakan Matriks

— Kita juga dapat melakukan operasi matematika pada matriks. Pada operasi matematika pada matriks proses yang terjadi bisa lebih kompleks dibanding pada vektor, dimana kita dapat melakukan operasi untuk memperoleh gambaran data pada tiap kolom atau baris.

Berikut adalah contoh operasi matematika sederhana pada matriks:

```
# mengalikan masing-masing elemen matriks dengan 2
my_data*2
```

```
##      col1 col2 col3
## row1   10    4   14
## row2   12    8    6
## row3   14   10    8
## row4   16   18   16
## row5   18   16   14
```

```
# memperoleh nilai log basis 2 pada masing-masing elemen matriks
log2(my_data)
```

```
##      col1 col2 col3
## row1 2.322 1.000 2.807
## row2 2.585 2.000 1.585
## row3 2.807 2.322 2.000
## row4 3.000 3.170 3.000
## row5 3.170 3.000 2.807
```

Seperti yang telah penulis jelaskan sebelumnya, kita juga dapat melakukan operasi matematika untuk memperoleh hasil penjumlahan elemen pada tiap baris atau kolom dengan menggunakan fungsi `rowSums()` untuk baris dan `colSums()` untuk kolom.

```
# Total pada tiap kolom
colSums(my_data)
```

```
## col1 col2 col3
##   35   28   29
```

```
# Total pada tiap baris
rowSums(my_data)
```

```
## row1 row2 row3 row4 row5
## 14   13   16   25   24
```

Jika kita tertarik untuk mencari nilai rata-rata tiap baris atau kolom kita juga dapat menggunakan fungsi `rowMeans()` atau `colMeans()`. Berikut adalah contoh penerapannya:

```
# Rata-rata tiap baris
rowMeans(my_data)

## row1 row2 row3 row4 row5
## 4.667 4.333 5.333 8.333 8.000
```

```
# Rata-rata tiap kolom
colMeans(my_data)
```

```
## col1 col2 col3
## 7.0  5.6  5.8
```

Kita juga dapat melakukan perhitungan statistika lainnya menggunakan fungsi `apply()`. Berikut adalah format sederhananya:

```
apply(x, MARGIN, FUN)
```

Note:

- `x` : data matriks
- `MARGIN` : Nilai yang dapat digunakan adalah 1 (untuk operasi pada baris) dan 2 (untuk operasi pada kolom)
- `FUN` : fungsi yang diterapkan pada baris atau kolom

untuk mengetahui fungsi (`FUN`) apa saja yang dapat diterapkan pada fungsi `apply()` jalankan sintaks bantuan berikut:

```
help(apply)
```

Berikut adalah contoh penerapannya:

```
# Rata-rata pada tiap baris
apply(my_data, 1, mean)
```

```
## row1 row2 row3 row4 row5
## 4.667 4.333 5.333 8.333 8.000
```

```
# Median pada tiap kolom
apply(my_data, 2, median)
```

```
## col1 col2 col3
##    7    5    7
```

2.9 Faktor

Dalam bahasa R , faktor merupakan vektor dengan level. Level disimpan sebagai R Character. Jika kita menggunakan SPSS maka factor ini akan sama dengan jenis data numerik atau ordinal.

Faktor merepresentasikan kategori atau grup pada data. Untuk membuat faktor pada R, kita dapat menggunakan fungsi `factor()`.

2.9.1 Membuat Variabel Faktor

Berikut adalah contoh sintaks pembuatan variabel faktor.

```
# membuat variabel faktor
faktor <- factor(c(1,2,1,2))
faktor
```

```
## [1] 1 2 1 2
## Levels: 1 2
```

Pada sintaks tersebut objek faktor terdiri atas dua buah kategori atau pada R disebut sebagai **factor levels**. Kita dapat mengecek factor levels menggunakan fungsi `levels()`.

```
levels(faktor)
```

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

Kita juga dapat memberikan label atau mengubah level pada faktor. Berikut adalah contoh bagaimana kita melakukannya:

```
# Ubah level
levels(faktor) <- c("baik","tidak_baik")
faktor
```

```
## [1] baik      tidak_baik baik      tidak_baik
## Levels: baik tidak_baik
```

```
# Ubah urutan level
faktor <- factor(faktor,
                  levels = c("tidak_baik","baik"))
faktor
```

```
## [1] baik      tidak_baik baik      tidak_baik
## Levels: tidak_baik baik
```

Note:

- Fungsi `is.factor()` dapat digunakan untuk mengecek apakah sebuah variabel adalah faktor. Hasil yang dimunculkan dapat berupa TRUE (jika faktor) atau FALSE (jika bukan)
- Fungsi `as.factor()` dapat digunakan untuk merubah sebuah variabel menjadi faktor.

```
# Cek jika objek faktor adalah faktor
is.factor(faktor)
```

```
## [1] TRUE
```

```
# Cek jika objek Jumlah adalah faktor
is.factor(Jumlah)
```

```
## [1] FALSE
```

```
# Ubah objek Jumlah menjadi faktor
as.factor(Jumlah)
```

```
##      Apel    Jeruk Rambutan     Salak
##      5       <NA>        6        7
## Levels: 5 6 7
```

2.9.2 Perhitungan Menggunakan Faktor

Jika kita ingin mengetahui jumlah masing-masing observasi pada masing-masing faktor, kita dapat menggunakan fungsi `summary()`. Berikut adalah contoh penerapannya:

```
summary(faktor)
```

```
## tidak_baik      baik
##          2          2
```

Pada contoh perhitungan menggunakan vektor kita telah membuat objek `pendapatan`. Pada objek tersebut kita ingin menghitung nilai rata-rata pendapatan berdasarkan objek faktor. Untuk melakukannya kita dapat menggunakan fungsi `tapply()`.

```
pendapatan
```

```
## Andi Joni Lina Rani
## 2000 1800 2500 3000
```

```
faktor
```

```
## [1] baik      tidak_baik baik      tidak_baik
## Levels: tidak_baik baik
```

```
# Rata-rata pendapatan dan simpan sebagai objek dengan nama:
# mean_pendapatan
mean_pendapatan <- tapply(pendapatan, faktor, mean)
mean_pendapatan
```

```
## tidak_baik      baik
##          2400      2250
```

```
# Hitung ukuran/panjang masing-masing grup
tapply(pendapatan, faktor, length)
```

```
## tidak_baik      baik
##              2          2
```

Untuk mengetahui jumlah masing-masing observasi masing-masing factor levels kita juga dapat menggunakan fungsi `table()`. Fungsi tersebut akan membuat frekuensi tabel pada masing-masing factor levels atau yang dikenal sebagai *contingency table*.

```
table(faktor)
```

```
## faktor
## tidak_baik      baik
##              2          2
```

```
# Cross-tabulation antara
# faktor dan pendapatan
table(pendapatan, faktor)
```

```
##           faktor
## pendapatan tidak_baik baik
##      1800        1   0
##      2000        0   1
##      2500        0   1
##      3000        1   0
```

2.10 Data Frames

Data frame merupakan kumpulan vektor dengan panjang sama atau dapat pula dikatan sebagai matriks yang memiliki kolom dengan jenis data yang berbeda-beda (numerik, karakter, logical). Pada data frame terdapat baris dan kolom. Baris disebut sebagai observasi, sedangkan kolom disebut sebagai variabel. Sehingga dapat dikatakan bahwa setiap observasi akan memiliki satu atau beberapa variabel.

2.10.1 Membuat Data Frame

Data frame dapat dibuat menggunakan fungsi `data.frame()`. Berikut adalah contoh cara membuat data frame:

```
# Membuat data frame
nama <- c("Andi", "Rizal", "Ani", "Ina")
pendapatan <- c(1000, 2000, 3500, 500)
tinggi <- c(160, 155, 170, 146)
usia <- c(35, 40, 25, 27)
menikah <- c(TRUE, FALSE, TRUE, TRUE)

data_teman <- data.frame(nama = nama,
                           gaji = pendapatan,
                           tinggi = tinggi,
```

```

menikah = menikah)

data_teman

##   nama gaji tinggi menikah
## 1 Andi 1000    160    TRUE
## 2 Rizal 2000    155   FALSE
## 3 Ani 3500    170    TRUE
## 4 Ina  500    146    TRUE

```

Untuk mengecek apakah objek `data_teman` merupakan data frame, kita dapat menggunakan fungsi `is.data.frame()`. Jika hasilnya TRUE, maka objek tersebut adalah data frame. Berikut adalah contoh penerapannya:

```
is.data.frame(data_teman)
```

```
## [1] TRUE
```

Note: untuk konversi objek menjadi data frame, kita dapat menjalankan fungsi `as.data.frame()`.

2.10.2 Subset Pada Data Frame

Subset pada data frame sebenarnya tidak berbeda dengan subset pada matriks. Bedanya adalah kita juga bisa melakukan subset langsung terhadap nama variabel menggunakan dollar sign. Untuk lebih memahaminya berikut adalah jenis subset pada data frame.

- Pengindeksan positif menggunakan nama dan lokasi.

```

# Subset menggunakan dollar sign
data_teman$nama

## [1] Andi Rizal Ani Ina
## Levels: Andi Ani Ina Rizal

# atau
data_teman[, "nama"]

## [1] Andi Rizal Ani Ina
## Levels: Andi Ani Ina Rizal

# subset baris 1 sampai 3 serta kolom 1 dan 3
data_teman[1:3, c(1,3)]


##   nama tinggi
## 1 Andi    160
## 2 Rizal    155
## 3 Ani     170

```

- Pengindeksan negatif

```
# Kecualikan kolom nama
data_teman[,-1]
```

```
##   gaji tinggi menikah
## 1 1000    160    TRUE
## 2 2000    155   FALSE
## 3 3500    170    TRUE
## 4  500    146    TRUE
```

- Pengideksan berdasarkan karakteristik

Kita ingin memilih data dengan kriteria teman yang telah menikah

```
data_teman[data_teman$menikah==TRUE, ]
```

```
##   nama gaji tinggi menikah
## 1 Andi 1000    160    TRUE
## 3  Ani 3500    170    TRUE
## 4  Ina  500    146    TRUE
```

```
# Tampilkan hanya kolom nama dan gaji untuk yang telah menikah
data_teman[data_teman$menikah==TRUE, 1:2]
```

```
##   nama gaji
## 1 Andi 1000
## 3  Ani 3500
## 4  Ina  500
```

kita juga dapat menggunakan fungsi `subset()` agar lebih mudah. Berikut adalah contoh penerapannya:

```
# subset terhadap teman yang berusia >=30 tahun
subset(data_teman, usia>=30)
```

```
##   nama gaji tinggi menikah
## 1 Andi 1000    160    TRUE
## 2 Rizal 2000    155   FALSE
```

Opsi lain adalah menggunakan fungsi `attach()` dan `detach()`. Fungsi `attach()` mengambil data frame dan membuat kolomnya dapat diakses hanya dengan memberikan nama mereka.

```
# attach data frame
attach(data_teman)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##   menikah, nama, tinggi
```

```
# ===== memulai data manipulation =====
data_teman[usia>=30]

##      nama gaji
## 1  Andi 1000
## 2 Rizal 2000
## 3  Ani 3500
## 4  Ina  500

# ===== mengakhiri data manipulation =====
# detach data frame

detach(data_teman)
```

2.10.3 Memperluas Data Frame

Kita dapat juga memperluas data frame dengan cara menambahkan variabel atau kolom baru pada data frame. Pada contoh kali ini penulis akan menambahkan kolom pendidikan terakhir pada objek `data_teman`. Berikut adalah sintaks yang digunakan.

```
# membuat vektor pendidikan
pendidikan <- c("S1", "S2", "D3", "D1")

# menambahkan variabel pendidikan pada data frame
data_teman$pendidikan <- pendidikan

# atau
cbind(data_teman, pendidikan=pendidikan)
```

2.10.4 Perhitungan Pada Data Frame

Perhitungan pada variabel numerik data frame pada dasarnya sama dengan perhitungan pada matriks. kita dapat menggunakan fungsi `rowSums()`, `colSums()`, `rowMeans()` dan `apply()`. Proses perhitungan dan manipulasi pada data frame akan dibahas pada sesi yang lain secara lebih detail.

2.11 List

List adalah kumpulan objek yang diurutkan, yang dapat berupa vektor, matriks, data frame, dll. Dengan kata lain, daftar dapat berisi semua jenis objek R.

2.11.1 Membuat List

List dapat dibuat menggunakan fungsi `list()`. Berikut disajikan contoh sebuah list sebuah keluarga:

```
# Membuat list keluarga
keluarga <- list(
  ayah = "Budi",
  usia_ayah = 48,
```

```

ibu = "Ani",
usia_ibu = "47",
anak = c("Andi", "Adi"),
usia_anak = c(15,10)
)

# Print
keluarga

## $ayah
## [1] "Budi"
##
## $usia_ayah
## [1] 48
##
## $ibu
## [1] "Ani"
##
## $usia_ibu
## [1] "47"
##
## $anak
## [1] "Andi" "Adi"
##
## $usia_anak
## [1] 15 10

# Nama elemen dalam list
names(keluarga)

## [1] "ayah"      "usia_ayah" "ibu"       "usia_ibu"
## [5] "anak"      "usia_anak"

# Jumlah elemen pada list
length(keluarga)

## [1] 6

# Subset berdasarkan nama
# mengambil elemen usia_ayah
keluarga$usia_ayah

## [1] 48

```

2.11.2 Subset List

Kita dapat memilih sebuah elemen pada list dengan menggunakan nama elemen atau indeks dari elemen tersebut. Berikut adalah contoh penerapannya:

```

# Subset berdasarkan nama
# mengambil elemen usia_ayah
keluarga$usia_ayah

## [1] 48

```

```
# Atau
keluarga[["usia_ayah"]]

## [1] 48

# Subset berdasarkan indeks
keluarga[[2]]

## [1] 48

# subset elemen pertama pada keluarga[[5]]
keluarga[[5]][1]

## [1] "Andi"
```

2.11.3 Memperluas List

Kita juga dapat menambahkan elemen pada list yang telah kita buat. Pada contoh list sebelumnya penulis akan menambahkan elemen keluarga yang lain seperti berikut:

```
# Menambahkan kakek dan nenek pada list
keluarga$kakek <- "Suprapto"
keluarga$nenek <- "Sri"

# Print
keluarga
```

```
## $ayah
## [1] "Budi"
##
## $usia_ayah
## [1] 48
##
## $ibu
## [1] "Ani"
##
## $usia_ibu
## [1] "47"
##
## $anak
## [1] "Andi" "Adi"
##
## $usia_anak
## [1] 15 10
##
## $kakek
## [1] "Suprapto"
##
## $nenek
## [1] "Sri"
```

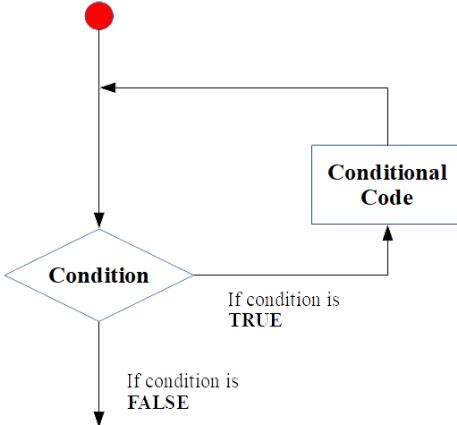


Figure 2.1: Diagram umum loop (sumber: Primartha, 2018).

Kita juga dapat menggabungkan beberapa list menjadi satu. Berikut adalah format sederhana bagaimana cara menggabungkan beberapa list menjadi satu:

```
list_baru <- c(list_a, list_b, list_c, ...)
```

2.12 Loop

Loop merupakan kode program yang berulang-ulang. *Loop* berguna saat kita ingin melakukan sebuah perintah yang perlu dijalankan berulang-ulang seperti melakukan perhitungan maupun melakukan visualisasi terhadap banyak variabel secara serentak. Hal ini tentu saja membantu kita karena kita tidak perlu menulis sejumlah sintaks yang berulang-ulang. Kita hanya perlu mengatur *statement* berdasarkan hasil yang kita harapkan.

Pada R bentuk *loop* dapat bermacam-macam (“*for loop*”, “*while loop*”, dll). R menyederhanakan bentuk *loop* ini dengan menyediakan sejumlah fungsi seperti `apply()`, `tapply()`, dll. Sehingga *loop* jarang sekali muncul dalam kode R. Sehingga R sering disebut sebagai *loopless loop*.

Meski *loop* jarang muncul bukan berarti kita tidak akan melakukannya. Terkadang saat kita melakukan komputasi statistik atau matematik dan belum terdapat paket yang mendukung proses tersebut, sering kali kita akan membuat sintaks sendiri berdasarkan algoritma metode tersebut. Pada algoritma tersebut sering pula terdapat *loop* yang diperlukan selama proses perhitungan. Secara sederhana diagram umum loop ditampilkan pada Gambar 2.1

2.12.1 For Loop

Mengulangi sebuah *statement* atau sekelompok *statement* sebanyak nilai yang ditentukan di awal. Jadi operasi akan terus dilakukan sampai dengan jumlah yang telah ditetapkan di awal atau dengan kata lain tes kondisi (Jika jumlah pengulangan telah cukup) hanya akan dilakukan di akhir. Secara sederhana bentuk dari *for loop* dapat dituliskan sebagai berikut:

```
for (value in vector){
  statements
}
```

Berikut adalah contoh sintaks penerapan *for loop*:

```
# Membuat vektor numerik
vektor <- c(1:5)

# loop
for(i in vektor){
  print(i)
}

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

Loop akan dimulai dari blok *statement for* sampai dengan `print(i)`. Berdasarkan *loop* pada contoh tersebut, *loop* hanya dilakukan sebanyak 5 kali sesuai dengan jumlah vektor yang ada.

2.12.2 While Loop

While loop merupakan loop yang digunakan ketika kita telah menetapkan *stop condition* sebelumnya. Blok *statement/kode* yang sama akan terus dijalankan sampai *stop condition* ini tercapai. *Stop condition* akan di cek sebelum melakukan proses *loop*. Berikut adalah pola dari *while loop* dapat dituliskan sebagai berikut:

```
while (test_expression){
  statement
}
```

Berikut adalah contoh penerapan dari *while loop*:

```
coba <- c("Contoh")
counter <- 1

# loop
while (counter<5){
  # print vektor
  print(coba)
  # tambahkan nilai counter sehingga proses terus berlangsung sampai counter = 5
  counter <- counter + 1
}

## [1] "Contoh"
## [1] "Contoh"
## [1] "Contoh"
## [1] "Contoh"
```

Loop akan dimulai dari blok *statement while* sampai dengan `counter <- 1`. *Loop* hanya akan dilakukan sepanjang nilai `counter < 5`.

2.12.3 Repeat Loop

Repeat loop akan menjalankan *statement/kode* yang sama berulang-ulang hingga *stop condition* tercapai. Berikut adalah pola dari *repeat loop*.

```
repeat {
  commands
  if(condition){
    break
  }
}
```

Berikut adalah contoh penerapan dari *repeat loop*:

```
coba <- c("contoh")
counter <- 1
repeat {
  print(coba)
  counter <- counter + 1
  if(counter < 5){
break
  }
}

## [1] "contoh"
```

Loop akan dimulai dari blok *statement while* sampai dengan *break*. *Loop* hanya akan dilakukan sepanjang nilai *counter* < 5. Hasil yang diperoleh berbeda dengan *while loop*, dimana kita memperoleh 4 buah kata “contoh”. Hal ini disebabkan karena *repeat loop* melakukan pengecekan *stop condition* tidak di awal *loop* seperti *while loop* sehingga berapapun nilainya, selama nilainya sesuai dengan *stop condition* maka *loop* akan dihentikan. Hal ini berbeda dengan *while loop* dimana proses dilakukan berulang-ulang sampai jumlahnya mendekati *stop condition*.

2.12.4 Break

Break sebenarnya bukan bagian dari *loop*, namun sering digunakan dalam *loop*. *Break* dapat digunakan pada *loop* manakala dirasa perlu, yaitu saat kondisi yang disyaratkan pada *break* tercapai.

Berikut adalah contoh penerapan *break* pada beberapa jenis *loop*.

```
# for loop
a = c(2,4,6,8,10,12,14)
for(i in a){
  if(i>8){
    break
  }
  print(i)
}

## [1] 2
## [1] 4
## [1] 6
## [1] 8
```

```
# while loop
a = 2
b = 4
while(a<7){
  print(a)
  a = a +1
  if(b+a>10){
    break
  }
}
```

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

```
# repeat loop
a = 1
repeat{
  print(a)
  a = a+1
  if(a>6){
    break
  }
}
```

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

2.13 Decision Making

Decision Making atau sering disebut sebagai *if then else statement* merupakan bentuk percabagan yang digunakan manakala kita ingin agar program dapat melakukan pengujian terhadap syarat kondisi tertentu. Pada Tabel 2.5 disajikan daftar percabangan yang digunakan pada R.

Table 2.5: Daftar percabangan pada R.

Statement	Keterangan
<i>if statement</i>	<i>if statement</i> hanya terdiri atas sebuah ekspresi Boolean, dan diikuti satu atau lebih <i>statement</i>
<i>if...else statement</i>	<i>if else statement</i> terdiri atas beberapa buah ekspresi Boolean. Ekspresi Boolean berikutnya akan dijalankan jika ekspresi *Boolan sebelumnya bernilai FALSE
<i>switch statement</i>	<i>switch statement</i> digunakan untuk mengevaluasi sebuah variabel beberapa pilihan

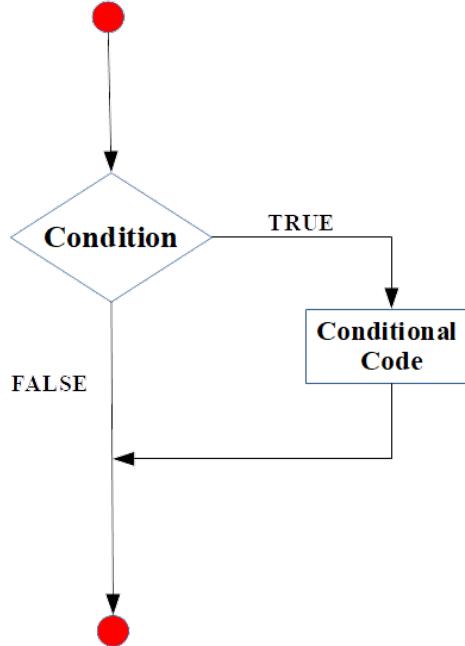


Figure 2.2: Diagram if statement (sumber: Primartha, 2018).

2.13.1 if statement

Pola *if statement* disajikan pada Gambar 2.2

Berikut adalah contoh penerapan *if statement*:

```

x <- c(1:5)
if(is.vector(x)){
  print("x adalah sebuah vector")
}

## [1] "x adalah sebuah vector"
  
```

2.13.2 if else statement

Pola dari *if else statement* disajikan pada Gambar 2.3

Berikut adalah contoh penerapan *if else statement*:

```

x <- c("Andi", "Iwan", "Adi")
if("Rina" %in% x){
  print("Rina ditemukan")
} else if("Adi" %in% x){
  print("Adi ditemukan")
} else{
  print("tidak ada yang ditemukan")
}

## [1] "Adi ditemukan"
  
```

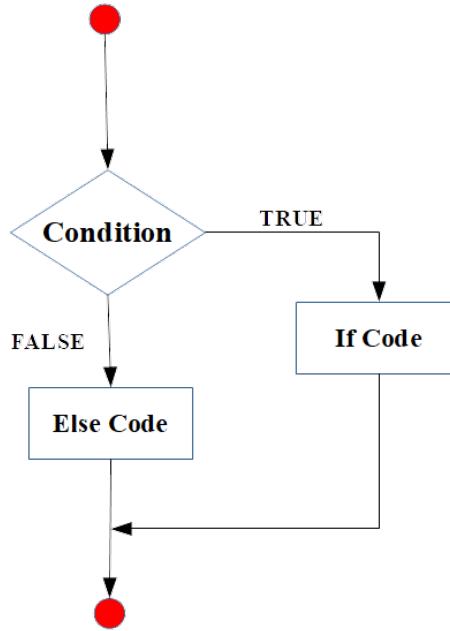


Figure 2.3: Diagram if else statement (sumber: Primartha, 2018).

2.13.3 switch statement

Pola dari *switch statement* disajikan pada Gambar 2.4

Berikut adalah contoh penerapan *switch statement*:

```

y = 3

x = switch(
  y,
  "Selamat Pagi",
  "Selamat Siang",
  "Selamat Sore",
  "Selamat Malam"
)

print(x)

## [1] "Selamat Sore"
  
```

2.14 Fungsi

Fungsi merupakan sekumpulan instruksi atau *statement* yang dapat melakukan tugas khusus. Sebagai contoh fungsi perkalian untuk menyelesaikan operasi perkalian, fungsi pemangkatan hanya untuk operasi pemangkatan, dll.

Pada R terdapat 2 jenis fungsi, yaitu: *build in function* dan *user define function*. *build in function* merupakan fungsi bawaan R saat pertama kita menginstall R. Contohnya adalah `mean()`, `sum()`, `ls()`, `rm()`, dll. Sedangkan *user define fuction* merupakan fungsi-fungsi yang dibuat sendiri oleh pengguna.

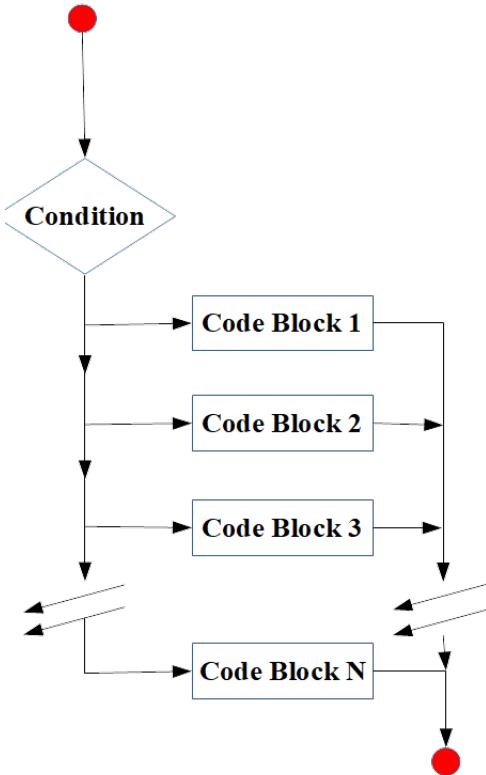


Figure 2.4: Diagram switch statement (sumber: Primartha, 2018).

Fungsi-fungsi buatan pengguna haruslah dideklarasikan (dibuat) terlebih dahulu sebelum dapat dijalankan. Pola pembentukan fungsi adalah sebagai berikut:

```
function_name <- function(argument_1, argument_2, ...){
  function body
}
```

Note:

- **function_name** : Nama dari fungsi R. R akan menyimpan fungsi tersebut sebagai objek
- **argument_1, argument_2, ...** : *Argument* bersifat opsional (tidak wajib). *Argument* dapat digunakan untuk memberi inputan kepada fungsi
- **function body** : Merupakan inti dari fungsi. Function body dapat terdiri atas 0 statement (kosong) hingga banyak statement.
- **return** : Fungsi ada yang memiliki *output* atau *return value* ada juga yang tidak. Jika fungsi memiliki *return value* maka *return value* dapat diproses lebih lanjut

Berikut adalah contoh penerapan *user define function*:

```
# Fungsi tanpa argument
bilang <- function(){
  print("Hello World!!")
}

# Print
bilang()
```

```
## [1] "Hello World!!"

# Fungsi dengan argumen
tambah <- function(a,b){
  print(a+b)
}

# Print
tambah(5,3)

## [1] 8

# Fungsi dengan return value
kali <- function(a,b){
  return(a*b)
}

# Print
kali(4,3)

## [1] 12
```

2.15 Referensi

1. Primartha, R. 2018. **Belajar Machine Learning Teori dan Praktik.** Penerbit Informatika : Bandung.
2. Rosadi,D. 2016. **Analisis Statistika dengan R.** Gadjah Mada University Press: Yogyakarta.
3. STHDA. **Easy R Programming Basics.** <http://www.sthda.com/english/wiki/easy-r-programming-basics>
4. Venables, W.N. Smith D.M. and R Core Team. 2018. **An Introduction to R.** R Manuals.
5. The R Core Team. 2018. **R: A Language and Environment for Statistical Computing.** R Manuals.

Chapter 3

Manajemen Data R

Data manajemen merupakan bagian penting dalam setiap proses analisa data. Proses import dan eksport data pada berbagai format penting untuk dipelajari. Selain itu, proses perapihan data sebelum analisa menjadi bagian yang harus ada pada awal proses analisa. Proses-proses tersebut akan kita ulas secara mendalam pada *chapter* ini. *Chapter* ini juga akan membahas bagaimana kita dapat melakukan sejumlah manipulasi data untuk memperoleh informasi lebih yang terkandung pada.

3.1 Import File

Pada sesi bagian ini penulis akan menjelaskan cara mengimport file pada R. File yang diimport ke dalam R terdiri atas file yang sering digunakan pada saat akan melakukan analisis data, antara lain: TXT, CSV, Excel, SPSS, SAS, dan STATA.

Pada bagian ini akan dijelaskan pula bagaimana melakukan import data menggunakan library `readr` serta kelebihan dari metode import data yang digunakan. Berikut adalah cara mengimport data berbagai format pada R.

Note: Pastikan kita telah mengatur lokasi *working directory* pada tempat dimana lokasi file yang akan kita baca berada untuk mempermudah dalam melakukan import file.

3.1.1 Import File Menggunakan Fungsi Bawaan R

Fungsi bawaan R secara umum hanya dapat membaca data dengan format TXT dan CSV. Pada RStudio fungsi ini bertambah dengan adanya library tambahan yang telah terinstall di RStudio untuk membaca file dengan format EXCEL, SPSS, SAS dan STATA.

Secara umum fungsi yang digunakan untuk membaca data dengan format tabel seperti TXT dan CSV adalah `fungsiread.table()`. Berikut adalah list fungsi dasar lainnya untuk membaca file dengan format TXT dan CSV pada R:

- `read.csv()`: untuk membaca file dengan format *comma separated value*(“.csv”).
- `read.csv2()`: varian yang digunakan jika pada file “.csv” yang akan dibaca mengandung koma (”,”) sebagai desimal dan semicolon (“;”) sebagai pemisah antar variabel atau kolom.
- `read.delim()`: untuk membaca file dengan format *tab-separated value*(“.txt”).
- `read.delim2()`: membaca file dengan format “.txt” dengan tanda koma (”,”) sebagai penunjuk bilangan desimal.

Masing-masing fungsi diatas dapat dituliskan kedalam R dengan format sebagai berikut:

```
# Membaca tabular data pada R
read.table(file, header = FALSE, sep = "", dec = ".")
# Membaca "comma separated value" files ("*.csv")
read.csv(file, header = TRUE, sep = ",", dec = ".", ...)
# atau gunakan read.csv2 jika tanda desimal pada data adalah "," dan pemisah kolom adalah ";"
read.csv2(file, header = TRUE, sep = ";", dec = ",", ...)
# Membaca TAB delimited files
read.delim(file, header = TRUE, sep = "\t", dec = ".", ...)
read.delim2(file, header = TRUE, sep = "\t", dec = ",", ...)
```

Note:

- **file:** nama file diakhiri dengan format file (misal: “nama_file.txt”) yang akan di import ke dalam file. Dapat pula diisi lokasi file tersebut berada, misal:(C:/Users/My PC/Documents/nama_file.txt atau .csv)
- **sep:** pemisah antar kolom. “\t” digunakan untuk tab-delimited file.
- **header:** nilai logik. jika TRUE, maka `read.table()` akan menganggap bahwa file yang akan dibaca pada baris pertama file merupakan header data.
- **dec:** karakter yang digunakan sebagai penunjuk desimal pada data.

Untuk info lebih lanjut terkait fungsi-fungsi tersebut dan contoh bagaimana menggunakananya, pembaca dapat mengakses fitur batuan dari fungsi tersebut menggunakan sintaks berikut:

```
# mengakses menu bantuan
?read.table
?read.csv
?read.csv2
?read.delim
?read.delim2
```

Misalkan penulis memiliki data pada file bernama “mtcars.csv” dengan desimal berupa titik pada datanya. Penulis ingin membaca file tersebut, maka penulis akan menuliskan sintaks berikut:

```
data <- read.csv("mtcars.csv")
```

Secara default perintah tersebut akan membaca baris pertama data sebagai header serta data berupa karakter menjadi factor. Untuk mencegah agar data berupa karakter menjadi faktor, perintah tersebut dapat ditambahkan parameter `stringAsFactor = FALSE`.

Kita juga dapat memilih file yang akan kita baca secara interaktif. Misal pada *working directory* terdapat beberapa file yang akan kita baca. Kita ingin melihat file dengan format tertentu yang hendak kita baca, namun kita malas mengecek file explorer pada windows. Untuk mengatasi masalah tersebut, kita dapat menggunakan fungsi `file.choose()` pada R. Fungsi tersebut akan menampilkan jendela windows explores sehingga kita dapat memilih file apa yang hendak dibaca. Berikut adalah contoh penerapannya:

```
data <- read.csv(file.choose())
```

Note: pastikan format file yang dibaca sama dengan fungsi import yang digunakan.

Kita juga dapat membaca file dari internet. Untuk melakukannya kita hanya perlu meng-copy url file tersebut. Berikut adalah contoh file yang dibaca dari internet:

```
# Membaca file dari internet
data <- read.delim("http://www.sthda.com/upload/boxplot_format.txt")

# mengecek 6 observasi awal
head(data)

##      Nom variable Group
## 1 IND1      10     A
## 2 IND2       7     A
## 3 IND3      20     A
## 4 IND4      14     A
## 5 IND5      14     A
## 6 IND6      12     A
```

3.1.2 Membaca File CSV dan TXT Menggunakan Library readr

Pada bagian sebelumnya kita telah belajar bagaimana cara membaca file dengan format CSV dan TXT menggunakan paket dasar R. Pada bagian ini penulis akan menjelaskan bagaimana cara membaca file dengan format TXT dan CSV pada R menggunakan paket **readr**.

readr dikembangkan oleh Hadley Wickham. paket **readr** memberikan solusi cepat dan ramah untuk membaca delimited file ke dalam R.

Dibandingkan dengan paket dasar R, **readr** memiliki kelebihan sebagai berikut:

- Mampu membaca file 10x lebih cepat dibandingkan pada paket bawaan R.
- Menampilkan *progress bar* yang bermanfaat jika proses pemuatan berlangsung agak lama.
- semua fungsi bekerja dengan cara yang persis sama dengan paket bawaan R.

Untuk dapat menggunakan **readr**, kita perlu menginstall paketnya terlebih dahulu. Untuk melakukannya jalankan sintaks berikut:

```
# Menginstall paket
install.packages("readr")

# Memuat paket
library(readr)
```

Berikut adalah format beberapa fungsi yang dapat digunakan:

```
# Fungsi umum (membaca TXT dan CSV) dapat juga membaca flat file dan tsv
read_delim(file, delim, col_names = TRUE)
# Membaca comma (",") separated values
read_csv(file, col_names = TRUE)
# Membaca semicolon (";") separated values
read_csv2(file, col_names = TRUE)
# Membaca tab separated values
read_tsv(file, col_names = TRUE)
```

Note:

- **file**: path file, koneksi atau raw vector. File yang berakhiran .gz, .bz2, .xz, atau .zip akan secara otomatis tidak terkompresi. File yang dimulai dengan “http://”, “https://”, “ftp://”, atau “ftps://” akan diunduh secara otomatis. File gz jarak jauh juga dapat diunduh & didekompresi secara otomatis.
- **delim**: karakter yang membatasi tiap nilai pada file.
- **col_names**: nilai logik. Jika TRUE, maka baris pertama akan menjadi header.

Berikut adalah contoh bagaimana cara membaca file menggunakan fungsi pada paket **readr**:

```
# Membaca file lokal
data <- read_csv("mtcars.csv")

# atau
data <- read_csv(file.choose())

# Membaca dari internet
data <- read_tsv("http://www.sthda.com/upload/boxplot_format.txt")
```

Kita juga dapat menspesifikasi jenis data pada kolom yang akan dibaca. Keuntungan dari penentuan jenis kolom (tipe data) akan memastikan data yang telah dibaca tidak salah berdasarkan jenis data pada masing-masing kolom.

Beberapa format jenis kolom yang tersedia pada **readr** adalah sebagai berikut:

- **col_integer()**: untuk menentukan integer (alias = “i”).
- **col_double()**: untuk menentukan kolom sebagai jenis data double (alias = “d”).
- **col_logical()**: untuk menentukan variabel logis (alias = “l”).
- **col_character()**: meninggalkan string apa adanya. Tidak mengonversinya menjadi faktor (alias = “c”).
- **col_factor()**: untuk menentukan variabel faktor (atau pengelompokan) (alias = “f”)
- **col_skip()**: untuk mengabaikan kolom (alias = “-” atau “_”)
- **col_date()** (alias = “D”), **col_datetime()** (alias = “T”) dan **col_time()** (“t”) untuk menentukan tanggal, waktu tanggal, dan waktu.

Berikut adalah contoh penerapannya:

```
data <- read_csv("my_file.csv", col_types = cols(
  x = "i", # kolom integer
  treatment = "c" # kolom karakter/string
))
```

3.1.3 Import File Excel Pada R

Keunggulan penggunaan excel sebagai format penyimpan data adalah kita dapat menyimpan banyak data dan memisahkannya pada lembar (*sheet*) yang berbeda sebagai suatu data yang independen dibandingkan pembacaan pada file csv yang hanya berisikan satu tabel data saja tiap file.

Pada R kita dapat melakukan pembacaan file menggunakan berbagai macam cara seperti menggunakan paket bawaan R maupun menggunakan library yang perlu kita install. Berikut adalah beberapa cara membaca file excel pada R.

- Mengkonversi terlebih dahulu satu sheet excel yang akan kita baca menjadi format “.csv” maupun “.txt” sehingga dapat dibaca seperti pada sub-bab 3.1.1.

- b. Menyalin data dari excel dan mengimport data pada R.

Cara ini sedikit mirip dengan cara sebelumnya, dimana kita perlu membuka file excel dan melakukan **select** dan **copy** (ctrl+c) tabel data yang hendak dibaca. Data tersebut selanjutnya akan tersimpan pada **clipboard**.

Data yang telah tersalin selanjutnya diimport ke R dengan mengetikkan sintaks berikut:

```
data <- read.table(file= "clipboard",
                   sep = "\t", header = TRUE)
```

Cara ini merupakan cara yang paling sering penulis gunakan. Kelemahan penggunaan cara ini adalah ketika kita melakukan proses **select** dan **copy** (ctrl+c) tabel yang jumlahnya sangat banyak dan terdapat teks-teks penjelasan terkait tabel data pada lembar kerja excel yang tidak ingin kita sertakan akan memakan waktu yang lebih lama pada proses **select**.

- c. Mengimport data menggunakan library readxl.

Paket **readxl**, yang dikembangkan oleh Hadley Wickham, dapat digunakan untuk dengan mudah mengimpor file Excel (xls | xlsx) ke R tanpa ada ketergantungan eksternal.

Untuk dapat menggunakan library **readxl** kita harus menginstallnya terlebih dahulu menggunakan sintaks berikut:

```
# Instal paket
install.packages("readxl")

# memuat paket
library(readxl)
```

Berikut adalah contoh cara mengimport data dengan format xls atau xlsx pada R.

```
# Tentukan sheet dengan nama sheet pada file
data <- read_excel("my_file.xlsx", sheet = "data")

# Tentukan sheet berdasarkan indeks sheet
data <- read_excel("my_file.xlsx", sheet = 2) # membaca sheet ke-2
```

- d. Mengimport data menggunakan library xlsx

Paket **xlsx**, solusi berbasis **java**, adalah salah satu paket R yang ampuh untuk membaca, menulis, dan memformat file Excel. Untuk dapat menggunakannya kita harus menginstall dan memuatnya terlebih dahulu. Berikut sintaks yang digunakan:

```
# Menginstall paket
install.packages("xlsx")

# Memuat paket
library(xlsx)
```

Terdapat dua buah fungsi yang disediakan pada paket tersebut yaitu **read.xlsx()** dan **read.xlsx2()**. Perbedaan keduanya adalah **read.xlsx2()** digunakan pada file data dengan ukuran yang besar serta proses pembacaan data yang lebih cepat dibandingkan dengan **read.xlsx()**. Format yang digunakan untuk kedua fungsi tersebut disajikan sebagai berikut:

```
read.xlsx(file, sheetIndex, header=TRUE)
read.xlsx2(file, sheetIndex, header=TRUE)
```

Note:

- **file**: nama atau lokasi file berada
- **sheetIndex**: Indeks dari sheet yang hendak dibaca
- **header**: nilai logik. Jika bernilai TRUE, maka baris pertama dari sheet menjadi header.

Berikut adalah contoh penggunaanya:

```
data <- read.xlsx(file.choose(), 1) # membaca sheet 1
```

Note: kita juga dapat membaca file dari internet seperti pada sub-bab 3.1.1.

3.1.4 Membaca File Dari Format Aplikasi Statistik

Untuk membaca file yang berasal dari format aplikasi statistik seperti SPSS, SAS, dan STATA kita perlu menginstal dan memuat paket-paket yang dibutuhkan sesuai dengan file yang akan kita install. Berikut adalah sintaks bagaimana cara mengimport file dari berbagai format aplikasi statistik.

```
# membaca file SPSS
install.packages("Hmisc") # menginstall paket
library(Hmisc) # memuat paket
# simpan SPSS dataset pada transport format
get file='c:\mydata.sav'.
export outfile='c:\mydata.por'.
data <- spss.get("c:\mydata.por", use.value.labels= TRUE)
# use.value.labels digunakan untuk mengubah label menjadi factor

# membaca file SAS
install.packages("Hmisc") # menginstall paket
library(Hmisc) # memuat paket
# simpan SAS dataset pada transport format
libname out xport 'c:/mydata.xpt';
data out.mydata;
set sasuser.mydata;
run;
data <- sasxport.get("c:/mydata.xpt")
# Variabel yang berupa karakter akan dikonversi menjadi factor

# membaca file STATA
install.packages("foreign") # menginstall paket
library(foreign) # memuat paket
data <- read.dta("c:/mydata.dta")
```

3.2 Eksport File

Setelah kita melakukan analisa dan telah memperoleh hasil yang kita inginkan dan memperoleh data frame berupa hasil prediksi suatu model atau data yang telah dibersihakan, kita ingin melakukan pelaporan dalam

bentuk file dengan format seperti EXCEL, CSV atau TXT. Untuk melakukannya kita perlu melakukan eksport data yang telah dihasilkan.

Pada bagian ini penulis akan menjelaskan bagaimana cara mengeksport data dari R kedalam format TXT, CSV, maupun EXCEL. Sebenarnya R memungkinkan untuk melakukan eksport dalam format lain seperti RDA maupun RDS yang tidak dibahas dalam buku ini karena berada diluar lingkup buku ini.

3.2.1 Eksport Data Menjadi Format TXT dan CSV

Terdapat dua cara untuk melakukan ekport data dari R menjadi format TXT atau CSV, yaitu melalui paket dasar R maupun menggunakan library `readr`. Kedua cara tersebut memiliki sejumlah kemiripan dari segi fungsi, namun berbeda dari segi kecepatan eksport.

Fungsi dasar yang digunakan pada R untuk melakukan eksport file kedalam format TXT dan CSv adalah `write.table()`. Format umum yang digunakan adalah sebagai berikut:

```
write.table(x, file, sep= " ", dec = ",",
            row.names = TRUE, col.names = TRUE)
```

Note:

- **x:** matriks atau data frame yang akan ditulisi.
- **file:** karakter yang menentukan nama file yang dihasilkan.
- **sep:** string pemisah bidang atau kolom, mis., `sep = "t"` (untuk nilai yang dipisahkan tab).
- **dec:** string yang akan digunakan sebagai pemisah desimal. Standarnya adalah `"."`.
- **row.names:** nilai logik yang menunjukkan apakah nama baris x harus ditulisi bersama dengan x, atau vektor karakter nama baris yang akan ditulisi.
- **col.names:** baik nilai logik yang menunjukkan apakah nama kolom x harus ditulisi bersama dengan x, atau vektor karakter nama kolom yang akan ditulisi. Jika `col.names = NA` dan `row.names = TRUE` ditambahkan nama kolom kosong, yang merupakan konvensi yang digunakan untuk file CSV untuk dibaca oleh spreadsheet.

Selain menggunakan fungsi tersebut, untuk eksport ke dalam format CSV juga dapa menggunakan fungsi `write.csv()` atau `write.csv2()`. Berikut adalah format yang digunakan:

```
write.csv(data, file="data.csv")
write.csv2(data, file="data.csv")
```

Secara penampakan kedua fungsi tersebut pada dasarnya sama dengan fungsi `write.table()`, bedanya adalah kedua fungsi tersebut spesifik digunakan untuk eksport file kedalam format CSV.

Note:

- `write.csv()` menggunakan `"."` sebagai titik desimal serta `,` sebagai pemisah antar kolom data.
- `write.csv2()` menggunakan `,` sebagai titik desimal serta `;` sebagai pemisah antar kolom data.

Misalkan kita ingin melakukan eksport data objek `mtcars` kedalam format CSV. Untuk melakukannya dapat dilakukan dengan sintaks berikut:

```
write.csv(mtcars, file="mtcars.csv", row.names = FALSE)
```

Note: Hasil eksport ditampilkan pada *working directory*

Kita juga dapat menggunakan fungsi `write_delim()` dari library `readr` untuk melakukan eksport data kedalam format CSV atau TXT. Berdasarkan format file yang hendak dihasilkan kita juga dapat menggunakan fungsi `write_csv()` atau `write_tsv()`. Berikut adalah penjelasan terkait kedua fungsi tersebut:

- `write_csv()`: untuk mengeksport kedalam format CSV.
- `write_tsv()`: untuk mengeksport kedalam format TXT.

Format sederhana ketiga fungsi fungsi tersebut adalah sebagai berikut:

```
# Fungsi umum
write_delim(x, path, delim = " ")
# Write comma (",") separated value files
write_csv(file, path)
# Write tab ("\t") separated value files
write_tsv(file, path)
```

Note:

- **x**: data frame yang akan ditulis
- **path**: path ke file hasil (dapat berupa nama file disertai ekstensi file yang akan dibuat)
- **delim**: Delimiter digunakan untuk memisahkan nilai. Harus karakter tunggal.

Berikut adalah contoh penerapan dari fungsi tersebut:

```
# memuat mtcars data
data(mtcars)
library(readr)

# eksport mtcars menjadi tsv atau txt
write_tsv(mtcars, path = "mtcars.txt")

# eksport mycars menjadi csv
write_csv(mtcars, path = "mtcars.csv")
```

3.2.2 Eksport Data Menjadi Format Excel

Untuk mengeksport data menjadi format EXCEL (“.xls” atau “.xlsx”) kita dapat menggunakan fungsi `write.xlsx()` dan `write.xlsx2()` dari library `xlsx`. Berikut adalah format sederhana yang digunakan:

```
write.xlsx(x, file, sheetName = "Sheet1",
           col.names = TRUE, row.names = TRUE, append = FALSE)
write.xlsx2(x, file, sheetName = "Sheet1",
           col.names = TRUE, row.names = TRUE, append = FALSE)
```

Note:

- **x**: sebuah data frame untuk ditulis ke dalam worksheet.
- **file**: path ke file output.
- **sheetName**: string karakter yang digunakan untuk nama sheet.
- **col.names, row.names**: nilai logik yang menentukan apakah nama kolom / nama baris x akan ditulis ke file.
- **append**: nilai logis yang menunjukkan apakah x harus ditambahkan ke file yang ada.

Berikut adalah contoh penerapannya:

```
library("xlsx")
# Menuliskan dataset pertama pada workbook
write.xlsx(USArrests, file = "myworkbook.xlsx",
           sheetName = "USA-ARRESTS", append = FALSE)
# Menambahkan dataset kedua pada workbook
write.xlsx(mtcars, file = "myworkbook.xlsx",
           sheetName="MTCARS", append=TRUE)
# Menambahkan dataset kedua pada workbook
write.xlsx(iris, file = "myworkbook.xlsx",
           sheetName="IRIS", append=TRUE)
```

3.3 Tibble Data Format

Tibble adalah data frame yang menyediakan metode print yang lebih bagus, berguna saat bekerja dengan kumpulan data besar. Pada bagian ini penulis akan menjelaskan penggunaan tibble sebagai alternatif kita dalam berinteraksi dengan data frame.

Untuk membuat tibble kita perlu menginstall dan memuat library **tibble** yang dikembangkan oleh **Hadley Wickham**. Berikut adalah sintaks yang digunakan:

```
# menginstall paket
install.packages("tibble")

# memuat paket
library(tibble)
```

3.3.1 Membuat Tibble

Untuk dapat membuat tibble kita dapat melakukan konversi data frame yang sudah ada menjadi tibble menggunakan fungsi **as_tibble()**. Berikut adalah contoh bagaimana membuat tibble menggunakan data **iris**:

```
## Warning: package 'tibble' was built under R version
## 3.5.3
```

```
# memuat data mtcars
data("iris")

# print
head(iris, 10)
```

```

##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1         5.1        3.5       1.4        0.2
## 2         4.9        3.0       1.4        0.2
## 3         4.7        3.2       1.3        0.2
## 4         4.6        3.1       1.5        0.2
## 5         5.0        3.6       1.4        0.2
## 6         5.4        3.9       1.7        0.4
## 7         4.6        3.4       1.4        0.3
## 8         5.0        3.4       1.5        0.2
## 9         4.4        2.9       1.4        0.2
## 10        4.9       3.1       1.5        0.1
##   Species
## 1   setosa
## 2   setosa
## 3   setosa
## 4   setosa
## 5   setosa
## 6   setosa
## 7   setosa
## 8   setosa
## 9   setosa
## 10  setosa

# konversi mtcars menjadi tibble
iris_tbl <- as_tibble(iris)

# print
iris_tbl

## # A tibble: 150 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
##       <dbl>      <dbl>      <dbl>      <dbl>
## 1         5.1        3.5       1.4        0.2
## 2         4.9        3.0       1.4        0.2
## 3         4.7        3.2       1.3        0.2
## 4         4.6        3.1       1.5        0.2
## 5         5.0        3.6       1.4        0.2
## 6         5.4        3.9       1.7        0.4
## 7         4.6        3.4       1.4        0.3
## 8         5.0        3.4       1.5        0.2
## 9         4.4        2.9       1.4        0.2
## 10        4.9       3.1       1.5        0.1
## # ... with 140 more rows, and 1 more variable:
## #   Species <fct>

```

Note: Kita dapat mengkonversi tibble menjadi data frame menggunakan fungsi `as.data.frame()`

Secara default saat kita print tibble, maka akan dimunculkan 10 observasi pertama. Pada data frame biasa jika kita print data tersebut maka seluruh observasi akan ditampilkan.

Penggunaan tibble ini cenderung menguntungkan saat kita bekerja dengan jumlah data yang besar dan ingin mengecek observasi yang ada. Hal ini berbeda dengan data frame biasa dimana untuk mengecek observasi

awal kita perlu menggunakan fungsi `head()` agar seluruh data tidak ditampilkan. Sehingga penggunaan tibble cenderung membuat proses analisa menjadi lebih rapi.

Kita juga dapat membuat tibble dari kumpulan sejumlah vektor menggunakan fungsi `tibble()`. `tibble()` akan secara otomatis mendaur ulang input dengan panjang 1 (variabel `y`), dan memungkinkan kita untuk merujuk ke variabel yang baru saja kita buat, seperti yang ditunjukkan pada sintaks berikut:

```
tibble(
  x = 1:20,
  y = 1,
  z = 2*x+5*y
)

## # A tibble: 20 x 3
##       x     y     z
##   <int> <dbl> <dbl>
## 1     1     1     7
## 2     2     1     9
## 3     3     1    11
## 4     4     1    13
## 5     5     1    15
## 6     6     1    17
## 7     7     1    19
## 8     8     1    21
## 9     9     1    23
## 10    10    1    25
## 11    11    1    27
## 12    12    1    29
## 13    13    1    31
## 14    14    1    33
## 15    15    1    35
## 16    16    1    37
## 17    17    1    39
## 18    18    1    41
## 19    19    1    43
## 20    20    1    45
```

Jika pembaca telah mulai familiar dengan fungsi `data.frame()`, perlu diingat bahwa `tibble()` melakukan lebih sedikit: tidak pernah mengubah jenis input (mis., tidak pernah mengubah string menjadi faktor!), tidak pernah mengubah nama variabel, dan tidak pernah membuat nama baris seperti yang biasa terjadi saat kita menggunakan fungsi `data.frame()`.

Cara lain yang dapat digunakan untuk membuat tibble adalah dengan menggunakan fungsi `tribble()` yang merupakan singkatan dari *transposed tibble*. `tribble()` dikustomisasi untuk entri data dalam kode: judul kolom didefinisikan oleh rumus (yaitu, mereka mulai dengan `~`), dan entri dipisahkan oleh koma. Hal ini memungkinkan untuk menata sejumlah kecil data dalam bentuk yang mudah dibaca. Berikut adalah contoh penerapannya:

```
tribble(
  ~x, ~y, ~z,
  #--/--/---
  "a", 2, 5,
  "b", 5, 7
)
```

```
## # A tibble: 2 x 3
##   x     y     z
##   <chr> <dbl> <dbl>
## 1 a      2     5
## 2 b      5     7
```

Penambahan komen (#/-/-) dilakukan untuk memperjelas posisi dari header sehingga meminimalisir kesalahan dalam input data.

3.3.2 Tibble vs Data Frame

terdapat dua buah perbedaan utama antara tibble dan data frame , yaitu: *printing* dan *subsetting*.

a. Printing

Tibbles memiliki metode print halus yang hanya menampilkan 10 baris pertama observasi, dan semua kolom yang sesuai dengan lebar layar. Ini membuatnya lebih mudah untuk bekerja dengan data besar. Selain namanya, setiap kolom melaporkan jenis datanya, fitur bagus yang dipinjam dari fungsi `str()`. Berikut adalah contohnya:

```
tibble(
  ~x, ~y, ~z,
  #--/---/-----
  "a", 2.1, FALSE,
  "b", 5.5, TRUE
)
```

```
## # A tibble: 2 x 3
##   x     y     z
##   <chr> <dbl> <lgl>
## 1 a      2.1 FALSE
## 2 b      5.5 TRUE
```

Tibbles dirancang agar kita tidak secara sengaja menampilkan data yang sangat banyak saat melakukan perintah `print()`. Tetapi terkadang kita membutuhkan lebih banyak output daripada tampilan default. Ada beberapa opsi yang dapat membantu.

Pertama, kita dapat secara eksplisit melakukan print data frame dan mengontrol jumlah baris (n) dan lebar tampilan. `width = Inf` akan menampilkan semua kolom. Berikut adalah contoh penerapannya

```
print(iris_tbl, n=15, width=Inf)
```

```
## # A tibble: 150 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
##       <dbl>        <dbl>        <dbl>        <dbl>
## 1         5.1         3.5         1.4         0.2
## 2         4.9         3.0         1.4         0.2
## 3         4.7         3.2         1.3         0.2
## 4         4.6         3.1         1.5         0.2
## 5         5.0         3.6         1.4         0.2
## 6         5.4         3.9         1.7         0.4
```

```

##   7       4.6      3.4      1.4      0.3
##   8       5        3.4      1.5      0.2
##   9       4.4      2.9      1.4      0.2
##  10      4.9      3.1      1.5      0.1
##  11      5.4      3.7      1.5      0.2
##  12      4.8      3.4      1.6      0.2
##  13      4.8        3      1.4      0.1
##  14      4.3        3      1.1      0.1
##  15      5.8        4      1.2      0.2
## # Species
## <fct>
## 1 setosa
## 2 setosa
## 3 setosa
## 4 setosa
## 5 setosa
## 6 setosa
## 7 setosa
## 8 setosa
## 9 setosa
## 10 setosa
## 11 setosa
## 12 setosa
## 13 setosa
## 14 setosa
## 15 setosa
## # ... with 135 more rows

```

Kita juga dapat mengontrol print default dengan melakukan pengaturan menggunakan fungsi `options()`. Berikut adalah contoh penerapannya:

- `options(tibble.print_max= n, tibble.print_min= m)`: jika terdapat lebih dari “m” baris, print hanya sejumlah “n” baris.
- `options(dplyr.print_min = Inf)`: untuk selalu menampilkan seluruh baris. Perlu diingat fungsi ini dapat digunakan saat kita telah memuat library `dplyr`.
- `options(tibble.width = Inf)`: menampilkan seluruh kolom tanpa mempedulikan lebar tampilan layar.

Cara terakhir untuk menampilkan seluruh observasi adalah dengan fungsi `view()`. Berikut adalah contoh penerapannya pada data `iris_tbl`:

```
view(iris_tbl)
```

b. Subsetting

Sejauh ini semua alat yang kita pelajari telah bekerja dengan data frame yang lengkap. Jika kita ingin mengeluarkan variabel tunggal, kita memerlukan beberapa alat baru, dollar sign (\$) dan [[]. [[dapat mengekstraksi berdasarkan nama atau posisi; \$ hanya mengekstraksi berdasarkan nama. Berikut adalah contoh penerapannya:

```
# print tibble
iris_tbl
```

```
## # A tibble: 150 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
##       <dbl>      <dbl>      <dbl>      <dbl>
## 1         5.1        3.5       1.4       0.2
## 2         4.9        3.0       1.4       0.2
## 3         4.7        3.2       1.3       0.2
## 4         4.6        3.1       1.5       0.2
## 5         5.0        3.6       1.4       0.2
## 6         5.4        3.9       1.7       0.4
## 7         4.6        3.4       1.4       0.3
## 8         5.0        3.4       1.5       0.2
## 9         4.4        2.9       1.4       0.2
## 10        4.9        3.1       1.5       0.1
## # ... with 140 more rows, and 1 more variable:
## #   Species <fct>
```

```
# subset berdasarkan nama kolom
iris_tbl$Sepal.Length
```

```
## [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8
## [13] 4.8 4.3 5.8 5.7 5.4 5.1 5.7 5.1 5.4 5.1 4.6 5.1
## [25] 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0
## [37] 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6
## [49] 5.3 5.0 7.0 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2
## [61] 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1
## [73] 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0
## [85] 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7
## [97] 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3
## [109] 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0
## [121] 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1 6.4 7.2 7.4 7.9
## [133] 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
## [145] 6.7 6.7 6.3 6.5 6.2 5.9
```

```
#subset berdasarkan posisi
iris_tbl[[1]]
```

```
## [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8
## [13] 4.8 4.3 5.8 5.7 5.4 5.1 5.7 5.1 5.4 5.1 4.6 5.1
## [25] 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0
## [37] 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6
## [49] 5.3 5.0 7.0 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2
## [61] 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1
## [73] 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0
## [85] 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7
## [97] 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3
## [109] 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0
## [121] 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1 6.4 7.2 7.4 7.9
## [133] 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
## [145] 6.7 6.7 6.3 6.5 6.2 5.9
```

Dibandingkan dengan data frame, tibble lebih ketat: tibble tidak pernah melakukan *partial matching*, dan mereka akan menghasilkan peringatan jika kolom yang kita coba akses tidak ada.

3.4 Merapikan Data

Sebelum memulai analisa terhadap data yang kita miliki, umumnya kita akan merapikan data yang akan kita gunakan. Tujuannya adalah agar data yang akan digunakan sudah siap untuk dilakukan analisa dengan software tertentu seperti R, dimana pada dataset perlu jelas antara variabel dan nilai (*value*), serta untuk mempermudah dalam memperoleh informasi pada data. Berikut adalah beberapa contoh dataset yang dapat pembaca cermati terkait manakah data yang telah rapi (*tidy data*) dan mana yang belum (*messy data*):

```
# Install paket dataset EDAWR
# install.packages("devtools")
# devtools::install_github("rstudio/EDAWR")

# hilangkan tanda # jika pembaca belum menginstall

library(EDAWR)

##
## Attaching package: 'EDAWR'

## The following objects are masked _by_ '.GlobalEnv':
##      a, b, y, z

# memuat dataset
storms <- EDAWR::storms
cases

##   country 2011 2012 2013
## 1     FR  7000  6900  7000
## 2     DE  5800  6000  6200
## 3     US 15000 14000 13000

pollution

##       city size amount
## 1 New York large    23
## 2 New York small    14
## 3 London large     22
## 4 London small     16
## 5 Beijing large    121
## 6 Beijing small     56
```

Sebelum kita melakukan analisa di dataset tersebut, kita harus tahu terlebih dahulu apa saja syarat suatu dataset dikatakan rapi (*tidy*). Berikut adalah syaratnya:

- Setiap variabel harus memiliki kolomnya sendiri
- Setiap observasi harus memiliki barisnya sendiri
- Setiap nilai berada pada sel tersendiri

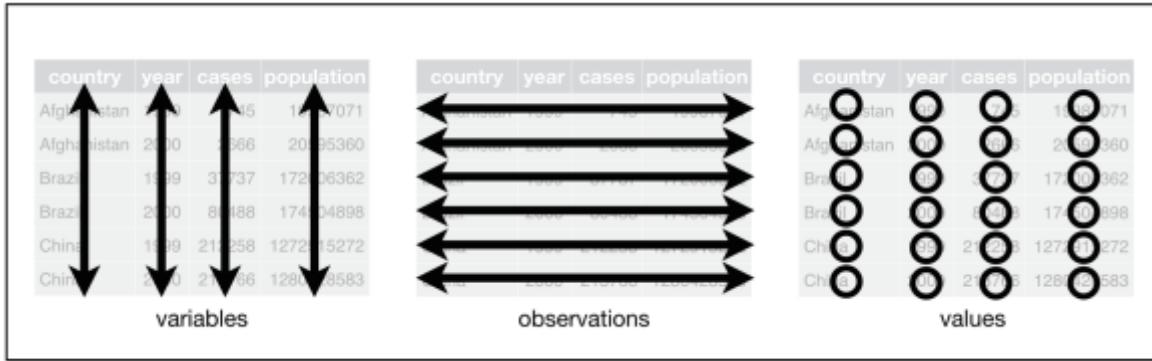


Figure 3.1: Visualisasi 3 rule tidy data

Ketiga syarat tersebut saling berhubungan sehingga jika salah satu syarat tersebut tidak terpenuhi, maka dataset belum bisa dikatakan *tidy*. Ketiga syarat tersebut dapat divisualisasikan melalui Gambar 3.1

Pada dataset `storms` terdapat 4 buah kolom dan 6 buah baris. Masing-masing kolom menyatakan variabel pada masing-masing observasi seperti nama badi , kecepatan angin, tekanan dan waktu . Ketiga syarat kerapihan data sudah terpenuhi pada data tersebut sehingga kita bisa melakukan analisa terhadap data tersebut, misalnya kecepatan angin dan tekanan pada masing-masing badi. Selain itu kita juga dapat dengan mudah menginput variabel baru pada dataset tersebut, misal: rasio (kecepatan angin/tekanan).

Berikut adalah contoh bagaimana kita dapat dengan mudah menarik nilai variabel pada masing-masing kolom dan membentuk variabel baru pada dataset tersebut:

```
# subset variabel
storms$storm

## [1] "Alberto" "Alex"      "Allison"  "Ana"       "Arlene"
## [6] "Arthur"

storms$wind

## [1] 110  45  65  40  50  45

storms$pressure

## [1] 1007 1009 1005 1013 1010 1010

storms$date

## [1] "2000-08-03" "1998-07-27" "1995-06-03" "1997-06-30"
## [5] "1999-06-11" "1996-06-17"

# membuat variabel baru
storms_new <- storms
storms_new$ratio <- storms_new$wind/storms_new$pressure
storms_new
```

```
##      storm wind pressure      date   ratio
## 1 Alberto  110     1007 2000-08-03 0.10924
## 2 Alex     45      1009 1998-07-27 0.04460
## 3 Allison  65      1005 1995-06-03 0.06468
## 4 Ana      40      1013 1997-06-30 0.03949
## 5 Arlene   50      1010 1999-06-11 0.04950
## 6 Arthur   45      1010 1996-06-17 0.04455
```

Pada dataset `cases` terdapat 3 buah kolom dan 3 baris. Pada kolom pertama berupa kode Negara, sedangkan kolom sisanya merupakan tahun. Jika kita perhatikan dengan seksama dataset tersebut merupakan sebuah *contingency table* dimana tabel tersebut menyatakan frekuensi kejadian pada tahun tertentu dan negara tertentu. Dataset tersebut belum dapat dikatakan *tidy* karena kolom 2011 sampai 2013 merupakan sebuah nilai dari observasi dan bukan sebuah variabel sehingga dataset tersebut masih tergolong dataset *messy*. Selain itu sangat sulit untuk dilakukan penarikan terhadap nilai pada setiap kolom serta pembentukan variabel baru sebagai pendukung analisa juga sulit dilakukan. Berikut adalah contoh melakukan penarikan nilai / subset pada masing variabel:

```
cases$country
```

```
## [1] "FR" "DE" "US"
```

```
names(cases[-1])
```

```
## [1] "2011" "2012" "2013"
```

```
unlist(cases[1:3, 2:4])
```

```
## 20111 20112 20113 20121 20122 20123 20131 20132 20133
## 7000  5800 15000  6900  6000 14000  7000  6200 13000
```

Pada dataset `pollution`terdapat 3 buah kolom dan 6 baris. Masing-masing kolom menyatakan lokasi berupa nama kota, keterangan ukuran partikel, serta nilai dari ukuran partikel. Beberapa dari kita mungkin menganggap dataset ini telah memenuhi syarat kerapihan data. Namun, coba kita cermati jika kita ingin membuat variabel baru terkait dengan berapa rentang ukuran partikel (range ukuran partikel) pada masing-masing kota. Hal tersebut tentu sangat sulit dilakukan pada dataset tersebut, namun dataset tersebut memungkinkan kita dengan mudah mengambil nilai dari masing-masing variabelnya seperti contoh berikut:

```
pollution$city
```

```
## [1] "New York" "New York" "London"    "London"
## [5] "Beijing"  "Beijing"
```

```
pollution$size
```

```
## [1] "large" "small" "large" "small" "large" "small"
```

```
pollution$amount
```

```
## [1] 23 14 22 16 121 56
```

Berdasarkan contoh-contoh tersebut pada pembahasan kali ini penulis akan menjelaskan bagaimana cara melakukan perapihan data menggunakan library `tidyr`. Sebelum kita melakukannya berikut adalah sintaks untuk menginstall library tersebut:

```
# memasang paket
install.packages("tidyr")

# memuat paket
library(tidyr)

## Warning: package 'tidyr' was built under R version
## 3.5.3
```

3.4.1 Gather

Pada dataset `cases` kolom 2011 sampai 2013 perlu dijadikan satu variabel yaitu tahun. Untuk melakukannya kita dapat menggunakan fungsi `gather()`. Secara sederhana fungsi tersebut dapat dituliskan dengan format sebagai berikut:

```
gather(data, key, value, ...)
```

Note:

- **data:** data frame
- **key, value:** nama kunci dan kolom nilai yang akan dibuat di output
- **...:** Spesifikasi kolom untuk dikumpulkan. Nilai yang diizinkan adalah:
 - nama variabel
 - jika kita ingin memilih semua variabel antara a dan e, gunakan `a:e`
 - jika kita ingin mengecualikan nama kolom y gunakan `-y`
 - untuk opsi lainnya, lihat: `dplyr::select()`

Berikut adalah contoh penerapannya pada dataset `cases`:

```
# Ubah dataset cases menjadi tibble simpan sebagai objek cases_new
library(tibble)
cases_tbl <- as_tibble(cases)

# print
cases_tbl

## # A tibble: 3 x 4
##   country `2011` `2012` `2013`
##   <chr>    <dbl>   <dbl>   <dbl>
## 1 FR        7000    6900    7000
## 2 DE        5800    6000    6200
## 3 US       15000   14000   13000

# gather
cases_new <- gather(cases_tbl,
                      # variabel kunci
```

```

key = "year",
# nilai variabel
value = "frequency",
# kecualikan kolom country
-country)

# print
cases_new

## # A tibble: 9 x 3
##   country year  frequency
##   <chr>    <chr>     <dbl>
## 1 FR      2011     7000
## 2 DE      2011     5800
## 3 US      2011    15000
## 4 FR      2012     6900
## 5 DE      2012     6000
## 6 US      2012    14000
## 7 FR      2013     7000
## 8 DE      2013     6200
## 9 US      2013    13000

```

Berdasarkan hasil yang diperoleh terlihat bahwa variabel tahun memiliki jenis data karakter. Jenis data ini masih belum sesuai sehingga perlu dikonversi agar menjadi jenis data numerik (*dbl = double*). Untuk melakukannya jalankan sintaks berikut:

```

# Ubah jenis variabel tahun menjadi numerik
cases_new$year <- as.numeric(cases_new$year)
cases_new

```

```

## # A tibble: 9 x 3
##   country year  frequency
##   <chr>    <dbl>     <dbl>
## 1 FR      2011     7000
## 2 DE      2011     5800
## 3 US      2011    15000
## 4 FR      2012     6900
## 5 DE      2012     6000
## 6 US      2012    14000
## 7 FR      2013     7000
## 8 DE      2013     6200
## 9 US      2013    13000

```

Data yang diperoleh sekarang telah rapi (*tidy*), sehingga sudah siap untuk dilakukan analisa data.

3.4.2 Spread

Fungsi `spread()` berkebalikan dengan `gather()`. Fungsi `gather()` menggabungkan beberapa kolom menjadi 2 buah kolom kunci sedangkan `spread()` merubah dua kolom menjadi beberapa kolom. Format sederhananya adalah sebagai berikut:

Note:

- **data**: data frame
- **key**: nama kolom yang akan dijadikan heading pada kolom baru
- **value**: nama kolom yang nilainya akan mengisi setiap sel

Pada contoh kasus pada data `pollution`, kita dapat memisahkan kolom 2 menjadi kolom baru yaitu kolom `big size` dan `small size`. Untuk melakukannya jalankan sintaks berikut:

```
# merubah objek pollution menjadi tibble
pollution_tbl <- as_tibble(pollution)
```

```
# print
pollution_tbl
```

```
## # A tibble: 6 x 3
##   city      size  amount
##   <chr>    <chr>  <dbl>
## 1 New York large    23
## 2 New York small    14
## 3 London    large    22
## 4 London    small    16
## 5 Beijing   large   121
## 6 Beijing   small    56
```

```
# spread
pollution_new <- spread(pollution_tbl,
                         key = size,
                         value = amount)
```

```
#print
pollution_new
```

```
## # A tibble: 3 x 3
##   city      large small
##   <chr>    <dbl> <dbl>
## 1 Beijing   121    56
## 2 London    22     16
## 3 New York  23     14
```

Terlihat bahwa data `pollution` tampak memenuhi syarat kerapihan data (*tidy*). Kita sekarang dapat meng-input variabel baru dan melakukan analisa terhadap data tersebut. Berikut adalah contoh penerapannya:

```
# input variabel range (large-small)
pollution_new$range <- pollution_new$large - pollution_new$small
```

```
# print
pollution_new
```

```
## # A tibble: 3 x 4
##   city      large small range
##   <chr>    <dbl> <dbl> <dbl>
## 1 Beijing   121    56    65
## 2 London    22     16     6
## 3 New York  23     14     9
```

Berdasarkan hasil yang diperoleh diketahui bahwa nilai range ukuran partikel terbesar berada di Kota Beijing.

3.4.3 Separate

Fungsi `separate()` merupakan fungsi yang digunakan untuk memisahkan sejumlah nilai pada sebuah kolom menjadi beberapa kolom berdasarkan karakter pemisah yang ada di dalam nilai suatu kolom. Fungsi ini berbeda dengan fungsi sebelumnya seperti `gather()` dan `spread()` yang menggabung atau memisahkan 2 atau beberapa kolom. Format sederhana fungsi `separate()` adalah sebagai berikut:

```
separate(data, col, into, sep = "[[:alnum:]]+", convert= TRUE)
```

Note:

- **data:** data frame.
- **col:** Nama kolom yang tidak dikutip.
- **into:** Vektor karakter menentukan nama variabel baru yang akan dibuat.
- **sep:** Pemisah antar kolom:
- Jika karakter, diartikan sebagai ekspresi reguler. Jika numerik, diartikan sebagai posisi untuk dibelah. Nilai-nilai positif mulai dari 1 di ujung kiri string; nilai negatif mulai dari -1 di ujung kanan string.
- **convert:** nilai logik. Jika bernilai TRUE maka kolom baru yang akan diperoleh akan dikonversi berdasarkan jenis data yang seharusnya.

Pada dataset `storms` kita ingin memisahkan kolom `date` menjadi beberapa kolom seperti `year`, `month`, dan `day`. Kita dapat menggunakan fungsi `separate()` untuk memisahkan nilai pada kolom tersebut berdasarkan karakter pemisah pada nilai kolom tersebut dalam hal ini adalah “-”. Berikut adalah cara melakukannya:

```
# merubah storms menjadi tibble
storms_tbl <- as_tibble(storms)

# print
storms_tbl

## # A tibble: 6 x 4
##   storm    wind pressure date
##   <chr>    <int>     <int> <date>
## 1 Alberto    110      1007 2000-08-03
## 2 Alex       45       1009 1998-07-27
## 3 Allison    65       1005 1995-06-03
## 4 Ana        40       1013 1997-06-30
## 5 Arlene     50       1010 1999-06-11
## 6 Arthur     45       1010 1996-06-17

# separate
storms_new <- separate(storms_tbl,
                       col = date,
                       into = c("year", "month", "days"),
                       sep = "-",
                       convert = TRUE)

# print
storms_new
```

```
## # A tibble: 6 x 6
##   storm    wind pressure year month  days
##   <chr>    <int>    <int> <int> <int>
## 1 Alberto    110     1007  2000     8     3
## 2 Alex       45      1009  1998     7    27
## 3 Allison    65      1005  1995     6     3
## 4 Ana        40      1013  1997     6    30
## 5 Arlene     50      1010  1999     6    11
## 6 Arthur     45      1010  1996     6    17
```

Berdasarkan hasil yang diperoleh terlihat bahwa data telah terpisah dengan benar yang ditunjukkan dari nilai yang terpisah dan jenis data yang dihasilkan.

3.4.4 Unite

Fungsi `unite()` merupakan kebalikan dari fungsi `separate()`, dimana fungsi ini menggabungkan sejumlah kolom menjadi 1 kolom. Format sederhana untuk melakukannya disajikan sebagai berikut:

```
unite(data, col, ..., sep = "_")
```

Note:

- **data**: data frame.
- **col**: nama kolom baru (tanpa tanda kutip) untuk ditambahkan.
- **sep**: pemisah yang akan digunakan pada antar nilai.

Pada dataset `storms_new` kita ingin menggabungkan kembali kolom `year`, `month`, dan `days` dengan karakter pemisah “/”. Berikut adalah cara melakukannya:

```
# unite
storms_old <- unite(storms_new,
                     col = "date",
                     year, month, days,
                     sep = "-")

# print
storms_old
```

```
## # A tibble: 6 x 4
##   storm    wind pressure date
##   <chr>    <int>    <int> <chr>
## 1 Alberto    110     1007 2000-8-3
## 2 Alex       45      1009 1998-7-27
## 3 Allison    65      1005 1995-6-3
## 4 Ana        40      1013 1997-6-30
## 5 Arlene     50      1010 1999-6-11
## 6 Arthur     45      1010 1996-6-17
```

```
# ubah jenis kolom menjadi date
storms_old$date <- as.Date(storms_old$date)
```

```
# print
storms_old
```

```
## # A tibble: 6 x 4
##   storm    wind pressure date
##   <chr>    <int>    <int> <date>
## 1 Alberto    110     1007 2000-08-03
## 2 Alex       45      1009 1998-07-27
## 3 Allison    65      1005 1995-06-03
## 4 Ana        40      1013 1997-06-30
## 5 Arlene     50      1010 1999-06-11
## 6 Arthur     45      1010 1996-06-17
```

3.5 Transformasi Data

Data frame merupakan struktur data utama dalam statistik dan dalam R. Struktur dasar data frame ialah ada satu observasi tiap baris dan setiap kolom mewakili variabel, ukuran, fitur, atau karakteristik pengamatan itu yang telah dijelaskan pada bagian sebelumnya. R memiliki implementasi internal data frame yang kemungkinan besar akan kita gunakan paling sering. Namun, ada paket di CRAN yang mengimplementasikan data frame layaknya basis data relasional yang memungkinkan kita untuk beroperasi pada data frame yang sangat besar.

Mengingat pentingnya mengelola dat frame, penting bagi kita untuk memiliki alat yang baik untuk melakukannya. R memiliki beberapa paket seperti fungsi `subset()` dan penggunaan operator “[” dan “\$” untuk mengekstrak himpunan bagian dari frame data. Namun, operasi lain, seperti pemfilteran, pengurutan, dan pengelompokan data, seringkali dapat menjadi operasi yang membosankan di R yang sintaksisnya tidak terlalu intuitif. Paket `dplyr` dirancang untuk mengurangi banyak masalah ini dan menyediakan serangkaian rutinitas yang dioptimalkan secara khusus untuk menangani data frame.

3.5.1 Paket dplyr

Paket `dplyr` dikembangkan oleh **Hadley Wickham** dari **RStudio** dan merupakan versi yang dioptimalkan dari paket `plyr`-nya. Paket `dplyr` tidak menyediakan fungsionalitas baru untuk R sendiri, dalam arti bahwa semua yang dilakukan `dplyr` sudah dapat dilakukan dengan fungsi basis R, tetapi sangat menyederhanakan fungsi yang ada di R.

Salah satu kontribusi penting dari paket `dplyr` adalah ia menyediakan “*grammar*” (khususnya, kata kerja) untuk manipulasi data dan untuk beroperasi pada data frame. Melalui *grammar* ini, kita dapat berkomunikasi dengan masuk akal apa yang telah kita lakukan terhadap data frame dapat pula dipahami orang lain (dengan asumsi mereka juga tahu *grammar*-nya). Hal ini berguna karena memberikan abstraksi untuk manipulasi data yang sebelumnya tidak ada. Kontribusi lain yang bermanfaat adalah bahwa fungsi `dplyr` sangat cepat, karena banyak operasi utama dikodekan dalam C++.

Pada bagian ini pembaca akan belajar **6** fungsi utama yang ada pada paket `dplyr`. Fungsi tersebut antara lain:

1. Mengambil sejumlah observasi berdasarkan nilainya (`filter()`).
2. Mengurutkan kembali baris data frame berdasarkan nilai pada sebuah atau beberapa variabel (`arrange()`).
3. Mengambil atau subset terhadap sebuah atau beberapa variabel berdasarkan nama variabel/kolom (`select()`).
4. Membuat variabel baru atau menambahkan kolom baru (`mutate()`).
5. Membuat ringkasan terhadap data frame (`summarize()`).
6. Mengelompokkan operasi berdasarkan grup data (`group_by()`).

Keseluruhan fungsi tersebut format fungsi yang seragam, yaitu:

1. Argumen pertama adalah data frame.
2. Argumen selanjutnya adalah deskripsi yang akan dilakukan terhadap data frame (filter, pengurutan kembali, membuat ringkasan, dll) menggunakan nama variabel (tanpa tanda kutip).
3. Hasil operasi yang diperoleh adalah data frame baru.

Untuk menginstall dan memuat paket `dplyr` jalankan sintaks berikut:

```
# Memasang paket
install.packages("dplyr")

# memuat paket
library(dplyr)
```

3.5.2 filter()

Fungsi `filter()` digunakan untuk mengekstrak himpunan bagian (subset) baris dari data frame. Fungsi ini mirip dengan fungsi `subset()` yang ada di R. Secara sederhana format fungsi `filter()` dapat dituliskan sebagai berikut:

```
filter(data, ....)
```

Note:

- **data** : data frame
- : Predikat logis didefinisikan dalam istilah variabel dalam **data**. Beberapa kondisi digabungkan dengan & (lihat Chapter 2 opeator relasi dan operator logika. Hanya baris tempat kondisi bernilai TRUE disimpan.

Misalkan kita akan melakukan filter terhadap data frame `pollution_tbl` terhadap variabel `size` dengan kriteria `large` dan `amount > 12`. Berikut adalah sintaks yang digunakan:

```
filter(pollution_tbl, size=="large" & amount > 12)

## # A tibble: 3 x 3
##   city     size  amount
##   <chr>    <chr> <dbl>
## 1 New York large     23
## 2 London    large     22
## 3 Beijing   large    121
```

Jika menggunakan paket dasar R:

```
subset(pollution_tbl, size=="large" & amount > 12)

## # A tibble: 3 x 3
##   city     size  amount
##   <chr>    <chr> <dbl>
## 1 New York large     23
## 2 London    large     22
## 3 Beijing   large    121
```

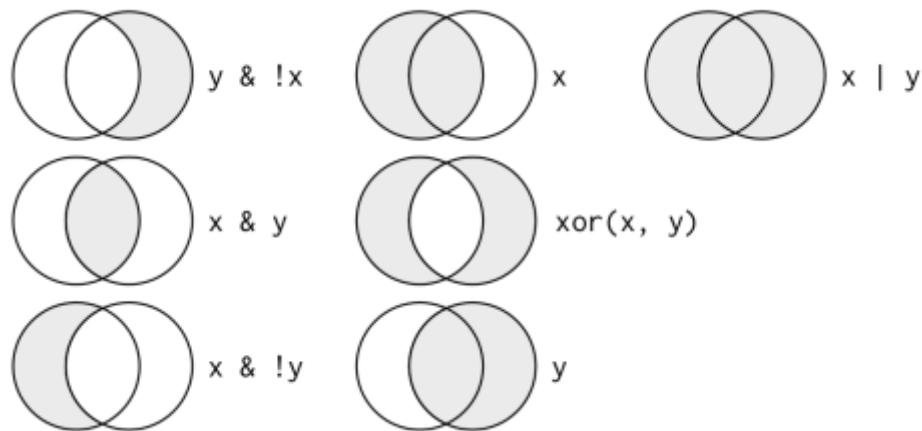


Figure 3.2: Diagram operasi Boolean

Operator “`>`” merupakan operator relasi (lihat chapter 2: operator relasi). Operator tersebut banyak digunakan untuk melakukan filter terhadap variabel/kolom yang mengandung nilai numerik.

Operator “`==`” merupakan operator logika (lihat chapter 2: operator logika). Operator tersebut digunakan untuk melakukan filter terhadap sejumlah syarat atau kondisi yang kita tetapkan. Jika nilai yang dihasilkan TRUE, maka hanya observasi tersebut yang akan ditampilkan. Untuk lebih memahami penerapan masing-masing operator logika pada proses filter perhatikan Gambar 3.2 berikut:

Note: Bagian yang diarsir adalah observasi yang akan ditampilkan pada output.

Salah satu bagian terpenting dan paling sering penulis gunakan pada fungsi ini memfilter *missing value* (melihat observasi yang mengandung *missing value* atau tidak melibatkan *missing value*). Berikut adalah contoh filter terhadap data pada `pollution_tbl` yang tidak mengandung *missing value* dan nilai `amount>0`.

```
filter(pollution_tbl, !(is.na(amount) | amount<=0))
```

Berdasarkan hasil yang diperoleh seluruh data tidak ada yang di drop sehingga dapat disimpulkan bahwa data tersebut tidak mengandung *missing value* dan nol.

3.5.3 `arrange()`

Fungsi `arrange()` bekerja mirip dengan fungsi `filter()` kecuali bahwa alih-alih memilih baris, fungsi ini mengubah urutan observasinya (mengurutkan dari yang terbesar atau sebaliknya). Dibutuhkan data frame dan sekumpulan nama kolom (atau ekspresi yang lebih rumit) untuk dipesan. Jika kita memberikan lebih dari satu nama kolom pada fungsi, setiap kolom tambahan akan digunakan untuk menentukan urutan nilai yang sama berdasarkan nilai kolom sebelumnya.

Fungsi `arrange()` mirip dengan fungsi `order()` pada paket dasar R. Format sederhana fungsi ini adalah sebagai berikut:

```
arrange(data, ....)
```

Note:

- **data** : data frame
- : daftar nama variabel yang tidak dikutip yang dipisahkan tanda koma, atau ekspresi yang melibatkan nama variabel. Gunakan `desc()` untuk mengurutkan variabel dalam urutan menurun.

Misalkan kita ingin melihat urutan mobil pada data `mtcars` berdasarkan penggunaan bahan bakar (`mpg`) dan bobot mobil (`wt`) tersebut. Berikut adalah sintaks yang digunakan:

```
data("mtcars")

# Ubah mtcars menjadi tibble
mtcars<- as_tibble(mtcars)

arrange(mtcars, mpg, wt)

## # A tibble: 32 x 11
##       mpg     cyl   disp     hp   drat     wt   qsec     vs
##       <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 10.4      8    472    205   2.93   5.25   18.0     0
## 2 10.4      8    460    215     3   5.42   17.8     0
## 3 13.3      8    350    245   3.73   3.84   15.4     0
## 4 14.3      8    360    245   3.21   3.57   15.8     0
## 5 14.7      8    440    230   3.23   5.34   17.4     0
## 6 15         8    301    335   3.54   3.57   14.6     0
## 7 15.2      8    304    150   3.15   3.44   17.3     0
## 8 15.2      8    276.   180   3.07   3.78    18      0
## 9 15.5      8    318    150   2.76   3.52   16.9     0
## 10 15.8     8    351    264   4.22   3.17   14.5     0
## # ... with 22 more rows, and 3 more variables:
## #   am <dbl>, gear <dbl>, carb <dbl>
```

Jika ingin urutan yang digunakan adalah dari yang terbesar ke terkecil untuk kedua variabel tersebut jalankan sintaks berikut:

```
arrange(mtcars, desc(mpg), desc(wt))

## # A tibble: 32 x 11
##       mpg     cyl   disp     hp   drat     wt   qsec     vs
##       <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 33.9      4    71.1    65   4.22   1.84   19.9     1
## 2 32.4      4    78.7    66   4.08   2.2    19.5     1
## 3 30.4      4    75.7    52   4.93   1.62   18.5     1
## 4 30.4      4   95.1    113   3.77   1.51   16.9     1
## 5 27.3      4     79     66   4.08   1.94   18.9     1
## 6 26         4   120.     91   4.43   2.14   16.7     0
## 7 24.4      4   147.     62   3.69   3.19    20      1
## 8 22.8      4   141.     95   3.92   3.15   22.9     1
## 9 22.8      4   108     93   3.85   2.32   18.6     1
## 10 21.5     4   120.     97   3.7    2.46   20.0     1
## # ... with 22 more rows, and 3 more variables:
## #   am <dbl>, gear <dbl>, carb <dbl>
```

Jika menggunakan fungsi `order()`:

```
attach(mtcars)
# urutan dari kecil ke besar
mtcars[order(mpg, wt), ]

## # A tibble: 32 x 11
##   mpg   cyl  disp    hp  drat    wt  qsec    vs
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 10.4     8   472   205  2.93  5.25  18.0    0
## 2 10.4     8   460   215   3   5.42  17.8    0
## 3 13.3     8   350   245  3.73  3.84  15.4    0
## 4 14.3     8   360   245  3.21  3.57  15.8    0
## 5 14.7     8   440   230  3.23  5.34  17.4    0
## 6 15        8   301   335  3.54  3.57  14.6    0
## 7 15.2     8   304   150  3.15  3.44  17.3    0
## 8 15.2     8   276.   180  3.07  3.78  18      0
## 9 15.5     8   318   150  2.76  3.52  16.9    0
## 10 15.8    8   351   264  4.22  3.17  14.5    0
## # ... with 22 more rows, and 3 more variables:
## #   am <dbl>, gear <dbl>, carb <dbl>

# urutan dari besar ke kecil
mtcars[order(-mpg, -wt), ]

## # A tibble: 32 x 11
##   mpg   cyl  disp    hp  drat    wt  qsec    vs
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 33.9     4   71.1   65  4.22  1.84  19.9    1
## 2 32.4     4   78.7   66  4.08  2.2   19.5    1
## 3 30.4     4   75.7   52  4.93  1.62  18.5    1
## 4 30.4     4   95.1  113  3.77  1.51  16.9    1
## 5 27.3     4   79     66  4.08  1.94  18.9    1
## 6 26       4  120.    91  4.43  2.14  16.7    0
## 7 24.4     4  147.    62  3.69  3.19  20      1
## 8 22.8     4  141.    95  3.92  3.15  22.9    1
## 9 22.8     4  108    93  3.85  2.32  18.6    1
## 10 21.5    4  120.    97  3.7   2.46  20.0    1
## # ... with 22 more rows, and 3 more variables:
## #   am <dbl>, gear <dbl>, carb <dbl>
```

Note: *missing value* akan selalu diurutkan pada observasi terakhir baik menggunakan urutan dari terbesar ke terkecil maupun sebaliknya.

3.5.4 select()

Fungsi `select()` dapat digunakan untuk memilih kolom dari data frame yang ingin kita fokuskan. Seringkali kita memiliki data frame yang besar yang berisi semua data, tetapi setiap analisis yang diberikan hanya menggunakan subset variabel atau pengamatan. Fungsi `select()` memungkinkan kita untuk mendapatkan beberapa kolom yang mungkin kita butuhkan.

Fungsi `select()` memiliki kesamaan dengan subset menggunakan tanda “[” dan “\$”. Perbedaannya adalah kita dapat melakukan hal lebih melalui fungsi ini seperti memilih berdasarkan kriteria tertentu menggunakan fungsi bantuan sebagai berikut:

1. `starts_with("abcd")`, pilih kolom yang memiliki awalan “abcd”.
2. `end_with("abcd")`, pilih kolom yang memiliki akhiran “abcd”.
3. `contains("abcd")`, pilih kolom yang mengandung nama “abcd”
4. `matches("(.)\\1")`, pilih variabel yang mengandung *regular expression*. Fungsi ini memilih variabel yang mengandung perulangan karakter.
5. `num_range("x", 1:3)`, cocokkan berdasarkan kolom dengan nama x1,x2,x3.

Berdasarkan fungsi bantuan tersebut, fungsi `select()` lebih powerfull dibandingkan dengan cara subset biasa serta lebih mudah dalam melakukannya. Berikut adalah format dari fungsi `select()`:

```
select(data, ....)
```

Note:

- `data` : data frame
- : Satu atau lebih ekspresi kutip yang dipisahkan oleh koma. kita dapat memperlakukan nama variabel seperti posisi, sehingga kita dapat menggunakan ekspresi seperti `x: y` untuk memilih rentang variabel. Nilai positif pilih variabel; nilai negatif drop variabel. Jika ekspresi pertama negatif, `select()` akan secara otomatis dimulai dengan semua variabel. Gunakan argumen bernama, mis. `new_name = old_name`, untuk mengganti nama variabel yang dipilih.

Berikut adalah contoh penerapan `select()` pada data frame `flights`.

```
# memasang paket
# install.packages("nycflights13")

# memuat data frame
library(nycflights13)

## Warning: package 'nycflights13' was built under R
## version 3.5.3

# data
flights

## # A tibble: 336,776 x 19
##   year month   day dep_time sched_dep_time dep_delay
##   <int> <int> <int>    <int>          <int>     <dbl>
## 1  2013     1     1      517            515       2
## 2  2013     1     1      533            529       4
## 3  2013     1     1      542            540       2
## 4  2013     1     1      544            545      -1
## 5  2013     1     1      554            600      -6
## 6  2013     1     1      554            558      -4
## 7  2013     1     1      555            600      -5
## 8  2013     1     1      557            600      -3
## 9  2013     1     1      557            600      -3
## 10 2013     1     1      558            600      -2
## # ... with 336,766 more rows, and 13 more variables:
## #   arr_time <int>, sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>,
```

```
## #   tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

```
# pilih kolom berdasarkan nama kolom
select(flights, year, month, day)
```

```
## # A tibble: 336,776 x 3
##       year   month   day
##   <int> <int> <int>
## 1  2013      1      1
## 2  2013      1      1
## 3  2013      1      1
## 4  2013      1      1
## 5  2013      1      1
## 6  2013      1      1
## 7  2013      1      1
## 8  2013      1      1
## 9  2013      1      1
## 10 2013      1      1
## # ... with 336,766 more rows
```

```
# pilih seluruh kolom dari year sampai day
select(flights, year:day)
```

```
## # A tibble: 336,776 x 3
##       year   month   day
##   <int> <int> <int>
## 1  2013      1      1
## 2  2013      1      1
## 3  2013      1      1
## 4  2013      1      1
## 5  2013      1      1
## 6  2013      1      1
## 7  2013      1      1
## 8  2013      1      1
## 9  2013      1      1
## 10 2013      1      1
## # ... with 336,766 more rows
```

```
# drop kolom dari year sampai day
select(flights, -(year:day))
```

```
## # A tibble: 336,776 x 16
##       dep_time sched_dep_time dep_delay arr_time
##   <int>          <int>     <dbl>    <int>
## 1      517           515      2       830
## 2      533           529      4       850
## 3      542           540      2       923
## 4      544           545     -1      1004
## 5      554           600     -6       812
## 6      554           558     -4       740
```

```

## 7      555       600      -5     913
## 8      557       600      -3     709
## 9      557       600      -3     838
## 10     558       600      -2     753
## # ... with 336,766 more rows, and 12 more variables:
## #   sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>,
## #   distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dttm>

```

```

# pilih kolom dengan akhiran time
select(flights, ends_with("time"))

```

```

## # A tibble: 336,776 x 5
##   dep_time sched_dep_time arr_time sched_arr_time
##       <int>           <int>     <int>           <int>
## 1      517            515     830            819
## 2      533            529     850            830
## 3      542            540     923            850
## 4      544            545    1004            1022
## 5      554            600     812            837
## 6      554            558     740            728
## 7      555            600     913            854
## 8      557            600     709            723
## 9      557            600     838            846
## 10     558            600     753            745
## # ... with 336,766 more rows, and 1 more variable:
## #   air_time <dbl>

```

```

# pilih kolom yang mengandung karakter "arr"
select(flights, contains("arr"))

```

```

## # A tibble: 336,776 x 4
##   arr_time sched_arr_time arr_delay carrier
##       <int>           <int>     <dbl> <chr>
## 1      830            819      11  UA
## 2      850            830      20  UA
## 3      923            850      33  AA
## 4     1004            1022     -18 B6
## 5      812            837     -25 DL
## 6      740            728      12  UA
## 7      913            854      19  B6
## 8      709            723     -14 EV
## 9      838            846      -8  B6
## 10     753            745       8  AA
## # ... with 336,766 more rows

```

Kita juga dapat menggunakan fungsi tambahan `everithing()` yang berguna jika kita ingin memindahkan variabel yang menjadi fokus kita ke awal data frame tanpa melakukan drop variabel. Berikut adalah contoh sintaksnya:

```
# pindahkan kolom yang mengandung time di awal
select(flights, contains("time"), everything())

## # A tibble: 336,776 x 19
##   dep_time sched_dep_time arr_time sched_arr_time
##       <int>           <int>     <int>           <int>
## 1      517            515     830            819
## 2      533            529     850            830
## 3      542            540     923            850
## 4      544            545    1004            1022
## 5      554            600     812            837
## 6      554            558     740            728
## 7      555            600     913            854
## 8      557            600     709            723
## 9      557            600     838            846
## 10     558            600     753            745
## # ... with 336,766 more rows, and 15 more variables:
## #   air_time <dbl>, time_hour <dttm>, year <int>,
## #   month <int>, day <int>, dep_delay <dbl>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>,
## #   distance <dbl>, hour <dbl>, minute <dbl>
```

3.5.5 mutate()

Fungsi `mutate()` ada untuk menghitung transformasi variabel dalam data frame. Seringkali, kita ingin membuat variabel baru yang berasal dari variabel yang ada dan fungsi `mutate()` menyediakan antarmuka yang bersih untuk melakukan itu. Format yang digunakan adalah sebagai berikut:

```
mutate(data, ....)
```

Note:

- **data** : data frame
- : Pasangan nama-nilai ekspresi, masing-masing dengan panjang 1 atau panjang yang sama dengan jumlah baris dalam grup (jika menggunakan `group_by()`) atau di seluruh input (jika tidak menggunakan grup). Nama setiap argumen akan menjadi nama variabel baru, dan nilainya akan menjadi nilai yang sesuai. Gunakan nilai NULL dalam mutasi untuk menjatuhkan drop variabel lama, sehingga variabel baru menimpa variabel yang ada dengan nama yang sama.

```
# subset data frame
flights_sml <- select(flights,
  year:day,
  ends_with("delay"),
  distance,
  air_time
)

# mutate()
mutate(flights_sml,
  gain = arr_delay - dep_delay,
```

```

  hours = air_time / 60,
  gain_per_hour = gain / hours
)

## # A tibble: 336,776 x 10
##   year month   day dep_delay arr_delay distance
##   <int> <int> <int>     <dbl>     <dbl>     <dbl>
## 1  2013     1     1       2       11     1400
## 2  2013     1     1       4       20     1416
## 3  2013     1     1       2       33     1089
## 4  2013     1     1      -1      -18     1576
## 5  2013     1     1      -6      -25      762
## 6  2013     1     1      -4       12      719
## 7  2013     1     1      -5       19     1065
## 8  2013     1     1      -3      -14      229
## 9  2013     1     1      -3       -8      944
## 10 2013     1     1      -2        8      733
## # ... with 336,766 more rows, and 4 more variables:
## #   air_time <dbl>, gain <dbl>, hours <dbl>,
## #   gain_per_hour <dbl>

```

Jika hanya ingin menyisakan variabel output fungsi `mutate()` pada data frame (variabel lain di drop), kita dapat menggunakan fungsi `transmute()`. Berikut adalah contoh sintaks yang digunakan:

```

transmute(flights,
  gain = arr_delay - dep_delay,
  hours = air_time / 60,
  gain_per_hour = gain / hours
)

```

```

## # A tibble: 336,776 x 3
##   gain hours gain_per_hour
##   <dbl> <dbl>      <dbl>
## 1  9  3.78       2.38
## 2  16 3.78       4.23
## 3  31 2.67      11.6
## 4  -17 3.05      -5.57
## 5  -19 1.93      -9.83
## 6  16  2.5        6.4
## 7  24  2.63       9.11
## 8  -11 0.883     -12.5
## 9  -5  2.33      -2.14
## 10 10  2.3        4.35
## # ... with 336,766 more rows

```

Adapaun fungsi-fungsi dan operator yang dapat digunakan pada `mutate()` untuk membuat variabel baru adalah sebagai berikut:

1. **Operator aritmatik** (+,-,*,/, $\hat{}$, $\%/\%$, $\%%$). operator aritmetik seperti $\%/\%$ dan $\%%$ sangat berguna dalam memecah integer menjadi beberapa bagian seperti hasil bagi tanpa sisa ($\%/\%$) dan sisa hasil bagi ($\%%$). Berikut adalah contoh penerapannya:

```
transmute(flights,
  dep_time,
  hour = dep_time %/%
    100,
  minute = dep_time %% 100
)

## # A tibble: 336,776 x 3
##   dep_time   hour minute
##       <int>   <dbl>   <dbl>
## 1      517     5     17
## 2      533     5     33
## 3      542     5     42
## 4      544     5     44
## 5      554     5     54
## 6      554     5     54
## 7      555     5     55
## 8      557     5     57
## 9      557     5     57
## 10     558     5     58
## # ... with 336,766 more rows
```

2. **Fungsi aritmetik** (`log()`,`sin()`,`cos()`,dll)
3. **Fungsi Offsets** (`lead()` dan `lag()`). memungkinkan kita untuk merujuk pada nilai-nilai memimpin atau tertinggal. Berikut adalah contoh penerapannya:

```
(x <- 1:10)

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

lag(x)

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

lead(x)

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

4. **Fungsi kumulatif** (`cumsum()`,`cumprod()`,`cummin()`,`cummax()`, dan `cummean()`). Jika kita membutuhkan agregat bergulir (mis., Jumlah yang dihitung di atas jendela bergulir). Berikut adalah contoh penerapannya:

```
x

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

cumsum(x)

## [1] 1 3 6 10 15 21 28 36 45 55
```

```
cummean(x)

## [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5
```

5. **Operator logik** (`<`, `<=`, `>`, `>=`, `!=`). Jika kita melakukan urutan operasi logis yang kompleks, seringkali ide yang baik untuk menyimpan nilai sementara dalam variabel baru sehingga kita dapat memeriksa bahwa setiap langkah berfungsi seperti yang diharapkan.
6. Rangking (`min_rank()`, `row_number()`, `dense_rank()`, `percent_rank()`, `cume_dist()` dan `ntile()`).

3.5.6 summarize() dan group_by()

Kita dapat membuat ringkasan data menggunakan fungsi `summarize()`. Fungsi tersebut akan merubah data frame menjadi sebuah baris berisi ringkasan data yang kita inginkan. Berikut adalah contoh penerapannya:

```
summarize(flights, delay = mean(dep_delay, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   delay
##   <dbl>
## 1 12.6
```

FUNGSI ini akan lebih berguna saat digunakan dengan fungsi `group_by()` sehingga dapat diperoleh ringkasan data pada setiap grup. berikut adalah contoh penerapannya:

```
by_day <- group_by(flights, year, month, day)
summarize(by_day, delay = mean(dep_delay, na.rm = TRUE))
```

```
## # A tibble: 365 x 4
## # Groups:   year, month [12]
##   year month   day delay
##   <int> <int> <int> <dbl>
## 1 2013     1     1  11.5
## 2 2013     1     2  13.9
## 3 2013     1     3  11.0
## 4 2013     1     4  8.95
## 5 2013     1     5  5.73
## 6 2013     1     6  7.15
## 7 2013     1     7  5.42
## 8 2013     1     8  2.55
## 9 2013     1     9  2.28
## 10 2013    1    10  2.84
## # ... with 355 more rows
```

3.5.7 Mengkombinasikan Beberapa Operasi Menggunakan Operator Pipe (%>%)

Operator pipa (`%>%`) sangat berguna untuk merangkai bersama beberapa fungsi `dplyr` dalam suatu urutan operasi. Perhatikan contoh sebelumnya dimana setiap kali kita ingin menerapkan lebih dari satu fungsi, urutannya akan dimulai dalam urutan panggilan fungsi bersarang yang sulit dibaca. Secara ringkas dapat kita tulis sebagai berikut:

```
third(second(first(x)))
```

Jika dituliskan menggunakan operator pipa akan menghasilkan sintak berikut:

```
x %>%
  first() %>%
  second() %>%
  third()
```

Dengan menuliskannya melalui cara tersebut kita dapat membacanya lebih mudah.

Misal kita ingin mengetahui hubungan antara variabel jarak (`dist`) terhadap rata-rata delay (`arr_delay`). Langkah-langkah untuk melakukannya dengan menggunakan operator pipa adalah sebagai berikut:

1. Kelompokkan penerbangan berdasarkan destinasiya (`group_by()`).
2. Hitung ringkasan data berdasarkan jarak, rata-rata delay, dan jumlah penerbangan.
3. Lakukan filter untuk membuang *noisy point* (jika diperlukan). Dalam hal ini jumlah penerbangan > 20 dan tujuan penerbangan Honolulu (“HNL”) adalah *outlier* atau *noisy point*.

Berikut adalah sintaks untuk melakukannya:

```
# Tanpa pipe operator
by_dest <- group_by(flights, dest)
delay <- summarize(by_dest,
  count = n(),
  dist = mean(distance, na.rm = TRUE),
  delay = mean(arr_delay, na.rm = TRUE))
delay <- filter(delay, count > 20, dest != "HNL")

# Dengan pipe operator
library(magrittr)

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:tidyর':
##
##     extract

delays <- flights %>%
  group_by(dest) %>%
  summarize(
    count = n(),
    dist = mean(distance, na.rm = TRUE),
    delay = mean(arr_delay, na.rm = TRUE)
  ) %>%
  filter(count > 20, dest != "HNL")

# Print
delays
```

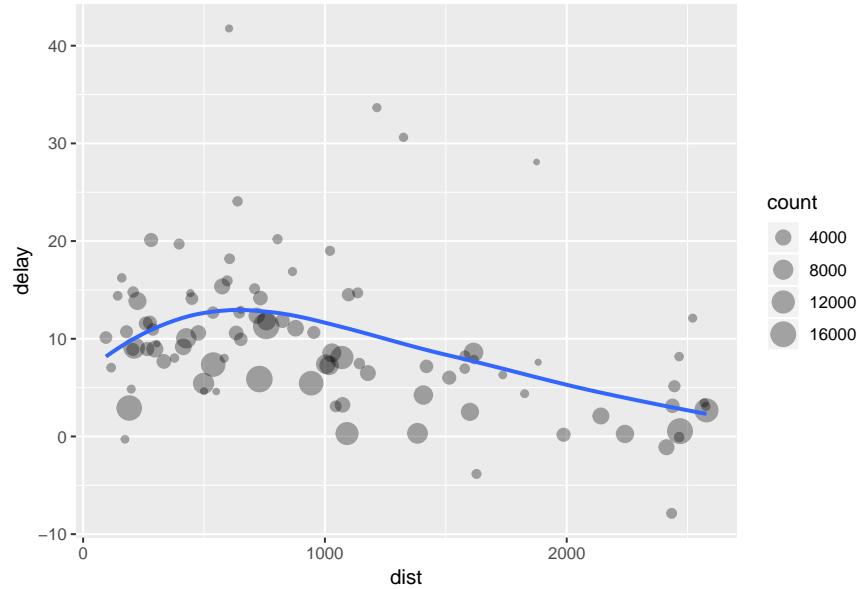


Figure 3.3: Jarak vs rata-rata delay

```

## # A tibble: 96 x 4
##   dest  count dist delay
##   <chr> <int> <dbl> <dbl>
## 1 ABQ     254 1826  4.38
## 2 ACK     265  199  4.85
## 3 ALB     439  143 14.4
## 4 ATL    17215  757. 11.3
## 5 AUS    2439 1514.  6.02
## 6 AVL     275  584.  8.00
## 7 BDL     443  116  7.05
## 8 BGR     375  378  8.03
## 9 BHM     297  866. 16.9
## 10 BNA    6333  758. 11.8
## # ... with 86 more rows

## Warning: package 'ggplot2' was built under R version
## 3.5.3

## 
## Attaching package: 'ggplot2'

## The following object is masked from 'mtcars':
## 
##   mpg

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

```

Berdasarkan Gambar 3.3, rata-rata delay meningkat seiring dengan pertambahan jarak penerbangan.

3.6 Referensi

1. Wickham, H. Grolemund G. 2016. **R For Data Science: Import, Tidy, Transform, Visualize, And Model Data.** O'Reilly Media, Inc.
2. Peng, R.D. 2015. **Exploratory Data Analysis with R.** Leanpub book.
3. Dplyr Documentation. <https://dplyr.tidyverse.org/>
4. Quick-R. **Data Input.** <https://www.statmethods.net/input/index.html>
5. Quick-R. **Data Management.** <https://www.statmethods.net/management/index.html>
6. STHDA. **Importing Data Into R .** <http://www.sthda.com/english/wiki/importing-data-into-r>
7. STHDA. **Exporting Data From R.** <http://www.sthda.com/english/wiki/exporting-data-from-r>

Visualisasi Data - R

Chapter 4

Visualisasi Data Menggunakan Fungsi Dasar R

Visualisasi data merupakan bagian yang sangat penting untuk mengkomunikasikan hasil analisa yang telah kita lakukan. Selain itu, komunikasi juga membantu kita untuk memperoleh gambaran terkait data selama proses analisa data sehingga membantu kita dalam memutuskan metode analisa apa yang dapat kita terapkan pada data tersebut.

R memiliki library visualisasi yang sangat beragam, baik yang merupakan fungsi dasar pada R maupun dari sumber lain seperti ggplot dan lattice. Seluruh library visualisasi tersebut memiliki kelebihan dan kekurangannya masing-masing.

Pada *chapter* ini kita tidak akan membahas seluruh library tersebut. Kita akab berfokus pada fungsi visualisasi dasar bawaan dari R. kita akan mempelajari mengenai jenis visualisasi data sampai dengan melakukan kustomisasi pada parameter grafik yang kita buat.

4.1 Visualisasi Data Menggunakan Fungsi `plot()`

Fungsi `plot()` merupakan fungsi umum yang digunakan untuk membuat plot pada R. Format dasarnya adalah sebagai berikut:

```
plot(x, y, type="p")
```

Note:

- **x dan y:** titik koordinat plot Berupa variabel dengan panjang atau jumlah observasi yang sama.
- **type:** jenis grafik yang hendak dibuat. Nilai yang dapat dimasukkan antara lain:
 - `type="p"` : membuat plot titik atau scatterplot. Nilai ini merupakan default pada fungsi `plot()`.
 - `type="l"` : membuat plot garis.
 - `type="b"` : membuat plot titik yang terhubung dengan garis.
 - `type="o"` : membuat plot titik yang ditimpa oleh garis.
 - `type="h"` : membuat plot garis vertikal dari titik ke garis $y=0$.
 - `type="s"` : membuat fungsi tangga.
 - `type="n"` : tidak membuat grafik plot sama sekali, kecuali plot dari axis. Dapat digunakan untuk mengatur tampilan suatu plot utama yang diikuti oleh sekelompok plot tambahan.

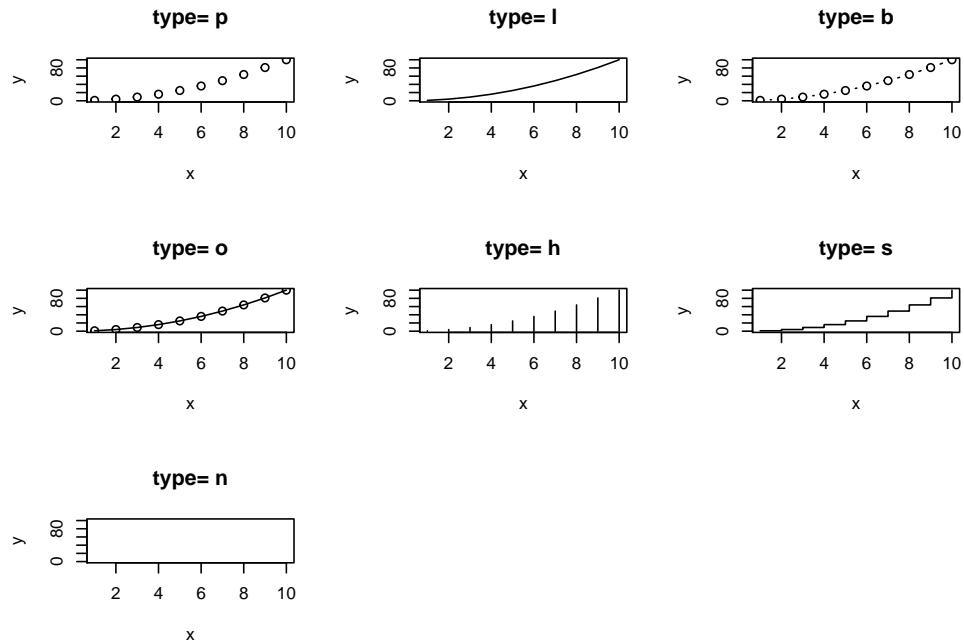


Figure 4.1: Plot berbagai jenis setting type

Untuk lebih memahaminya berikut penulis akan sajikan contoh untuk masing-masing grafik tersebut. Berikut adalah contoh sintaks dan hasil plot yang disajikan pada Gambar 4.1:

```
# membuat vektor data
x <- c(1:10); y <- x^2

# membagi jendela grafik menjadi 4 baris dan 2 kolom
par(mfrow=c(3,3))

# loop
type <- c("p", "l", "b", "o", "h", "s", "n")
for (i in type){
  plot(x,y, type= i,
    main= paste("type=", i))
}
```

Pada contoh selanjutnya akan dilakukan plot terhadap dataset `trees`. Untuk memuatnya jalankan sintaks berikut:

```
library(tibble)

# memuat dataset
trees <- as_tibble(trees)

# print
trees
```

A tibble: 31 x 3

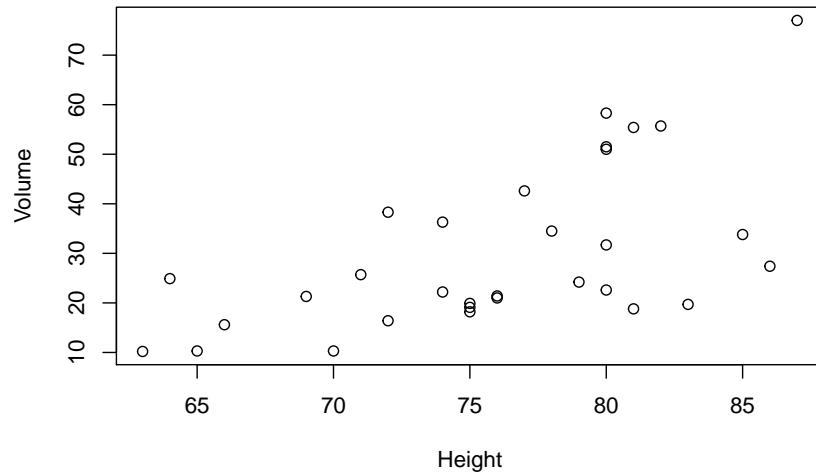


Figure 4.2: Scatterplot Height vs Volume

```
##      Girth Height Volume
##      <dbl>  <dbl> <dbl>
## 1    8.3     70   10.3
## 2    8.6     65   10.3
## 3    8.8     63   10.2
## 4   10.5     72   16.4
## 5   10.7     81   18.8
## 6   10.8     83   19.7
## 7    11      66   15.6
## 8    11      75   18.2
## 9   11.1     80   22.6
## 10   11.2     75   19.9
## # ... with 21 more rows
```

Pada dataset tersebut kita ingin membuat scatterplot untuk melihat korelasi antara variabel `Height` dan `Volume`. Untuk melakukannya jalankan sintaks berikut:

```
plot(trees$Height, trees$Volume)
```

```
# atau
with(trees, plot(Height, Volume))
```

Kita juga dapat menggunakan formula untuk membuat scatterplot pada Gambar 4.2. Berikut adalah contoh sintaks yang digunakan:

```
x <- trees$Height
y <- trees$Volume

plot(y~x)
```

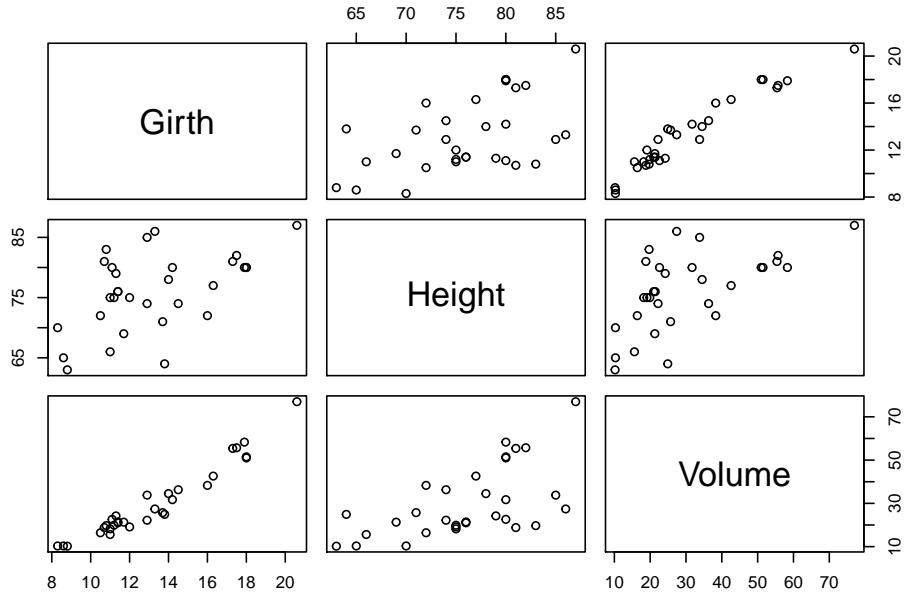


Figure 4.3: Matriks scatterplot dataset trees

Fungsi `plot()` juga dapat digunakan untuk membentuk matriks scatterplot. Untuk membuatnya kita hanya perlu memasukkan seluruh dataset kedalam fungsi `plot()`. Berikut adalah sintaks dan output yang dihasilkan berupa Gambar 4.3:

```
plot(trees)
```

Selain itu jika kita memasukkan objek `lm()` yang merupakan fungsi untuk melakukan operasi regresi linier pada fungsi `plot()`, output yang dihasilkan berupa plot diagnostik yang berguna untuk menguji asumsi model regresi linier. Berikut adalah contoh sintaks dan output yang dihasilkan pada Gambar 4.4:

```
# membagi jendela grafik menjadi 2 baris dan 2 kolom
par(mfrow=c(2,2))

# plot
plot(lm(Volume~Height, data=trees))
```

Selain objek-objek tersebut, fungsi `plot()` akan banyak digunakan dalam analisis statistika kita pada chapter lainnya.

4.2 Matriks Scatterplot

Pada bagian sebelumnya kita telah belajar bagaimana membuat matriks scatterplot menggunakan fungsi `plot()`. Pada bagian ini kita akan belajar cara membuat matriks scatterplot menggunakan fungsi `pairs()`. Secara umum format fungsi dituliskan sebagai berikut:

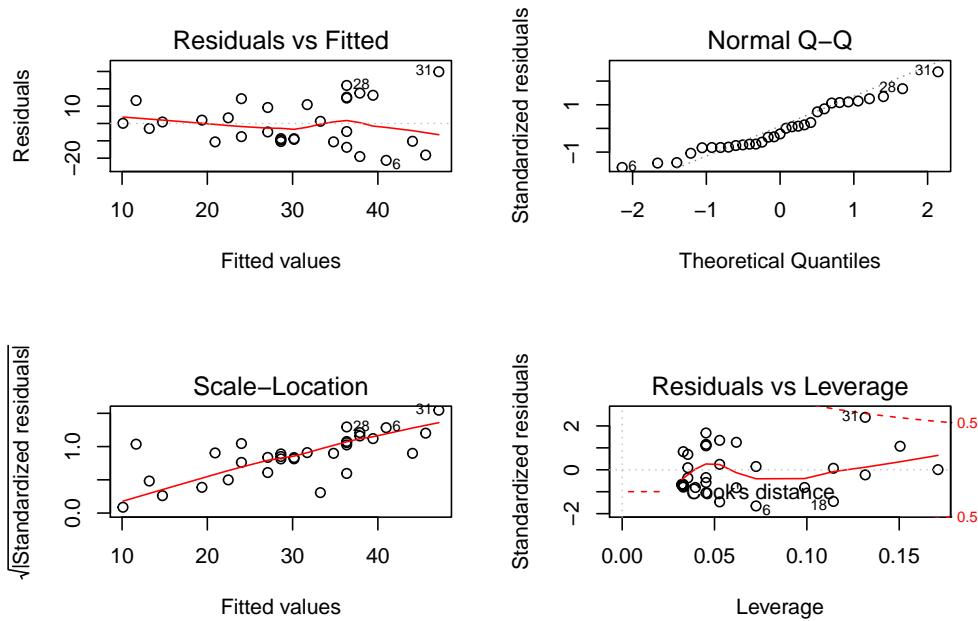


Figure 4.4: Plot diagnostik regresi linier

```
pairs(data, lower.panel=NULL)
```

Note:

- **data:** data frame
- **lower.panel:** menampilkan atau tidak menampilkan panel bawah

Untuk lebih memahami penggunaan fungsi tersebut, berikut akan disajikan contoh penggunaannya pada dataset `iris`. Sebelum melakukannya jalankan sintaks berikut untuk memuat dataset:

```
# memuat dataset irir
iris <- as_tibble(iris)

# print
iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
	<dbl>	<dbl>	<dbl>	<dbl>
## 1	5.1	3.5	1.4	0.2
## 2	4.9	3	1.4	0.2
## 3	4.7	3.2	1.3	0.2
## 4	4.6	3.1	1.5	0.2
## 5	5	3.6	1.4	0.2
## 6	5.4	3.9	1.7	0.4
## 7	4.6	3.4	1.4	0.3
## 8	5	3.4	1.5	0.2

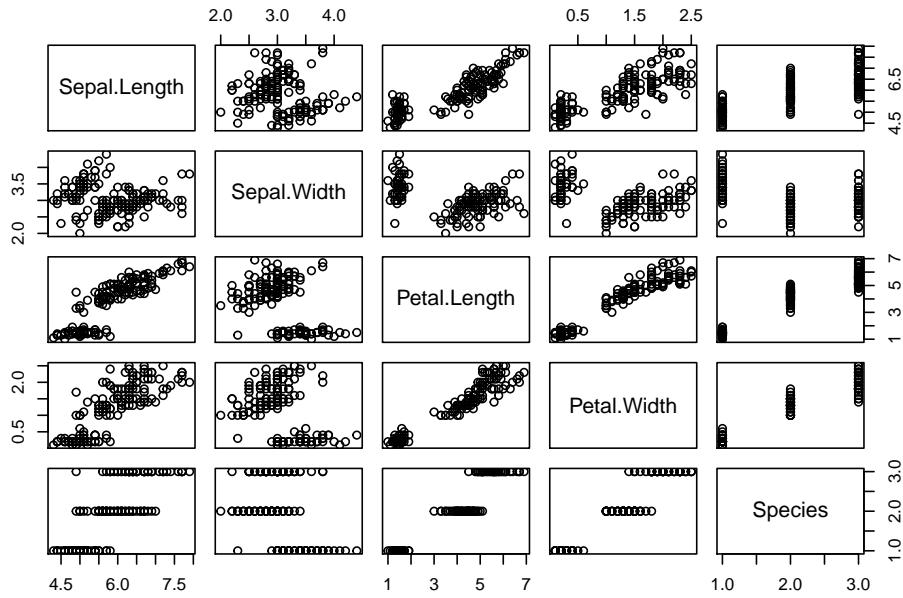


Figure 4.5: Matriks scatterplot iris

```
##   9      4.4      2.9      1.4      0.2
## 10     4.9      3.1      1.5      0.1
## # ... with 140 more rows, and 1 more variable:
## #   Species <fct>
```

Untuk membuat matriks scatterplot kita hanya perlu memasukkan objek `iris` kedalam fungsi `pairs()`. Berikut adalah sintaks yang digunakan dan output yang dihasilkan pada Gambar 4.5:

```
pairs(iris)
```

Kita dapat melakukan drop terhadap panel bawah grafik tersebut. Untuk melakukannya kita perlu memasukkan parameter `lower.panel=NULL`. Output yang dihasilkan akan tampak seperti pada Gambar 4.6.

```
pairs(iris, lower.panel=NULL)
```

Kita dapat merubah warna titik berdasarkan factor `Species`. Langkah pertama yang perlu dilakukan adalah melakukan drop variabel `Species` pada dataset dan memasukkan objek baru tanpa variabel tersebut kedalam fungsi `pairs()`. Warna berdasarkan grup diberikan dengan menambahkan parameter `col=` pada fungsi `pairs()`. Berikut adalah contoh penerapannya dan output yang dihasilkan pada Gambar 4.7:

```
# drop variabel Species
# simpan dataset baru pada objek iris2
iris2 <- iris[,1:4]

# print
iris2
```

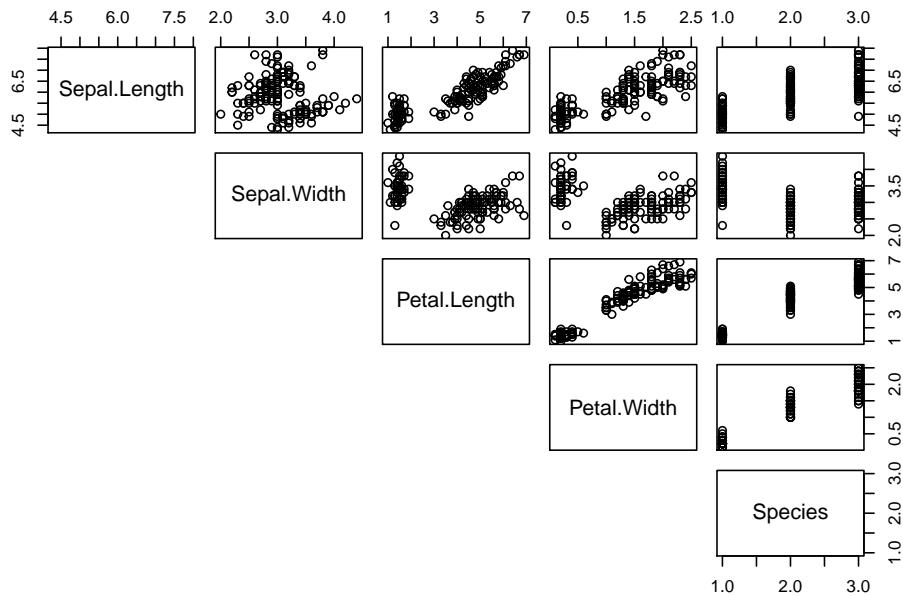


Figure 4.6: Matriks scatterplot iris tanpa panel bawah

```
## # A tibble: 150 x 4
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
##       <dbl>      <dbl>      <dbl>      <dbl>
## 1         5.1        3.5       1.4       0.2
## 2         4.9        3.0       1.4       0.2
## 3         4.7        3.2       1.3       0.2
## 4         4.6        3.1       1.5       0.2
## 5         5.0        3.6       1.4       0.2
## 6         5.4        3.9       1.7       0.4
## 7         4.6        3.4       1.4       0.3
## 8         5.0        3.4       1.5       0.2
## 9         4.4        2.9       1.4       0.2
## 10        4.9        3.1       1.5       0.1
## # ... with 140 more rows
```

```
# spesifikasi vektor warna titik berdasarkan spesies
my_col <- c("#00AFBB", "#E7B800", "#FC4E07")

# plot
pairs(iris2, lower.panel=NULL,
      # spesifikasi warna
      col= my_col[iris$Species])
```

Kita juga dapat mengganti panel bawah menjadi nilai korelasi antar variabel. Untuk melakukannya kita perlu mendefinisikan sebuah fungsi untuk panel bawah dan panel atas (jika ingin warna titik berdasarkan faktor). Setelah fungsi panel bawah dan atas didefinisikan, langkah selanjutnya adalah melakukan memasukkan nilainya kedalam fungsi `pairs()`. Berikut adalah sintaks yang digunakan serta output yang dihasilkan pada Gambar 4.8:

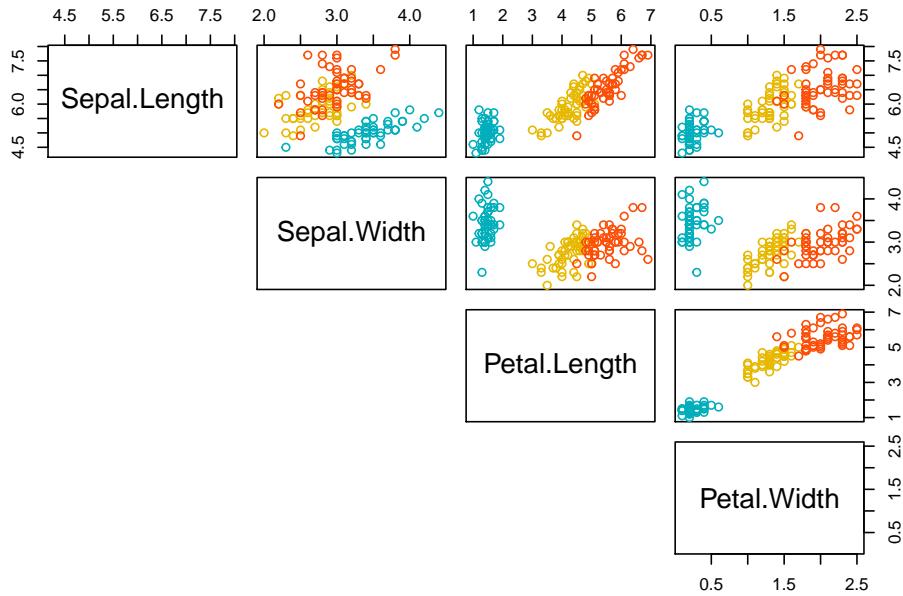


Figure 4.7: Matriks scatterplot iris tanpa panel bawah

```
# membuat fungsi untuk menghitung
# nilai korelasi yang ditempatkan pada panel bawah
panel.cor <- function(x, y){
  # definisi parameter grafik
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  # menghitung koefisien korelas
  r <- round(cor(x, y), digits=2)
  # menambahkan text berdasarkan koefisien korelasi
  txt <- paste0("R = ", r)
  # mengatur besar text sesuai besarnya nilai korelasi
  cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * abs(r))
}

# kustomisasi panel atas agar
# warna titik berdasarkan factor
my_col <- c("#00AFBB", "#E7B800", "#FC4E07")
upper.panel<-function(x, y){
  points(x,y, col = my_col[iris$Species])
}

pairs(iris2,
      lower.panel= panel.cor,
      upper.panel= upper.panel)
```

Jika kita tidak ingin nilai korelasi ditampilkan di panel bawah, kita dapat merubahnya sehingga dapat tampil pada panel atas bersamaan dengan scatterplot. Untuk melakukannya kita perlu mendefinisikan fungsi pada

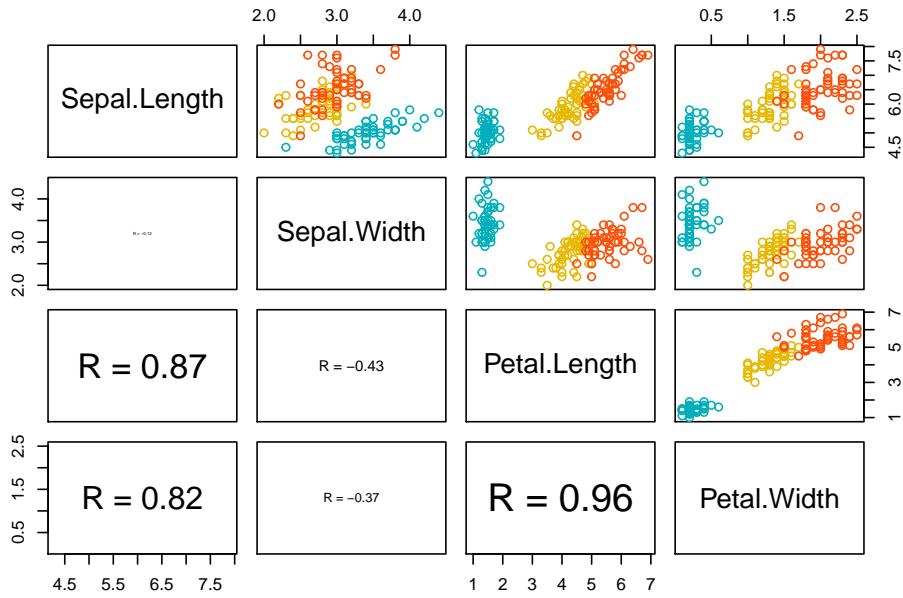


Figure 4.8: Matriks scatterplot iris dengan koefisien korelasi

panel atas dan memasukkannya pada parameter `upper.panel=`. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 4.9:

```
# kustomisasi panel atas
upper.panel<-function(x, y){
  points(x,y, col=c("#00AFBB", "#E7B800", "#FC4E07")[iris$Species])
  r <- round(cor(x, y), digits=2)
  txt <- paste0("R = ", r)
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  text(0.5, 0.9, txt)
}

# plot
pairs(iris2, lower.panel = NULL,
      upper.panel = upper.panel)
```

4.3 Box plot

Box plot pada R dapat dibuat menggunakan fungsi `boxplot()`. Berikut adalah sintaks untuk membuat boxplot variabel `Sepal.Length` pada dataset `iris` dan output yang dihasilkan pada Gambar 4.10:

```
boxplot(iris$Sepal.Length)
```

Boxplot juga dapat dibuat berdasarkan variabel factor. Hal ini berguna untuk melihat perbedaan ditribusi data pada masing-masing grup. Pada sintaks berikut dibuat boxplot berdasarkan variabel `Species`. Output yang dihasilkan disajikan pada Gambar 4.11:

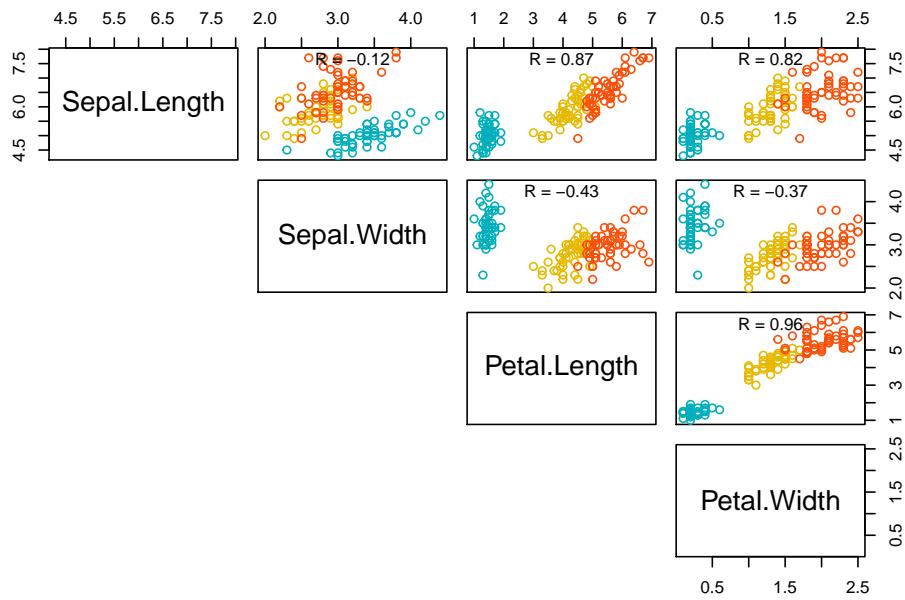


Figure 4.9: Matriks scatterplot iris dengan koefisien korelasi di panel atas

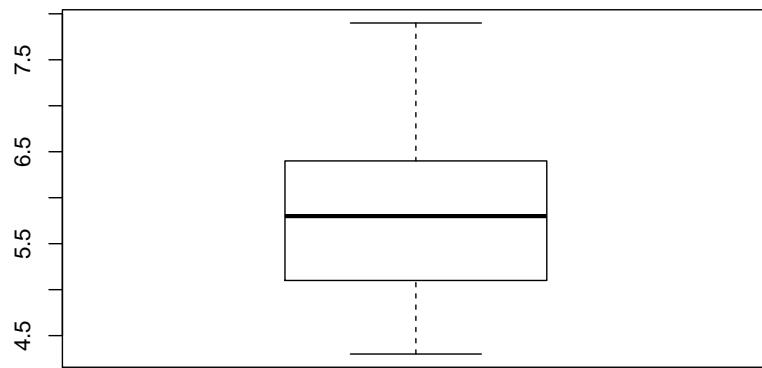


Figure 4.10: Boxplot variabel Sepal.Length

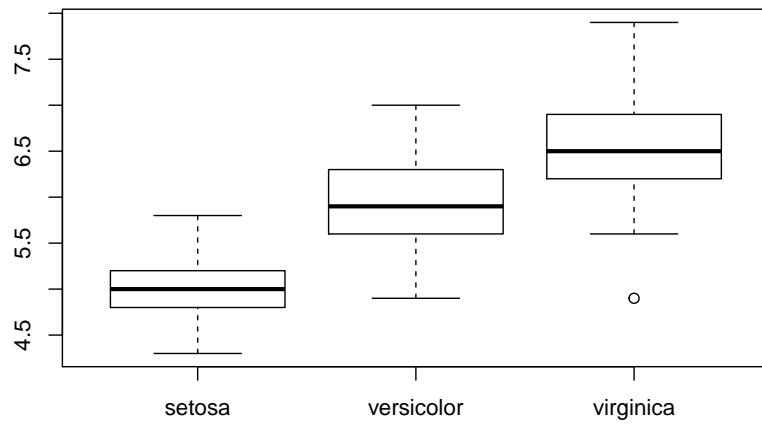


Figure 4.11: Boxplot berdasarkan variabel species

```
boxplot(iris$Sepal.Length~iris$Species)
```

Kita juga dapat mengubah warna outline dan box pada boxplot. Berikut adalah contoh sintaks yang digunakan untuk melakukannya dan output yang dihasilkan disajikan pada Gambar 4.12:

```
boxplot(iris$Sepal.Length~iris$Species,
       # ubah warna outline menjadi steelblue
       border = "steelblue",
       # ubah warna box berdasarkan grup
       col= c("#999999", "#E69F00", "#56B4E9"))
```

Kita juga dapat membuat boxplot pada *multiple group*. Data yang digunakan untuk contoh tersebut adalah dataset ToothGrowth. Berikut adalah sintaks untuk memuat dataset tersebut:

```
# memuat dataset sebagai tibble
ToothGrowth <- as_tibble(ToothGrowth)

# print
ToothGrowth
```

```
## # A tibble: 60 x 3
##       len supp dose
##   <dbl> <fct> <dbl>
## 1     4.2 VC    0.5
## 2    11.5 VC    0.5
## 3     7.3 VC    0.5
## 4     5.8 VC    0.5
## 5     6.4 VC    0.5
## 6    10.0 VC    0.5
## 7    11.2 VC    0.5
```

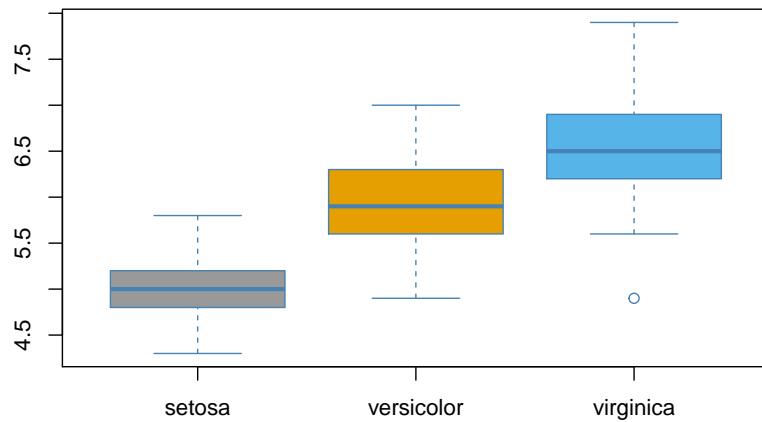


Figure 4.12: Boxplot dengan warna berdasarkan spesies

```

## 8 11.2 VC 0.5
## 9 5.2 VC 0.5
## 10 7 VC 0.5
## # ... with 50 more rows

# ubah variable dose menjadi factor
ToothGrowth$dose <- as.factor(ToothGrowth$dose)

# print
ToothGrowth

## # A tibble: 60 x 3
##       len supp dose
##   <dbl> <fct> <dbl>
## 1 4.2   VC    0.5
## 2 11.5  VC    0.5
## 3 7.3   VC    0.5
## 4 5.8   VC    0.5
## 5 6.4   VC    0.5
## 6 10.0  VC    0.5
## 7 11.2  VC    0.5
## 8 11.2  VC    0.5
## 9 5.2   VC    0.5
## 10 7.0   VC    0.5
## # ... with 50 more rows

```

Contoh sintaks dan output boxplot *multiple group* disajikan pada Gambar 4.13:

```
boxplot(len ~ supp*dose, data = ToothGrowth,
       col = c("white", "steelblue"))
```

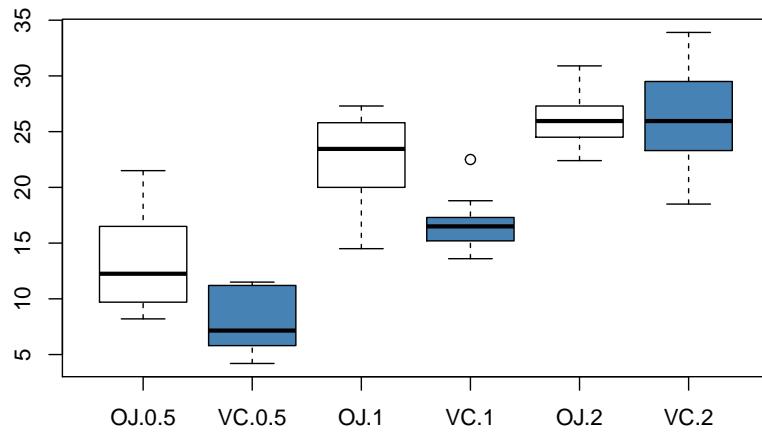


Figure 4.13: Boxplot multiple group

4.4 Bar Plot

Barplot pada R dapat dibuat menggunakan fungsi `barplot()`. Untuk lebih memahaminya berikut disajikan contoh barplot menggunakan dataset `VADeaths`. Untuk memuatnya jalankan sintaks berikut:

```
VADeaths
```

```
##      Rural Male Rural Female Urban Male Urban Female
## 50-54      11.7       8.7    15.4      8.4
## 55-59      18.1      11.7    24.3     13.6
## 60-64      26.9      20.3    37.0     19.3
## 65-69      41.0      30.9    54.6     35.1
## 70-74      66.0      54.3    71.1     50.0
```

Contoh bar plot untuk variabel `Rural Male` disajikan pada Gambar 4.14:

```
par(mfrow=c(1,2))
barplot(VADeaths[, "Rural Male"], main="a")
barplot(VADeaths[, "Rural Male"], main="b", horiz=TRUE)
```

```
par(mfrow=c(1,1))
```

Kita dapat mengubah warna pada masing-masing bar, baik outline bar maupun box pada bar. Selain itu kita juga dapat mengubah nama grup yang telah dihasilkan sebelumnya. Berikut sintaks untuk melakukannya dan output yang dihasilkan pada Gambar 4.15:

```
barplot(VADeaths[, "Rural Male"],
# ubah warna outline menjadi steelblue
border="steelblue",
```

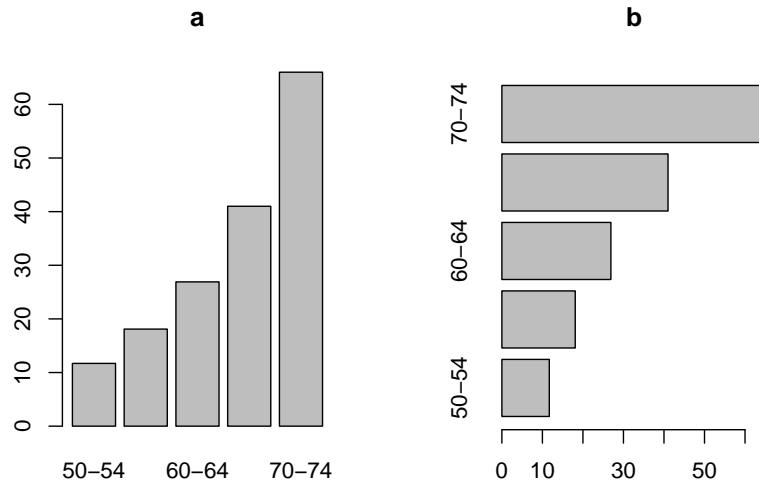


Figure 4.14: a. bar plot vertikal; b. bar plot horizontal

```
# ubah wana box
col= c("grey", "yellow", "steelblue", "green", "orange"),
# ubah nama grup dari A sampai E
names.arg = LETTERS[1:5],
# ubah orientasi menjadikan horizontal
horiz=TRUE)
```

Untuk bar plot dengan *multiple group*, tersedia dua pengaturan posisi yaitu *stacked bar plot*(menunjukkan proporsi penyusun pada masing-masing grup) dan *grouped bar plot*(melihat perbedaan individual pada masing-masing grup). Pada Gambar 4.16 dan Gambar 4.17 , disajikan kedua jenis bar plot tersebut.

```
# staked
barplot(VADeaths,
        col = c("lightblue", "mistyrose", "lightcyan",
               "lavender", "cornsilk"),
        legend = rownames(VADeaths))
```

```
# grouped
barplot(VADeaths,
        col = c("lightblue", "mistyrose", "lightcyan",
               "lavender", "cornsilk"),
        legend = rownames(VADeaths), beside = TRUE)
```

4.5 Line Plot

Line plot pada R dapat dibentuk menggunakan fungsi `plot()`. Selain itu fungsi `lines()` dapat pula digunakan untuk menambahkan line plot pada grafik. Berikut adalah sintaks untuk membuat line plot dan outputnya pada Gambar 4.18:

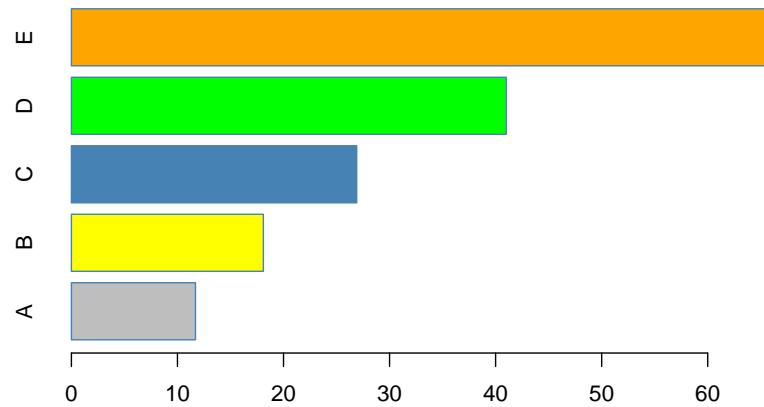


Figure 4.15: Kustomisasi bar plot

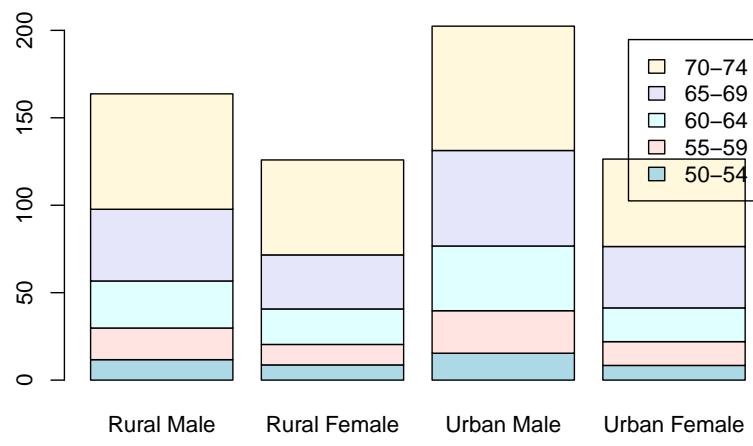


Figure 4.16: Stacked bar plot

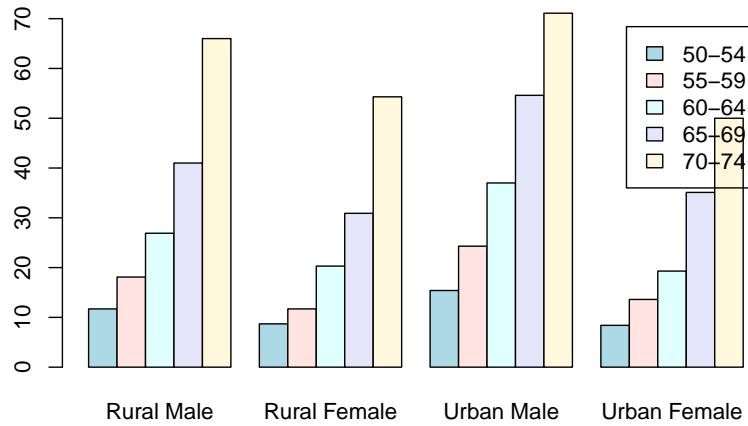


Figure 4.17: Grouped bar plot

```
# Membuat vektor data
x <- c(1:20)
y <- 2*x
z <- x^2

# Membuat line plot x vs y
plot(y~x, type="b",
      lty=1,
      col="blue")

# Menambahkan line plot x vs z
lines(z~x, type="o",
      lty=2,
      col="red")

# Menambahkan legend
legend("topleft", legend=c("Line 1", "Line 2"),
      col=c("red", "blue"), lty = 1:2, cex=0.8)
```

4.6 Pie Chart

Pie chart digunakan untuk membuat visualisasi proporsi pada sebuah data. Pie chart pada R dibuat menggunakan fungsi `pie()`. Berikut adalah sintaks untuk membuat pie chart dan output yang dihasilkan pada Gambar 4.19:

```
par(mar = c(0, 1, 0, 1))
pie(
  c(280, 60, 20),
```

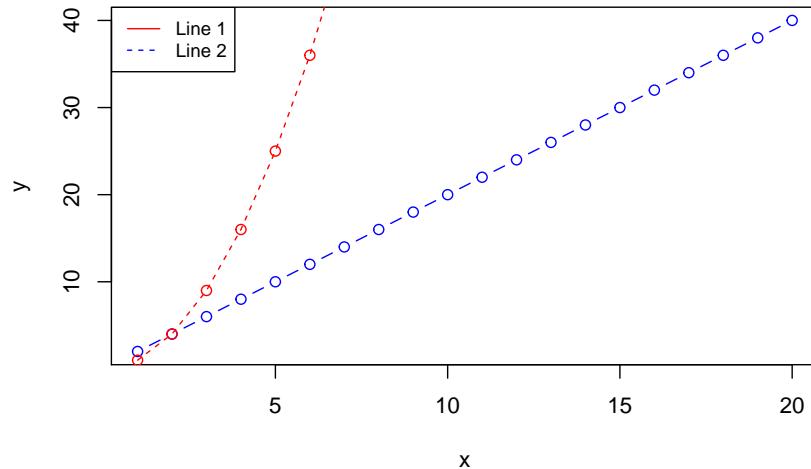


Figure 4.18: Line plot

```
c('Sky', 'Sunny side of pyramid', 'Shady side of pyramid'),
  col = c('#0292D8', '#F7EA39', '#C4B632'),
  init.angle = -50, border = NA
)
```

4.7 Histogram dan Density Plot

Fungsi `hist()` dapat digunakan untuk membuat histogram pada R. Secara sederhana fungsi tersebut didefinisikan sebagai berikut:

```
hist(x, breaks="Sturges")
```

Note:

- **x:** vektor numerik
- **breaks:** *breakpoints* antar sel histogram.

Pada dataset `trees` akan dibuat histogram variabel `Height`. Untuk melakukannya jalankan sintaks berikut:

```
hist(trees$Height)
```

Output yang dihasilkan disajikan pada Gambar 4.20:

Density plot pada R dapat dibuat menggunakan fungsi `density()`. Berbeda dengan fungsi `hist()`, fungsi ini tidak langsung menghasilkan grafik densitas. Fungsi `density()` hanya menghitung kernel densitas pada data. Densitas yang telah dihitung selanjutnya diplotkan menggunakan fungsi `plot()`. Berikut adalah sintaks dan output yang dihasilkan pada Gambar 4.21:

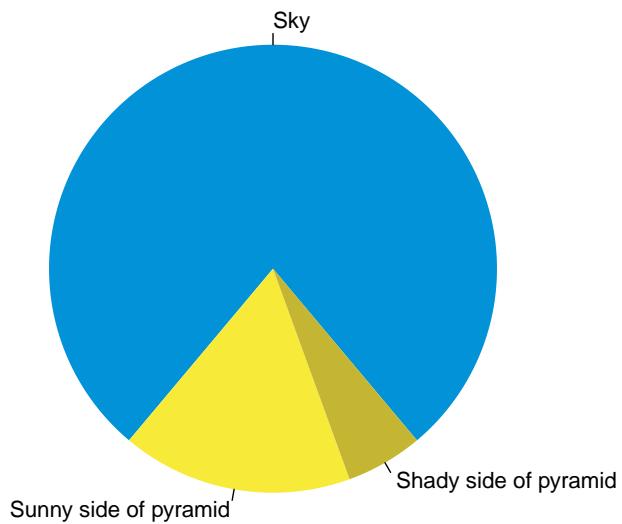


Figure 4.19: Pie chart

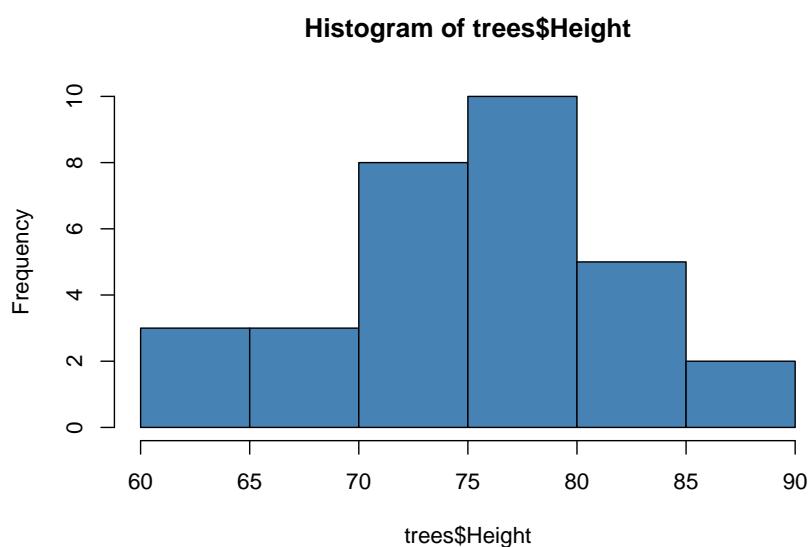


Figure 4.20: Histogram

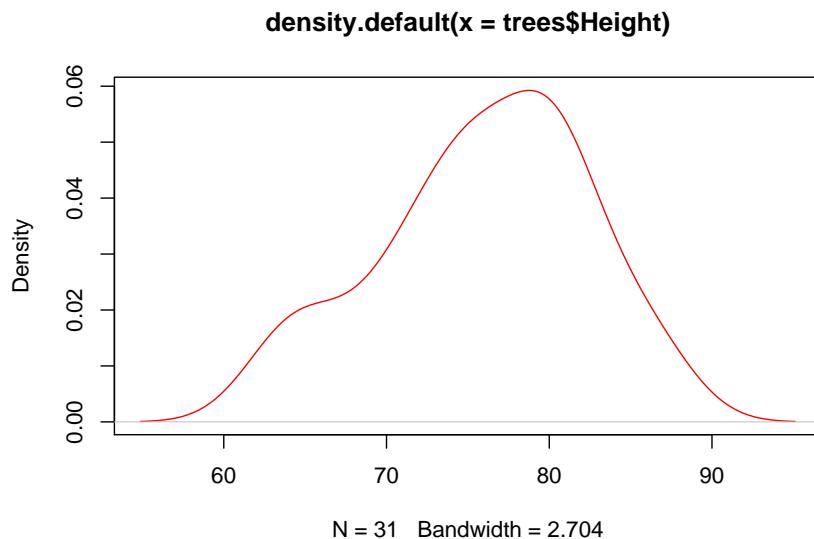


Figure 4.21: Density plot

```
# menghitung kernel density
dens <- density(trees$Height)

# plot densitas dengan outline merah
plot(dens,col="red")
```

Kita juga dapat menambahkan grafik densitas pada histogram sehingga mempermudah pembacaan pada histogram. Untuk melakukannya kita perlu mengubah kernel histogram dari frekuensi menjadi density dengan menambahkan argumen `freq=FALSE` pada fungsi `hist()`. Selanjutnya tambahkan fungsi `polygon()` untuk memplotkan grafik densitas. Berikut adalah sintak dan output yang dihasilkan pada Gambar 4.22:

```
# menghitung kernel density
dens <- density(trees$Height)

# histogram
hist(trees$Height, freq=FALSE, col="steelblue")

# tambahkan density plot
polygon(dens, border="red")
```

4.8 QQ Plot

QQ plot digunakan untuk mengecek distribusi suatu data apakah berdistribusi normal atau tidak. Pada R QQ plot dibuat menggunakan 2 fungsi yaitu: `qnorm()` dan `qqline()`. Fungsi `qnorm()` digunakan untuk memproduksi normal QQ plot suatu variabel. Sedangkan fungsi `qqline()` digunakan untuk membuat garis referensi distiribusi normal. Suatu distribusi dikatakan normal jika titik observasi yang dihasilkan mengikuti garis referensi tersebut.

Berikut adalah cara membuat QQ plot menggunakan variabel `Volume` pada dataset `trees`. Output yang dihasilkan disajikan pada Gambar 4.23.

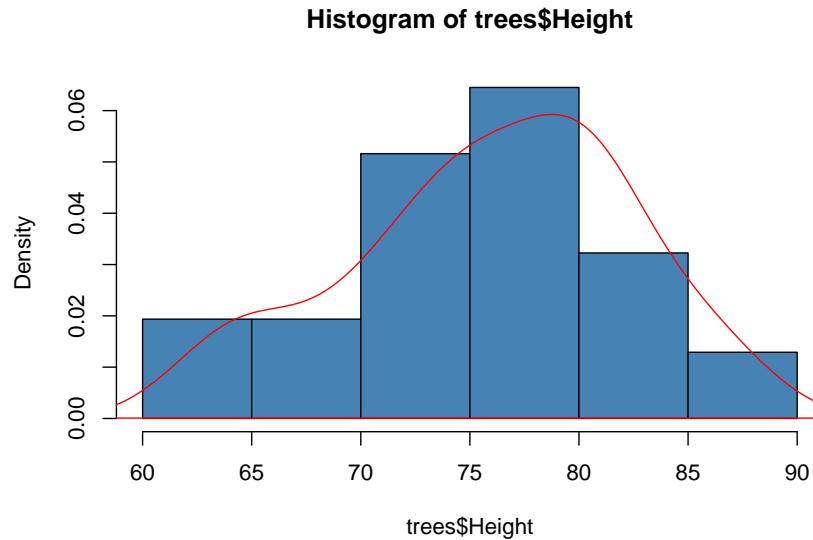


Figure 4.22: Density plot dan histogram

```
qqnorm(trees$Volume)
qqline(trees$Volume, col="red")
```

4.9 Dot Chart

Fungsi `dotchart()` pada R digunakan untuk membuat dot chart. Format yang digunakan adalah sebagai berikut:

```
dotchart(x, labels = NULL, groups = NULL,
         gcolor = par("fg"), color = par("fg"))
```

Note:

- **x:** vektor atau matriks numerik.
- **labels:** vektor label untuk tiap titik.
- **groups:** grouping variabel yang mengindikasikan bagaimana **x** dikelompokkan.
- **gcolor:** warna yang digunakan pada label grup dan nilai observasi.
- **color:** warna yang digunakan untuk titik dan label.

Pada contoh berikut disajikan cara membuat dot chart pada dataset `mtcars` untuk melihat mobil yang paling hemat bahan bakar berdasarkan variabel `mpg` dan jumlah silinder (`cyl`). Berikut sintaks yang digunakan dan output yang dihasilkan pada Gambar 4.24:

```
# mengurutkan dataset mtcars berdasarkan variabel mpg
mtcars <- mtcars[order(mtcars$mpg), ]

# mengubah variabel cyl menjadi faktor
grps <- as.factor(mtcars$cyl)
```

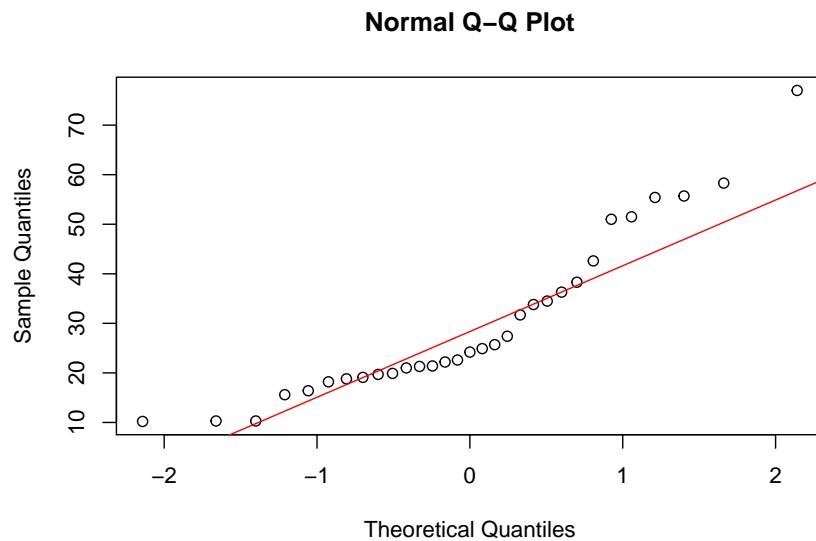


Figure 4.23: QQ plot

```
# membuat vektor warna berdasarkan jumlah grup
my_cols <- c("#999999", "#E69F00", "#56B4E9")

# plot
dotchart(mtcars$mpg, labels = row.names(mtcars),
          groups = grps, gcolor = my_cols,
          color = my_cols[grps],
          cex = 0.6, pch = 19, xlab = "mpg")
```

4.10 Kustomisasi Parameter Grafik

Pada bagian ini penulis akan menjelaskan cara untuk kustomisasi parameter grafik seperti:

- menambahkan judul, legend, teks, axis, dan garis.
- mengubah skala axis, simbol plot, jenis garis, dan warna.

4.10.1 Menambahkan Judul

Pada grafik di R, kita dapat menambahkan judul dengan dua cara, yaitu: pada plot melalui parameter dan melalui fungsi plot(). Kedua cara tersebut tidak berbeda satu sama lain pada parameter input.

Untuk menambahkan judul pada plot secara langsung, kita dapat menggunakan argumen tambahan sebagai berikut:

- main:** teks untuk judul.
- xlab:** teks untuk keterangan axis X.
- ylab:** teks untuk keterangan axis y.

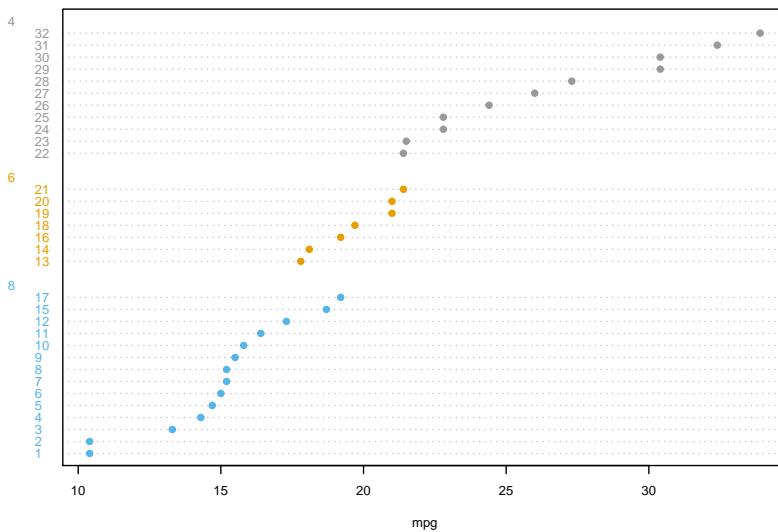


Figure 4.24: Dot chart

- d. **sub**: teks untuk sub-judul.

Berikut contoh sintaks penerapan masing-masing argumen tersebut beserta dengan output yang dihasilkan pada Gambar 4.25:

```
# menambahkan judul
barplot(c(2,5), main="Main title",
        xlab="X axis title",
        ylab="Y axis title",
        sub="Sub-title")
```

kita juga dapat melakukan kustomisasi pada warna, *font style*, dan ukuran font judul. Untuk melakukan kustomisasi pada warna pada judul, kita dapat menambahkan argumen sebagai berikut:

- col.main**: warna untuk judul.
- col.lab**: warna untuk keterangan axis.
- col.sub**: warna untuk sub-judul

Untuk kustomisasi font judul, kita dapat menambahkan argumen berikut:

- font.main**: *font style* untuk judul.
- font.lab**: *font style* untuk keterangan axis.
- font.sub**: *font style* untuk sub-judul.

Note:

Nilai yang dapat dimasukkan antara lain:

- 1: untuk teks normal.
- 2: untuk teks cetak tebal.
- 3: untuk teks cetak miring.

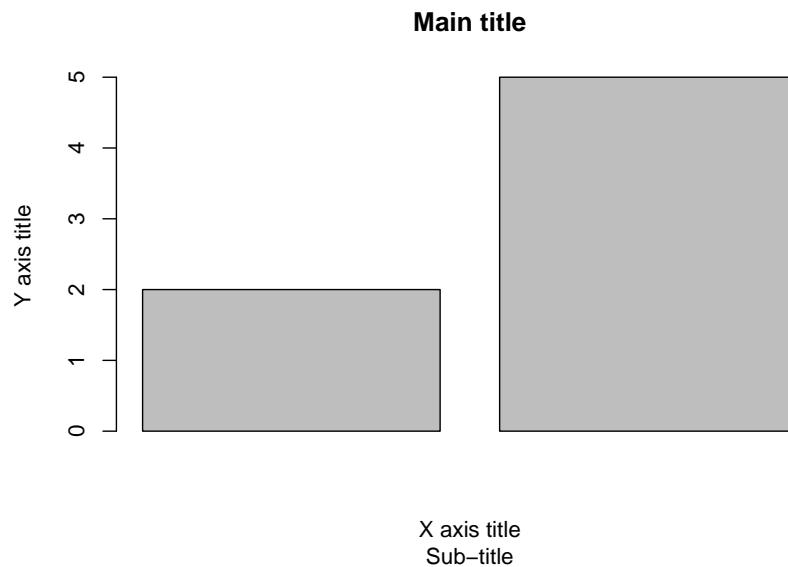


Figure 4.25: Menambahkan Judul

- 4: untuk teks cetak tebal dan miring.
- 5: untuk font simbol.

Sedangkan untuk ukuran font, kita dapat menambahkan variabel berikut:

- a. **cex.main**: ukuran teks judul.
- b. **cex.lab**: ukuran teks keterangan axis.
- c. **cex.sub**: ukuran teks sub-judul.

Berikut sintaks penerapan seluruh argumen tersebut beserta output yang dihasilkan pada Gambar 4.26:

```
# menambahkan judul
barplot(c(2,5),
        # menambahkan judul
        main="Main title",
        xlab="X axis title",
        ylab="Y axis title",
        sub="Sub-title",
        # kustomisasi warna font
        col.main="red",
        col.lab="blue",
        col.sub="black",
        # kustomisasi font style
        font.main=4,
        font.lab=4,
        font.sub=4,
        # kustomisasi ukuran font
        cex.main=2,
        cex.lab=1.7,
        cex.sub=1.2)
```

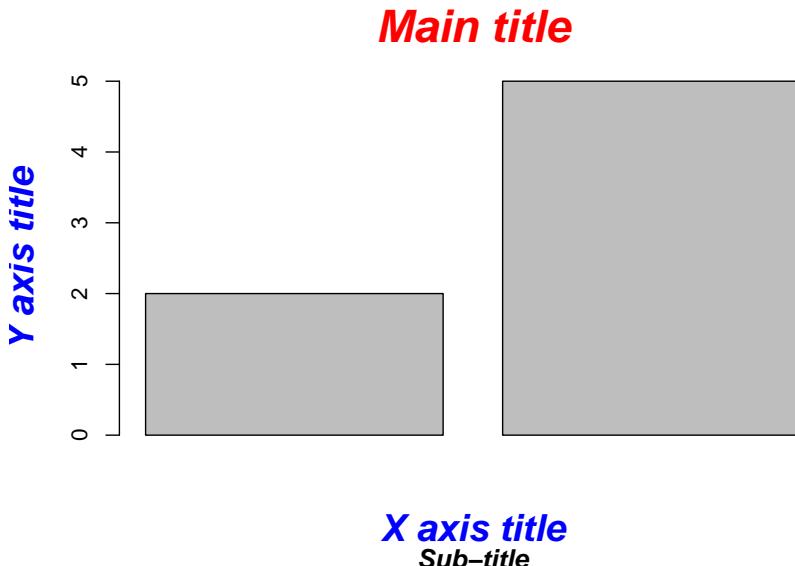


Figure 4.26: Menambahkan Judul (2)

Kita telah belajar bagaimana menambahkan judul langsung pada fungsi plot. Selain cara tersebut, telah penulis jelaskan bahwa kita dapat menambahkan judul melalui fungsi `title()`. argumen yang dimasukkan pada dasarnya tidak berbeda dengan ketika kita menambahkan judul secara langsung pada plot. Berikut adalah contoh sintaks dan output yang dihasilkan pada Gambar 4.27:

```
# menambahkan judul
barplot(c(2,5,8))

# menambahkan judul
title(main="Main title",
      xlab="X axis title",
      ylab="Y axis title",
      sub="Sub-title",
      # kustomisasi warna font
      col.main="red",
      col.lab="blue",
      col.sub="black",
      # kustomisasi font style
      font.main=4,
      font.lab=4,
      font.sub=4,
      # kustomisasi ukuran font
      cex.main=2,
      cex.lab=1.7,
      cex.sub=1.2)
```

4.10.2 Menambahkan Legend

Fungsi `legend()` pada R dapat digunakan untuk menambahkan legend pada grafik. Format sederhananya adalah sebagai berikut:

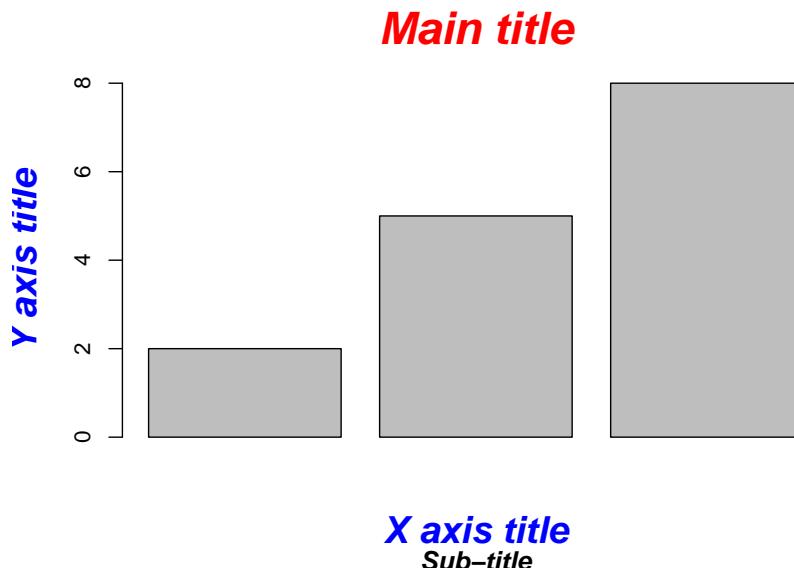


Figure 4.27: Menambahkan Judul (3)

```
legend(x, y=NULL, legend, fill, col, bg)
```

Note:

- **x** dan **y**: koordinat yang digunakan untuk posisi legend.
- **legend**: teks pada legend
- **fill**: warna yang digunakan untuk mengisi box disamping teks legend.
- **col**: warna garis dan titik disamping teks legend.
- **bg**: warna latar belakang legend box.

Berikut adalah contoh sintaks dan ouput penerapan argumen disajikan pada Gambar 4.28:

```
# membuat vektor numerik
x <- c(1:10)
y <- x^2
z <- x*2

# membuat line plot
plot(x,y, type="o", col="red", lty=1)

# menambahkan line plot
lines(x,z, type="o", col="blue", lty=2)

# menambahkan legend
legend(1, 95, legend=c("Line 1", "Line 2"),
       col=c("red", "blue"), lty=1:2, cex=0.8)
```

Kita dapat menambahkan judul, merubah font, dan merubah warna backgroud pada legend. Argumen yang ditambahkan pada legend adalah sebagai berikut:

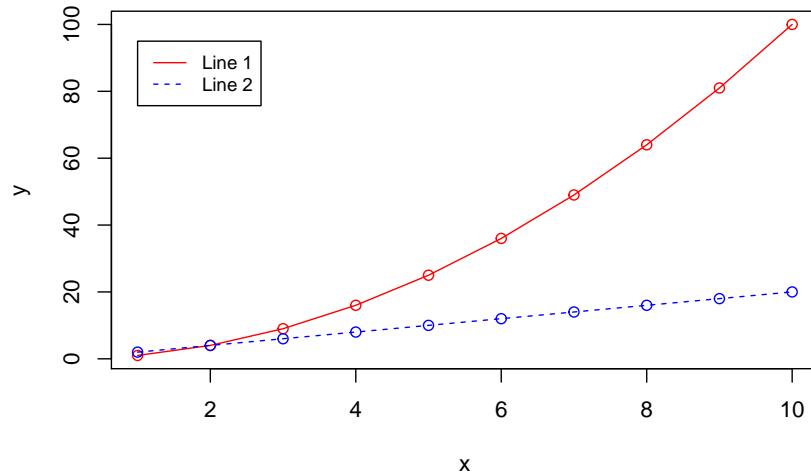


Figure 4.28: Menambahkan legend

- a. **title**: Judul legend
- b. **text.font**: integer yang menunjukkan *font style* pada teks legend. Nilai yang dapat dimasukkan adalah sebagai berikut:
 - 1: normal
 - 2: cetak tebal
 - 3: cetak miring
 - 4: cetak tebal dan miring.
- c. **bg**: warna background legend box.

Berikut adalah penerapan sintaks dan output yang dihasilkan pada Gambar 4.29:

```
# membuat line plot
plot(x,y, type="o", col="red", lty=1)

# menambahkan line plot
lines(x,z, type="o", col="blue", lty=2)

# menambahkan legend
legend(1, 95, legend=c("Line 1", "Line 2"),
       col=c("red", "blue"), lty=1:2, cex=0.8,
       title="Line types", text.font=4, bg='lightblue')
```

Kita dapat melakukan kustomisasi pada border dari legend melalui argumen `box.lty`=(jenis garis), `box.lwd`=(ukuran garis), dan `box.col`=(warna box). Berikut adalah penerapan argumen tersebut beserta output yang dihasilkan pada Gambar 4.30:

```
# membuat line plot
plot(x,y, type="o", col="red", lty=1)

# menambahkan line plot
```

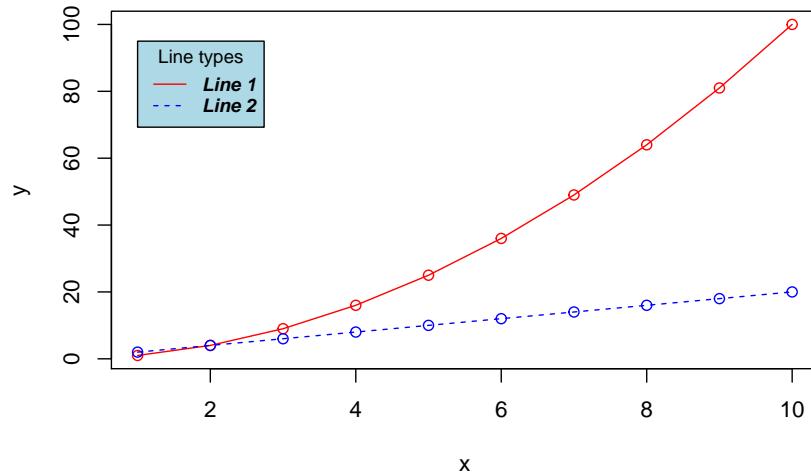


Figure 4.29: Menambahkan legend (2)

```
lines(x,z, type="o", col="blue", lty=2)

# menambahkan legend
legend(1, 95, legend=c("Line 1", "Line 2"),
       col=c("red", "blue"), lty=1:2, cex=0.8,
       title="Line types", text.font=4, bg='white',
       box.lty=2, box.lwd=2, box.col="steelblue")
```

Selain menggunakan koordinat, kita juga dapat melakukan kustomisasi posisi legend menggunakan *keyword* seperti: bottomright“, “bottom“, “bottomleft“, “left“, “topleft“, “top“, “topright“, “right” and “center”. Sejumlah kustomisasi legend berdasarkan *keyword* disajikan pada Gambar 4.31:

```
# plot
plot(x,y, type = "n")

# posisi kiri atas, inset =0.05
legend("topleft",
       legend = "(x,y)",
       title = "topleft, inset = .05",
       inset = 0.05)
# posisi atas
legend("top",
       legend = "(x,y)",
       title = "top")
# posisi kanan atas inset = .02
legend("topright",
       legend = "(x,y)",
       title = "topright, inset = .02",
       inset = 0.02)
# posisi kiri
legend("left",
```

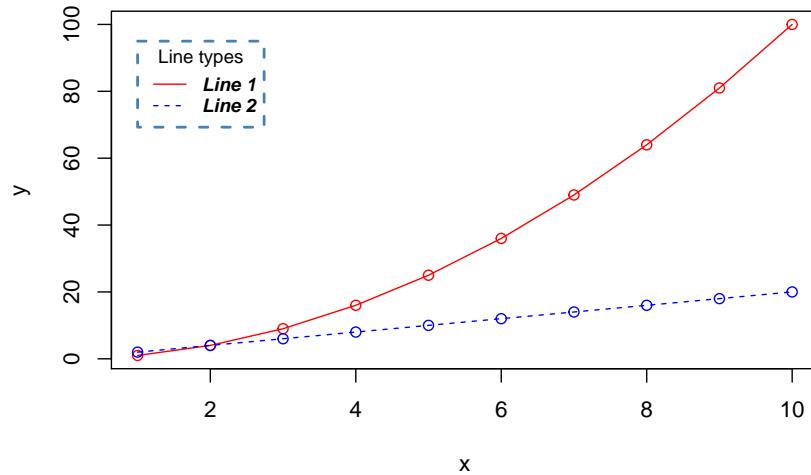


Figure 4.30: Menambahkan legend (3)

```

legend = "(x,y)",
title = "left")
# posisi tengah
legend("center",
      legend = "(x,y)",
      title = "center")
# posisi kanan
legend("right",
      legend = "(x,y)",
      title = "right")
# posisi kiri bawah
legend("bottomleft",
      legend = "(x,y)",
      title = "bottomleft")
# posisi bawah
legend("bottom",
      legend = "(x,y)",
      title = "bottom")
# posisi kanan bawah
legend("bottomright",
      legend = "(x,y)",
      title = "bottomright")

```

4.10.3 Menambahkan Teks Pada Grafik

Teks pada grafik dapat kita tambahkan baik sebagai keterangan yang menunjukkan label suatu observasi, keterangan tambahan disekitar bingkai grafik, maupun sebuah persamaan yang ada pada bidang grafik. Untuk menambahkannya kita dapat menggunakan dua buah fungsi yaitu: `text()` dan `mtext()`.

Fungsi `text()` berguna untuk menambahkan teks di dalam bidang grafik seperti label titik observasi dan

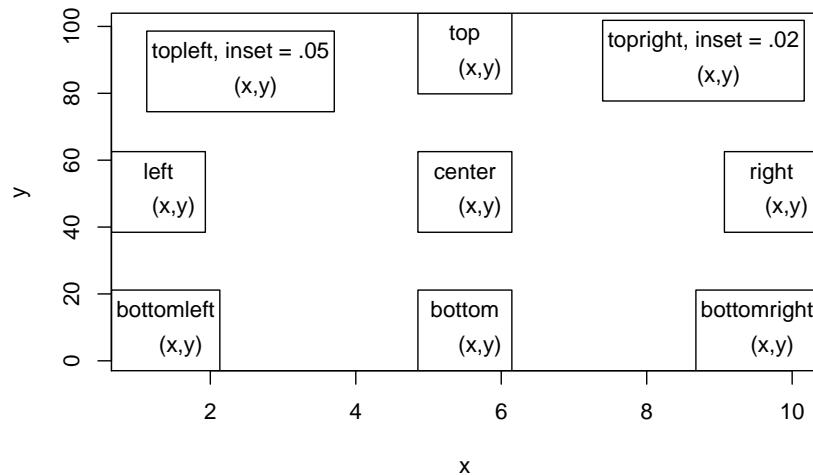


Figure 4.31: Kustomisasi posisi legend

persamaan di dalam bidang grafik. Format yang digunakan adalah sebagai berikut:

```
text(x, y, labels)
```

Note:

- **x** dan **y**: vektor numerik yang menunjukkan koordinat posisi teks.
- **labels**: vektor karakter yang menunjukkan teks yang hendak dituliskan.

Berikut adalah contoh sintaks untuk memberi label pada sejumlah data yang memiliki kriteria yang kita inginkan dan output yang dihasilkan pada Gambar 4.32:

```
# tandai observasi yang memiliki nilai
# mpg < 15 dan wt > 5
d <- mtcars[mtcars$wt >= 5 & mtcars$mpg <= 15, ]

# plot
plot(mtcars$wt, mtcars$mpg, main="Milage vs. Car Weight",
      xlab="Weight", ylab="Miles/(US gallon")

# menambahkan text
text(d$wt, d$mpg, row.names(d),
      cex=0.65, pos=3,col="red")
```

Sedangkan sintaks berikut adalah contoh bagaimana menambahkan persamaan kedalam bidang grafik dan output yang dihasilkan pada Gambar 4.33:

```
plot(1:10, 1:10,
     main="text(...)\nexamples\n~~~~~")
```

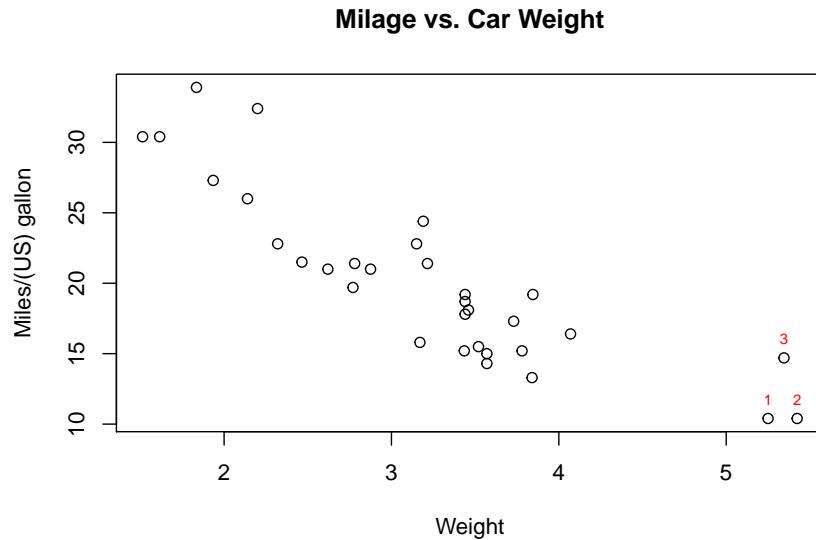


Figure 4.32: Menambahkan teks

```
text(4, 9, expression(hat(beta) == (X^t * X)^{-1} * X^t * y))
text(7, 4, expression(bar(x) == sum(frac(x[i], n), i==1, n)))
```

Fungsi `mtext()` berguna untuk menambahkan teks pada frame sekitar bidang grafik. Format yang digunakan adalah sebagai berikut:

```
mtext(text, side=3)
```

Note:

- `text`: teks yang akan ditulis.
- `side`: integer yang menunjukkan lokasi teks yang akan ditulis. Nilai yang dapat dimasukkan antara lain:
 - **1**: bawah
 - **2**: kiri
 - **3**: atas
 - **4**: kanan.

Berikut adalah contoh penerapan dan output yang dihasilkan pada Gambar 4.34:

```
plot(1:10, 1:10,
     main="mtext(... examples\n~~~~~")
mtext("Magic function", side=3)
```

4.10.4 Menambahkan Garis Pada Plot

Fungsi `abline()` dapat digunakan untuk menambahkan garis pada plot. Garis yang ditambahkan dapat berupa garis vertikal, horizontal, maupun garis regresi. Format yang digunakan adalah sebagai berikut:

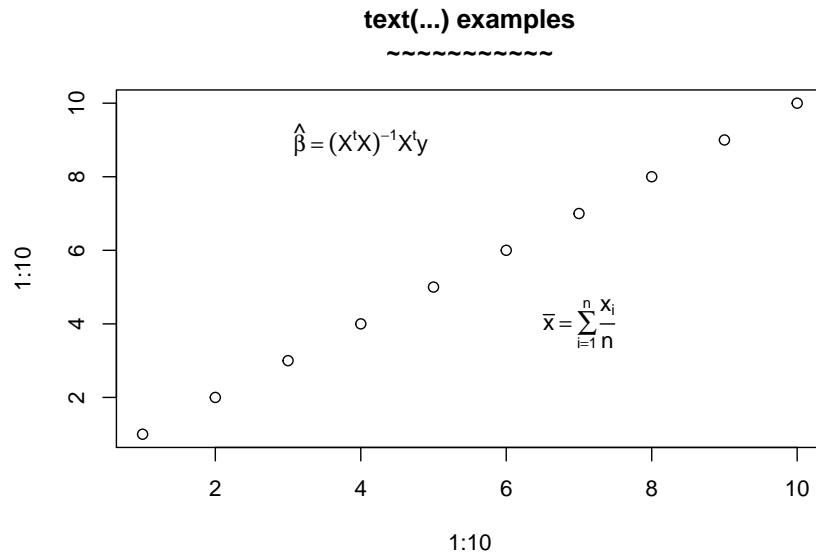


Figure 4.33: Menambahkan teks (2)

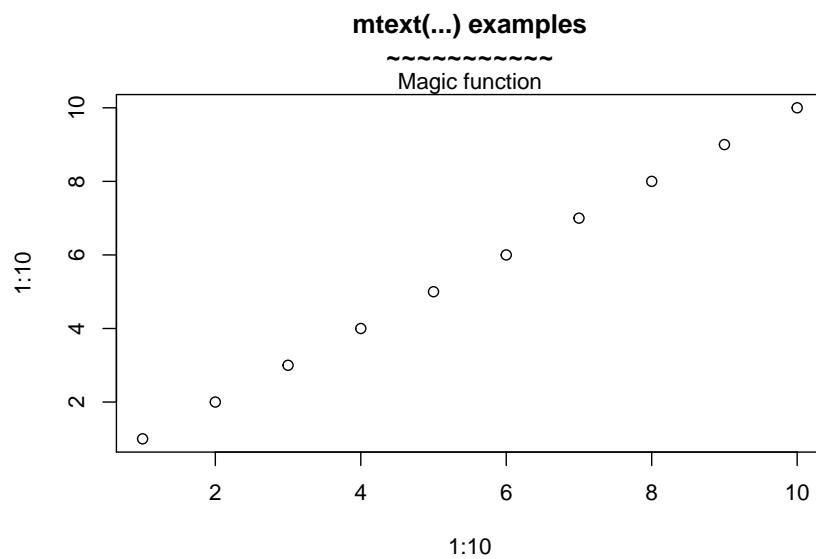


Figure 4.34: Menambahkan teks (3)

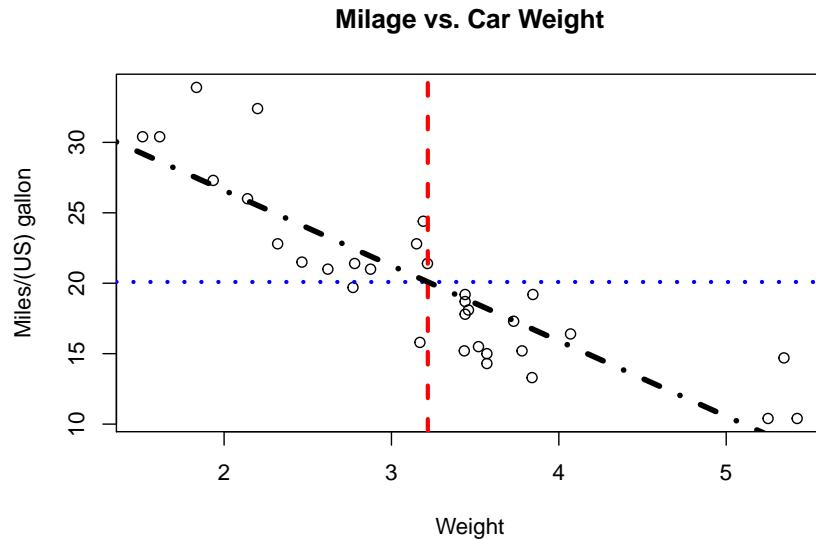


Figure 4.35: Menambahkan garis

```
abline(v=y)
```

Berikut adalah contoh sintaks bagaimana menambahkan garis pada sebuah plot dan output yang dihasilkan disajikan pada Gambar 4.35:

```
# membuat plot
plot(mtcars$wt, mtcars$mpg, main="Milage vs. Car Weight",
      xlab="Weight", ylab="Miles/(US) gallon")

# menambahkan garis vertikal di titik rata-rata weight
abline(v=mean(mtcars$wt), col="red", lwd=3, lty=2)

# menambahkan garis horizontal di titik rata-rata mpg
abline(h=mean(mtcars$mpg), col="blue", lwd=3, lty=3)

# menambahkan garis regresi
abline(lm(mpg~wt, data=mtcars), lwd=4, lty=4)
```

4.10.5 Merubah Simbol plot dan Jenis Garis

Simbol plot (jenis titik) dapat diubah dengan menambahkan argumen `pch=` pada plot. Nilai yang dimasukkan pada argumen tersebut adalah integer dengan kemungkinan nilai sebagai berikut:

- `pch = 0,square`
- `pch = 1,circle (default)`
- `pch = 2,triangle point up`
- `pch = 3,plus`
- `pch = 4,cross`
- `pch = 5,diamond`

- pch = 6,triangle point down
- pch = 7,square cross
- pch = 8,star
- pch = 9,diamond plus
- pch = 10,circle plus
- pch = 11,triangles up and down
- pch = 12,square plus
- pch = 13,circle cross
- pch = 14,square and triangle down
- pch = 15, filled square
- pch = 16, filled circle
- pch = 17, filled triangle point-up
- pch = 18, filled diamond
- pch = 19, solid circle
- pch = 20,bullet (smaller circle)
- pch = 21, filled circle blue
- pch = 22, filled square blue
- pch = 23, filled diamond blue
- pch = 24, filled triangle point-up blue
- pch = 25, filled triangle point down blue

Untuk lebih memahami bentuk simbol tersebut, penulis akan menyajikan sintaks yang menampilkan seluruh simbol tersebut pada satu grafik. Output yang dihasilkan disajikan pada Gambar 4.36:

```
generateRPointShapes<-function(){
  # menentukan parameter plot
  oldPar<-par()
  par(font=2, mar=c(0.5,0,0,0))
  # produksi titik axis
  y=rev(c(rep(1,6),rep(2,5), rep(3,5), rep(4,5), rep(5,5)))
  x=c(rep(1:5,5),6)
  # plot seluruh titik dan label
  plot(x, y, pch = 0:25, cex=1.5, ylim=c(1,5.5), xlim=c(1,6.5),
        axes=FALSE, xlab="", ylab="", bg="blue")
  text(x, y, labels=0:25, pos=3)
  par(mar=oldPar$mar,font=oldPar$font )
}

# Print
generateRPointShapes()
```

Pada R kita juga dapat mengatur jenis garis yang akan ditampilkan pada plot dengan menambahkan argumen `lty=` (*line type*) pada fungsi plot. Nilai yang dapat dimasukkan adalah nilai integer. Keterangan masing-masing nilai tersebut adalah sebagai berikut:

- lty = 0, blank
- lty = 1, solid (default)
- lty = 2, dashed
- lty = 3, dotted
- lty = 4, dotdash
- lty = 5, longdash
- lty = 6, twodash

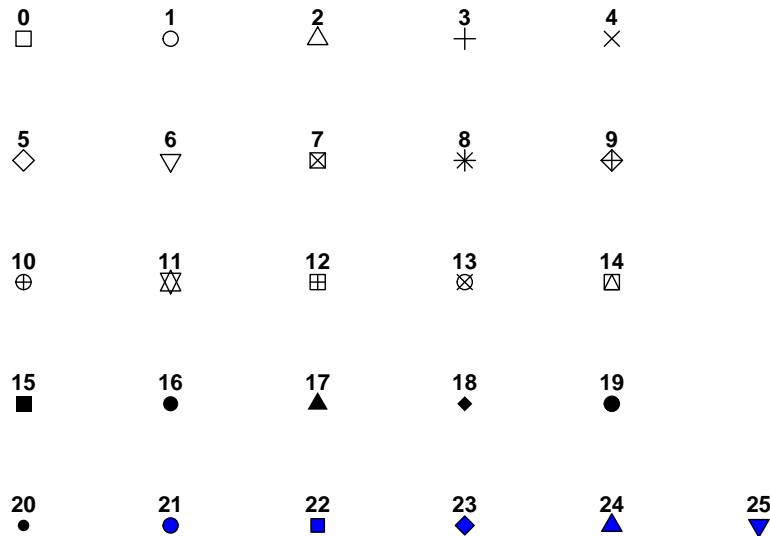


Figure 4.36: Symbol plot

Untuk lebih memahaminya, pada sintaks berikut disajikan plot seluruh jenis garis tersebut beserta output yang dihasilkannya pada Gambar 4.37:

```
generateRLineTypes<-function(){
  oldPar<-par()
  par(font=2, mar=c(0,0,0,0))
  plot(1, pch="", ylim=c(0,6), xlim=c(0,0.7), axes = FALSE ,xlab="", ylab="")
  for(i in 0:6) lines(c(0.3,0.7), c(i,i), lty=i, lwd=3)
  text(rep(0.1,6), 0:6,
       labels=c("0.'blank'", "1.'solid'", "2.'dashed'", "3.'dotted'",
               "4.'dotdash'", "5.'longdash'", "6.'twodash'"))
  par(mar=oldPar$mar,font=oldPar$font )
}
generateRLineTypes()
```

4.10.6 Mengatur Axis Plot

Kita dapat melakukan pengaturan lebih jauh terhadap axis, seperti: menambahkan axis tambahan pada atas dan bawah frame, mengubah rentang nilai axis, serta kustomisasi *tick mark* pada nilai axis. Hal ini diperlukan karena fungsi grafik dasar R tidak dapat mengatur axis secara otomatis saat plot baru ditambahkan pada plot pertama dan rentang nilai plot baru lebih besar dibanding plot pertama, sehingga sebagian nilai plot baru tidak ditampilkan pada hasil akhir.

Untuk menambahkan axis pada R kita dapat menambahkan fungsi `axis()` setelah plot dilakukan. Format yang digunakan adalah sebagai berikut:

```
axis(side, at=NULL, labels=TRUE)
```

Note:

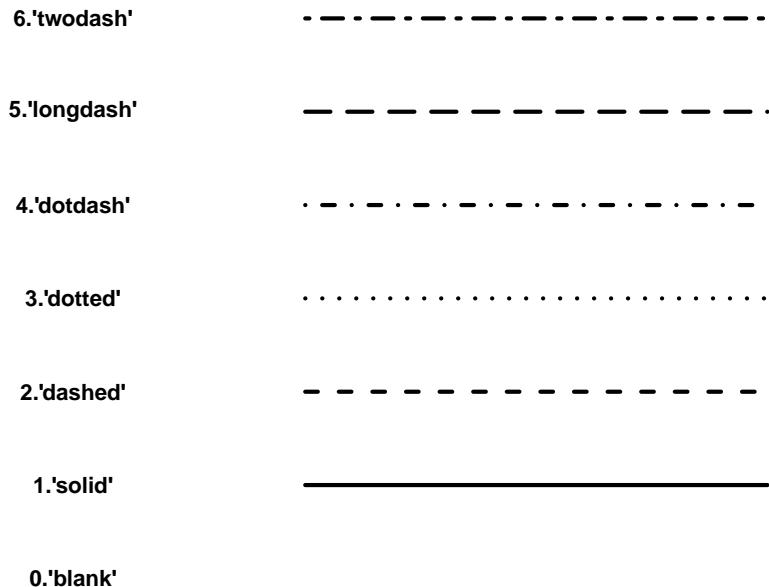


Figure 4.37: Line type

- **side:** nilai integer yang mengidikasikan posisi axix yang hendak ditambahkan. Nilai yang dapat dimasukkan adalah sebagai berikut:
 - **1:** bawah
 - **2:** kiri
 - **3:** atas
 - **4:** kanan.
 - **at:** titik dimana *tick-mark* hendak digambarkan. Nilai yang dapat dimasukkan sama dengan **side**.
 - **labels:** Teks label *tick-mark*. Dapat juga secara logis menentukan apakah anotasi harus dibuat pada *tick mark*.

Berikut contoh sintaks penerapan fungsi tersebut dan output yang dihasilkan pada Gambar 4.38:

```
# membuat vektor numerik
x <- c(1:4)
y <- x^2

# plot
plot(x, y, pch=18, col="red", type="b",
      frame=FALSE, xaxt="n") # Remove x axis

# menambahkan axis
# bawah
axis(1, 1:4, LETTERS[1:4], col.axis="blue")
# atas
axis(3, col = "darkgreen", lty = 2, lwd = 0.5)
# kanan
axis(4, col = "violet", col.axis = "dark violet", lwd = 2)
```

Kita dapat mengubah rentang nilai pada axis menggunakan fungsi `xlim()` dan `ylim()` yang menyatakan vektor nilai maksimum dan minimum rentang. Selain itu kita dapat juga melakukan transformasi baik pada sumbu x dan sumbu y. Berikut adalah argumen yang dapat ditambahkan pada fungsi grafik:

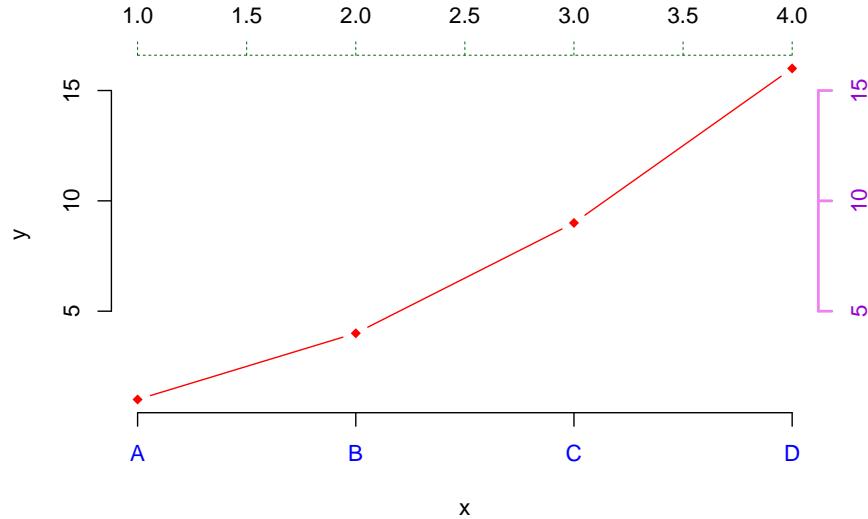


Figure 4.38: Menambahkan axis

- **xlim**: limit nilai sumbu x dengan format: `xlim(min, max)`.
- **ylim**: limit nilai sumbu x dengan format: `ylim(min, max)`.

Untuk transformasi skala log, kita dapat menambahkan argumen berikut:

- **log="x"**: transformasi log sumbu x.
- **log="y"**: transformasi log sumbu y.
- **log="xy"**: transformasi log sumbu x dan y.

Berikut adalah contoh sintaks penerapan argumen tersebut beserta output yang dihasilkan pada Gambar 4.39:

```
# membagi jendela grafik menjadi 1 baris dan 3 kolom
par(mfrow=c(1,3))

# membuat vektor numerik
x<-c(1:10); y<-x*x

# simple plot
plot(x, y)

# plot dengan pengaturan rentang skala
plot(x, y, xlim=c(1,15), ylim=c(1,150))

# plot dengan transformasi skala log
plot(x, y, log="y")
```

Kita dapat melakukan kustomisasi pada *tick mark*. Kustomisasi yang dapat dilakukan adalah merubah warna, *font style*, ukuran font, orientasi, serta menyembunyikan *tick mark*.

Argumen yang ditambahkan adalah sebagai berikut:

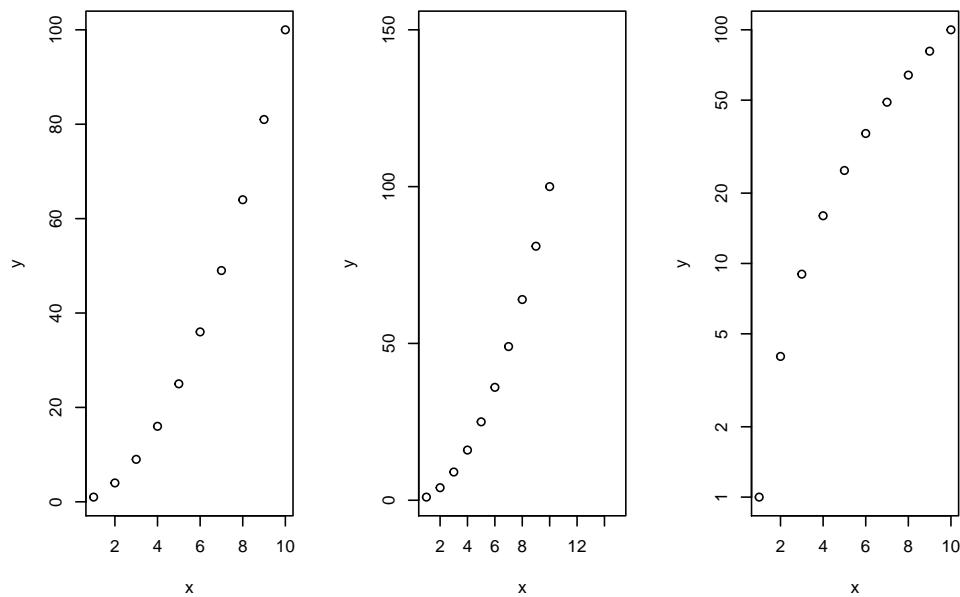


Figure 4.39: Mengubah rentang dan skala axis

- **col.axis:** warna *tick mark*.
- **font.axis:** integer yang menunjukkan *font style*. Sama dengan pengaturan judul.
- **cex.axis:** pengaturan ukuran *tick mark*.
- **las:** mengatur orientasi *tick mark*. Nilai yang dapat dimasukkan adalah sebagai berikut:
 - **0:** paralel terhadap posisi axis (default)
 - **1:** selalu horizontal
 - **2:** selalu perpendikular dengan posisi axis
 - **3:** selalu vertikal
- **xaxt** dan **yaxt:** karakter untuk menunjukkan apakah axis akan ditampilkan atau tidak. nilai dapat berupa “n”(sembunyikan) dan “s”(tampilkan).

Berikut adalah contoh penerapan argumen tersebut beserta output pada Gambar 4.40:

```
# membuat vektor numerik
x<-c(1:10); y<-x*x

# plot
plot(x,y,
      # warna
      col.axis="red",
      # font style
      font.axis=2,
      # ukuran
```

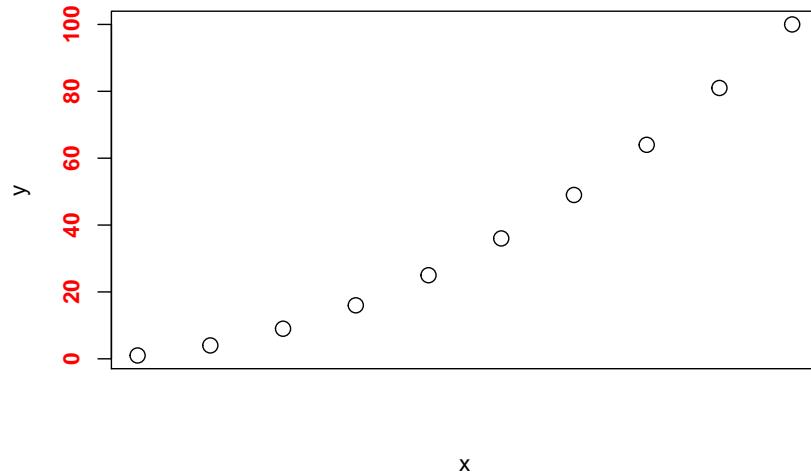


Figure 4.40: Kustomisasi tick mark

```
cex=1.5,
# orientasi
las=3,
# sembunyikan sumbu x
xaxt="n")
```

4.10.7 Mengatur Warna

Pada fungsi dasar R, warna dapat diatur dengan mengetikkan nama warna maupun kode hexadesimal. Selain itu kita juga dapat menambahkan warna lain melalui library lain yang tidak dijelaskan pada chapter ini.

Untuk penggunaan warna hexadesima kita perlu mengetikkan "#" yang diikuti oleh 6 kode warna. Untuk mempelajari kode-kode dan warna yang dihasilkan, silahkan pembaca mengunjungi situs <http://www.visibone.com/>.

Pada sintaks berikut disajikan visualisasi nama-nama warna bawaan yang ada pada R. Output yang dihasilkan disajikan pada Gambar 4.41:

```
showCols <- function(cl=colors(), bg = "grey",
                      cex = 0.75, rot = 30) {
  m <- ceiling(sqrt(n <- length(cl)))
  length(cl) <- m*m; cm <- matrix(cl, m)
  require("grid")
  grid.newpage(); vp <- viewport(w = .92, h = .92)
  grid.rect(gp=gpar(fill=bg))
  grid.text(cm, x = col(cm)/m, y = rev(row(cm))/m, rot = rot,
            vp=vp, gp=gpar(cex = cex, col = cm))
}

# print 60 nama warna pertama
showCols(bg="gray20", cl=colors() [1:60], rot=30, cex=0.9)
```

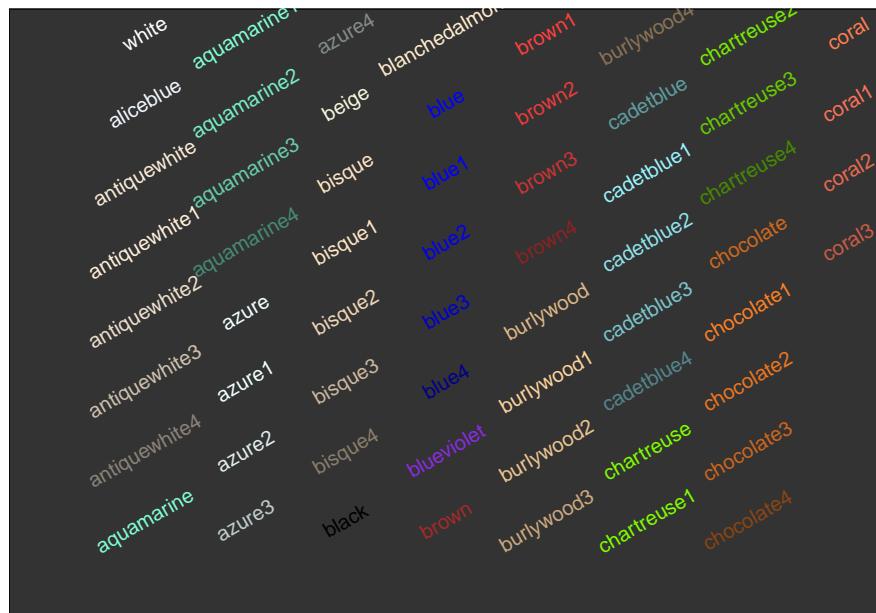


Figure 4.41: Nama warna

```
## Loading required package: grid
```

4.11 Alternatif Library Dasar Lain

Kita juga dapat melakukan visualisasi menggunakan library lain yang memiliki tampilan mirip dengan fungsi visualisasi dasar R. Bedanya adalah library-library ini memberikan fungsi tambahan sehingga visualisasi yang dihasilkan menjadi lebih praktis.

4.11.1 Scatterplot Menggunakan Library car

Library **car** menyediakan alternatif lain visualisasi menggunakan scatterplot. Berikut adalah contoh sintaks dan output yang dihasilkan pada Gambar 4.42:

```
# memasang paket
# install.packages("car")

# memuat paket
library(car)

## Warning: package 'car' was built under R version 3.5.3

# plot
scatterplot(Volume~Height, data=trees)
```

Pada grafik tersebut terkandung beberapa elemen penting, yaitu:

- titik observasi

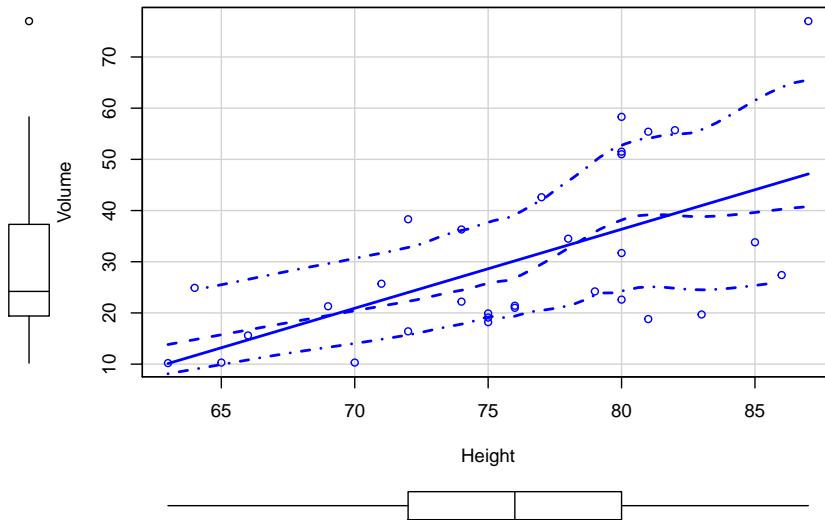


Figure 4.42: Enhanced scatterplot

- garis regresi (garis lurus)
- non-parametric regression smooth (*dashed line*)
- garis smoothed conditional (*point dashed line*)
- box plot masing-masing variabel.

4.11.2 Matriks Scatterplot Menggunakan Library psych

FUNGSI `pairs.panels()` PADA LIBRARY `psych` DAPAT DIGUNAKAN UNTUK MEMBUAT MATRIKS SCATTERPLOT. GRAFIK YANG DIHASILKAN JUGA LEBIH RINGKAS DAN MENAMPILKAN FUNKSIONAL LAIN PADA BAGIAN DIAGONAL LAIN BERUPA HISTOGRAM DAN DENSITY PLOT YANG DAPAT MENUNJUKKAN DISTRIBUSI DARI VARIABEL YANG ADA. SELAIN ITU PADA FUNKSIONALITAS GRAFIK JUGA DAPAT DITINGKATKAN DENGAN PENAMBAHAN NILAI KORELASI ANTAR VARIABEL YANG SCARA DEFAULT DITAMBAHKAN PADA PANEL ATAS. BERIKUT ADALAH CONTOH SINTAKS DAN OUTPUT YANG DIHASILKAN PADA GAMBAR 4.43:

```
# memasang paket
# install.packages("psych")

# memuat paket
library(psych)

## Warning: package 'psych' was built under R version
## 3.5.3

# plot
pairs.panels(trees,
  method = "pearson", # metode korelasi
  hist.col = "grey",
  density = TRUE, # menampilkan plot densitas
  ellipses = FALSE, # menampilkan correlation ellipses
```

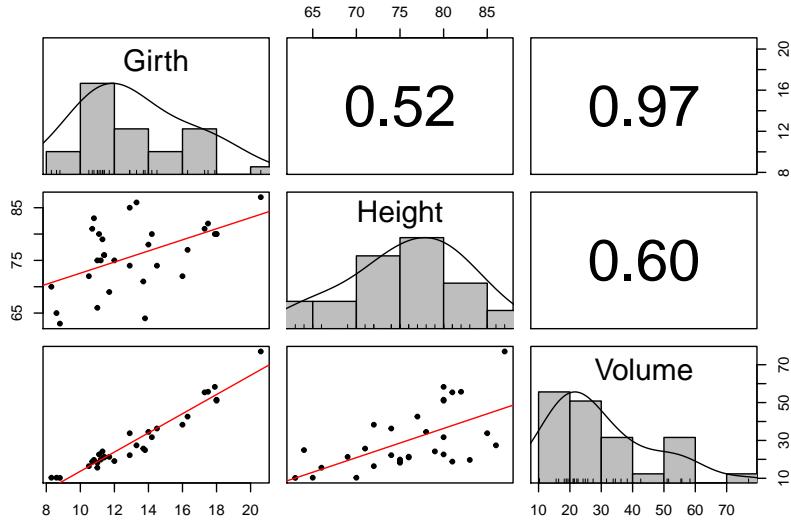


Figure 4.43: Enhanced scatterplot matrices

```
lm = TRUE # menampilkan garis regresi linier
)
```

4.11.3 Box Plot Menggunakan Library gplots

Fungsi `boxplot2()` pada paket `gplots` memberikan fungsionalitas lebih dibandingkan box plot yang dihasilkan dari fungsi dasar R. Plot yang dihasilkan akan menampilkan jumlah observasi pada tiap box. Berikut adalah contoh sintaks penerapan dan output yang dihasilkan pada Gambar 4.44:

```
# memasang paket
# install.packages("gplots")

# memuat paket
library(gplots)

## Warning: package 'gplots' was built under R version
## 3.5.3

# plot
boxplot2(len ~ dose, data = ToothGrowth)
```

4.11.4 QQ Plot Menggunakan Library car

Fungsi `qqPlot()` pada library `car` dapat pula digunakan untuk membuat qq plot. Kelebihannya adalah qqplot yang dihasilkan akan dilengkapi dengan garis referensi yang memudahkan dalam membaca apakah data masih dalam rentang distribusi normal atau tidak. Selain itu, untuk membuatnya juga hanya diperlukan satu perintah saja. Hal ini tentu berbeda ketika kita menggunakan fungsi dasar R. Berikut adalah contoh sintaks penerapan dan output yang dihasilkan pada Gambar 4.45:

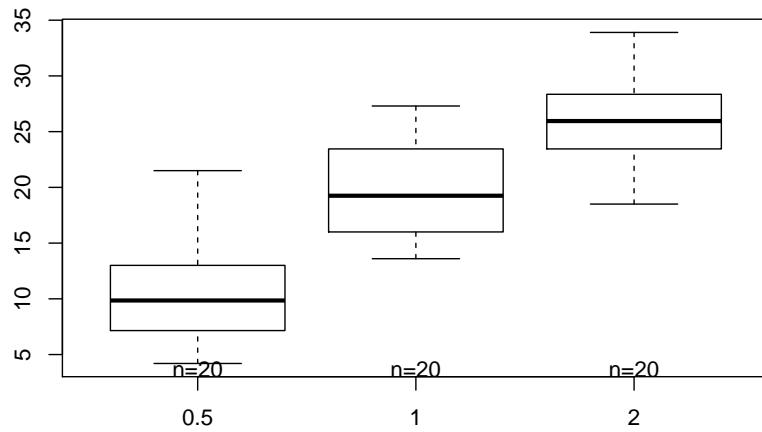


Figure 4.44: Enhanced box plot

```
# memasang paket
# install.packages("car")

# memuat paket
library(car)

# plot
qqPlot(trees$Height)

## [1] 3 20
```

4.11.5 Plot Group Means Menggunakan Library gplots

Plot ini akan sering kita gunakan saat melakukan analisis statistik menggunakan anova baik anova satu arah maupun dua arah. Plot ini berguna untuk melihat adanya interaksi antar faktor saat melakukan analisis anova dua arah. Berikut adalah contoh sintaks penerapan dan output yang dihasilkan pada Gambar 4.46:

```
# memasang paket
# install.packages("gplots")

# memuat paket
library(gplots)

# plot
plotmeans(len ~ dose, data = ToothGrowth)
```

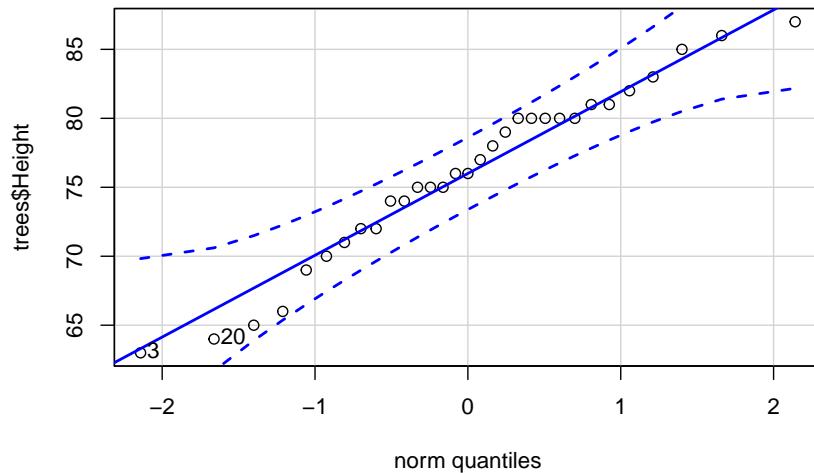


Figure 4.45: Enhanced qq plot

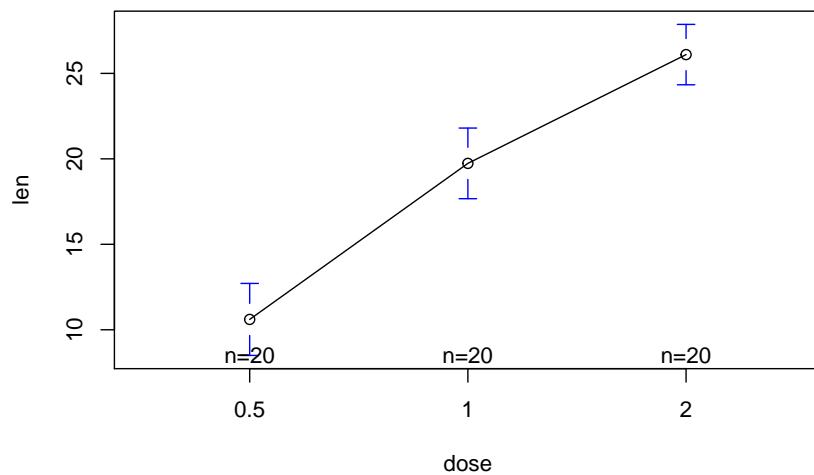


Figure 4.46: Plot group means

4.12 Referensi

1. Maindonald, J.H. 2008. **Using R for Data Analysis and Graphics Introduction, Code and Commentary.** Centre for Mathematics and Its Applications Australian National University.
2. Scherber, C. 2007. **An introduction to statistical data analysis using R.** R_Manual Goettingen.
3. Venables, W.N. Smith D.M. and R Core Team. 2018. **An Introduction to R.** R Manuals.
4. STHDA. **R Base Graphs.** <http://www.sthda.com/english/wiki/r-base-graphs>

Chapter 5

Visualisasi Data Menggunakan GGPlot

Library `ggplot2` merupakan implementasi dari *The Grammar of Graphics* yang ditulis oleh **Leland Wilkinson**. `ggplot2` merupakan library yang dikembangkan oleh **Hadley Wickham** ketika ia sedang menempuh kuliah di **Iowa State University** dan masih dikembangkan hingga sekarang.

`ggplot2` merupakan paket visualisasi yang powerfull. Kita dapat menggunakannya bersamaan dengan *piping operator* yang disediakan oleh paket `dplyr` sehingga menambah kemudahan kita dalam melakukan analisis data.

Grafik `ggplot2` terdiri dari sejumlah komponen kunci. Berikut adalah sejumlah komponen kunci yang membentuk grafik `ggplot2`.

- **data frame:** menyimpan semua data yang akan ditampilkan di plot.
- **aesthetic mapping:** menggambarkan bagaimana data dipetakan ke warna, ukuran, bentuk, lokasi. Dalam plot diberikan pada fungsi `aes()`
- **geoms:** objek geometris seperti titik, garis, bentuk.
- **facets:** menjelaskan bagaimana plot bersyarat / panel harus dibangun.
- **stats:** transformasi statistik seperti binning, quantiles, smoothing.
- **scales:** skala apa yang digunakan oleh *aesthetic map* (contoh: pria = merah, wanita = biru).
- **coordinate system:** menggambarkan sistem di mana lokasi geom akan digambarkan.

Sebelum kita mulai memcoba melakukan visualisasi data menggunakan `ggplot2`, kita perlu menginstall dan memuat terlebih dahulu library `ggplot2`. Berikut adalah sintaks yang digunakan untuk menginstall dan memuat paket `ggplot2`:

```
# memasang paket
# install.packages('ggplot2')

# memuat paket
library(ggplot2)
```

Dataset yang akan kita gunakan adalah dataset `gapminder`. Dataset ini berisi data demografi penduduk dari berbagai negara dan benua. Untuk dapat menggunakannya kita perlu menginstall dan memuatnya terlebih dahulu. Berikut adalah sintaks untuk menginstall dan memuat dataset tersebut:

```
# memasang paket
# install.packages("gapminder")

# memuat paket
library(gapminder)

## Warning: package 'gapminder' was built under R version
## 3.5.3

# memuat paket dplyr dan tibble
library(dplyr)
library(tibble)

# melihat struktur dataset
glimpse(gapminder)

## Observations: 1,704
## Variables: 6
## $ country    <fct> Afghanistan, Afghanistan, Afghan...
## $ continent   <fct> Asia, Asia, Asia, Asia, Asia, As...
## $ year        <int> 1952, 1957, 1962, 1967, 1972, 19...
## $ lifeExp     <dbl> 28.80, 30.33, 32.00, 34.02, 36.0...
## $ pop         <int> 8425333, 9240934, 10267083, 1153...
## $ gdpPercap   <dbl> 779.4, 820.9, 853.1, 836.2, 740....
```

```
# melihat variabel year
unique(gapminder$year)
```

```
## [1] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997
## [11] 2002 2007
```

Dataset gapminder memiliki 6 variabel dan 1704 observasi. 20 observasi pertama dataset gapminder dapat dilihat pada Tabel 5.1

5.1 Scatterplot

Scatterplot dapat dibuat pada ggplot2 menggunakan fungsi `geom_point()`. Format sederhananya dituliskan sebagai berikut:

```
ggplot(data, aes(...))+
  geom_point(size, color, shape)
```

Berikut adalah contoh sederhana scatterplot variabel `lifeExp` terhadap variabel `gdpPercap`. Output yang dihasilkan disajikan pada Gambar 5.1:

```
ggplot(gapminder, aes(gdpPercap, lifeExp))+
  geom_point()
```

Table 5.1: 20 observasi pertama dataset gapminder

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.80	8425333	779.4
Afghanistan	Asia	1957	30.33	9240934	820.9
Afghanistan	Asia	1962	32.00	10267083	853.1
Afghanistan	Asia	1967	34.02	11537966	836.2
Afghanistan	Asia	1972	36.09	13079460	740.0
Afghanistan	Asia	1977	38.44	14880372	786.1
Afghanistan	Asia	1982	39.85	12881816	978.0
Afghanistan	Asia	1987	40.82	13867957	852.4
Afghanistan	Asia	1992	41.67	16317921	649.3
Afghanistan	Asia	1997	41.76	22227415	635.3
Afghanistan	Asia	2002	42.13	25268405	726.7
Afghanistan	Asia	2007	43.83	31889923	974.6
Albania	Europe	1952	55.23	1282697	1601.1
Albania	Europe	1957	59.28	1476505	1942.3
Albania	Europe	1962	64.82	1728137	2312.9
Albania	Europe	1967	66.22	1984060	2760.2
Albania	Europe	1972	67.69	2263554	3313.4
Albania	Europe	1977	68.93	2509048	3533.0
Albania	Europe	1982	70.42	2780097	3630.9
Albania	Europe	1987	72.00	3075321	3738.9

Kita dapat mengubah warna, jenis, dan ukuran titik pada scatterplot. Pengubahan warna dan jenis titik berguna untuk menunjukkan grup data pada grafik. Sedangkan perubahan ukuran titik sangat berguna untuk menunjukkan nilai variabel lain khususnya variabel kontinyu pada sebuah titik. Berikut adalah contoh penerapannya. Output yang dihasilkan disajikan pada Gambar 5.2 sampai dengan Gambar 5.4:

```
ggplot(gapminder, aes(gdpPercap, lifeExp, color=continent))+
  geom_point()+
  # merubah sumbu x kedalam fungsi log
  scale_x_log10()
```

```
ggplot(gapminder, aes(gdpPercap, lifeExp, shape=continent))+
  geom_point()+
  # merubah sumbu x kedalam fungsi log
  scale_x_log10()
```

```
ggplot(gapminder, aes(gdpPercap, lifeExp,
                      size=pop, color=continent))+
  geom_point()+
  # merubah sumbu x kedalam fungsi log
  scale_x_log10()
```

Untuk menunjukkan asosiasi antara dua variabel kontinyu kita juga dapat menambahkan garis regresi dan confidence interval garis regresinya. Fungsi yang digunakan adalah `geom_smooth()`. Secara default fungsi tersebut akan membuat garis loess regression pada grafik. Agar dapat membuat garis regresi linier kita perlu menambahkan argumen `method="lm"`. Selain itu, jika kita tidak ingin menampilkan garis confidence interval kita dapat menambahkan argumen `se=False`. Format sederhananya disajikan pada sintaks berikut:

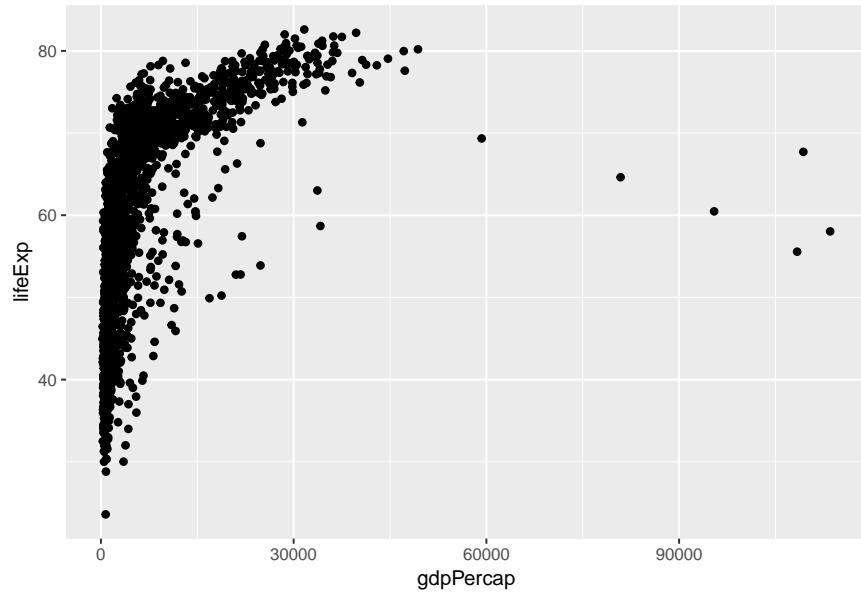


Figure 5.1: Scatterplot lifeExp vs gdpPerCap

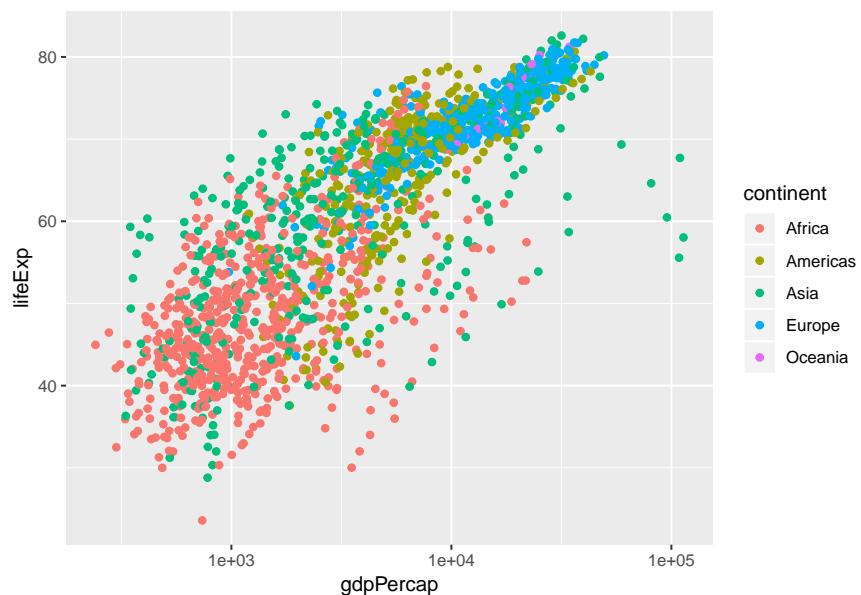


Figure 5.2: Scatterplot lifeExp vs gdpPerCap tiap benua (1)

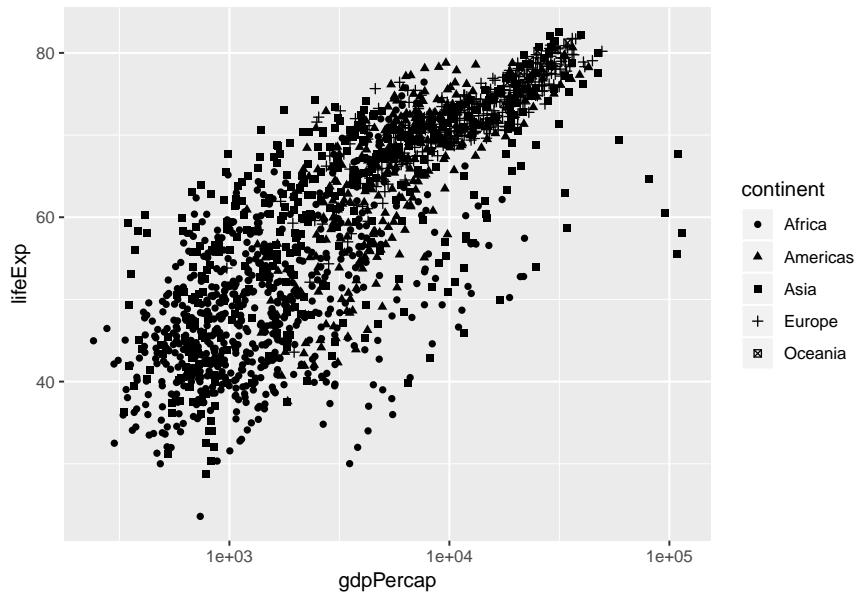


Figure 5.3: Scatterplot lifeExp vs gdpPerCap tiap benua (2)

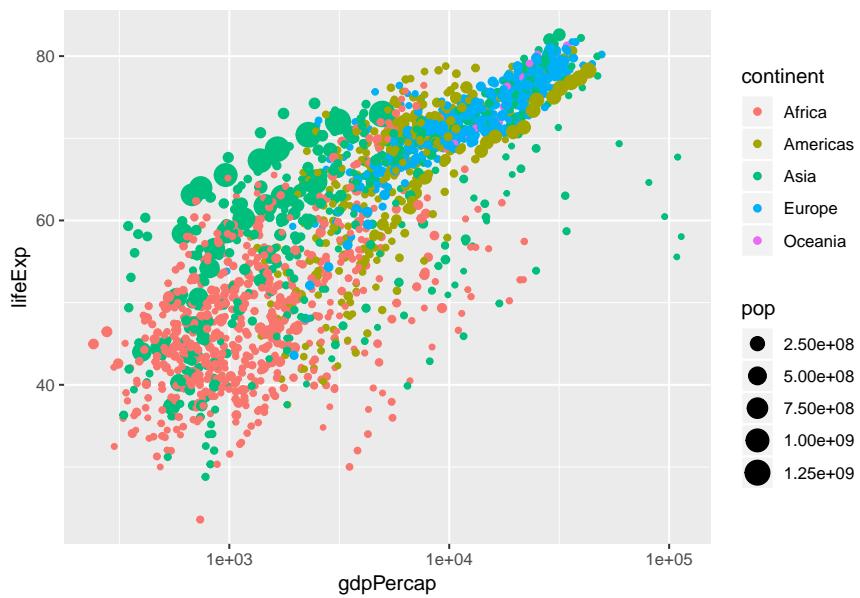


Figure 5.4: Scatterplot lifeExp vs gdpPerCap dan populasi tiap negara dan benua

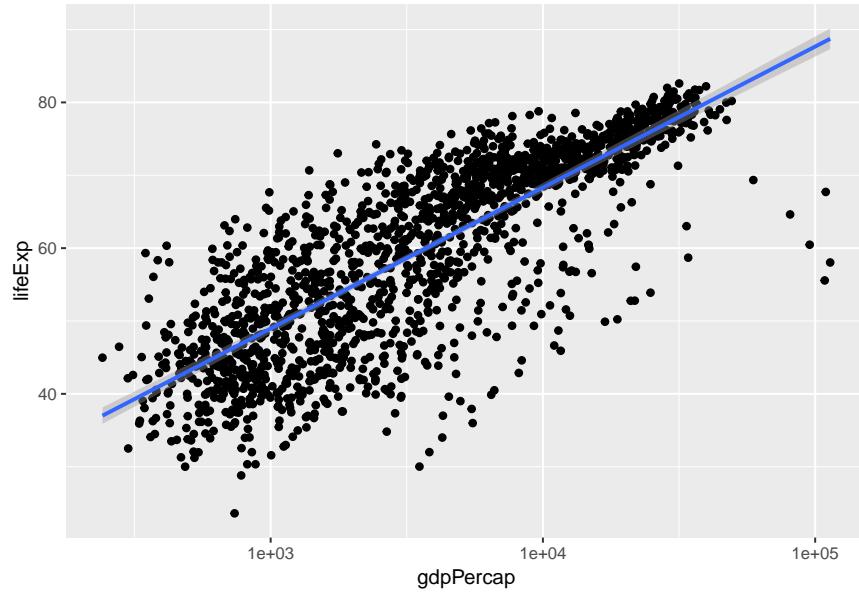


Figure 5.5: Scatterplot lifeExp vs gdpPercap dengan garis penghalusan regresi linier

```
geom_smooth(method="auto", se=TRUE, fullrange=FALSE, level=0.95)
```

Note:

- **method:** metode penghalusan yang digunakan. Nilai yang dapat dimasukkan adalah lm, glm, gam, loess, rlm.
- method="loess": merupakan nilai default pada fungsi dan menghasilkan metode penghalusan loess regression.
- method="lm": menghasilkan metode penghalusan regresi linier. Kita juga dapat melakukan spesifikasi terhadap fungsi persamaan regresi yang digunakan dengan menambahkan argumen formula=y~x....
- **se:** nilai logis. Jika TRUE garis confidence interval akan ditampilkan sepanjang garis penghalusan.
- **fullrange:** nilai logis. Jika TRUE kecokongan mencakup seluruh plot.
- **level:** level confidence interal yang digunakan. Secara default bernilai 0.95.

Berikut adalah contoh sintaks penerapan pada variabel gdpPercap dan lifeExp. Output yang dihasilkan disajikan pada Gambar 5.5:

```
ggplot(gapminder, aes(gdpPercap, lifeExp))+
  geom_point()+
  # merubah sumbu x kedalam fungsi log
  scale_x_log10()+
  # menambahkan smoothing method
  geom_smooth(method="lm", level=0.99)
```

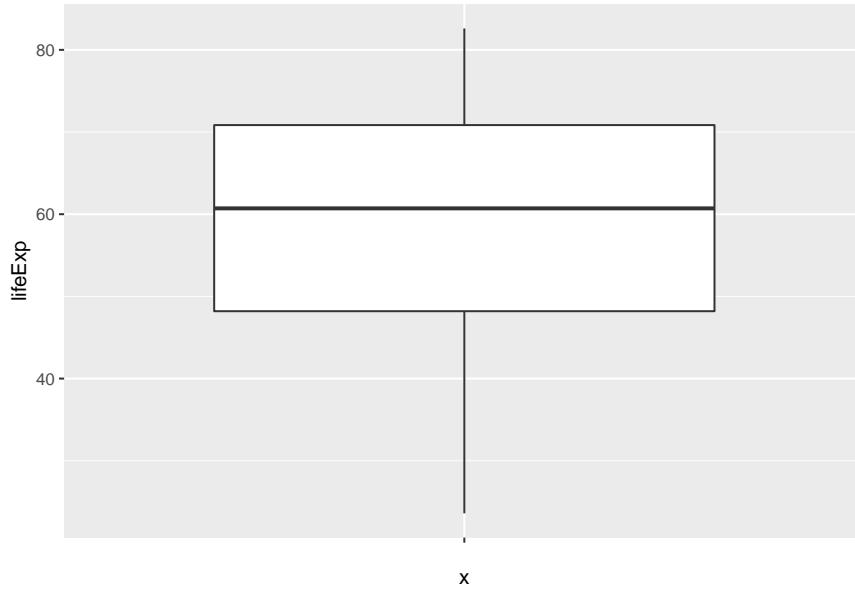


Figure 5.6: Box plot variabel lifeExp

5.2 Box Plot dan Violin Plot

Box plot merupakan visualisasi yang powerful dalam menggambarkan distribusi data, melihat adanya outlier, serta membandingkan distribusi antar data. Format visualisasi dapat dituliskan sebagai berikut:

```
ggplot(data, aes(...))+
  geom_boxplot(geom_boxplot(outlier.colour="black",
                            outlier.shape=16,
                            outlier.size=2,
                            notch=FALSE))
```

Note:

- **outlier.colour, outlier.shape, outlier.size:** Warna, bentuk dan ukuran untuk titik-titik outlier.
- **notch:** nilai logis. Jika TRUE, buat **notched box plot**. *Notch* menunjukkan *confidence interval* di sekitar median yang biasanya didasarkan pada median $\pm 1,58 \cdot \frac{(IQR)}{\sqrt{(n)}}$. *Notch* digunakan untuk membandingkan kelompok; jika takik dua kotak tidak tumpang tindih, ini adalah bukti kuat bahwa median berbeda.

Berikut merupakan contoh visualisasi variabel `lifeExp` pada dataset `gapminder`. Output yang dihasilkan disajikan pada Gambar 5.6:

```
ggplot(gapminder, aes("", lifeExp))+
  geom_boxplot()
```

Kita dapat melakukan visualisasi bagi setiap kelompok data. Pada sintaks berikut visualisasi dilakukan untuk variabel `lifeExp` pada tiap `continent`. Pada contoh berikut akan ditampilkan cara menambahkan titik rata-rata dan warna pada masing-masing grup. Output yang dihasilkan disajikan pada Gambar 5.7:

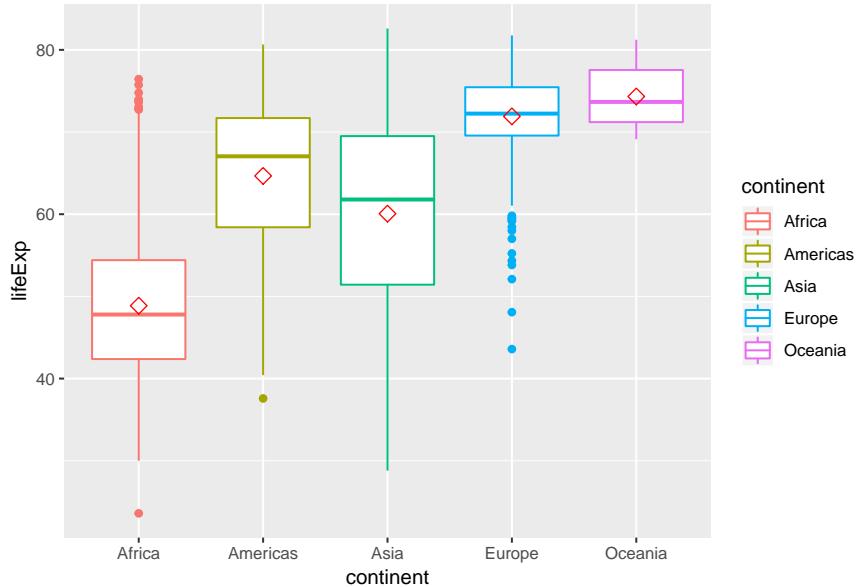


Figure 5.7: Box plot variabel lifeExp pada tiap continent

```
ggplot(gapminder, aes(continent, lifeExp, color=continent))+
  geom_boxplot()+
  stat_summary(fun.y=mean, geom="point",
              shape=23, size=3, color="red")
```

Misalkan kita ingin mengetahui perubahan distribusi dari variabel `lifeExp` pada masing-masing `continent` pada tahun 1952 dan 2007. Untuk melakukannya kita perlu melakukan subset pada dataset `gapminder` untuk memfilter data pada tahun 1952 dan 2007. Data selanjutnya dilakukan input kedalam fungsi `ggplot()`. Berikut adalah contoh sintaks yang digunakan. Output yang dihasilkan disajikan pada Gambar 5.8:

```
gapminder %>%
  filter(year==1952 | year==2007) %>%
  ggplot(aes(continent, lifeExp, fill=factor(year)))+
  geom_boxplot(notch=TRUE)
```

Berdasarkan Gambar 5.8 terlihat bahwa usia harapan hidup pada tiap benua meningkat sejak tahun 1952 sampai 2007. Selain itu, peningkatan tersebut bersifat signifikan yang ditunjukkan dari tidak adanya *notch* yang saling overlap pada masing-masing benua.

Untuk lebih detailnya kita akan coba melakukan visualisasi pada benua Asia untuk melihat perubahan variabel `lifeExp`. Berikut adalah sintaks yang digunakan dan output yang dihasilkan disajikan pada Gambar 5.9:

```
gapminder %>%
  filter(continent=="Asia") %>%
  ggplot(aes(factor(year), lifeExp))+
  geom_boxplot()
```

Violin plot memiliki kesamaan dengan box plot. Perbedaannya terletak pada violin plot tidak hanya menyajikan data titik-titikkuartil data, namun violin plot juga menampilkan kernel probabilitas distibusi data. Fungsi yang digunakan untuk membuatnya adalah `geom_violin()`.

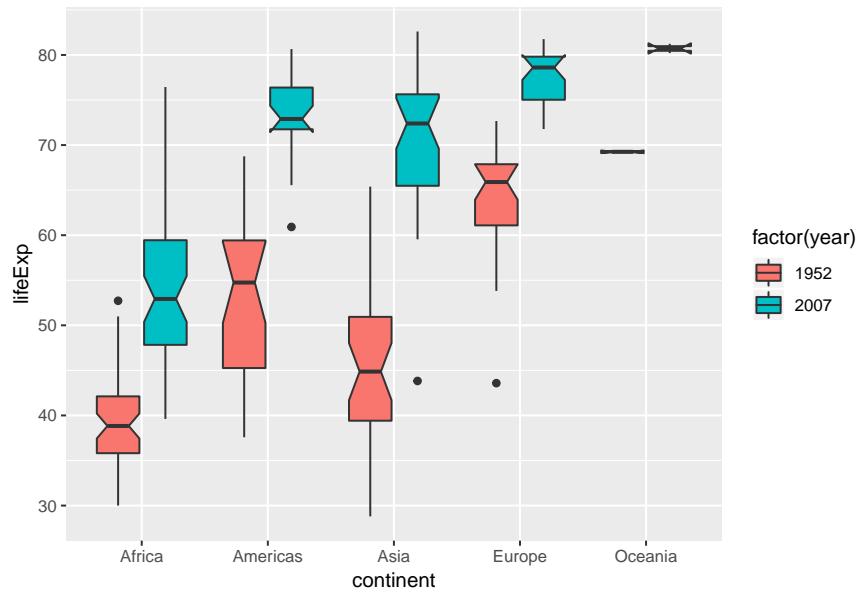


Figure 5.8: Box plot variabel lifeExp pada tiap continent (1952 dan 2007)

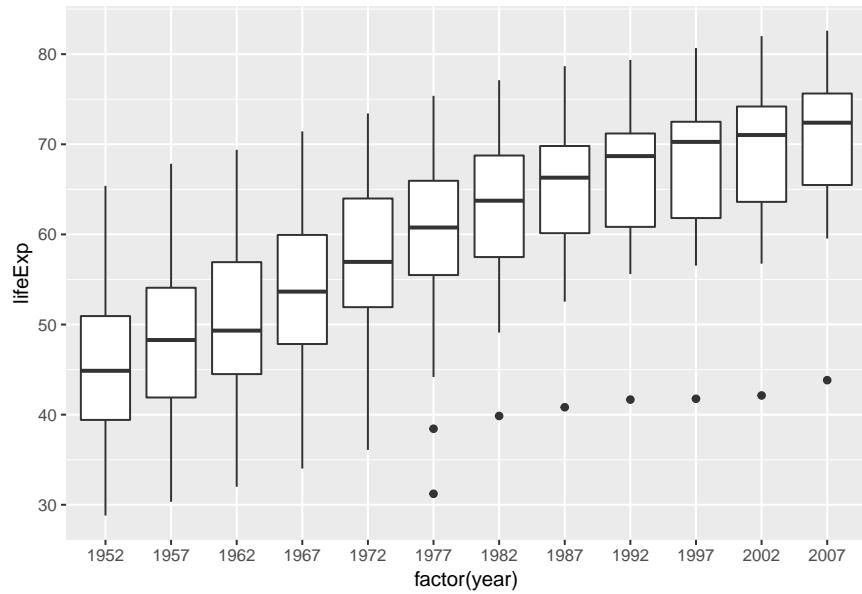


Figure 5.9: Box plot variabel lifeExp Benua Asia

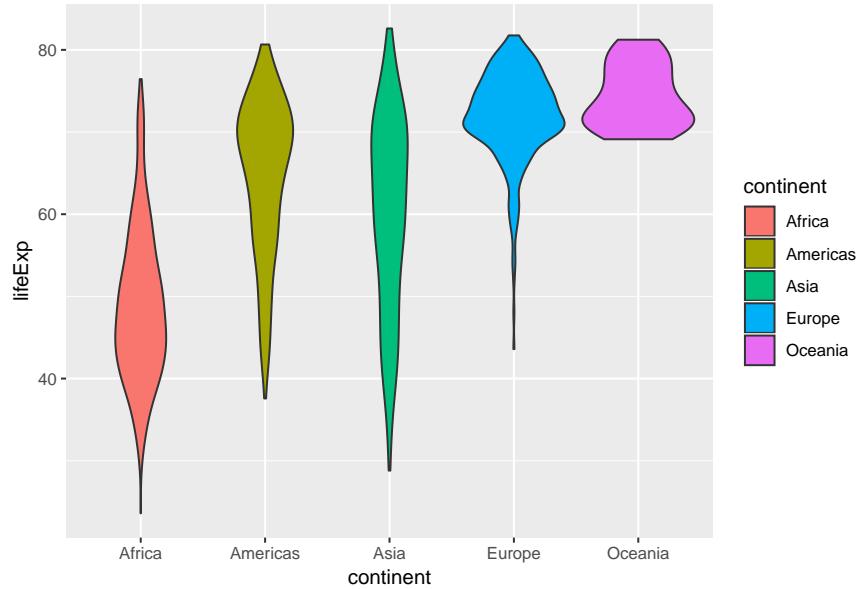


Figure 5.10: Violin plot variabel lifeExp pada masing-masing benua

Pada dataset `gapminder` kita ingin mevisualisasikan distribusi `lifeExp` pada masing-masing `continent`. Berikut adalah contoh sintaks untuk membuat visualisasi dasar violin plot. Output yang dihasilkan disajikan pada Gambar 5.10:

```
gapminder %>%
  ggplot(aes(continent, lifeExp, fill=continent)) +
  # violin plot
  geom_violin()
```

Kita juga dapat melakukan modifikasi terhadap violin plot tersebut seperti penambahan titik kuartil, titik mean dan modifikasi terhadap warna tampilakannya. COntoh sintaksnya dan output disajikan pada Gambar 5.11:

```
gapminder %>%
  ggplot(aes(continent, lifeExp, fill=continent)) +
  # violin plot
  geom_violin() +
  # menambahkan boxplot dengan lebar 0.1
  geom_boxplot(width=0.1, fill="white") +
  # menambahkan titik mean
  stat_summary(fun.y=mean, geom="point",
              # ukuran dan jenis titik
              size=1, shape=23,
              # warna titik
              color="red", fill="white")
```

5.3 Bar Plot

Pada `ggplot2` bar plot dapat dibuat menggunakan fungsi `geom_bar()`. Untuk membuat bar plot, langkah pertama yang perlu dilakukan adalah membuat tabulasi data variabel terlebih dahulu. Berikut adalah

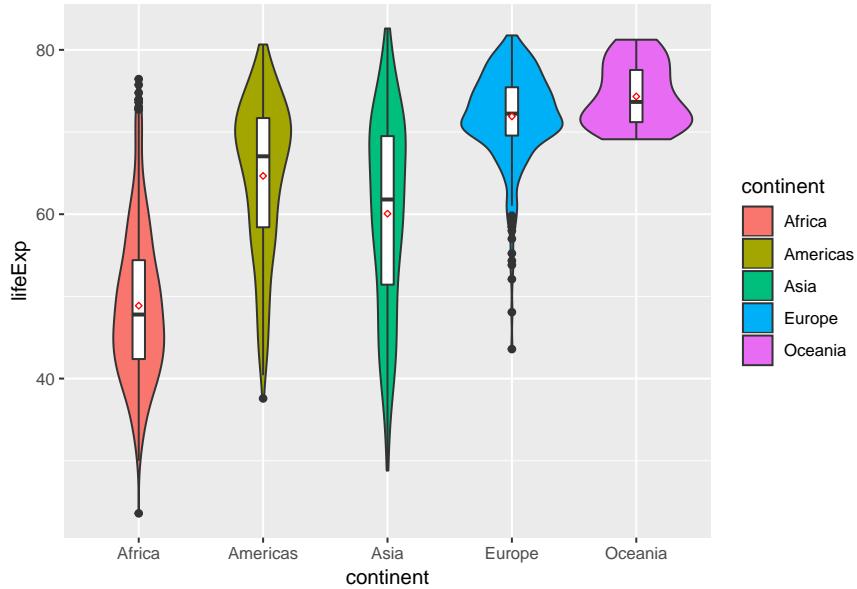


Figure 5.11: Violin plot variabel lifeExp pada masing-masing benua (2)

contoh sintaks untuk membuat bar plot dari rata-rata lifeExp pada masing-masing continent. Output yang dihasilkan disajikan pada Gambar 5.12:

```
gapminder %>%
  # kelompokkan berdasarkan continent
  group_by(continent)%>%
  # membuat ringkasan data
  summarize(mean_lifeExp=mean(lifeExp))%>%
  # urutkan dari yang terbesar
  arrange(desc(mean_lifeExp))%>%
  # plot
  ggplot(aes(continent, mean_lifeExp))+ 
  # membuat bar plot berdasarkan nilai observasi
  geom_bar(stat="identity")
```

Kita juga dapat membuat bar plot dengan garis confidence interval. Untuk melakukannya kita perlu terlebih dahulu menghitung standard error dari data. Standard error selanjutnya digunakan untuk menghitung nilai atas dan bawah dari nilai rata-rata. Berikut adalah contoh visualisasi bar plot dengan confidence interval (Gambar 5.13):

```
gapminder %>%
  # kelompokkan berdasarkan continent
  group_by(continent)%>%
  # membuat ringkasan data
  summarize(mean_lifeExp=mean(lifeExp),
           n=n(), sd=sd(lifeExp),
           se=sd/sqrt(n))%>%
  # plot
  ggplot(aes(continent, mean_lifeExp))+ 
  # membuat bar plot
  geom_bar(stat="identity", color="white")+
```

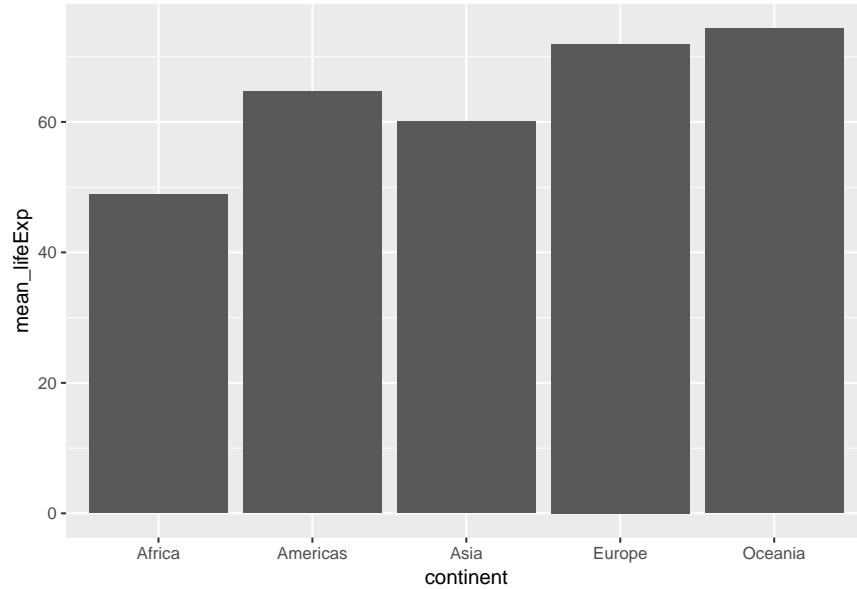


Figure 5.12: Bar plot rata-rata lifeExp masing-masing benua

```
# menambahkan error bar
geom_errorbar(aes(ymin=mean_lifeExp-se,
                   ymax=mean_lifeExp+se),
               width=0.2)
```

Kita juga dapat melakukannya pada visualisasi data beberapa grup. Berikut adalah contoh sintaks dan output (Gambar 5.14) bar plot dengan beberapa grup:

```
gapminder %>%
  # filter data tahun 1952 dan 2007
  filter(year==1952 | year==2007) %>%
  # Ubah year menjadi faktor
  mutate(year=as.factor(year)) %>%
  # kelompokkan berdasarkan continent
  group_by(continent, year) %>%
  # membuat ringkasan data
  summarize(mean_lifeExp=mean(lifeExp),
            n=n(), sd=sd(lifeExp),
            se=sd/sqrt(n)) %>%
  # plot
  ggplot(aes(continent, mean_lifeExp,
             fill=year)) +
  # membuat bar plot
  geom_bar(stat="identity",
           position=position_dodge()) +
  # menambahkan error bar
  geom_errorbar(aes(ymin=mean_lifeExp-se,
                    ymax=mean_lifeExp+se),
                width=0.2,
                position=position_dodge(0.9))
```

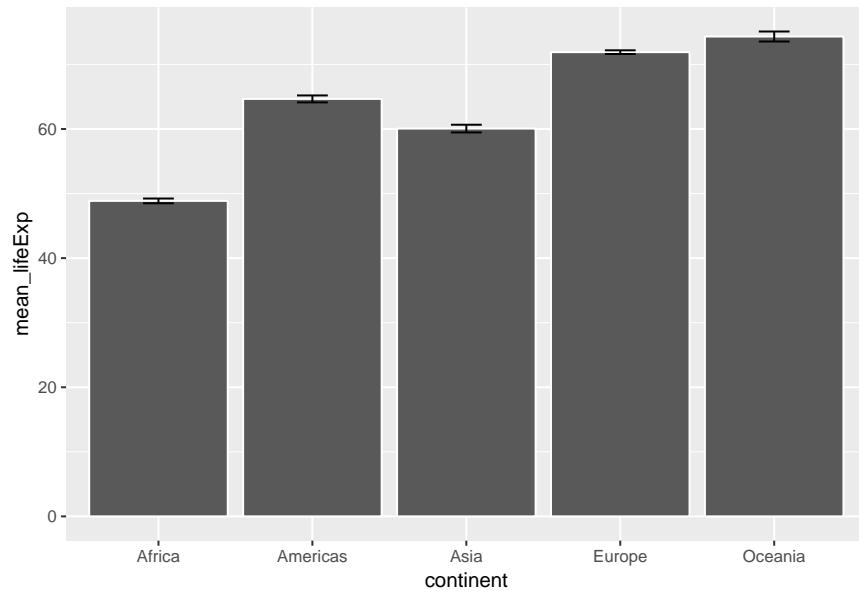


Figure 5.13: Bar plot rata-rata lifeExp masing-masing benua dengan confidence interval

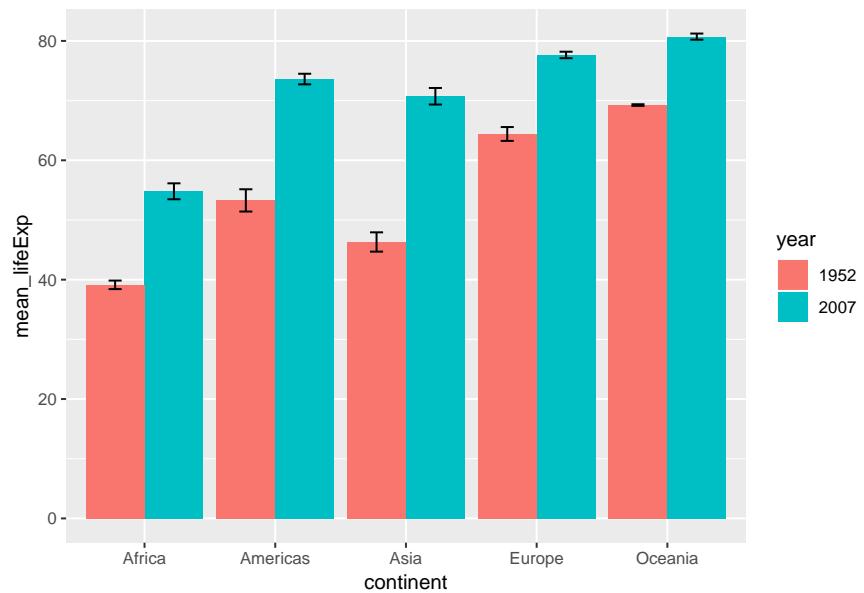


Figure 5.14: Bar plot rata-rata lifeExp masing-masing benua (1952 dan 2007) dengan confidence interval

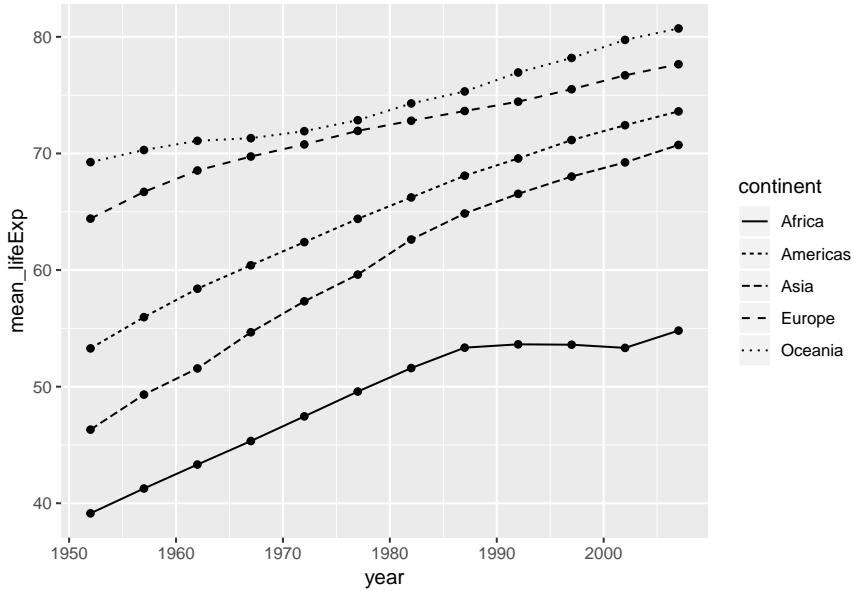


Figure 5.15: Line plot lifeExp masing-masing benua

5.4 Line Plot

Line plot dapat digunakan untuk menunjukkan adanya perubahan pada selang waktu tertentu. Pada `ggplot2`, line plot dapat dibuat menggunakan fungsi `geom_line()`. Berikut adalah contoh sintaks dan grafik (Gambar 5.15) untuk membuat line plot:

```
gapminder%>%
  # kelompokkan data berdasarkan year dan continent
  group_by(year,continent)%>%
  # ringkasan data
  summarize(mean_lifeExp=mean(lifeExp))%>%
  # plot
  ggplot(aes(year, mean_lifeExp,
             linetype=continent))+ 
  # membuat line plot
  geom_line()+
  # menambahkan point
  geom_point()
```

Kita juga dapat menambahkan error bar pada line plot. Berikut adalah contoh sintak dan grafik (Gambar 5.16) yang dihasilkan:

```
gapminder%>%
  # filter benua asia
  filter(continent=="Asia")%>%
  # kelompokkan data berdasarkan year dan continent
  group_by(year)%>%
  # ringkasan data
  summarize(mean_lifeExp=mean(lifeExp),
           sd=sd(lifeExp))%>%
  # plot
```

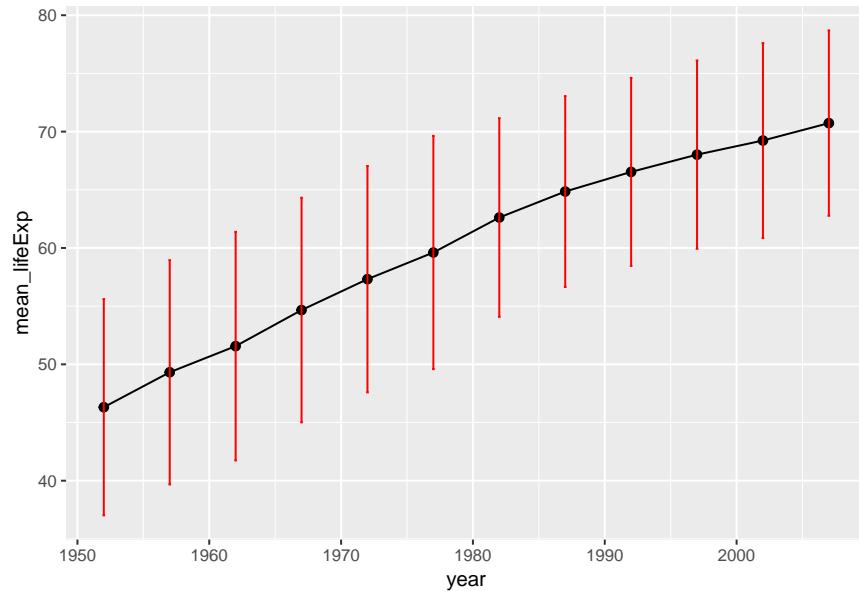


Figure 5.16: Histogram lifeExp

```
ggplot(aes(year, mean_lifeExp))+
  # membuat line plot
  geom_line()+
  # menambahkan point
  geom_point(size=2)+ 
  # menambahkan error bar
  geom_errorbar(aes(ymin=mean_lifeExp-sd,
                     ymax=mean_lifeExp+sd),
                width=0.2, color="red")
```

5.5 Pie Chart

Pie chart pada ggplot2 dapat dibuat menggunakan fungsi `geom_bar()` dan `coord_polar()`. Berikut adalah contoh sintaks yang digunakan dan output (Gambar 5.17) yang dihasilkan:

```
total <- sum(gapminder$pop)
gapminder%>%
  # kelompokkan berdasarkan continent
  group_by(continent)%>%
  # ringkas data
  summarize(pop=sum(as.numeric(pop)), percent=(pop/total)*100)%>%
  ggplot(aes(x="", percent, fill=continent))+ 
  geom_bar(stat="identity")+
  coord_polar("y", start=0)
```

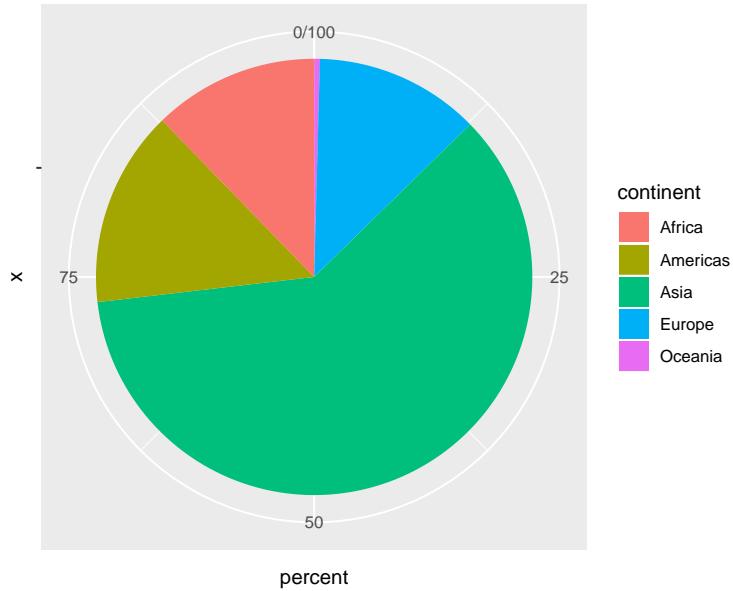


Figure 5.17: Pie chart pop

5.6 Histogram dan Desity Plot

Histogram pada `ggplot2` dapat dibuat dengan fungsi `geom_histogram()`. Berikut adalah sintaks untuk membuat histogram pada variabel `lifeExp`. Output yang dihasilkan disajikan pada Gambar 5.18:

```
gapminder %>%
  ggplot(aes(lifeExp)) +
  geom_histogram()
```

Kita dapat membuat grafik histogram berdasarkan grup data. Pada contoh sebelumnya dibuat histogram berdasarkan variabel `continent`. Berikut adalah sintaks dan output yang dihasilkan pada Gambar 5.19:

```
gapminder %>%
  ggplot(aes(lifeExp, fill=continent)) +
  geom_histogram(alpha=0.5,
                 # atur posisi agar sesuai grup
                 position="identity",
                 color="black")
```

Density plot dapat dibuat dengan menggunakan fungsi `geom_density()`. Berikut adalah contoh sintaks untuk membuat density plot variabel `lifeExp`. Output yang dihasilkan disajikan pada Gambar 5.20:

```
gapminder %>%
  ggplot(aes(lifeExp)) +
  geom_density()
```

Kita juga dapat membuat grafik density berdasarkan grup data. Pada contoh sebelumnya dibuat density plot berdasarkan variabel `continent`. Berikut adalah sintaks dan output yang dihasilkan pada Gambar 5.21:

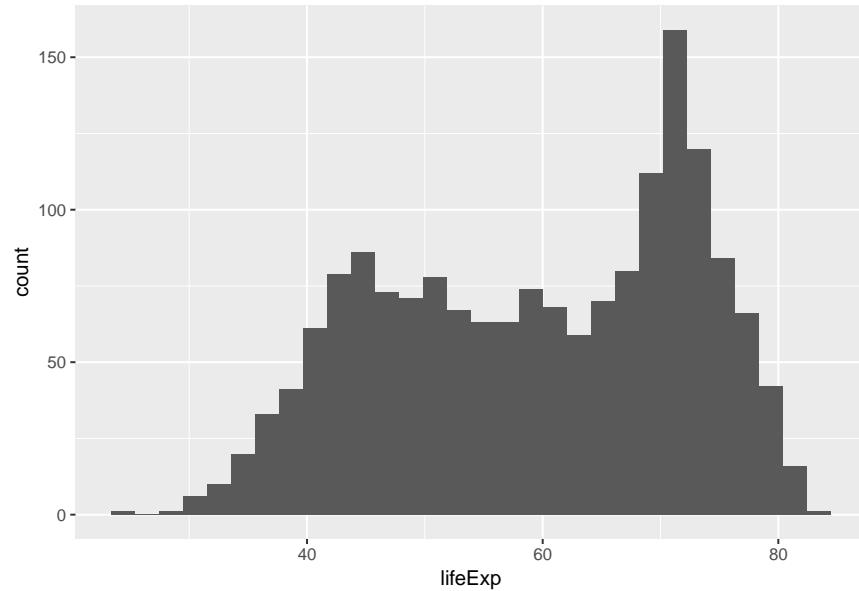


Figure 5.18: Histogram lifeExp

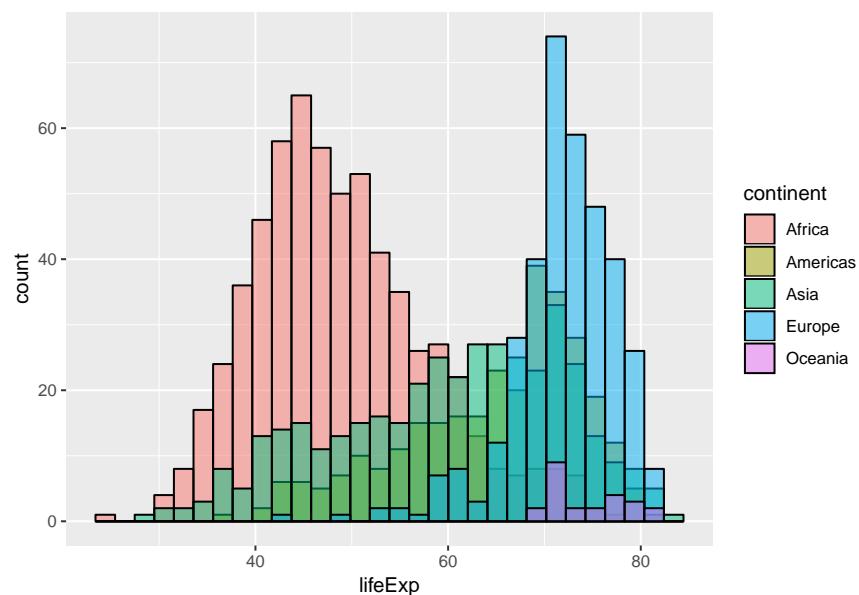


Figure 5.19: Histogram lifeExp berdasarkan benua

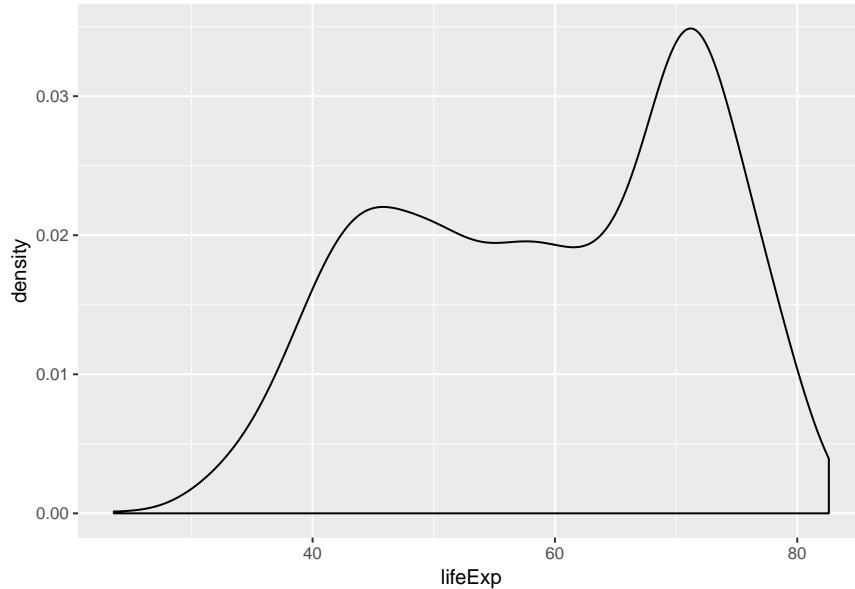


Figure 5.20: Density plot lifeExp

```
gapminder %>%
  ggplot(aes(lifeExp, fill=continent)) +
  geom_density(alpha=0.5,
    # atur posisi agar sesuai grup
    position="identity",
    color="black")
```

Jika dinginkan kita juga dapat menambahkan density plot pada histogram. Pada Gambar 4.20 ditambahkan density plot sehingga dihasilkan output seperti Gambar 5.22.

```
gapminder %>%
  ggplot(aes(lifeExp)) +
  geom_histogram(aes(y=..density..),
    # spesifikasi warna bar
    color="black", fill="white") +
  geom_density(fill="red", alpha=0.3)
```

5.7 QQ Plot

QQ plot pada paket `ggplot2` dapat dibuat dengan menggunakan fungsi `stat_qq()`. Berikut adalah contoh sintaks untuk melakukannya. Output yang dihasilkan disajikna pada Gambar 5.23.

```
ggplot(gapminder, aes(sample=lifeExp)) +
  # qq plot
  stat_qq() +
  # garis referensi
  stat_qq_line()
```

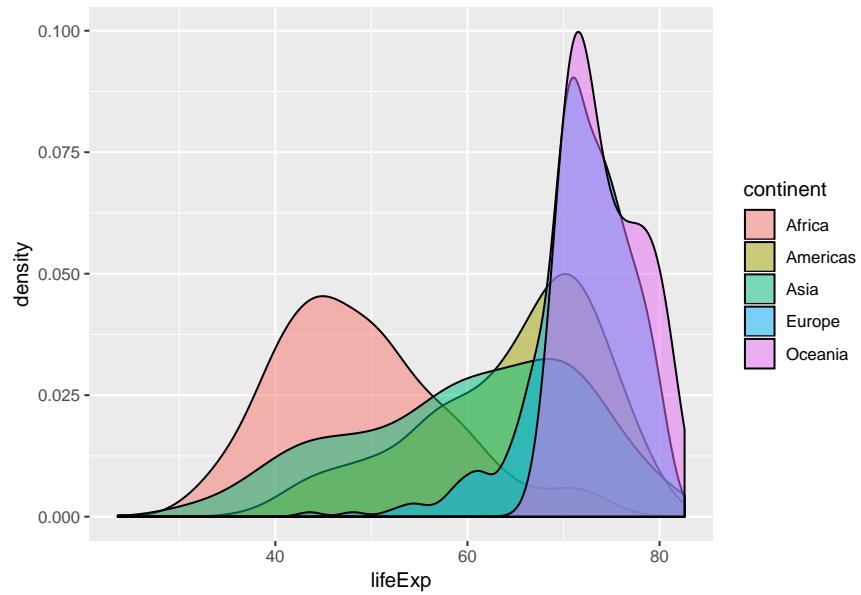


Figure 5.21: Density plot lifeExp berdasarkan benua

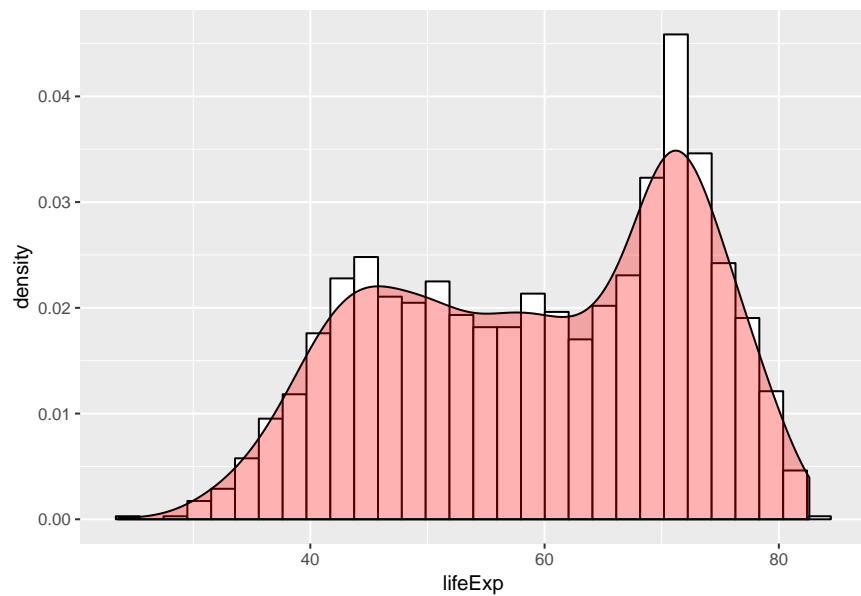


Figure 5.22: histogram dan density plot lifeExp

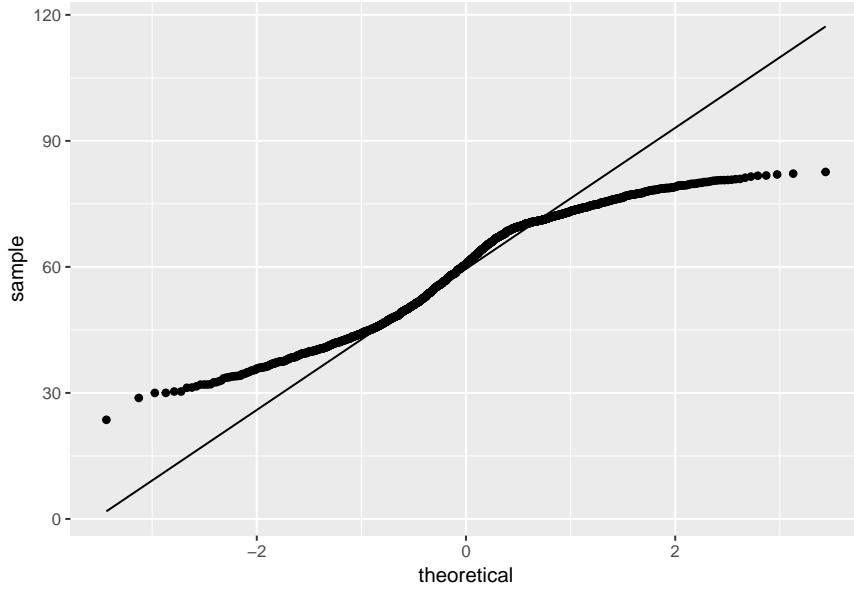


Figure 5.23: QQ plot variabel lifeExp

5.8 Dot Plot

Dot plot dapat dibuat menggunakan fungsi `geom_dotplot` atau `geom_jitter()`. Perbedaan keduanya adalah `geom_jitter()` menambahkan *noise* pada plot sehingga mencegah terjadinya *overplotting*. Berikut adalah contoh sintaks untuk membuat dotplot pada multiple group dan output yang dihasilkan pada Gambar 5.24:

```
gapminder %>%
  filter(year==1952 | year==2007) %>%
  ggplot(aes(continent, lifeExp, fill=factor(year)))+
  geom_dotplot(binaxis="y",
    # spesifikasi posisi plot
    stackdir="center",
    position=position_dodge(0.8),
    size=0.1)
```

```
## Warning: Ignoring unknown parameters: size
```

Kita juga dapat menambahkan plot dari dari plot yang sudah ada seperti box plot atau violin plot. Berikut adalah contoh sintaks dan output yang dihasilkan pada Gambar 5.25:

```
gapminder %>%
  filter(year==1952 | year==2007) %>%
  ggplot(aes(continent, lifeExp, fill=factor(year)))+
  # box plot dibawah
  geom_boxplot(position=position_dodge(0.8))+
```

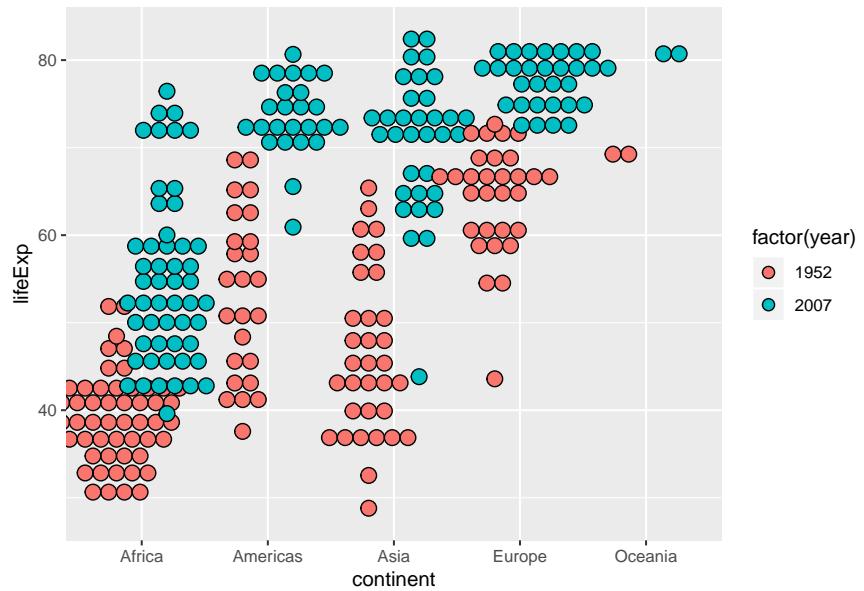


Figure 5.24: Dot plot variabel lifeExp masing-masing benua (1952-2007)

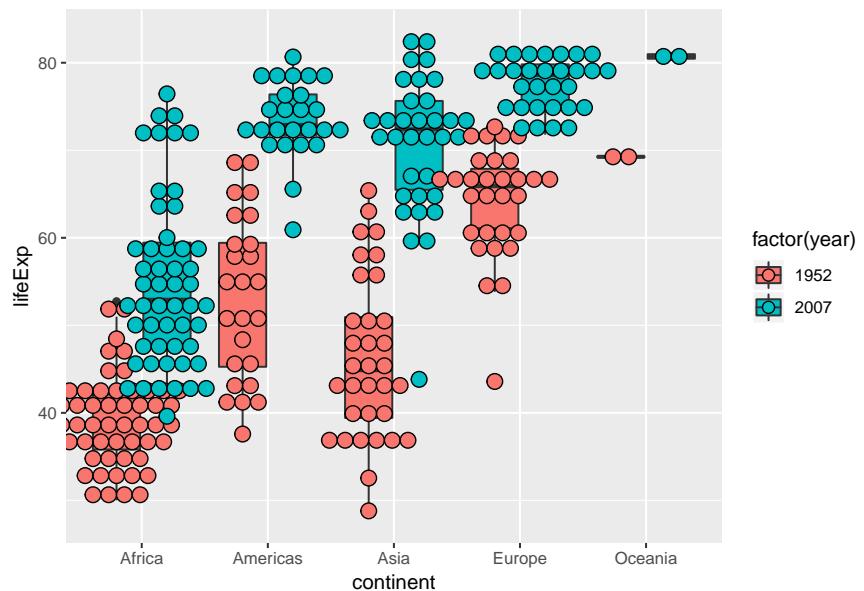


Figure 5.25: Dot plot variabel lifeExp masing-masing benua (1952-2007) (2)

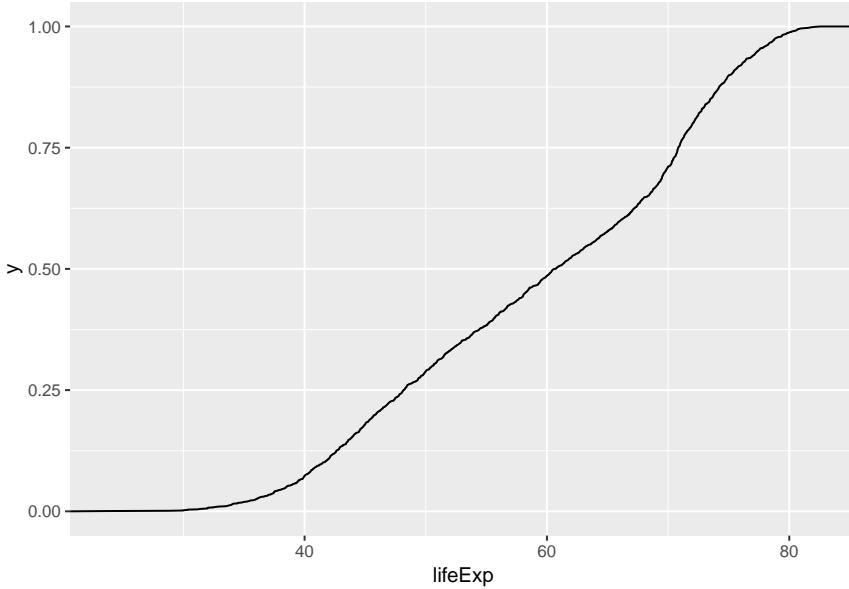


Figure 5.26: ECDF plot variabel lifeExp

5.9 ECDF Plot

Empirical Cumulative Density FUnction (ECDF) plot merupakan grafik yang digunakan untuk menggambarkan distribusi suatu data. Dari grafik ini kita dapat mengetahui fraksi suatu data baik yang terendah maupun yang tertinggi. ECDF pada `ggplot2` dapat dibuat dengan dua cara yaitu dengan `geom_line()` dan `stat_ecdf()`. Jika menggunakan fungsi `geom_line()` kita perlu membuat fraksi kumulatif dari variabel yang akan kita plotkan. Sedangkan dengan menggunakan `stat_ecdf()`, kita tidak perlu melakukannya karena fungsi tersebut akan secara otomatis memproses data kita. Berikut adalah sintaks dan output (Gambar 5.26) contoh ecdf:

```
ggplot(gapminder, aes(lifeExp))+
  stat_ecdf(geom="line")
```

5.10 Parameter Grafik

Pada bagian ini penulis akan menjelaskan bagaimana cara mengatur parameter grafik seperti judul grafik, legend, warna, tema, dll. Pengaturan parameter grafik pada `ggplot2` sebenarnya jauh lebih sederhana dibandingkan dengan fungsi dasar visualisasi R. Selain itu, kita dapat membuat tampilan grafik kita jauh lebih menarik dengan membuat tema kustom pada grafik kita.

5.10.1 Merubah Judul Grafik, Keterangan Axis dan Legend

Untuk merubah judul grafik dan keterangan axis kita dapat melakukannya melalui dua cara. Cara pertama adalah dengan memasukkan mengubahnya satu persatu menggunakan fungsi `ggtitle()` (judul grafik), `xlab()` (keterangan sumbu x), dan `ylab()` (keterangan pada sumbu y). Cara kedua adalah dengan menggunakan fungsi `labs()` dimana selain dapat mengubah judul grafik dan keterangan axis fungsi tersebut dapat juga digunakan untuk mengubah keterangan legend.

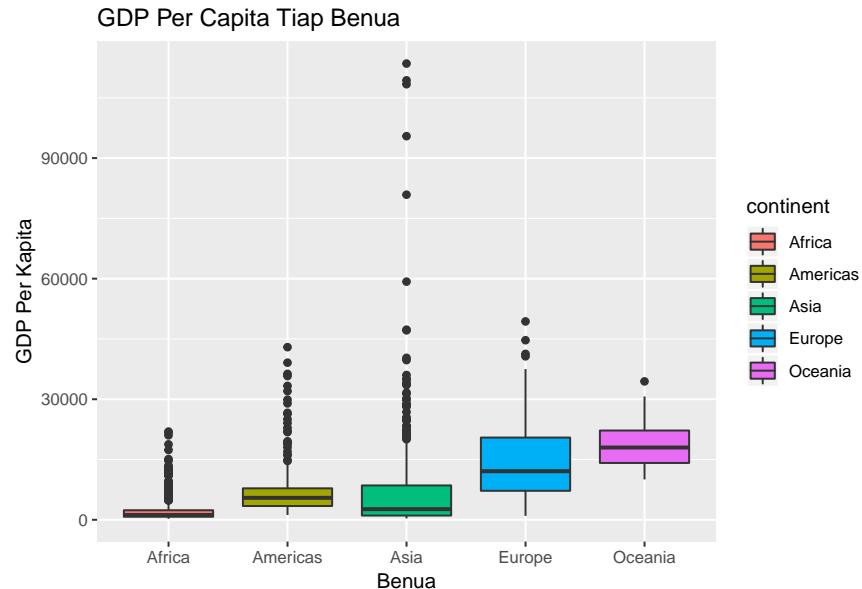


Figure 5.27: Mengubah judul grafik dan keterangan axis

Pada sintaks berikut penulis akan memberikan contoh bagaimana mengubah judul grafik dan keterangan axis menggunakan dua cara tersebut. Output yang dihasilkan disajikan pada Gambar 5.27.

```
# Cara 1
ggplot(gapminder, aes(continent, gdpPercap, fill=continent))+
  # membuat box plot
  geom_boxplot()+
  # menambahkan judul
  ggtitle("GDP Per Capita Tiap Benua")+
  # mengubah keterangan axis
  xlab("Benua")+
  ylab("GDP Per Kapita")

# cara 2
ggplot(gapminder, aes(continent, gdpPercap, fill=continent))+
  # membuat box plot
  geom_boxplot()+
  # kustomisasi judul dan keterangan axis
  labs(title="GDP Per Capita Tiap Benua",
       x="Benua", y="GDP Per Kapita")
```

Pada Gambar 5.27 kita belum mengubah keterangan legend. Berikut adalah sintaks untuk mengubah keterangan legend pada grafik tersebut beserta output yang disajikan pada Gambar 5.28.

```
# cara 2
ggplot(gapminder, aes(continent, gdpPercap,
                      # warna box berdasarkan benua
                      fill=continent))+
  # membuat box plot
  geom_boxplot()+
  # kustomisasi judul dan keterangan axis
```

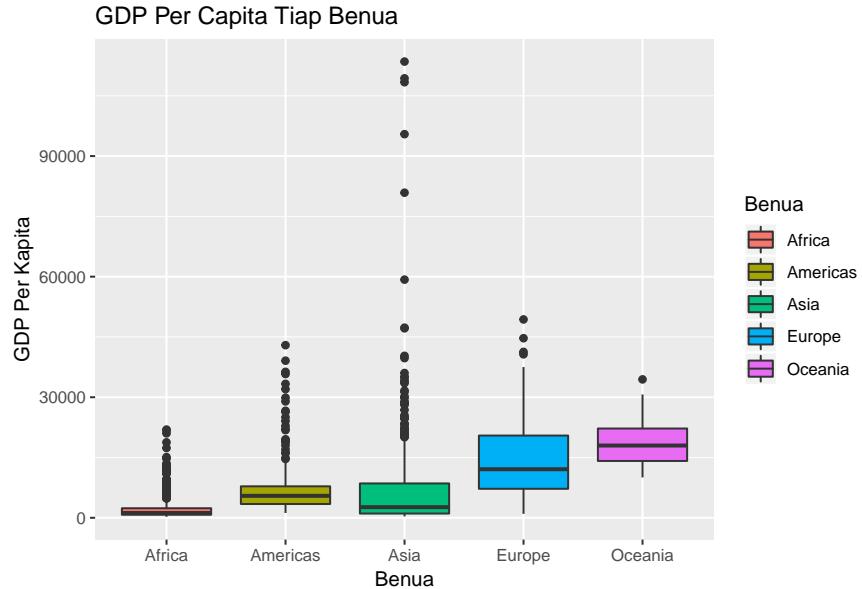


Figure 5.28: Mengubah keterangan legend pada grafik

```
labs(title="GDP Per Capita Tiap Benua",
      x="Benua", y="GDP Per Kapita",
      # mengubah keterangan legend
      fill="Benua")
```

Judul, keterangan axis, dan keterangan legend dapat dikustomisasi menggunakan fungsi `theme()` dan `element_text()`. Berikut adalah format yang digunakan:

```
# Judul
<ggplot> + theme(plot.title = element_text(family, face, colour, size))
# keterangan sumbu x
<ggplot> + theme(axis.title.x = element_text(family, face, colour, size))
# keterangan sumbu y
<ggplot> + theme(axis.title.y = element_text(family, face, colour, size))
# keterangan legend
<ggplot> + theme(axis.title.y = element_text(family, face, colour, size))
```

Note:

- **family**: font family.
- **face**: tampilan font. Nilai yang dapat digunakan antara lain: “plain”, “italic”, “bold” dan “bold.italic”.
- **colour**: warna teks.
- **size**: ukuran teks

Berikut adalah contoh penerapan fungsi tersebut pada grafik Gambar 5.28. Output yang dihasilkan disajikan pada Gambar 5.29.

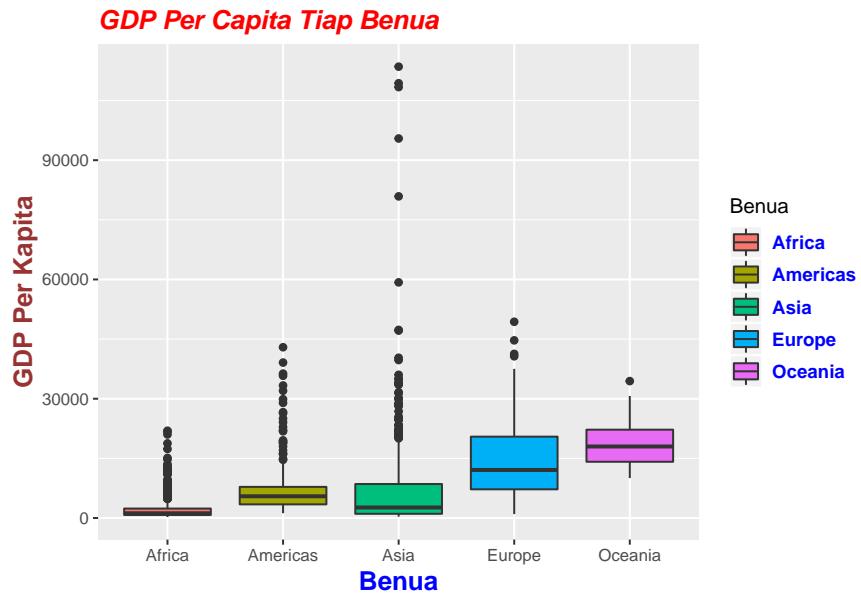


Figure 5.29: Kustomisasi judul grafik dan keterangan axis

```
# cara 2
ggplot(gapminder, aes(continent, gdpPercap,
                      # warna box berdasarkan benua
                      fill=continent))+
  # membuat box plot
  geom_boxplot()+
  # kustomisasi judul dan keterangan axis
  labs(title="GDP Per Capita Tiap Benua",
       x="Benua", y="GDP Per Kapita",
       # mengubah keterangan legend
       fill="Benua")+
  theme(
    plot.title = element_text(color="red", size=14, face="bold.italic"),
    axis.title.x = element_text(color="blue", size=14, face="bold"),
    axis.title.y = element_text(color="#993333", size=14, face="bold"),
    legend.text = element_text(colour="blue", size=10, face="bold")
  )
```

5.10.2 Merubah Tampilan dan Posisi Legend

Posisi legend dapat diubah dengan menambahkan argumen `legend.position` pada fungsi `theme()`. Posisi legend dapat diubah dengan memasukkan nilai berupa karakter seperti “left”, “top”, “right”, dan “bottom”. Selain itu, posisi legend dapat dispesifikasi menggunakan vektor numerik `c(x,Y)`. Nilai x dan y berkisar antara 0 sampai 1. Nilai `c(0,0)` menandakan posisi legend pada bagian kiri bawah dan `c(0,1)` menyatakan kiri atas.

Penggunaan karakter dan vektor numerik akan menghasilkan output posisi legend yang berbeda. Jika menggunakan karakter posisi legend akan diubah diluar bidang plot. Sedangkan vektor numerik akan mengubah posisi legend menjadi ada pada bidang plot. Untuk lebih memahaminya berikut disajikan dua

GDP Per Capita Tiap Benua

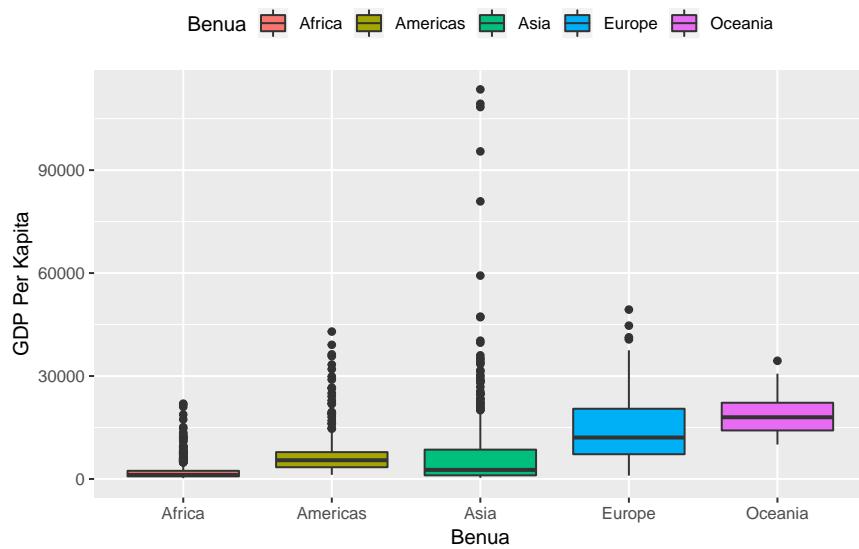


Figure 5.30: Kustomisasi posisi legend berdasarkan karakter

buah gambar. Gambar 5.30 menyajikan pengaturan legend menggunakan karakter, sedangkan Gambar 5.31 menyajikan pengaturan legend menggunakan vektor numerik.

```
# cara 2
ggplot(gapminder, aes(continent, gdpPercap,
                      # warna box berdasarkan benua
                      fill=continent))+
  # membuat box plot
  geom_boxplot()+
  # kustomisasi judul dan keterangan axis
  labs(title="GDP Per Capita Tiap Benua",
       x="Benua", y="GDP Per Kapita",
       # mengubah keterangan legend
       fill="Benua")+
  theme(legend.position="top")
```

```
# cara 2
ggplot(gapminder, aes(continent, gdpPercap,
                      # warna box berdasarkan benua
                      fill=continent))+
  # membuat box plot
  geom_boxplot()+
  # kustomisasi judul dan keterangan axis
  labs(title="GDP Per Capita Tiap Benua",
       x="Benua", y="GDP Per Kapita",
       # mengubah keterangan legend
       fill="Benua")+
  theme(legend.position=c(0.9,0.75))
```

Pada fungsi `theme()` kita juga dapat merubah background dari legend box menggunakan argumen `legend.background` dan `element_rect`. Selain itu kita juga dapat mengubah orientasi dari legend yang

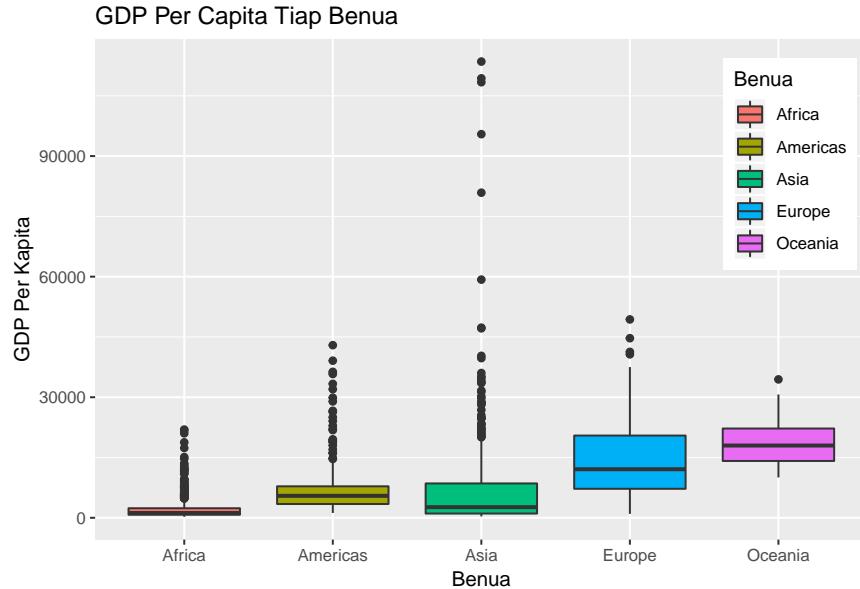


Figure 5.31: Kustomisasi posisi legend berdasarkan vektor numerik

semula vertikal menjadi horizontal dengan menambahkan argumen `legend.box`. Berikut adalah contoh sintaks penerapannya. Output yang dihasilkan disajikan pada Gambar 5.32.

```
# cara 2
ggplot(gapminder, aes(continent, gdpPercap,
# warna box berdasarkan benua
fill=continent,
# warna outline berdasarkan benua
color=continent))+

# membuat box plot
geom_boxplot()+
# kustomisasi judul dan keterangan axis
labs(title="GDP Per Capita Tiap Benua",
x="Benua", y="GDP Per Kapita",
# mengubah keterangan legend
fill="Benua (fill)",
color="Benua (outline)")+
theme(legend.position="bottom",
# mengubah tampilan legend box
legend.background = element_rect(fill="lightblue",
size=0.5, linetype="solid",
colour ="darkblue"),
# mengubah orientasi legend
legend.box= "horizontal")
```

Kita dapat juga menghilangkan legend baik seluruh legend maupun legend spesifik. Pada Gambar 5.33 dan Gambar 5.34 disajikan contoh cara menghilangkan seluruh legend maupun sebagian legend.

```
# Menghilangkan seluruh legend
ggplot(gapminder, aes(continent, gdpPercap,
# warna box berdasarkan benua
```

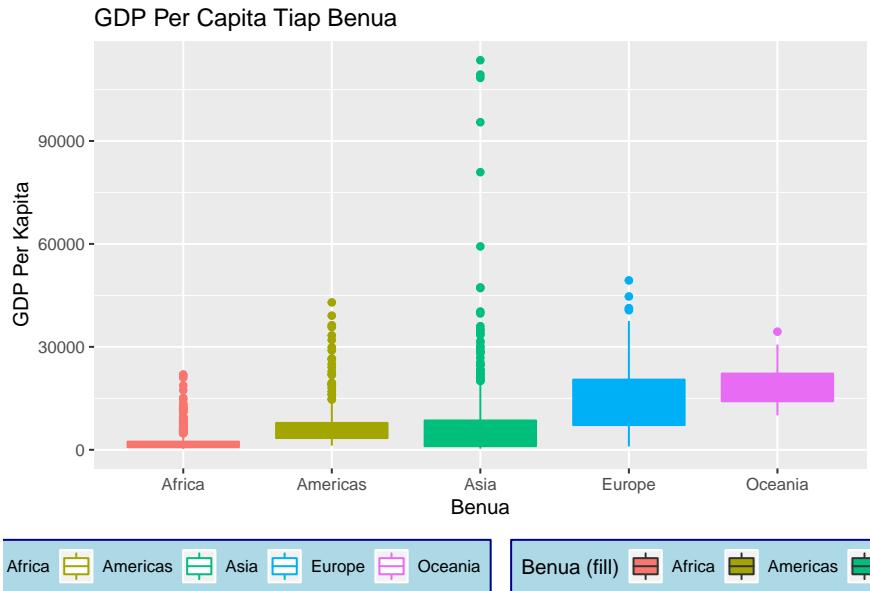


Figure 5.32: Kustomisasi tampilan legend

```
fill=continent,
# warna outline berdasarkan benua
color=continent))+

# membuat box plot
geom_boxplot()+
# kustomisasi judul dan keterangan axis
labs(title="GDP Per Capita Tiap Benua",
x="Benua", y="GDP Per Kapita",
# mengubah keterangan legend
fill="Benua")+
theme(legend.position="none")

# Menghilangkan seluruh legend
ggplot(gapminder, aes(continent, gdpPercap,
# warna box berdasarkan benua
fill=continent,
# warna outline berdasarkan benua
color=continent))+

# membuat box plot
geom_boxplot()+
# kustomisasi judul dan keterangan axis
labs(title="GDP Per Capita Tiap Benua",
x="Benua", y="GDP Per Kapita",
# mengubah keterangan legend
fill="Benua (fill)",
color="Benua (outline)")+
theme(legend.position="bottom",
# mengubah tampilan legend box
legend.background = element_rect(fill="lightblue",
size=0.5, linetype="solid"),
```

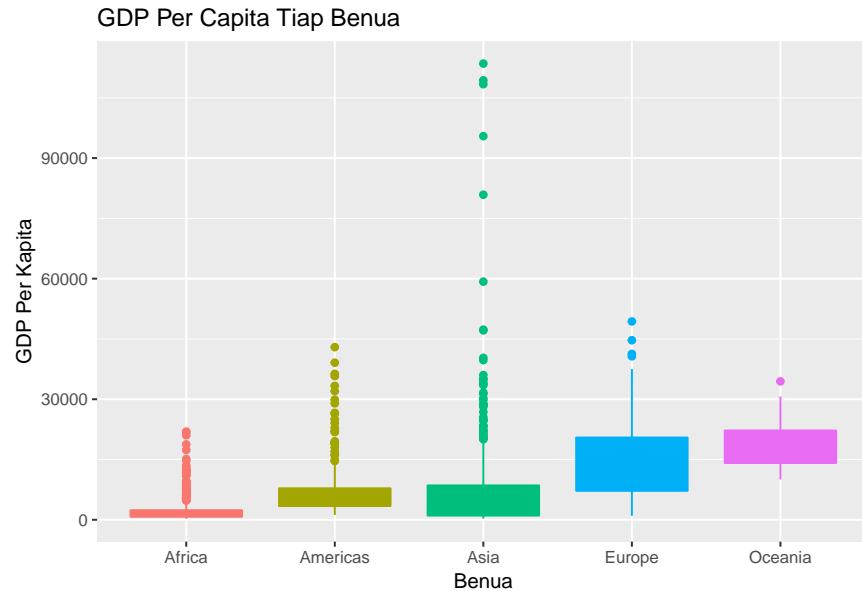


Figure 5.33: Menghilangkan seluruh legend

```
colour = "darkblue"))+
# Menghilangkan legend Benua (outline)
guides(color=FALSE)
```

5.10.3 Merubah Warana Pada Grafik Secara Otomatis dan Manual

Kita dapat merubah warna grafik baik secara otomatis dan manual. Secara otomatis warna dapat diubah dengan memasukkan nama variabel kedalam argumen `fill` dan `color`. Namun, jika kita inginkan kita dapat memasukkan kode warna untuk memperoleh warna yang seragam pada seluruh kelompok data.

Pada contoh sintaks berikut diberikan contoh bagaimana merubah warna pada seluruh grup data dengan satu warna yang seragam. Output yang dihasilkan disajikan pada Gambar 5.35:

```
ggplot(gapminder, aes(continent, lifeExp))+
# spesifikasi warna tunggal
geom_boxplot(color="darkred", fill="#A4A4A4")
```

Selain itu, kita dapat mengubah warna berdasarkan grup baik secara otomatis maupun manual. Berikut adalah contoh sintaks warna berdasarkan grup secara otomatis. Output yang dihasilkan disajikan pada Gambar 5.36.

```
ggplot(gapminder, aes(continent, gdpPercap,
# warna berdasarkan grup
fill=continent))+
geom_boxplot()
```

Kita dapat mengatur pecahayaan (l) dan intensitas warna (c) dari warna yang kita tampilkan menggunakan fungsi `scale_fill_hue()`. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.37.

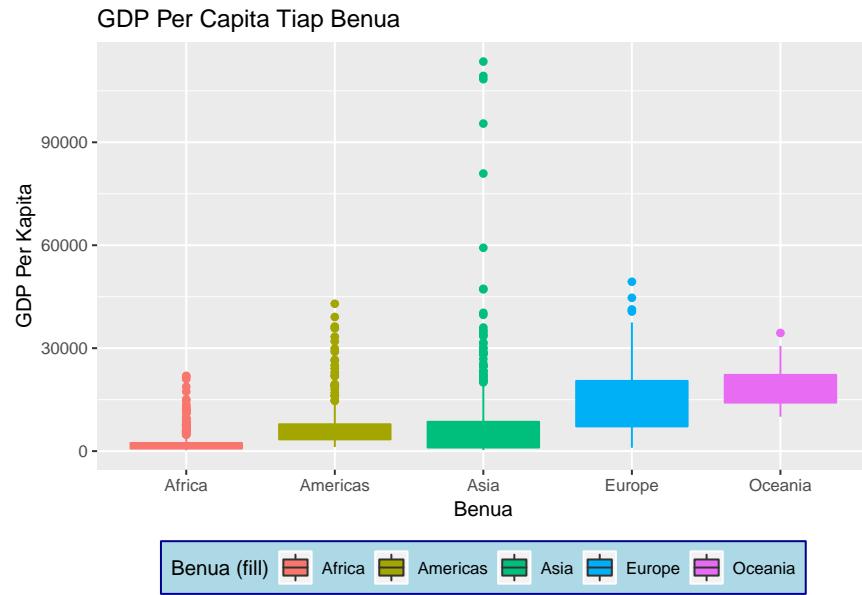


Figure 5.34: Menghilangkan sebagian legend legend

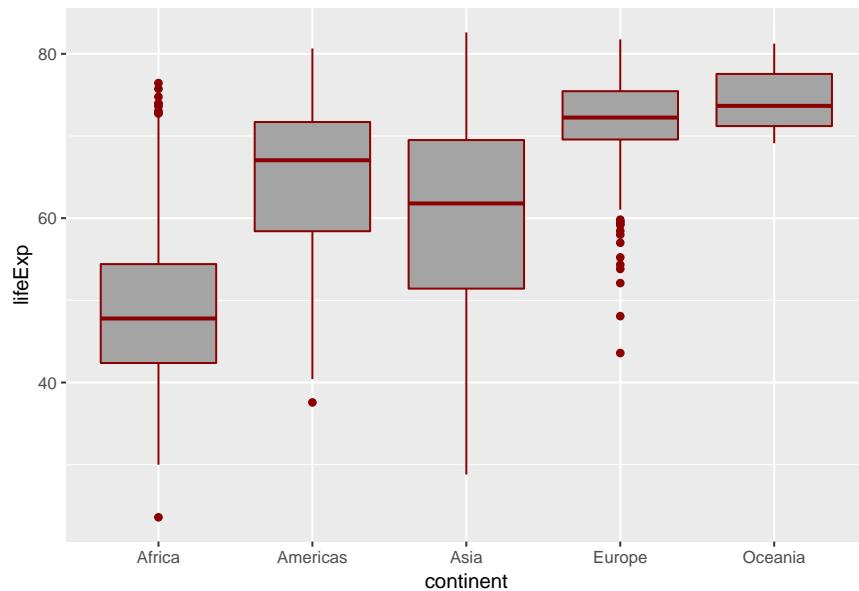


Figure 5.35: Merubah warna grup berdasarkan satu warna

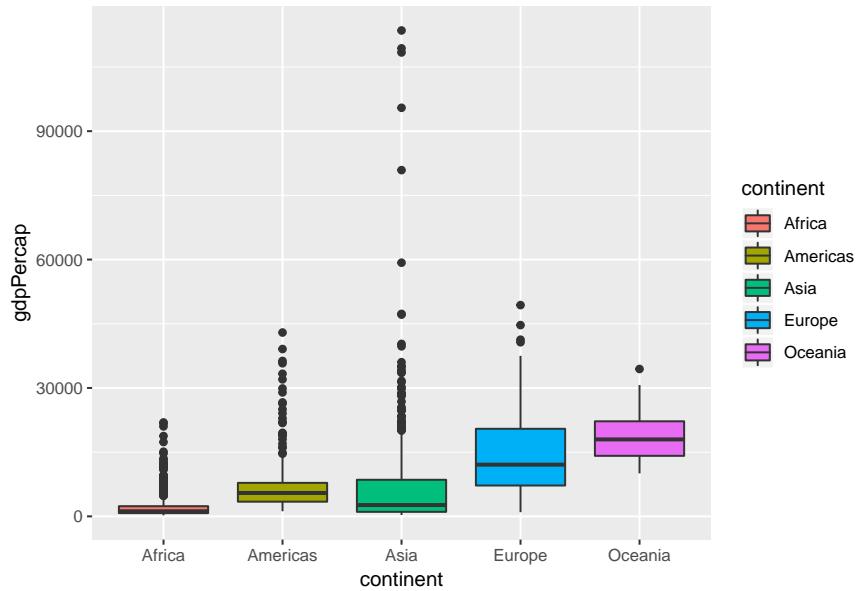


Figure 5.36: Merubah warna grup secara otomatis

```
ggplot(gapminder, aes(continent, gdpPercap,
# warna berdasarkan grup
fill=continent))+  
geom_boxplot()+
# merubah l dan c
scale_color_hue(l=40, c=35)
```

Jika kita tidak menginginkan warna yang secara otomatis ditampilkan oleh `ggplot2`, kita dapat mengubahnya secara manual menggunakan fungsi `scale_fill_manual()` (untuk box plot, bar plot, dll) dan `scale_color_manual()` (untuk line plot, dot plot dan scatterplot). Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.38.

```
ggplot(gapminder, aes(continent, gdpPercap,
# warna berdasarkan grup
fill=continent))+  
geom_boxplot()+
# merubah warna secara manual
scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9",
"#B47846", "#B4464B"))
```

Jika kita tidak hafal dengan kode hexadesimal warna tersebut kita dapat juga menggunakan palet warna. Contoh palet warna yang akan digunakan adalah dari library `RColorBrewer`. Berikut adalah contoh sintaks untuk menginstal dan memuat paket tersebut:

```
# memasang paket
# install.packages("RColorBrewer")

# memuat paket
library(RColorBrewer)
```

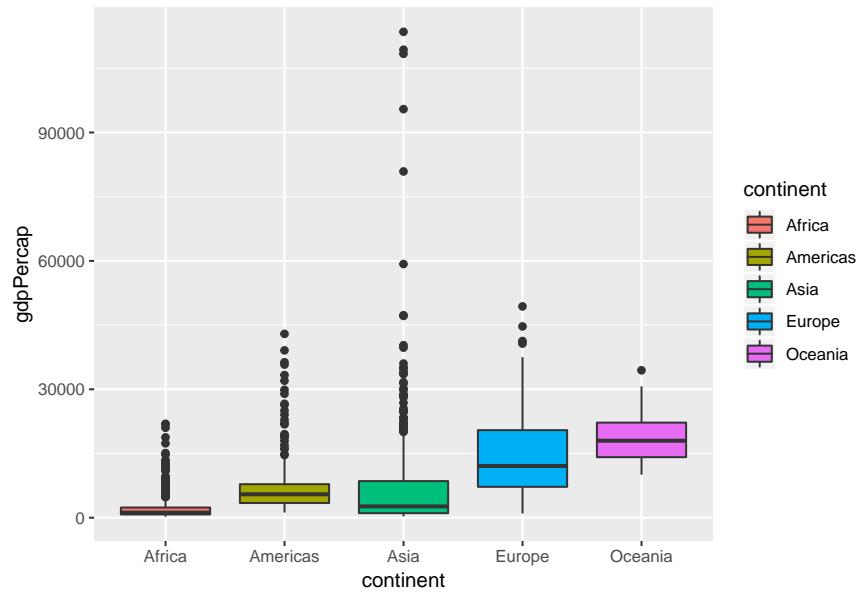


Figure 5.37: Merubah pencahayaan dan intensitas warna

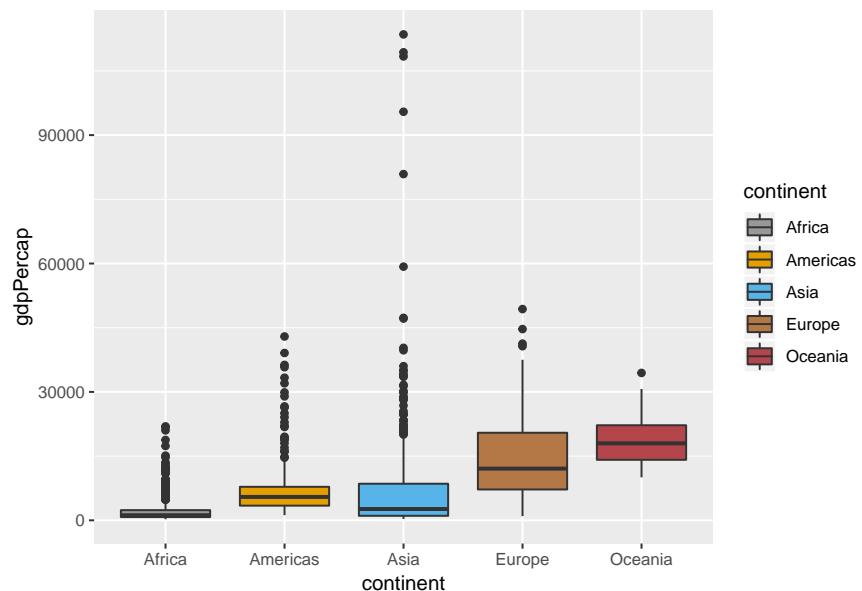


Figure 5.38: Merubah warna secara manual

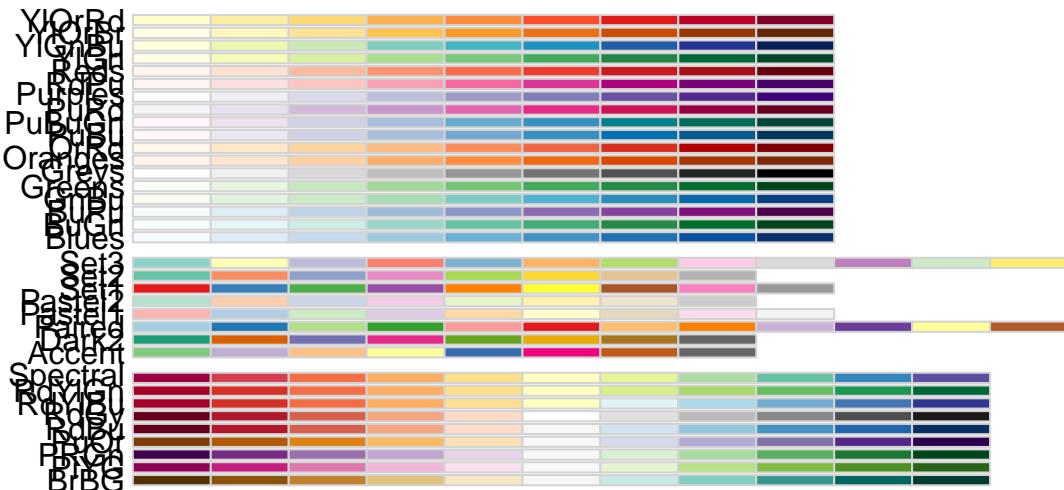


Figure 5.39: Palet warna RColorBrewer

Pada sintak berikut penulis akan menampilkan seluruh palet warna pada pekt tersebut. Output yang dihasilkan disajikan pada Gambar 5.39.

```
display.brewer.all()
```

Pada Gambar 5.39 terdapat 3 jenis warna antara lain:

- Sequential palettes**, digunakan untuk menunjukkan urutan dari rendah ke tinggi atau gradien. Nama palet yang ada antara lain: Blues, BuGn, BuPu, GnBu, Greens, Greys, Oranges, OrRd, PuBu, PuBuGn, Purples, RdPu, Reds, YlGn, YlGnBu, YlOrBr, dan YlOrRd.
- Diverging palettes**, digunakan untuk menunjukkan perubahan pada data yang memiliki nilai positif dan negatif. Palet yang tersedia antara lain: BrBG, PiYG, PRGn, PuOr, RdBu, RdGy, RdYlBu, RdYlGn, dan Spectral.
- Qualitative palettes**, digunakan untuk merepresentasikan variabel nominal atau kategori karena tidak menunjukkan besaran atau perbedaan nilai antar grup. Palete yang tersedia antara lain: Accent, Dark2, Paired, Pastel1, Pastel2, Set1, Set2, dan Set3.

Pada contoh sintaks berikut disajikan contoh penerapan dan output yang dihasilkan pada Gambar 5.40.

```
ggplot(gapminder, aes(continent, gdpPercap,
# warna berdasarkan grup
fill=continent)) +
```

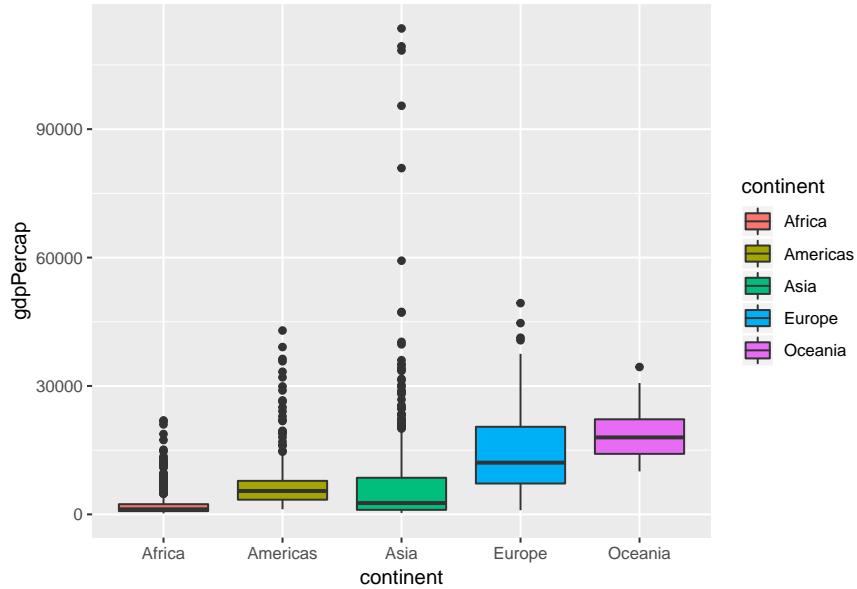


Figure 5.40: Merubah warna menggunakan palet

```
geom_boxplot()+
# merubah warna menggunakan palet
scale_color_brewer(palette="Dark2")
```

Jika kita tidak menginginkan warna-warna terang, kita dapat menggunakan fungsi `scale_color_grey()` (untuk line plot, dot plot, dan scatterplot) dan `scale_fill_grey()` (untuk bar plot, histogram, box plot, dll). Fungsi tersebut akan memberikan warna palet gray pada plot. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.41.

```
ggplot(gapminder, aes(continent, gdpPercap,
# warna berdasarkan grup
fill=continent))+
geom_boxplot()+
# merubah warna menggunakan palet
scale_fill_grey()
```

5.10.4 Kustomisasi Titik

Untuk mengubah jenis titik pada scatterplot, outlier pada box plot, dan dot plot, kita dapat menambahkan argumen `shape` pada fungsi geometrinya. Nilai yang mungkin dimasukkan berupa nilai diskrit yang berkisar antara 0 sampai 25. Selain itu, ukuran dari titik dapat diinput dengan menambahkan argumen `size`. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.42.

```
ggplot(gapminder, aes(gdpPercap, lifeExp))+
# spesifikasi jenis, ukuran dan warna titik
geom_point(shape=4, size=2, color="blue")
```

Untuk data dengan multiple group, kita dapat mengubah jenis, ukuran dan warna secara otomatis dengan memasukkan nama variabel kedalam argumen `shape`, `size` dan `color`. Sedangkan secara manual kita

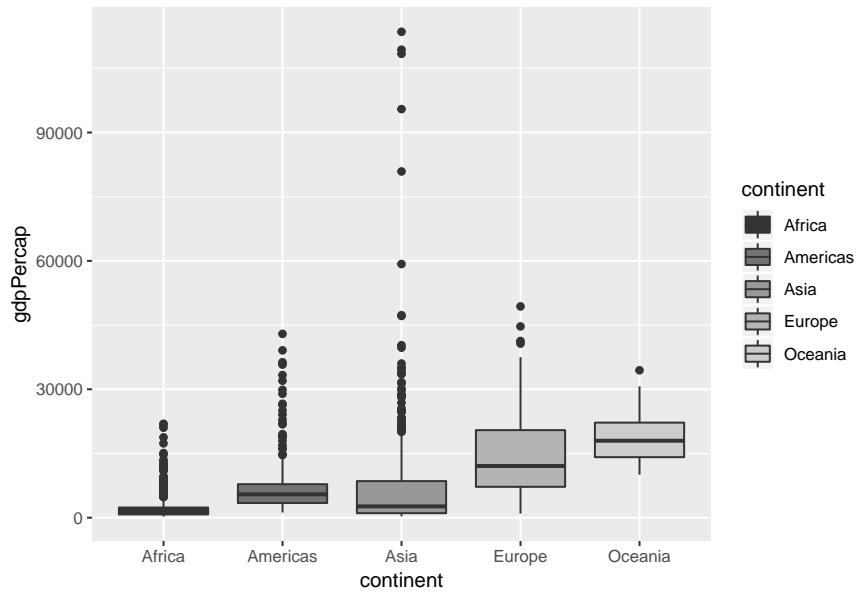


Figure 5.41: Merubah warna menggunakan palet gray

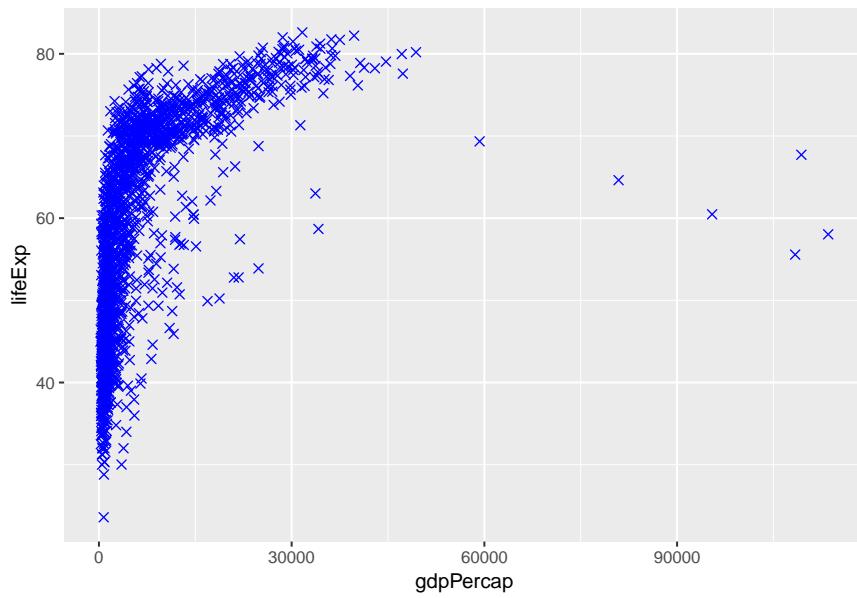


Figure 5.42: Kustomisasi jenis, ukuran dan warna titik

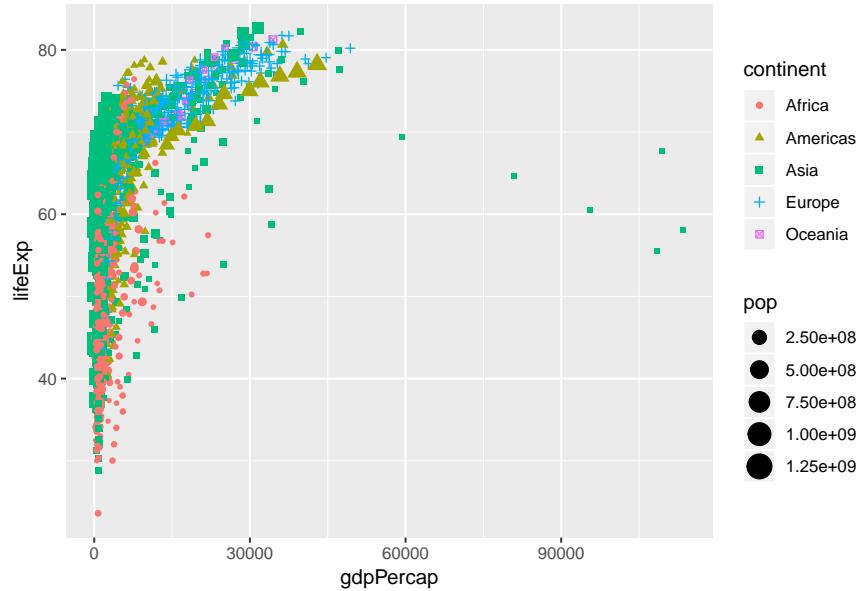


Figure 5.43: Kustomisasi jenis, ukuran dan warna titik untuk multiple group secara otomatis

dapat menambahkan fungsi `scale_shape_manual()` (jenis titik), `scale_color_manual()` (warna titik), dan `scale_size_manual()` (ukuran titik). Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.43 dan Gambar 5.44.

```
# cara otomatis
ggplot(gapminder, aes(gdpPercap, lifeExp,
# spesifikasi jenis, ukuran dan warna
shape=continent, color=continent,
size=pop))+

geom_point()

# cara manual
ggplot(gapminder, aes(gdpPercap, lifeExp,
# spesifikasi jenis, ukuran dan warna
shape=continent, color=continent,
size=pop))+

geom_point()+
scale_shape_manual(values=c(1:5))+
scale_color_manual(values=c("#999999", "#E69F00", "#56B4E9",
"#B47846", "#B4464B"))
```

5.10.5 Kustomisasi Jenis Garis

Jenis, warna dan ukuran garis dapat diatur dengan menambahkan argumen `linetype`, `size` dan `color`. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.45.

```
gapminder%>%
  filter(continent=="Asia")%>%
  group_by(year)%>%
```

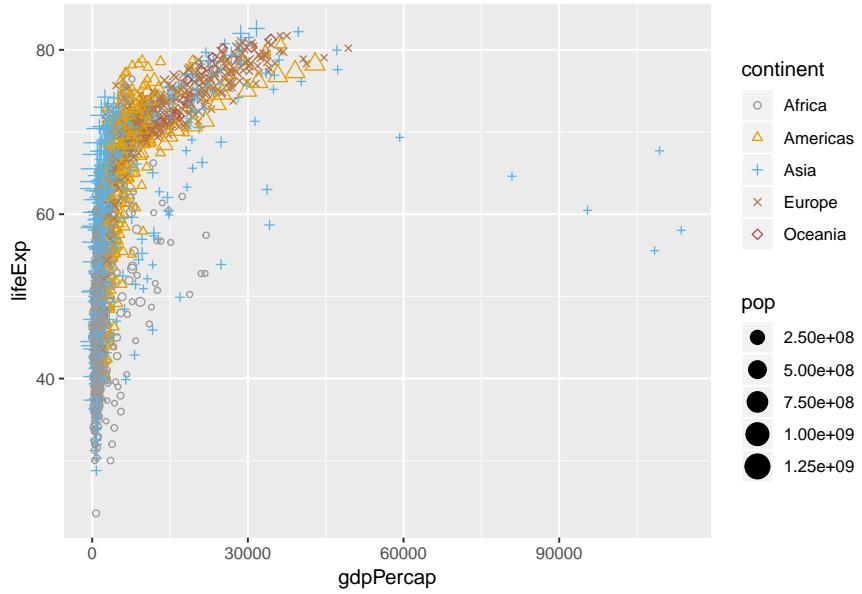


Figure 5.44: Kustomisasi jenis, ukuran dan warna titik untuk multiple group secara manual

```
summarize(mean_pop=mean(pop))%>%
# plot
ggplot(aes(year, mean_pop))++
  geom_line(linetype="dashed", color="blue",
            size=1)++
  geom_point(shape=1, color="red")
```

Untuk data dengan multiple group, kita dapat mengubah jenis garis, warna dan ukuran secara manual maupun secara otomatis. Secara otomatis kita dapat menginputkan nama variabel kedalam argumen `linetype`, `size` dan `color`. Secara manual, kita dapat mengubah jenis, warna dan ukuran menggunakan fungsi `scale_linetype_manual()` (jenis garis), `scale_color_manual()` (warna garis), dan `scale_size_manual()` (ukuran garis). Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.46 dan Gambar 5.47.

```
# cara otomatis
gapminder%>%
  filter(continent %in% c("Asia", "Africa"))%>%
  group_by(year, continent)%>%
  summarize(mean_pop=mean(pop))%>%
# plot
  ggplot(aes(year, mean_pop,
             linetype=continent,
             color=continent))++
  geom_line()
  geom_point(shape=1, color="red")
```

```
# cara manual
gapminder%>%
  filter(continent %in% c("Asia", "Africa"))%>%
  group_by(year, continent)%>%
```

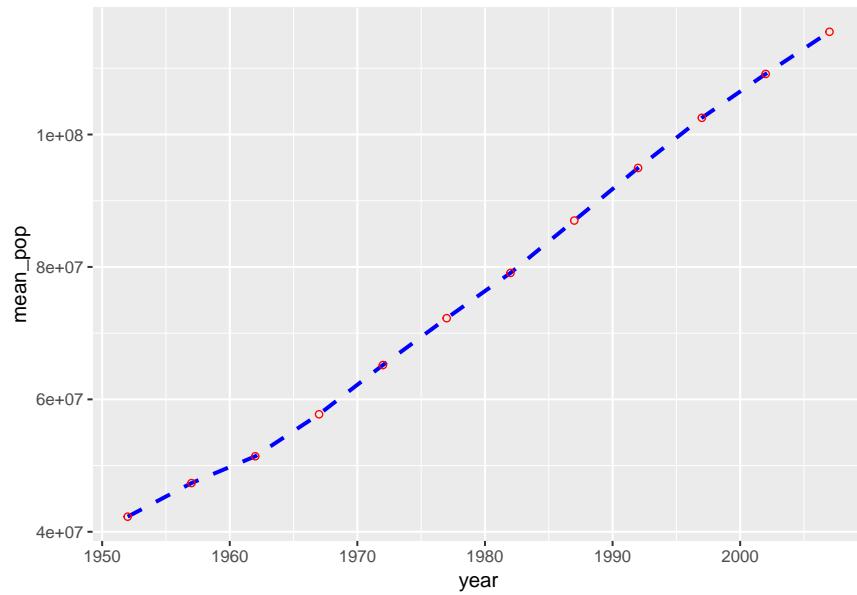


Figure 5.45: Kustomisasi jenis, ukuran dan warna garis

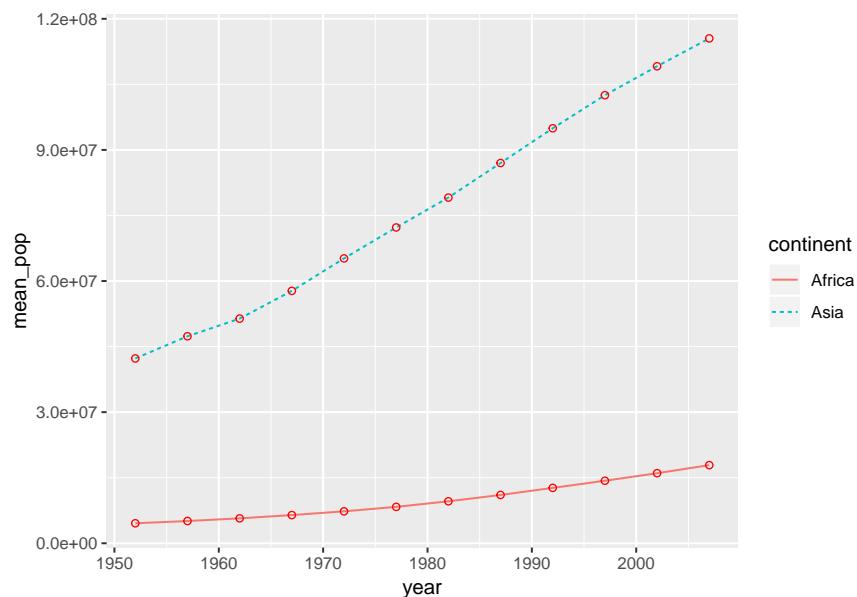


Figure 5.46: Kustomisasi jenis, ukuran dan warna garis untuk multiple group secara otomatis

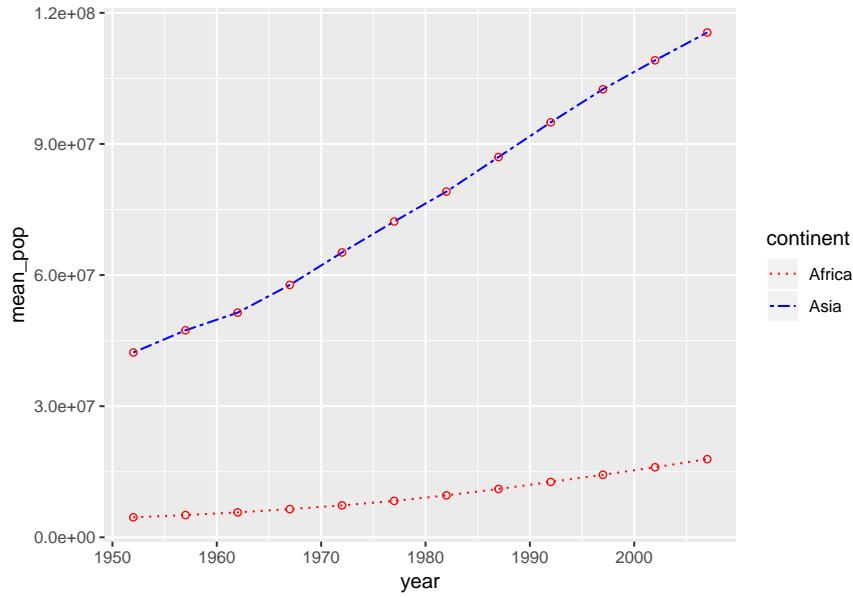


Figure 5.47: Kustomisasi jenis, ukuran dan warna garis untuk multiple group secara manual

```
summarize(mean_pop=mean(pop))%>%
# plot
ggplot(aes(year, mean_pop,
           linetype=continent,
           color=continent))+ 
  geom_line()+
  geom_point(shape=1, color="red")+
  scale_linetype_manual(values=c("dotted", "twodash"))+
  scale_color_manual(values=c("red","blue"))
```

5.10.6 Menambahkan Label Pada Titik Observasi dan Bidang Plot

Pada artikel ini penulis akan menjelaskan bagaimana kita dapat menambahkan teks pada plot. Fungsi-fungsi yang dapat digunakan antara lain:

- `geom_text()`: menambahkan teks secara langsung pada plot.
- `geom_label()`: menambahkan teks dengan kotak disekelilingnya.
- `annotate()`: menambahkan teks tertentu pada bagian tertentu bidang plot.
- `annotation_custom()`: menambahkan anotasi statik yang sama pada setiap panel.

Misal kita akan membuat plot antara variabel pop vs gdpPercap seperti yang ditunjukkan pada Gambar 5.48 berikut:

```
ggplot(gapminder, aes(gdpPercap, pop))+
  geom_point()
```

Misalkan kita ingin menandai negara yang memiliki $\text{gdpPercap} > 50000$. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.49.

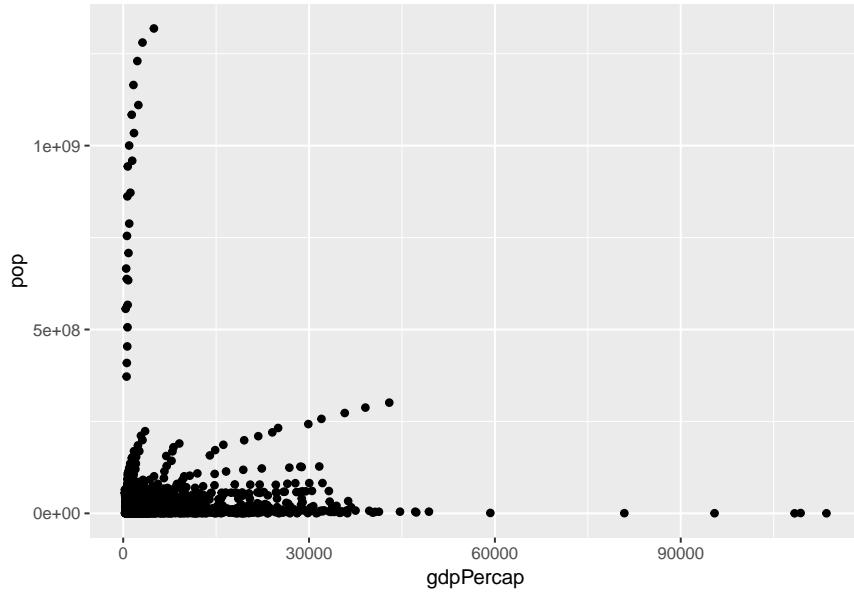


Figure 5.48: Scatterplot variabel pop vs gdpPercap

```
ggplot(gapminder, aes(gdpPercap, pop))+
  geom_point(shape=1)+
  geom_label(
    # subset data sesua kriteria
    data=subset(gapminder,gdpPercap>50000),
    # label berdasarkan kriteria
    aes(label=country),
    # ukuran teks
    size = 3)
```

Selain teks yang menunjukkan observasi, kita dapat menambahkan anotasi pada grafik. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.50.

```
ggplot(gapminder, aes(gdpPercap, pop))+
  geom_point(shape=1)+
  # menambahkan label sesuai kriteria data
  geom_label(
    # subset data sesua kriteria
    data=subset(gapminder,gdpPercap>50000),
    # label berdasarkan kriteria
    aes(label=country),
    # ukuran teks
    size = 3)+
  annotate(geom="text", x=90000,
          y=2e+08, label="outlier",
          color="red")
```

Kita dapat pula menambahkan teks statik yang sama pada setiap panel. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.51.

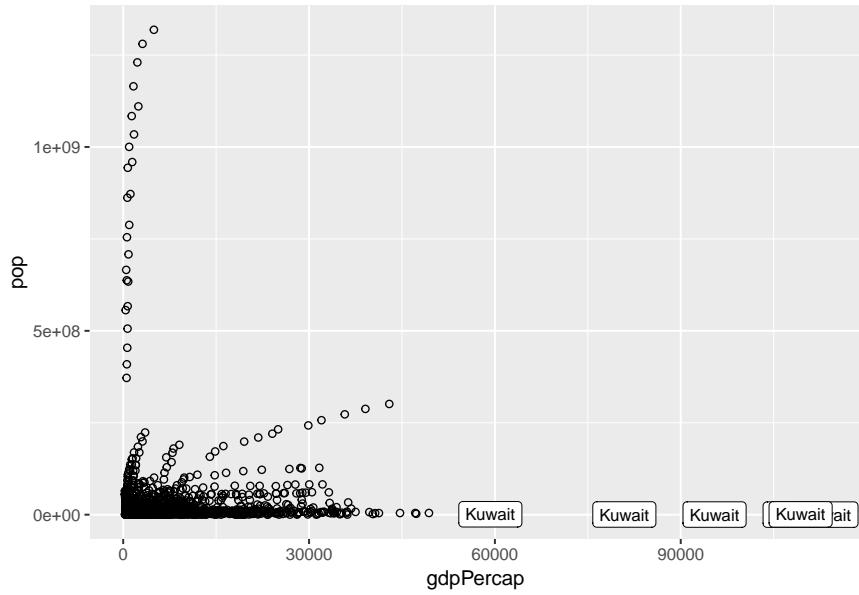


Figure 5.49: Scatterplot variabel pop vs gdpPercap dengan label

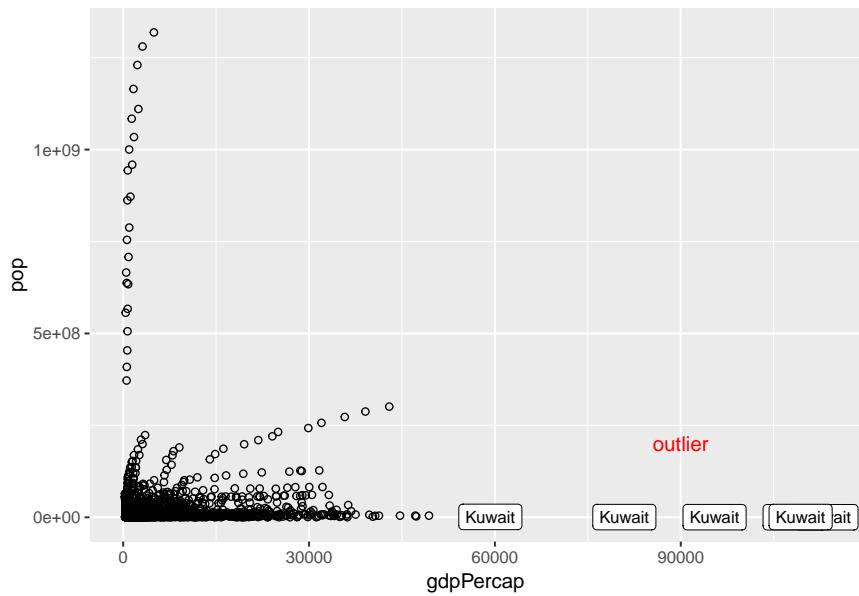


Figure 5.50: Scatterplot variabel pop vs gdpPercap dengan label dan notasi

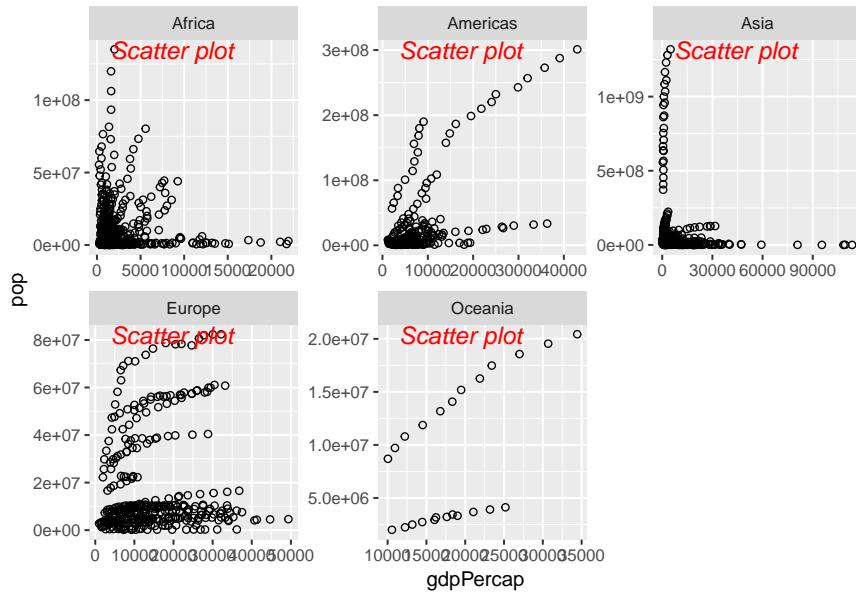


Figure 5.51: Scatterplot variabel pop vs gdpPercap dengan label dan notasi pada tiap panel

```
library(grid)

# membuat teks
d <- grob <- grobTree(textGrob("Scatter plot", x=0.1, y=0.95, hjust=0,
  gp=gpar(col="red", fontsize=13, fontface="italic")))

# plot
ggplot(gapminder, aes(gdpPercap, pop))+
  geom_point(shape=1)+ # menambahkan anotasi
  annotation_custom(d)+ # membagi plot menjadi beberapa panel
  facet_wrap(~continent, scales="free")
```

5.10.7 Kustomisasi Tema Pada Plot

Kita dapat melakukan kustomisasi tema plot untuk membuat tampilan plot kita lebih menarik. Pada bagian ini penulis akan membahas tema yang dapat digunakan serta cara untuk melakukan edit terhadap tema yang telah ada sebelumnya.

Tema-tema yang telah terpasang secara default pada paket `ggplot2` antara lain:

- `theme_gray`: background dengan warna abu-abu dengan garis grid putih.
- `theme_bw`: background putih dan garis grid berwarna abu-abu.
- `theme_linedraw`: garis hitam di sekeliling bidang plot.
- `theme_light`: garis grid dan axis berwarna abu-abu terang.
- `theme_minimal`: tidak memiliki frame disekeliling bidang plot.
- `theme_classic`: tidak ada garis grid dan axis.
- `theme_void`: tema kosong.
- `theme_dark`: background gelap.

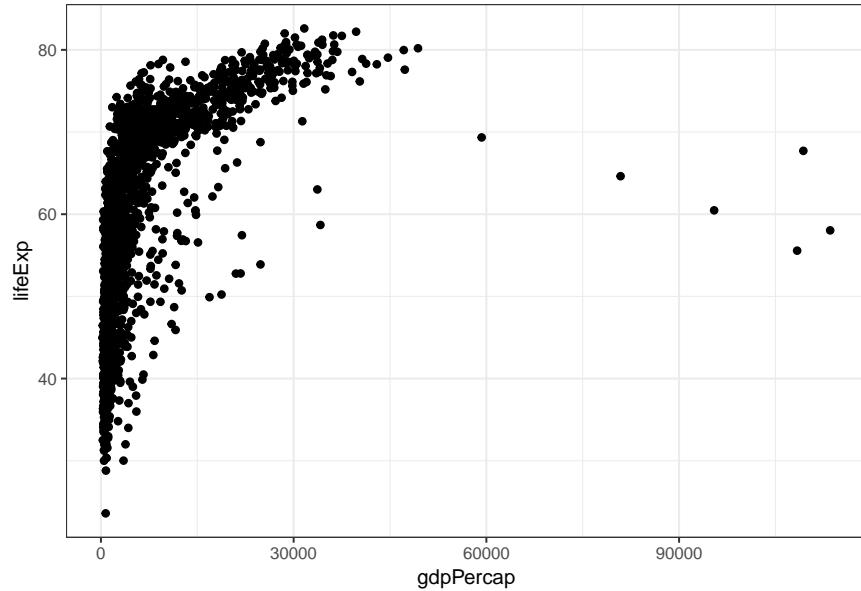


Figure 5.52: Scatterplot dengan tema black and white

Pada contoh berikut disajikan sebagian contoh penerapan tema pada plot. Output yang dihasilkan pada Gambar 5.52.

```
ggplot(gapminder, aes(gdpPercap, lifeExp)) +
  geom_point() +
  theme_bw()
```

Kita juga dapat menggunakan tema kustom yang terdapat pada library `ggthemes`. Berikut adalah sintaks yang digunakan untuk menginstall dan memuat paket tersebut:

```
# Memasang paket
install.packages("ggthemes")
```

```
# memuat paket
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version
## 3.5.3
```

tema-tema yang tersedia pada paket tersebut antara lain:

- **theme_tufte**: tema minimalis.
- **theme_economist**: tema yang digunakan pada majalah Economist.
- **theme_stata**: tema yang digunakan pada visualisasi progra stata.
- **theme_wsj**: tema yang digunakan pada Wall Street Journal.
- **theme_cal**: tema yang digunakan pada LibreOffice Calc dan Google Docs.
- **theme_hc**: tema yang didasarkan pada Highcharts JS.

Pada contoh berikut disajikan sebagian contoh penerapan tema pada plot. Output yang dihasilkan pada Gambar 5.53.

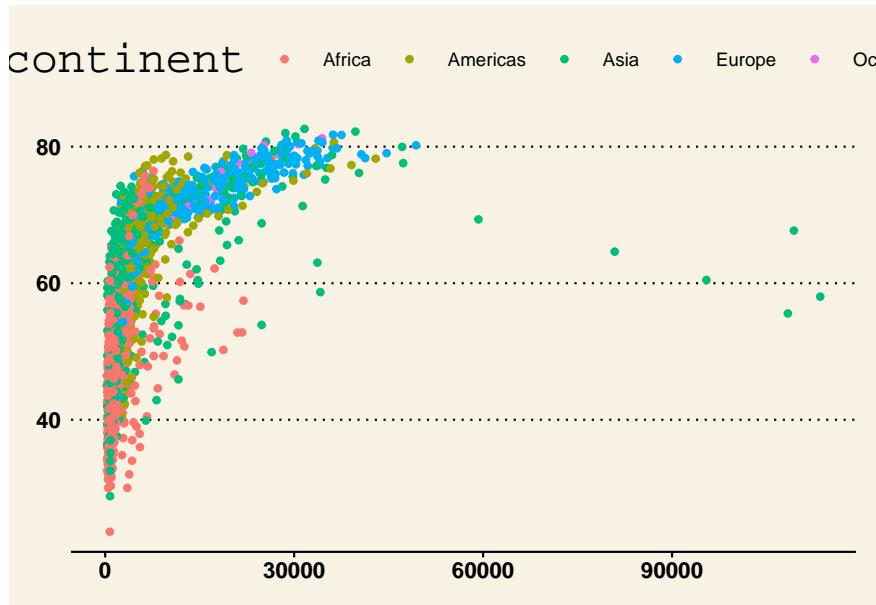


Figure 5.53: Scatterplot dengan tema Wall Street Journal

```
ggplot(gapminder, aes(gdpPercap, lifeExp,
                      color=continent)) +
  geom_point() +
  theme_wsj()
```

Kita dapat juga membuat tema kustom berdasarkan tema yang telah ada. Untuk melakukannya kita hanya perlu merubah sejumlah argument default yang ada pada fungsi tema dan menamai tema sesuai dengan yang kita inginkan menggunakan *user define function*. Berikut adalah contoh argumen yang dapat diubah pada `theme_wsj`.

```
theme_wsj
```

```
## function (base_size = 12, color = "brown", base_family = "sans",
##           title_family = "mono")
## {
##   colorhex <- ggthemes::ggthemes_data$wsj$bg[color]
##   theme.foundation(base_size = base_size, base_family = base_family) +
##     theme(line = element_line(linetype = 1, colour = "black"),
##           rect = element_rect(fill = colorhex, linetype = 0,
##                               colour = NA), text = element_text(colour = "black"),
##           title = element_text(family = title_family, size = rel(2)),
##           axis.title = element_blank(), axis.text = element_text(face = "bold",
##                       size = rel(1)), axis.text.x = element_text(colour = NULL),
##           axis.text.y = element_text(colour = NULL), axis.ticks = element_line(colour = NULL),
##           axis.ticks.y = element_blank(), axis.ticks.x = element_line(colour = NULL),
##           axis.line = element_line(), axis.line.y = element_blank(),
##           legend.background = element_rect(), legend.position = "top",
##           legend.direction = "horizontal", legend.box = "vertical",
##           panel.grid = element_line(colour = NULL, linetype = 3),
##           panel.grid.major = element_line(colour = "black"))
```

```

##             panel.grid.major.x = element_blank(), panel.grid.minor = element_blank(),
##             plot.title = element_text(hjust = 0, face = "bold"),
##             plot.margin = unit(c(1, 1, 1, 1), "lines"), strip.background = element_rect()
## }
## <bytecode: 0x0000000018b1d858>
## <environment: namespace:ggthemes>

```

Berdasarkan output yang disajikan kita dapat merubah sejumlah argumen seperti base size, color, base_family, dll.

5.10.8 Penskalaan dan Transformasi Axis

Pada bagian ini penulis akan menjelaskan bagaimana cara melakukan modifikasi terhadap sumbu x dan y seperti menetapkan limit nilai maksimum dan minimum axis serta melakukan transformasi pada tiap axis.

Untuk mengatur rentang nilai axis, kita dapat melakukannya dengan fungsi sebagai berikut:

- **xlim()** dan **ylim()**: mengatur limit aksis sumbu x dan y.
- **expand_limits()**: mengatur limit sumbu x dan y sekaligus dapat mengatur intercept kedua sumbu tersebut.
- **scale_x_continuous()** dan **scale_y_continuous()**: megatur limit axis termasuk axis tick dan label.

Pada contoh berikut akan disajikan cara mengatur limit axis dengan menggunakan **xlim()** dan **ylim()** serta menggunakan **expand_limits()**. Output yang dihasilkan disajikan pada Gambar 5.54.

```

gapminder%>%
  filter(continent=="Europe")%>%
  ggplot(aes(gdpPerCap, lifeExp))+
  geom_point()+
  theme_wsj(base_size=7)+
  labs(title="GDP per Capita vs Life Expectancy",
       y="Life Expectancy",
       x="GDP per Capita (US Dollar)")+
  # mengatur limit axis
  expand_limits(x=c(0, 55000), y=c(0, 90))

```

```

# atau
gapminder%>%
  filter(continent=="Europe")%>%
  ggplot(aes(gdpPerCap, lifeExp))+
  geom_point()+
  theme_wsj(base_size=7)+
  labs(title="GDP per Capita vs Life Expectancy",
       y="Life Expectancy",
       x="GDP per Capita (US Dollar)")+
  # mengatur limit axis
  xlim(0,55000)+
  ylim(0,90)

```

Kita juga dapat menggunakan fungsi **scale_x_continuous()** dan **scale_y_continuous()** untuk mengatur limit axis ,*axis tick* dan label. Format yang digunakan adalah sebagai berikut:

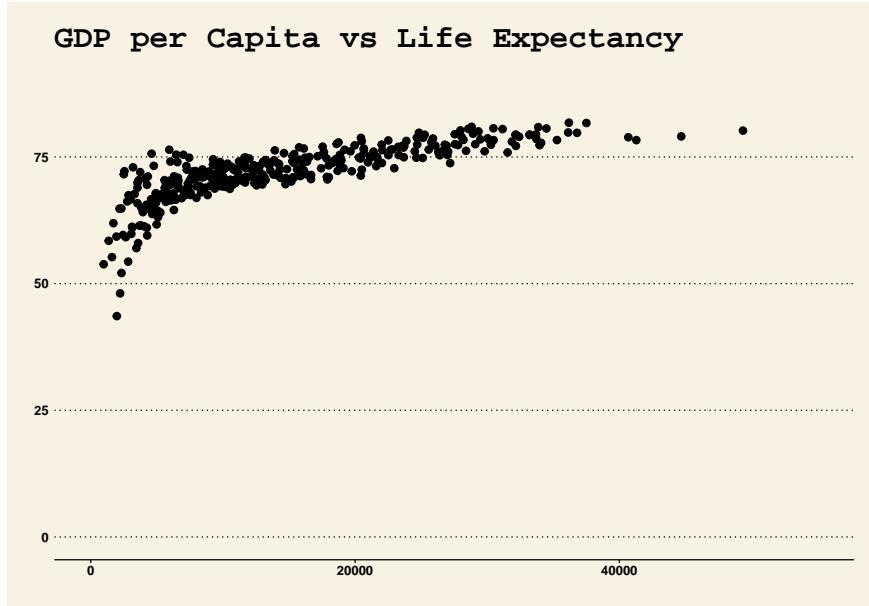


Figure 5.54: Scatterplot dengan axis limits

```
scale_x_continuous(name, breaks, labels, limits, trans)
scale_y_continuous(name, breaks, labels, limits, trans)
```

Note:

- **name:** label axis sumbu x dan y.
- **breaks:** untuk mengontrol jeda dalam panduan (*axis tick*, garis grid, ...). Di antara nilai-nilai yang mungkin, adalah sebagai berikut:
 - NULL: menyembunyikan seluruh breaks.
 - waiver(): komputasi break default.
 - vektor numerik atau karakter untuk menspesifikasi break yang akan ditampilkan.
- **labels:** label axis. Nilai yang dapat dimasukkan antara lain;
 - NULL: tanpa label.
 - waiver(): label default.
 - vektor karakter yang digunakan untuk spesifikasi label break.
- **limits:** vektor numerik untuk spesifikasi limit sumbu x dan y.
- **trans:** transformasi axis. Nilai yang dapat digunakan adalah “log2”, “log10”, dll.

Pada contoh berikut disajikan contoh mengatur limit axis dan label axis menggunakan fungsi `scale_x_continuous()` dan `scale_y_continuous()`. Grafik yang dihasilkan akan tampak seperti Gambar 5.55.

```
# atau
gapminder %>%
  filter(continent == "Asia") %>%
  ggplot(aes(gdpPerCap, lifeExp)) +
  geom_point() +
```

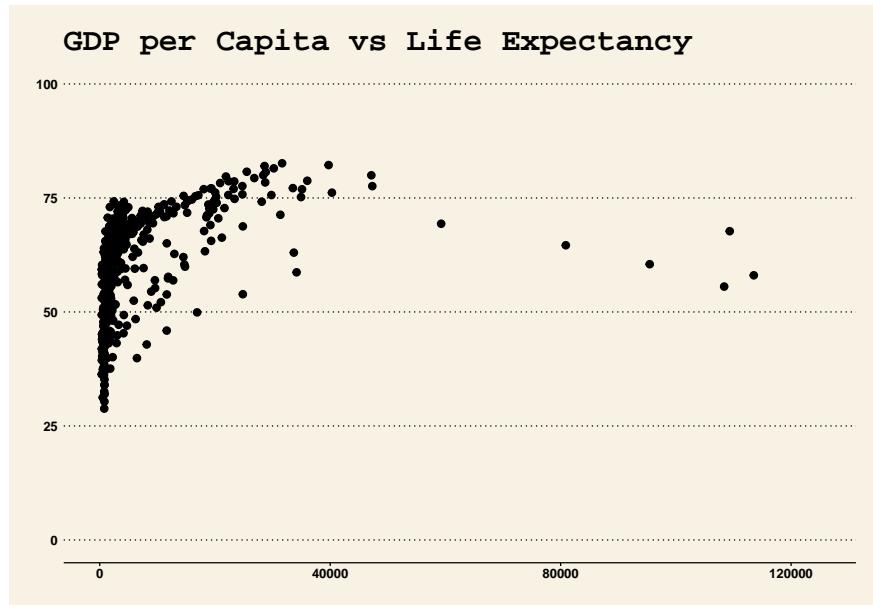


Figure 5.55: Scatterplot dengan axis limits (2)

```
theme_wsj(base_size=7) +
  ggtitle("GDP per Capita vs Life Expectancy") +
  # spesifikasi limit dan label axis
  scale_x_continuous(name="GDP per Capita",
                      limits=c(0, 125000)) +
  scale_y_continuous(name="Life Expectancy",
                      limits=c(0,100))
```

Transformasi axis dapat dilakukan dengan fungsi bawaan dari ggplot2. Fungsi transformasi bawaan berupa transformasi log dan sqrt. Berikut adalah fungsi bawaan untuk transformasi tersebut:

- `scale_x_log10()` dan `scale_y_log10()`: transformasi log basis 10.
- `scale_x_sqrt()` dan `scale_y_sqrt()`: transformasi akar kuadrat.
- `scale_x_reverse()` dan `scale_x_reverse()`: membalikkan koordinat.
- `coord_trans(x="log10", y="log10")`: memungkinkan transformasi untuk kedua axis sesuai fungsi yang diinputkan pada sumbu x dan sumbu y seperti "log2", "log10", "sqrt", dll.
- `scale_x_continuous(trans="log2")` dan `scale_y_continuous(trans="log2")`: nilai lain yang dapat diinputkan adalah "log10".

Pada contoh berikut disajikan contoh transformasi sumbu x menggunakan fungsi `scale_x_log10()`. Grafik yang dihasilkan akan tampak seperti Gambar 5.56.

```
# atau
gapminder %>%
  filter(continent=="Europe") %>%
  ggplot(aes(gdpPercap, lifeExp)) +
  geom_point() +
  theme_wsj(base_size=7) +
  labs(title="log(GDP per Capita) vs Life Expectancy",
       y="Life Expectancy",
```

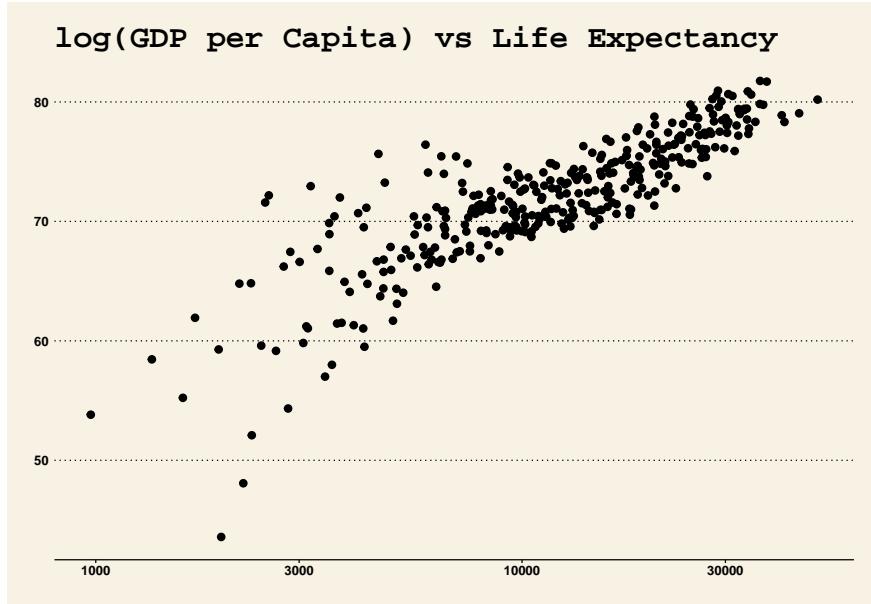


Figure 5.56: Scatterplot dengan transformasi axis

```
x="GDP per Capita (US Dollar)")+
# transformasi sumbu x
scale_x_log10()
```

Tick mark pada axis juga dapat kita atur menggunakan fungsi `scale_x_continuous()` dan `scale_y_continuous()`. Untuk mengubah format dan label *tick mark* kita perlu menginstall dan memuat library `scales` yang berfungsi untuk mengakses fungsi pada argumen `break`. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.57.

```
# memasang paket
# install.packages("scales")

# memuat paket
library(scales)

## Warning: package 'scales' was built under R version
## 3.5.3

# plot
ggplot(gapminder, aes(gdpPercap, lifeExp))+
  geom_point()+
  theme_bw()+
  # kustomisasi tick mark sumbu y
  scale_y_continuous(trans= log2_trans(),
                     breaks=trans_breaks("log2", function(x) 2^x),
                     labels= trans_format("log2", math_format(2^.x)))+
  # kustomisasi sumbu x
  scale_x_continuous(labels = dollar)
```

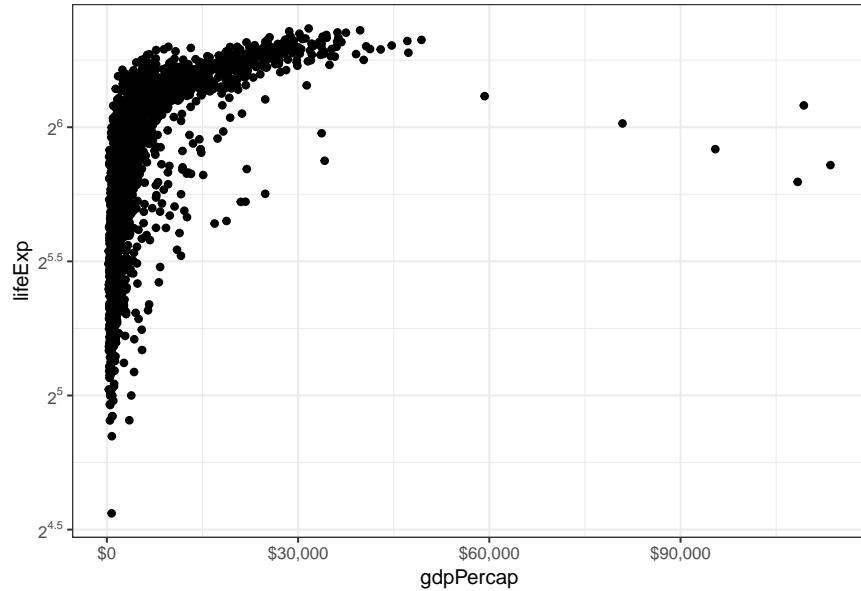


Figure 5.57: Scatterplot dengan transformasi tick mark axis

5.10.9 Kustomisasi Tick Mark Axis

Pada bagian ini pembaca akan mempelajari bagaimana melakukan kustomisasi tampilan *tick mark*. Selain itu kita juga akan belajar bagaimana melakukan pengaturan pada garis axis.

Warna, ukuran font, dan tampilan font (*font style*) pada *tick mark* dapat diubah menggunakan fungsi `theme()` dan `element_text()`. Format yang digunakan adalah sebagai berikut:

```
# x axis tick mark labels
<plot> + theme(axis.text.x = element_text(family, face, colour, size, angle))
# y axis tick mark labels
<plot> + theme(axis.text.y = element_text(family, face, colour, size, angle))
```

Note:

- **family:** *font family*, seperti: “sans”, “times new roman”, dll.
- **face:** *font face*, nilai yang mungkin adalah “plain”, “italic”, “bold” dan “bold.italic”.
- **color:** warna teks.
- **size:** ukuran teks dalam satuan pts.
- **angle:** sudut kemiringan teks berkisar antara 0 sampai 360.

Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.58.

```
ggplot(gapminder, aes(continent, gdpPercap,
                      fill=continent)) +
  geom_boxplot() +
  theme_economist() +
  scale_fill_economist() +
  # kustomisasi tick mark
  theme(axis.text.x = element_text(face="bold",
                                    color="#993333",
```

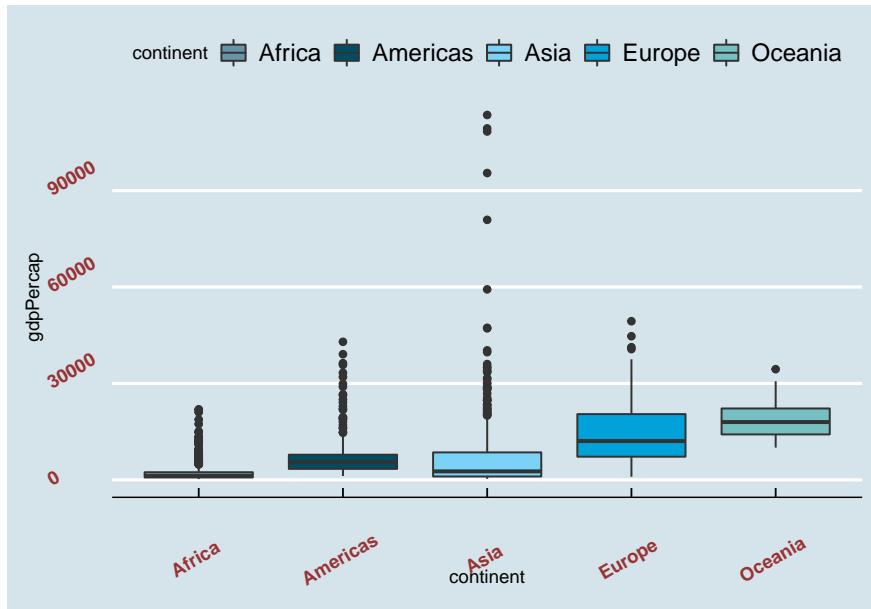


Figure 5.58: Mengubah tampilan dari tick mark

```

        size=10,
        angle=30),
axis.text.y = element_text(face="bold",
                           color="#993333",
                           size=10,
                           angle=30))

```

Untuk menonaktifkan *tick mark* pada plot kita dapat menggunakan fungsi `element_blank()`. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.59.

```

ggplot(gapminder, aes(continent, gdpPercap,
                      fill=continent)) +
  geom_boxplot() +
  theme_stata() +
  scale_fill_stata() +
  # menyembunyikan tick mark dan tick mark label
  theme(axis.text.x=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks=element_blank())

```

Kita dapat melakukan pengaturan terhadap garis axis menggunakan argumen `axis.lines` dan fungsi `element_line`. Berikut adalah format yang digunakan:

```
<plot> + theme(axis.line = element_line(color, size, linetype,
                                         lineend, color))
```

Note:

- `color`: warna garis.
- `size`: ukuran garis.

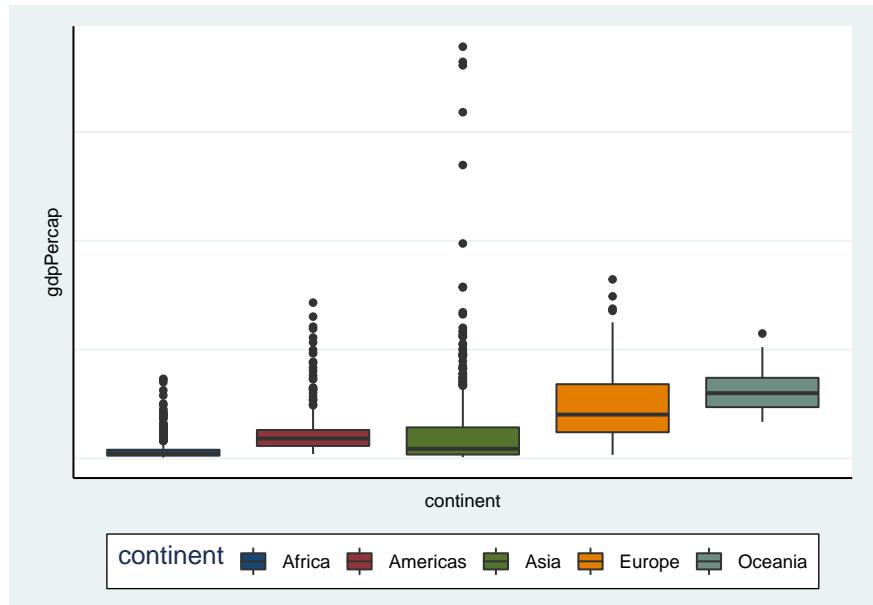


Figure 5.59: Menyembunyikan tampilan dari tick mark

- **linetype:** jenis garis.
- **lineend:** akhir dari garis. Nilai yang dapat dimasukkan antara lain: “round”, “butt” atau “square”.

Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.60.

```
ggplot(gapminder, aes(continent, gdpPercap,
                      fill=continent))+
  geom_boxplot()+
  theme_wsj()+
  scale_fill_wsj()+
  # kustomisasi garis axis
  theme(axis.line = element_line(colour = "darkblue",
                                  size = 1, linetype = "solid"))
```

Kita dapat mengatur *tick* pada axis baik yang memiliki skala diskrit maupun kontinyu. Fungsi yang digunakan adalah `scale_x_continuous()` dan `scale_y_continuous()` untuk *tick* dengan nilai kontinyu dan `scale_x_discrete()` dan `scale_y_discrete()`.

Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.61.

```
ggplot(gapminder, aes(continent, lifeExp,
                      fill=continent))+
  geom_boxplot()+
  theme_gdocs()+
  scale_fill_gdocs()+
  # kustomisasi tick mark
  scale_y_continuous(
    # nilai dari 0 sampai 100 tiap 10 tick
    breaks=seq(0,100,10))
```

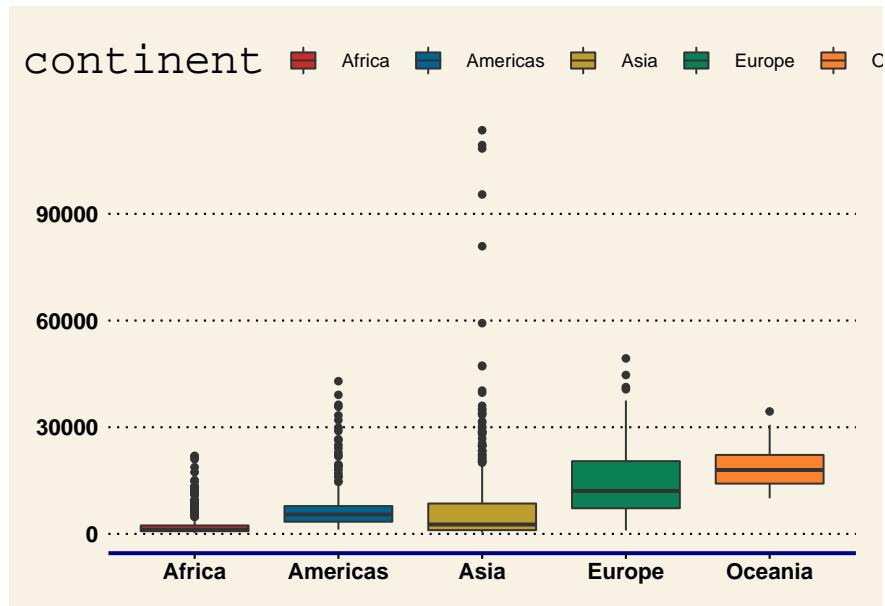


Figure 5.60: Kustomisasi tampilan dari garis axis

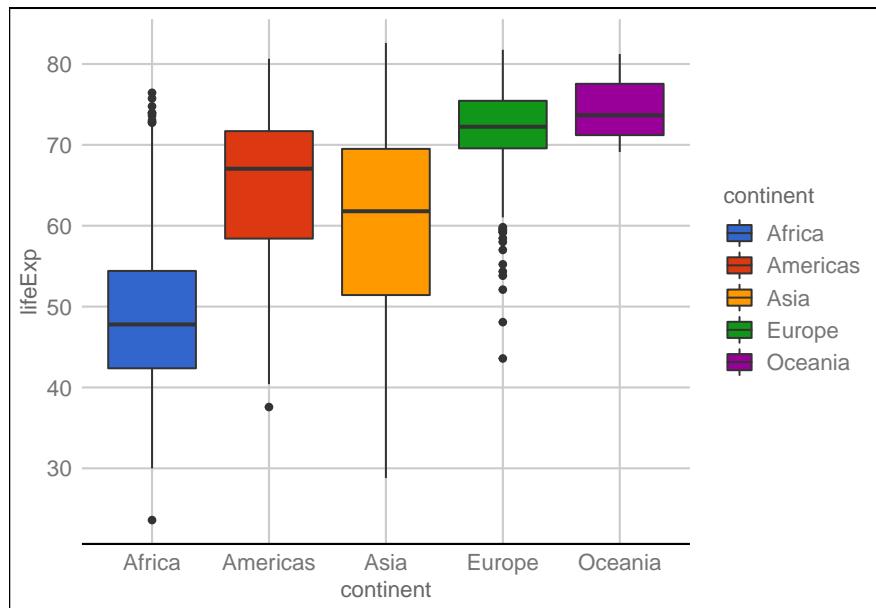


Figure 5.61: Kustomisasi tick mark

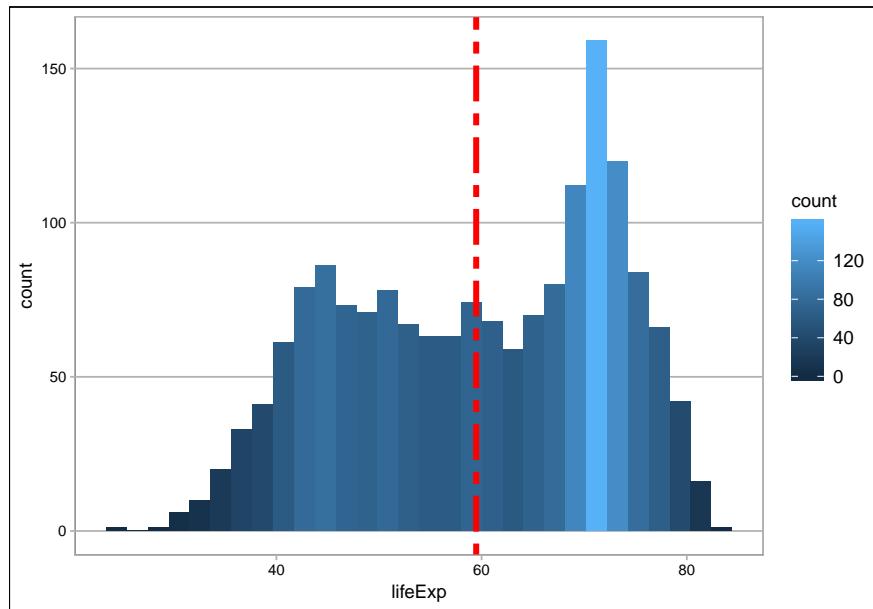


Figure 5.62: Penerapan vline

5.10.10 Menambahkan Garis Lurus Pada Plot

Fungsi yang dapat digunakan untuk menambahkan garis lurus antara lain:

- **geom_hline()**: menambahkan garis horizontal.
- **geom_abline()**: menambahkan garis regresi.
- **geom_vline()**: menambahkan garis vertikal.
- **geom_segment()**: menambahkan garis segmen.

Format yang digunakan untuk fungsi `geom_hline()` dan `geom_vline()` adalah sebagai berikut:

```
geom_hline(yintercept, linetype, color, size)
geom_vline(xintercept, linetype, color, size)
```

Berikut adalah contoh penerapan kedua fungsi tersebut yang disajikan pada Gambar 5.62 dan Gambar 5.63:

```
ggplot(gapminder, aes(lifeExp, fill=..count..))+
  geom_histogram()+
  theme_calc()+
  # menambahkan garis vertikal
  geom_vline(xintercept=mean(gapminder$lifeExp),
             linetype="twodash",
             color="red",
             size=1.5)
```

```
ggplot(gapminder, aes(continent, lifeExp,
                      fill=continent))+
  geom_boxplot()+
  theme_calc()+
```

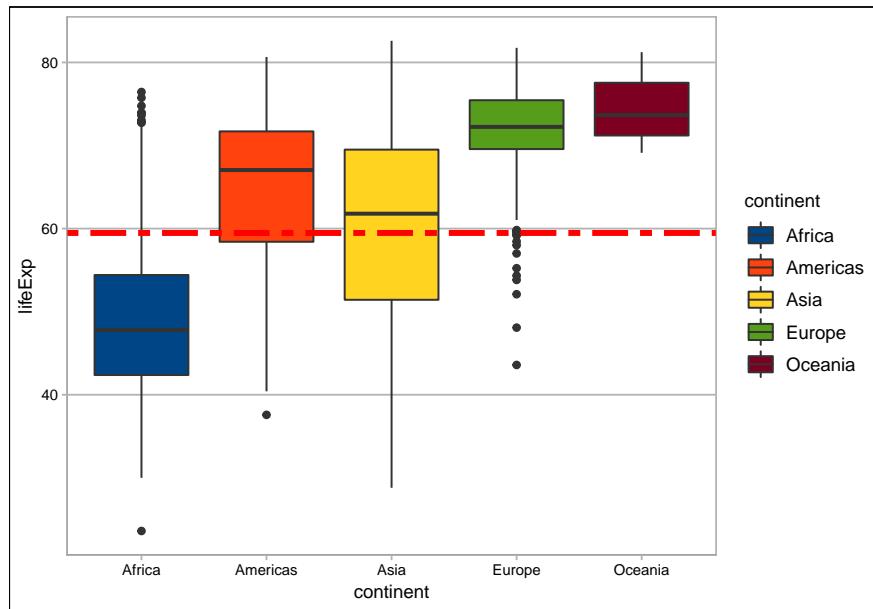


Figure 5.63: Penerapan hline

```
scale_fill_calc()+
# menambahkan garis horizontal
geom_hline(yintercept=mean(gapminder$lifeExp),
linetype="twodash",
color="red",
size=1.5)
```

Selain menggunakan fungsi `geom_smooth()`, garis regresi dapat ditambahkan melalui fungsi ‘`geom_abline()`’. Format yang digunakan adalah sebagai berikut:

```
geom_abline(intercept, slope, linetype, color, size)
```

Untuk membuat garis regresi kita perlu membuat model regresi terlebih dahulu menggunakan fungsi `lm()`. Berikut adalah contoh model yang dibuat beserta koefisien regresinya.

```
# membuat model regresi
mod <- lm(lifeExp~gdpPercap, data=gapminder)

# print model
mod

##
## Call:
## lm(formula = lifeExp ~ gdpPercap, data = gapminder)
##
## Coefficients:
## (Intercept)      gdpPercap
##      5.40e+01    7.65e-04
```

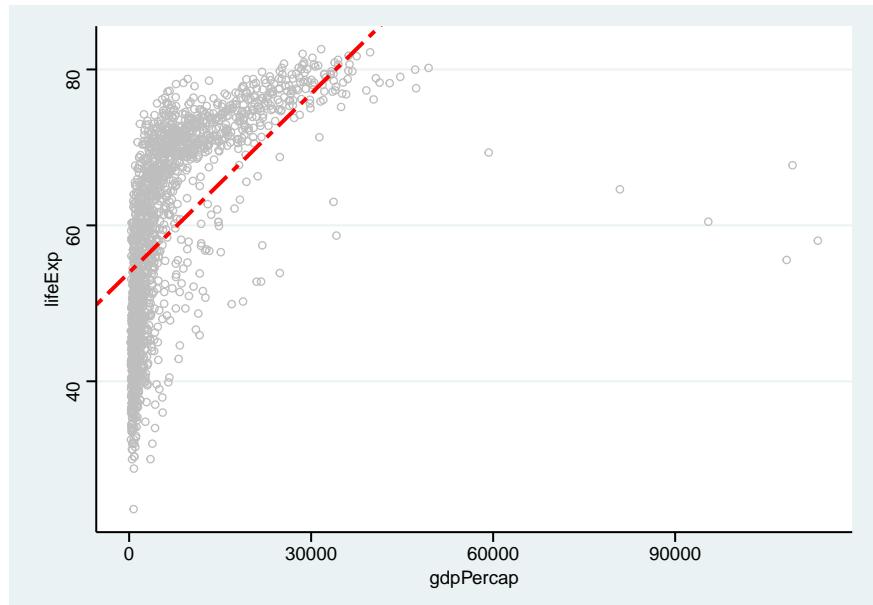


Figure 5.64: Penerapan abline

```
# koefisien regresi model
coef <- coefficients(mod)

# print koefisien
coef
```

```
## (Intercept)    gdpPercap
##      5.396e+01   7.649e-04
```

Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.64 untuk membuat plot regresi linier.

```
ggplot(gapminder, aes(gdpPercap, lifeExp))+
  geom_point(shape=1, color="grey")+
  theme_stata()+
  # menambahkan garis regresi
  geom_abline(intercept=5.395556e+01,
              slope=7.648826e-04,
              linetype="twodash",
              color="red",
              size=1)
```

Kita dapat menambahkan garis segment untuk menunjukkan sebuah observasi. Format yang digunakan adalah sebagai berikut:

```
geom_segment(aes(x, y, xend, yend))
```

Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.65 untuk membuat garis segmen.

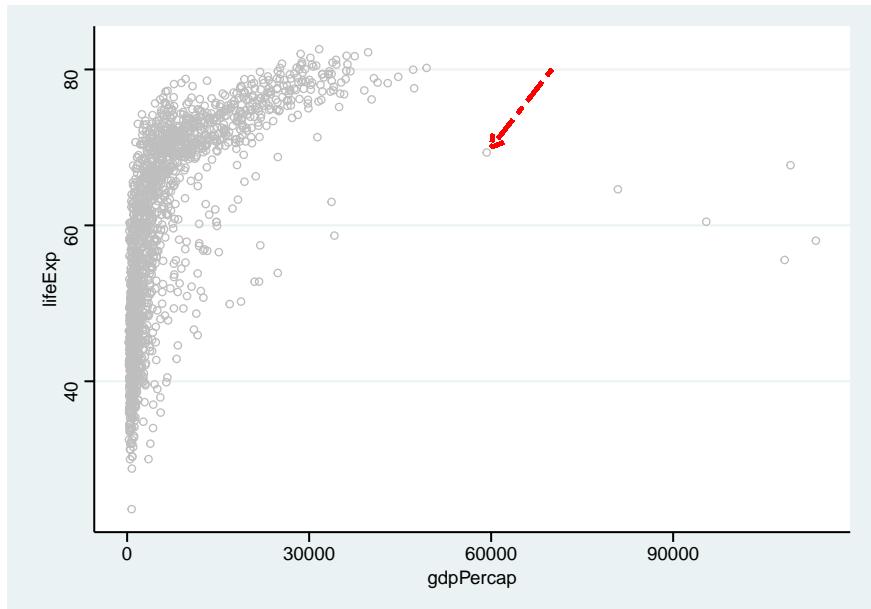


Figure 5.65: Penerapan garis segmen

```
library(grid)
ggplot(gapminder, aes(gdpPercap, lifeExp))+
  geom_point(shape=1, color="grey")+
  theme_stata()+
  # menambahkan tanda panah
  geom_segment(x=70000, y=80,
               xend=60000, yend=70,
               arrow=arrow(length=unit(0.1, "inches")),
               linetype="twodash",
               color="red",
               size=1)
```

5.10.11 Melakukan Rotasi Pada Grafik

Rotasi grafik atau pembalikan axis dapat dilakukan menggunakan fungsi berikut:

- **coord_flip()**: untuk membuat plot horizontal. Rotasi axis sehingga sumbu x dapat menjadi sumbu y dan sebaliknya.
- **scale_x_reverse()** dan **scale_x_reverse()**: pembalikan skala pada axis.

Misalkan kita ingin membuat plot horizontal pada box plot sehingga mempermudah kita dalam melakukan perbandingan terhadap masing-masing grup. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.66.

```
ggplot(gapminder, aes(continent, lifeExp,
                      fill=continent))+
  geom_boxplot()+
  theme_economist()+
  scale_fill_economist()
```

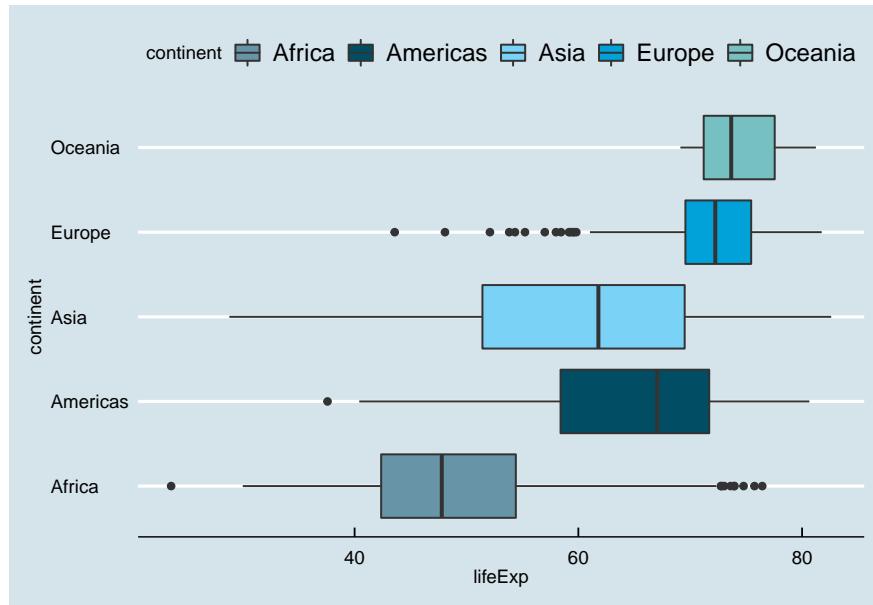


Figure 5.66: Rotasi axis

```
# rotasi axis
coord_flip()
```

Kita dapat juga melakukan pembalikan skala pada axis sehingga skala yang semula berawal dari min ke max menjadi sebaliknya. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.67.

```
ggplot(gapminder, aes(lifeExp, fill=..count..))+
  geom_histogram()+
  theme_wsj()+
  # pembalikan sumbu y
  scale_y_reverse()
```

5.10.12 Facet

Facet digunakan untuk membagi plot menjadi panel matriks. Setiap panel menunjukkan setiap kelompok data. Fungsi facet yang dapat digunakan antara lain:

- `facet_grid()`
- `facet_wrap()`

Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.68 dan Gambar 5.69 untuk membuat facet pada satu variabel.

```
ggplot(gapminder, aes(lifeExp, fill=..count..))+
  geom_histogram()+
  theme_gdocs()+
  facet_grid(.~continent)
```

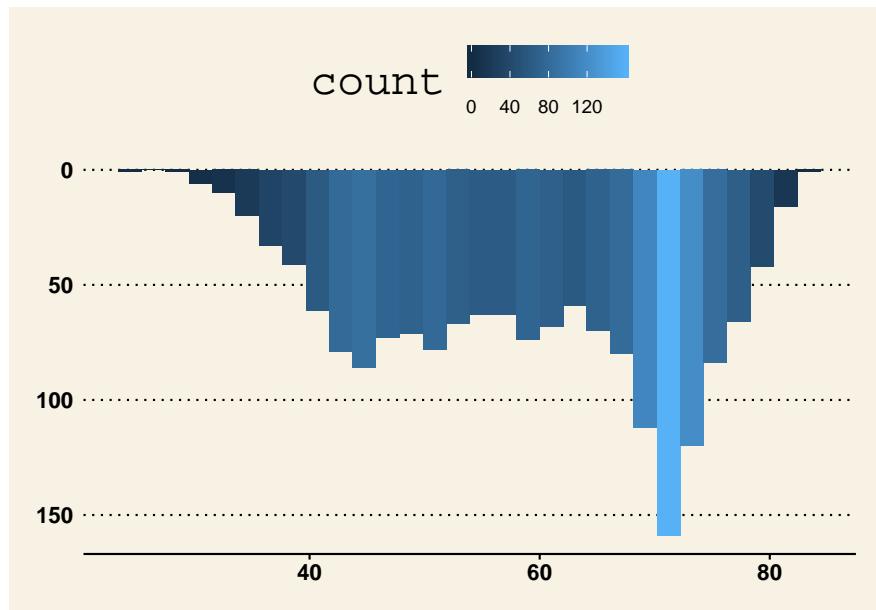


Figure 5.67: Pembalikan sumbu y

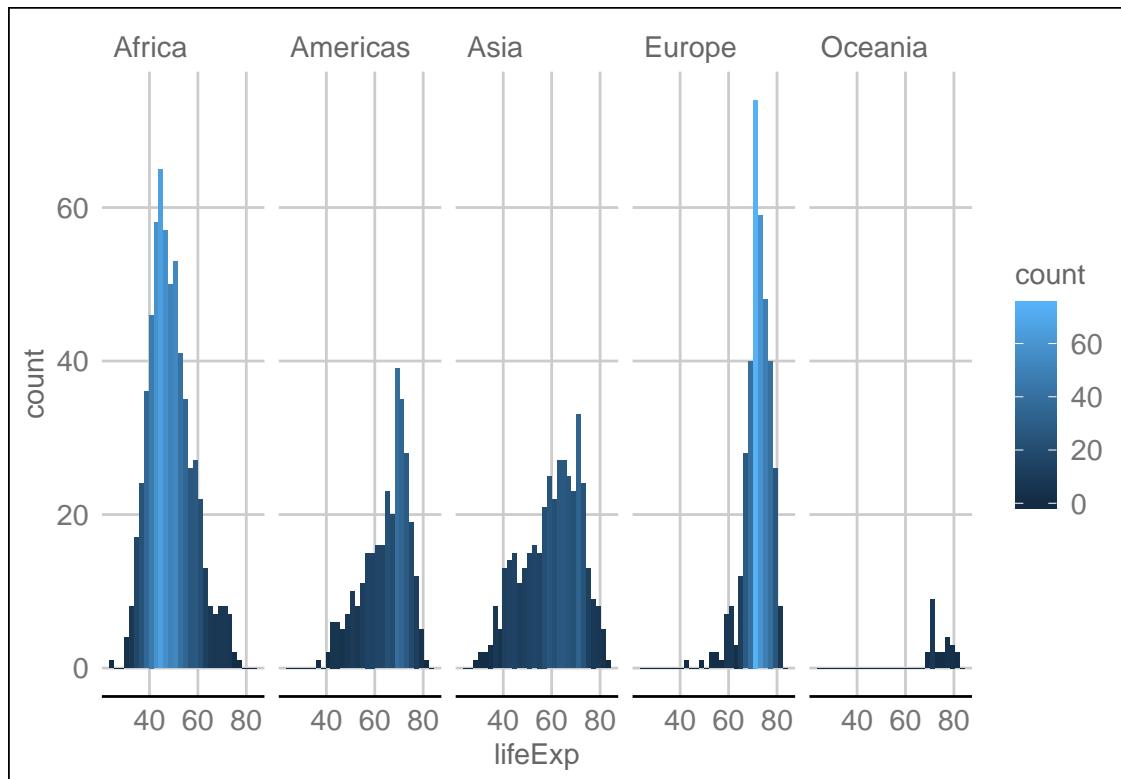


Figure 5.68: Facet horizontal satu variabel

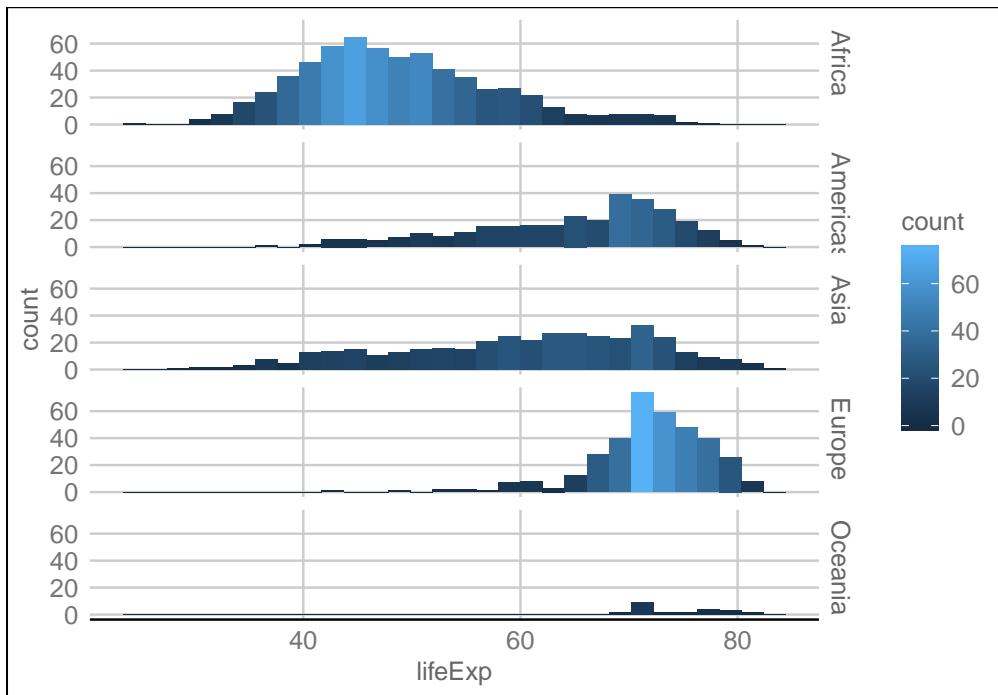


Figure 5.69: Facet vertikal satu variabel

```
ggplot(gapminder, aes(lifeExp, fill=..count..))+
  geom_histogram()+
  theme_gdocs()+
  facet_grid(continent~.)
```

Kita dapat pula melakukan facet terhadap dua buah variabel. Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.70 untuk membuat facet pada dua variabel.

```
gapminder%>%
  filter(year==1952|year==2007,
        continent %in% c("Asia", "Americas"))%>%
  ggplot(aes(continent, lifeExp,
             fill=factor(year)))+
  geom_boxplot()+
  theme_stata()+
  scale_fill_stata()+
  facet_grid(continent~factor(year))
```

Kita dapat mengatur skala dari axis menggunakan argument sebagai berikut:

- **free**: skala akan disesuaikan berdasarkan pada setiap axis.
- **free_x**: skala pada sumbu x akan dibiarkan menyesuaikan secara bebas.
- **free_y**: skala pada sumbu y akan dibiarkan menyesuaikan secara bebas.
- **fixed** (default): skala axis diseragamkan pada seluruh panel.

Berikut adalah sintaks yang digunakan beserta output yang dihasilkan pada Gambar 5.71 untuk membuat facet pada dua variabel dengan skala bebas pada sumbu y.

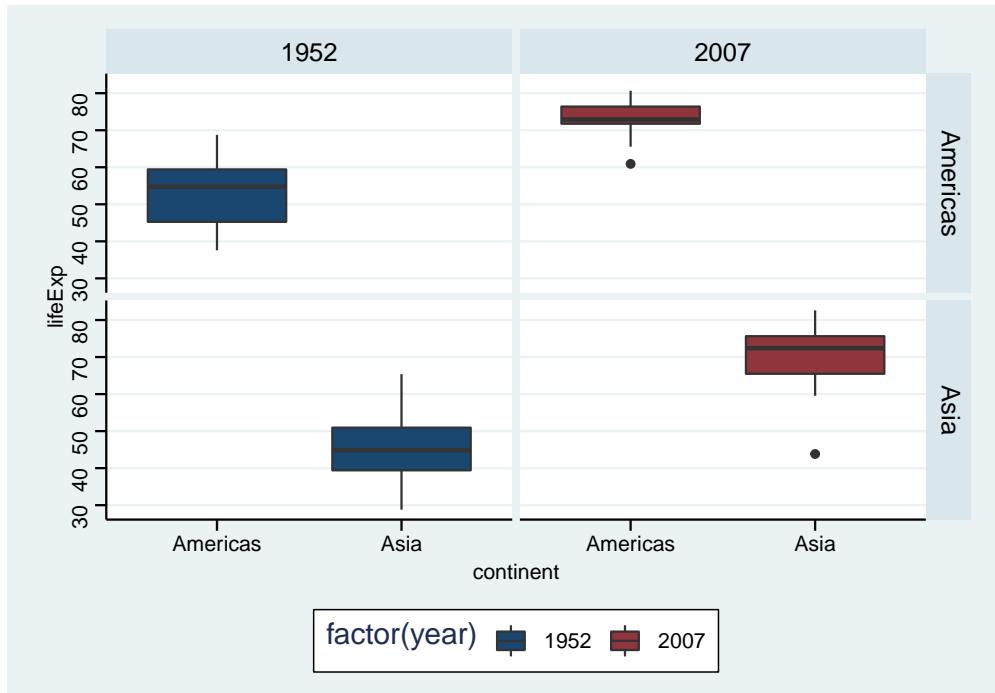


Figure 5.70: Facet dua variabel

```
gapminder%>%
  filter(year==1952 | year==2007,
        continent %in% c("Asia", "Americas"))%>%
  ggplot(aes(continent, lifeExp,
             fill=factor(year)))+
  geom_boxplot()+
  theme_stata()+
  scale_fill_stata()+
  facet_grid(continent~factor(year), scales="free_y")
```

5.11 Referensi

1. Wickham, H. Grolemund G. 2016. **R For Data Science: Import, Tidy, Transform, Visualize, And Model Data**. O'Reilly Media, Inc.
2. Peng, R.D. 2015. **Exploratory Data Analysis with R**. Leanpub book.
3. GGPLOT2 Documentation. <https://ggplot2.tidyverse.org/>
4. STHDA. ggplot2 - Essentials. <https://www.sthda.com/english/wiki/ggplot2-essentials>

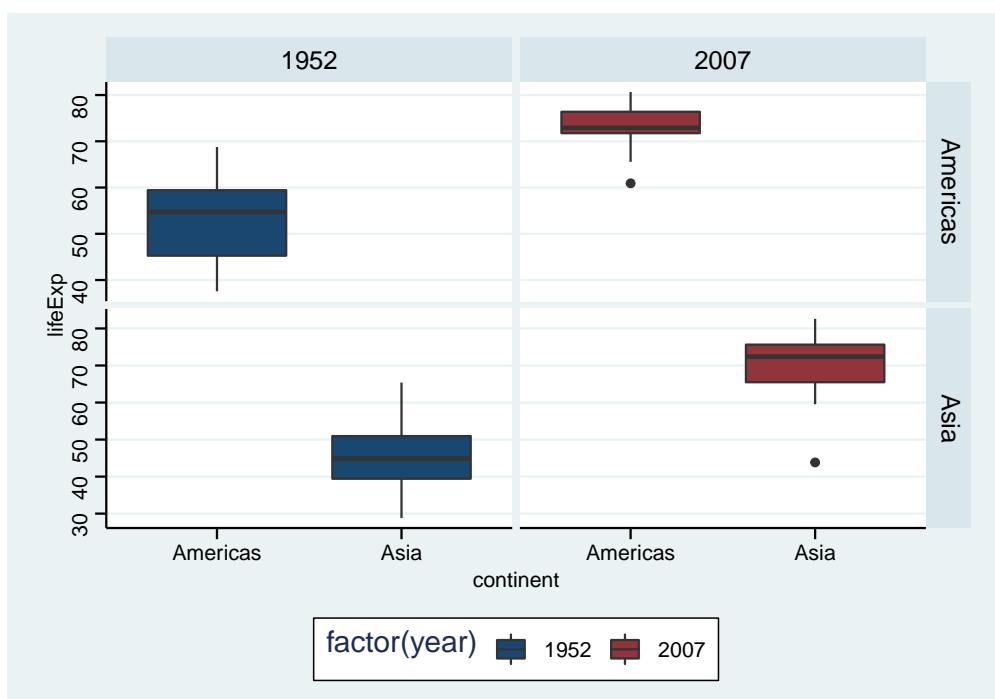


Figure 5.71: Facet dua variabel dengan skala bebas pada sumbu y

Statistika Deskriptif - R

Chapter 6

Ringkasan Numerik

Pada bidang lingkungan kita sering kali menemui sebuah pernyataan “konsentrasi rata-rata TSS pada sungai tersebut adalah 30 mg/l” atau “kedalaman penampang saluran tersebut berkisar antara 1 sampai 2 meter”. Kedua pernyataan tersebut merupakan sebuah penyampaian informasi terkait karakteristik data yang ada. Pernyataan yang pertama menyatakan karakteristik nilai pemusatan data, sedangkan yang kedua menyatakan karakteristik sebaran suatu data.

Karakteristik lain yang sering digunakan untuk menjelaskan suatu data adalah bentuk distibusi suatu data dan estimasi nilai ekstrim seperti nilai maksimum dan minimum suatu data. Seluruh karakteristik data tersebut perlu dihitung untuk memperoleh informasi numerik pada data.

Pada chapter ini kita akan membahas terkait metode untuk membuat ringkasan dan deksripsi data. Pembahasan akan terdiri dari ukuran nilai pemusatan data, ukuran sebaran atau variabilitas data dan bentuk distribusi data. Selain itu kita akan membahas nilai ekstrim yang ada pada sebuah data dan transformasi data.

6.1 Ukuran Pemusatan Data

Nilai rata-rata (mean) dan nilai tengah (median) merupakan dua nilai yang paling umum digunakan untuk menyatakan lokasi pemusatan data meskipun kedua nilai bukanlah satu atau dua ukuran yang tersedia. Apa sajakah properti dari kedua ukuran tersebut dan kapan salah satu atau keduanya dapat digunakan bersamaan?.

6.1.1 Pengukuran Klasik-Mean

Nilai mean (\bar{X}) diperoleh dengan menjumlahkan seluruh data dan membaginya dengan jumlah observasinya yang dapat dituliskan seperti Persamaan (6.1):

$$\bar{X} = \sum_{i=1}^n \frac{X_i}{n} \quad (6.1)$$

Nilai mean yang disimbolkan dengan “X bar” merupakan nilai mean untuk sampel. Nilai mean untuk populasi disimbolkan oleh huruf Yunani “mu atau μ ”.

Pada Persamaan (6.1), jika data terdiri dari banyak grup maka nilai rata-rata dihitung berdasarkan jumlah nilai observasi dikali dengan bobotnya. Nilai mean tersebut disebut sebagai *weighted mean* yang dapat ditulis berdasarkan Persamaan Persamaan (6.2).

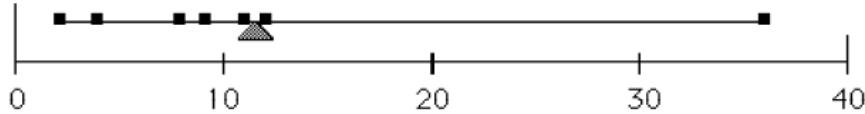


Figure 6.1: Nilai mean (segitiga) sebagai titik kesetimbangan pada data.

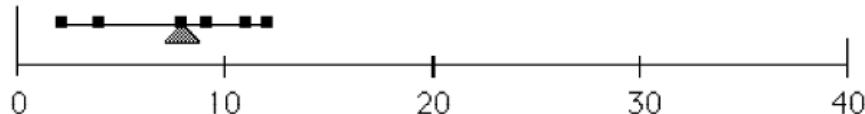


Figure 6.2: Pergeseran nilai mean (segitiga) ke kiri setelah penghilangan outlier.

$$\bar{X} = \sum_{i=1}^n \bar{X}_i \cdot \frac{n_i}{n} \quad (6.2)$$

dimana \bar{X}_i merupakan nilai rata-rata grup ke- i dan $\frac{n_i}{n}$ merupakan bobot pengali yang berupa rasio antara observasi grup ke- i dengan keseluruhan observasi.

Kita biasanya akan berhadapan dengan nilai observasi yang baru sehingga nilai mean yang telah ada akan ikut berubah. Perubahan nilai mean tersebut disebabkan karena setiap observasi yang disertakan dalam perhitungan mean memiliki pengaruhnya masing-masing. Jika observasi tersebut cenderung ekstrim besar maka nilai mean akan bergeser menuju kearahnya begitu juga sebaliknya.

Pengaruh dari sebuah nilai observasi ke- j atau X_j dapat dilihat dengan menghitung seluruh observasi secara bersamaan kecuali observasi ke- j pada sebuah grup. Dapat dituliskan pada Persamaan (6.4)

$$\bar{X} = \bar{X}_{(j)} \cdot \frac{(n-1)}{n} + X_j \cdot \frac{1}{n} \quad (6.3)$$

$$\bar{X} = \bar{X}_{(j)} + (X_j - \bar{X}_{(j)}) \cdot \frac{1}{n} \quad (6.4)$$

dimana $\bar{X}_{(j)}$ adalah nilai mean seluruh observasi kecuali X_j . Setiap observasi yang mempengaruhi nilai mean keseluruhan (\bar{X}) didefinisikan oleh $(X_j - \bar{X}_{(j)})$ sebagai jarak antara observasi tersebut dengan nilai rata-rata yang tidak termasuk observasi tersebut di dalamnya. Sehingga seluruh nilai observasi tidak memiliki pengaruh yang sama terhadap nilai rata-rata seluruh observasi.

Outlier merupakan observasi yang memiliki nilai yang ekstrim tinggi atau rendah dibanding seluruh observasi yang ada sehingga memiliki pengaruh yang besar terhadap nilai mean keseluruhan (\bar{X}). Pengaruhnya yang sangat besar terhadap nilai rata-rata keseluruhan akan menyebabkan nilai rata-rata akan bergeser ke arah *outlier* tersebut. Selain itu penampilan dari distribusi frekuensi yang terbentuk akan terlihat memiliki ekor yang panjang.

Untuk lebih memahami pengaruh observasi terhadap nilai rata-rata, disajikan dua buah gambar yaitu: Gambar 6.1 dan Gambar 6.2

Pada Gambar 6.1 disajikan 7 buah data konsentrasi TSS di suatu sungai. Nilai rata-rata TSS pada sungai tersebut adalah 11 mg/l. Jika kita amati sebagian besar data (6 observasi) berada pada interval nilai konsentrasi TSS 2 sampai 12 mg/l. Observasi yang lain terletak jauh dari mayoritas observasi lainnya yaitu sebesar 37 mg/l. Observasi yang berbeda secara ekstrim dari nilai secara umum pada suatu data disebut

Table 6.1: Data Debit Sampel (m³/detik)

observasi	debit
1	457
2	185
3	133
4	160
5	119
6	115
7	101
8	58
9	68
10	50
11	65
12	128

sebagai *outlier*. Nilai *outlier* tersebut menyebabkan nilai rata-rata yang terbentuk tidak representatif terhadap keseluruhan data yang ada dan cenderung menggeser nilai rata-rata mendekati nilai *outlier* tersebut. Nilai observasi yang ekstrim biasanya muncul dari adanya kesalahan perlakuan terhadap sampel seperti botol sampel yang digunakan tidak bersih atau prosedur analisa yang dilakukan tidak standar sehingga memungkinkan adanya partikulat udara yang terukur pada proses penimbangan.

Salah satu cara untuk menangani adanya *outlier* tersebut adalah dengan menghapus observasi yang merupakan *outlier*. Pada Gambar 6.2 terlihat bahwa penghapusan *outlier* telah menggeser nilai rata-rata ke kiri. Nilai rata-rata yang baru tersebut jika diperhatikan dari Gambar 6.2 lebih menggambarkan keseluruhan data yang ada. Tidak terlihat adanya nilai yang berada jauh jaraknya dari nilai rata-rata yang baru.

Pada contoh tersebut dapat kita simpulkan bahwa nilai mean sangat sensitif terhadap adanya *outlier*. Pada prakteknya nilai mean tidaklah berdiri sendiri selama proses analisa. Nilai mean memerlukan nilai lain seperti median untuk menganalisa apakah data yang diperoleh tidak simetris yang dapat mengindikasikan adanya outlier.

Pada R untuk menghitung nilai rata-rata, kita dapat menggunakan fungsi `mean()`. Format fungsi yang digunakan dituliskan pada persamaan berikut:

```
mean(x, trim = 0, na.rm = FALSE)
```

Note:

- **x**: objek atau vektor numerik.
- **trim**: menyatakan fraksi data (berkisar antara 0 sampai 0,5) yang perlu dilakukan pemotongan (*trim*) pada observasi awal dan akhir **x** (yang telah diurutkan) sebelum nilai mean dihitung. **na.rm**: nilai logis yang menyatakan apakah *missing value* perlu disertakan dalam perhitungan atau tidak. Jika disertakan maka output yang akan dihasilkan adalah NA.

Analisa Nilai Mean Grup Data Tunggal (Single Group)

Untuk lebih memahami penerapannya pada R, pada Tabel 6.1 berikut disajikan data terkait debit air suatu sungai.

Data pada Tabel 6.1 dapat divisualisasikan seperti pada Gambar 6.3:

Berdasarkan Gambar 6.3, terdapat *outlier* yang ditunjukkan pada debit sungai yang lebih besar dari 400 m³/detik. Hasil tersebut dapat terjadi salah satunya karena adanya kondisi ekstrim seperti banjir yang menyebabkan sungai meluap atau terjadi kesalahan pengukuran dari alat ukur yang ada di lapangan.



Figure 6.3: Visualisasi debit sungai pada sampel

Untuk menghitung nilai rata-rata debit pada data tersebut, masukkan variabel `debit` yang telah penulis simpan sebagai objek `sungai` kedalam fungsi `mean()` seperti berikut:

```
mean(sungai$debit)
```

```
## [1] 136.6
```

Berdasarkan hasil yang diperoleh, dapat dilihat bahwa nilai rata-rata debit pada sungai tersebut adalah $136.5833 \text{ m}^3/\text{detik}$.

Kita dapat menghitung nilai mean dengan terlebih dahulu menghilangkan *outlier* pada data. Untuk melakukannya kita perlu melakukan subset terhadap data tanpa *outlier* di dalamnya sebelum data tersebut dimasukkan kedalam fungsi `mean()`. Berikut sintaks yang digunakan untuk melakukan hal tersebut:

```
# memuat paket
library(dplyr)

# melakukan filter terhadap data
sungai_subset<-sungai%>%
  filter(debit<=400)

# menghitung mean
mean(sungai_subset$debit)
```

```
## [1] 107.5
```

Berdasarkan hasil yang diperoleh terlihat bahwa nilai rata-rata yang baru lebih kecil dari yang sebelumnya (bergeser ke kiri) dengan nilai debit sungai yang baru sebesar $107.4545 \text{ m}^3/\text{detik}$. Hal ini terjadi karena pengaruh dari data *outlier* yang telah dihilangkan.

Analisa Nilai Rata-Rata Berdasarkan Grup Data

Pada contoh sebelumnya kita telah melakukan perhitungan nilai mean untuk studi kasus grup tunggal. Pada contoh ini akan disajikan contoh kasus perhitungan nilai mean untuk data berkelompok.

Dataset pada contoh kasus ini diambil dari buku **Statistical Methods in Water Resources**. Data yang digunakan adalah data konsentrasi TDS dan Uranium di airtanah dengan perbedaan konsentrasi bikarbonat dalam air tanah yaitu $\leq 50\%$ (0) dan $> 50\%$ (1). Dataset yang digunakan disajikan pada Tabel 6.2.

Note: data yang digunakan dapat diunduh pada link berikut [google.drive](#). Simpan dataset tersebut pada *working directory* pembaca agar mudah dalam proses membaca data.

```
# memuat library
library(readxl)
```

```
## Warning: package 'readxl' was built under R version
## 3.5.3
```

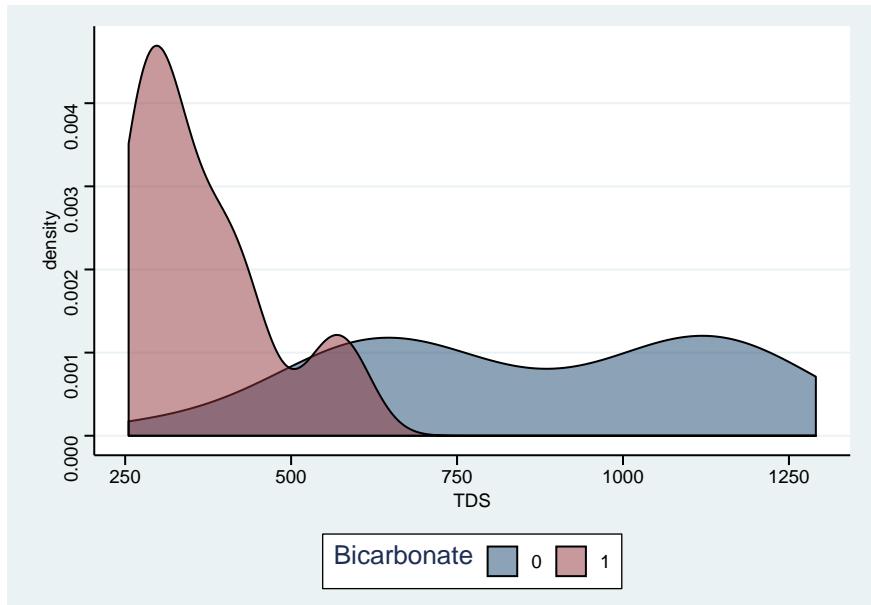


Figure 6.4: Visualisasi konsentrasi TDS pada air tanah

```
# memuat data excel
data_gw <- read_excel("hhappc.xls", sheet="appc16")

# membuang kolom ke-4
data_gw<-data_gw %>%
  select(TDS, Uranium, Bicarbonate) %>%
  mutate(Bicarbonate=as.factor(Bicarbonate))
```

Visualisasi data Tabel 6.2, disajikan pada Gambar 6.4 dan Gambar 6.5:

Pada dataset tersebut kita ingin melihat apakah terdapat perbedaan antara konsentrasi TDS dan uranium pada kondisi kesadahan bikarbonat $\leq 50\%$ dan $> 50\%$. Untuk melakukannya pada R kita perlu mengelompokkan data tersebut terlebih dahulu berdasarkan variabel bikarbonat. Setelah itu nilai rata-rata dapat dihitung. Berikut sintaks yang digunakan:

```
data_gw %>%
  group_by(Bicarbonate) %>%
  summarize(TDS = mean(TDS), Uranium = mean(Uranium))
```

```
## # A tibble: 2 x 3
##   Bicarbonate    TDS  Uranium
##   <fct>        <dbl>    <dbl>
## 1 0             864.     3.47
## 2 1             364.     5.16
```

Berdasarkan hasil yang diperoleh konsentrasi TDS dan Uranium dipengaruhi oleh kesadahan airtanah. Pada konsentrasi Bikarbonat $> 50\%$ konsentrasi TDS akan lebih rendah sedangkan konsentrasi Uranium sebaliknya. Untuk menguji apakah nilai tersebut berbeda signifikan, kita perlu melakukan uji hipotesis yang akan dibahas pada Chapter selanjutnya.

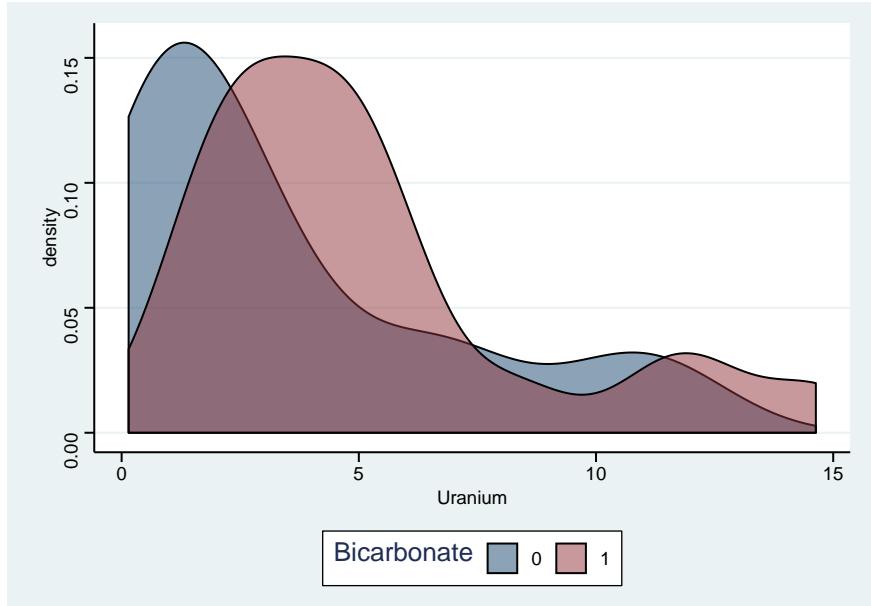


Figure 6.5: Visualisasi konsentrasi Uranium pada air tanah

6.1.2 Median Sebagai Ukuran Pemusatan Data yang Resistan

Median atau persentil 50 (P_{50}) merupakan nilai pusat dari distribusi suatu data yang telah dirangkinkan berdasarkan besar nilai observasinya. Untuk data dengan jumlah observasi ganjil median adalah titik tengah yang memiliki jumlah observasi yang sama baik di atas nilai media maupun di bawahnya. Untuk data dengan jumlah observasi genap, media merupakan rata-rata dari dua titik observasi pusat. Untuk memperoleh median dari suatu distribusi data, langkah pertama yang perlu dilakukan adalah mengurutkan data dari observasi dengan nilai terkecil sampai dengan yang besar sehingga x_1 merupakan observasi terkecil hingga x_n merupakan observasi terbesar. Persamaan (6.5) (untuk data ganjil) dan Persamaan (6.6) (untuk data genap) merupakan persamaan untuk menghitung median berdasarkan jumlah observasi yang ada.

$$\text{Median}(P_{0.5}) = \frac{X_{(n+1)}}{2} \quad (6.5)$$

$$\text{Median}(P_{0.5}) = \frac{1}{2} \cdot \left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}\right)+1} \right) \quad (6.6)$$

Median hanya dipengaruhi minimal oleh besarnya nilai observasi tunggal, yang ditentukan semata-mata oleh urutan relatif observasi. Resistensi terhadap efek dari perubahan nilai atau kehadiran pengamatan terpencil (*outlier*) sering merupakan sifat yang diinginkan. Meski demikian median memiliki kelemahan utama yaitu kurang representatif dalam mendeskripsikan rata-rata dari data dibandingkan mean. Hal ini disebabkan karena median tidak menggunakan seluruh nilai yang ada pada data.

Analisa Nilai Median Grup Data Tunggal (Single Group)

Kita akan menggunakan kembali data pada Tabel 6.1 untuk menghitung median data tersebut. Pada R median dihitung menggunakan fungsi `median()`. Format yang digunakan adalah sebagai berikut:

```
median(x, na.rm = FALSE)
```

Note:

- **x:** objek atau vektor numerik.
- **na.rm:** nilai logis yang menyatakan apakah *missing value* perlu disertakan dalam komputasi atau tidak.

Untuk data pada Tabel 6.1, median dapat dihitung menggunakan sintaks berikut:

```
median(sungai$debit)
```

```
## [1] 117
```

Berdasarkan hasil komputasi diperoleh median debit sungai sebesar $117 \text{ m}^3/\text{detik}$. Nilai tersebut tidak berbeda jauh dengan nilai mean tanpa *outlier* data sungai sebesar $107.4545 \text{ m}^3/\text{detik}$.

Jika kita melakukan perhitungan menggunakan menggunakan data `sungai_subset` (tanpa *outlier*), maka diperoleh $115 \text{ m}^3/\text{detik}$ yang nilainya juga tidak bergeser jauh dengan median sebelumnya yang membuktikan bahwa median resisten terhadap *outlier*.

Analisa Nilai Median Berdasarkan Grup Data

Pada contoh ini kita akan menggunakan kembali data pada Tabel 6.2. Sintaks berikut adalah cara menghitung median untuk data berkelompok:

```
data_gw %>%
  group_by(Bicarbonate) %>%
  summarize(TDS=median(TDS), Uranium=median(Uranium))
```

```
## # A tibble: 2 x 3
##   Bicarbonate     TDS Uranium
##       <fct>     <dbl>    <dbl>
## 1 0             819.     1.94
## 2 1             327.     4.46
```

Pada median TDS kita tidak menemui perbedaan dengan nilai rata-ratanya. Hal ini disebabkan karena bentuk distribusinya yang relatif simetris. Sedangkan pada Uranium distribusi yang terbentuk memiliki kemencengan (*skewness*) positif. Hal ini menyebabkan nilai mean yang terbentuk akan sangat dipengaruhi oleh observasi dengan nilai ekstrim yang dimiliki.

6.1.3 Ukuran Pemusatan Data Lainnya

Ukuran pemusatan data lainnya yang kurang sering digunakan adalah modus, rata-rata geometrik (*geometric mean*), dan *trimmed mean*. Modus merupakan nilai observasi yang sering muncul. Jika kita visualisasikan menggunakan histogram maka modus merupakan bar tertinggi pada histogram. Modus lebih dapat diaplikasikan pada data berkelompok yang nilai observasinya merupakan integer (*finite number*) dibanding data dengan nilai kontinyu. Modus sangat mudah diperoleh, namun sangat buruk sebagai ukuran pemusatan data untuk jenis data kontinyu karena sering bergantung pengelompokan data yang sewenang-wenang atau semaunya.

Geometric mean sering digunakan untuk distribusi data memiliki bentuk kemencengan positif. *Geometric mean* merupakan rata-rata logaritmik yang diubah kembali ke unit asalnya. Untuk menghitungnya digunakan Persamaan (6.7).

$$GM = \exp(\bar{Y}) \quad (6.7)$$

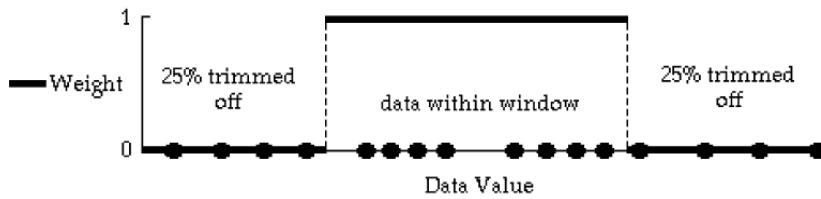


Figure 6.6: Jendela diagram trimmed mean.

dimana

$$Y_i = \ln(X_i) \quad (6.8)$$

Untuk data yang memiliki kemencengan positif, *geometric mean* biasanya cukup dekat dengan median. Bahkan, ketika logaritma data simetris, *geometric mean* adalah estimasi median. Ini karena median dan *geometric mean* sama. Ketika ditransformasikan kembali ke satuan asli, rerata geometris terus menjadi estimasi untuk median, tetapi bukan merupakan estimasi untuk rerata.

Pada R *geometric mean* dapat kita hitung menggunakan sintaks fungsi yang kita buat sendiri:

```
geomean <- function(x){
  y = log(x)
  GM = exp(mean(y))
  return(GM)
}
```

Data pada Tabel 6.1 merupakan data dengan kemencengan positif. Nilai *geometric mean* data tersebut dihitung menggunakan sintaks berikut:

```
geomean(sungai$debit)
```

```
## [1] 112.4
```

Berdasarkan hasil komputasi diperoleh nilai *geometric mean* debit sungai sebesar $112.4315\text{ m}^3/\text{detik}$. Nilai yang diperoleh tidak berbeda dengan nilai median sebesar $117\text{ m}^3/\text{detik}$.

Kompromi antara median dan mean tersedia dengan memotong beberapa observasi terendah dan tertinggi, dan menghitung mean dari apa yang tersisa. Perkiraan pemusatan data seperti itu tidak dipengaruhi oleh observasi yang paling ekstrem (dan mungkin anomali), seperti mean. Namun mereka memungkinkan besarnya sebagian besar nilai untuk mempengaruhi estimasi, tidak seperti median. Estimator ini disebut “*trimmed mean*”, dan persentase data yang diinginkan dapat dipangkas. Pemangkasan yang paling umum adalah menghapus 25 persen dari data di setiap ujung - rata-rata yang dihasilkan dari 50 persen pusat data biasanya disebut “*trimmed mean*”, tetapi lebih tepatnya 25 persen *trimmed mean*. “*trimmed mean 0%*” adalah mean sampel itu sendiri, sementara memangkas semua kecuali 1 atau 2 nilai pusat menghasilkan median. Persentase pemangkasan harus secara eksplisit dinyatakan saat digunakan. *Trimmed mean* adalah estimator yang resistan, karena tidak sangat dipengaruhi oleh *outlier*, dan bekerja dengan baik untuk berbagai macam bentuk distribusi (normal, lognormal, dll). Ini dapat dianggap sebagai rata-rata tertimbang (*weighted mean*), di mana data di luar ‘jendela’ cutoff diberi bobot 0, dan mereka yang berada di dalam jendela bobot 1,0 (lihat Gambar 6.6).

Pada R *trimmed mean* dapat dihitung dengan spesifikasi argumen `trim` pada fungsi `mean()`. Pada data debit sungai (Tabel 6.1) dihitung *trimmed mean* dengan data yang dipangkas adalah 5% di kedua ujung observasi atau `trim=0.1`.

```
mean(sungai$debit, trim=0.1)
```

```
## [1] 113.2
```

Nilai yang diperoleh sekarang mendekati nilai median dan *geometric mean* yaitu sebesar $113.2 \text{ m}^3/\text{detik}$.

6.2 Ukuran Sebaran Data

Saat kita mengetahui kedalaman rata-rata sungai, kita pasti ingin mengetahui berapa interval atau variasi dari kedalamannya. Kita tidak cukup hanya dengan mengetahui nilai pemasaran datanya saja, kita juga perlu mengetahui seberapa besar variasi atau variabilitas datanya.

Variabilitas suatu data diukur dengan melihat sebaran data dari nilai rata-ratanya (mean). Semakin besar sebaran suatu data, semakin tidak berarti nilai rata-ratanya karena nilai rata-ratanya bisa sangat berbeda dari sejumlah nilai pada datanya.

6.2.1 Pengukuran Klasik (Varian dan Simpangan Baku)

Varian sampel dan nilai akar dari varian sampel (Simpangan Baku) merupakan ukuran penyebaran data klasik. Sama dengan mean varian dan simpangan baku dipengaruhi oleh *outlier*. Semakin besar nilai keduanya, semakin besar variabilitas datanya. Kedua ukuran tersebut dinyatakan pada Persamaan (6.9) dan Persamaan (6.10).

Varian Sampel

$$s^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{(n-1)} \quad (6.9)$$

simpangan baku

$$s = \sqrt{s^2} \quad (6.10)$$

Kedua nilai tersebut dihitung berdasarkan kuadrat deviasi nilai observasi dari rata-ratanya, sehingga jika pada data terdapat *outlier* maka nilai outlier akan memperbesar deviasi data dari nilai mean. Ketika *outlier* hadir, pengukuran menjadi tidak stabil. Hal ini akan memberi kesan sebaran data menjadi jauh lebih besar daripada yang ditunjukkan oleh mayoritas nilai pada data.

Varian dan simpangan baku pada R dihitung menggunakan fungsi `var()` (varian) dan `sd()`. Format yang digunakan adalah sebagai berikut:

```
var(x, na.rm = FALSE)
sd(x, na.rm = FALSE)
```

Note:

- **x**: objek atau vektor numerik.
- **na.rm**: nilai logis yang menyatakan apakah *missing value* perlu disertakan dalam komputasi atau tidak.

Analisa Varian dan simpangan baku Grup Tunggal

Kita akan menggunakan kembali data pada Tabel 6.1 untuk menghitung varian dan simpangan baku data tersebut. Berikut adalah sintaks untuk melakukannya:

```
# varian data sungai
var(sungai$debit)

## [1] 11926

# simpangan baku data sungai
sd(sungai$debit)
```

```
## [1] 109.2
```

Sekarang mari kita bandingkan dengan data yang tidak menyertakan outlier.

```
# varian data sungai
var(sungai_subset$debit)

## [1] 1919

# simpangan baku data sungai
sd(sungai_subset$debit)

## [1] 43.8
```

Berdasarkan hasil yang diperoleh terlihat bahwa nilai varian dan simpangan baku data dengan *outlier* jauh lebih besar dibanding data tanpa *outlier*.

Analisa Varian dan simpangan baku Multi Grup

Pada contoh ini kita akan menggunakan kembali data pada Tabel 6.2. Sintaks berikut adalah cara menghitung varian dan simpangan baku untuk data berkelompok:

```
data_gw %>%
  group_by(Bicarbonate) %>%
  summarize(var_TDS=var(TDS), var_Uranium=var(Uranium),
           sd_TDS=sd(TDS), sd_Uranium=sd(Uranium))

## # A tibble: 2 x 5
##   Bicarbonate var_TDS var_Uranium sd_TDS sd_Uranium
##   <fct>        <dbl>      <dbl>    <dbl>     <dbl>
## 1 0            79471.     13.0     282.      3.61
## 2 1            10559.     13.5     103.      3.68
```

Jika kita perhatikan nilai varian dan simpangan baku Uranium pada dua kondisi kesadahan memiliki nilai yang nyaris sama. Hal sebaliknya terjadi pada variabel TDS yang menunjukkan perbedaan pada dua ukuran sebaran datanya. TDS pada kesadahan >50% memiliki varian dan simpangan baku yang lebih kecil dibanding kondisi kesadahan satunya, yang menunjukkan data pada kondisi kesadahan >50% lebih tidak tersebar dibanding kesadahan satunya.

6.2.2 Ukuran Sebaran Data yang Resisten Terhadap Outlier

Simpangan kuartil atau *interquartile range* (IQR) merupakan ukuran sebaran data yang resisten dan paling sering digunakan. IQR mengukur kisaran 50% pusat data sehingga pengukuran tidak dipengaruhi oleh adanya outlier pada 25% pada data pada setiap ujungnya. Untuk visualisasinya kita dapat melihat kembali pada ambar 6.6.

IQR didefinisikan sebagai persentil ke-75 dikurangi dengan persentil ke-25. Persentil ke-75, ke-50 (median) dan ke-25 membagi data menjadi empat tempat berukuran sama. Persentil ke-75 (P_{75}), juga disebut kuartil atas, adalah nilai yang melebihi tidak lebih dari 75% data dan dilampaui oleh tidak lebih dari 25 persen data. Persentil ke-25 (P_{25}) atau kuartil lebih rendah adalah nilai yang melebihi tidak lebih dari 25% dari data dan dilampaui oleh tidak lebih dari 75%. Dengan mempertimbangkan data yang telah diurutkan dari yang terkecil ke yang terbesar: $X_i, i = 1, \dots, n$. Persentil (P_j) dihitung berdasarkan Persamaan (6.11).

$$P_j = X_{(n+1) \cdot j} \quad (6.11)$$

dimana n merupakan ukuran sampel X_j , dan j merupakan fraksi data yang kurang dari atau sama dengan nilai persentil (untuk persentil ke-25, 50, dan 75, $j = .25, .50, \text{ dan } .75$).

Pada R, IQR dapat dihitung secara langsung menggunakan fungsi `IQR()` atau secara tidak langsung menggunakan fungsi `quantile()`. Penggunaan fungsi `quantile()` digunakan untuk mencari persentil dari data. Telah dijelaskan sebelumnya bahwa IQR merupakan selisih dari persentil 75 dan persentil 25. Format yang digunakan untuk menghitung IQR adalah sebagai berikut:

```
# secara langsung
IQR(x, na.rm=FALSE)

# secara tidak langsung
quantile(x, 3/4)-quantile(x, 1/4)

# atau
quantile(x, .75)-quantile(x, .25)
```

Note:

- **x**: objek atau vektor numerik.
- **na.rm**: nilai logis yang menyatakan apakah *missing value* perlu disertakan dalam komputasi atau tidak.

Pada Tabel 6.1, kita dapat menghitung IQR dari data. Berikut adalah contoh sintaks yang digunakan:

```
IQR(sungai$debit)
```

```
## [1] 72.5
```

Salah satu penaksir penyebaran yang resisten selain IQR adalah *Median Absolute Deviation*, atau MAD. MAD dihitung dengan pertama-tama mendaftar nilai absolut dari semua selisih $|d|$ antara masing-masing pengamatan dan median. Median dari nilai absolut ini adalah MAD yang ditulis berdasarkan Persamaan (6.12).

$$MAD (X_i) = \text{median } |d| \quad (6.12)$$

dimana

$$d_i = X_i - \text{median}(X_i) \quad (6.13)$$

Pada R, MAD tidak dapat dihitung secara langsung. Kita perlu membuat *user defined function* untuk dapat digunakan sewaktu-waktu. Berikut adalah fungsi yang dibuat:

```
MAD <- function(x){
  # median data
  m = median(x)
  # MAD
  d = abs(x-m)
  mad = mean(d)
  # print
  return (mad)
}
```

Pada Tabel 6.1, kita dapat menghitung MAD dari data menggunakan fungsi yang telah dibuat. Berikut adalah contoh sintaks yang digunakan:

```
MAD(sungai$debit)
```

```
## [1] 60.42
```

6.3 Ringkasan Data Menggunakan Fungsi `summary()` dan `stat.desc()`

Ringkasan data menggunakan fungsi `summary()` akan memberikan ringkasan data seperti nilai mean, kuartil, nilai minimum dan maksimum, serta *missing value*. Jika data berupa variabel tunggal maka output yang dihasilkan berupa nilai-nilai yang telah penulis sebutkan sebelumnya. Berikut adalah contoh sintaks yang digunakan:

```
summary(sungai$debit)
```

```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##   50.0   67.2  117.0  136.6  139.8  457.0
```

Jika objek yang diinputkan kedalam fungsi tersebut adalah data frame, maka ringkasan data akan diberikan pada setiap kolom dengan ketentuan berikut:

- jika kolom berupa variabel numerik maka output yang diperoleh berupa mean, median, min, max dan kuartil.
- jika kolom berupa factor maka output yang dihasilkan berupa rekapan jumlah observasi pada masing-masing grup.

Berikut adalah contoh sintaks penerapannya:

```
summary(data_gw)
```

```
##          TDS      Uranium   Bicarbonate
##  Min.    : 255   Min.    : 0.147   0:23
##  1st Qu.: 323   1st Qu.: 1.558   1:21
##  Median  : 560   Median   : 3.093
##  Mean    : 626   Mean    : 4.276
##  3rd Qu.: 853   3rd Qu.: 5.807
##  Max.    :1291   Max.    :14.634
```

Ringkasan data lain dapat dilakukan dengan menggunakan fungsi `stat.desc()` dari library `pastecs`. Kelebihan dari ringkasan data menggunakan fungsi ini adalah kita tidak hanya memperoleh ringkasan data dengan output seperti diatas, namun kita juga memperoleh output berupa nilai *standard error* (SE), *confidence interval* (CI), dan koefisien variasi (coef.var) yang merupakan hasil bagi dari simpangan baku dibagi dengan nilai rata-rata.

Berikut adalah sintak yang digunakan untuk menghasilkan ringkasan data menggunakan fungsi `stat.desc()`:

```
# memasang paket
install.packages("pastecs")
```

```
# memuat paket
library(pastecs)
```

```
## Warning: package 'pastecs' was built under R version
## 3.5.3
```

```
# ringkasan data
stat.desc(data_gw)
```

```
##          TDS      Uranium   Bicarbonate
##  nbr.val    4.400e+01  44.0000      NA
##  nbr.null   0.000e+00  0.0000      NA
##  nbr.na     0.000e+00  0.0000      NA
##  min        2.552e+02  0.1473      NA
##  max        1.291e+03 14.6342      NA
##  range      1.035e+03 14.4869      NA
##  sum        2.753e+04 188.1604      NA
##  median     5.602e+02  3.0934      NA
##  mean       6.257e+02  4.2764      NA
##  SE.mean    4.986e+01  0.5572      NA
##  CI.mean.0.95 1.005e+02  1.1238      NA
##  var        1.094e+05 13.6623      NA
##  std.dev    3.307e+02  3.6963      NA
##  coef.var   5.286e-01  0.8643      NA
```

6.4 Ukuran Kemencengan Data

Ketika data memiliki kemencengan, nilai mean tidak sama dengan median, tetapi bergeser ke arah ekor distribusi. Jadi untuk kemencengan positif, nilai mean melebihi lebih dari 50% dari data, seperti pada Gambar 6.7 dan Gambar 6.8. Simpangan baku juga meningkat dengan data di bagian ekor. Data yang menceng juga mempertanyakan penerapan tes hipotesis yang didasarkan pada asumsi bahwa data memiliki distribusi normal. Tes-tes ini, yang disebut tes parametrik, mungkin bernilai dipertanyakan ketika diterapkan pada data seperti data sumber daya air, karena data seringkali tidak normal atau bahkan simetris.

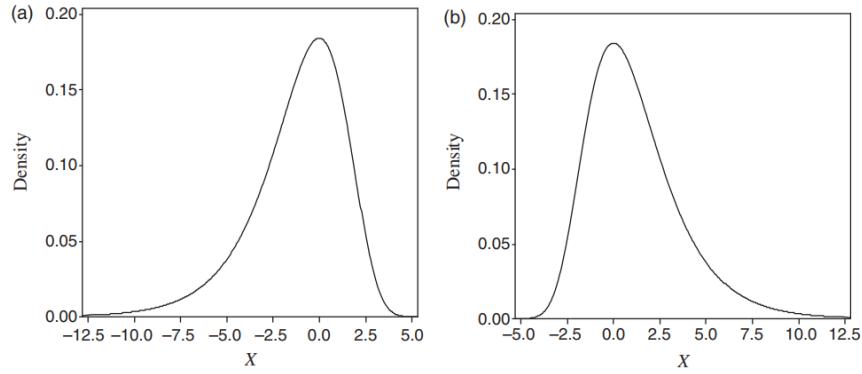


Figure 6.7: a) Kemencengan negatif, b) Kemencengan positif.

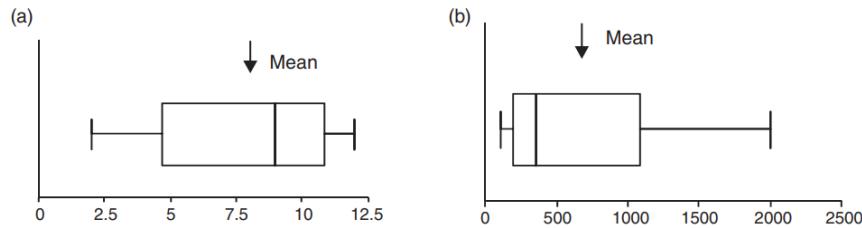


Figure 6.8: Box plot untuk data dengan a) Kemencengan negatif, b) Kemencengan positif.

6.4.1 Ukuran Kemencengan Klasik

Koefisien kemencengan (g) merupakan ukuran kemencengan yang sering digunakan. Koefisien kemencengan dituliskan pada Persamaan (6.14).

$$g = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \frac{(x_i - \bar{X})^3}{s^3} \quad (6.14)$$

Kemencengan positif (ekor panjang kekanan) memiliki nilai g positif sedangkan kemencengan negatif (ekor panjang kekiri) memiliki nilai g negatif. Sekali lagi, Pengaruh beberapa *outlier* adalah penting - suatu distribusi simetris yang memiliki satu *outlier* akan menghasilkan ukuran kemencengan (g) yang besar (dan mungkin menyesatkan).

Pada R kita dapat menghitung sendiri koefisien kemencengan (g) menggunakan *user define function*. Berikut adalah contoh sintaks fungsi yang dibuat:

```
skew <- function(x){
  ave = mean(x)
  n = length(x)
  sd = sd(x)
  g=(n/((n-1)*(n-2)))*sum(((x-ave)^3)/(sd^3))
  return(g)
}
```

Pada contoh sebelumnya dengan menggunakan fungsi yang telah dibuat diperoleh koefisien kemencengan sebagai berikut:

```
skew(data_gw$Uranium)
```

```
## [1] 1.184
```

6.4.2 Ukuran Kemencengan yang Resisten

Ukuran kemencengan yang lebih resisten adalah *quartile skew coefficient* (qs). Merupakan ukuran kemenengan didasarkan pada ketiga nilai kuartil data seperti yang ditunjukkan pada Persamaan (6.15) yang menyatakan perbedaan pada jarak kuartil atas dan bawah terhadap median dibagi dengan IQR.

$$qs = \frac{(P_{.75} - P_{.50}) - (P_{.75} - P_{.25})}{P_{.75} - P_{.25}} \quad (6.15)$$

Kemencengan positif akan memiliki nilai qs positif dan begitupun sebaliknya. Pada R kita dapat menghitung nilai qs menggunakan *user define function*. Berikut adalah contoh sintaks fungsi yang dibuat:

```
qs <- function(x){
  p75 = quantile(x, 3/4)
  p50 = median(x)
  p25 = quantile(x, 1/4)
  skew = ((p75-p50)-(p50-p25))/(p75-p25)
  return(skew)
}
```

Pada contoh sebelumnya dengan menggunakan fungsi yang telah dibuat diperoleh koefisien kemencengan sebagai berikut:

```
qs(data_gw$Uranium)
```

```
##      75%
## 0.2772
```

6.5 Outlier

Outlier merupakan pengamatan yang nilainya sangat berbeda dari yang lain dalam kumpulan data, sering menimbulkan kekhawatiran atau alarm. Meskipun sebenarnya kita tidak perlu khawatir dengan adanya *outlier*. *Outlier* sering ditangani dengan membuangnya sebelum mendeskripsikan data, atau sebelum beberapa prosedur uji hipotesis chapter-chapter selanjutnya. Sekali lagi, mereka seharusnya tidak perlu dikhawatirkan. *Outlier* mungkin merupakan poin paling penting dalam kumpulan data dan harus diselidiki lebih lanjut.

Untuk lebih memahami kenapa *outlier* begitu penting pada data kita berikut merupakan contoh kasus dari asal kata *outlier*. Misalkan bahwa data pada “lubang” ozon Antartika, suatu daerah dengan konsentrasi ozon yang sangat rendah, telah dikumpulkan selama kurang lebih 10 tahun sebelum penemuan aktualnya. Namun, rutinitas pengecekan data otomatis selama pemrosesan data menyertakan instruksi untuk menghapus “*outlier*”. Definisi *outlier* didasarkan pada konsentrasi ozon yang ditemukan pada pertengahan garis lintang. Dengan demikian semua data yang tidak biasa ini tidak pernah dilihat atau dipelajari selama beberapa waktu. Jika *outlier* dihapus, risiko diambil hanya dengan melihat apa yang diharapkan dilihat. Jika hal tersebut dilakukan maka anomali yang terjadi pada atmosfer dapat luput kita pelajari.

Berdasarkan kasus tersebut kita perlu dengan baik mempertimbangkan apakah *outlier* pada data perlu dihapus atau tidak. Jika berkaitan dengan pembuatan model, penghapusan *outlier* merupakan sesuatu yang

dapat memperbaiki akurasi dari model. Namun, pada sebuah penelitian terkadang diperlukan informasi lebih lanjut mengapa terdapat *outlier* pada data sehingga kita dapat memperoleh pengetahuan baru dari proses pencarian tersebut.

Outlier dapat terjadi karena tiga hal, yaitu:

1. Kesalahan pengukuran atau perekaman data.
2. Observasi dari populasi tidak sama dengan sebagian besar data seperti misalnya data debit banji akibat jebolnya sebuah bendungan akan berbeda dengan debit banjir akibat presipitasi.
3. Kejadian langka pada sebuah populasi yang sedikit memiliki kemencengangan pada distribusinya.

Metode grafis seperti box plot sangat membantu dalam mengidentifikasi *outlier*. Setiap kali *outlier* terjadi, pertama-tama verifikasi bahwa tidak ada penyalinan, titik desimal, atau kesalahan nyata lainnya yang telah dibuat. Jika tidak, tidak mungkin untuk menentukan apakah titik itu valid. Upaya yang dilakukan untuk verifikasi, seperti menjalankan kembali sampel di laboratorium, akan tergantung pada manfaat yang diperoleh versus biaya verifikasi. Kejadian masa lalu mungkin tidak dapat diduplikasi. Jika tidak ada kesalahan yang dapat dideteksi dan diperbaiki, ** *outlier* tidak boleh dibuang hanya berdasarkan fakta bahwa mereka tampak tidak biasa**. *Outlier* sering dibuang untuk membuat data cocok dengan distribusi teoretis yang sudah terbentuk sebelumnya seperti distribusi normal. Tidak ada alasan untuk menganggap bahwa mereka seharusnya dibuang! Seluruh rangkaian data dapat muncul dari distribusi yang memiliki kemencengangan, dan mengambil logaritma atau transformasi lain dapat menghasilkan data yang cukup simetris. Bahkan jika tidak ada transformasi yang mencapai simetri, *outlier* tidak perlu dibuang. Daripada menghilangkan data aktual (dan mungkin sangat penting) untuk menggunakan prosedur analisis yang membutuhkan simetri atau normalitas, prosedur yang tahan terhadap *outlier* harus digunakan. Jika menghitung rata-rata tampak bernilai kecil karena *outlier*, median telah terbukti menjadi ukuran lokasi yang lebih tepat untuk data yang memiliki kemencengangan. Jika melakukan uji-t (dijelaskan pada chapter selanjutnya) tampaknya tidak valid karena set data yang tidak normal, gunakan *rank-sum test* sebagai gantinya.

Singkatnya, biarkan panduan data prosedur analisis yang digunakan, daripada mengubah data untuk menggunakan beberapa prosedur yang memiliki persyaratan terlalu ketat untuk situasi yang dihadapi.

6.6 Transformasi Data

Transformasi data dilakukan untuk memenuhi tiga tujuan, antara lain:

1. membuat data lebih simetris,
2. membuat data lebih linier, dan
3. membuat data memiliki varian yang konsisten.

Beberapa ilmuwan lingkungan takut bahwa dengan mentransformasikan data, hasilnya diperoleh yang sesuai dengan gagasan yang telah terbentuk sebelumnya. Oleh karena itu, transformasi adalah metode untuk **melihat apa yang ingin kita lihat** dari data. Namun dalam kenyataannya, masalah serius dapat terjadi ketika prosedur dengan asumsi simetri, linieritas, atau homoseksualitas (varians konstan) digunakan pada data yang tidak memiliki karakteristik yang diperlukan ini. Transformasi dapat menghasilkan karakteristik ini, dan dengan demikian penggunaan variabel yang diubah memenuhi tujuan.

Satu unit pengukuran tidak lebih valid secara apriori daripada yang lainnya. Sebagai contoh, logaritma negatif konsentrasi ion hidrogen (pH), sama validnya dengan sistem pengukuran dengan konsentrasi ion hidrogen itu sendiri. Transformasi seperti akar kuadrat kedalaman air pada sumur sumur, atau akar kubik volume curah hujan, seharusnya tidak mengandung stigma lebih daripada pH. Skala pengukuran ini mungkin lebih sesuai untuk analisis data daripada unit aslinya. Hoaglin (1988) telah menulis artikel yang bagus tentang transformasi tersembunyi, secara konsisten diterima begitu saja, yang umum digunakan oleh semua orang. Oktaf dalam musik adalah transformasi frekuensi logaritmik. Setiap kali piano dimainkan,

Use	θ	Transformation	Name	Comment
		\cdot		higher powers can be used
		\cdot		
for (-) skewness	3	x^3	cube	
	2	x^2	square	
	1	x	original units	no transformation
	1/2	\sqrt{x}	square root	commonly used
	1/3	$\sqrt[3]{x}$	cube root	commonly used
for (+) skewness	0	$\log(x)$	logarithm	commonly used. Holds the place of x^0
	-1/2	$-1/\sqrt{x}$	reciprocal root	the minus sign preserves order of observations
	-1	$-1/x$	reciprocal	
	-2	$-1/x^2$		
		\cdot		
		\cdot		lower powers can be used

Figure 6.9: Ladder of power

transformasi logaritmik digunakan! Begitu pula dengan skala Richter untuk gempa bumi, mil per galon untuk konsumsi bensin, f-stop untuk eksposur kamera, dll. semua menggunakan transformasi. Dalam ilmu analisis data, keputusan yang menggunakan skala pengukuran harus ditentukan oleh data, bukan dengan kriteria yang ditentukan sebelumnya. Tujuan penggunaan transformasi adalah untuk kesimetrian, linieritas, dan homoskedastisitas. Selain itu, penggunaan banyak teknik tahan seperti persentil dan prosedur uji non-parametrik (akan dibahas kemudian) tidak berbeda dengan skala pengukuran. Hasil *rank-sum test*, setara nonparametrik dari uji-t, akan persis sama apakah unit asli atau logaritma dari unit tersebut digunakan.

Untuk membuat distribusi asimetris menjadi lebih simetris, data dapat diubah atau diekspresikan kembali menjadi unit baru. Unit-unit baru ini mengubah jarak antara pengamatan pada plot garis. Efeknya adalah memperluas atau mengecilkan jarak ke pengamatan ekstrem di satu sisi median, membuatnya lebih pada setiap sisinya. Transformasi yang paling umum digunakan dalam lingkungan adalah logaritma, seperti Log debit air, konduktivitas hidrolik, atau konsentrasi sering diambil sebelum analisis statistik dilakukan.

Transformasi data biasanya melibatkan fungsi power seperti pada fungsi $y = x^\theta$, dimana x merupakan data yang belum ditransformasi, y adalah data yang telah ditransformasi, dan θ merupakan power eksponensial. Pada Gambar 6.9 nilai θ di-list kedalam "ladder of powers" (Velleman dan Hoaglin, 1981 dalam Helsel dan Hirsch, 2002), sebuah struktur yang berguna untuk menentukan nilai θ yang tepat.

Seperti yang dapat dilihat dari *ladder of powers*, setiap transformasi dengan θ kurang dari 1 dapat digunakan untuk membuat data dengan kemencengan positif lebih simetris. Dengan membuat box plot atau plot Q-Q dari data yang diubah kita dapat mengetahui apakah transformasi yang telah dilakukan sesuai. Jika transformasi logaritmik memberikan kompensasi yang berlebihan untuk kemiringan yang tepat dan menghasilkan distribusi yang sedikit kiri (kemencengan negatif), transformasi 'lebih ringan' dengan θ lebih dekat ke 1, seperti transformasi kuadrat atau akar kubik, harus digunakan. Transformasi dengan $\theta > 1$ akan membantu membuat data yang condong ke kiri lebih simetris.

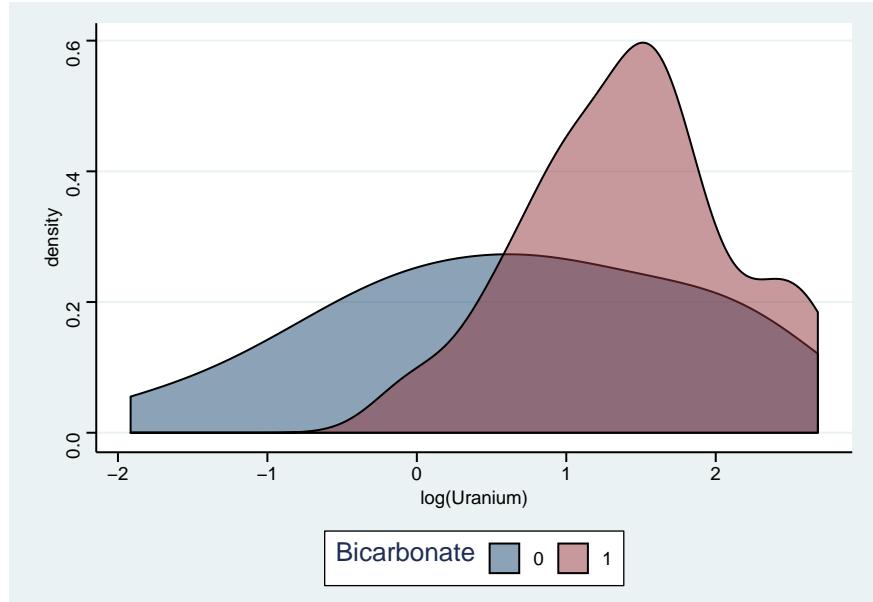


Figure 6.10: Visualisasi konsentrasi Uranium hasil transformasi pada air tanah

Namun, kecenderungan untuk mencari transformasi ‘terbaik’ harus dihindari. Misalnya, ketika berhadapan dengan beberapa set data yang serupa, mungkin lebih baik untuk menemukan satu transformasi yang bekerja cukup baik untuk semua, daripada menggunakan yang sedikit berbeda untuk masing-masingnya. Harus diingat bahwa setiap set data adalah sampel dari populasi yang lebih besar, dan sampel lain dari populasi yang sama kemungkinan akan menunjukkan transformasi ‘terbaik’ yang sedikit berbeda. Penentuan ‘terbaik’ dalam ketelitian tinggi adalah pendekatan yang jarang sepadan dengan usaha.

Pada Gambar 6.5 kosentrasi distribusi Uranium pada tiap grup memiliki kemencenggan positif. Untuk membuatnya simetris kita perlu melakukan transformasi yang sesuai jenis transformasi yang dilakukan dapat dimulai dari akar kuadrat sampai invers akar kuadrat (berdasarkan Gambar 6.9). Pada contoh ini kita akan mencoba melakukan trasnformasi logaritmik. Berikut adalah contoh visualisasi hasil transformasinya (lihat Gambar 6.10):

Berdasarkan hasil transformasi, kita telah memperoleh ditribusi yang cukup simetris untuk kedua grup data tersebut. Pembaca dapat mencobanya menggunakan transformasi lainnya sendiri.

6.7 Referensi

1. Damanhuri, E. 2011. **Statitika Lingkunga**. Penerbit ITB.
2. Helsel, D.R., Hirsch, R.M. 2002. **statistical Methods in Water Resources**. USGS.
3. Ofungwu, J. 2014. **Statistical Applications For Environmental Analysis and Risk Assessment**. John Wiley & Sons, Inc.
4. Rosadi, D. 2015. **Analisis Statistika dengan R**. Gadjah Mada University Press.
5. STHDA. **Descriptive Statistics and Graphics**. <http://www.sthda.com/english/wiki/descriptive-statistics-and-graphs>

Table 6.2: Kosentrasi TDS dan Uranium dalam berbagai kondisi kesadahan

TDS	Uranium	Bicarbonate
682.6	0.9315	0
819.1	1.9380	0
303.8	0.2919	0
1151.4	11.9042	0
582.4	1.5674	0
1043.4	2.0623	0
634.8	3.8858	0
1087.2	0.9772	0
1123.5	1.9354	0
688.1	0.4367	0
1174.5	10.1142	0
599.5	0.7551	0
1240.8	6.8559	0
538.4	0.4806	0
607.8	1.1452	0
705.9	6.0876	0
1290.6	10.8823	0
526.1	0.1473	0
784.7	2.6741	0
953.1	3.0918	0
1149.3	0.7592	0
1074.2	3.7101	0
1116.6	7.2446	0
301.2	5.7129	1
265.4	4.7366	1
295.9	2.8057	1
442.4	5.6290	1
342.7	3.0950	1
361.3	3.5774	1
262.1	1.7711	1
546.2	11.2724	1
273.9	4.9807	1
281.4	4.0833	1
588.9	14.6342	1
574.1	12.3835	1
307.1	1.5291	1
409.4	4.4647	1
327.1	2.4574	1
425.7	6.3042	1
310.1	4.5441	1
289.8	0.9672	1
408.2	2.1568	1
383.0	8.3810	1
255.2	2.7957	1

Chapter 7

Ekplorasi Data Menggunakan Grafik

Pada Chapter 4 dan 5 kita telah belajar bagaimana cara membuat grafik menggunakan R. Sejauh ini kita belum belajar kegunaan dari masing-masing grafik yang telah kita pelajari. Pada Chapter ini kita tidak lagi akan membahas bagaimana membuat grafik menggunakan R. Kita akan fokus terhadap fungsi grafik tersebut dalam analisa kita. Secara umum grafik dibuat untuk memvisualisasikan distribusi, perbedaan antar sampel, korelasi dan asosiasi antar sampel, serta ukuran sampel.

Penulis dan pembaca pasti sepakat bahwa visualisasi data merupakan tahapan awal yang perlu kita lakukan sebelum memutuskan untuk melakukan analisa data seperti uji hipotesis dan modeling. Angka yang ditampilkan dalam ringkasan data tidaklah cukup untuk melihat data terutama kaitannya dengan pengecekan terhadap asumsi model.

Pada Gambar 7.1 disajikan delapan buah scatterplot dengan koefisien korelasi yang sama persis. Komputasi statistik tanpa melihat pada visualisasi data akan menyebabkan misinterpretasi pada data. Grafik memberikan ringkasan visual data dengan cepat dan lengkap dibandingkan penyajian data dalam tabel angka.

Grafik sangat penting untuk dua tujuan:

1. untuk memberikan wawasan bagi analis ke dalam data di bawah pengawasan, dan
2. untuk mengilustrasikan konsep-konsep penting ketika mempresentasikan hasil kepada orang lain.

Tugas pertama disebut **Analisis Data Eksplorasi (EDA)**, dan merupakan subjek Chapter ini. Prosedur EDA seringkali merupakan (atau seharusnya) menjadi ‘pandangan pertama’ pada data. Pola dan teori tentang bagaimana sistem berperilaku dikembangkan dengan mengamati data melalui grafik. Ini adalah prosedur induktif - data dirangkum dibanding dilakukan pengujian. Hasil mereka memberikan panduan untuk pemilihan prosedur pengujian hipotesis deduktif yang tepat.

Setelah analisis selesai, temuan harus dilaporkan kepada orang lain. Apakah laporan tertulis atau presentasi lisan, analis harus meyakinkan audiens bahwa kesimpulan yang dicapai didukung oleh data. Tidak ada cara yang lebih baik untuk melakukan ini selain melalui grafik. Banyak metode grafis yang sama yang merangkum informasi dengan ringkas untuk analis juga akan memberikan wawasan tentang data untuk pembaca atau audiens.

7.1 Grafik Untuk Melihat Ditrbusi Data

Analisis yang umumnya dilihat pada distribusi data adalah apakah data berdistribusi normal atau tidak. Hal ini akan mempengaruhi jenis analisis statistika yang digunakan pada data. Terdapat beberapa grafik yang dapat digunakan untuk melihat bentuk ditribusi data. Grafik-grafik tersebut antara lain: *stem and*

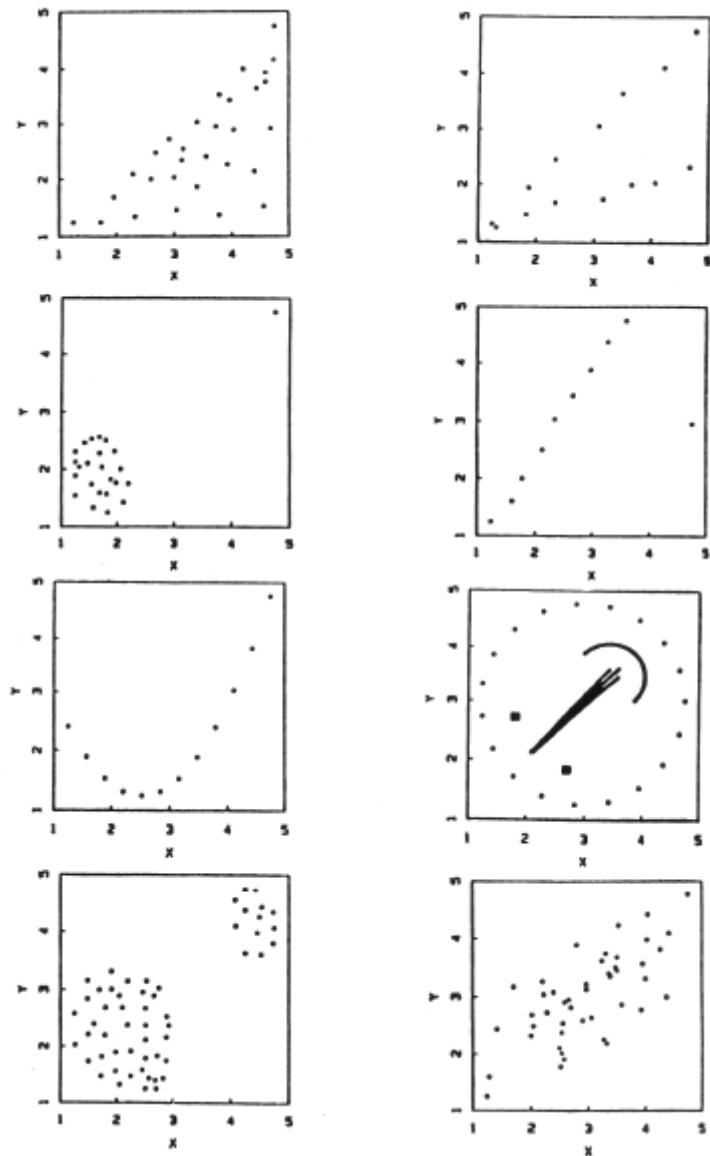


Figure 7.1: Scatterplot dengan koefisien korelasi $r=0,7$.

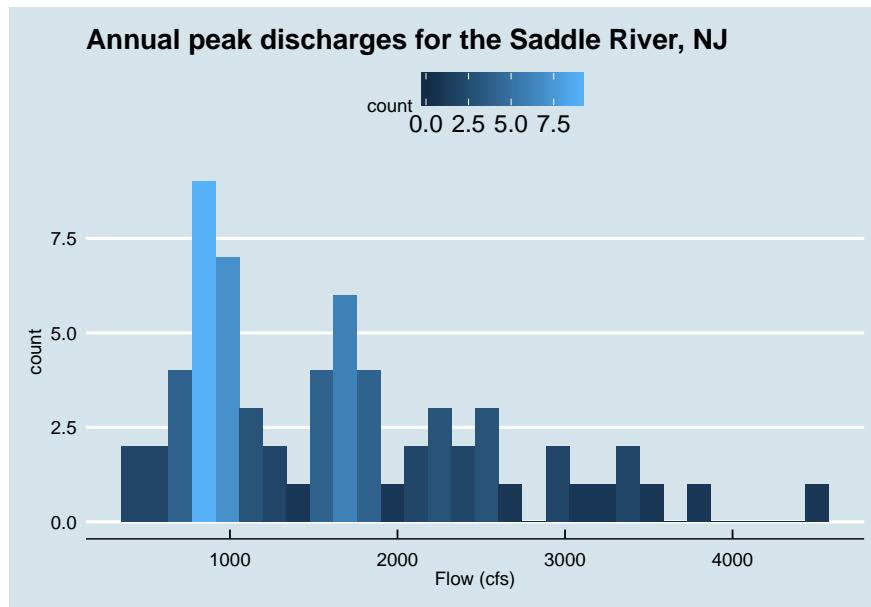


Figure 7.2: Histogram dengan bin.width=default debit sungai Saddle

leaf, histogram, density plot, QQ-plot, serta box plot atau violin plot. Pada analisis distribusi *stem and leaf* kurang populer untuk digunakan. Hal ini disebabkan karena visualisasinya kurang cocok diterapkan pada data dengan jumlah observasi besar. Selain itu, kita juga tidak bisa melakukan perbandingan antar grup menggunakan jenis visualisasi tersebut.

7.1.1 Histogram

Histogram adalah grafik yang sudah dikenal, dan konstruksinya dirinci dalam berbagai teks pengantar tentang statistik. Batang digambar dengan tinggi n_i , atau fraksi n_i/n , dari data yang termasuk dalam salah satu dari beberapa kategori atau interval (Gambar 7.2). Iman dan Conover (1983) mengemukakan bahwa untuk ukuran sampel n , jumlah interval k harus bilangan bulat terkecil sehingga $2^k n$.

Histogram sangat berguna untuk menggambarkan perbedaan besar dalam bentuk data seperti apakah data simetris seperti distribusi normal atau memiliki kemencengan. Histogram tidak dapat digunakan untuk penilaian yang lebih tepat karena tampilan dipengaruhi oleh jumlah batang yang digunakan. Untuk lebih memahaminya perhatikan Gambar 7.2 dan Gambar 7.3. Kedua histogram tersebut tampak berbeda meskipun data input yang diberikan sama. Pada Gambar 7.2 kita akan melihat bahwa debit dengan kejadian terbanyak terjadi pada rentang 800-900 cfs, sedangkan pada Gambar 7.3 kita melihat bahwa debit dengan kejadian terbanyak terjadi pada 800-1200 cfs.

```
# memuat library
library(readxl)
library(ggplot2)
library(ggthemes)

# memuat data excel
sungai <- read_excel("hhappc.xls", sheet="appc1")
```

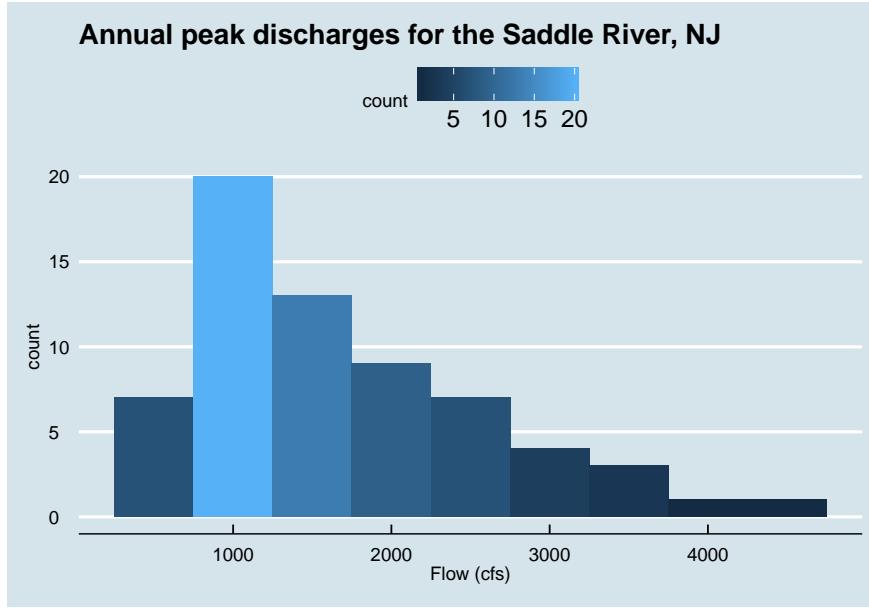


Figure 7.3: Histogram dengan bin.width=500 debit sungai Saddle

7.1.2 Density Plot

Density plot memecahkan masalah yang dimiliki histogram dalam melihat grafik dengan menyajikan data bukan dari jumlah kejadian atau observasi, namun data disajikan berdasarkan frekuensi relatif data (density) yang digambarkan dalam bentuk *smooth curve*. Contoh density plot dapat dilijat pada Gambar 7.4. Dari grafik yang dihasilkan seakan tampak jelas bahwa distibusi data memiliki kemencengan positif dengan frekuensi relatif debit terbanyak berada pada debit 1000 cfs.

7.1.3 QQ-plot

Kita telah mempelajari sebelumnya pada Chapter 5 bahwa QQ-plot data digunakan untuk mengecek apakah data yang kita miliki berdistribusi normal atau tidak. Contoh QQ-plot dapat dilijat pada Gambar 7.5. Pada grafik yang dihasilkan terlihat bahwa data tidak berdistribusi normal. Hal ini terlihat dari sebagian observasi pada debit <1000 cfs yang tidak mengikuti garis referensi.

```
ggplot(sungai, aes(sample=Flow))+
  # qq plot
  stat_qq()+
  # garis referensi
  stat_qq_line()+
  theme_economist()
```

Bentuk lain yang dapat digunakan untuk menguji kecocokan suatu distribusi dengan distribusi normal adalah grafik ECDF. Berbeda dengan QQ-plot, grafik ini dapat digunakan untuk melakukan pengecekan distribusi secara umum.

7.1.4 Box Plot dan Violin Plot

Grafik lain yang dapat digunakan untuk menggambarkan distribusi data adalah box plot dan violin plot. Box plot memberikan cara yang simpel untuk melihat ditribusi data seperti melihat posisi sejumlah kuartil,

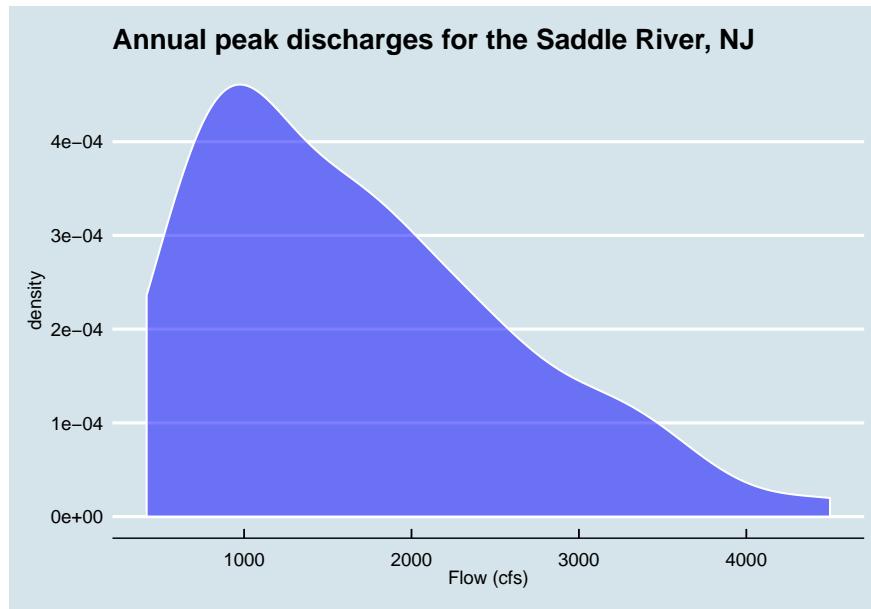


Figure 7.4: Density plot debit sungai Saddle

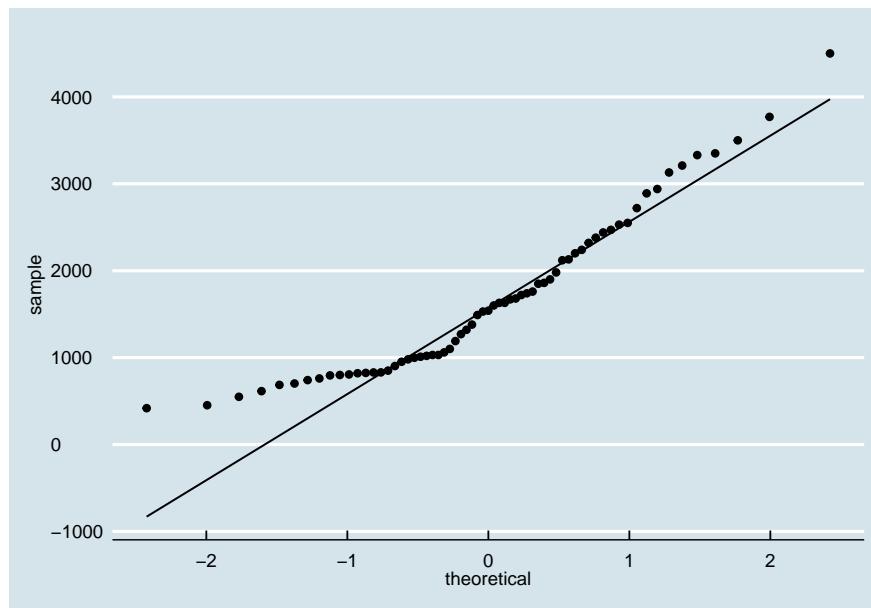


Figure 7.5: QQ plot debit sungai Saddle

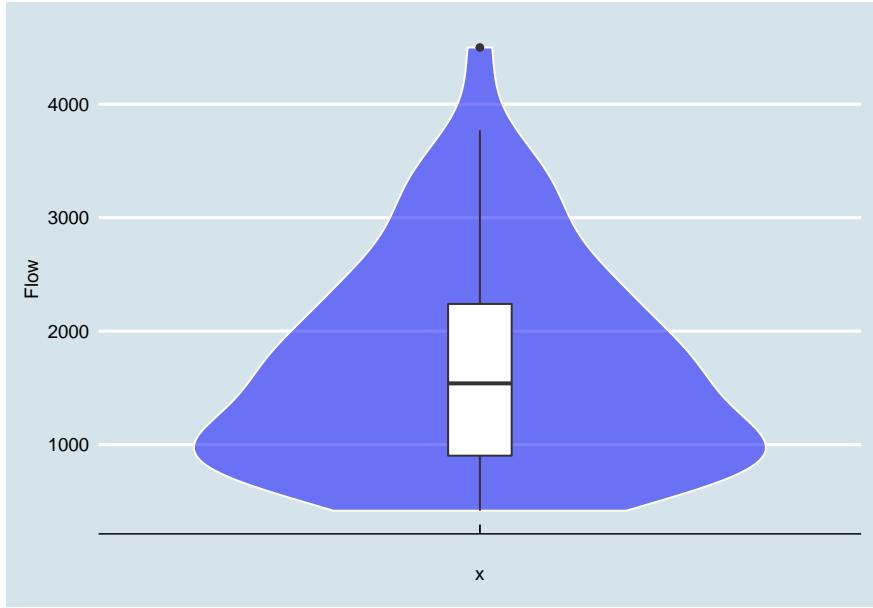


Figure 7.6: Box plot dan violin plot debit sungai Saddle

nilai minimum dan maksimum. Selain itu kita juga dapat melihat adanya outlier pada data.

Kita dapat menambah fungsionalitas dari box plot ini dengan menambahkan violin plot. Pada Chapter 5 kita telah belajar bahwa kita dapat menambahkan box plot pada violin plot atau sebaliknya sehingga memudahkan dalam mendeskripsikan bentuk distribusi data. Jika dengan box plot kita tidak dapat melihat secara baik bentuk dari data yang sesungguhnya karena hanya menampilkan lokasi sejumlah kuartil. Pada violin plot kita dapat melihat bentuk data yang ada melalui tampilan dua denisty plot (tampak seperti biola) yang digambarkan. Kekurangannya adalah kita tidak dapat melihat observasi mana yang menjadi outlier, sehingga kedua grafik ini biasa digambarkan secara bersamaan. Berikut adalah contoh box plot dan violin plot dari data debit sungai Saddle (Gambar 7.6).

```
ggplot(sungai, aes(x="", y=Flow))+
  geom_violin(fill="blue", alpha=0.5, color="white")+
  geom_boxplot(width=0.1)+
  theme_economist()
```

Berdasarkan grafik yang dihasilkan pada Gambar 7.6 kita dapat melihat bahwa ditribusi data debit sungai memiliki kemencengan positif. Hal ini terjadi karena terdapat satu *outlier* pada data yang disebabkan karena nilai observasinya diluar dari nilai maksimum data yang ditetapkan sebagai $\max = Q3 + 1,5 * IQR$.

7.2 Grafik Untuk Melihat Beda Distribusi Data Antar Grup

Grafik yang telah dijelaskan sebelumnya seperti box plot, violin plot, histogram, dan density plot merupakan grafik yang bagus untuk memvisualisasikan beda distribusi data antar grup untuk data numerik. Untuk data berupa kategori kita dapat menggunakan bar plot. Pada penerapannya bar plot juga dapat memvisualisasikan ringkasan data seperti nilai mean dan sebarannya pada data.

Pada contoh ini penulis hanya akan memberikan contoh penerapan menggunakan box plot dan bar plot menggunakan data konsentrasi Antrazine yang diukur pada bulan Juni dan September. Untuk melakukannya kita perlu memuat data dan melakukan transformasi terhadap datanya terlebih dahulu.

```
# memuat data excel
atrazine <- read_excel("hhappc.xls", sheet="appc4")

# print
head(atrazine)

## # A tibble: 6 x 2
##   June_atrazine Sept_atrazine
##       <dbl>        <dbl>
## 1      0.38        2.66
## 2      0.04        0.63
## 3     -0.01        0.59
## 4      0.03        0.05
## 5      0.03        0.84
## 6      0.05        0.580

# transformasi data
library(tidyr)
atrazine <- gather(atrazine,
                    key="month",
                    value="concentration")

# print
head(atrazine)

## # A tibble: 6 x 2
##   month      concentration
##   <chr>          <dbl>
## 1 June_atrazine    0.38
## 2 June_atrazine    0.04
## 3 June_atrazine   -0.01
## 4 June_atrazine    0.03
## 5 June_atrazine    0.03
## 6 June_atrazine    0.05
```

Pada data konsentrasi Atrazine tersebut terdapat nilai negatif yang dalam hal ini merupakan kesalahan dalam pengukuran dari alat. Untuk membersihkannya kita dapat membuat nilai observasi tersebut menjadi NA.

```
atrazine$concentration[atrazine$concentration<0] <- NA

head(atrazine)

## # A tibble: 6 x 2
##   month      concentration
##   <chr>          <dbl>
## 1 June_atrazine    0.38
## 2 June_atrazine    0.04
## 3 June_atrazine    NA
## 4 June_atrazine    0.03
## 5 June_atrazine    0.03
## 6 June_atrazine    0.05
```

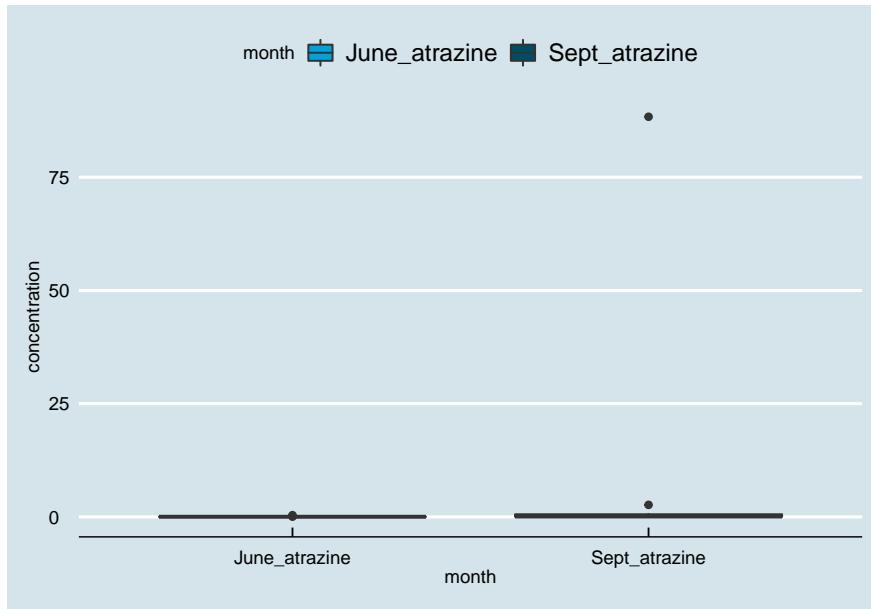


Figure 7.7: Box plot konsentrasi Atrazine pada bulan Juni dan September

Selanjutnya kita akan memvisualisasikan beda antara distribusi data pada kedua bulan menggunakan box plot (Gambar 7.7). Konsentrasi rata-rata Atrazine akan divisualisasikan menggunakan bar plot (Gambar 7.8).

```
ggplot(atrazine, aes(month, concentration, fill=month))+
  geom_boxplot()+
  theme_economist()+
  scale_fill_economist()

library(dplyr)
atrazine %>%
  group_by(month) %>%
  summarize(mean_atrazine=mean(concentration, na.rm=TRUE)) %>%
  ggplot(aes(month, mean_atrazine, fill=month))+
  geom_bar(stat="identity")+
  theme_economist()+
  scale_fill_economist()
```

Pada visualisasi yang dihasilkan terdapat perbedaan signifikan antara distribusi dan nilai rata-rata konsentrasi Atrazine pada dua periode tersebut. Hal ini disebabkan karena terdapat sebuah outlier pada periode September yang menyebabkan nilai rata-rata yang dihasilkan bergeser jauh kearah outlier. Pembaca dapat membuat visualisasi data pada data tersebut tanpa *outlier* dengan terlebih dahulu melakukan filter terhadap *outlier*.

7.3 Grafik Untuk Memvisualisasikan Korelasi Antar Variabel

Scatterplot dapat digunakan untuk memvisualisasikan korelasi antar dua variabel. Pada bagian ini akan diberikan contoh visualisasi antara variabel konsentrasi TDS dan Uranium pada air tanah.

7.4. GRAFIK YANG DIGUNAKAN UNTUK MEMVISUALISASIKAN ASOSIASI ANTAR VARIABEL

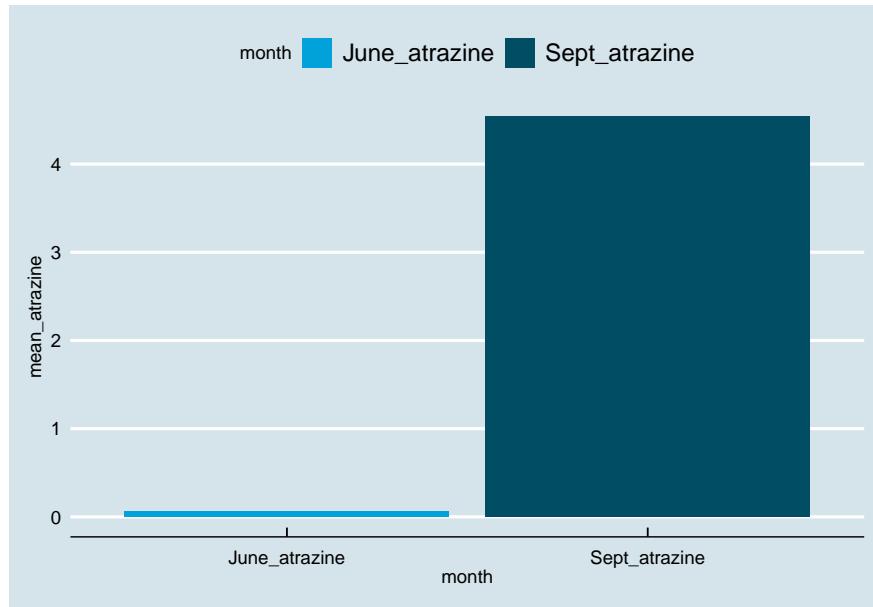


Figure 7.8: Bar plot konsentrasi Atrazine pada bulan Juni dan September

Untuk melakukannya kita perlu memuat terlebih dahulu dataset yang digunakan. Visualisasi data disajikan pada Gambar 7.9.

```
# memuat data excel
gw <- read_excel("hhappc.xls", sheet="appc16")

ggplot(gw, aes(TDS, Uranium))+
  geom_point()+
  geom_smooth(method="lm")+
  theme_economist()
```

Berdasarkan grafik yang dihasilkan terdapat hubungan linier antara konsentrasi TDS dan Uranium pada air-tanah. Meningkatnya konsentrasi TDS pada air tanah juga menyebabkan peningkatan konsentrasi Uranium pada airtanah.

7.4 Grafik Yang Digunakan Untuk Memvisualisasikan Asosiasi Antar Variabel

Asosiasi antar variabel kategori dapat dilakukan baik dengan pie chart maupun dengan bar plot. Pie chart kurang sering digunakan untuk visualisasi *multiple group* sehingga bar plot lebih sering digunakan.

Pada contoh kali ini penulis akan melihat terdapat asosiasi antara musim dan strata terhadap jumlah Corbicula di sungai Tennessee. Untuk melakukannya kita perlu memuat terlebih dahulu dataset yang digunakan. Visualisasi data disajikan pada Gambar 7.10.

```
# memuat data excel
corbicula<- read_excel("hhappc.xls", sheet="appc8")

# print
head(corbicula)
```

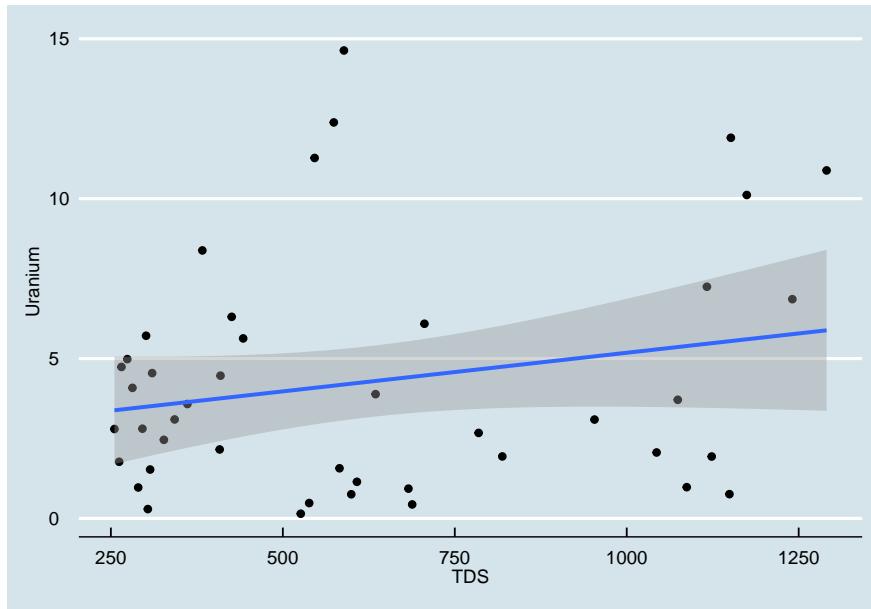


Figure 7.9: Scatterplot hubungan antara konsentrasi TDS dan Uranium pada airtanah

```
## # A tibble: 6 x 4
##   Year Season Strata Corbicula
##   <dbl> <chr>   <dbl>     <dbl>
## 1 1969 Winter     1        25
## 2 1969 Winter     1        20
## 3 1969 Winter     1        30
## 4 1969 Spring      1        9
## 5 1969 Spring      1        8
## 6 1969 Spring      1        9

corbicula %>%
  mutate(Season=as.factor(Season),
        Strata=as.factor(Strata)) %>%
  group_by(Season,Strata) %>%
  summarize(Corbicula=mean(Corbicula)) %>%
  ggplot(aes(Season, Corbicula, fill=Strata))+
  geom_bar(stat="identity",position=position_dodge2())+
  theme_economist()+
  scale_fill_economist()
```

Berdasarkan grafik yang dihasilkan terdapat pengaruh musim dan strata terhadap jumlah corbicula di sungai Tennessee. Jumlah tertinggi berada saat musim semi pada strata 3, sedangkan terendah berada pada musim dingin juga pada strata 3.

7.5 Grafik Yang Digunakan Untuk Memvisualisasikan Ukuran Sampel dan Perubahan Sepanjang Waktu

Untuk memvisualisasikan perubahan sepanjang waktu, kita dapat menggunakan line plot. Pada data corbicula kita ingin memvisualisasikan perubahan jumlah corbicula rata-rata pada setiap tahun. Visualisasi dari data disajikan pada Gambar 7.11.

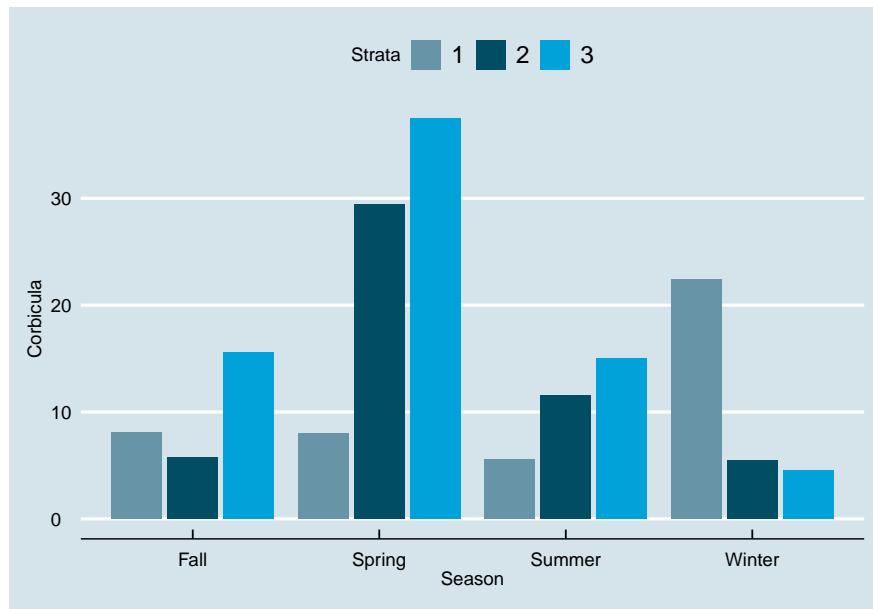


Figure 7.10: Bar plot Jumlah rata-rata corbicula pada sungai Tennessee

```
corbicula %>%
  group_by(Year) %>%
  summarize(Corbicula=mean(Corbicula)) %>%
  ggplot(aes(Year, Corbicula)) +
  geom_line() +
  geom_point(shape=1) +
  theme_economist()
```

Berdasarkan garfik yang dihasilkan dapat disimpulkan bahwa jumlah rata-rata corbicula menurun setiap tahunnya.

7.6 Referensi

1. Gardener, M. 2012. **Statistics for Ecologists Using R and Excel**-Data collection, exploration, analysis and presentation. Pelagic Publishing.
2. Helsel, D.R., Hirsch, R.M. 2002. **statistical Methods in Water Resources**. USGS.
3. Ofungwu, J. 2014. **Statistical Applications For Environmental Analysis and Risk Assessment**. John Wiley & Sons, Inc.
4. Peck, R.Devore, J.L. 2012. **Statistics The Exploration & Analysis of Data-** Seventh Edition. Brooks/Cole.

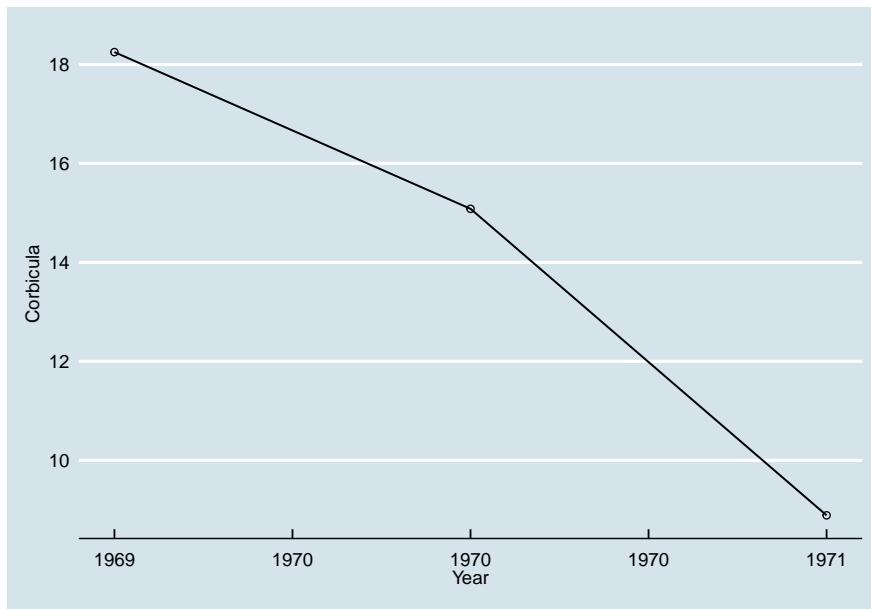


Figure 7.11: Line plot perubahan jumlah rata-rata corbicula di sungai Tennessee

Probabilitas dan Distribusi Probabilitas

Chapter 8

Probabilitas

Probabilitas merupakan kemungkinan suatu peristiwa akan terjadi. Probabilitas memiliki rentang nilai dari 0 sampai dengan 1. Probabilitas 0 artinya suatu peristiwa (*event*) mustahil atau tidak pernah terjadi, sedangkan probabilitas 1 menunjukkan suatu peristiwa yang selalu terjadi.

Contoh sederhana dari probabilitas dalam kehidupan sehari-hari adalah ketika kita melempar koin ke udara untuk melihat kemungkinan sisi yang akan tampak saat koin tersebut jatuh ke tanah. Peristiwa yang mungkin akan terjadi adalah mata uang akan menampilkan sisi depan (*head*) atau sisi belakang (*tail*). Kemungkinan untuk mendapatkan *tail* maupun *head* adalah sama yaitu 0,5.

Pada contoh pelemparan koin, kita misalkan kejadian munculnya sisi *head* adalah A , sedangkan peluang munculnya selain sisi *head* (sisi lainnya) adalah A' . Secara sederhana peluang munculnya suatu kejadian A pada contoh tersebut dapat dituliskan kedalam Persamaan (8.1) dan Persamaan (8.1).

$$P(A) + P(A') = 1 \quad (8.1)$$

dimana

$$P_A = \frac{\text{Jumlah peristiwa } A}{\text{Jumlah peristiwa yang mungkin terjadi}} \quad (8.2)$$

Pada contoh pelemparan koin, kita ingin mengetahui peluang munculnya *head* pada pelemparan koin. Jumlah peristiwa yang mungkin terjadi saat pelemparan koin ada 2 yaitu munculnya *head* atau *tail*. Peluang munculnya sisi *head* dapat dihitung menggunakan Persamaan (8.1) seperti berikut:

$$P_{\text{head}} = \frac{\text{Jumlah peristiwa head}}{\text{Jumlah peristiwa yang mungkin terjadi}} = \frac{1}{2} = 0,5$$

Probabilitas suatu peristiwa dapat dibedakan kedalam 3 kategori, yaitu:

1. **Probabilitas apriori:** probabilitas yang ditentukan sebelumnya tanpa perlu melakukan suatu eksperimen atau kita dapat memperkirakan sebelumnya peristiwa apa saja yang dapat terjadi. Contoh: pelemparan koin, pelemparan dadu,dll.
2. **Probabilitas frekuensi relatif (empiris):** probabilitas yang ditentukan berdasarkan fakta setelah kejadian. Contoh: Berdasarkan hasil survei 80 dari 100 orang responden mahasiswa sadar akan pentingnya memilah sampah, sehingga peluang seorang mahasiswa sadar akan pentingnya pemilahan sampah berdasarkan hasil survei tersebut adalah $P_A = \frac{80}{100} = 0,8$.
3. **Probabilitas subyektif:** probabilitas yang dilakukan berdasarkan pertimbangan perseorangan (seorang ahli atau orang yang berpengalaman). Contoh: probabilitas 10 kantong kompos memiliki berat < 1 kg menurut seorang penjual berdasarkan pengalamannya adalah 0,1 atau dari 10 kantong kompos terdapat satu kantong yang beratnya < 1 kg.

8.1 Aturan Dasar Probabilitas

Secara umum terdapat dua buah aturan dasar yang digunakan dalam perhitungan probabilitas yaitu aturan penjumlahan dan aturan perkalian. Kedua aturan tersebut akan penulis bahas secara detail pada bagian ini.

Sebelum kita membahas keduanya sebaiknya kita bahas terlebih dahulu pengertian umum yang merupakan elemen dasar dalam memahami konsep probabilitas. Berikut adalah istilah-istilah yang digunakan dalam probabilitas:

1. **Ruang sampel (*sample space*)**: gabungan dari semua kemungkinan, dan kemungkinan secara individual yang disebut sebagai titik sampel. Suatu peristiwa didefinisikan sebagai sub-himpunan (*subset*) dari ruang sampel. Ruang sampel bisa bersifat diskrit atau kontinu, yang dapat bernilai berhingga (*finite*) maupun tak berhingga. Peristiwa dalam pelemparan koin merupakan contoh ruang sampel berhingga. Contoh lainnya adalah pada pelemparan 2 buah dadu. Ruang sampel yang mungkin terbentuk merupakan kombinasi dari keenam masing-masing mata dadu. Berikut adalah contoh sintaks R untuk menghasilkan ruang sampel pada 2 buah dadu:

```
# install.packages("prob")
library(prob)

# ruang sampel 2 buah dadu
rolldie(2)
```

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

```
## 29 5 5
## 30 6 5
## 31 1 6
## 32 2 6
## 33 3 6
## 34 4 6
## 35 5 6
## 36 6 6
```

Berdasarkan sintaks tersebut terdapat 36 ruang sampel pada pelemparan 2 buah dadu. Ruang sampel yang dihasilkan dapat ditulis $Ruang sampel S = \{(X_1, X_2) | 1 \leq X_1 \leq 6; 1 \leq X_2 \leq 6\}$.

2. **Peristiwa mustahil (*impossible event*)**: dinyatakan dengan ϕ , merupakan peristiwa yang tidak memiliki titik sampel. Dengan demikian, peristiwa tersebut mempunyai himpunan kosong.
3. **Peristiwa tertentu (*certain event*)**: dinyatakan dengan S , merupakan semua peristiwa yang mengandung semua titik sampel dalam ruang sampel.
4. **Peristiwa komplementer (*complementary event*)**: Untuk suatu peristiwa dalam ruang sampel S , peristiwa komplementer dinyatakan dengan E yang mencakup semua titik sampel dalam S yang tidak terkandung dalam E .

Setelah pembaca memahami seluruh istilah tersebut, kita akan kembali menjelaskan kedua aturan dasar perhitungan probabilitas yaitu aturan penjumlahan dan perkalian.

Aturan penjumlahan merupakan aturan yang digunakan untuk menghitung suatu peristiwa A atau peristiwa lain yaitu peristiwa B yang akan terjadi dan ditulis sebagai $P(A \text{ atau } B)$ atau $P(A \cup B)$. Terdapat dua buah aturan penjumlahan yaitu:

1. Aturan penjumlahan peristiwa *mutually exclusive*.
2. Aturan penjumlahan untuk peristiwa *not mutually exclusive*.

Aturan selanjutnya adalah aturan perkalian yaitu aturan yang digunakan untuk menghitung bahwa peristiwa A dan peristiwa B akan terjadi bersamaan dan ditulis sebagai $P(A \text{ dan } B)$ atau $P(A \cap B)$. Aturan ini terdiri atas:

1. Aturan perkalian peristiwa *independent* (bebas).
2. Aturan perkalian peristiwa *dependent* (tidak bebas).

8.1.1 Peristiwa *Mutually Exclusive*

Peristiwa *mutually exclusive* merupakan suatu kondisi dimana peristiwa peristiwa satu tidak memungkinkan terjadinya peristiwa lainnya (tidak mungkin terjadi bersamaan). Terjadinya peristiwa A atau B merupakan penjumlahan kemungkinan terjadinya kedua peristiwa tersebut. Probabilitas peristiwa *mutually exclusive* dapat dituliskan menggunakan Persamaan (8.3).

$$P(A \cup B) = P(A) + P(B) \quad (8.3)$$

Untuk memudahkan pembaca memahami peristiwa *mutually exclusive* bayangkan pembaca diminta melemparkan sebuah dadu. Pembaca diminta untuk menentukan peluang munculnya angka 1 atau 6 pada dadu. Kedua peristiwa tersebut tidak mungkin terjadi bersamaan karena hanya dilakukan menggunakan satu dadu. Selain itu, jumlah himpunan masing-masing peristiwa pertama dan kedua hanyalah satu sehingga tidak memungkinkan adanya irisan pada kedua peristiwa tersebut. Untuk menghitungnya kita dapat langsung menggunakan Persamaan (8.3).

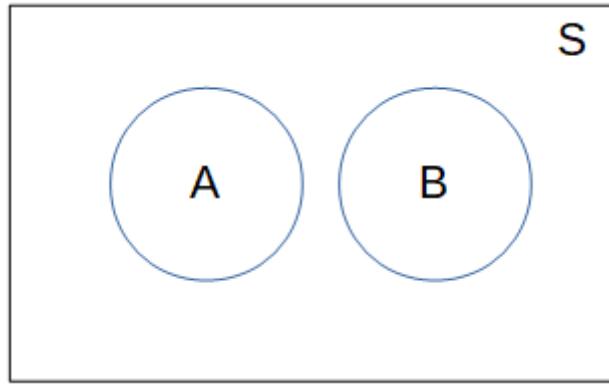


Figure 8.1: Diagram venn peristiwa mutually exclusive

$$P(1 \cup 6) = P(1) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Peristiwa pada contoh soal tersebut dapat digambarkan menggunakan diagram venn yang ditunjukkan pada Gambar 8.1).

Jika pembaca ingin menggunakan R untuk menghitung probabilitas peristiwa *mutually exclusive*, pembaca dapat menggunakan fungsi `Prob()` pada library `prob` untuk menghitung secara langsung probabilitas dari *subset* data. Berikut adalah contoh sintak untuk menghitung probabilitas munculnya angka 1 atau 6 dari pelemparan sebuah dadu:

```
# menentukan ruang sampel (S)
S <- rollDie(1, makespace=TRUE)
```

```
# print
S
```

```
##   X1   probs
## 1   1 0.1667
## 2   2 0.1667
## 3   3 0.1667
## 4   4 0.1667
## 5   5 0.1667
## 6   6 0.1667
```

```
# membuat subset peristiwa 1 dan 2
P1 <- subset(S, X1==1)
P6 <- subset(S, X1==6)
```

```
# menghitung probabilitas gabungan
P1$probs + P6$probs
```

```
## [1] 0.3333
```

```
# atau
Prob(P1) + Prob(P6)
```

```
## [1] 0.3333
```

Persamaan (8.3) dapat diperluas tidak hanya berlaku pada dua buah peristiwa. Jika jenis peristiwa A yang ada sebanyak n , maka Persamaan (8.3) dapat dituliskan kembali menjadi Persamaan (8.4).

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) \quad (8.4)$$

Kumpulan peristiwa yang terjadi $\{A_1, A_2, \dots, A_n\}$ pada ruang sampel S disebut sebagai *partisi* S jika A_1, A_2, \dots, A_n merupakan peristiwa dan $A_1 \cup A_2 \cup \dots \cup A_n = S$. Sehingga probabilitas seluruh partisi tersebut dapat dituliskan pada Persamaan (8.5).

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) = P(S) = 1 \quad (8.5)$$

8.1.2 Peristiwa *Not Mutually Exclusive*

Bila dua buah peristiwa tidak *mutually exclusive*, maka kedua peristiwa tersebut dapat terjadi secara bersamaan atau memiliki himpunan yang saling beririsan jika ditinjau dari pembahasan pelemparan dadu sebelumnya. Probabilitas suatu peristiwa yang tidak *mutually exclusive* dapat dituliskan berdasarkan Persamaan (8.6).

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (8.6)$$

Untuk memahami peristiwa yang tidak *mutually exclusive*, pembaca dapat membayangkan kembali melempar sebuah dadu. Pembaca diminta menghitung probabilitas keluar angka ganjil pada dadu atau angka prima pada dadu. Kedua peristiwa tersebut memiliki himpunnanya masing-masing. Untuk peristiwa angka ganjil himpunan yang terjadi adalah ganjil= $\{1, 3, 5\}$, sedangkan untuk angka prima adalah prima= $\{1, 2, 3, 5\}$. Kedua peristiwa tersebut memiliki irisan himpunan yaitu saat mata dadu menunjukkan angka 1, 3, dan 5. Nilai probabilitas kedua peristiwa tersebut tidak bisa dihitung dengan langsung menjumlahkan probabilitas keduanya masing-masing karena terdapat satu peristiwa yang merupakan bagian dari peristiwa lain sehingga peristiwa tersebut sebagian perlu dihilangkan dari probabilitas salah satunya seperti yang ditulikan pada Persamaan (8.6). Berdasarkan persamaan tersebut probabilitas yang peristiwa tersebut adalah sebagai berikut:

$$P(\text{ganjil} \cup \text{prima}) = P(\text{ganjil}) + P(\text{prima}) - P(\text{ganjil} \cap \text{prima}) = \frac{3}{6} + \frac{4}{6} - \frac{3}{6} = \frac{3+4-3}{6} = \frac{2}{3}$$

Peristiwa *not mutually exclusiv* dapat digambarkan menggunakan diagram venn yang ditunjukkan pada Gambar 8.2).

Pada R peristiwa tersebut dapat dihitung menggunakan sintaks berikut:

```
# kita akan menggunakan kembali objek S pada sintaks sebelumnya
# melakukan subset pada masing-masing peristiwa
ganjil <- subset(S, X1 %in% c(1, 3, 5))
prima <- subset(S, X1 %in% c(1, 2, 3, 5))

# menghitung irisan kedua peristiwa
irisan <- intersect(ganjil, prima)

# menghitung probabilitas yang terbentuk
Prob(ganjil)+Prob(prima)-Prob(irisan)
```

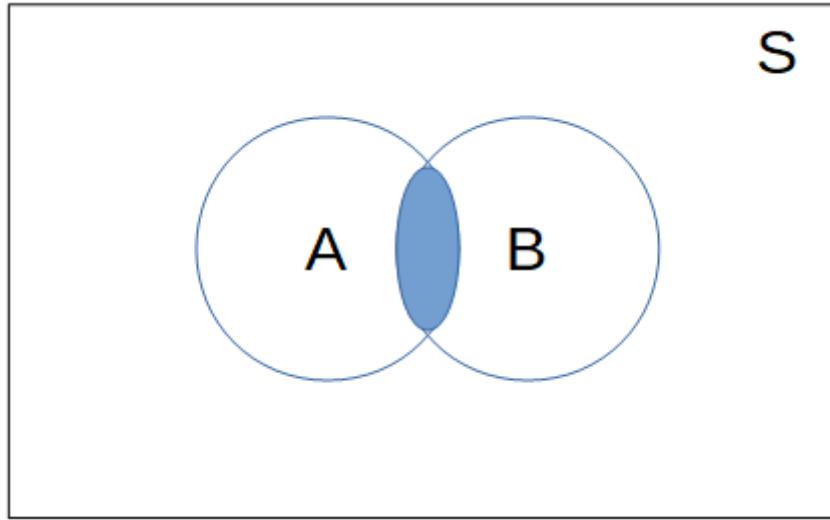


Figure 8.2: Diagram venn peristiwa not mutually exclusive

```
## [1] 0.6667
```

Untuk contoh yang lain misalkan seorang konsultan pengendalian kerugian diberikan data kerugian klien akibat kebakaran. Terdapat 250 kasus kebakaran dengan sejumlah penyebab. Penyebab utama disebabkan oleh membuang putung rokok sembarangan sebanyak 108 kasus, peralatan memasak sebanyak 95 kasus, pembakaran sebanyak 12 kasus, dan sumber kebakaran tidak diketahui sebanyak 35 kasus. Konsultan pengendalian kerugian ingin mengetahui berapa probabilitas untuk memilih klaim kebakaran dari kelompok dengan penyebab utama akibat aktivitas merokok sembarangan atau akibat pembakaran. Karena konsultan menentukan probabilitas “satu atau yang lain,” ia akan menentukan probabilitas berdasarkan peristiwa majemuk. Konsultan kemudian harus menentukan apakah peristiwa tersebut *mutually exclusive* atau tidak. Untuk melakukannya ia harus menjawab pertanyaan “Dapatkan seseorang melakukan klaim bahwa peristiwa kebakaran dapat disebabkan oleh aktivitas merokok dan pembakaran yang dilakukan secara bersamaan?”. Konsultan menentukan bahwa ini tidak mungkin. Oleh karena itu, peristiwa-peristiwa tersebut *mutually exclusive* dan probabilitas dari kedua peristiwa yang terjadi pada saat yang sama adalah nol. Probabilitas selanjutnya dapat dihitung menggunakan Persamaan (8.6).

$$P(\text{merokok} \cup \text{pembakaran}) = P(\text{merokok}) + P(\text{pembakaran}) - P(\text{merokok} \cap \text{pembakaran})$$

$$P(\text{merokok} \cup \text{pembakaran}) = \left(\frac{108}{250}\right) + \left(\frac{12}{250}\right) - 0 = 0,48 \text{ atau } 48\%$$

Persamaan (8.6) dapat diperluas tidak hanya menggunakan dua buah peristiwa tapi dapat dihitung nilai probabilitasnya untuk lebih dari dua peristiwa. Pada Persamaan (8.7) disajikan persamaan untuk menghitung probabilitas untuk 3 buah peristiwa.

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \quad (8.7)$$

Untuk peristiwa yang lebih banyak kita perlu menggambarkan terlebih dahulu diagram venn dari ruang sampel yang akan kita gunakan.

8.1.3 Peristiwa *Dependent*

Peristiwa *dependent* terjadi bila probabilitas terjadinya satu peristiwa (peristiwa A) dipengaruhi oleh probabilitas terjadinya peristiwa lainnya (peristiwa B) atau $P(B | A)$. Peristiwa ini merupakan probabilitas kondisional karena terjadinya B dipengaruhi oleh terjadinya A. Pendekatan yang digunakan dituliskan pada Persamaan (8.8).

$$P(A | B) = \frac{P(A \cap B)}{P(A)} \text{ dimana } P(A) > 0 \quad (8.8)$$

Untuk memahami probabilitas kondisional bayangkan pembaca harus melakukan survey terkait studi AMDAL di suatu kota. Responden yang digunakan merupakan seseorang yang telah menyelesaikan kuliahnya atau telah memperoleh gelar sarjana. Kategorisasi terhadap populasi dilakukan berdasarkan jenis kelamin dan status pekerjaan dengan jumlah yang proporsional dengan jumlah populasinya yang dapat dilihat pada Tabel 8.1. Sampel diambil dari populasi tersebut sesuai dengan proporsi jenis kelamin dan status pekerjaan. Pada studi ini ingin diketahui manfaat dari pembangunan industri pendirian industri baru bagi kota tersebut.

Table 8.1: Populasi orang yang telah menyelesaikan masa studinya di suatu kota.

Jenis Kelamin	Bekerja	Belum Bekerja	Total
Laki-Laki	460	40	500
Perempuan	140	260	400
Total	600	300	900

Proses survey dilakukan dengan metode wawancara. Responden yang telah dilakukan wawancara selanjutnya tidak boleh diwawancara lagi sehingga pada jumlah keseluruhan sampel terus berkurang. Hitunglah probabilitas kondisional dari pengambilan responden laki-laki akibat pengambilan responden seseorang yang telah bekerja?.

Berdasarkan contoh tersebut terdapat dua buah peristiwa yaitu peristiwa responden yang telah bekerja (dilambangkan dengan E) dan responden laki-laki (dilambangkan dengan M) atau dapat dituliskan sebagai berikut:

M : seorang laki-laki yang terpilih. E : seseorang yang dipilih dan telah bekerja.

Probabilitas kondisional dari pengambilan responden laki-laki akibat pengambilan responden seseorang yang telah bekerja selanjutnya dihitung seperti berikut:

$$P(M | E) = \frac{460}{600} = \frac{23}{30}$$

Misalkan $n(A)$ merupakan notasi yang menyatakan jumlah elemen dari suatu set A. Dengan menggunakan notasi tersebut, dimana setiap orang dewasa yang telah menyelesaikan studinya memiliki kesempatan yang sama untuk dipilih sebagai responden dalam penelitian dapat dituliskan sebagai berikut:

$$P(M | E) = \frac{n(E \cap M)}{n(E)} = \frac{\frac{n(E \cap M)}{n(S)}}{\frac{n(E)}{n(S)}}$$

$$P(M | E) = \frac{P(E \cap M)}{P(E)}$$

Persamaan yang dihasilkan sesuai dengan Persamaan (8.8), dimana $P(E \cap M)$ dan $P(E)$ dihitung berdasarkan besarnya ruang sampel S. Untuk memverifikasi hasil yang telah diperoleh sebelumnya, kita dapat melakukan perhitungan seperti berikut:

$$P(E) = \frac{600}{900}$$

serta

$$P(E \cap M) = \frac{460}{900} = \frac{23}{45}$$

Sehingga

$$P(E \mid M) = \frac{\frac{23}{45}}{\frac{2}{3}} = \frac{23}{30}$$

Berdasarkan hasil yang diperoleh telah dapat dibuktikan bahwa probabilitas kondisional dari pengambilan responden laki-laki akibat pengambilan responden seseorang yang telah bekerja sebesar $\frac{23}{30}$. Probabilitas lainnya dapat pembaca hitung sendiri untuk lebih memperdalam pengetahuan pembaca mengenai probabilitas kondisional.

Pada R dengan menggunakan contoh soal sebelumnya kita dapat melakukan perhitungan probabilitas kondisional pengambilan sampel laki-laki akibat dari pengambilan sampel seseorang yang telah bekerja. Sintaks yang digunakan adalah sebagai berikut:

```
# membuat data frame
S <- data.frame("jenis_kelamin"=c("laki-laki","perempuan"), "bekerja"=c(460,140), "belum_bekerja"=c(40,20))

# reshaping
library(tidyr)
S<-gather(S, key="status_pekerjaan", value="frekuensi", -jenis_kelamin)

# melakukan subset dan menghitung probabilitas
# peluang responden merupakan pegawai
E <- subset(S, status_pekerjaan=="bekerja")
P_E <- sum(E$frekuensi)/sum(S$frekuensi)
# peluang responden laki-laki dan bekerja
E_M <- subset(S, status_pekerjaan=="bekerja"&jenis_kelamin=="laki-laki")
P_E_M <- sum(E_M$frekuensi)/sum(S$frekuensi)

# Probabilitas kondisional
P_E_M/P_E

## [1] 0.7667
```

8.1.4 Peristiwa *Independent*

Untuk menentukan probabilitas dua atau lebih peristiwa akan terjadi bersamaan, perlu ditentukan terlebih dahulu apakah peristiwa-peristiwa tersebut bersifat bebas. Misalnya dalam melempar 2 buah dadu, probabilitas munculnya angka 1 pada dadu pertama adalah $\frac{1}{6}$ dan probabilitas munculnya angka 2 pada dadu kedua juga sama dengan dadu pertama. Jika kita menginginkan kedua nilai tersebut muncul bersamaan pada saat pelemparan, maka probabilitas kejadiannya adalah hasil perkalian kedua probabilitas peristiwa pada

masing-masing dadu yaitu $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. Pendekatan perhitungan probabilitas untuk peristiwa *independent* dapat dituliskan pada Persamaan (8.9).

$$P(A \cap B) = P(A) \cdot P(B) \quad (8.9)$$

Dengan menggunakan contoh soal sebelumnya, kita akan menentukan probabilitas responden penelitian kita adalah laki-laki (L) dan bekerja (E). Berdasarkan Persamaan (8.9), probabilitas terpilihnya jenis responden tersebut adalah sebagai berikut:

$$P(L \cap E) = \frac{500}{900} \cdot \frac{300}{900} = \frac{5}{27}$$

Dengan menggunakan R sintaks yang digunakan adalah sebagai berikut:

```
# subset responden laki-laki
L <- subset(S, jenis_kelamin=="laki-laki")
E <- subset(S, status_pekerjaan=="bekerja")

# probabilitas
P_L <- sum(L$frekuensi)/sum(S$frekuensi)
P_E <- sum(E$frekuensi)/sum(S$frekuensi)

# Probabilitas peristiwa independen
P_L * P_E

## [1] 0.3704
```

8.2 Teori Bayes

Teori Bayes memberikan formula probabilitas suatu peristiwa yang tergantung pada kontribusi dan ragam pada tahap sebelumnya. Formula tersebut dapat dituliskan pada Persamaan (8.10).

$$P(B_k | A) = \frac{P(B_k) \cdot P(A | B_k)}{\sum_{i=1}^n P(B_i) \cdot P(A | B_i)} \text{ dimana } k=1,2,\dots,n \quad (8.10)$$

Untuk membuktikan persamaan tersebut, kita akan menggunakan Persamaan (8.8) dengan melihat $P(B_k \cap A)$ dengan dua cara yang berbeda. Untuk lebih mudahnya, misalkan nilai $P(B_k) > 0$ untuk seluruh k , sehingga:

$$P(A) \cdot P(B_k | A) = P(B_k \cap A) = P(B_k) \cdot P(A | B_k) \quad (8.11)$$

sejak nilai $P(A) > 0$ kita dapat membaginya untuk mendapatkan

$$P(B_k | A) = \frac{P(B_k) \cdot P(A | B_k)}{P(A)} \quad (8.12)$$

Sekarang ingat kembali bahwa $\{B_k\}$ adalah partisi, teorema probabilitas total probabilitas total memberikan penyebut pada persamaan terakhir menjadi

$$P(A) = \sum_{k=1}^n P(B_k \cap A) = \sum_{k=1}^n P(B_k) \cdot P(A | B_k) \quad (8.13)$$

Apa artinya? Biasanya dalam aplikasinya kita diberikan (atau tahu) probabilitas apriori $P(B_k)$. Kita keluar dan mengumpulkan sejumlah data yang kita gunakan untuk mewakili peristiwa A. Kita ingin tahu: bagaimana kita dapat memperbaharui $P(B_k | A)$ menjadi $P(B_k | A)$? Jawabannya adalah dengan teori Bayes.

Untuk memahaminya misalkan sebuah instalasi air menggunakan tawas sebagai koagulannya. Tawas ini disuplai dari 4 perusahaan pemasok bahan kimia. Spesifikasi yang diinginkan adalah paling tidak tawas tersebut mengandung kadar efektif 60%. Data tentang perusahaan pemasok dan kegagalan untuk memenuhi standar yang diinginkan adalah:

- Perusahaan 1: memasok 20% dengan kegagalan 1 dalam 20 atau kegagalan = 0,05,
- Perusahaan 2: memasok 60% dengan kegagalan 1 dalam 10 atau kegagalan = 0,10,
- Perusahaan 3: memasok 15% dengan kegagalan 1 dalam 10 atau kegagalan = 0,10,
- Perusahaan 4: memasok 5% dengan kegagalan 1 dalam 20 atau kegagalan = 0,05.

Bila dari stok tawas digudang tersebut direksi ingin mengetahui berapa kemungkinan terjadinya kegagalan pada stok tawas dari perusahaan 1, dengan menggunakan teori Bayes kita dapat menghitungnya seperti berikut:

$$P(B_1 | A) = \frac{0,20 \cdot 0,05}{(0,6 \cdot 0,1 + 0,15 \cdot 0,1 + 0,05 \cdot 0,05)} = 0,114$$

8.3 Ekspektasi Matematis

Misalkan Menteri Kesehatan Republik Indonesia merilis hasil studi yang menyatakan usia harapan hidup masyarakat Indonesia adalah 70 tahun. Ini tidak berarti saat kita berusia 65 tahun kita akan meninggal 5 tahun berikutnya. Pengertian usia harapan hidup ini didasarkan pada probabilitas yaitu ekspektasi matematis yang dituliskan pada Persamaan (8.14).

$$E(X) = \sum_{i=1}^n x_i \cdot P(X_i) \quad (8.14)$$

Misalkan terdapat eksperimen yang menghasilkan i buah peristiwa, dan masing-masing mempunyai probabilitas terjadi: $p_1, p_2, p_3, \dots, p_i$.

sehingga: $p_1 + p_2 + p_3 + \dots + p_k = 1$ maka ekspektasinya adalah $E = p_1 \cdot x_1 + p_2 \cdot x_2 + p_3 \cdot x_3 + \dots + p_i \cdot x_i$. Hasil perjumlahan tersebut akan menghasilkan Persamaan (8.14).

Untuk memahami penerapan ekspektasi matematis, misalkan sebuah konsultan sedang menyiapkan proposal untuk sebuah proyek. Biaya untuk menyiapkan proposal adalah 5 juta rupiah, sedang keuntungan kotor bila proyek ini diperoleh adalah:

- 50 juta rupiah dengan probabilitas 0,20
- 30 juta rupiah dengan probabilitas 0,50
- 10 juta rupiah dengan probabilitas 0,20
- 0 rupiah dengan probabilitas 0,10.

Bila kemungkinan mendapatkan proyek tersebut adalah 0,30, maka keuntungan yang diharapkan adalah:

- Probabilitas memperoleh keuntungan 45 juta rupiah (keuntungan kotor-modal)=probabilitas mendapatkan proyek x keuntungan proyek tersebut= $0,30 \times 0,20 = 0,06$
- Probabilitas memperoleh keuntungan 25 juta = $0,30 \times 0,50 = 0,15$
- Probabilitas memperoleh keuntungan 5 juta = $0,30 \times 0,20 = 0,06$

- Probabilitas memperoleh kerugian 5 juta = $(0,30 \times 0,10) + 0,70 = 0,73$

Maka ekspektasinya = $(45 \text{ juta} \times 0,06) + (5 \text{ juta} \times 0,15) + (5 \text{ juta} \times 0,06) - (5 \text{ juta} \times 0,73) = 3,1 \text{ juta}$. Dengan demikian perusahaan tersebut dapat memutuskan apakah akan meneruskan membuat proposal tersebut, dengan kemungkinan merugi sebesar 5 juta rupiah (biaya membuat proposal) dan kemungkinan untung 3 juta rupiah.

8.4 Referensi

1. Damanhuri, E. 2011. **Statitika Lingkunga**. Penerbit ITB.
2. Kerns, G.Jay. 2018. **Introduction to Probability and Statistics Using R Third Edition**. GNU Free Documentation License.
3. Janicak, C.A. 2007. **Applied Statistics in Occupational Safety and Health**. Government Institutes.
4. Walpole, E. R., Myers, H.M., Myers, S.L., Keying Ye. 2011. **Probability & Statistics for Engineering & Scientists Ninth Edition**. Prentice Hall.

Chapter 9

Distribusi Probabilitas

Distribusi probabilitas merupakan sebuah fungsi yang menggambarkan kemungkinan memperoleh sejumlah nilai dalam suatu variabel acak. Dengan kata lain distribusi probabilitas menjelaskan bahwa nilai yang muncul pada sampel acak akan bervariasi berdasarkan distribusi probabilitas yang menyertainya.

Untuk memahaminya misalkan kita melakukan suatu sampling dengan cara survey terhadap sejumlah responden untuk mengetahui produksi sampah harianya. Keseluruhan nilai timbulan yang diperoleh selanjutnya disebut sebagai distribusi timbulan sampah. Distribusi tersebut berguna saat kita mengetahui hasil mana yang paling mungkin, sebaran nilai potensial serta kemungkinan hasil yang berbeda.

9.1 Properti Umum dari Distribusi Probabilitas

Distribusi probabilitas menjelaskan kemungkinan suatu peristiwa atau nilai muncul. Ahli statistika menjelaskan distribusi probabilitas kedalam Persamaan (9.1).

$$P(x) = \text{kemungkinan suatu variabel acak mengandung nilai } x \quad (9.1)$$

Jumlah seluruh probabilitas adalah 1. Selain itu retang nilai probabilitas berkisar antara 0 sampai 1, dimana hal ini telah penulis jelaskan pada Chapter sebelumnya.

Distribusi probabilitas menjelaskan sebaran nilai variabel acak. Akibatnya, jenis variabel menentukan jenis distribusi probabilitas. Untuk variabel acak tunggal, ahli statistik membagi distribusi menjadi dua jenis berikut:

- **Distribusi probabilitas yang diskrit**

Fungsi probabilitas diskrit dikenal sebagai fungsi massa probabilitas, dimana kita dapat mengasumsikan sejumlah nilai diskrit. Misalnya pelemparan dadu serta perhitungan sebuah peristiwa seperti flu di suatu daerah merupakan fungsi tersendiri. Kedua contoh tersebut merupakan contoh peristiwa diskrit karena tidak ada nilai antara, misalnya pada dadu tidak ada nilai antara 1 dan 2, dan seterusnya. Pada perhitungan jumlah peristiwa flu juga tidak ada nilai antara orang terserang dlu dan tidak. Contoh lainnya adalah perhitungan jumlah buku diperpustakaan yang diperiksa tiap jam. Kita dapat menghitung jumlah buku perjam seperti 21 buku atau 22 buku, tetapi kita tidak dapat menghitung jumlah buku pada nilai antara kedua nilai tersebut. Distribusi probabilitas diskrit terdiri atas:

- a. Binomial
- b. Hypergeometric

- c. Poisson
- d. Geometric
- e. Multinomial

- **Distribusi probabilitas yang bersifat kontinu**

Fungsi probabilitas kontinu dikenal juga sebagai fungsi densitas probabilitas. Suatu distribusi dikatakan sebagai distribusi kontinu jika nilai yang terkandung dalam distribusi tersebut tidak terbatas serta skala yang digunakan dapat pula mengandung nilai desimal. Contoh suatu pengukuran yang menghasilkan distribusi kontinu adalah tinggi, berat, suhu, dll. Distribusi probabilitas kontinu terdiri atas:

- a. Normal
- b. Binomial
- c. Uniform
- d. Loh Normal
- e. Gamma, dll.

9.2 Distribusi Binomial dan Multinomial

Suatu percobaan sering dilakukan dengan proses yang berulang-ulang. Tiap proses percobaan yang dilakukan akan menghasilkan dua luaran yaitu **sukses** atau **gagal**. Untuk memahaminya misalkan pembaca melakukan pelemparan sebuah koin. Jika yang keluar adalah bagian kepala (*head*) maka proses tersebut dikatakan sukses, jika sebaliknya maka gagal. Proses penentuan sukses dan gagal tersebut tergantung pada sudut pandang kita melihatnya. Proses percobaan demikian disebut sebagai **Bernoulli process** (proses bernouli). Setiap percobaan yang dilakukan disebut sebagai **Bernoulli trial**.

Bernoulli process memiliki ciri-ciri sebagai berikut:

- Eksperimen terdiri atas sejumlah perulangan percobaan.
- Setiap perulangan percobaan menghasilkan luaran yang diklasifikasikan sebagai **sukses** atau **gagal**.
- Probabilitas sukses (dinotasikan p) konstan dari setiap perulangan.
- Setiap perulangan percobaan bersifat independen.

Untuk memahami **Bernoulli process** kita akan menggunakan contoh pelemparan dadu. Misalkan kita akan melakukan percobaan pelemparan dadu sebanyak 3 kali. Probabilitas munculnya bagian angka 1 pada dadu adalah $\frac{1}{6}$. Probabilitas ini konstan pada setiap perulangan. Jika kita menginginkan ketiga perulangan tersebut menghasilkan angka 1. Probabilitasnya merupakan hasil kali probabilitas pada tiap giliran pelemparan seperti berikut:

$$P(111) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{216}$$

9.2.1 Distribusi Binomial

Jumlah X keberhasilan atau jumlah percobaan sukses dalam percobaan Bernoulli disebut **variabel acak binomial**. Distribusi probabilitas dari variabel acak diskrit ini disebut distribusi binomial, dan nilainya akan dinotasikan dengan $b(x; n, p)$ karena mereka bergantung pada jumlah percobaan dan probabilitas keberhasilan pada percobaan yang diberikan.

Bernoulli trial dapat menghasilkan percobaan yang sukses dengan probabilitas p dan percobaan gagal dengan probabilitas $q = 1 - p$. Sehingga distribusi probabilitas binomial percobaan sukses untuk variabel acak X dengan jumlah n percobaan yang independen dinyatakan kedalam Persamaan (9.2).

$$b(x; n, p) = (nC_x) p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n. \quad (9.2)$$

dimana

$$nC_x = \frac{n!}{x!(n-x)!} \quad (9.3)$$

Dengan menggunakan contoh sebelumnya kita dapat menghitung probabilitas keluarnya angka 1 pada dadu untuk 3 kali percobaan (n) atau seluruh percobaan menghasilkan angka 1 ($n = x = 3$) adalah:

$$\begin{aligned} b\left(3; 3, \frac{1}{6}\right) &= \left(\frac{3!}{3!(3-3)!}\right) \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{(3-3)} \\ b\left(3; 3, \frac{1}{6}\right) &= (1) \cdot \left(\frac{1}{6}\right)^3 \cdot (1) = \frac{1}{216} \end{aligned}$$

Kita juga dapat melakukan perhitungan tersebut menggunakan R untuk mengetahui probabilitas munculnya angka 1 pada dadu untuk 3 kali percobaan. Sintak yang digunakan adalah sebagai berikut:

```
dbinom(x=3, # jumlah kejadian sukses
        size=3, # jumlah percobaan
        prob=1/6) # probabilitas kejadian
```

[1] 0.00463

Probabilitas Kumulatif Binomial

Pada kondisi lain kita tidak hanya tertarik dengan probabilitas munculnya suatu peristiwa. Kita terkadang tertarik untuk menghitung probabilitas kumulatif dari suatu peristiwa. Misalkan probabilitas munculnya nomor dadu < 4 atau pada perhitungan kendaraan per jam yang melalui suatu jalan kita tertarik menghitung peluang kendaraan yang melintas tiap jam < 50 kendaraan. Untuk menghitung probabilitas yang demikian kita perlu melakukan akumulasi probabilitas yang memenuhi kriteria yang telah kita tentukan sebelumnya. Probabilitas kumulatif distribusi binomial berdasarkan kondisi tertentu dinyatakan pada Persamaan (9.4).

$$B(r, n, p) = \sum_{x=0}^r b(x; n, p) \quad (9.4)$$

dimana r adalah kondisi probabilitas yang kita inginkan yang dapat dituliskan sebagai $P(X < r)$ atau $P(X > r)$. Kondisi lain yang dapat kita gunakan adalah $P(a \leq X \leq b)$.

Untuk memahaminya kita akan membuat sebuah contoh kasus. Misalkan kita diminta untuk melakukan analisis kerusakan *diffused aerator* pada suatu instalasi air limbah. Jumlah aerator total pada instalasi tersebut adalah 10 buah. Probabilitas sebuah aerator rusak sebesar $\frac{1}{10}$. Tentukan berapa probabilitas jika (a) setidaknya 3 aerator tersebut tidak rusak? (b) 4 sampai 5 buah aerator rusak? serta (c) sebanyak 5 aerator rusak?.

- (a) jika setidaknya 3 aerator tidak rusak

$$P(X \leq 3) = \sum_{x=0}^3 b(x; 10, 0.1)$$

$$P(X \leq 3) = b(0; 10, 0.1) + b(1; 10, 0.1) + b(2; 10, 0.1) + b(3; 10, 0.1)$$

$$P(X \leq 3) = 0,987$$

Kita juga dapat memperoleh nilai tersebut dengan melihat tabel statistika yang dapat pembaca lihat pada tautan [berikut](#)

(b) jika 4 sampai 5 aerator rusak

$$P(4 \leq X \leq 5) = \sum_{x=4}^5 b(x; 10, 0.1) = \sum_{x=0}^5 b(x; 10, 0.1) - \sum_{x=0}^3 b(x; 10, 0.1)$$

$$P(4 \leq X \leq 5) = 1 - 0,987 = 0,013$$

(c) jika tepat 5 aerator rusak

$$P(X = 5) = b(5; 10, 0.1) = \sum_{x=0}^5 b(x; 10, 0.1) - \sum_{x=0}^4 b(x; 10, 0.1)$$

$$P(X = 5) = 1 - 0,998 = 0,002$$

Untuk melakukan perhitungannya pada R kita dapat menggunakan fungsi `pbinom()`. Fungsi tersebut akan menghitung probabilitas bedasarkan nilai kondisi yang telah kita masukkan. Berikut adalah sintaks yang digunakan:

```
# a) jika setidaknya 3 aerator tidak rusak
pbinom(q=3,
       size=10, # jumlah percobaan
       prob=0.1) # probabilitas sukses
```

```
## [1] 0.9872
```

```
# b) jika setidaknya 4 sampai 5 aerator rusak
pbinom(q=5,size=10,prob=0.1)-pbinom(q=3,size=10,prob=0.1)
```

```
## [1] 0.01265
```

```
# c) jika tepat 5 aerator rusak
dbinom(x=5, # jumlah kejadian sukses
       size=10, # jumlah percobaan
       prob=0.1) # probabilitas sukses
```

```
## [1] 0.001488
```

Menghitung Nilai Rata-Rata dan Varians Distribusi Binomial

Kita sudah mengetahui bahwa distribusi probabilitas binomial hanya bergantung pada nilai n , p , dan q . Berdasarkan tersebut nilai mean, dan varians dari distribusi probabilitasnya juga bergantung pada ketiga nilai tersebut. Nilai mean dituliskan pada Persamaan (9.5), sedangkan varians dari distribusi probabilitas distuliskan pada Persamaan (9.6).

$$\mu = np \quad (9.5)$$

$$\sigma^2 = npq \quad (9.6)$$

9.2.2 Multinomial Eksperimen dan Distribusi Multinomial

Eksperimen Binomial (*Binomial process*) dapat menjadi **eksperimen multinomial** jika kita menginginkan luaran dari percobaan yang dilakukan memiliki lebih dari satu hasil. Contoh dari eksperimen multinomial ini misalnya adalah penarikan kartu dari seperangkam kartu. Kita dapat mengagap penarikan kartu dengan pengembalian sebagai eksperimen multinomial jika luaran yang diinginkan adalah 4 jenis kartu dalam set kartu tersebut.

Secara umum, jika percobaan yang diberikan dapat menghasilkan salah satu k dari hasil yang mungkin E_1, E_2, \dots, E_k dengan probabilitas yang dihasilkan sebesar p_1, p_2, \dots, p_k , maka **distribusi multinomial** akan memberikan nilai probabilitas yang dinyatakan E_1 terjadi sebanyak x_1 kali, E_2 terjadi sebanyak x_2 kali, sampai dengan E_k terjadi sebanyak x_k kali dalam n percobaan yang independent, dimana

$$x_1 + x_2 + \dots + x_k = n$$

Selanjutnya fungsi probabilitas dituliskan sebagai berikut:

$$f(x_1, x_2, \dots, x_k; p_1, p_2, \dots, p_k)$$

Seperti yang telah kita ketahui bersama bahwa nilai $p_1 + p_2 + \dots + p_k = 1$. Sejak percobaan yang dilakukan independen, maka setiap percobaan yang menghasilkan x_1 yang merupakan luaran dari E_1 , x_2 yang merupakan luaran dari E_2 sampai dengan x_k yang merupakan luaran dari E_k akan terjadi dengan probabilitas $p^{x_1} p^{x_2} \dots p^{x_k}$. Jumlah urutan yang menghasilkan hasil yang serupa untuk percobaan n adalah sama dengan jumlah partisi n item ke dalam k grup dengan x_1 di grup pertama, x_2 di grup kedua, sampai dengan x_k di grup k . Kondisi ini dapat dituliskan seperti berikut:

$$\binom{n}{x_1, x_2, \dots, x_k} = \frac{n!}{x_1! x_2! \dots x_k!}$$

Sejak seluruh partisi bersifat *mutually exclusive* dan terjadi dengan probabilitas yang setara, kita dapat memperoleh distribusi multinomial dengan mengalikan probabilitas tiap luaran spesifik dengan jumlah total partisinya. Persamaan distribusi multinomial yang diperoleh dituliskan kedalam Persamaan (9.7).

$$f(x_1, x_2, \dots, x_k; p_1, p_2, \dots, p_k) = \binom{n}{x_1, x_2, \dots, x_k} p^{x_1} p^{x_2} \dots p^{x_k} \quad (9.7)$$

dimana

$$\sum_{i=1}^k x_i = n \text{ dan } \sum_{i=1}^k p_i = 1 \quad (9.8)$$

Untuk memahami penerapan distribusi multinomial, penulis akan memberikan sebuah studi kasus. Probabilitas sejenis pompa memiliki umur ekonomis 2 tahun sebesar 0,30, antara 2-4 tahun adalah 0,50, dan 4-5 tahun adalah 0,20. Hitunglah probabilitas 8 buah pompa dimana pompa dengan umur ekonomis 2 tahun sebanyak 2 buah, 2-4 tahun sebanyak 5 buah, dan antara 4-5 sebanyak 1 buah.

Untuk menyelesaikannya kita perlu mendata jumlah kemunculan disertai dengan probabilitas kejadian pada tiap grup seperti berikut:

$$n = 8; x_1 = 2 \text{ dengan } p_1 = 0,30; x_2 = 5 \text{ dengan } p_2 = 0,50, \text{ dan } x_3 = 1 \text{ dengan } p_3 = 0,20$$

dengan menggunakan Persamaan (9.7), maka probabilitasnya dapat dihitung seperti berikut:

$$f(2, 5, 1; 0.30, 0.50, 0.20) = \binom{8}{2, 5, 1} (0,3)^2 (0,5)^5 (0,2)^1$$

$$f(2, 5, 1; 0.30, 0.50, 0.20) = 0,0945$$

Pada R kita dapat menggunakan fungsi `dmultinom()` untuk menghitung probabilitas distribusi multinomial. Komponen dari fungsi tersebut adalah sebagai berikut:

```
dmultinom(x, size, prob)
```

Note:

- **x:** vektor numerik
- **size:** jumlah percobaan atau perulangan
- **prob:** vektor numerik probabilitas tiap grup hasil.

Berikut adalah sintaks untuk menghitung probabilitas multinomial pada contoh kasus di atas:

```
dmultinom(c(2,5,1), # jumlah kejadian tiap grup
           size=8, # jumlah percobaan
           prob=c(0.3,0.5,0.2)) # probabilitas masing-masing luaran
```

```
## [1] 0.0945
```

9.3 Distribusi Hipergeometris

Distribusi hipergeometris didasarkan pada eksperimen hipergeometris yang memiliki asumsi sebagai berikut:

1. Sampel dengan ukuran n diambil secara acak tanpa pengembalian (*sampel without replacement*) dari populasi berukuran N .
2. Pada populasi, k didefinisikan sebagai observasi yang **sukses**, sedangkan $N - k$ didefiniskan sebagai observasi yang **gagal**.

Melalui asumsi tersebut, kita dapat menemukan perbedaan antara distribusi hipergeometris dengan distribusi binomial. Perbedaan yang paling mendasar adalah metode sampling yang digunakan, dimana distribusi binomial mengasumsikan sampel dengan pengembalian (*sample with replacement*), sedangkan distribusi hipergeometris mengasumsikan sebaliknya.

Distribusi probabilitas hipergeometris untuk variabel acak X dengan jumlah sampel n dari populasi terpilih dengan ukuran populasi N , dimana k merupakan observasi **sukses** dan $N - k$ merupakan observasi yang gagal dapat dituliskan berdasarkan Persamaan (9.9).

$$h(x; N, n, k) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}, \quad \max\{0, n - (N - k)\} \leq x \leq \min\{n, k\} \quad (9.9)$$

Range dari x dapat ditentukan dari 3 koefisien binomial, dimana x dan $n - x$ tidak lagi lebih besar dari k dan $N - k$ dan nilai keduanya tidak boleh lebih kecil dari 0. Biasanya ketika kedua nilai k dan $N - k$ lebih besar dari ukuran sampel n , range dari variabel acak hipergeometris akan menjadi $x = 0, 1, \dots, n$.

Untuk memahami penerapan distribusi hipergeometris, kita dapat menerapkannya dalam sebuah contoh kasus. Misalkan pembaca ditugaskan untuk melakukan sortir terhadap 40 sak kompos yang akan dijual dengan berat rata-rata 2 kg. Pembaca diberi tahu bahwa 3 dari seluruh kompos tersebut memiliki berat kurang dari 2 kg sehingga tidak dapat dijual. Tugas pembaca adalah menemukan ketiga sak kompos tersebut. Untuk mempermudah proses tersebut pembaca melakukan sampling secara acak dengan jumlah sampling 5 buah kompos tanpa pengembalian. Hitunglah berapa peluang pada tiap sampling tersebut pembaca menemukan 1 kompos yang memiliki berat lebih kecil dari 2 kg tersebut?.

Berdasarkan studi kasus tersebut kita dapat menyimpulkan bahwa proses sampling yang dilakukan adalah dengan menggunakan prosedur sampling tanpa pengembalian, sehingga distribusi hipergeometris dapat diterapkan dengan nilai $n = 5$, $N = 40$, $k = 3$, dan $x = 1$. Dengan menggunakan Persamaan (9.9), peluang ditemukan 1 sak kompos yang tidak sesuai adalah sebagai berikut:

$$h(1; 40, 5, 3) = \frac{\binom{3}{1} \binom{40-3}{5-1}}{\binom{40}{5}} = 0,3011$$

Berdasarkan hasil perhitungan diketahui bahwa peluang untuk menemukan 1 sak kompos dari 3 sak kompos yang tidak sesuai sebesar 30% pada tiap kali sampling.

Pada R distribusi probabilitas hipergeometris dapat dihitung menggunakan fungsi `dhyper()`. Format yang digunakan pada fungsi tersebut adalah sebagai berikut:

```
dhyper(x, m, n, k)
```

Note:

- **x:** vektor numerik yang menyatakan observasi sukses pada tiap sampling
- **m:** jumlah observasi sukses
- **n:** jumlah observasi gagal
- **k:** ukuran sampel

Berikut adalah sintaks untuk menghitung probabilitas hipergeometris contoh kasus di atas:

```
dhyper(x=1, # observasi sukses tiap sampling
       m=3, # observasi sukses (k)
       n=37, # observasi gagal (N-k)
       k=5) # sampel
```

```
## [1] 0.3011
```

Probabilitas Kumulatif Hipergeometris

Pada contoh kasus sebelumnya, sak kompos yang tidak memenuhi kriteria bisa saja saat sampling tidak hanya ditemukan 1 sak yang tidak memenuhi, bisa dua, tiga atau sama sekali tidak ada yang ditemukan sak yang tidak memenuhi kriteria. Kondisi tersebut mengharuskan kita menghitung probabilitas kumulatif dari suatu kondisi seperti $P(X < r)$, $P(X > r)$, atau $P(a < X < b)$ yang dapat dituliskan pada Persamaan (9.10).

$$H(r; N, n, k) = \sum_{x=0}^r h(x; N, n, k) \quad (9.10)$$

Pada contoh sebelumnya, hitunglah probabilitas jika paling banyak 2 sak kompos yang tidak memenuhi kriteria ditemukan pada sampel?.

Untuk melakukannya kita perlu menghitung probabilitas hipergeometris untuk kondisi saat $x = 0, 1, 2$. Dengan menggunakan Persamaan (9.10), nilai probabilitas yang dihasilkan adalah sebagai berikut:

$$P(X \leq 2) = b(0; 40, 5, 3) + b(1; 40, 5, 3) + b(2; 40, 5, 3) = 0,999$$

Pada R probabilitas kumulatif dapat dihitung menggunakan fungsi `phyper()`. Format fungsi yang digunakan adalah sebagai berikut:

```
phyper(q, m, n, k, lower.tail=TRUE)
```

Note:

- **q:** vektor numerik yang menyatakan observasi maksimum yang sukses saat sampling
- **m:** jumlah observasi sukses
- **n:** jumlah observasi gagal
- **k:** ukuran sampel
- **lower.tail:** probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE

Dengan menggunakan sintaks tersebut, probabilitas kumulatif dapat dihitung sebagai berikut:

```
phyper(q=2, # probabilitas <=2
       m=3, # observasi sukses
       n=37, # jumlah gagal
       k=5) # jumlah sampel
```

```
## [1] 0.999
# atau
dhyper(0,3,37,5)+dhyper(1,3,37,5)+dhyper(2,3,37,5)
```

```
## [1] 0.999
```

Menghitung Nilai Rata-Rata dan Varians Distribusi Hipergeometris

Nilai rata-rata dan varians distribusi hipergeometris dituliskan kedalam Persamaan (9.11):

$$\mu = \frac{nk}{N} \text{ dan } \sigma^2 = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N} \left(1 - \frac{k}{N}\right) \quad (9.11)$$

Bila nilai $n \ll N$, maka pendekatan distribusi binomial dapat dilakukan dengan pendekatan n dan $p = \frac{k}{N}$. Pendekatan yang dilakukan akan cukup baik bila $n \leq 0,1N$.

9.4 Distribusi Binomial Negatif dan Distribusi Geometris

Mari kita bayangkan percobaan di mana sifat-sifatnya (propertinya) sama dengan percobaan binomial, dengan pengecualian bahwa uji coba akan diulangi sampai sejumlah keberhasilan terjadi. Oleh karena itu, alih-alih probabilitas x keberhasilan dalam n percobaan, di mana n tetap, kita sekarang tertarik pada probabilitas bahwa keberhasilan k terjadi pada percobaan ke- x . Eksperimen semacam ini disebut eksperimen **binomial negatif**.

Sifat-sifat dari percobaan binomial negatif adalah sebagai berikut:

1. Percobaan terdiri atas sejumlah x perulangan.
2. Setiap percobaan memiliki dua hasil (**sukses** dan **gagal**).
3. Probabilitas sukses dinotasikan dengan p yang sama pada setiap percobaan.
4. Setiap percobaan bersifat independen yang berarti hasil dari sebuah percobaan tidak akan mempengaruhi hasil percobaan lainnya.
5. Percobaan dilakukan secara terus-menerus sampai dengan sejumlah k sukses terjadi, dimana k ditentukan terlebih dahulu.

Untuk memahaminya misalkan kita menguji sebuah obat dengan memberikannya kepada pasien yang sakit. Keberhasilan obat tersebut Obat dinyatakan sukses jika secara efektif memberikan efek pemulihan bagi pasien. Probabilitas obat tersebut melakukannya berdasarkan hasil studi yang telah dilakukan sebesar 60%. Kita tertarik untuk mengetahui probabilitas pasien kelima yang mengalami efek penyembuhan dimana pasien ini merupakan pasien ketujuh yang diberikan obat tersebut. Untuk melakukannya kita definisikan kejadian sukses dengan simbol S dan gagal dengan simbol F , urutan yang mungkin dari ketujuh pasien berdasarkan respon terhadap obat adalah $S F S S S F S$ yang probabilitas kejadian berdasarkan urutan tersebut adalah sebagai berikut:

$$(0,6)(0,4)(0,6)(0,6)(0,6)(0,4)(0,6) = (0,6)^5(0,4)^2$$

Kita dapat mendaftar sejumlah luaran yang mungkin pada kejadian tersebut mengatur ulang F dan S kecuali untuk hasil terakhir, yang harus menjadi keberhasilan kelima. Jumlah total luaran yang mungkin sama dengan jumlah partisi dari enam (7-1) percobaan pertama menjadi dua kelompok dengan 2 buah gagal dan 4 buah sukses menjadi kelompok tersendiri. Hal ini dapat dilakukan berdasarkan $\binom{6}{4} = 15$ cara yang *mutually exclusive*. Oleh karena itu, jika X mewakili hasil dimana keberhasilan kelima terjadi, maka

$$P(X=7) = \binom{6}{4} (0.6)^2 (0,4)^2 = 0,1866$$

9.4.1 Distribusi Binomial Negatif

Berdasarkan contoh kasus tersebut, kita dapat mendefinisikan formula untuk distribusi probabilitas binomial negatif. Jika percobaan yang bersifat independen dan berulang dapat menghasilkan keberhasilan (kejadian sukses) dengan probabilitas p dan kegagalan dengan probabilitas $q = 1 - p$, maka distribusi probabilitas variabel acak X , jumlah percobaan di mana keberhasilan k terjadi dinyatakan pada Persamaan (9.12):

$$b^*(x; k, p) = \binom{x-1}{k-1} p^k q^{x-k}, \quad x = k, k+1, k+2, \dots \quad (9.12)$$

Untuk lebih memahami penerapan dari distribusi binomial negatif. Misalkan pada suatu evan NBA (Championship series) atau final antar juara wilayah, dimana pada pertandingan puncak kedua tim dari dua wilayah akan melakukan 7 pertandingan terakhir. Suatu tim dinyatakan juara jika berhasil meraih 4 kemenangan dari 7 pertandingan yang ada. Anggaplah tim A dan B berhadapan satu sama lain. Probabilitas A memenangkan suatu pertandingan terhadap tim B sebesar 0,55, tentukan:

- Berapakah probabilitas tim A memenangkan kejuaraan pada pertandingan ke-6 dari 7 pertandingan yang ada?
- Berapakah probabilitas A memenangkan kejuaraan?

Dengan menggunakan Persamaan (9.12), probabilitas kemenangan tim A terhadap tim B dapat dihitung sebagai berikut:

- **Tim A juara pada pertandingan ke-6 ($x = 6$, $k = 4$, dan $p = 0,55$)**

$$b^*(6; 4, 0.55) = \binom{6-1}{4-1} 0,55^4 (1 - 0,55)^{6-4} = 0,1853$$

- **Tim A menjuarai kejuaraan**

Tim A dapat menjuarai kejuaraan jika telah memenangkan 4 dari 7 pertandingan. Kemungkinan Tim A dapat memenangkan pertandingan tersebut dapat terjadi pada pertandingan ke-4 (menang berturut-turut), pertandingan ke-6, dan pertandingan ke-7.

$$b^*(4; 4, 0.55) + b^*(5; 4, 0.55) + b^*(6; 4, 0.55) + b^*(7; 4, 0.55)$$

$$0,0915 + 0,1647 + 0,1853 + 0,1668 = 0,6083$$

Pada R kita dapat menghitung probabilitas binomial negatif menggunakan fungsi `dnbnom()`. Format fungsi tersebut adalah sebagai berikut:

```
dnbinom(x, size, prob)
```

Note:

- **x:** jumlah observasi gagal
- **size:** jumlah observasi sukses
- **prob:** probabilitas kejadian sukses

Dengan menggunakan fungsi `dnbinom()`, probabilitas Tim A menang pada pertandingan ke-6 dapat dihitung sebagai berikut:

```
dnbinom(x=2, # jumlah observasi gagal
        size=4, # jumlah observasi sukses
        prob=0.55) # probabilitas sukses
```

```
## [1] 0.1853
```

Pada pertanyaan kedua soal dapat kita impulkan bahwa kita hendak mencari probabilitas kumulatif kemenangan Tim A. Untuk melakukannya kita dapat menggunakan fungsi `pnbinom()`. Format fungsi tersebut adalah sebagai berikut:

```
pnbinom(q, size, prob, lower.tail = TRUE)
```

Note:

- **q**: jumlah observasi gagal minimum
- **size**: jumlah observasi sukses maksimum
- **prob**: probabilitas sukses
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE

Dengan menggunakan fungsi tersebut, probabilitas kumulatif kemenangan Tim A adalah sebagai berikut:

```
pbinom(q=3, # jumlah observasi gagal min
       size=4, # jumlah observasi sukses maks
       prob=0.55) # probabilitas sukses
```

```
## [1] 0.6083
```

```
# atau
dnbinom(x=0,size=4,prob=0.55) +
dnbinom(x=1,size=4,prob=0.55) +
dnbinom(x=2,size=4,prob=0.55) +
dnbinom(x=3,size=4,prob=0.55)
```

```
## [1] 0.6083
```

9.4.2 Distribusi Geometris

Pada kenyataannya kita hanya tertarik terhadap probabilitas kejadian sukses pertama kali akan terjadi. Probabilitas kejadian tersebut merupakan kejadian pada **distribusi probabilitas geometris**. Distribusi ini merupakan kasus khusus distribusi binomial negatif.

Jika suatu percobaan independen dapat menghasilkan kejadian sukses dengan probabilitas p dan gagal dengan probabilitas $q = 1 - p$, maka distribusi probabilitas variabel acak X , jumlah percobaan dimana sukses pertama terjadi didefinisikan pada Persamaan (9.13):

$$g(x; p) = pq^{x-1}, \quad x = 1, 2, 3, \dots \quad (9.13)$$

Agar pembaca lebih memahaminya, misalkan suatu pabrik memiliki probabilitas 1 dari 100 barang produksinya merupakan produk cacat. Pemeriksaan dilakukan pada setiap barang tersebut. Tentukan probabilitas barang ke-5 hasil pengecekan merupakan barang yang cacat?

Dengan menggunakan Persamaan (9.13) probabilitas barang ke-5 merupakan produk gagal sebagai berikut:

$$g(5; 0.01) = (0, 01)(0, 99)^{5-1} = 0, 0096$$

Pada R probabilitas tersebut dapat dihitung menggunakan fungsi `dgeom()`. Format fungsi tersebut adalah sebagai berikut:

```
dgeom(x, prob)
```

Note:

- **x**: vektor numerik observasi dimana kejadian sukses terjadi pertama kali

- **prob:** probabilitas kejadian sukses

Dengan menggunakan fungsi tersebut, probabilitas geometris dapat dihitung seperti berikut:

```
dgeom(x=5, # observasi kejadian sukses pertama terjadi
      prob=0.01) # probabilitas kejadian sukses
```

```
## [1] 0.00951
```

Terkadang kita tertarik untuk mempelajari probabilitas kejadian sukses pertama kali berdasarkan suatu rentang observasi atau dapat didefinisikan $P(X < r)$, $P(X > r)$, atau $P(a < X < b)$ yang dapat dituliskan pada Persamaan (9.14).

$$G(r; p) = \sum_{x=0}^r g(x, p) \quad (9.14)$$

Berdasarkan contoh sebelumnya hitunglah probabilitas barang cacat pertama kali ditemukan pada observasi kurang dari sama dengan observasi ke-5?

$$P(P \leq 5) = g(0; 0.01) + g(1; 0.01) + \dots + g(5; 0.01) = 0.0585$$

Pada R fungsi yang digunakan untuk menghitung probabilitas kumulatif distribusi probabilitas geometris adalah `pgeom()`. Format yang digunakan adalah sebagai berikut:

```
pgeom(q, prob, lower.tail = TRUE)
```

Note:

- **q:** batas observasi minimum terjadi
- **prob:** probabilitas sukses
- **lower.tail:** probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE

Berdasarkan hal tersebut, maka probabilitas dapat dihitung seperti berikut:

```
pgeom(5, 0.01)
```

```
## [1] 0.05852
```

```
# atau
dgeom(0, 0.01) +
  dgeom(1, 0.01) +
  dgeom(2, 0.01) +
  dgeom(3, 0.01) +
  dgeom(4, 0.01) +
  dgeom(5, 0.01)
```

```
## [1] 0.05852
```

Nilai rata-rata dan varians distribusi probabilitas geometris disajikan pada Persamaan (9.15).

$$\mu = \frac{1}{p} \text{ dan } \sigma^2 = \frac{1-p}{p^2} \quad (9.15)$$

9.5 Distribusi Poisson

Distribusi probabilitas Poisson menggambarkan berapa kali suatu peristiwa terjadi pada sebuah interval yang spesifik. Interval dapat berupa waktu, jarak, area, atau volume.

Distribusi Poisson didasarkan pada dua asumsi. Asumsi pertama menjelaskan bahwa probabilitas proporsional terhadap panjang interval. Asumsi yang kedua adalah interval bersifat independen. Dengan kata lain semakin panjang suatu interval, semakin besar probabilitas, dan jumlah kejadian pada sebuah interval tidak mempengaruhi interval lainnya. Distribusi ini juga merupakan bentuk terbatas dari distribusi binomial dimana probabilitas keberhasilan sangat kecil dengan ukuran sampel n besar.

Probabilitas Poisson memiliki karakteristik sebagai berikut:

1. variabel acak merupakan berapa kali suatu peristiwa terjadi selama interval yang ditentukan. 2. probabilitas suatu peristiwa proporsional terhadap ukuran interval.
2. interval tidak tumpang tindih dan bersifat independen

Distribusi probabilitas Poisson pada variabel acak X , merepresentasikan jumlah luaran yang terjadi pada interval waktu yang diberikan atau wilayah yang spesifik dan dinotasikan sebagai t . Distribusi probabilitas dituliskan pada Persamaan (9.16):

$$p(x; \lambda t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}, \quad x = 0, 1, 2, \dots \quad (9.16)$$

dimana λ merupakan rata-rata jumlah luaran per satuan waktu, jarak, area, atau volume dan $e = 2,71828\dots$

Kumulatif probabilitas Poisson dituliskan berdasarkan Persamaan (9.17):

$$P(r; \lambda t) = \sum_{x=0}^r p(x; \lambda t) \quad (9.17)$$

dengan nilai rata-rata dan varians distribusinya disajikan pada Persamaan (9.18).

$$\mu \text{ dan } \sigma = \lambda t \quad (9.18)$$

Untuk lebih memahami penerapan kedua persamaan tersebut, misalkan selama melakukan eksperimen laboratorium, rata-rata jumlah partikel radioaktif yang melewati couter pada 1 milidetik sebesar 4. Berapa probabilitas 6 partikel memasuki counter pada milidetik yang diberikan?

Contoh kasus tersebut dapat diselesaikan menggunakan Persamaan (9.16) dengan nilai $x = 6$ dan $\lambda t = 4$ seperti berikut:

$$p(6; 4) = \frac{e^{-4} (4)^6}{6!} = 0,1042$$

Kita dapat menggunakan fungsi `dpois()` pada R untuk menghitung probabilitas Poisson. Format yang digunakan adalah sebagai berikut:

```
dpois(x, lambda)
```

Note:

- **x:** vektor numerik
- **lambda:** jumlah rata-rata luaran

Probabilitas 6 partikel memasuki counter berdasarkan fungsi tersebut adalah sebagai berikut:

```
dpois(x=6, lambda=4)
```

```
## [1] 0.1042
```

Contoh kasus tersebut juga dapat diselesaikan menggunakan Persamaan (9.17) dengan terlebih dahulu menghitung selisih $P(X \leq 6)$ terhadap $P(X \leq 5)$. Pada R fungsi yang digunakan adalah `ppois()`. Format fungsi tersebut adalah sebagai berikut:

```
ppois(q, lambda, lower.tail = TRUE)
```

Note:

- **q**: vektor numerik
- **lambda**: jumlah rata-rata luaran
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE

Dengan menggunakan fungsi tersebut, hasil yang diperoleh adalah sebagai berikut:

```
ppois(6,4)-ppois(5,4)
```

```
## [1] 0.1042
```

9.6 Distribusi Uniform

Distribusi uniform merupakan distribusi kontinu yang paling sederhana yang ditandai dengan fungsi densitas yang datar serta probabilitas yang seragam sepanjang interval tertutup. Distribusi ini juga disebut sebagai distribusi persegi panjang sebab bentuk distribusinya yang menyerupai persegi panjang. Fungsi densitas untuk distribusi uniform disajikan pada Persamaan (9.19).

$$f(x; A, B) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases} \quad (9.19)$$

Nilai mean dan varians dari distribusi uniform disajikan pada Persamaan (9.20).

$$\mu = \frac{A+B}{2} \text{ dan } \sigma^2 = \frac{(B-A)^2}{12} \quad (9.20)$$

Untuk memudahkan pemahaman pembaca mengenai distribusi ini, berikut penulis sajikan visualisasi distribusi uniform dengan nilai minimum=1 dan maksimum=3. Variabel acak dibuat menggunakan fungsi `runif()`. Format fungsi yang digunakan adalah sebagai berikut:

```
runif(n, min = 0, max = 1)
```

Note:

- **n**: jumlah data atau panjang variabel acak

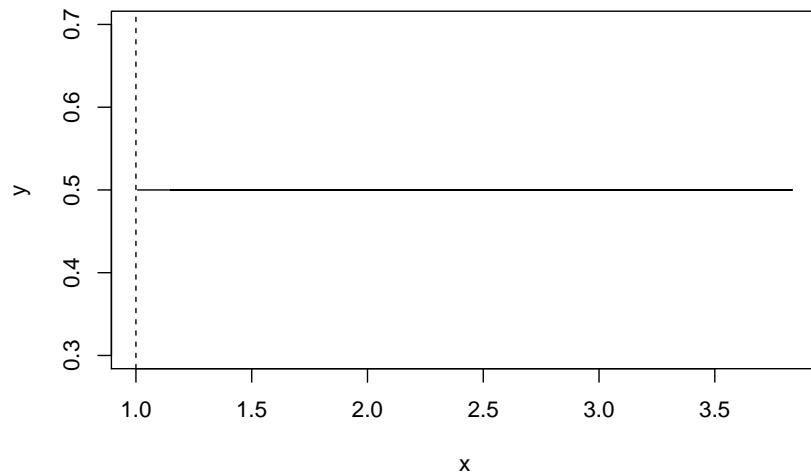


Figure 9.1: Distribusi uniform dengan nilai min 1 dan max 3

- **min:** nilai minimum variabel acak
- **max:** nilai maksimum variabel acak

Visualisasi disajikan pada Gambar 9.1.

Berdasarkan Gambar 9.1, probabilitas distribusinya adalah sebagai berikut:

$$f(x; A, B) = \begin{cases} \frac{1}{3} & 0 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

Kita juga dapat menghitung probabilitas suatu nilai melalui distribusi tersebut. Sebagai contoh, hitunglah probabilitas nilai $X \geq 3$?

$$P[X \geq 3] = \int_3^4 dx = \frac{1}{4}$$

Jika contoh tersebut divisualisasikan, maka akan tampak seperti pada Gambar 9.2.

Pada R terdapat 2 buah fungsi untuk menghitung probabilitas distribusi unifofm. Fungsi pertama adalah `dunif()` dan yang kedua adalah `punif()`. Fungsi pertama akan menghasilkan probabilitas (*likelihood*) dari suatu nilai yang kita inginkan, sedangkan fungsi kedua adalah fungsi probabilitas kumulatif yang akan menghasilkan nilai berdasarkan rentang yang dimasukkan (rentang satu arah bisa \leq atau \geq).

Format fungsi `dunif()` adalah sebagai berikut:

```
dunif(x, min = 0, max = 1)
```

Note:

- **n:** jumlah data atau panjang variabel acak
- **min:** nilai minimum variabel acak

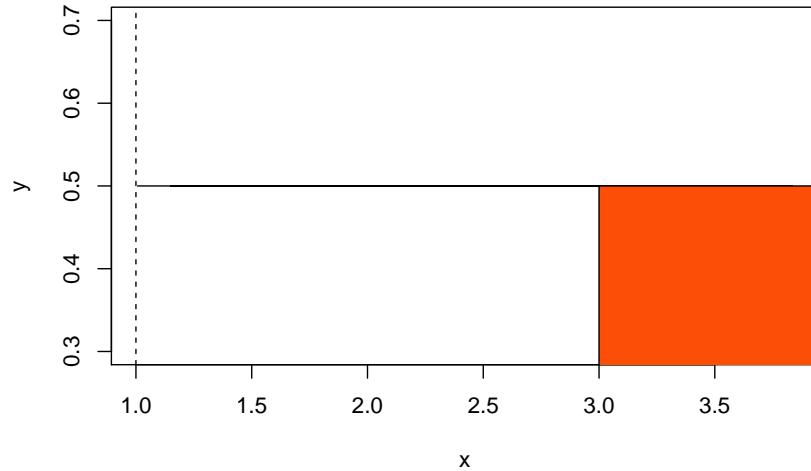


Figure 9.2: Probabilitas distribusi uniform pada rentang nilai x 3 sampai 4

- **max:** nilai maksimum variabel acak

Format fungsi `punif` adalah sebagai berikut:

```
punif(q, min = 0, max = 1, lower.tail = TRUE)
```

Note:

- **q:** vektor numerik
- **min:** nilai minimum variabel acak
- **max:** nilai maksimum variabel acak
- **lower.tail:** probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

Nilai probabilitas berdasarkan contoh soal sebelumnya merupakan contoh kasus probabilitas kumulatif sehingga digunakan fungsi `punif()` untuk menghitung probabilitasnya.

```
punif(3, min=1, max=4, lower.tail=FALSE)
```

```
## [1] 0.3333
```

9.7 Distribusi Normal

Distribusi kontinu yang paling sering digunakan dalam analisa statistik adalah **distribusi normal** atau disebut juga sebagai distribusi Gauss. Distribusi ini dicirikan dari bentuknya yang mirip dengan lonceng. Secara umum fungsi densitas distribusi normal disajikan pada Persamaan (9.21).

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty \quad (9.21)$$

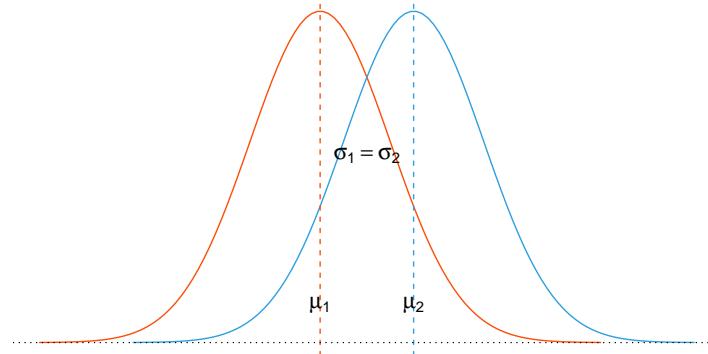


Figure 9.3: Distribusi normal dengan nilai mean sama dan simpangan baku berbeda.

dimana $\pi = 3,14159\dots$ dan $e = 2,71828\dots$

Berdasarkan persamaan di atas terdapat dua parameter penting dalam distribusi normal yaitu nilai mean μ dan simpangan baku σ . Kedua nilai tersebut akan mempengaruhi bentuk dari distribusi normal yang terbentuk. Pada contoh selanjutnya akan diberikan visualisasi mengenai bentuk distribusi normal dengan berbagai variasi mean dan simpangan baku.

- **Distribusi normal dengan μ berbeda dan σ yang sama.**

Pada Gambar 9.3 disajikan visualisasi dua buah distribusi normal dengan nilai μ sama dan σ berbeda.

- **Distribusi normal dengan μ sama dan σ yang berbeda.**

Pada Gambar 9.4 disajikan visualisasi dua buah distribusi normal dengan nilai μ berbeda dan σ sama. Perbedaan σ menyebabkan bentuk distribusi yang lebih datar. σ kecil membuat bentuk distribusi yang lebih lancip (sebaran data kecil), sedangkan σ akan berlaku sebaliknya.

- **Distribusi normal dengan μ berbeda dan σ yang berbeda.**

Pada Gambar 9.5 disajikan visualisasi dua buah distribusi normal dengan nilai μ berbeda dan σ yang berbeda pula. Perbedaan tersebut menyebabkan perbedaan letak distribusi normal serta bentuk distribusinya.

Berdasarkan visualisasi di atas, sifat-sifat dasar distribusi normal adalah sebagai berikut:

1. Modus distribusi normal berada pada titik horizontal dimana kurva berada pada posisi maksimum atau $x = \mu$.
2. Kurva berbentuk simetris terhadap nilai mean μ .
3. Kurva memiliki titik infleksi pada $x = \mu \pm \sigma$, yang cekung kebawah jika $\mu - \sigma < X < \mu + \sigma$ dan cekung ke atas pada titik diluar rentang tersebut.
4. Kurva normal mendekati sumbu horizontal asimtotik saat kita melanjutkan ke arah mana pun dari rata-rata.

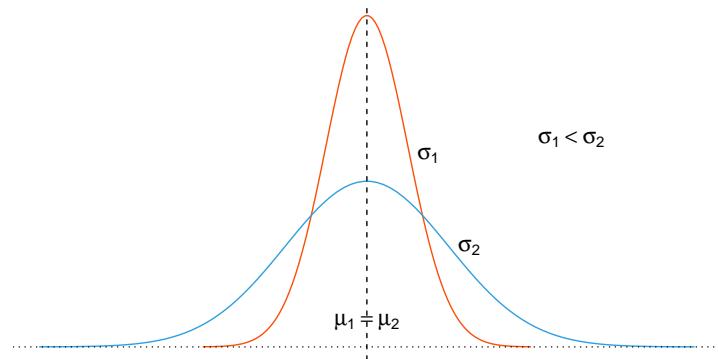


Figure 9.4: Distribusi normal dengan nilai mean sama dan simpangan baku berbeda.

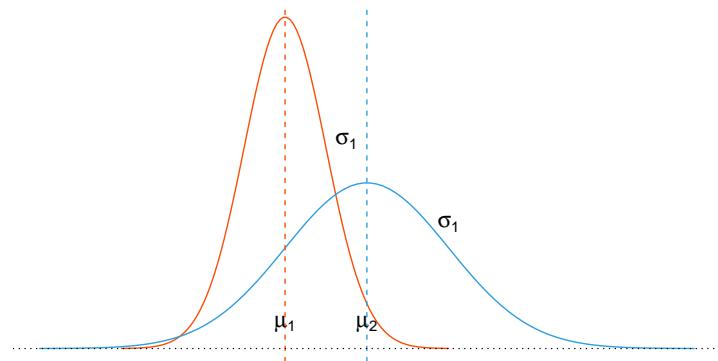


Figure 9.5: Distribusi normal dengan nilai mean berbeda dan simpangan baku berbeda.

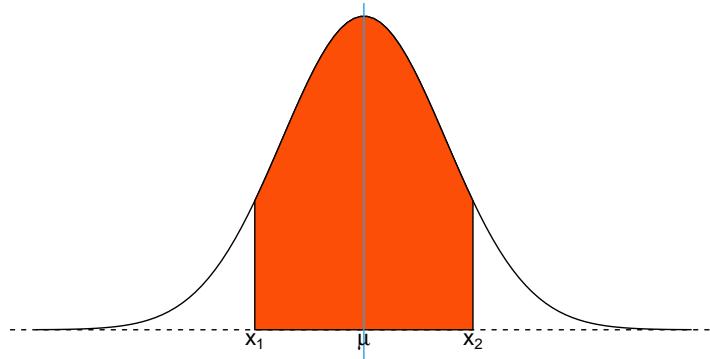


Figure 9.6: Luas area di bawah kurva normal.

5. Luas total di bawah kurva dan di atas sumbu horizontal adalah 1.

Sifat lain yang dimiliki oleh distribusi normal adalah sebagai berikut:

1. Sekitar 68% luas dibawah kurva normal berada pada kisaran 1σ dari nilai μ .
2. Sekitar 95% luas dibawah kurva normal berada pada kisaran 2σ dari nilai μ .
3. Sekitar 99,7% luas dibawah kurva normal berada pada kisaran 3σ dari nilai μ .

Secara kolektif, titik-titik ini dikenal sebagai **aturan empiris** atau **aturan 68-95-99,7**. Hal ini jelas, mengingat distribusi normal sebagian besar hasil akan berada dalam 3σ dari μ .

9.7.1 Luas Area Di Bawah Kurva Normal

Untuk memperoleh luas area dibawah distribusi normal kita perlu membuat batasan pada kurva tersebut. Dua koordinat pembatas dapat kita definisikan sebagai x_1 dan x_2 . Luas area yang diarsir (lihat Gambar 9.6) selanjutnya dihitung dengan cara mengintegalkan area yang diarsir dengan batasan dua koordinat sebelumnya atau dapat ditulis sebagai berikut:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} n(x; \mu, \sigma) dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

Menghitung luas area di bawah kurva normal bukanlah pekerjaan yang mudah dilakukan sehingga terkadang kita memerlukan alat bantu untuk melakukan proses perhitungan. Alat bantu yang umum digunakan adalah Tabel distribusi normal standard (distribusi normal dengan $\mu 0$ dan $\sigma^2 1$) yang dapat pembaca unduh pada tautan [berikut](#). Untuk dapat menggunakan Tabel tersebut kita perlu mengubah nilai x_1 dan x_2 pada Gambar 9.6 menjadi Z . Untuk melakukannya kita dapat menggunakan Persamaan (9.22).

$$Z = \frac{X - \mu}{\sigma} \tag{9.22}$$

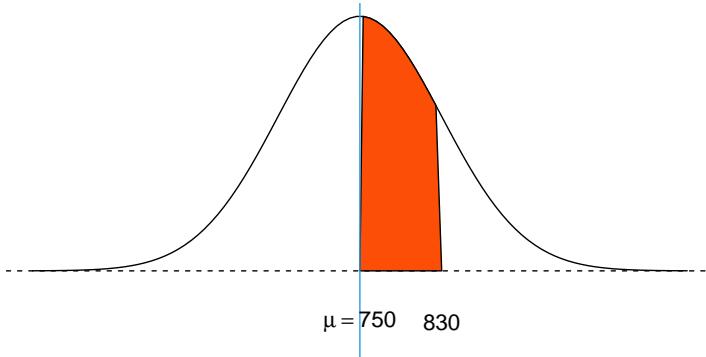


Figure 9.7: Luas area masa layan lampu antara 750 sampai 830 jam.

Jika kita lihat berdasarkan Gambar 9.6, luas area yang diblok dapat dihitung dengan Tabel distribusi normal. Luas area yang dicari dihitung sebagai berikut:

$$P(x_1 < z \leq x_2) = P(z < x_2) + P(z > x_1)$$

Agar pembaca lebih memahami penerapan distribusi ini, misalkan kita diminta untuk menghitung probabilitas masa layan lampu. Rerata masa layan lampu suatu produk adalah 750 jam dengan simpangan baku 80 jam. Distribusi dari masa layan lampu tersebut diasumsikan mengikuti distribusi normal. Hitunglah probabilitas:

- Lampu memiliki masa layan antara 750 sampai 830 jam?
- Lampu dengan masa layan tepat 830 jam?
- Lampu dengan masa layan lebih dari atau sama dengan 830 jam

Masa layan lampu antara 750 sampai 830 jam

Distribusi dan luasan yang dicari dapat digambarkan berdasarkan Gambar 9.7 berikut:

Nilai rentang perlu dikonversi kedalam nilai distribusi normal standard menggunakan Persamaan (9.22). Berikut adalah proses perhitungannya:

$$Z_{750} = \frac{750 - 750}{80} = 0 \quad \text{dan} \quad Z_{830} = \frac{830 - 750}{80} = 1$$

Dengan menggunakan Tabel distribusi normal, probabilitas yang diinginkan dapat dihitung sehingga diperoleh nilai probabilitas sebagai berikut:

$$P(z < 1) = 0,3413$$

Masa layan lampu tepat 830 jam

Pertanyaan dapat dijawab menggunakan Persamaan (9.21). Hasil yang diperoleh adalah sebagai berikut:

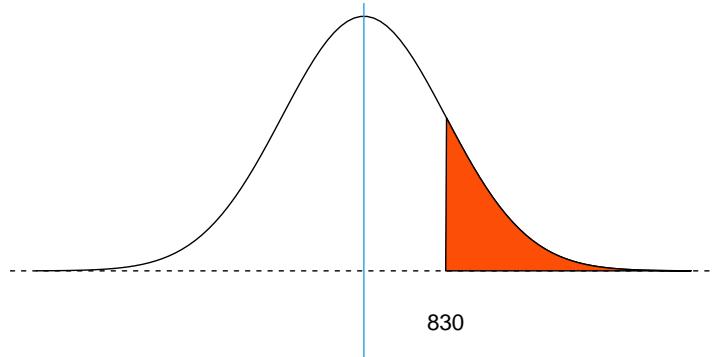


Figure 9.8: Luas area masa layan lampu lebih dari atau sama dengan 830 jam.

$$n(750; 750, 80) = \frac{1}{\sqrt{2\pi(80)}} e^{-\frac{1}{2(80)^2}((750)-(750))^2} = 0.003$$

Masa layan lampu lebih dari atau sama dengan 830 jam

Distribusi dan luasan yang dicari dapat digambarkan berdasarkan Gambar 9.8 berikut:

Probabilitas berdasarkan luas area tersebut dapat dihitung seperti berikut:

$$P(z > 1) = 0,5 - 0,3413 = 0,1578$$

Pada R probabilitas distribusi normal dapat dihitung menggunakan dua buah fungsi yaitu `dnorm()` (probabilitas distribusi normal) dan `pnorm()` (probabilitas kumulatif distribusi normal). Format yang digunakan adalah sebagai berikut:

```
dnorm(x, mean = 0, sd = 1)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)
```

Note:

- `x,p`: vektor numerik atau kuantil
- `mean`: rata-rata populasi
- `sd`: simpangan baku populasi

Berikut adalah contoh penyelesaian contoh soal masa layan lampu menggunakan sintaks R:

```
# masa layan 750 sampai 830 jam
pnorm(q=830, mean=750, sd=80) - pnorm(q=750, mean=750, sd=80)
```

```
## [1] 0.3413
```

```
# masa layan 830 jam
dnorm(x=830, mean=750, sd=80)

## [1] 0.003025

# masa layan lebih dari sama dengan 830 jam
pnorm(q=830, mean=750, sd=80, lower.tail=FALSE)

## [1] 0.1587
```

9.7.2 Uji Kecocokan Distribusi Normal

Uji kecocokan distribusi data apakah distribusi tersebut berdistribusi normal dapat dilakukan dengan dua cara yaitu: analisis grafik dan analisis numerik. Analisis grafik dilakukan dengan menggunakan grafik yaitu histogram, density plot, QQ-plot dan ECDF. Dalam Chapter sebelumnya telah penulis jelaskan bahwa ECDF dapat digunakan untuk menguji kecocokan distribusi secara umum. Metode lain yang dapat digunakan adalah metode numerik menggunakan metode Shapiro-Wilk (SW), Shapiro-Francia (SF), dll. Dalam buku ini penulis hanya akan menjelaskan uji kecocokan distribusi normal menggunakan metode Shapiro-Wilk.

Kita akan menggunakan dataset `airquality` yang merupakan data pengukuran kualitas udara New York bulan Mei sampai September 1973. Pertama-tama kita perlu melihat ringkasan data dari dataset tersebut. Berikut adalah sintaks untuk melakukannya:

```
library(tibble)

glimpse(airquality)

## Observations: 153
## Variables: 6
## $ Ozone    <int> 41, 36, 12, 18, NA, 28, 23, 19, 8, ...
## $ Solar.R  <dbl> 190, 118, 149, 313, NA, NA, 299, 9...
## $ Wind     <dbl> 7.4, 8.0, 12.6, 11.5, 14.3, 14.9, ...
## $ Temp     <int> 67, 72, 74, 62, 56, 66, 65, 59, 61...
## $ Month    <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5...
## $ Day      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, ...
```

Kita akan mengecek distribusi dari konsentrasi ozon apakah berdistribusi normal atau tidak, kita dapat memvisualisasikannya menggunakan grafik dalam contoh ini adalah grafik QQ-plot dan ECDF. Berikut adalah visualisasi yang dihasilkan

```
# install.packages("ggpubr")
library(ggplot2)
library(dplyr)
library(ggpubr)
```

Berdasarkan kedua grafik tersebut terlihat bahwa titik observasi tidak mengikuti garis referensi distribusi normal sehingga dapat disimpulkan bahwa distribusi konsentrasi ozon tidak berdistribusi normal.

Metode lain yang digunakan untuk melakukan uji kecocokan terhadap distribusi normal adalah dengan menggunakan metode Shapiro-Wilk (SW). Metode ini merupakan metode nonparametrik yang memiliki power yang cukup besar untuk ukuran sampel relatif kecil (< 2000). Perintah yang digunakan pada R untuk melakukan uji SW adalah `shapiro.test()`. Format yang digunakan adalah sebagai berikut:

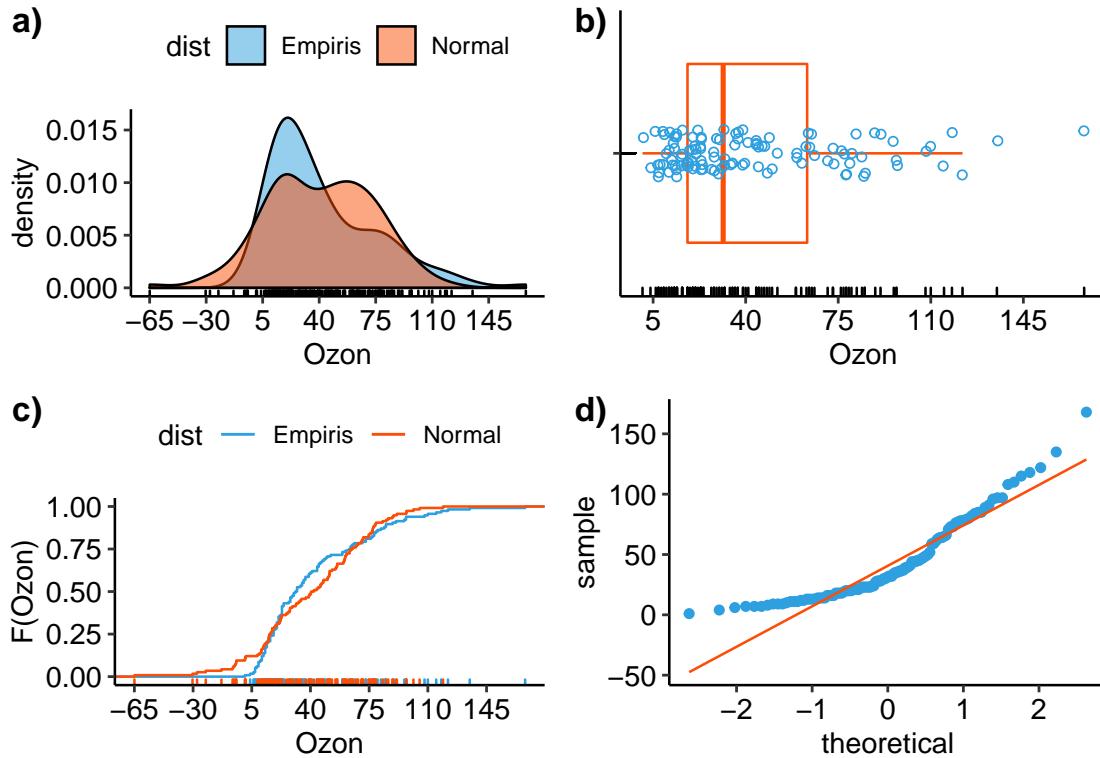


Figure 9.9: Visualisasi distribusi konsentrasi ozon Kota New York a)density plot, b)boxplot, c)ecdf, d)qq-plot

```
shapiro.test(x)
```

Note:

x: vektor numerik.

Pengujian ini merupakan suatu bentuk pengujian statistik sehingga melibatkan dua buah hipotesis. Pengujian statistik tidak akan dijelaskan secara detail pada Chapter ini. Hipotesis yang digunakan adalah sebagai berikut

H_0 : Sampel berdistribusi normal

H_1 : Sampel data tidak berdistribusi normal

Berikut adalah uji kecocokan yang dilakukan pada R:

```
shapiro.test(airquality$Ozone)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: airquality$Ozone  
## W = 0.88, p-value = 3e-08
```

Berdasarkan hasil perhitungan diperoleh nilai p-value sebesar 2.79e-08. Dengan menggunakan tingkat kepercayaan 95% (error=5%), dapat disimpulkan bahwa distribusi ozon tidak berdistribusi normal (p-value < error).

9.7.3 Pendekatan Distribusi Binomial Menggunakan Distribusi Normal

Nilai probabilitas distribusi diskrit seperti distribusi binomial dapat dihitung dengan menggunakan pendekatan distribusi binomial. Pendekatan ini berlaku jika jumlah sampel yang digunakan sangat besar. Dengan menggunakan pendekatan ini, kita dapat menghitung probabilitas kumulatif distribusi binomial menggunakan Tabel distribusi normal.

Jika X merupakan variabel acak binomial dengan mean $\mu = np$ dan varians $\sigma^2 = npq$, maka bentuk batas distribusi z dengan pendekatan distribusi normal standard dituliskan kedalam Persamaan (9.23).

$$Z = \frac{X - np}{\sqrt{npq}} \quad (9.23)$$

dimana $n \rightarrow \infty$ merupakan distribusi normal standard $n(z; 0, 1)$. Nilai z yang diperoleh selanjutnya dapat digunakan untuk mencari luasan dibawah kurva normal menggunakan Tabel distribusi normal.

Untuk memahami penerapannya, misalkan diketahui probabilitas suatu pompa air rusak disuatu kawasan adalah 0,4. Jika pada kawasan tersebut terdapat 100 buah pompa. Hitunglah probabilitas jumlah pompa rusak di kawasan tersebut kurang dari 30? ?

Pada kasus tersebut kejadian sukses didefinisikan jika terdapat pompa yang rusak. Untuk menghitungnya pertama-tama kita perlu menghitung nilai mean dan simpangan baku dari populasinya.

$$\mu = np = (100)(0,4) = 40 \quad \text{dan} \quad \sigma = \sqrt{npq} = \sqrt{(100)(0,4)(0,6)} = 4,899$$

Selanjutnya dihitung nilai z dengan nilai $X = 29,5$ berdasarkan Persamaan (9.23).

$$Z = \frac{29,5 - 40}{\sqrt{4,899}} = -2,14$$

Dengan menggunakan tabel distribusi normal standard, probabilitas dari kejadian tersebut adalah sebagai berikut (lihat Gambar 9.8):

$$P(X < 30) \approx P(Z < -2,14) = 0,0162.$$

Pada R probabilitas dari peristiwa tersebut dapat dihitung seperti berikut:

```
pnorm(q=29.5,mean=40,sd=4.899)
```

```
## [1] 0.01604
```

9.8 Distribusi Gamma dan Eksponensial

Distribusi eksponensial dan gamma merupakan distribusi yang berperan penting dalam menjelaskan teori antrian dan masalah reliabilitas. Waktu antara kedatangan di fasilitas layanan dan waktu kegagalan komponen bagian dan sistem kelistrikan sering dimodelkan dengan baik oleh distribusi eksponensial. Hubungan antara gamma dan eksponensial memungkinkan gamma digunakan dalam jenis masalah yang serupa.

Variabel acak kontinu X memiliki distribusi gamma, dengan parameter α dan β . Fungsi densitas dari distribusi tersebut dituliskan kedalam Persamaan (9.24).

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (9.24)$$

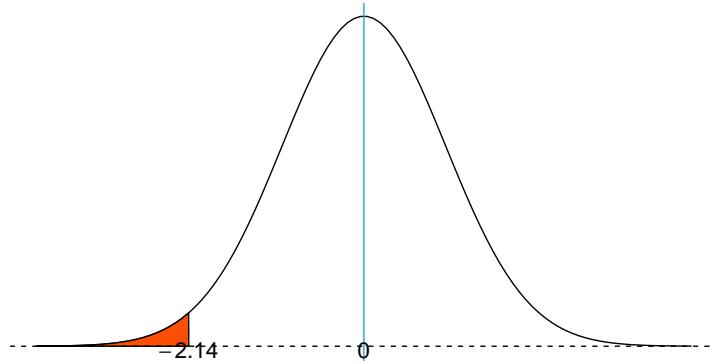


Figure 9.10: Luas area jumlah pompa rusak kurang dari 30.

dimana $\Gamma(\alpha)$ merupakan fungsi gamma yang dituliskan pada Persamaan (9.25)

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \quad (9.25)$$

Pada Gambar 9.11 disajikan visualisasi distribusi gamma dengan variasi α dan β .

Distribusi eksponensial merupakan kasus khusus dari distribusi gamma dengan nilai $\alpha = 1$. Distribusi probabilitas eksponensial dituliskan kedalam Persamaan (9.26).

$$f(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (9.26)$$

dimana $\beta > 0$.

Persamaan-persamaan berikut merupakan cara untuk menghitung mean dan varians dari kedua distribusi tersebut. Persamaan (9.27) merupakan persamaan untuk menghitung kedua nilai tersebut untuk distribusi gamma, sedangkan Persamaan (9.28) digunakan pada distribusi eksponensial.

$$\mu = \alpha\beta \quad \text{dan} \quad \sigma^2 = \alpha\beta^2 \quad (9.27)$$

$$\mu = \beta \quad \text{dan} \quad \sigma^2 = \beta^2 \quad (9.28)$$

Misalkan dalam suatu kawasan terdapat sistem penyediaan air minum lingkup kecil dengan komponen utama pompa, dimana waktu dalam tahun dimana pompa tersebut gagal beroperasi (rusak) disimbolkan sebagai T . Variabel acak T dimodelkan dengan cukup baik menggunakan distribusi eksponensial dengan waktu sampai pompa tersebut gagal berfungsi $\beta = 5$. Jika 5 buah pompa dipasang pada sistem yang berbeda, berapa probabilitas setidaknya 2 buah pompa masih berfungsi hingga akhir tahun ke-8?

Probabilitas suatu pompa masih dapat berfungsi setelah 8 tahun dari kasus tersebut dapat dituliskan sebagai berikut:

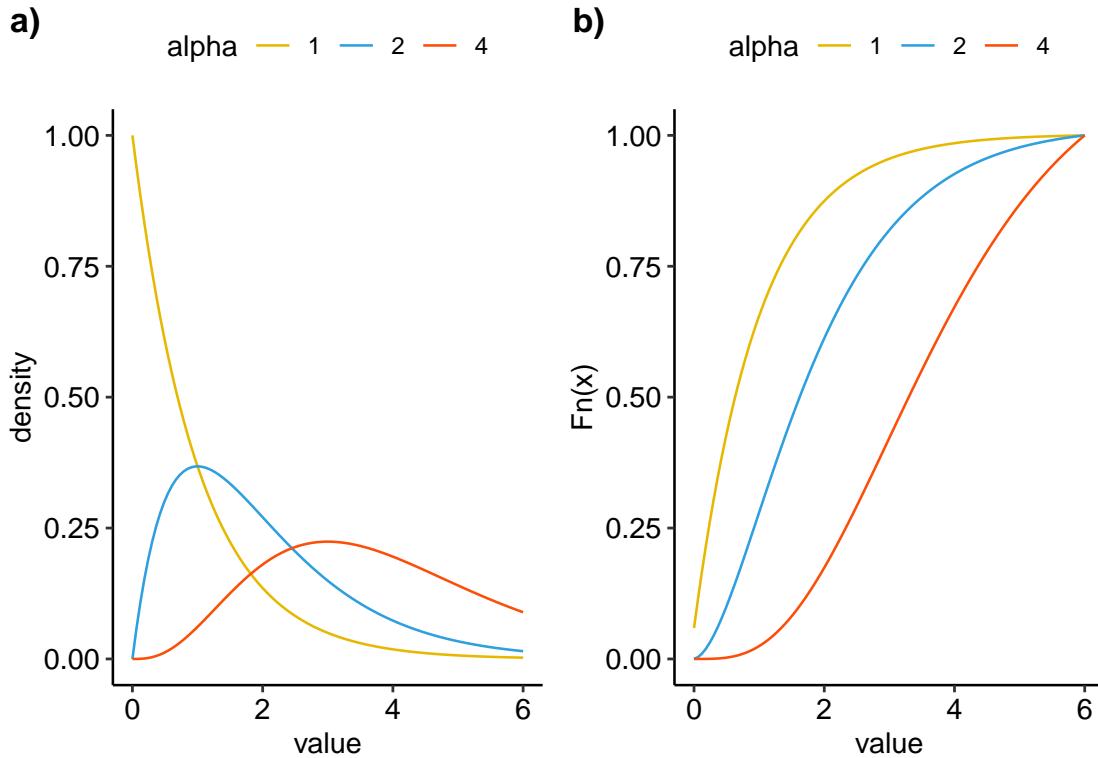


Figure 9.11: Visualisasi distribusi gamma dengan variasi alpha dengan beta 1 a) density plot, b)ecdf

$$P(T > 8) = \frac{1}{5} \int_0^{\infty} e^{-\frac{t}{5}} dt = e^{-\frac{8}{5}} \approx 0,2.$$

Probabilitas sedikitnya 2 buah pompa yang masih beroperasi setelah 8 tahun dapat dihitung menggunakan probabilitas kumulatif distribusi binomial seperti berikut:

$$P(X \geq 2) = \sum_{x=2}^5 b(x; 5, 0.2) = 1 - \sum_{x=0}^1 b(x; 5, 0.2) = 1 - 0,7373 = 0,2627$$

Pada R probabilitas gamma dapat dihitung menggunakan 2 fungsi yaitu `dgamma()` dan `pgamma()` (probabilitas kumulatif). Format fungsi tersebut adalah sebagai berikut:

```
dexp(x, rate=1)
pexp(q, rate=1, lower.tail = TRUE)
```

Note:

- **x,p**: vektor numerik atau kuantil
- **rate**: nilai 1/beta
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

Berikut adalah sintaks yang digunakan untuk menghitung probabilitas pada contoh soal tersebut:

```
# probabilitas pompa masih berfungsi lebih dari 8 tahun
pexp(q=8, rate=1/5, lower.tail=FALSE)
```

```
## [1] 0.2019
```

```
# probabilitas sedikitnya 2 pompa yang masih berfungsi
pbinom(q=1, size=5, prob=0.2018965, lower.tail=FALSE)
```

```
## [1] 0.2666
```

Contoh lainnya misalkan pada pengujian toksikan terhadap hewan uji untuk menentukan dosis mematikan pada suatu hewan uji. Toksikan yang digunakan merupakan logam berat yang ada di perairan. Berdasarkan dosis tertentu, hasil percobaan menentukan bahwa *survival time*, dalam minggu dari hewan uji, memiliki distribusi gamma dengan $\alpha = 5$ dan $\beta = 10$. Hitunglah probabilitas hewan uji tidak dapat selamat tidak lebih dari 60 minggu?

Diberikan variabel acak X yang menyatakan *survival time* (waktu hingga mati). Probabilitas yang terjadi dapat dituliskan sebagai berikut:

$$P(X \leq 60) = \frac{1}{\beta^5} \int_0^{\infty} \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(5)} dx$$

Integral pada persamaan diatas dapat diselesaikan dengan menggunakan **incomplete gamma function**, yang menjadikan persamaan di atas menjadi fungsi distribusi kumulatif untuk distribusi gamma yang dapat dituliskan kembali seperti berikut:

$$F(x; \alpha) = \int_0^x \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)} dy$$

jika diberikan $y = \frac{x}{\beta}$ dan $x = \beta y$, persamaan tersebut dapat dituliskan lagi menjadi berikut:

$$P(X \leq 60) = \int_0^6 \frac{y^4 e^{-y}}{\Gamma(5)} dy$$

yang dapat dituliskan sebagai $F(6; 5)$ pada tabel **incomplete gamma function** yang dapat pembaca lihat pada buku yang ditulis oleh Ronald E. Walpole (2012) pada Appendix A.23. Berdasarkan tabel tersebut diperoleh nilai probabilitas sebagai berikut:

$$P(X \leq 60) = F(6; 5) = 0,715$$

Pada R probabilitas distribusi gamma dapat dihitung menggunakan fungsi `dgamma()` dan `pgamma()` (probabilitas kumulatif). Format fungsi tersebut adalah sebagai berikut:

```
dgamma(x, shape, scale)
pgamma(q, shape, scale, lower.tail = TRUE)
```

Note:

- **x,p**: vektor numerik atau kuantil
- **shape**: nilai alpha
- **scale**: nilai beta

- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

Berikut adalah nilai probabilitas dari contoh kasus tersebut:

```
pgamma(q=60, shape=5, scale=10)
```

```
## [1] 0.7149
```

9.9 Distribusi Chi-Square, Student's t, dan Snedecor's F.

Pada sub-chapter kali ini penulis akan menjelaskan sejumlah distribusi probabilitas yang memegang peranan penting dalam statistika inferensi. Distribusi ini tidak akan penulis jelaskan secara detail karena akan terdapat porsi tersendiri pada chapter selanjutnya.

9.9.1 Distribusi Chi-Square

Distribusi chi-square merupakan kasus khusu lain dari distribusi gamma, dimana nilai α dan β dari distribusi ini masing-masing adalah $\alpha = \nu/2$ dan $\beta = 2$ dengan nilai ν merupakan integer positif. Distribusi ini memiliki parameter tunggal yaitu ν yang disebut sebagai *degrees of freedom* (derajat kebebasan). Fungsi densitas distribusi ini disajikan pada Persamaan (9.29).

$$f(x; \nu) = \begin{cases} \frac{1}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}} & x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (9.29)$$

dimana ν adalah integer positif.

Pada Gambar 9.12 disajikan visualisasi distribusi chi-square dengan variasi ν .

Pada R terdapat dua buah fungsi yang digunakan untuk menghitung probabilitas distribusi chi-square yaitu `dchisq()` dan `pchisq()` (probabilitas kumulatif). Format fungsi tersebut adalah sebagai berikut:

```
dgamma(x, df)
pgamma(q, df, lower.tail = TRUE)
```

Note:

- **x,p**: vektor numerik atau kuantil
- **df**: derajat kebebasan ($n-1$)
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

9.9.2 Distribusi Student's t

Distribusi student's t merupakan kasus lain dari distribusi gamma. Fungsi densitas probabilitasnya disajikan pada Persamaan (9.30).

$$\frac{\Gamma\left[\frac{(r+1)}{2}\right]}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{x^2}{r}\right)^{-\frac{(r+1)}{2}}, \quad -\infty < x < \infty \quad (9.30)$$

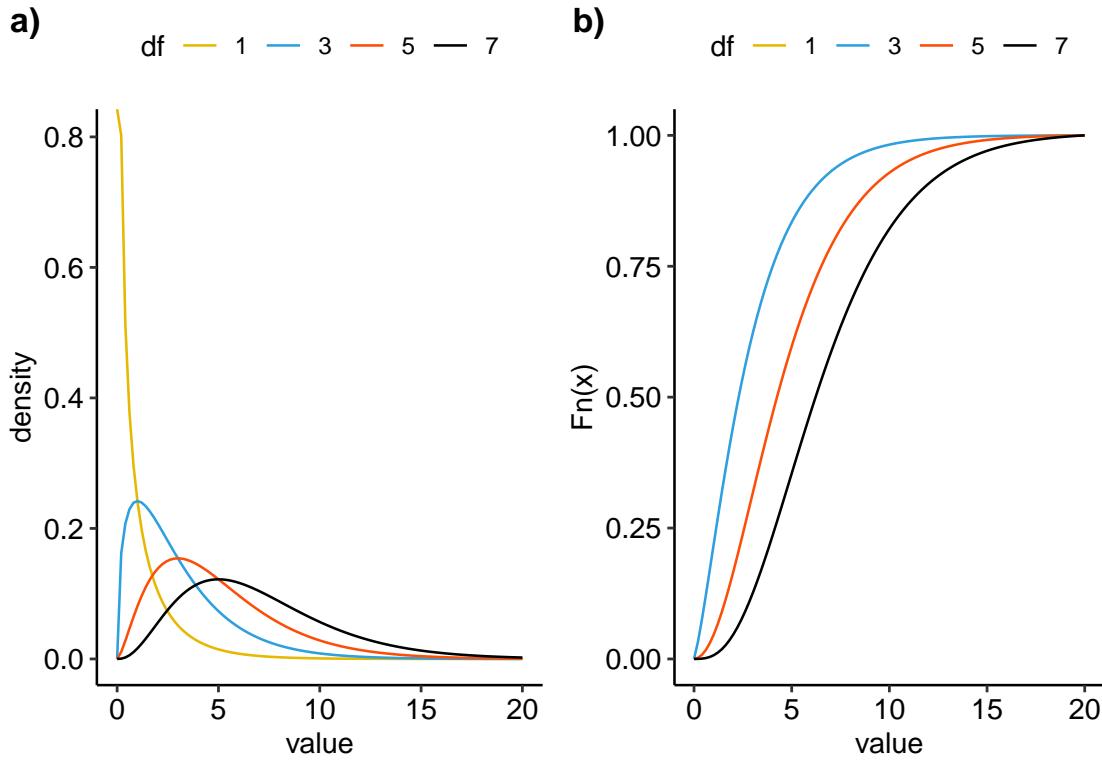


Figure 9.12: Visualisasi distribusi chi-square dengan variasi derajat kebebasan a) density plot, b)ecdf

dimana r merupakan derajat kebebasan atau $(n - 1)$.

Pada Gambar 9.13 disajikan visualisasi distribusi t dengan variasi r .

Jika kita perhatikan dengan seksama terlihat bahwa peningkatan derajat kebebasan akan membuat distribusi yang dihasilkan semakin mendekati kurva normal.

Pada R terdapat dua fungsi yang berguna untuk menghitung probabilitas distribusi t yaitu `dt()` dan `pt()` (probabilitas kumulatif). Format fungsi yang digunakan adalah sebagai berikut:

```
dt(x, df)
pt(q, df, lower.tail = TRUE)
```

Note:

- **x,p**: vektor numerik atau kuantil
- **df**: derajat kebebasan ($n-1$)
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

9.9.3 Distribusi Snedecor's F

Fungsi densitas probabilitas distribusi F disajikan pada Persamaan (9.31).

$$f(x; df_1, df_2) = \begin{cases} \frac{\Gamma\left[\frac{(df_1+df_2)}{2}\right]}{\Gamma\left(\frac{df_1}{2}\right)\Gamma\left(\frac{df_2}{2}\right)} \left(\frac{df_1}{df_2}\right)^{\frac{df_1}{2}} x^{\frac{df_1}{2}-1} \left(1 + \frac{df_1}{df_2}x\right)^{-\frac{(df_1+df_2)}{2}} & x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (9.31)$$

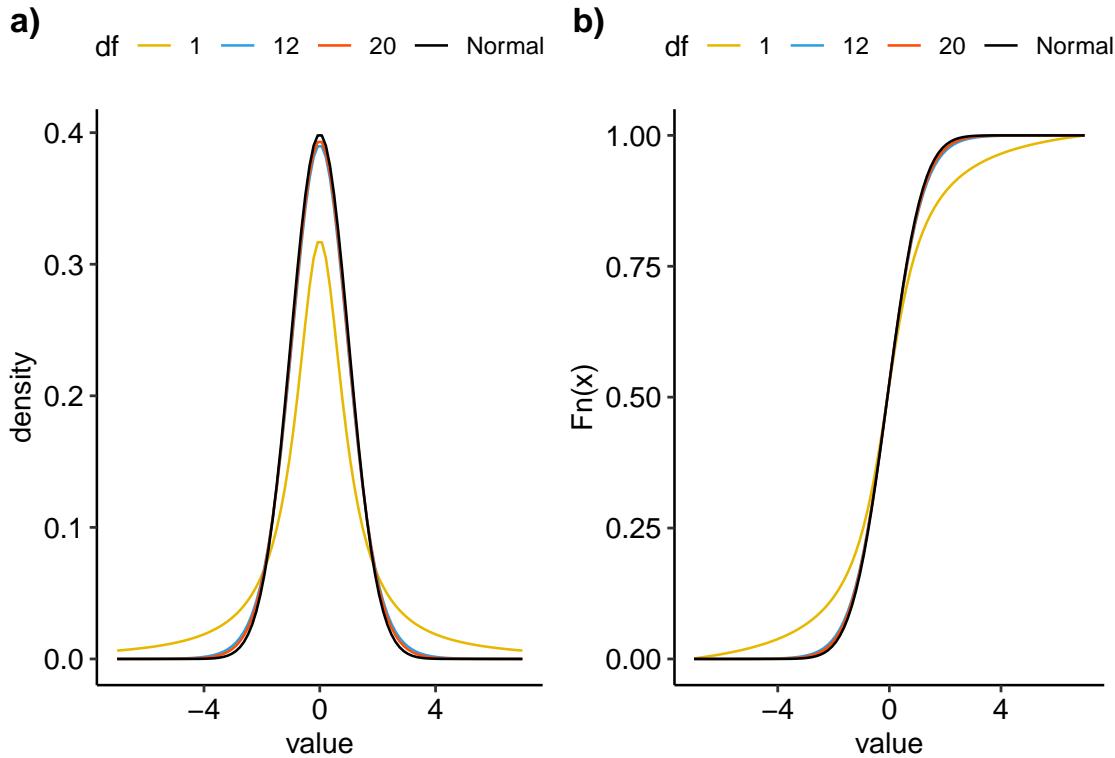


Figure 9.13: Visualisasi distribusi t dengan variasi derajat kebebasan a) density plot, b)ecdf

dimana df_1 dan df_2 merupakan derajat kebebasan.

Pada Gambar 9.14 disajikan visualisasi distribusi F dengan variasi df_1 dan df_2 .

Pada R terdapat dua fungsi yang berguna untuk menghitung probabilitas distribusi F yaitu `df()` dan `pf()` (probabilitas kumulatif). Format fungsi yang digunakan adalah sebagai berikut:

```
dt(x, df1, df2)
pt(q, df1, df2, lower.tail = TRUE)
```

Note:

- **x,p**: vektor numerik atau kuantil.
- **df1,df2**: derajat kebebasan.
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

9.10 Distribusi Kontinu Lainnya

9.10.1 Distribusi Beta

Distribusi ini merupakan perluasan dari distribusi uniform. Distribusi ini didasarkan pada fungsi beta yang disajikan pada Persamaan (9.32).

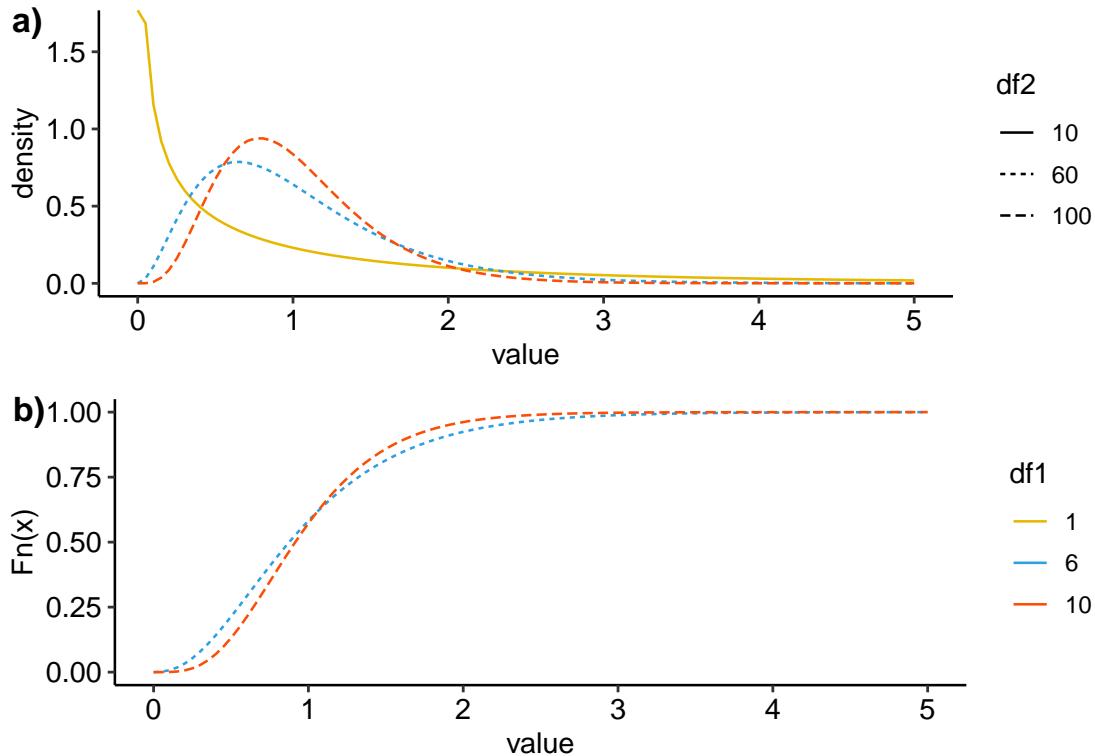


Figure 9.14: Visualisasi distribusi F dengan variasi derajat kebebasan a) density plot, b)ecdf

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)}, \quad \text{untuk } \alpha, \beta > 0 \quad (9.32)$$

dimana $\Gamma(\alpha)$ merupakan fungsi gamma.

Suatu variabel acak kontinu X memiliki distribusi beta dengan parameter $\alpha > 0$ dan $\beta > 0$ jika densitas fungsiya diberikan pada Persamaan (9.33).

$$f(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad (9.33)$$

Nilai mean dan varians distribusi beta dituliskan kedalam Persamaan (9.34).

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \text{dan} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} \quad (9.34)$$

Pada Gambar 9.15 disajikan visualisasi distribusi beta dengan variasi α dan β .

Pada R terdapat dua fungsi yang berguna untuk menghitung probabilitas distribusi beta yaitu `dbeta()` dan `pbeta()` (probabilitas kumulatif). Format fungsi yang digunakan adalah sebagai berikut:

```
dbeta(x, shape1, shape2)
pbeta(q, shape1, shape2, lower.tail = TRUE)
```

Note:

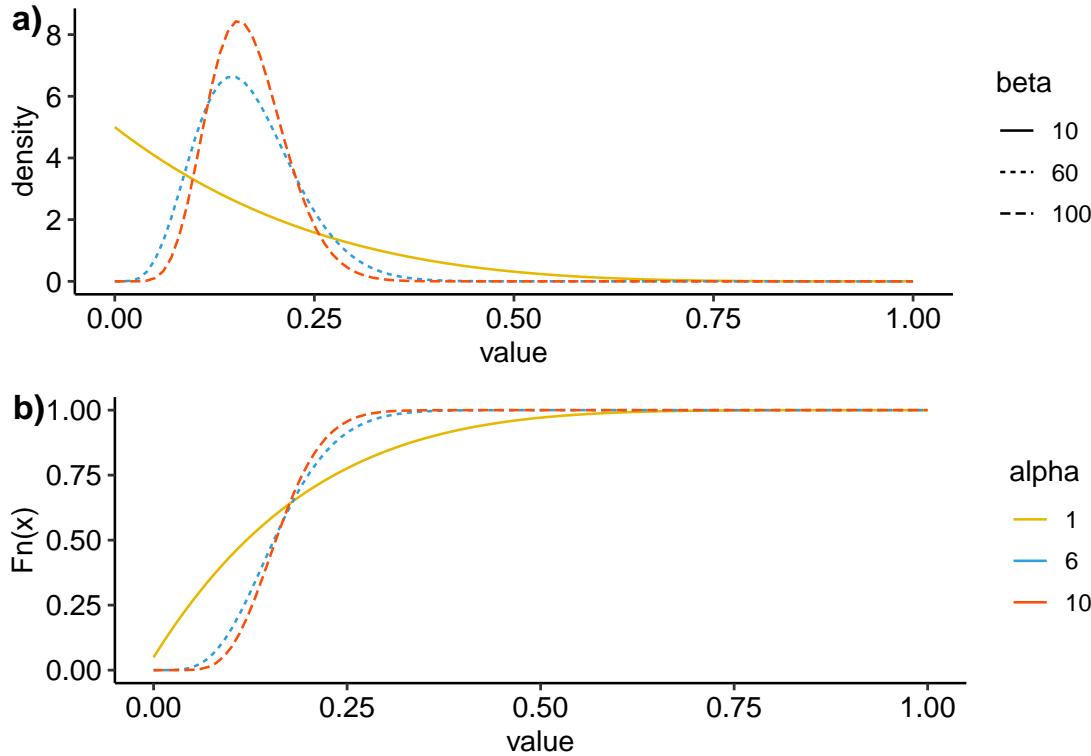


Figure 9.15: Visualisasi distribusi beta dengan variasi derajat kebebasan a) density plot, b)ecdf

- **x,p**: vektor numerik atau kuantil.
- **shape1**: alpha.
- **shape2**: beta.
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

9.10.2 Distribusi Lognormal

Distribusi lognormal telah digunakan pada berbagai aplikasi yang luas. Distribusi berlaku dalam kasus di mana transformasi log natural menghasilkan distribusi normal.

Suatu variabel acak X berdistribusi lognormal jika variabel acak $Y = \ln(X)$ berdistribusi normal dengan nilai mean μ dan simpangan baku σ . Fungsi densitas yang digunakan disajikan pada Persamaan (9.35).

$$f(x; \mu, \sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (9.35)$$

Nilai mean dan varians distribusi lognormal dihitung menggunakan Persamaan (9.36).

$$\mu = e^{\mu + \frac{\sigma^2}{2}} \quad \text{dan} \quad \sigma^2 = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1) \quad (9.36)$$

Pada Gambar 9.16 disajikan visualisasi distribusi lognormal dengan variasi μ dan σ .

Pada R, fungsi utama yang digunakan untuk menghitung probabilitas distribusi lognormal adalah `dlnorm()` dan `pnorm()` (probabilitas kumulatif). Format fungsi tersebut adalah sebagai berikut:

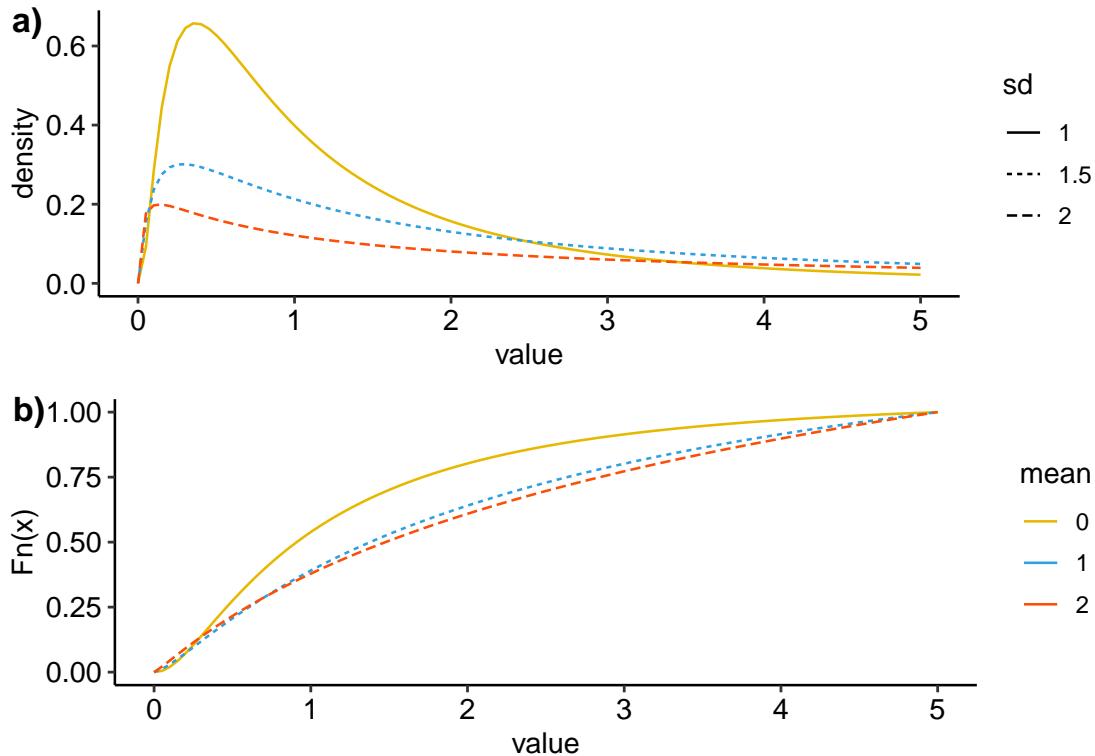


Figure 9.16: Visualisasi distribusi beta dengan variasi mean dan sd a) density plot, b)ecdf

```
dlnorm(x, meanlog = 0, sdlog = 1)
plnorm(q, meanlog = 0, sdlog = 1, lower.tail = TRUE)
```

Note:

- **x,p**: vektor numerik atau kuantil.
- **meanlog**: mean dalam bentuk logaritmik.
- **sdlog**: simpangan baku dalam bentuk logaritmik.
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

9.10.3 Distribusi Cauchy

Distribusi Cauchy merupakan kasus khusus dari distribusi t ketika nilai r atau derajat kebebasannya adalah 1. Distribusi ini tampak mirip dengan distribusi normal, namun dengan ekor yang lebih panjang. Parameter utama dari distribusi probabilitas ini adalah β dan m atau median. Fungsi densitas probabilitasnya dituliskan kedalam Persamaan (9.37).

$$f(x; m, \beta) = \frac{1}{\beta \pi} \left[1 + \left(\frac{x - m}{\beta} \right)^2 \right]^{-1}, \quad -\infty < x < \infty \quad (9.37)$$

Pada Gambar 9.17 disajikan visualisasi distribusi Cauchy dengan variasi m dan β .

Pada R, fungsi utama yang digunakan untuk menghitung probabilitas distribusi Cauchy adalah `dcauchy()` dan `pcauchy()` (probabilitas kumulatif). Format fungsi tersebut adalah sebagai berikut:

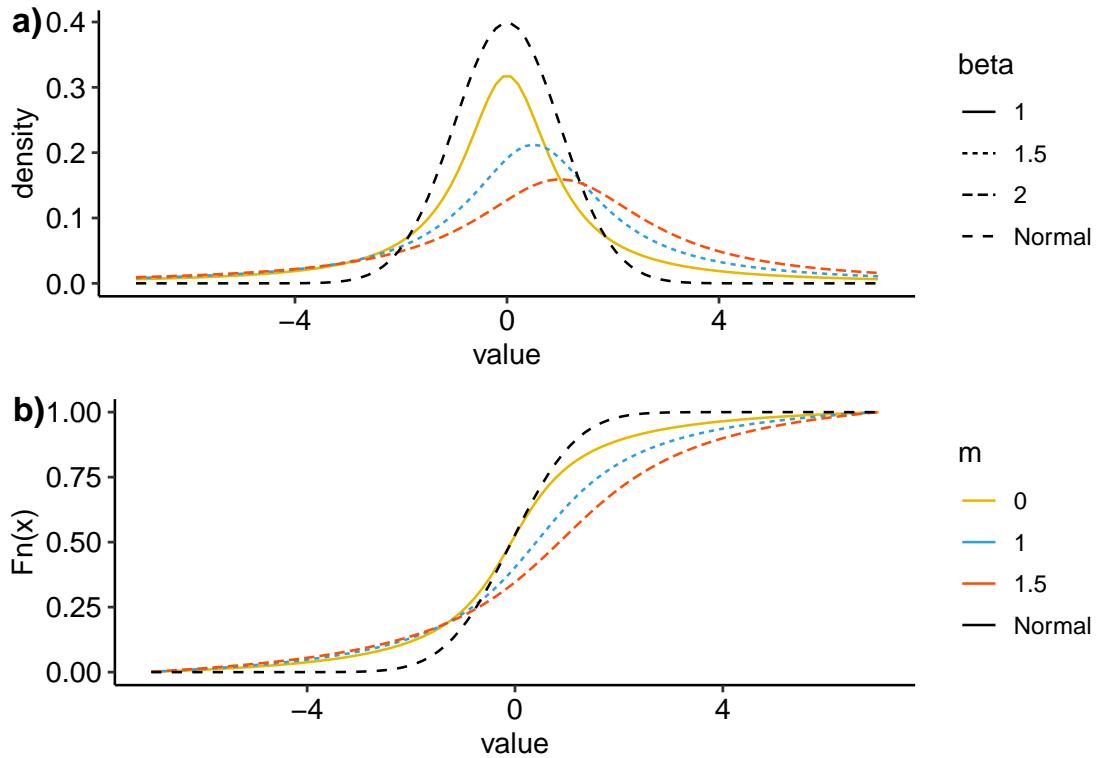


Figure 9.17: Visualisasi distribusi t dengan variasi m dan beta a) density plot, b)ecdf

```
dcauchy(x, location = 0, scale = 1)
pcauchy(q, location = 0, scale = 1, lower.tail = TRUE)
```

Note:

- **x,p**: vektor numerik atau kuantil.
- **location**: median.
- **scale**: beta.
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

9.10.4 Distribusi Logistik

Distribusi logistik sering diterapkan untuk memodelkan pertumbuhan populasi berdasarkan asumsi tertentu seperti keterbatasan lahan atau ruang atau bahkan makanan. Fungsi densitas probabilitas distribusi logistik disajikan pada Persamaan (9.38).

$$f(x; \mu, \sigma) = \frac{1}{\sigma} \exp\left(-\frac{x-\mu}{\sigma}\right) \left[1 + \exp\left(-\frac{x-\mu}{\sigma}\right)\right]^{-2}, \quad -\infty < x < \infty \quad (9.38)$$

Pada Gambar 9.18 disajikan visualisasi distribusi Cauchy dengan variasi m dan β .

Pada R, fungsi utama yang digunakan untuk menghitung probabilitas distribusi Logistik adalah `dlogis()` dan `plogis()` (probabilitas kumulatif). Format fungsi tersebut adalah sebagai berikut:

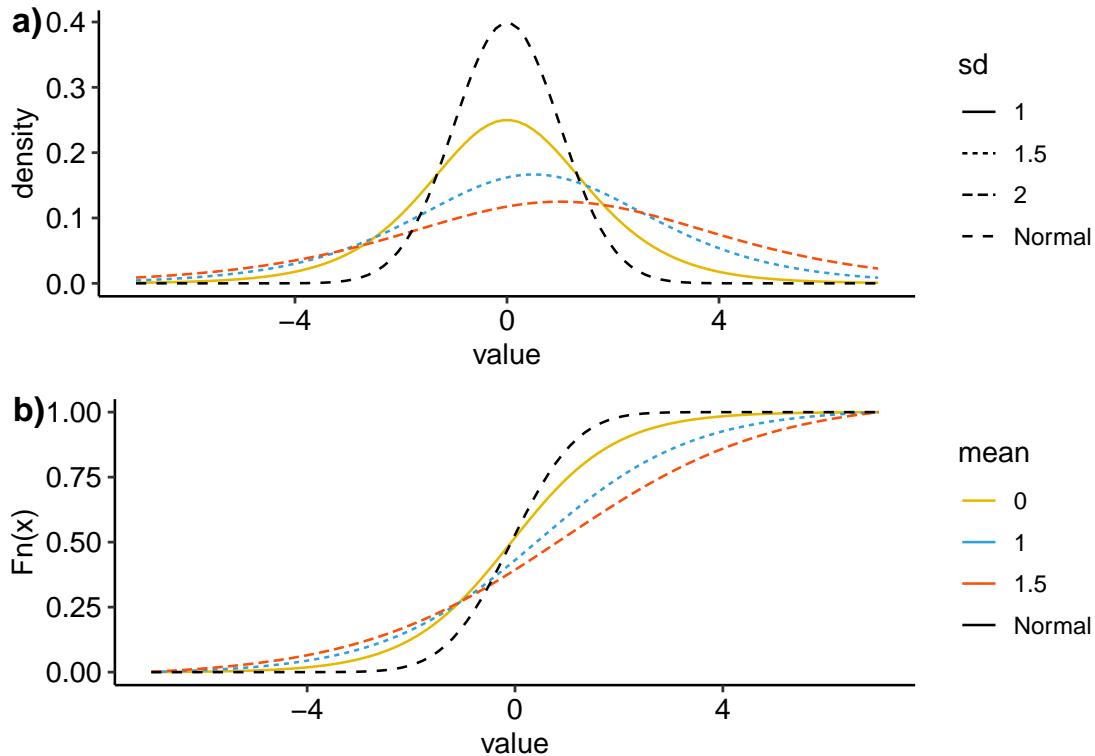


Figure 9.18: Visualisasi distribusi logistik dengan variasi mean dan simpangan baku, a) density plot, b)ecdf

```
dlogis(x, location = 0, scale = 1)
plogis(q, location = 0, scale = 1, lower.tail = TRUE)
```

Note:

- **x,p**: vektor numerik atau kuantil.
- **location**: mean.
- **scale**: simpangan baku
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

9.10.5 Distribusi Weibull

Distribusi lain yang digunakan secara intensif selain distribusi gamma dan eksponensial untuk memperkirakan probabilitas kegagalan suatu proses adalah distribusi Weibull. Fungsi densitas probabilitas distribusi Weibull disajikan pada Persamaan (9.39).

$$f(x; \alpha, \beta) = \begin{cases} f(x; \alpha, \beta) = \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right) & x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (9.39)$$

Pada Gambar 9.19 disajikan visualisasi distribusi gamma dengan variasi α dan β .

Pada R probabilitas distribusi Weibull dapat dihitung menggunakan fungsi `dweibull()` dan `pweibull()` (probabilitas kumulatif). Format fungsi tersebut adalah sebagai berikut:

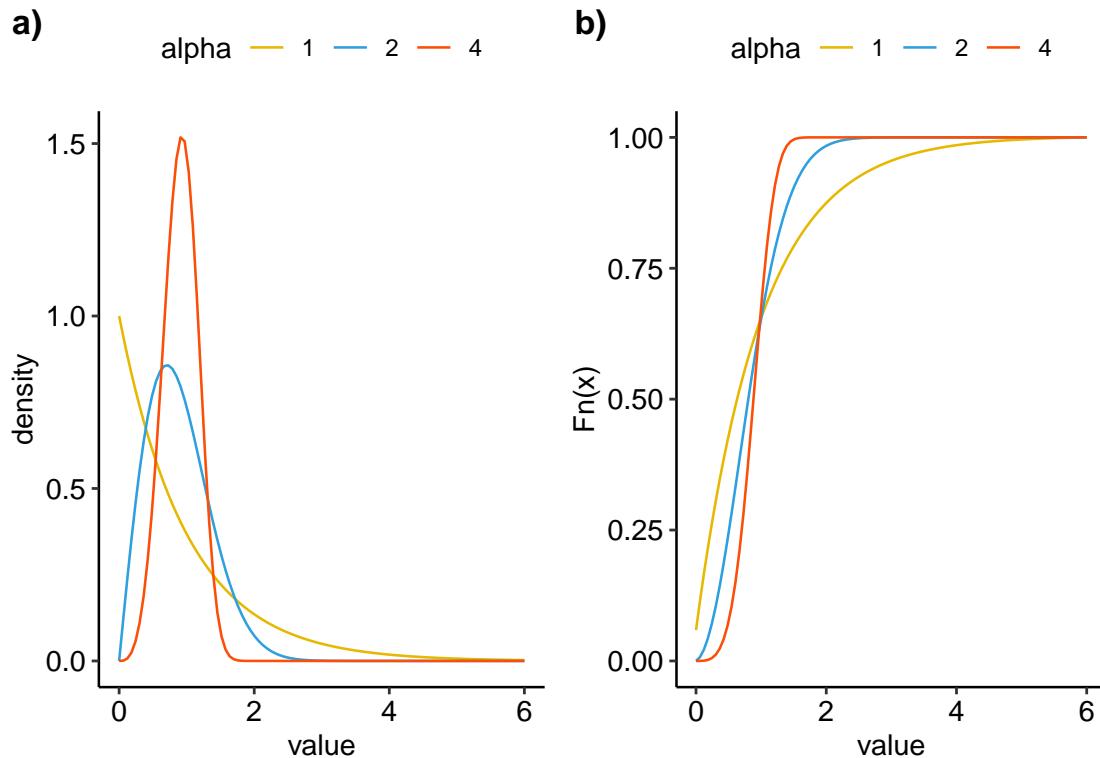


Figure 9.19: Visualisasi distribusi weibull dengan variasi alpha dengan beta 1 a) density plot, b)ecdf

```
dweibull(x, shape, scale)
pweibull(q, shape, scale, lower.tail = TRUE)
```

Note:

- **x,p**: vektor numerik atau kuantil
- **shape**: nilai alpha
- **scale**: nilai beta
- **lower.tail**: probabilitas dihitung dari ujung bawah. Nilai yang mungkin adalah TRUE atau FALSE.

9.11 Referensi

1. Chi Yau. 2014. **R Tutorial with Bayesian Statistics Using OpenBUGS**. Amazon Kindle
2. Damanhuri, E. 2011. **Statitika Lingkunga**. Penerbit ITB.
3. Janicak, C.A. 2007. **Applied Statistics in Occupational Safety and Health**. Government Institutes.
4. Kerns, G.Jay. 2018. **Introduction to Probability and Statistics Using R Third Edition**. GNU Free Documentation License.
5. King, William B. **PROBILITY DISTRIBUTIONS, QUANTILES, CHECKS FOR NORMALITY**. <http://ww2.coastal.edu/kingw/statistics/R-tutorials/prob.html>.
6. Ofungwu, J. 2014. **Statistical Applications For Environmental Analysis and Risk Assessment**. John Wiley & Sons, Inc.
7. Quick-R. **Probability Plots** . <https://www.statmethods.net/advgraphs/probability.html>

8. STAT TREK. **Binomial Probability Distribution.** <https://stattrek.com/probability-distributions/binomial.aspx?tutorial=prob>.
9. _____. **Hypergeometric Distribution.** <https://stattrek.com/probability-distributions/hypergeometric.aspx?tutorial=prob>.
10. _____. **Multinomial Distribution.** <https://stattrek.com/probability-distributions/multinomial.aspx?tutorial=prob>.
11. _____. **Negative Binomial Distribution.** <https://stattrek.com/probability-distributions/negative-binomial.aspx?tutorial=prob>.
12. _____. **Poisson Distribution.** <https://stattrek.com/probability-distributions/poisson.aspx?tutorial=prob>.
13. UBC. **Probability DIistribution.** <https://www.zoology.ubc.ca/~schluter/R/probability/>.
14. Walpole, E. R., Myers, H.M., Myers, S.L., Keying Ye. 2011. **Probability & Statistics for Engineering & Scientists Ninth Edition.** Prentice Hall.

Statistika Inferensi - R

Chapter 10

Penaksiran Secara Statistika

Pada Chapter 6-Ringkasan Numerik kita telah belajar beberapa atribut kunci dari data seperti \bar{X} dan s . Kedua nilai tersebut disebut sebagai nilai estimasi sampel dari populasi (untuk mean μ dan simpangan baku σ). Pada Chapter ini kita akan melakukan eksplorasi lebih jauh lagi mengenai interval estimasi (*interval estimate*) yang akan menyinggung kedua nilai tersebut lebih jauh.

10.1 Definisi Interval Estimasi

Median sampel dan mean sampel menyatakan titik pemasukan data. Estimasi menggunakan kedua nilai tersebut disebut sebagai estimasi titik (*point estimation*). Estimasi titik sendiri tidak menggambarkan reliabilitas atau kurangnya reliabilitas (variabilitas) dari estimasi ini. Sebagai contoh, anggaplah terdapat dua data X dan Y dengan mean 5 dengan jumlah observasi yang sama. Data Y memiliki nilai mean 5 dengan sangat sedikit variasi didalamnya, sedangkan data X jauh lebih bervariasi. Perkiraan titik 5 untuk X jauh lebih tidak dapat diandalkan dibandingkan dengan untuk Y karena variabilitas yang lebih besar dalam data X. Dengan kata lain, lebih banyak kehati-hatian diperlukan ketika menyatakan bahwa 5 memperkirakan mean populasi sebenarnya X daripada ketika menyatakan ini untuk Y. Melaporkan hanya perkiraan mean sampel (poin) 5 gagal memberikan petunjuk tentang perbedaan ini.

Sebagai alternatif untuk estimasi titik, estimasi interval adalah interval yang memiliki probabilitas yang dinyatakan mengandung nilai populasi sebenarnya. interval lebih lebar untuk set data yang memiliki variabilitas lebih besar. Jadi dalam contoh di atas, interval antara 4,7 dan 5,3 mungkin memiliki kemungkinan 95% untuk mengandung mean populasi Y yang sebenarnya (tidak diketahui). Butuh interval yang jauh lebih luas, katakanlah antara 2,0 dan 8,0, untuk memiliki probabilitas yang sama untuk mengandung rerata sebenarnya dari X. Karena itu, perbedaan keandalan dari dua estimasi dengan jelas dinyatakan menggunakan estimasi interval. Estimasi interval dapat memberikan dua informasi yang estimasi poin tidak dapat berikan, antara lain:

1. Pernyataan probabilitas atau kemungkinan bahwa interval berisi nilai populasi sebenarnya (keandalannya).
2. Pernyataan kemungkinan bahwa satu titik data dengan besaran tertentu berasal dari populasi yang diteliti.

Estimasi interval untuk poin pertama disebut sebagai interval kepercayaan (*confidence interval*), sedangkan yang kedua disebut sebagai interval prediksi (*prediction interal*). Meskipun salin terkait, pembaca perlu berhati-hati sebab kedua definisi tersebut sering kali tertukar satu sama lain.

10.2 Interpretasi Interval Estimasi

Untuk lebih memahami mengenai definisi interval estimasi pada sub-chapter ini akan diberikan contoh yang diambil dari buku **statistical Methods in Water Resources** karya Helsel dan Hirsch (2012). Misalkan mean populasi sebenarnya μ konsentrasi dalam akuifer adalah 10. Selain itu, anggaplah bahwa varians populasi sebenarnya σ^2 sama dengan 1. Karena nilai-nilai ini dalam praktiknya tidak pernah diketahui, sampel diambil untuk memperkirakannya dengan mean sampel x dan sampel varian s^2 . Dana yang cukup tersedia untuk mengambil 12 sampel air (kira-kira satu per bulan) selama satu tahun, dan hari-hari di mana pengambilan sampel terjadi dipilih secara acak. Dari 12 sampel ini xx dan s^2 dihitung. Meskipun pada kenyataannya hanya satu set 12 sampel akan diambil setiap tahun, menggunakan komputer 12 hari dapat dipilih beberapa kali untuk menggambarkan konsep perkiraan interval. Untuk masing-masing dari 10 set independen dari 12 sampel, interval kepercayaan pada mean dihitung dengan menggunakan persamaan yang diberikan pada Tabel 10.1 dan Gambar 10.1.

Table 10.1: Sepuluh interval kepercayaan 90% sekitar nilai mean sebenarnya sebesar 10 (Data berdistribusi normal dan Tanda plus menyatakan data tidak disertakan dalam nilai mean sebenarnya)

No.	N	Mean	St.Dev	90% Interval kepercayaan
1	12	10,06	1,11	9,46 sampai 10,64
2	12	10,60	0,81	+ 10,18 sampai 11,02
3	12	9,95	1,26	9,29 sampai 10,60
4	12	10,18	1,26	9,52 sampai 10,83
5	12	10,17	1,33	9,48 sampai 10,85
6	12	10,22	1,19	9,60 sampai 10,84
7	12	9,71	1,51	8,92 sampai 10,49
8	12	9,90	1,01	9,38 sampai 10,43
9	12	9,95	0,10	9,43 sampai 10,46
10	12	9,88	1,37	9,17 sampai 10,59

Kesepuluh interval pada contoh di atas disebut dengan dengan **interval kepercayaan 90%** dari nilai mean sesungguhnya. Nilai mean sebenarnya akan terdapat pada interval tersebut dengan probabilitas 90%. Sehingga berdasarkan Tabel 10.1 terdapat 9 interval yang memiliki nilai mean sesungguhnya didalamnya (probabilitas 90%). Jika kita sekali lagi melakukan sampling dengan jumlah sampling yang sama pada interval nilai baru yang dihasilkan akan mengandung nilai mean sesungguhnya dan dapat juga tidak. Probabilitas interval tersebut mengandung nilai mean sesungguhnya disebut sebagai **tingkat kepercayaan** (*confidence level*). Probabilitas nilai interval tidak mengandung mean sesungguhnya disebut sebagai **alpha level** (α) yang dituliskan berdasarkan Persamaan (10.1).

$$\alpha = 1 - \text{confidence level} \quad (10.1)$$

Lebar interval kepercayaan adalah fungsi dari bentuk distribusi data (variabilitas dan kemencengannya), ukuran sampel, dan tingkat kepercayaan yang diinginkan. Ketika tingkat kepercayaan meningkat, lebar interval juga meningkat, karena interval yang lebih besar lebih mungkin mengandung nilai sebenarnya daripada interval yang lebih pendek. Dengan demikian interval kepercayaan 95% akan lebih luas daripada interval 90% untuk data yang sama.

Interval kepercayaan simetris pada rata-rata biasanya dihitung dengan asumsi data mengikuti distribusi normal. Jika tidak, distribusi rerata itu sendiri akan mendekati normal sepanjang ukuran sampel besar (katakanlah 50 pengamatan atau lebih besar). Interval kepercayaan dengan asumsi normalitas kemudian akan memasukkan mean sebenarnya $(1 - \alpha)\%$ dari waktu. Dalam contoh di atas, data dihasilkan dari distribusi normal, sehingga ukuran sampel kecil 12 tidak menjadi masalah. Namun ketika data memiliki kemencengan dan ukuran sampel di bawah 50 atau lebih, interval kepercayaan simetris tidak akan mengandung

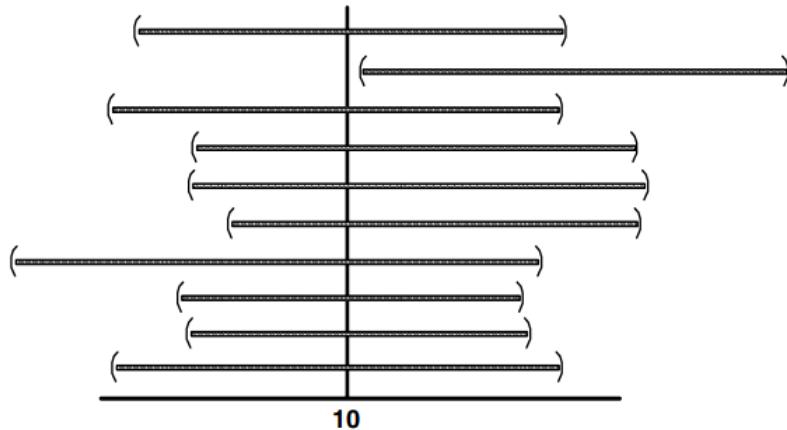


Figure 10.1: Sepuluh interval kepercayaan 90 persen data dengan nilai mean sebenarnya 10 (Helsel dan Hirsch, 2002)

rata-rata $(1 - \alpha)\%$ sepanjang waktu. Dalam contoh di bawah ini, interval kepercayaan simetris secara salah dihitung untuk data yang miring (Gambar 10.2). Hasil (Gambar 10.3 dan Tabel 10.2) menunjukkan bahwa interval kepercayaan kehilangan nilai sebenarnya dari 1 lebih sering daripada yang seharusnya. Semakin besar skewness, semakin besar ukuran sampel harus sebelum interval kepercayaan simetris dapat diandalkan. Sebagai alternatif, interval kepercayaan asimetris dapat dihitung untuk situasi umum data yang memiliki kemencenggan.

Table 10.2: Sepuluh interval kepercayaan 90% sekitar nilai mean sebenarnya sebesar 1 (Data tidak berdistribusi normal dan Tanda plus menyatakan data tidak disertakan dalam nilai mean sebenarnya)

No.	N	Mean	St.Dev	90% Interval kepercayaan
1	12	0,784	0,320	+ 0,618 sampai 0,950
2	12	0,811	0,299	+ 0,656 sampai 0,966
3	12	1,178	0,700	0,815 sampai 1,541
4	12	1,030	0,459	0,792 sampai 1,267
5	12	1,079	0,573	0,782 sampai 1,376
6	12	0,833	0,363	0,644 sampai 1,021
7	12	0,789	0,240	+ 0,644 sampai 0,913
8	12	1,159	0,815	0,736 sampai 1,581
9	12	0,822	0,365	+ 0,633 sampai 0,992
10	12	0,837	0,478	0,589 sampai 1,085

10.3 Interval Kepercayaan Median

Interval kepercayaan median populasi dapat dihitung tanpa perlu mengikuti asumsi distribusi tertentu. Sehingga nilai median dapat digunakan untuk memperkirakan nilai pusat data untuk distribusi data yang tidak berdistribusi normal.

10.3.1 Interval Estimasi Median Metode Nonparametrik

Interval estimasi nonparametrik untuk median populasi sebenarnya dihitung menggunakan distribusi binomial. Pertama, tingkat signifikansi yang diinginkan α dinyatakan, error yang dapat diterima tidak termasuk median yang sebenarnya. Satu-setengah $(\alpha/2)$ dari error ini digunakan untuk setiap akhir interval (Gambar 10.4). Tabel distribusi binomial memberikan nilai kritis bawah dan atas x' dan x pada setengah tingkat alfa yang diinginkan $(\alpha/2)$. Nilai-nilai kritis ini ditransformasikan ke dalam rangking R_l dan R_u yang sesuai dengan titik data C_l dan C_u di ujung interval kepercayaan.

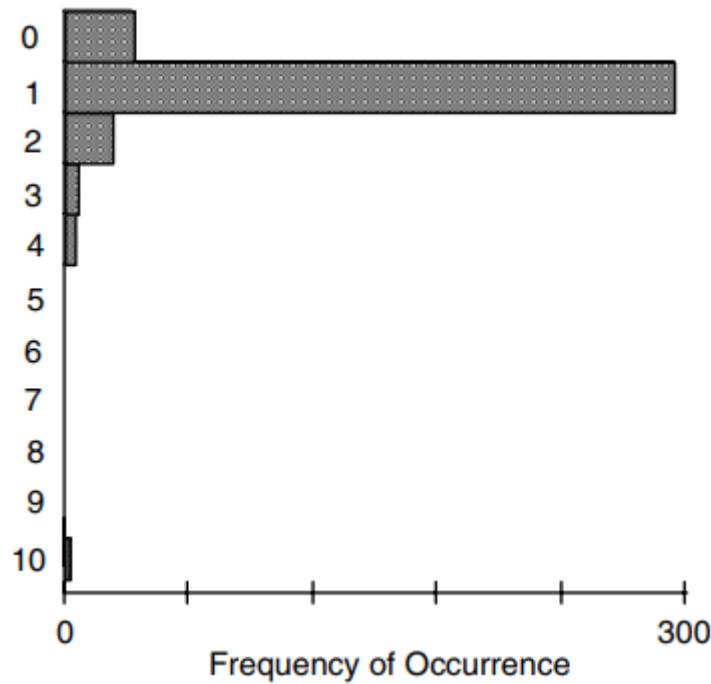


Figure 10.2: Histogram data dengan nilai mean populasi 1 dan simpangan baku populasi 0.75 (Helsel dan Hirsch, 2002)

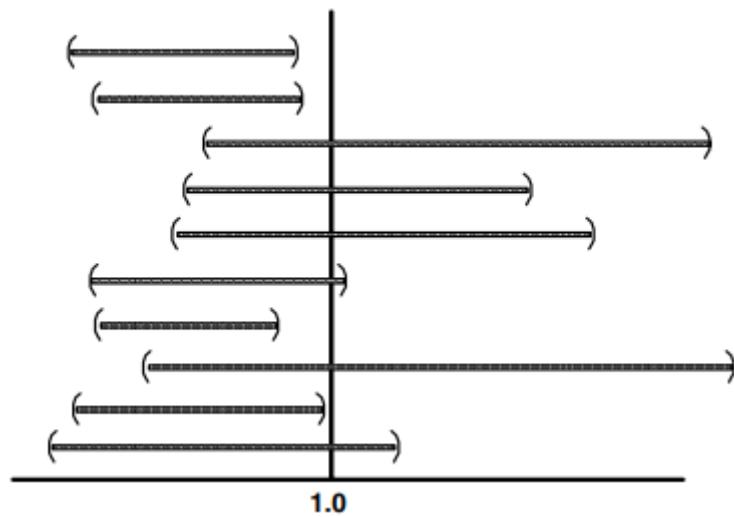


Figure 10.3: Sepuluh interval kepercayaan 90 persen data dengan nilai mean sebenarnya (Helsel dan Hirsch, 2002)

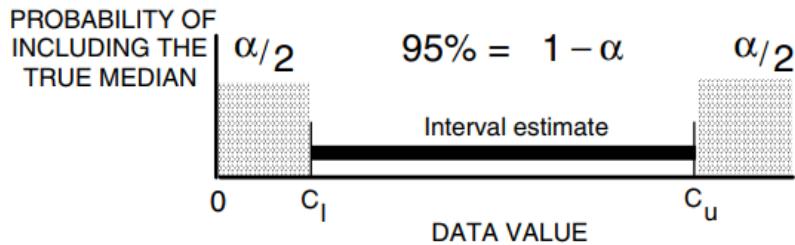


Figure 10.4: Probabilitas median populasi P50 pada dua sisi interval estimasi (Helsel dan Hirsch, 2002)

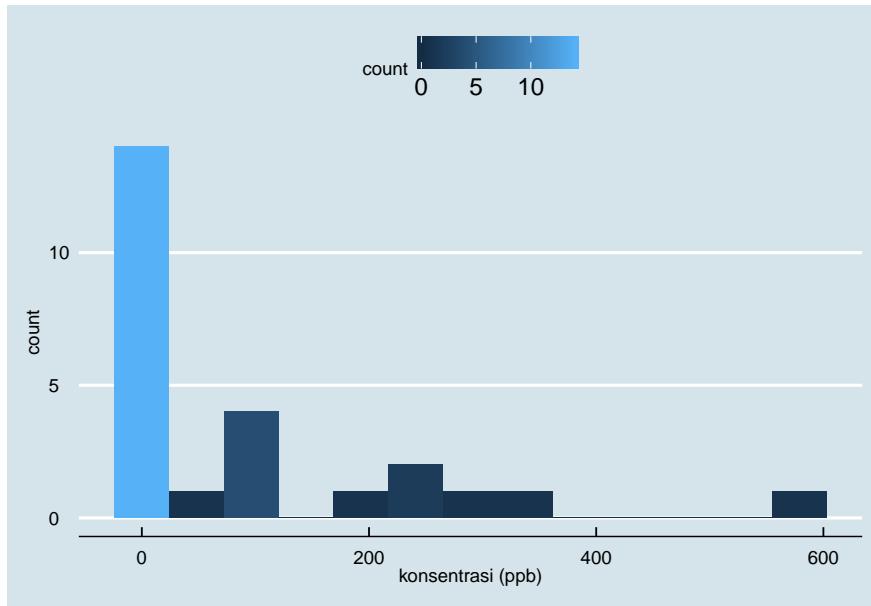


Figure 10.5: Distribusi konsentrasi arsenik dalam air tanah

Interval nonparametrik tidak selalu dapat secara tepat menghasilkan tingkat kepercayaan yang diinginkan ketika ukuran sampel kecil. Ini karena mereka terpisah, melompat dari satu nilai data ke yang berikutnya di ujung interval. Namun, tingkat kepercayaan yang dekat dengan yang diinginkan tersedia untuk semua kecuali ukuran sampel terkecil.

Untuk lebih memahaminya diberikan data 25 pengukuran konsentrasi arsenik di air tanah dalam ppb yang disajikan pada Tabel 10.3.

Visualisasi Tabel 10.3 ditunjukkan pada Gambar 10.5. Berdasarkan gambar tersebut terlihat bahwa data memiliki kemenceng yang positif sehingga penaksiran rata-rata populasi menggunakan nilai mean tidak dapat dilakukan.

Berdasarkan data pada Tabel 10.3, median konsentrasi arsenik $\hat{C}_{0.5}=19$ yang berada pada urutan data ke-13 dari data yang telah diurutkan dari yang terkecil ke yang terbesar. Untuk menentukan interval kepercayaan 95% median konsentrasi arsenik $C_{0.5}$, nilai kritis berdasarkan nilai error mendekati $\alpha/2=0,025$ adalah $x'=7$. Untuk lebih memahaminya pembaca dapat mengunduh tabel distribusi binomial pada laman [berikut](#). Nilai $x'=7$ diperoleh menggunakan Tabel distribusi binomial dengan $n=25$ dan $p=0,5$ yang ditampilkan pada Gambar 10.6 dengan nilai probabilitas sebesar 0,022 (mendekati 0,025) yang setara dengan area yang diarsir pada Gambar 10.4.

Berdasarkan Persamaan (10.2) dan Persamaan (10.3), rangking R_l pada observasi yang menyatakan batas kepercayaan bawah (*lower confidence limit*) adalah 8 ($R_l=7+1$) dan R_u yang menyatakan batas kepercayaan

Table 10.3: Konsentrasi Arsenik dalam air tanah (ppb)

observasi	konsentrasi
1	1.3
2	1.5
3	1.8
4	2.6
5	2.8
6	3.5
7	4.0
8	4.8
9	8.0
10	9.5
11	12.0
12	14.0
13	19.0
14	23.0
15	41.0
16	80.0
17	100.0
18	110.0
19	120.0
20	190.0
21	240.0
22	250.0
23	300.0
24	340.0
25	580.0

atas (*upper confidence level*) adalah $25 - 7 = 18$. Berdasarkan nilai probabilitas $x' = 0,022$, maka nilai alpha yang sesungguhnya sebesar $\alpha = 2 * 0,022 = 0,044$. Nilai tersebut setara dengan tingkat kepercayaan $1 - 0,044$ atau 95,6%. Nilai interval kepercayaan median antara observasi ke-8 dan 18 adalah $C_l = 4,8 \leq C_{0.5} \leq 110 = C_u$ pada $\alpha = 0,044$. Nilai asimetrik disekitar $\hat{C}_{0.5} = 19$ mencerminkan kemencengangan pada data.

Jika pembaca ingin melakukan perhitungan pada R, pembaca harus membuat fungsi sebagai berikut:

```
med_npCI <- function(x,alpha){
  # mengurutkan data
  x_sort=sort(x)
  # menghitung jumlah observasi
  n=length(x)
  # menghitung median data
  med = median(x, na.rm=TRUE)
  # loop untuk mencari nilai probabilitas terdekat
  # dengan alpha
  for(i in 1:n){
    b = pbinom(i,n,0.5)
    if(b>alpha/2){
      break
    }
  }
  # mengambil x'
  x_i=i-1
```

n	x	p									
		0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
25	0	0.277	0.072	0.017	0.004	0.001	0.000	0.000	0.000	0.000	0.000
	1	0.642	0.271	0.093	0.027	0.007	0.002	0.000	0.000	0.000	0.000
	2	0.873	0.537	0.254	0.098	0.032	0.009	0.002	0.000	0.000	0.000
	3	0.966	0.764	0.471	0.234	0.096	0.033	0.010	0.002	0.000	0.000
	4	0.993	0.902	0.682	0.421	0.214	0.090	0.032	0.009	0.002	0.000
	5	0.999	0.967	0.838	0.617	0.378	0.193	0.083	0.029	0.009	0.002
	6	1.000	0.991	0.930	0.780	0.561	0.341	0.173	0.074	0.026	0.007
	7	1.000	0.996	0.975	0.891	0.727	0.512	0.306	0.154	0.069	0.022
	8	1.000	1.000	0.992	0.953	0.851	0.677	0.467	0.274	0.134	0.054
	9	1.000	1.000	0.998	0.983	0.929	0.811	0.630	0.425	0.242	0.115
	10	1.000	1.000	1.000	0.994	0.970	0.902	0.771	0.586	0.384	0.212
	11	1.000	1.000	1.000	0.998	0.989	0.956	0.875	0.732	0.543	0.345
	12	1.000	1.000	1.000	1.000	0.997	0.983	0.940	0.846	0.694	0.500
	13	1.000	1.000	1.000	1.000	0.999	0.994	0.975	0.922	0.817	0.655
	14	1.000	1.000	1.000	1.000	1.000	0.998	0.991	0.966	0.904	0.788
	15	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.997	0.997	0.997

Figure 10.6: Lokasi probabilitas x berdasarkan tabel distribusi binomial

```
# menghitung Rl dan Ru
rl=x_i+1
ru=n-x_i
# menghitung true confidence level
CL=1-2*(pbinom(x_i,n,0.5))
# menghitung Lower dan Upper CL
LCL=x_sort[rl]
UCL=x_sort[ru]
# menggabungkannya dalam satu data
data=data.frame("median"=med,
                 "True CL %"=CL*100,
                 "Lower CL "=LCL,
                 "Upper CL "=UCL)
return(data)
}
```

Note:

- **x:** vektor numerik
- **alpha:** alpha level yang digunakan

Fungsi yang telah dibuat tersebut selanjutnya dapat pembaca gunakan saat akan menghitung interval kepercayaan median populasi menggunakan metode Nonparametrik. Berikut penerapannya menggunakan data pada Tabel 10.3 yang telah disimpan kedalam objek `gwardat`.

```
med_npCI(x=gwardat$konsentrasi, alpha=0.05)
```

```
##   median True.CL.. Lower.CL Upper.CL
## 1      19     95.67      4.8     110
```

Alternatif lain yang dapat digunakan untuk menghitung interval kepercayaan jika sampel cukup besar $n > 20$ menggunakan metode Nonparametrik adalah dengan menggunakan pendekatan tabel distribusi normal untuk memperkirakan distribusi binomial. Dengan menggunakan pendekatan ini, hanya sebagian kecil tabel distribusi binomial ($n=20$) yang diperlukan untuk melakukannya. Nilai kritis $z_{\alpha/2}$ dari tabel distribusi normal menentukan rangking atas dan bawah observasi yang menyatakan awal dan akhir nilai interval kepercayaan yang dinyatakan pada Persamaan (10.4) dan Persamaan (10.5). Pembulatan diperlukan dalam proses ini sebab nilai ranking harus berupa integer.

$$R_l = \frac{n - z_{\frac{\alpha}{2}} \cdot \sqrt{n}}{2} \quad (10.4)$$

$$R_l = \frac{n - z_{\frac{\alpha}{2}} \cdot \sqrt{n}}{2} + 1 \quad (10.5)$$

Menggunakan contoh data pada Tabel 10.3, dengan 95% interval kepercayaan ($z_{\alpha/2}=1,96$) median dapat dihitung seperti berikut:

$$R_l = \frac{25 - 1,96 \cdot \sqrt{25}}{2} = 7,6$$

$$R_l = \frac{25 + 1,96 \cdot \sqrt{25}}{2} + 1 = 18,4$$

Berdasarkan hasil perhitungan diperoleh rangking bawah adalah data ke-8 dan rangking atas adalah 18. Kedua data dibulatkan berdasarkan integer terdekat. Nilai interval kepercayaan median yang diperoleh sama dengan metode sebelumnya sebab rangking batas bawah dan atasnya yang seragam.

Jika pembaca ingin menggunakan R, maka fungsi yang sebelumnya telah kita buat dapat dimodifikasi seperti berikut:

```
med_norCI <- function(x, alpha){
  # mengurutkan data dari yang terkecil
  x_sort=sort(x)
  # menghitung jumlah observasi
  n = length(x)
  # menghitung median data
  med = median(x, na.rm=TRUE)
  # menghitung Rl dan Ru
  rl=round((n-abs(qnorm(alpha/2))*sqrt(n))/2,0)
  ru=round(((n+abs(qnorm(alpha/2))*sqrt(n))/2)+1,0)
  # menghitung Lower dan Upper CL
  LCL=x_sort[rl]
  UCL=x_sort[ru]
  # menggabungkannya dalam satu data
  data=data.frame("median"=med,
                  "Lower CL"=LCL,
                  "Upper CL"=UCL)
  return(data)
}
```

Fungsi `med_norCI()` sama dengan fungsi `med_npCI()`. Perbedaannya terletak pada penggunaan distribusi normal pada proses penentuan titik kritisnya.

Dengan menggunakan fungsi `med_norCI()`, rentang kepercayaan median dapat dihitung seperti berikut:

Note:

- **x:** vektor numerik
- **alpha:** alpha level yang digunakan

```
med_norCI(x=gwardat$konsentrasi, alpha=0.05)
```

```
##   median Lower.CL Upper.CL
## 1      19     4.8     110
```

Jika kita tidak ingin menggunakan vektor dalam fungsi, kita dapat juga menggunakan data frame sebagai inputnya. Kelebihannya adalah kita dapat menghitung rentang kepercayaan seluruh kolom dalam satu kali proses. Hal ini tentunya akan menghemat waktu yang digunakan. Berikut adalah contoh sintaks fungsi untuk menghitung interval kepercayaan median menggunakan distribusi normal dengan metode nonparametrik yang digunakan:

```
med_norCI <- function(df, alpha){
  # membuat matrik untuk menyimpan
  # hasil loop
  med = rep(NA, ncol(df))
  LCL = rep(NA, ncol(df))
  UCL = rep(NA, ncol(df))
  var = rep(NA, ncol(df))
  # looping
  for(i in 1:ncol(df)){
    # mengurutkan data
    x_sort = sort(df[, i])
    # mengambil nama kolom dataset
    var[i] = colnames(df[i])
    # menghitung jumlah observasi
    n = length(x_sort)
    # menghitung median data
    med[i] = median(x_sort, na.rm=TRUE)
    # menghitung Lower dan Upper CL
    LCL[i]=x_sort[(round((n-abs(qnorm(alpha/2))*sqrt(n))/2,0))]
    UCL[i]=x_sort[(round(((n+abs(qnorm(alpha/2))*sqrt(n))/2)+1,0))]
  }
  # menggabungkannya dalam satu data
  data=data.frame("variabel"=var,
                  "median"=med,
                  "Lower CL"=LCL,
                  "Upper CL"=UCL)
  return(data)
}
```

Note:

- **df**: data frame
- **alpha**: alpha level yang digunakan

Fungsi tersebut dapat menghitung sekaligus Interval kepercayaan median dengan metode nonparametrik. Pembaca dapat mencobanya dengan menggunakan dataset yang pembaca miliki. Pembaca dapat mengabaikan peringatan yang muncul dan berfokus pada hasil yang diperoleh. Sebagai contoh jalankan fungsi tersebut menggunakan dataset **airquality** berikut:

```
med_norCI(df=airquality, alpha=0.05)
```

```
##   variabel median Lower.CL Upper.CL
## 1 Ozone     31.5    23.0    39.0
## 2 Solar.R   205.0   187.0   225.0
## 3 Wind       9.7     9.2    10.3
## 4 Temp      79.0    77.0    81.0
## 5 Month      7.0     7.0     7.0
## 6 Day        16.0   13.0    18.0
```

10.3.2 Metode Parametrik Interval Estimasi Median

Telah dijelaskan pada chapter 6 bahwa rata-rata geometrik merupakan merupakan nilai rata-rata yang digunakan untuk mengestimasi median sampel untuk data yang memiliki kemencengan dengan transformasi yang digunakan agar data simetris adalah transformasi logaritmik $y = \ln(x)$. Pada metode ini data diasumsikan memiliki distribusi lognormal (kemencengan positif). Rerata dan interval geometris akan lebih efisien (interval lebih pendek) dari median dan interval kepercayaannya ketika data benar-benar lognormal. Median sampel dan intervalnya lebih tepat dan lebih efisien jika logaritma data masih menunjukkan kemencengan dan/atau *outlier*. Estimasi media menggunakan metode parametrik dituliskan kedalam Persamaan (10.6) dan Persamaan (10.7).

$$GM_x = \exp(\bar{y}) \quad (10.6)$$

dimana $y = \ln(x)$ dan \bar{y} =mean sampel y .

$$\exp\left(\bar{y} - t_{(\frac{\alpha}{2}, n-1)} \sqrt{\frac{s_y^2}{n}}\right) \leq GM_x \leq \exp\left(\bar{y} + t_{(\frac{\alpha}{2}, n-1)} \sqrt{\frac{s_y^2}{n}}\right) \quad (10.7)$$

dimana s_y^2 = varians sampel y pada unit log natural.

Pada Tabel 10.3, untuk menghitung interval keyakinan median menggunakan pendekatan mean geometrik GM_x kita perlu mentransformasi datanya terlebih dahulu sehingga menjadi bentuk natural log. hasil transformasi disajikan pada Tabel 10.4.

visualisasi distribusi yang baru disajikan pada Gambar 10.7.

Nilai mean dari data tersebut adalah 3,17 dengan simpangan baku sebesar 1,96. Berdasarkan Gambar 10.7, kita telah memperoleh distribusi yang simetris.

Dengan menggunakan Persamaan (10.6) dan Persamaan (10.7) selanjutnya dapat dihitung interval kepercayaannya dengan derajat kepercayaan 95%.

$$GM_x = \exp(3,17) = 23,8$$

$$\exp\left(3,17 - 2,064 \cdot \sqrt{\frac{1,96^2}{25}}\right) \leq GM_x \leq \exp\left(3,17 + 2,064 \cdot \sqrt{\frac{1,96^2}{25}}\right)$$

$$\exp(2,36) \leq GM_x \leq \exp(3,98)$$

$$10,6 \leq GM_x \leq 53,5$$

Dengan menggunakan R dapat dikerjakan menggunakan fungsi sebagai berikut:

Table 10.4: Transformasi logaritmik konsentrasi Arsenik dalam air tanah (ppb)

observasi	konsentrasi
1	0.2624
2	0.4055
3	0.5878
4	0.9555
5	1.0296
6	1.2528
7	1.3863
8	1.5686
9	2.0794
10	2.2513
11	2.4849
12	2.6391
13	2.9444
14	3.1355
15	3.7136
16	4.3820
17	4.6052
18	4.7005
19	4.7875
20	5.2470
21	5.4806
22	5.5215
23	5.7038
24	5.8289
25	6.3630

```

med_gm <- function(x, alpha){
  x = log(x)
  # rata-rata geometrik
  gm = exp(mean(x, na.rm=TRUE))
  # menghitung derajat kebebasan
  df = length(x)-1
  # menghitung batas bawah dan atas
  LCL = exp(mean(x, na.rm=TRUE)-qt(1-alpha/2,df)*sqrt(var(x, na.rm=TRUE)/length(x)))
  UCL = exp(mean(x, na.rm=TRUE)+qt(1-alpha/2,df)*sqrt(var(x, na.rm=TRUE)/length(x)))
  # menggabungkan hasil
  data=data.frame("GM"=gm,
                  "Lower CL"=LCL,
                  "Upper CL"=UCL)
  return(data)
}
  
```

Note:

- **x**: vektor numerik
- **alpha**: alpha level yang digunakan

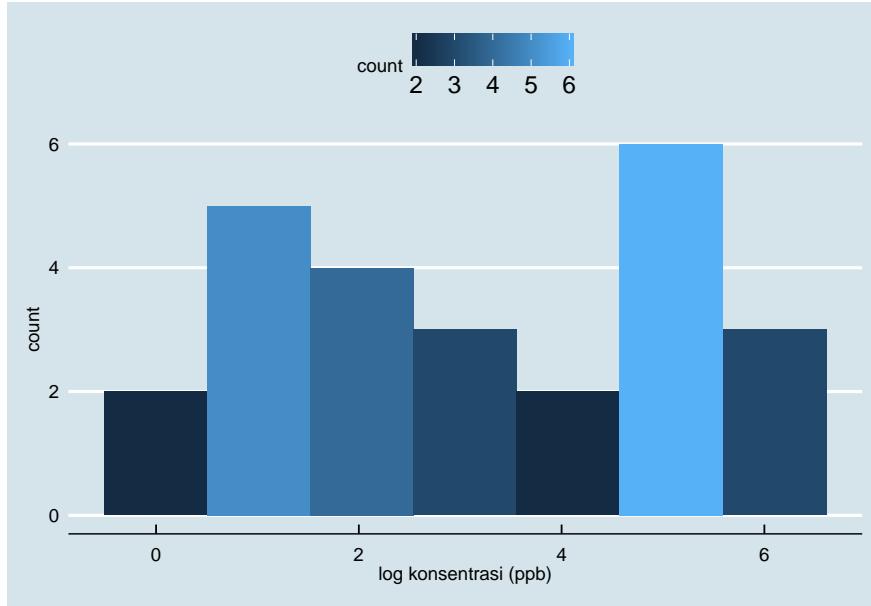


Figure 10.7: Distribusi logaritmik konsentrasi arsenik dalam air tanah

```
med_gm(x=gwardat$konsentrasi, alpha=0.05)
```

```
##      GM Lower.CL Upper.CL
## 1 23.87     10.63    53.6
```

Fungsi `med_gm()` dapat dilakukan sejumlah modifikasi seperti penggunaan data frame sebagai input serta proses transformasi yang dilakukan didalam fungsi yang ada sekaligus. Berikut adalah contoh fungsi yang digunakan untuk input berupa data frame dan proses transformasi termasuk didalamnya:

```
med_gm <- function(df, alpha){
  # membuat vektor untuk menyimpan hasil loop
  var = rep(NA, ncol(df))
  gm = rep(NA, ncol(df))
  LCL = rep(NA, ncol(df))
  UCL = rep(NA, ncol(df))
  # looping
  for(i in 1:ncol(df)){
    # mengambil nama kolom
    var[i] = colnames(df[i])
    # transformasi variabel (logaritmik)
    x = log(df[,i])
    # rata-rata geometrik
    gm[i] = exp(mean(x, na.rm=TRUE))
    # menghitung derajat kebebasan
    d = length(x)-1
    # menghitung batas bawah dan atas
    LCL[i] = exp(mean(x, na.rm=TRUE)-qt(1-alpha/2,d)*sqrt(var(x, na.rm=TRUE)/length(x)))
    UCL[i] = exp(mean(x, na.rm=TRUE)+qt(1-alpha/2,d)*sqrt(var(x,na.rm=TRUE)/length(x)))
  }
  # menggabungkan hasil
}
```

```

data=data.frame("Variabel"=var,
                "GM"=gm,
                "Lower CL"=LCL,
                "Upper CL"=UCL)
return(data)
}

```

Note:

- **df:** data frame
- **alpha:** alpha level yang digunakan

Untuk menguji fungsi tersebut jalankan fungsi tersebut menggunakan dataset yang pembaca miliki. Dalam contoh ini ak diberikan contoh penerapannya menggunakan dataset `airquality`. Jalankan fungsi berikut untuk memperoleh hasilnya.

```
med_gm(airquality, 0.05)
```

```

##   Variabel      GM Lower.CL Upper.CL
## 1   Ozone  30.524  26.583  35.049
## 2 Solar.R 149.561 131.409 170.220
## 3   Wind   9.273   8.697   9.888
## 4   Temp   77.284  75.740  78.859
## 5 Month    6.847   6.623   7.079
## 6   Day   12.270  10.728  14.034

```

Pembaca perlu berhati-hati dalam menentukan apakah akan menggunakan metode Nonparametrik atau parametrik. Jika data berdistribusi lognormal kita dapat menggunakan metode parametrik.

10.4 Interval Kepercayaan Mean

Estimasi interval juga dapat dihitung untuk mean populasi sebenarnya μ . Hal ini sangat sesuai jika pusat data menjadi fokus dalam analisa statistik. Interval simetris di sekitar sampel rata-rata X paling sering dihitung. Untuk ukuran sampel besar, interval simetris secara memadai menggambarkan variasi rata-rata, terlepas dari bentuk distribusi data. Ini karena distribusi rata-rata sampel akan mendekati dengan distribusi normal ketika ukuran sampel semakin besar, meskipun data mungkin tidak terdistribusi secara normal. Untuk ukuran sampel yang lebih kecil, rata-rata tidak akan didistribusikan secara normal kecuali jika data itu sendiri terdistribusi secara normal. Ketika data meningkat kemencengannya, lebih banyak data diperlukan sebelum distribusi rata-rata dapat didekati secara memadai oleh distribusi normal. Untuk distribusi yang sangat miring atau data yang mengandung *outlier*, mungkin diperlukan lebih dari 100 pengamatan sebelum nilai rata-rata tidak akan terpengaruh oleh nilai terbesar untuk mengasumsikan bahwa distribusinya akan simetris.

10.4.1 Interval Kepercayaan Mean Untuk Distribusi Yang Simetris

Interval kepercayaan mean untuk distribusi simetris dihitung menggunakan tabel distribusi *student's t* yang tersedia dalam buku teks statistik dan perangkat lunak. Tabel ini dimasukkan untuk menemukan nilai kritis untuk t pada setengah tingkat alfa yang diinginkan. Pada buku lain sering dijelaskan bahwa distribusi t hanya digunakan untuk sampel kecil (beberapa menyebutkan $n < 30$), sedangkan untuk distribusi besar digunakan distribusi normal. Penggunaan distribusi normal jarang digunakan dalam prakiraan. Hal ini

disebabkan karena pada proses perhitungan diperlukan nilai simpangan baku σ . Pada kenyataannya pada pengukuran dilapangan kita sering sekalin melakukan estimasi terhadap simpangan baku melalui sampel s karena kita tidak mengetahui nilai simpangan baku populasinya sehingga pada buku ini akan digali lebih jauh metode estimasi interval menggunakan persamaan distribusi t.

Lebar interval kepercayaan adalah fungsi dari nilai-nilai kritis dari tabel distribusi t, simpangan baku data, dan ukuran sampel. Ketika data memiliki kemecengan atau mengandung *outlier*, asumsi di balik interval t dan distribusi normal tidak berlaku. Interval simetris yang dihasilkan akan sangat luas sehingga sebagian besar pengamatan akan dimasukkan di dalamnya. Ini juga dapat mencapai di bawah nol di ujung bawah. Titik akhir negatif dari interval kepercayaan untuk data yang tidak dapat menjadi negatif adalah sinyal yang jelas bahwa asumsi interval kepercayaan simetris tidak diperlukan. Untuk data tersebut, dengan asumsi distribusi lognormal seperti yang dijelaskan dalam sub chapter sebelumnya (interval kepercayaan median) akan lebih tepat. Interval kepercayaan dihitung menggunakan Persamaan (10.8).

$$\bar{x} - t_{(\frac{\alpha}{2}, n-1)} \cdot \sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{x} + t_{(\frac{\alpha}{2}, n-1)} \cdot \sqrt{\frac{s^2}{n}} \quad (10.8)$$

Untuk lebih memahami cara penerapannya, kita akan menggunakan kembali data pada Tabel 10.4. Langkah pertama yang perlu dilakukan adalah menghitung mean sampel \bar{x} dan simpangan baku sampel s . Berdasarkan hasil perhitungan diperoleh nilai $\bar{x} = 98.352$ dan $s = 144.685$. Dengan menggunakan Persamaan (10.8), interval estimasi mean dengan tingkat kepercayaan 95% dapat dihitung sebagai berikut:

$$3,17 - t_{(0.25, 24)} \cdot \sqrt{\frac{1,96^2}{25}} \leq \mu \leq 3,17 + t_{(0.25, 24)} \cdot \sqrt{\frac{1,96^2}{25}}$$

$$2,36 \leq \mu \leq 3,98$$

Berdasarkan hasil yang diperoleh terdapat 95% peluang nilai mean populasi μ terletak pada interval 2,36 sampai 3,98. Perlu diingat bahwa metode parametrik sangat sensitif dengan adanya *outlier* sehingga jika pembaca ingin menggunakan pastikan terlebih dahulu tidak ada *outlier* pada data dengan cara melakukan visualisasi data.

Pada R kita dapat menggunakan fungsi `stat.desc()` untuk menhitung statistika deskriptif serta interval kepercayaan mean-nya. Berikut adalah sintaks yang digunakan:

```
# memuat paket
library(pastecs)

# ringkasan data
r=stat.desc(gwardat2$konsentrasi)
r

##      nbr.val      nbr.null      nbr.na          min
##      25.0000      0.0000      0.0000      0.2624
##      max      range          sum          median
##      6.3630      6.1007     79.3167      2.9444
##      mean      SE.mean CI.mean.0.95          var
##      3.1727      0.3919     0.8089      3.8400
##      std.dev      coef.var
##      1.9596      0.6176

# batas bawah (LCL)
mean(gwardat2$konsentrasi)-r[[11]]
```

```
## [1] 2.364

# batas atas (UCL)
mean(gwardat2$konsentrasi)+r[[11]]

## [1] 3.982
```

Selain itu , kita juga dapat menghitung interval kepercayaan mean menggunakan fungsi `t.test()`. Fungsi ini pada dasarnya dilakukan untuk melakukan uji hipotesis terhadap satu rata-rata. Untuk lebih tahu mengenai fungsi tersebut jalankan sintaks bantuan berikut:

```
?t.test
```

Untuk menghitung interval kepercayaan mean jalankan sintaks berikut:

```
t.test(gwardat$konsentrasi, conf.level= 0.95)

##
##  One Sample t-test
##
## data: gwardat$konsentrasi
## t = 3.4, df = 24, p-value = 0.002
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   38.63 158.08
## sample estimates:
## mean of x
##      98.35
```

10.4.2 Interval Kepercayaan Mean Untuk Distribusi Yang Asimetris

Mean dan interval kepercayaan dapat dihitung dengan mengasumsikan distribusi data mengikuti distribusi logaritmik $y = \ln(x)$. Metode ini berguna untuk jenis data yang memiliki bentuk distribusi data yang memiliki kemencenggan positif (perlu transformasi logaritmik agar simetris).Metode ini memberikan perkiraan rata-rata yang lebih dapat diandalkan (varians lebih rendah) daripada perhitungan rata-rata sampel biasa tanpa transformasi log.

Untuk memperkirakan rata-rata populasi μ_x dalam unit aslinya, anggap datanya berdistribusi normal. Satu-setengah varians logaritma ditambahkan ke \bar{y} (rata-rata log) sebelum eksponensial. Karena varians sampel s_y^2 hanya perkiraan varians sebenarnya dari logaritma, estimasi sampel rata-rata akan menjadi bias. Namun, untuk sampel dengan s_y^2 kecil dan ukuran sampel besar bias dapat diabaikan. Interval kepercayaan dapat dituliskan berdasarkan Persamaan (10.9).

$$\mu_x = \exp(\bar{y} + 0,5 \cdot s_y^2) \quad (10.9)$$

dimana $y = \ln(x)$, \bar{y} = mean sampel dan s_y^2 = varians sampel y dalam unit log natural.

Interval kepercayaan sekitar μ_x bukan estimasi interval yang dihitung untuk rata-rata geometri dalam Persamaan (10.7). Interval kepercayaan tidak dapat dihitung hanya dengan mengekspansi interval sekitar \bar{y} . Interval kepercayaan yang tepat dalam satuan asli untuk rata-rata data lognormal dapat dihitung. Untuk lebih jelasnya pembaca dapat melihatnya pada situs <http://jse.amstat.org/v13n1/olsson.html>.

Metode Cox dapat digunakan untuk menghitung interval keyakinan dengan nilai estimasi rata-rata menggunakan Persamaan (10.9). Persamaan yang digunakan dapat dituliskan sebagai berikut (Persamaan (10.10)).

$$\ln(\mu_x) = \bar{Y} + \frac{s_y^2}{2} \pm z_{\left(\frac{\alpha}{2}\right)} \sqrt{\frac{s_y^2}{n} + \frac{s_y^4}{2(n-1)}} \quad (10.10)$$

Persamaan (10.10) dapat dimodifikasi dengan menggunakan distribusi t dibanding menggunakan distribusi normal. Penggunaan distribusi t akan memperbaiki kelemahan penggunaan distribusi normal pada sampel yang berukuran kecil.

Data Tabel 10.3 dapat kita gunakan untuk menghitung rata-rata menggunakan Persamaan (10.10). Hal ini disebabkan karena data yang ada memiliki kemencengan positif sehingga dapat dianggap bahwa transformasi logaritmik dapat membentuk distribusi ini menjadi lebih simetris.

Berdasarkan hasil perhitungan diperoleh nilai $\bar{Y} = 3.173$ dan $s_y^2 = 1.96$. Sehingga nilai interval selanjutnya dapat dihitung menggunakan Persamaan (10.10) dengan interval keyakinan 95%.

$$\ln(\mu_x) = 3,17 + \frac{1,96^2}{2} \pm 1,96 \sqrt{\frac{1,96^2}{25} + \frac{1,96^4}{2(25-1)}}$$

$$\ln(\mu_x) = 5,10 \pm 1,33$$

Sehingga

$$\exp(5,10 - 1,33) \leq \mu_x \leq \exp(5,10 + 1,33)$$

$$43,38 \leq \mu_x \leq 620,17$$

Nilai interval yang dihasilkan sangat panjang sehingga nilai rata-rata yang dihasilkan tidak dapat diandalkan untuk memperkirakan lokasi nilai mean populasi.

Pada contoh berikut akan disajikan sintaks untuk menghitung interval kepercayaan mean data pada Tabel 10.3 berdasarkan Persamaan (10.10) dan sitribusi yang digunakan adalah distribusi t. Pembaca dapat memodifikasi sintaks berikut jika ingin menggunakan distribusi normal.

```
mean_asci<-function(x,alpha){
  m=mean(x, na.rm=TRUE)
  # mean data hasil transformasi logaritmik
  ave = mean(log(x), na.rm=TRUE)
  # simpangan baku data hasil transformasi
  sd = sd(log(x), na.rm=TRUE)
  # jumlah observasi
  n = length(x)
  # derajat kebebasan
  df = n-1
  # interval keyakinan satu sisi
  re = 1-(alpha/2)
  # CI menggunakan distribusi t
  LCL = exp(ave+(0.5*sd^2)-qt(re,df)*sqrt(((sd^2)/n)+((sd^4)/(2*df))))
  UCL = exp(ave+(0.5*sd^2)+qt(re,df)*sqrt(((sd^2)/n)+((sd^4)/(2*df))))
  # menggabungkan hasil
  data = data.frame("Mean"=m,
```

```

    "Lower CL"=LCL,
    "Upper CL"=UCL)
return(data)
}

```

Note:

- **x**: vektor numerik
- **alpha**: alpha level yang digunakan

```
mean_asci(x=gwardat$konsentrasi, alpha=0.05)
```

```
##      Mean Lower.CL Upper.CL
## 1 98.35     40.11   660.9
```

Jika pembaca ingin menggunakan data frame sebagai input yang digunakan selain vektor, fungsi tersebut dapat dimodifikasi seperti berikut:

```

mean_asci<-function(df,alpha){
  # membuat vektor untuk menyimpan hasil loop
  var = rep(NA, ncol(df))
  m = rep(NA, ncol(df))
  LCL = rep(NA, ncol(df))
  UCL = rep(NA, ncol(df))
  # looping
  for(i in 1:ncol(df)){
    # mengambil nama kolom
    var[i] = colnames(df[i])
    # menghitung mean data
    m[i]=mean(df[,i], na.rm=TRUE)
    # mean data hasil transformasi logaritmik
    ave = mean(log(df[,i]), na.rm=TRUE)
    # simpangan baku data hasil transformasi
    sd = sd(log(df[,i]), na.rm=TRUE)
    # jumlah observasi
    n = length(df[,i])
    # derajat kebebasan
    d = n-1
    # interval keyakinan satu sisi
    re = 1-(alpha/2)
    # CI menggunakan distribusi t
    LCL[i] = exp(ave+(0.5*sd^2)-qt(re,d)*sqrt(((sd^2)/n)+((sd^4)/(2*d))))
    UCL[i] = exp(ave+(0.5*sd^2)+qt(re,d)*sqrt(((sd^2)/n)+((sd^4)/(2*d))))
  }
  # menggabungkan hasil
  data = data.frame("Variabel"=var,
                    "Mean"=m,
                    "Lower CL"=LCL,
                    "Upper CL"=UCL)
  return(data)
}

```

Note:

- **df:** data frame
- **alpha:** alpha level yang digunakan

Untuk menguji fungsi tersebut, pembaca dapat memasukkan data frame yang pembaca miliki kedalam persamaan tersebut. Berikut adalah contoh sintaks yang digunakan untuk menghitung interval kepercayaan mean pada dataset `airquality`. Pembaca dapat menjalankannya pada komputer pembaca.

```
mean_asci(airquality, 0.05)
```

```
##   Variabel      Mean Lower.CL Upper.CL
## 1 Ozone    42.129   37.744   52.209
## 2 Solar.R 185.932  178.856  241.060
## 3 Wind     9.958    9.404   10.747
## 4 Temp     77.882   76.341   79.498
## 5 Month    6.993    6.766    7.237
## 6 Day      15.804   14.945   20.433
```

10.5 Interval Prediksi Nonparametrik

Pertanyaan yang sering diajukan adalah apakah satu pengamatan baru kemungkinan berasal dari distribusi yang sama dengan data yang dikumpulkan sebelumnya, atau sebagai alternatif dari distribusi yang berbeda. Pertanyaan dapat dievaluasi dengan menentukan apakah pengamatan baru di luar interval prediksi yang dihitung dari data yang ada. Interval prediksi mengandung $100 \cdot (1 - \alpha)$ persen dari distribusi data, sementara $100 \cdot \alpha$ persen berada di luar interval. Jika pengamatan baru datang dari distribusi yang sama dengan data yang diukur sebelumnya, ada kemungkinan $100 \cdot \alpha$ persen bahwa pengamatan baru tersebut akan berada di luar interval prediksi. Karena pengamatan baru tersebut berada di luar interval tidak “membuktikan” pengamatan baru itu berbeda, hanya saja sepertinya begitu. Seberapa besar kemungkinan ini tergantung pada pilihan α yang ditentukan oleh peneliti.

Interval prediksi dihitung dengan tujuan yang berbeda dari interval kepercayaan. Interval prediksi terkait dengan nilai data individu yang berlawanan dengan ringkasan statistik seperti nilai mean. Interval prediksi lebih luas daripada interval kepercayaan, karena pengamatan individu lebih bervariasi daripada ringkasan statistik yang dihitung dari beberapa pengamatan. Tidak seperti interval kepercayaan, interval prediksi memperhitungkan variabilitas titik data tunggal di sekitar median atau rata-rata, di samping kesalahan dalam memperkirakan pusat distribusi. Ketika $\text{mean} \pm 2$ simpangan baku secara keliru digunakan untuk memperkirakan lebar interval prediksi, data baru dinyatakan berasal dari populasi yang berbeda lebih sering daripada yang seharusnya.

10.5.1 Interval Prediksi Nonparametrik Dua Sisi

Interval prediksi tingkat kepercayaan nonparametrik α secara sederhana dinyatakan sebagai interval antara persentil distribusi $\alpha/2$ dan $1 - (\frac{\alpha}{2})$ (Gambar 10.8). Interval ini mengandung $100 \cdot (1 - \alpha)$ data, sedangkan $100 \cdot \alpha$ persen berada di luar interval. Oleh karena itu jika titik data tambahan baru berasal dari distribusi yang sama dengan data yang diukur sebelumnya, ada kemungkinan $100 \cdot \alpha$ persen bahwa itu akan berada di luar interval prediksi. Interval akan mencerminkan bentuk data yang dikembangkannya, dan tidak ada asumsi tentang bentuk bentuk yang perlu dibuat. Interval prediksi nonparametrik dua sisi dinyatakan berdasarkan Persamaan (10.11).

$$PI_{np} = X_{\frac{\alpha}{2} \cdot (n+1)} \text{ sampai dengan } X_{[1 - (\frac{\alpha}{2})] \cdot (n+1)} \quad (10.11)$$



Figure 10.8: Prediksi interval dua sisi (Helsel dan Hirsch, 2002)

Kita akan kembali menggunakan data pada Tabel 10.3. Dengan menggunakan tingkat kepercayaan 90% kita diminta untuk menentukan interval prediksi dari konsentrasi arsenik pada data tersebut tanpa mengasumsikan distribusi dari data.

Untuk melakukannya kita perlu menentukan observasi ke-2,5 dan 97,5 (berdasarkan nilai $\alpha/2$) dengan rangking observasi berdasarkan Persamaan (10.11) adalah $(0,05 * 26)$ atau rangking observasi antara observasi 1 (R_1) dan 2 (R_2) dan $(0,95 * 26)$ rangking observasi antara observasi 24 (R_{24}) dan 25 (R_{25}). Dengan menggunakan interpolasi linier pada observasi ke-1, 2 , 24 dan 25, interval prediksi yang diperoleh adalah sebagai berikut:

$$X_1 + \left(\frac{R_{(0.05 \cdot 26)} - R_1}{R_2 - R_1} \right) \cdot (X_2 - X_1) \text{ sampai dengan } X_{24} + \left(\frac{R_{(0.95 \cdot 26)} - R_{24}}{R_{25} - R_{24}} \right) \cdot (X_{25} - X_{24})$$

$$1,3 + \left(\frac{1,3 - 1}{2 - 1} \right) \cdot (1,5 - 1,3) \text{ sampai dengan } 340 + \left(\frac{24,5 - 24}{25 - 24} \right) \cdot (580 - 340)$$

$$1,4 \text{ sampai dengan } 508 \text{ ppb}$$

Observasi baru diluar rentang tersebut akan dianggap berasal dari distribusi yang berbeda dengan tingkat error sebesar 10% ($\alpha=10\%$).

Dengan menggunakan R pembaca dapat menghitung interval prediksi menggunakan fungsi berikut:

```
PInp <- function(x, alpha){
  # mengurutkan data
  x_sort = sort(x)
  # jumlah observasi
  n = length(x)
  # menghitung alpha masing-masing sisi
  err <- alpha/2
  # menentukan rangkin observasi sesuai alpha
  rl = err*(n+1)
  ru = (1-err)*(n+1)
  # menentukan observasi untuk interpolasi linier
  rl_1= ceiling(rl) # bulatkan ke bawah
  rl_2= floor(rl) # bulatkan ke atas
  ru_1= ceiling(ru)
  ru_2= floor(ru)}
```

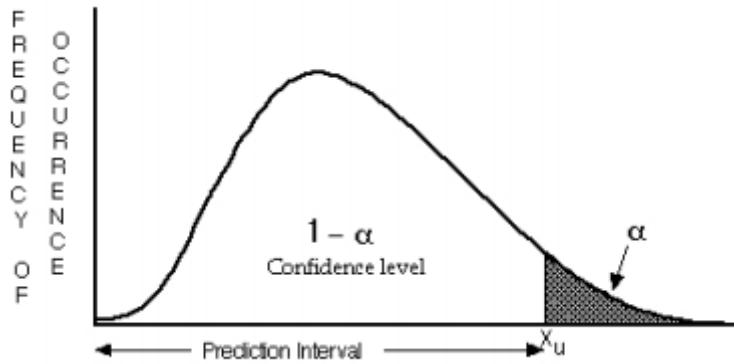


Figure 10.9: Prediksi interval satu sisi (Helsel dan Hirsch, 2002)

```
# menentukan interval prediksi
LPI = round(x_sort[rl_1]+((rl-rl_1)/(rl_2-rl_1))*(x_sort[rl_2]-x_sort[rl_1]),1)
UPI = round(x_sort[ru_1]+((ru-ru_1)/(ru_2-ru_1))*(x_sort[ru_2]-x[ru_1]),1)
# menggabungkan hasil
data = data.frame("Lower PI"=LPI,
                   "Upper PI"=UPI)
return(data)
}
```

Note:

- **x**: vektor numerik
- **alpha**: alpha level yang digunakan

```
PInp(x=gwardat$konsentrasi, alpha=0.1)
```

```
##   Lower.PI Upper.PI
## 1      1.4     508
```

10.5.2 Interval Prediksi Nonparametrik Satu Sisi

Interval prediksi satu sisi digunakan jika kita ingin mengecek apakah pengamatan baru lebih besar dari data yang ada, atau lebih kecil dari data yang ada, tetapi tidak keduanya. Keputusan untuk menggunakan interval satu sisi harus didasarkan sepenuhnya pada pertanyaan yang menarik. Seharusnya tidak ditentukan setelah melihat data dan memutuskan bahwa pengamatan baru cenderung hanya lebih besar, atau hanya lebih kecil, daripada informasi yang ada. Interval satu sisi menggunakan α dibanding $\alpha/2$ sebagai nilai error, menempatkan semua error di satu sisi interval (Gambar 10.9). Interval prediksi dituliskan berdasarkan Persamaan (10.12).

$$PI_{np} : x_{baru} < X_{\alpha \cdot (n+1)} \text{ atau } x_{baru} > X_{[1-\alpha] \cdot (n+1)} \quad (10.12)$$

Untuk memahami penerapannya, misalkan kita memiliki nilai arsenik baru dengan konsentrasi 355 ppb. Kita perlu menentukan apakah nilai tersebut lebih besar dari sebagian besar data yang ada.

Dengan menggunakan Persamaan (10.12) dan $\alpha=0.1$ atau tingkat kepercayaan 90%, interval prediksi satu sisi atau data teratas dari persentil ke-90 dari data yang ada adalah $X_{0.9} \cdot 26 = X_{23,4}$. Dengan menggunakan interpolasi linier pada observasi data dengan rangking ke-23 (R_{23}) dan 24 (R_{24}) diperoleh:

$$X_{23} + 0,4 \cdot (X_{24} - X_{23}) = 300 + 0,4 \cdot 40 = 316 \text{ ppb}$$

Berdasarkan data yang diperoleh diketahui bahwa batas atas dari interval prediksi adalah $316 < 355$ ppb, sehingga disimpulkan bahwa konsentrasi 355 ppb lebih besar dari sebagian besar data yang ada.

Dengan menggunakan R interval prediksi menggunakan satu sisi dapat dihitung menggunakan fungsi berikut:

```
PInp_os <- function(x, obs, alpha, side){
  # mengurutkan data dari yang terkecil
  x_sort = sort(x)
  # jumlah observasi
  n = length(x)
  # rangking observasi
  ru = (1-alpha)*(n+1)
  ru_1 = ceiling(ru)
  ru_2 = floor(ru)
  rl = alpha*(n+1)
  rl_1 = ceiling(rl)
  rl_2 = floor(rl)
  # perhitungan interval atas dan bawah
  PIup = x_sort[ru_1]+((ru-ru_1)/(ru_2-ru_1))*(x_sort[ru_2]-x_sort[ru_1])
  PIdown = x_sort[rl_1]+((rl-rl_1)/(rl_2-rl_1))*(x_sort[rl_2]-x_sort[rl_1])
  # decision making
  if((side=="upper") & (PIup<obs)){
    cat("PI =", PIup, ", observasi baru=", obs)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi lebih besar dibandingkan sebagian besar nilai yang ada")
  } else if((side=="lower") & (PIdown>obs)){
    cat("PI =", PIdown, ", observasi baru=", obs)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi lebih kecil dibandingkan sebagian besar nilai yang ada")
  } else if(side==""){
    print("side belum ditentukan tentukan apakah lower atau upper")
  } else{
    cat("batas bawah =", PIdown, ", batas atas =", PIup)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi sama dengan sebagian besar nilai yang ada")
  }
}
```

Note:

- **x**: vektor numerik
- **alpha**: alpha level yang digunakan
- **obs**: observasi baru yang akan dibandingkan
- **side**: untuk memilih jenis uji satu sisi yang digunakan. nilai yang mungkin adalah **Upper** (membandingkan dengan limit atas) dan **Lower** (membandingkan dengan limit bawah)

```
PInp_os(x=gwardat$konsentrasi, obs=355, alpha=0.1, side="upper")
```

```
## PI = 316 ,observasi baru= 355
## -----
## Kesimpulan:
## nilai observasi lebih besar dibandingkan sebagian besar nilai yang ada
```

10.6 Interval Prediksi Parametrik

Interval prediksi parametrik juga digunakan untuk menentukan apakah pengamatan baru kemungkinan berasal dari distribusi yang berbeda dari data yang dikumpulkan sebelumnya. Namun, pada metode parametrik asumsi bentuk dari distribusi data akan diperhitungkan. Asumsi ini memberikan lebih banyak informasi untuk membangun interval, asalkan asumsi tersebut valid. Jika data tidak mengikuti distribusi yang diajukan, interval prediksi mungkin tidak akurat.

10.6.1 Interval Prediksi Distribusi Simetris

Asumsi yang digunakan untuk melakukan perhitungan interval prediksi untuk distribusi data yang simetris adalah data haruslah berdistribusi normal. Interval prediksi selanjutnya dibentuk secara simetris pada kedua sisi nilai mean. Interval ini lebih lebar rentangnya dibandingkan dengan interval kepercayaan nilai mean. Persamaan matematis yang digunakan untuk menghitungnya dituliskan pada Persamaan (10.13).

$$PI = \bar{X} - t_{(\frac{\alpha}{2}, n-1)} \cdot \sqrt{s^2 + \left(\frac{s^2}{n}\right)} \text{ sampai } \bar{X} + t_{(\frac{\alpha}{2}, n-1)} \cdot \sqrt{s^2 + \left(\frac{s^2}{n}\right)} \quad (10.13)$$

Untuk interval satu sisi Persamaan (10.13), menjadi Persamaan (10.14).

$$PI = \bar{X} - t_{(\alpha, n-1)} \cdot \sqrt{s^2 + \left(\frac{s^2}{n}\right)} \text{ sampai } \bar{X} + t_{(\alpha, n-1)} \cdot \sqrt{s^2 + \left(\frac{s^2}{n}\right)} \quad (10.14)$$

Untuk lebih memahaminya misalkan terdapat hasil pengukuran baru konsentrasi arsenik sebesar 350 ppb dengan menggunakan data pada Tabel 10.3 sebagai pembanding. Buktikan bahwa observasi baru tersebut berasal dari distribusi yang sama dengan $\alpha=5\%$.

Dengan menggunakan Persamaan (10.13), interval prediksi dapat dihitung sebagai berikut:

$$PI = 98,4 - t_{(0.025, 24)} \cdot \sqrt{144,7^2 + \frac{144,7^2}{25}} \text{ sampai } 98,4 + t_{(0.025, 24)} \cdot \sqrt{144,7^2 + \frac{144,7^2}{25}}$$

$$PI = 98,4 - 2,064 \cdot 147,6 \text{ sampai } 98,4 + 2,064 \cdot 147,6$$

$$PI = -206,25 \text{ sampai } 403,05$$

Berdasarkan hasil perhitungan yang dilakukan terlihat bahwa limit interval prediksi yang dihasilkan terdapat nilai negatif. Konsentrasi negatif mengindikasikan bahwa data yang digunakan tidaklah simetris sehingga penggunaan interval prediksi untuk data yang simetris tidak dapat digunakan pada data tersebut. Metode perhitungan interval prediksi untuk data asimetris lebih cocok untuk digunakan.

Pada R interval prediksi disekitar nilai mean dapat dihitung menggunakan fungsi berikut:

```

PI_sim <- function(x, obs, alpha, side){
  # menghitung nilai mean
  ave = mean(x, na.rm=TRUE)
  # menghitung nilai varians data
  var = var(x, na.rm=TRUE)
  # menghitung df
  n = length(x)
  df = n-1
  # perhitungan rentang satu sisi
  pi_l1 = ave-qt((1-alpha), df)*sqrt(var+(var/n))
  pi_u1 = ave+qt((1-alpha), df)*sqrt(var+(var/n))
  # perhitungan rentang dua sisi
  pi_l2 = ave-qt((1-alpha/2), df)*sqrt(var+(var/n))
  pi_u2 = ave+qt((1-alpha/2), df)*sqrt(var+(var/n))
  # decision making
  if(side=="upper" & obs>pi_u1){
    cat("PI Atas =",pi_u1,",observasi baru=",obs)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi lebih besar dibandingkan sebagian besar nilai yang ada")
  }else if(side=="lower" & obs<pi_l1){
    cat("PI Bawah =",pi_l1,",observasi baru=",obs)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi lebih kecil dibandingkan sebagian besar nilai yang ada")
  }else if(side=="two side" & obs>pi_u2){
    cat("PI Bawah =",pi_l2,",observasi baru=",obs, ",PI Atas =",pi_u2)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi lebih besar dibandingkan sebagian besar nilai yang ada")
  }else if(side=="two side" & obs<pi_l2){
    cat("PI Bawah =",pi_l2,",observasi baru=",obs, ",PI Atas =",pi_u2)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi lebih kecil dibandingkan sebagian besar nilai yang ada")
  }else if(side=="upper" & obs<pi_u1){
    cat("PI Atas =",pi_u1,",observasi baru=",obs)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi sama dengan sebagian besar nilai yang ada")
  }else if(side=="lower" & obs>pi_l1){
    cat("PI Bawah =",pi_l1,",observasi baru=",obs)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi sama dengan sebagian besar nilai yang ada")
  }else{
    cat("PI Bawah =",pi_l2, ",observasi baru=",obs,",PI Atas =",pi_u2)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi sama dengan sebagian besar nilai yang ada")
  }
}

```

Note:

- **x**: vektor numerik
- **alpha**: alpha level yang digunakan
- **obs**: observasi baru yang akan dibandingkan
- **side**: untuk memilih jenis uji digunakan. nilai yang mungkin adalah **upper** (membandingkan dengan limit atas uji satu sisi), **lower** (membandingkan dengan limit bawah uji satu sisi) dan **two side** (uji dua sisi).

```
# interval prediksi satu sisi
PI_sim(x = gwardat$konsentrasi, obs = 350, alpha=0.05, side=2)
```

```
## PI Bawah = -206.2 ,observasi baru= 350 ,PI Atas = 402.9
## -----
## Kesimpulan:
## nilai observasi sama dengan sebagian besar nilai yang ada
```

10.6.2 Interval Prediksi Untuk Distribusi Data Yang Tidak Simetris

Untuk distribusi data yang tidak simetris, data perlu dilakukan transformasi terlebih dahulu sebelum dilakukan. Data di lingkungan khusunya parameter di air cenderung memiliki bentuk distribusi tidak simetris (cenderung memiliki kemencengan positif). Transformasi logaritmik biasanya dapat digunakan untuk data tersebut agar bentuknya dapat simetris dan dapat memenuhi asumsi normalitas pada data. Data yang telah dilakukan transformasi selanjutnya dihitung menggunakan Persamaan (10.13) untuk interval prediksi dua sisi dan Persamaan (10.14) untuk interval prediksi satu sisi. Hasil perhitungan selanjutnya dilakukan transformasi kembali sesuai dengan invers dari transformasinya dalam hal ini menggunakan transformasi eksponensial (jika transformasi awalnya adalah natural log). Untuk data dengan bentuk distribusi logaritmik (kemencengan positif), interval prediksi yang digunakan disajikan pada Persamaan (10.15) (dua sisi) dan Persamaan (10.16) (satu sisi).

$$PI = \exp \left(\bar{X} - t_{(\frac{\alpha}{2}, n-1)} \cdot \sqrt{s^2 + \left(\frac{s^2}{n} \right)} \right) \text{ sampai } \exp \left(\bar{X} + t_{(\frac{\alpha}{2}, n-1)} \cdot \sqrt{s^2 + \left(\frac{s^2}{n} \right)} \right) \quad (10.15)$$

$$PI = \exp \left(\bar{X} - t_{(\alpha, n-1)} \cdot \sqrt{s^2 + \left(\frac{s^2}{n} \right)} \right) \text{ sampai } \exp \left(\bar{X} + t_{(\alpha, n-1)} \cdot \sqrt{s^2 + \left(\frac{s^2}{n} \right)} \right) \quad (10.16)$$

dimana $y = \ln(x)$, \bar{y} adalah nilai rata-rata dari transformasi logaritmik data, dan s_y^2 adalah varians dari transformasi logaritmik data.

Dengan menggunakan contoh soal sebelumnya misalkan terdapat observasi baru konsentrasi arsenik sebesar 350 ppb. Kita perlu menentukan apakah observasi baru tersebut berasal dari distribusi yang sama berdasarkan data pada Tabel 10.3.

Berdasarkan hasil visualisasi diketahui bahwa distribusi data yang dihasilkan memiliki bentuk kemencengan positif sehingga interval prediksi asimetris dapat digunakan. Dengan menggunakan $\alpha=5\%$ prediksi interval dua sisi dapat dihitung menggunakan Persamaan (10.15).

$$\begin{aligned} \ln(PI) &= 3,71 - t_{(0.025, 24)} \cdot \sqrt{1,96^2 + \frac{1,96^2}{25}} \text{ sampai } 3,71 + t_{(0.025, 24)} \cdot \sqrt{1,96^2 + \frac{1,96^2}{25}} \\ \ln(PI) &= 3,71 - 2,064 \cdot 2,11 \text{ sampai } 3,71 + 2,064 \cdot 2,11 \end{aligned}$$

$$\ln(PI) = -1,19 \text{ sampai } 7,53$$

$$PI = 0,31 \text{ sampai } 1476.07$$

Berdasarkan hasil yang diperoleh diketahui bahwa observasi baru berada diantara rentang tersebut. Rentang yang dihasilkan cukup besar yang disebabkan karena tingkat kepercayaan yang digunakan juga besar (95%). Pembaca dapat juga menggunakan tingkat kepercayaan yang lain seperti 99% dan 90%. Semakin besar alpha yang digunakan interval prediksi yang dihasilkan semakin kecil. Namun perlu diingat bahwa semakin kecil rentangnya maka error (alpha) juga semakin besar.

Pembaca juga dapat menggunakan bentuk transformasi lain untuk membentuk data yang lebih simetris dan memenuhi asumsi distribusi normal. Bentuk transformasi lain akan mengubah bentuk persamaan yang digunakan. Transformasi kuadrat misalnya akan mengubah transformasi pada persamaan (10.15) dan Persamaan (10.16) menjadi akar kuadrat.

Pada R interval prediksi dengan bentuk transformasi data logaritmik dapat dituliskan sebagai berikut:

```
PI_asim <- function(x, obs, alpha, side){
  # transformasi logaritmik (kemencengan positif)
  x_trans = log(x)
  # menghitung nilai mean
  ave = mean(x_trans)
  # menghitung nilai varians data
  var = var(x_trans)
  # menghitung df
  n = length(x)
  df = n-1
  # perhitungan rentang satu sisi
  pi_l1 = exp(ave-qt((1-alpha), df)*sqrt(var+(var/n)))
  pi_u1 = exp(ave+qt((1-alpha), df)*sqrt(var+(var/n)))
  # perhitungan rentang dua sisi
  pi_l2 = exp(ave-qt((1-alpha/2), df)*sqrt(var+(var/n)))
  pi_u2 = exp(ave+qt((1-alpha/2), df)*sqrt(var+(var/n)))
  # decision making
  if(side=="upper" & obs>pi_u1){
    cat("PI Atas =",pi_u1,",observasi baru=",obs)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi lebih besar dibandingkan sebagian besar nilai yang ada")
  }else if(side=="lower" & obs<pi_l1){
    cat("PI Bawah =",pi_l1,",observasi baru=",obs)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi lebih kecil dibandingkan sebagian besar nilai yang ada")
  }else if(side=="two side" & obs>pi_u2){
    cat("PI Bawah =",pi_l2,",observasi baru=",obs, ",PI Atas =",pi_u2)
    cat("\n-----")
    cat("\nKesimpulan:")
    cat("\nnilai observasi lebih besar dibandingkan sebagian besar nilai yang ada")
  }else if(side=="two side" & obs<pi_l2){
    cat("PI Bawah =",pi_l2,",observasi baru=",obs, ",PI Atas =",pi_u2)
    cat("\n-----")
    cat("\nKesimpulan:")}
```

```

cat("\nnilai observasi lebih kecil dibandingkan sebagian besar nilai yang ada")
}else if(side=="upper" & obs<pi_u1){
  cat("PI Atas =",pi_u1,",observasi baru=",obs)
  cat("\n-----")
  cat("\nKesimpulan:")
  cat("\nnilai observasi sama dengan sebagian besar nilai yang ada")
}else if(side=="lower" & obs>pi_l1){
  cat("PI Bawah =",pi_l1,",observasi baru=",obs)
  cat("\n-----")
  cat("\nKesimpulan:")
  cat("\nnilai observasi sama dengan sebagian besar nilai yang ada")
}else{
  cat("PI Bawah =",pi_l2, ",observasi baru=",obs,"PI Atas =",pi_u2)
  cat("\n-----")
  cat("\nKesimpulan:")
  cat("\nnilai observasi sama dengan sebagian besar nilai yang ada")
}
}

```

Note:

- **x**: vektor numerik
- **alpha**: alpha level yang digunakan
- **obs**: observasi baru yang akan dibandingkan
- **side**: untuk memilih jenis uji digunakan. nilai yang mungkin adalah **upper** (membandingkan dengan limit atas uji satu sisi), **lower** (membandingkan dengan limit bawah uji satu sisi) dan **two side** (uji dua sisi).

```

# interval prediksi satu sisi
PI_asim(x = gwardat$konsentrasi, obs = 350, alpha=0.05, side=2)

```

```

## PI Bawah = 0.386 ,observasi baru= 350 ,PI Atas = 1476
## -----
## Kesimpulan:
## nilai observasi sama dengan sebagian besar nilai yang ada

```

10.7 Interval Kepercayaan Persentil (Interval Toleransi)

Kuantil atau persentil telah digunakan secara tradisional dalam sumber daya air untuk menggambarkan frekuensi kejadian banjir. Dengan demikian banjir 100 tahun adalah persentil ke-99 (0,99 kuantil) dari distribusi puncak banjir tahunan. Besarnya banjir yang diperkirakan hanya akan dilampaui sekali dalam 100 tahun. Banjir 20 tahun besarnya besarnya yang diperkirakan hanya akan dilampaui sekali dalam 20 tahun (5 kali dalam 100 tahun), atau merupakan persentil ke-95 dari puncak tahunan. Demikian pula, banjir 2 tahun adalah median atau persentil ke-50 dari puncak tahunan. Persentil banjir ditentukan dengan asumsi bahwa aliran puncak mengikuti distribusi yang ditentukan seperti distribusi Log Pearson type III atau distribusi Gumbel.

Interval kepercayaan persentil berbeda dengan interval kepercayaan median. Hal yang paling jelas terlihat adalah interval kepercayaan persentil mengukur interval kepercayaan pada setiap persentil data yang ada, sedangkan interval kepercayaan media hanya mengukur pada lokasi pusat data atau persentil ke-50.

Interval kepercayaan persentil juga disebut sebagai interval toleransi. Nilai persentil digunakan sebagai **koefisien cakupan** dari interval toleransi. Pada chapter ini akan dibahas lebih jauh mengenai metode perhitungan interval toleransi baik dengan metode parametrik maupun dengan metode nonparametrik.

10.7.1 Interval Kepercayaan Nonparametrik Persentil

Metode perhitungan interval kepercayaan nonparametrik persentil mirip dengan perhitungan interval kepercayaan median. Kita akan menggunakan kembali tabel binomial jika sampel kita kecil untuk menentukan limit atas dan bawah yang merupakan nilai kritis dari alpha yang telah kita tetapkan. Nilai kritis ini selanjutnya akan ditransformasikan kedalam bentuk rangking pada data yang menunjukkan titik observasi ujung pada interval kepercayaan.

Tabel binomial dimasukkan pada kolom dengan nilai p , persentil yang diinginkan interval kepercayaannya. Jadi untuk interval kepercayaan pada persentil ke-75, kolom $p = 0,75$ digunakan. Cari pada baris kolom tersebut sampai n dengan probabilitas mendekati alpha level ($\alpha/2$) ditemukan. Nilai kritis x_l bawah adalah bilangan bulat yang sesuai dengan probabilitas p^* . Nilai kritis kedua x_u juga diperoleh dengan melanjutkan pada kolom tersebut sampai menemukan probabilitas $p' = (1 - \frac{\alpha}{2})$. Nilai kritis x_l dan x_u digunakan untuk menghitung rangking R_l dan R_u yang sesuai dengan nilai data di ujung atas dan bawah limit kepercayaan (Persamaan (10.17) dan Persamaan (10.18)). Level interval kepercayaan yang dihasilkan akan sama dengan $(p' - p)$.

$$R_l = x_l + 1 \quad (10.17)$$

$$R_u = x_u \quad (10.18)$$

Untuk memahami mengenai penerapan interval kepercayaan persentil diberikan sebuah contoh kita diminta untuk menentukan 95% interval kepercayaan nilai persentil ke-20 ($C_{0.20}$) data konsentrasi arsenik pada Tabel 10.3 (p=0,2).

Berdasarkan data pada Tabel 10.3, nilai persentil ke-20 ($\bar{C}_{0.20} = 3.36\text{ppb}$, yaitu data yang berada pada rangking $0,2*(26)=5,2$ atau dua per sepuluh jarak antara data ke-5 dan ke-6. Untuk menentukan rentang kepercayaan persentil ke-20 sebenarnya dari data, kita perlu menggunakan kembali tabel binomial dengan menginputkan nilai $p=0,2$. nilai kritis x_l diperoleh dengan mencari probabilitas data pada kolom $p=0,2$ yang mendekati nilai $\alpha/2=0,025$ adalah 1 ($p'=0,027$, error probabilitas sisi bawah distribusi). Dengan menggunakan Persamaan (10.17), diperoleh $R_l = 2$. Dengan cara sama untuk sisi atas distribusi nilai kritis atas x_u diperoleh dengan menginputkan nilai $p=0,20$ dengan nilai probabilitas mendekati $1 - \frac{\alpha}{2} = 0,975$ diperoleh sebesar 9 ($p'=0,983$, error probabilitas sisi atas distribusi). Sehingga rentang kepercayaan 95,6% ($0,983-0,027=0,956$) untuk persentil ke-20 berada pada range data dengan rangkin ke-2 dan ke-9, atau

$$1,5 \leq C_{0.20} \leq 8 \text{ pada } \alpha = 0,044$$

Jika data yang kita miliki cukup besar dengan jumlah sampel $n > 20$ (sebagian buku menyebutkan $n > 30$), kita dapat menggunakan distribusi normal untuk memperkirakan rentang kepercayaan persentil. Persamaan yang digunakan untuk menentukan batas atas dan bawah disajikan pada Persamaan (10.19) dan Persamaan (10.20).

$$R_l = np + z_{\frac{\alpha}{2}} \cdot \sqrt{np(1-p)} + 0,5 \quad (10.19)$$

$$R_u = np + z_{[1-\frac{\alpha}{2}]} \cdot \sqrt{np(1-p)} + 0,5 \quad (10.20)$$

Dengan menggunakan contoh sebelumnya kita dapat menghitung kembali rentang kepercayaan 95% persentil ke-20 menggunakan Persamaan (10.19) dan Persamaan (10.20) diperoleh rangking data batas bawah dan atas sebagai berikut:

$$R_l = 25 \cdot 0,2 + (-1,96) \cdot \sqrt{25 \cdot 0,2(1-0,2)} + 0,5 = 5 - 1,96 \cdot 2 + 0,5 = 1,6$$

$$R_u = 25 \cdot 0,2 + 1,96 \cdot \sqrt{25 \cdot 0,2(1-0,2)} + 0,5 = 5 + 1,96 \cdot 2 + 0,5 = 9,4$$

Berdasarkan hasil perhitungan diperoleh rangking data batas bawah dan batas atas secara berurutan adalah data ke-2 (batas bawah) dan data-9 (batas atas). Hasil yang diperoleh ini sama dengan yang telah diperoleh sebelumnya.

Pada R kita dapat membentuk fungsi untuk menghitung interval kepercayaan persentil sesuai dengan yang kita inginkan seperti lokasi persentil yang ingin kita uji serta jenis uji yang digunakan berdasarkan jumlah sampel yang kita inputkan. Selain itu rentang kepercayaan ini dapat pula digunakan untuk menghitung rentang kepercayaan median atau persentil ke-50.

```
CI_npPercent <- function(x, p, alpha){
  # jumlah observasi
  n = length(x)
  # mengurutkan data
  x = sort(x)
  # membuat vektor yang akan menyimpan
  # hasil loop
  bl = rep(NA,n)
  bu = rep(NA,n)
  # decision makin
  if(n<= 20){
    # looping
    for(i in 1:n){
      bl[i] = pbinom(i, n, p)
      if(b>alpha/2){
        break
      }
    }
    for(i in 1:n){
      bu[i] = pbinom(i, n, p)
      if(b>(1-(alpha/2))){
        break
      }
    }
    # menghitung selisih terhadap alpha
    dbl = abs(alpha-bl)
    dbu = abs(alpha-bu)
    # mencari titik kritis
    min_bl = which.min(dbl)
    min_bu = which.min(dbu)
    # menhitung rangking nilai bawah dan atas
    rl = min_bl + 1
    ru = min_bu
    # mencari data sesuai rangking bawah dan atas
    LCI = x[rl]
    UCI = x[ru]
  }else{
    # menghitung rangking nilai bawah dan atas
    rl = (n*p)+qnorm(alpha/2)*sqrt((n*p)*(1-p))+0.5
    ru = (n*p)+(qnorm(1-(alpha/2))*sqrt((n*p)*(1-p)))+0.5
    # mencari data sesuai rangking bawah dan atas
    LCI = x[floor(rl)]+
```

```

((rl-floor(rl))/(ceiling(rl)-floor(rl)))*(x[ceiling(rl)]-x[floor(rl)])
UCI = x[floor(ru)]+
((ru-floor(ru))/(ceiling(ru)-floor(ru)))*(x[ceiling(ru)]-x[floor(ru)])
}
cat("Lower CI=", LCI, " <= C(" , p, ") <= ",
"Upper CI=", UCI)
}

```

Note:

- **x**: vektor numerik.
- **p**: persentil yang ingin dicari. Nilai berkisar antara 0 sampai 1.
- **alpha**: alpha level yang digunakan.

Pembaca dapat menjalankan fungsi tersebut menggunakan data pada contoh soal sebelumnya yaitu mencari interval kepercayaan persentil ke-20. Jalankan sintaks berikut untuk mengetahui hasil yang diperoleh.

```
CI_npPercent(x= gwardat$konsentrasi, p= 0.2, alpha=0.05)
```

10.7.2 Uji Nonparametrik Untuk Persentil

Pengujian persentil dilakukan untuk mengecek apakah sebuah persentil berbeda (lebih besar atau lebih kecil) dibandingkan dengan sejumlah nilai. Sebagai contoh misalkan terdapat median kualitas harian suatu parameter tidak boleh melebihi standar yang berlaku sebesar X_0 ppb. Contoh lain dalam bidang hidrologi periode ulang hujan (PUH) 10 tahun atau persentil ke-90 dari debit puncak tahunan suatu kawasan dapat dilakukan pengujian apakah nilai yang ada dilapangan berbeda dengan PUH 10 tahunan yang telah kita hitung untuk digunakan dalam mendesain saluran drainase. Pembahasan pengujian persentil tersebut tidak akan sampai menyinggung pengujian hipotesis yang akan dibahas pada Chapter selanjutnya. Pembahasan akan berkisar membandingkan suatu nilai dengan interval kepercayaan seperti yang telah dijelaskan pada pembahasan terkait interval prediksi.

Pengujian apakah sebuah nilai X_0 berbeda dengan sejumlah rentang nilai yang ditetapkan dapat dilakukan dengan pengujian dua sisi dan satu sisi. Pengujian satu sisi melihat apakah suatu nilai X_0 berada diluar interval kepercayaan persentil atau diantara nilai batas bawah X_l dan nilai batas atasnya X_u (lihat Gambar 10.10). Sedangkan pengujian satu sisi melihat apakah suatu nilai lebih besar atau lebih kecil (tergantung apakah pengujian satu sisi sebelah atas distribusi atau sebelah bawah distribusi) dari interval kepercayaan persentil yang digunakan (lihat Gambar 10.11 dan Gambar 10.12).

Untuk menghitungnya secara nonparametrik menggunakan Persamaan (10.19) dan Persamaan (10.20) untuk jumlah sampel kecil sedangkan untuk sampel besar kita dapat menggunakan Persamaan (10.19) dan Persamaan (10.20).

Dengan menggunakan kembali data pada Tabel 10.3 kita akan menghitung apakah kadar arsenik kualitas air tanah tersebut melebihi baku mutu arsenik pada air minum dengan baku mutu konsentrasi arsenik tidak melampaui 300 ppb. Dengan menggunakan nilai $\alpha=0,05$ dan batas bawah persentil yang digunakan sebagai acuan pembanding adalah persentil ke-90 dapat dihitung sebagai berikut:

$$\bar{C}_{0.90} = (25 + 1) \cdot 0,9 = 23,4 \text{ (data point)} = 300 + 0,4(340 - 300) = 316 \text{ ppb}$$

Karena jumlah sampel lebih besar dari 20, maka kita dapat menghitung batas atas data menggunakan Persamaan (10.20).

$$R_l = np + z_{0,05} \sqrt{np(1-p)} = 25 \cdot 0,9 + (-1,64) \cdot \sqrt{2,25} + 0,5 = 20,5$$

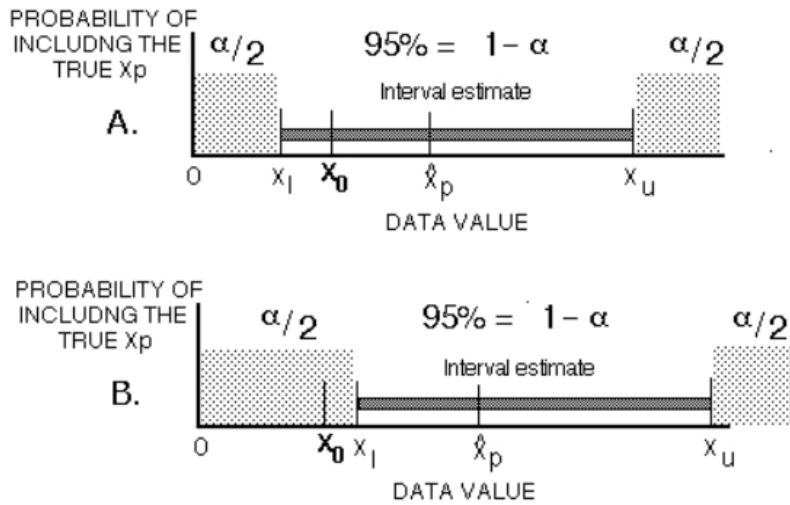


Figure 10.10: Interval estimasi persentil X_p sebagai penguji apakah $X_p=X_0$. A) X_0 didalam interval estimasi sehingga X_p tidak berbeda secara signifikan dari X_0 , B) X_0 berada diluar rentang estimasi sehingga X_p berbeda secara signifikan dari X_0 . (Helsel dan Hirsch, 2002)

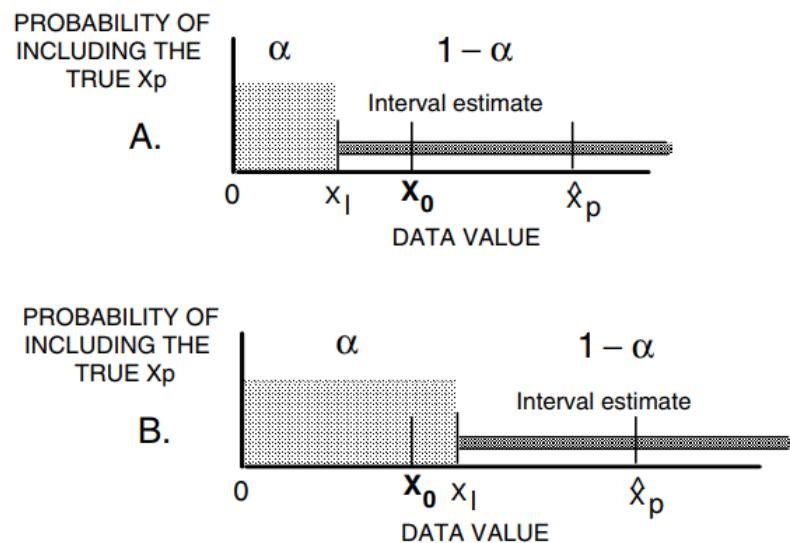


Figure 10.11: Interval estimasi persentil X_p sebagai penguji apakah $X_p>X_0$. A) X_0 didalam interval estimasi sehingga X_p tidak signifikan lebih besar dari X_0 , B) X_0 berada diluar rentang estimasi sehingga X_p signifikan lebih besar dari X_0 . (Helsel dan Hirsch, 2002)

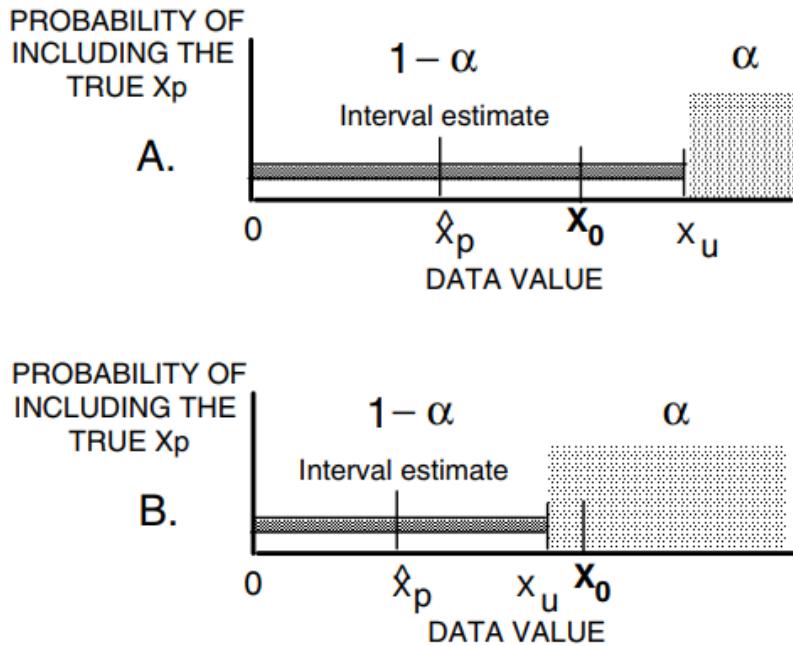


Figure 10.12: Interval estimasi persentil X_p sebagai penguji apakah $X_p > X_0$. A) X_0 didalam interval estimasi sehingga X_p tidak signifikan lebih kecil dari X_0 , B) X_0 berada diluar rentang estimasi sehingga X_p signifikan lebih kecil dari X_0 . (Helsel dan Hirsch, 2002)

Kita dapat membulatkan hasilnya menjadi observasi 20 atau 21. Interpolasi linier dapat dilakukan sehingga diperoleh nilai observasi sebesar 215 ppb. Nilai ini lebih kecil dibandingkan $X_0=300$ ppb, sehingga baku mutu arsenik belum terlampaui oleh kualitas air tanah tersebut.

Kita dapat menggunakan fungsi `CI_npPercent()` untuk menghitung rentang persentil yang kita inginkan. Untuk pengujian satu sisi nilai *alpha* yang akan diinputkan perlu dikali oleh dua karena fungsi tersebut pada dasarnya digunakan untuk menghitung rentang kepercayaan persentil secara nonparametrik (nilai *alpha* dibagi pada kedua sisi). Berikut adalah contoh sintaks untuk menguji apakah sampel yang kita miliki melebihi baku mutu (persentil ke-90 dan *alpha*=5%):

```
CI_npPercent(gwardat$konsentrasi, 0.9, 0.1)
```

10.7.3 Interval Kepercayaan Parametrik Untuk Persentil

Interval kepercayaan untuk persentil juga dapat dihitung dengan mengasumsikan bahwa data mengikuti distribusi tertentu. Asumsi distribusi digunakan karena sering ada data yang tidak cukup untuk menghitung persentil dengan presisi yang diperlukan. Menambahkan informasi yang terkandung dalam distribusi akan meningkatkan ketepatan estimasi selama asumsi distribusi masuk akal. Namun ketika distribusi yang diasumsikan tidak sesuai dengan data dengan baik, estimasi yang dihasilkan kurang akurat, dan lebih menyentak, daripada jika tidak ada yang diasumsikan. Sayangnya, situasi di mana asumsi paling dibutuhkan ketika ukuran sampel yang kecil, adalah situasi yang sama di mana sulit untuk menentukan apakah data mengikuti distribusi yang diasumsikan.

Pada interval kepercayaan parametrik asumsi terhadap kecocokan data terhadap suatu distribusi perlu diperhatikan. Data di lingkungan umumnya memiliki bentuk distribusi mengikuti distribusi lognormal. Selain itu, distribusi yang sering sekali digunakan adalah distribusi Pearson Tipe III dan Gumbel. Kedua pendekatan distribusi tersebut akan mempengaruhi metode perhitungan yang digunakan. Sehingga pengetahuan yang lebih baik mengenai distribusi tersebut diperlukan. Pada buku ini kita hanya akan membahas mengenai

perhitungan interval kepercayaan menggunakan distribusi lognormal. Pembaca dapat membaca mengenai penerapan distribusi lainnya melalui jurnal yang ditulis oleh Wei dan Song (2019).

Perhitungan estimasi titik dan interval untuk persentil dengan asumsi distribusi lognormal dapat dilakukan dengan mudah. Pertama sampel rata-rata \bar{y} dan sampel simpangan baku s_y logaritma dihitung. Estimasi titik kemudian dihitung menggunakan Persamaan (10.21).

$$X_p = \exp(\bar{y} + z_p \cdot s_y) \quad (10.21)$$

dimana z_p merupakan kuantil ke- p dari distribusi normal standard dan $y = \ln(x)$.

Estimasi interval untuk median sebelumnya diberikan pada Persamaan (10.7) dengan asumsi bahwa data mengikuti distribusi lognormal. Untuk persentil lainnya, interval kepercayaan dihitung menggunakan distribusi t non-sentral (Stedinger, 1983). Tabel distribusi itu ditemukan dalam artikel Stedinger, dengan list yang lebih lengkap terdapat pada perpustakaan online atau perangkat lunak matematika. Interval kepercayaan pada X_p dihitung menggunakan Persamaan (10.22).

$$CI(X_p) = \exp\left(\bar{y} + \zeta_{\frac{\alpha}{2}} \cdot s_y, \bar{y} + \zeta_{[1-\frac{\alpha}{2}]} \cdot s_y\right) \quad (10.22)$$

dimana $\zeta\alpha/2$ merupakan $\alpha/2$ dari kuantil distribusi t non-sentral untuk persentil dengan ukuran sampel n yang diinginkan.

Untuk lebih memahami penerapannya pembaca dapat mengerjakan contoh soal pada bagian sebelumnya. Dengan menggunakan estimasi interval 90% kita perlu menentukan interval estimasi persentil ke-90 dari data konsentrasi arsenik dengan asumsi distribusi yang digunakan berupa distribusi lognormal.

Dengan menggunakan Persamaan (10.21) estimasi titik persentil ke-90 dapat dihitung.

$$C_{0,90} = \exp(3,17 + 1,28 \cdot 1,96) = 292,6 \text{ ppb}$$

Nilai tersebut lebih rendah dibanding estimasi konsentrasi sebelumnya asumsi data mengikuti distribusi lognormal dengan konsentrasi persentil ke-90 arsenik sebesar 316 ppb.

Dengan menggunakan Persamaan (10.22), interval kepercayaan 90% dapat dihitung.

$$\exp(3,17 + 0,898 \cdot 1,96) < C_{0,90} < \exp(3,17 + 1,838 \cdot 1,96)$$

$$138,4 < C_{0,90} < 873,5$$

10.7.4 Uji Parametrik Untuk Persentil

Seperti pada bagian sebelumnya kita ingin melihat apakah persentil dari sekumpulan data berbeda dengan nilai tertentu (dapat berupa baku mutu). Pengujian dilakukan dengan melihat apakah nilai tertentu tersebut berada diantara interval kepercayaan persentil dari data (uji dua sisi), lebih besar atau lebih kecil dari batas bawah atau batas atas interval kepercayaan persentil (uji satu sisi). Langkah pengujian dilakukan sama dengan sebelumnya dengan menghitung terlebih dulu batas atas atau batas bawah persentil data yang selanjutnya dibandingkan dengan nilai tertentu.

Dengan menggunakan hasil dari perhitungan sebelumnya, dengan menggunakan alpha=0,05 kita perlu menentukan apakah batas bawah interval kepercayaan melampaui baku mutu arsenik sebesar 300 ppb (uji satu sisi). Berdasarkan hasil perhitungan diperoleh batas bawah interval kepercayaan persentil ke-90 sebesar 138,4 ppb atau lebih kecil dibandingkan batas yang ditentukan, sehingga disimpulkan bahwa konsentrasi arsenik persentil ke-90 pada data tidak melampaui baku mutu yang ditentukan.

10.8 Interval Kepercayaan Menggunakan Metode Bootstrap

Bootstrap merupakan metode inferensi populasi menggunakan data sampel. Metode ini dikembangkan oleh Bradley Efron pada tahun 1979. Jika pembaca ingin lebih mengenal metode ini pembaca dapat membaca makalahnya di tautan [berikut](#). Pembaca dapat membaca makalah tersebut secara gratis.

Bootstrap mengandalkan pengambilan sampel dengan pengembalian dari data sampel. Teknik ini dapat digunakan untuk memperkirakan standard error (se) dari setiap statistik dan untuk memperoleh interval kepercayaan (CI) untuk itu. Bootstrap sangat berguna ketika CI tidak memiliki bentuk tertutup, atau memiliki bentuk yang sangat rumit.

Misalkan kita memiliki sejumlah sampel dengan ukuran n : $X = \{x_1, x_2, \dots, x_n\}$ dan kita tertarik dengan CI dari beberapa statistik data $T = t(X)$. Metode ini sangat mudah untuk dikerjakan. Kita hanya perlu mengulang sejumlah R kali skema berikut: Untuk pengulangan ke- i , sampling dengan pengembalian n data dari data sampel yang tersedia. Namai sampel baru tersebut sebagai sampel bootstrap ke- i , X_i , dan hitung statistik (mean, median atau persentil) yang ingin dihitung interval kepercayaannya.

Sebagai hasilnya, kita akan mendapatkan nilai R dari statistik yang telah kita hitung: T_1, T_2, \dots, T_R . Kita dapat menyebutnya sebagai realisasi bootstrap dari T atau distribusi bootstrap dari T . Berdasarkan hal tersebut, kita dapat menghitung CI untuk T .

Pada contoh kali ini penulis akan menyajikan cara melakukan bootstrap untuk menghitung interval kepercayaan pada mean, median, dan persentil. Data yang penulis gunakan adalah data `gwardat` pada Tabel 10.3.

Bootstrap pada R dilakukan dengan menggunakan library `boot`. Berikut adalah contoh pertama menghitung interval kepercayaan median:

```
# memuat paket
library(boot)

##
## Attaching package: 'boot'

## The following object is masked from 'package:psych':
##   logit

## The following object is masked from 'package:car':
##   logit

# membuat hasil yang diperoleh lebih random
# dan reproducible
set.seed(100)

# membuat fungsi boot
med.boot.func <- function(x,i){
  median(x[i])
}

# melakukan bootstrap
median.boot <- boot(gwardat$konsentrasi,
                      # memasukkan fungsi boot
                      med.boot.func,
```

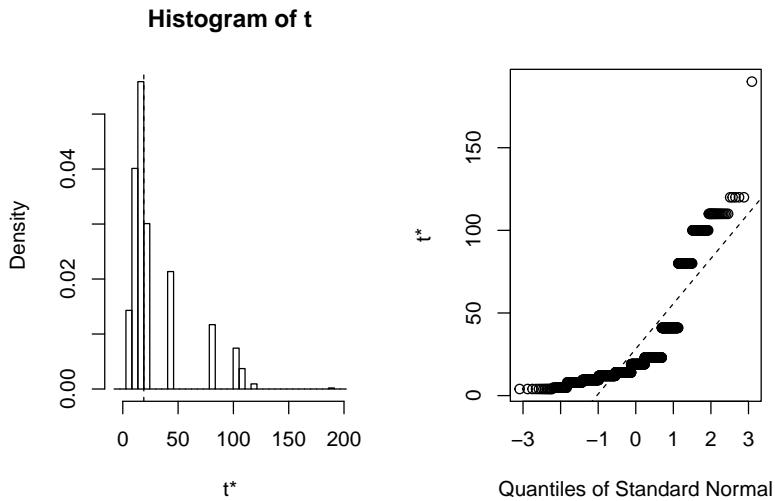


Figure 10.13: Distribusi bootstrap median

```
# menentukan jumlah replikasi
R=1000)

# print hasil
median.boot

## 
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
## Call:
## boot(data = gwardat$konsentrasi, statistic = med.boot.func, R = 1000)
##
## 
## Bootstrap Statistics :
##      original   bias   std. error
## t1*       19    9.279     27.31
```

Berdasarkan hasil yang diperoleh diketahui bahwa median dataset original sebesar 19. Pada hasil juga diperoleh nilai bias bootstrap. Nilai bias tersebut merupakan selisih dari nilai rata-rata 1000 median hasil bootstrap dikurangi dengan median sampel keseluruhan (original median). Standard error merupakan standar deviasi dari 1000 median yang terhitung.

Untuk mengetahui distribusi sampling yang telah kita lakukan. kita dapat melihat distribusi dengan mengeplotkan semua sampel bootstrap pada histogram dan QQ-plot dengan menjalankan sintaks berikut:

Berdasarkan Gambar 10.13 diketahui bahwa median tidak berdistribusi normal. Hal ini ditunjukkan dari distribusi pada histogram yang membentuk kemencenggan positif. Selain itu, distribusi data pada QQ-plot juga tidak mengikuti garis referensi yang ada sehingga dapat disimpulkan bahwa distribusi median bootstrap tidak berdistribusi normal.

Untuk menghitung interval kepercayaan median dengan tingkat kepercayaan 95%, kita dapat menggunakan fungsi `boot.ci()`. Berikut adalah sintaks yang digunakan:

```
boot.ci(boot.out=median.boot, conf=0.95)

## Warning in boot.ci(boot.out = median.boot, conf =
## 0.95): bootstrap variances needed for studentized
## intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = median.boot, conf = 0.95)
##
## Intervals :
## Level      Normal          Basic
## 95%   (-43.81,  63.25 )  (-72.00,  33.20 )
##
## Level      Percentile       BCa
## 95%   ( 4.8, 110.0 )  ( 4.8, 100.0 )
## Calculations and Intervals on Original Scale
```

Terdapat 4 buah hasil dari 4 buah metode yang dihasilkan sebagai output fungsi tersebut. Metode Normal mengasumsikan distribusi median bootstrap berdistribusi normal. Berdasarkan hasil visualisasi yang diperoleh diketahui bahwa distribusi data cenderung memiliki kemencengangan positif sehingga metode ini tidak dapat digunakan, Selain itu, rentang yang dihasilkan juga terdapat nilai negatif yang mustahik dihasilkan pada konsentrasi arsenik (sebagian besar parameter lingkungan memiliki nilai terkecil ≤ 0). Metode *Basic* memiliki asumsi bahwa tidak ada bias antara median data dengan median rata-rata hasil bootstrap. Berdasarkan hasil perhitungan yang dilakukan terlihat bahwa terdapat bias pada hasil bootstrap yang cukup besar sehingga metode ini tidak dapat diterapkan. Metode ketiga adalah metode persentil. Metode ini mengasumsikan bahwa distribusi median simetris, yang telah disinggung sebelumnya bahwa distribusi median tidak simetris sehingga metode ini tidak dapat diterapkan pada kasus ini. Metode terakhir adalah metode BCa (*bias-corrected and accelerated*) mengoreksi bias dan membuat lebih sedikit asumsi. Metode ini akan sering banyak kita gunakan pada data kita. Berdasarkan seluruh hasil yang telah diperoleh dapat kita simpulkan bahwa metode BCa cukup baik dalam menjelaskan interval kepercayaan median. Selain itu, metode persentil juga mempunyai hasil yang relatif mirip dengan BCa meskipun dari asumsi yang digunakan keempat metode tersebut tidak ada yang terpenuhi.

Pesan peringatan dalam *output* hasil menunjukkan bahwa interval kepercayaan kelima, disebut sebagai *studentized interval*, tidak dapat dihitung karena varians untuk sampel bootstrap tidak disediakan. Interval kepercayaan *studentized* berusaha untuk mengoreksi bias dengan “*studentizing*” setiap median yang dihitung (misal: mengurangi rata-rata median dan kemudian membaginya dengan standard error). Tampaknya tidak ada rumus umum untuk menghitung standard error untuk median, tetapi ada pedoman dalam literatur untuk memperkirakan kesalahan standar ketika populasi data yang mendasarinya diasumsikan terdistribusi secara normal. Dalam contoh ini, interval BCa tampaknya cukup baik. Perhatikan bahwa dalam beberapa kasus, mungkin ada peringatan bahwa “beberapa interval BCa mungkin tidak stabil.” Dalam hal itu, hasil BCa harus diabaikan jika intervalnya tampak tidak masuk akal, dan pilihan dibuat dari opsi lain yang tersisa, berdasarkan pada Ulasan histogram dan plot probabilitas estimasi bootstrap.

Setelah pembaca memahami prosedur melakukan bootstrap pada median, pembaca dapat melakukan bootstrap pada persentil dan mean. Untuk bootstrap mean jalankan sintaks berikut:

```
# membuat hasil yang diperoleh lebih random
# dan reproducible
set.seed(100)
```

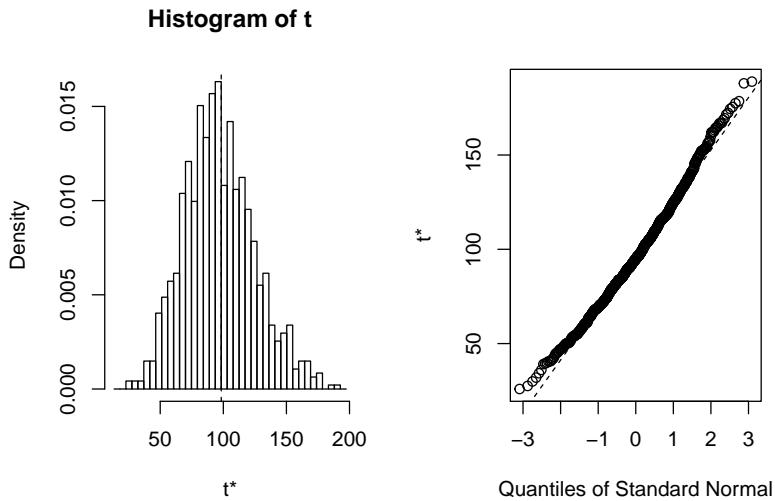


Figure 10.14: Distribusi bootstrap mean

```
# membuat fungsi boot
mean.boot.func <- function(x,i){
  mean(x[i])
}

# melakukan bootstrap
mean.boot <- boot(gwardat$konsentrasi,
  # memasukkan fungsi boot
  mean.boot.func,
  # menentukan jumlah replikasi
  R=1000)

# print hasil
mean.boot

## 
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
## 
## Call:
## boot(data = gwardat$konsentrasi, statistic = mean.boot.func,
##       R = 1000)
##
## 
## Bootstrap Statistics :
##      original   bias   std. error
## t1*    98.35  -1.397     27.84
```

Distribusi mean bootstrap disajikan pada Gambar 10.14

Berdasarkan hasil yang diperoleh, distribusi mean bootstrap mengikuti distribusi normal. Untuk memperoleh interval kepercayaan mean bootstrap jalankan sintaks berikut:

```
boot.ci(boot.out=mean.boot, conf=0.95)

## Warning in boot.ci(boot.out = mean.boot, conf = 0.95):
## bootstrap variances needed for studentized intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = mean.boot, conf = 0.95)
##
## Intervals :
## Level      Normal          Basic
## 95%   ( 45.18, 154.32 )  ( 39.21, 149.05 )
##
## Level      Percentile       BCa
## 95%   ( 47.66, 157.50 )  ( 56.43, 175.06 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

Berdasarkan hasil yang diperoleh dapat dilihat bahwa metode BCa tidak dapat digunakan karena hasil yang diperoleh tidak stabil. Ketiga metode lainnya dapat digunakan karena asumsi normalitas (simetri) terpenuhi. Selain itu bias yang dihasilkan relatif kecil sehingga metode *Basic* juga dapat digunakan.

Bootstrap terakhir yang kita lakukan adalah untuk memperoleh interval kepercayaan persentil dalam hal ini penulis akan mencobanya dengan persentil ke-90. Berikut alah sintaks yang digunakan:

```
# membuat hasil yang diperoleh lebih random
# dan reproducible
set.seed(100)

# membuat fungsi boot
p90.boot.func <- function(x,i){
  quantile(x[i], probs=0.9)
}

# melakukan bootstrap
p90.boot <- boot(gwardat$konsentrasi,
                  # memasukkan fungsi boot
                  p90.boot.func,
                  # menentukan jumlah replikasi
                  R=1000)

# print hasil
p90.boot

## 
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
## 
## Call:
## boot(data = gwardat$konsentrasi, statistic = p90.boot.func, R = 1000)
##
```

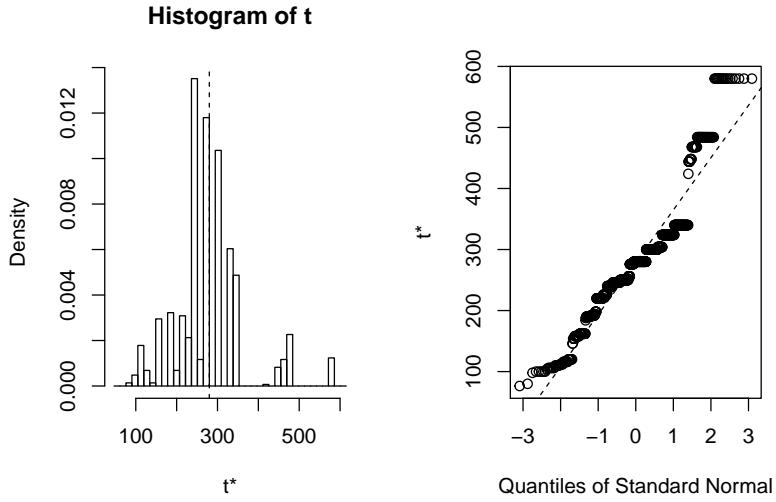


Figure 10.15: Distribusi bootstrap persentil 90

```
## Bootstrap Statistics :
##      original   bias   std. error
## t1*       280 -0.9156     85.66
```

Visualisasi distribusi bootstrap persentil 90 disajikan pada Gambar 10.15

Bentuk visualisasi distribusi bootstrap persentil ke-90 yang dihasilkan pada Gambar 10.15 terlihat sedikit memiliki kemencengan positif. Untuk menghitung interval kepercayaan 95% persentil ke-90 jalankan sintaks berikut:

```
boot.ci(boot.out=p90.boot, conf=0.95)
```

```
## Warning in boot.ci(boot.out = p90.boot, conf = 0.95):
## bootstrap variances needed for studentized intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = p90.boot, conf = 0.95)
##
## Intervals :
## Level      Normal          Basic
## 95%  (113.0, 448.8 )  ( 76.0, 448.0 )
##
## Level      Percentile        BCa
## 95%  (112, 484 )    (120, 580 )
## Calculations and Intervals on Original Scale
```

Berdasarkan keempat hasil yang diperoleh, metode BCa cukup baik digunakan sebab distribusi bootstrap yang dihasilkan tidak memenuhi asumsi ketiga metode lainnya.

10.9 Kegunaan Lain Dari Interval Kepercayaan

Selain digunakan untuk menghitung interval estimasi, interval kepercayaan dapat pula digunakan sebagai pendekripsi adanya *outlier*, kontrol kualitas, dan penentuan ukuran sampel yang akan digunakan pada suatu penelitian agar hasil yang diperoleh lebih presisi. Pembahasan juga akan disertai apakah normalitas pada distribusi data akan mempengaruhi performa dari ketiga jenis kegunaan interval kepercayaan.

10.9.1 Implikasi Non-Normalitas Pada Pendekripsi *Outlier*

Outlier merupakan pengamatan yang tampak berbeda karakteristiknya dibandingkan sebagian besar pengamatan yang ada. Pengamatan ini seringkali dihapus dari prosedur analisis yang mengharuskan distribusi suatu data mengikuti distribusi normal. Penghapusan observasi ini bisa jadi tidak baik dilakukan sebab bisa saja observasi tersebut valid. Observasi yang bersifat *outlier* bisa jadi merupakan observasi terpenting sebab bisa saja dapat memberikan gambaran penting pada suatu kondisi ekstrim atau hubungan kausatif yang penting. Penghapusan ini tidak perlu dilakukan selama prosedur pengukuran yang sejenis tersedia dan tidak mengharuskan suatu distribusi data mengikuti distribusi tertentu, meskipun terdapat kelebihan dan kekurangan yang perlu kita perhatikan.

Untuk menghapus observasi yang kita identifikasi sebagai *outlier*, aturan atau tes perlu dilakukan seperti tes yang diajukan oleh Beckman dan Cook (1983). Tes yang paling umum didasarkan pada interval-t, dan mengasumsikan data mengikuti distribusi normal. Biasanya Persamaan (10.13) untuk interval pediksi data yang mengikuti distribusi normal disederhanakan dengan mengabaikan nilai $\frac{s^2}{n}$. Nilai diluar interval prediksi tersebut selanjutnya dapat dinyatakan sebagai *outlier*. Uji lain yang dapat dilakukan adalah dengan visualisasi menggunakan box plot. Nilai diluar $Q1 - 1,5IQR$ atau $Q3 + 1,5IQR$ dinyatakan sebagai *outlier*. Namun pada Chapter ini tidak akan dibahas lebih lanjut sebab pada Chapter 7 telah dibahas mengenai deteksi *outlier* melalui visualisasi data.

10.9.2 Implikasi Non-Normalitas Pada Kontrol Kualitas

Presentasi visual interval kepercayaan yang digunakan secara luas dalam proses industri adalah kontrol chart. Sejumlah kecil produk disampel dari total kemungkinan pada titik waktu tertentu, dan rerata dihitung. Pengambilan sampel diulang pada interval reguler atau acak, tergantung pada desain, menghasilkan serangkaian cara sampel. Ini digunakan untuk membangun satu jenis kontrol chart, xbar chart. Chart ini secara visual mendekripsi ketika rata-rata sampel masa depan menjadi berbeda dari yang digunakan untuk membangun grafik. Keputusan perbedaan didasarkan pada melebihi interval kepercayaan parametrik di sekitar rata-rata yang telah dijelaskan di bagian lain Chapter ini.

Misalkan laboratorium kimia mengukur larutan standar yang sama beberapa kali selama sehari untuk menentukan apakah peralatan dan operator menghasilkan hasil yang konsisten. Untuk serangkaian pengukuran n pada interval waktu m , ukuran sampel total $N = N \cdot M$. Perkiraan konsentrasi terbaik untuk standar itu adalah rata-rata keseluruhan yang dihitung menggunakan Persamaan (10.23).

$$\bar{X} = \sum_{i=1}^N \frac{X_i}{N} \quad (10.23)$$

\bar{X} diplot sebagai garis tengah grafik. Interval kepercayaan pada rata-rata tersebut dijelaskan oleh Persamaan (10.8), menggunakan ukuran sampel n yang tersedia untuk menghitung setiap nilai rata-rata individu. Batas interval tersebut juga diplotkan sebagai garis paralel pada chart kontrol kualitas. Nilai rata-rata yang akan berada diluar batas plot rata-rata ini hanya sebesar $\alpha \cdot 100\%$ dari waktu jika rata-rata berdistribusi normal. Observasi yang berada di luar batas lebih sering daripada ini diambil untuk menunjukkan bahwa sesuatu dalam proses telah berubah.

Jika n besar (katakanlah 30 atau lebih), Teorema Limit Pusat menyatakan bahwa rata-rata akan terdistribusi secara normal meskipun data yang mendasarinya mungkin tidak. Namun jika n jauh lebih kecil, seperti yang sering terjadi, berarti mungkin tidak mengikuti pola ini. Khususnya, untuk data yang memiliki kemencengan (data dengan *outlier* hanya pada satu sisi), distribusi di sekitar rata-rata mungkin masih memiliki kemencengan. Hasilnya adalah nilai besar untuk simpangan baku, dan pita kepercayaan yang lebar. Oleh karena itu, chart akan memiliki kekuatan yang lebih rendah untuk mendeteksi observasi yang mulai menjauh dari nilai rata-rata yang diharapkan daripada jika data tidak memiliki kemencengan.

Chart kontrol juga diproduksi untuk menggambarkan varians proses. Chart kontrol juga menggunakan nilai range (R chart) atau simpangan baku (S chart). Kedua grafik bahkan lebih sensitif terhadap perubahan data dari kondisi normal dibandingkan \bar{X} chart. Keduanya akan mengalami kesulitan dalam mendeteksi perubahan varian ketika data yang mendasarinya tidak normal, dan ukuran sampel n untuk setiap rata-rata kecil.

10.9.3 Implikasi Non-Normalitas Terhadap Desain Sampling

Persamaan t-interval juga digunakan untuk menentukan jumlah sampel yang diperlukan untuk memperkirakan rata-rata dengan tingkat presisi yang ditentukan. Namun, persamaan tersebut membutuhkan data untuk kira-kira mengikuti distribusi normal. Persamaan tersebut harus mempertimbangkan *power* serta lebar interval. Artinya kita harus memutuskan apakah mean adalah karakteristik yang paling tepat untuk mengukur data yang memiliki kemencengan.

Untuk memperkirakan ukuran sampel telah cukup untuk menentukan interval estimasi dengan lebar spesifik dapat menggunakan Persamaan (10.24).

$$n = \left(\frac{t_{\frac{\alpha}{2}, n-1} \cdot s}{\Delta} \right)^2 \quad (10.24)$$

dimana s merupakan simpangan baku sampel dan Δ merupakan setengah lebar interval yang diinginkan. Seperti yang telah dibahas di atas, untuk ukuran sampel kurang dari 30 hingga 50 dan bahkan lebih besar dengan data yang sangat menceng, perhitungan ini mungkin memiliki kesalahan besar. Perkiraan s akan tidak akurat, dan akan sangat sangat meningkat nilainya karena kemencenga dan/atau *outlier* apapun. Karenanya, estimasi n yang dihasilkan akan besar. Sebagai contoh, Hakanson (1984) memperkirakan jumlah sampel yang diperlukan untuk memberikan lebar interval yang masuk akal untuk karakteristik sedimen sungai dan danau, termasuk kimia sedimen. Berdasarkan koefisien variasi yang dilaporkan dalam artikel, data untuk sedimen sungai cukup menceng, seperti yang mungkin diharapkan. Ukuran sampel yang diperlukan untuk sungai dihitung pada 200 dan lebih tinggi.

Sebelum menggunakan persamaan sederhana seperti itu, data yang menceng harus ditransformasi terlebih dahulu sehingga lebih simetris, jika bukan berdistribusi normal. Misalnya, logaritma akan secara drastis menurunkan taksiran ukuran sampel untuk data miring, setara dengan Persamaan (10.15). Ukuran sampel akan dihasilkan yang memungkinkan median (rata-rata geometris) diperkirakan dalam faktor toleransi multiplikasi sama dengan $\pm 2\Delta$ dalam satuan log.

Masalah kedua dengan Persamaan (10.24) untuk memperkirakan ukuran sampel, bahkan ketika data mengikuti distribusi normal, ditunjukkan oleh Kupper dan Hafner (1989). Mereka menunjukkan Persamaan (10.24) meremehkan ukuran sampel sebenarnya yang diperlukan untuk tingkat presisi tertentu, bahkan untuk perkiraan $n \leq 40$. Hal ini disebabkan karena Persamaan (10.24) tidak mengakui bahwa simpangan baku s hanya estimasi dari nilai sebenarnya σ . Mereka menyarankan menambahkan probabilitas toleransi ke Persamaan (10.24), mirip dengan *statement of power*. Maka perkiraan lebar interval setidaknya sekecil lebar interval yang diinginkan untuk beberapa persentase tertentu (katakanlah 90 atau 95%) dari waktu tersebut. Misalnya, ketika n akan sama dengan 40 berdasarkan Persamaan (10.24), lebar interval yang dihasilkan akan lebih kecil dari lebar yang diinginkan 2Δ hanya sekitar 42% dari waktu. Ukuran sampel seharusnya menjadi 53 untuk memastikan lebar interval berada dalam kisaran toleransi 90% dari waktu.

Ukuran sampel yang diperlukan untuk estimasi interval median atau untuk melakukan tes nonparametrik dari Chapter selanjutnya dapat diturunkan tanpa asumsi normalitas yang diperlukan di atas untuk interval-t. Noether (1987) menjelaskan estimasi ukuran sampel yang lebih kuat ini, yang memasukkan pertimbangan *power* sehingga lebih valid daripada Persamaan (10.24). Namun, baik estimasi normalitas distribusi data atau nonparametric mempertimbangkan efek penting dan sering diamati dari musiman atau tren, dan karenanya mungkin tidak pernah memberikan perkiraan yang cukup akurat untuk menjadi sesuatu yang lebih dari sekadar panduan kasar.

10.10 Referensi

1. Bachman, L. J. 1984. **Field and laboratory analyses of water from the Columbia Aquifer in Eastern Maryland.** Ground Water 22. 460-467.
2. Damanhuri, E. 2011. **Statitika Lingkunga.** Penerbit ITB.
3. Deryto, T. tanpa tahun. **Bootstrap in R.** Datacamp. <https://www.datacamp.com/community/tutorials/bootstrap-r>
4. Efron, B. 1979. **Bootstrap Methods: Another Look at The Jackknife.** The Annals of Statistics. Vol:7(1). 1-26.
5. Helsel, D.R., Hirsch, R.M. 2002. **statistical Methods in Water Resources.** USGS.
6. Kupper, L. L., and K. B. Hafner. 1989. **How appropriate are popular sample size formulas?.** American Statistician 43. 101-105.
7. Noether, G. E. 1987. **Sample size determination for some common nonparametric tests.** Journal American Statistical Assoc.. 82, 645-647.
8. Ofungwu, J. 2014. **Statistical Applications For Environmental Analysis and Risk Assessment.** John Wiley & Sons, Inc.
9. Ting Wei, Songbai Song. 2019. **Confidence interval estimation for precipitation quantiles based on principle of maximum entropy.** Entropy. 21.