PENGANTAR METODE STATISTIKA DALAM METEOROLOGI DAN KLIMATOLOGI

Dr. Nurtiti Sunusi, S.Si, M.Si

PENGANTAR METODE STATISTIKA

DALAM METEOROLOGI DAN KLIMATOLOGI

Jilid 1

Dr. Nurtiti Sunusi, S.Si, M.Si.

PT GLOBAL EKSEKUTIF TEKNOLOGI

PENGANTAR METODE STATISTIKA

DALAM METEOROLOGI DAN KLIMATOLOGI

Jilid 1

Penulis :
Nurtiti Sunusi

ISBN : 978-623-198-229-2

Editor : Dr. Neila Sulung, S.Pd., Ns., M.Kes.
Penyunting : Mila Sari, S.ST, M.Si
Desain Sampul dan Tata Letak : Atyka Trianisa, S.Pd

Penerbit : PT GLOBAL EKSEKUTIF TEKNOLOGI
Anggota IKAPI No. 033/SBA/2022

Redaksi :
Jl. Pasir Sebelah No. 30 RT 002 RW 001
Kelurahan Pasie Nan Tigo Kecamatan Koto Tangah
Padang Sumatera Barat
Website : www.globaleksekutifteknologi.co.id
Email : globaleksekutifteknologi@gmail.com

Cetakan pertama, April 2023

KATA PENGANTAR

Alhamdulilahi Rabbil ‘Alamiin. Puji syukur ke hadirat Tuhan Yang Maha Kuasa, yang telah memberikan Rahmat-Nya sehingga buku Pengantar Metode Statistika dalam Ilmu Metereologi dan Klimatologi beserta soal dan pembahasannya dapat diselesaikan dengan sebaik-baiknya.

Buku ini memberikan informasi secara lengkap mengenai materi dasar yang berkaitan dengan konsep dasar statistika dalam ilmu atmosfer khususnya meteOrologi dan klimatologi. Buku ini juga dilengkapi dengan implementasi program R dalam analisis data eksplorasi yang disusun secara sederhana sehingga dapat dengan mudah dipahami oleh pengguna.

Penyusun meyakini bahwa dalam pembuatan buku Pengantar Metode Statistika dalam Ilmu Atmosfer ini masih jauh dari sempurna. Oleh karena itu penyusun mengharapkan kritik dan saran yang membangun guna penyempurnaan buku ini dimasa yang akan datang.

Akhir kata, penyusun mengucapkan banyak terima kasih kepada semua pihak yang telah membantu baik secara langsung maupun tidak langsung.

Makassar, April 2023

Penyusun

DAFTAR ISI

DAFTAR GAMBAR

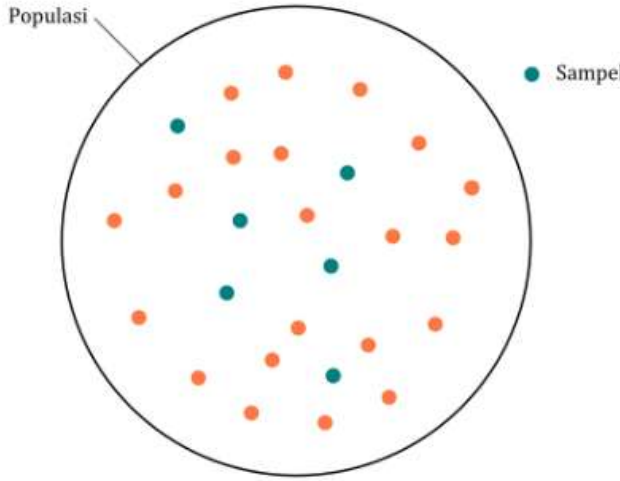
DAFTAR TABEL

BAB 1
PENGANTAR

1.1 Apa itu Statistik?

Buku ini berkaitan dengan penggunaan metode statistik dalam ilmu atmosfer, khususnya dalam berbagai spesialisasi dalam meteorologi dan klimatologi. Statistik adalah sebuah kumpulan data, angka, atau informasi, sedangkan statistika adalah ilmu yang mempelajari bagaimana data atau angka tersebut dikumpulkan, diolah, dan dianalisis untuk menghasilkan sebuah informasi yang bisa digunakan dalam pengambilan keputusan. Dalam kehidupan di era Big Data ini, statistik sangat berperan signifikan dalam membantu memudahkan kehidupan sehari-hari manusia. Contohnya, peranan statistik dapat dilihat dalam kegiatan ilmiah, kegiatan proses belajar mengajar, dan dalam kegiatan ilmu pengetahuan. Tidak cuma itu, penelitian dan statistik adalah dua hal yang tidak bisa dipisahkan. Walaupun terdapat jenis penelitian yang tidak membutuhkan peran ilmu statistika, namun untuk bisa menghasilkan sebuah kesimpulan yang dapat digeneralisasikan ke sebuah populasi yang lebih luas, ilmu statistika sangat diperlukan. Dalam pengaplikasiannya, statistik memang mempunyai kaitan dan manfaat langsung dengan banyak hal dalam kehidupan manusia. Manfaat atau kegunaan statistik tentu saja tidak terbatas pada kegiatan ilmu pengetahuan seperti penelitian saja. Statistik juga dimanfaatkan secara luas, baik dalam bidang ilmu alam maupun dalam bidang bisnis, industri, dan ekonomi. Pada artikel berikut ini, kita akan mengetahui peranan statistik di berbagai bidang.

Data sendiri dapat didefinisikan sebagai kelompok informasi yang mewakili atribut kualitatif atau kuantitatif suatu variabel atau serangkaian variabel. Dalam istilah awam, data dalam statistik dapat berupa serangkaian informasi yang menggambarkan entitas tertentu. Contoh data dapat berupa usia siswa di kelas yang diberikan. Ketika Anda mengumpulkan usia itu, itu menjadi data Anda.



Gambar 0.1. Gambaran umum populasi dan sampel, warna hijau menandakan sampel yang diambil dari populasi.

Satu set dalam statistik disebut sebagai populasi. Meskipun istilah ini biasanya digunakan untuk merujuk pada jumlah orang di tempat tertentu, dalam statistik, populasi merujuk pada seluruh rangkaian tempat Anda mengumpulkan data. Statistik sendiri secara sadar atau tidaknya banyak digunakan dalam kehidupan sehari- hari meski dalam bentuk sederhana sekalipun.

Penggunaan teknis analisis statistika ini terbukti mampu memberi bantuan yang penting dalam banyak kegiatan berskala besar seperti pada perusahaan maupun pemerintahan. Contohnya dalam kegiatan riset atau penelitian seperti bidang ekonomi, akademik, pengambilan keputusan, metode statistik mampu membantu memberikan gambaran sebuah permasalahan serta prediksi juga rekomendasi terhadap kemungkinan-kemungkinan yang ada. Dengan adanya buku ini, diharapkan akan membantu memberikan apresiasi setidaknya dalam kaitannya dengan ilmu atmosfer khususnya klimatologi dan metereologi.

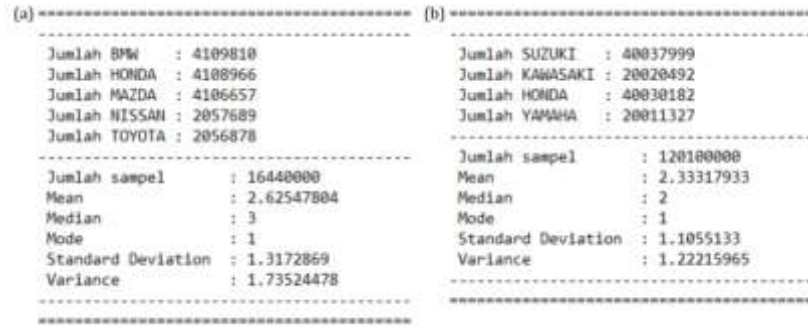
Pada dasarnya, statistik berkaitan dengan ketidakpastian. Mengevaluasi dan mengukur ketidakpastian, serta membuat kesimpulan dan perkiraan dalam menghadapi ketidakpastian, semuanya adalah bagian dari statistik. Sehingga, tidak mengherankan jika statistik memiliki banyak peran dalam ilmu atmosfer karena ketidakpastian perilaku atmosfer yang membuat atmosfer menjadi menarik. Misalnya, banyak orang yang tertarik dengan perkiraan cuaca, dimana justru karena ketidakpastian yang melekat pada masalah tersebut yang membuatnya menarik. Jika memungkinkan untuk membuat perkiraan yang sempurna bahkan satu hari ke depan (yaitu, jika tidak ada ketidakpastian yang terlibat), praktik meteorologi akan sangat membosankan.

1.2 Statistik Deskriptif dan Inferensial

Statistika adalah metode di mana kita mengumpulkan, menganalisis, meringkas, dan mengambil kesimpulan dari data eksperimental atau nyata. Untuk menyelesaikan proses tersebut, statistika dipecah menjadi 2 jenis: Deskriptif dan Inferensial. Apa perbedaan dari kedua jenis statistika tersebut?

Statistika Deskriptif adalah sebuah cara untuk mengatur, merepresentasikan, dan mendeskripsikan kumpulan data menggunakan tabel, grafik, dan banyak parameter numerik lainnya. Berikut beberapa istilah:

Mean	:	Nilai rata-rata dari data yang bertipe numerik/angka.
Median	:	Nilai tengah yang membagi data terurut menjadi 2 bagian sama besar.
Mode	:	Nilai yang paling sering muncul dari data.
Standard Deviation:	:	Rata-rata kuadrat dari selisih/jarak setiap observasi dengan nilai mean.
Correlation	:	Nilai yang menggambarkan keeratan hubungan antara dua variabel numerik. Belum tentu menggambarkan hubungan sebab akibat.
Variance	:	Nilai yang menggambarkan seberapa bervariasi/beragamnya suatu data bertipe numerik/angka.



Gambar 0.2. Statistika Deskriptif dari data simulai kendaraan motor dengan 5 brand (a) dan kendaraan mobil dengan 4 brand (b).

Implementasi dalam program R

Sintaks:

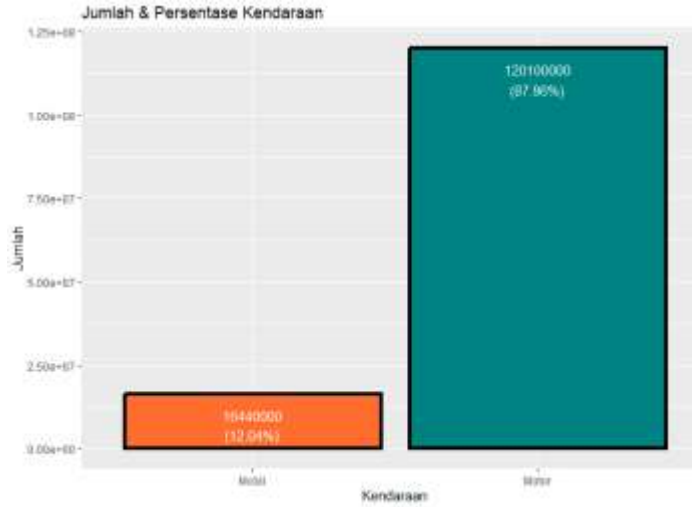
```
> rm(list = ls())     ## Untuk memastikan data tidak tertumpuk ##
> set.seed(8)         ## Setting Random                         ##
>
> ## Function Untuk menghitung jumlah Type ##
> sum_of_type <- function(y){
+   y     <- as.array(y)
+   u_y   <- unique(y)
+   new_y <- c()
+   sum_y <- c()
+
+   message('========================================')
+   message('----------------------------------------')
+   for(i in 1:length(u_y)){
+     sum_y[i] <- 0
+     for (j in 1:length(y)) {
+       if(u_y[i] == y[j]){
+         sum_y[i] <- sum_y[i] + 1
+         new_y[j] <- i
+       }
+     }
+     message(' Jumlah ',u_y[i],' : ', sum_y[i])
+   }
```

```
+
+   y_n       <- length(new_y)
+   y_mean    <- mean(new_y)
+   y_median  <- median(new_y)
+         y_modus          <-   new_y[which.max(tabulate(match(new_y,
unique(new_y))))]
+   y_sd      <- sd(new_y)
+   y_var     <- var(new_y)
+
+   ## Output ##
+   message('----------------------------------------')
+   message(' Jumlah sampel       : ', y_n)
+   message(' Mean                : ', round(y_mean,8))
+   message(' Median              : ', y_median)
+   message(' Mode                : ', y_modus)
+   message(' Standard Deviation  : ', round(y_sd,8))
+   message(' Variance            : ', round(y_var,8))
+   message('----------------------------------------')
+   message('========================================')
+
+   output        <- NULL
+   output$N      <- y_n
+   output$mean   <- y_mean
+   output$median <- y_median
+   output$modus  <- y_modus
+   output$sd     <- y_sd
+   output$var    <- y_var
+   output$uniq   <- u_y
+   output$data   <- data.frame(Uniq = u_y, Jumlah = sum_y)
+   output$new_y  <- new_y
+
+   return(output)
+ }
> ## data simulasi mobil sebanyak 16.44 jt##
> dt_mobil   <- c()
> rd_mobil   <- round(runif(16440000 , min = 1, max = 5),0)
> ## membuat 5 brand mobil ##
> for (i in 1:length(rd_mobil)) {
+   if(rd_mobil[i] == 1){
+     dt_mobil[i] <- 'TOYOTA'
+   } else if(rd_mobil[i] == 2){
+     dt_mobil[i] <- 'HONDA '
+   } else if(rd_mobil[i] == 3){
+     dt_mobil[i] <- 'BMW   '
```

```
+   }  else if(rd_mobil[i] == 4){
+     dt_mobil[i] <- 'MAZDA '
+   } else {
+     dt_mobil[i] <- 'NISSAN'
+   }
+ }
> dt_summary_mobil  <- sum_of_type(dt_mobil)
========================================
----------------------------------------
 Jumlah BMW    : 4109810
 Jumlah HONDA  : 4108966
 Jumlah MAZDA  : 4106657
 Jumlah NISSAN : 2057689
 Jumlah TOYOTA : 2056878
----------------------------------------
 Jumlah sampel       : 16440000
 Mean                : 2.62547804
 Median              : 3
 Mode                : 1
 Standard Deviation  : 1.3172869
 Variance            : 1.73524478
----------------------------------------
========================================
>
> ## data simulasi motor sebanyak 120.10 jt##
> dt_motor   <- c()
> rd_motor   <- round(runif(120100000 , min = 1, max = 4),0)
> ## membuat 4 brand motor ##
> for (i in 1:length(rd_motor)) {
+   if(rd_motor[i] == 1){
+     dt_motor[i] <- 'YAMAHA  '
+   } else if(rd_motor[i] == 2){
+     dt_motor[i] <- 'HONDA   '
+   } else if(rd_motor[i] == 3){
+     dt_motor[i] <- 'SUZUKI  '
+   } else {
+     dt_motor[i] <- 'KAWASAKI'
+   }
+ }
> dt_summary_motor  <- sum_of_type(dt_motor)
========================================
----------------------------------------
 Jumlah SUZUKI   : 40037999
 Jumlah KAWASAKI : 20020492
```

```
 Jumlah HONDA    : 40030182
 Jumlah YAMAHA   : 20011327
----------------------------------------
 Jumlah sampel       : 120100000
 Mean                : 2.33317933
 Median              : 2
 Mode                : 1
 Standard Deviation  : 1.1055133
 Variance            : 1.22215965
----------------------------------------
========================================
```

Statistika Deskriptif sangat penting bagi Data Scientists untuk melihat sebuah pattern dari data. Tetapi, jenis statistika ini hanya bisa digunakan untuk sampel data yang sedang dipelajari, tidak bisa digunakan untuk melakukan generalisasi atau mengambil kesimpulan tentang populasi atau kelompok lainnya. Contoh dari Statistika Deskriptif adalah koleksi data di dalam kota yang menggunakan internet atau memiliki mobil. Misalkan, menurut BPS (Badan Pusat Statistik), jumlah kendaraan mobil penumpang di Indonesia pada tahun 2018 mencapai 16,44 juta. Jumlah kendaraan sepeda motor mencapai 120,10 juta. Data deskriptif ini bisa dibuat dalam bentuk tabel maupun grafik.



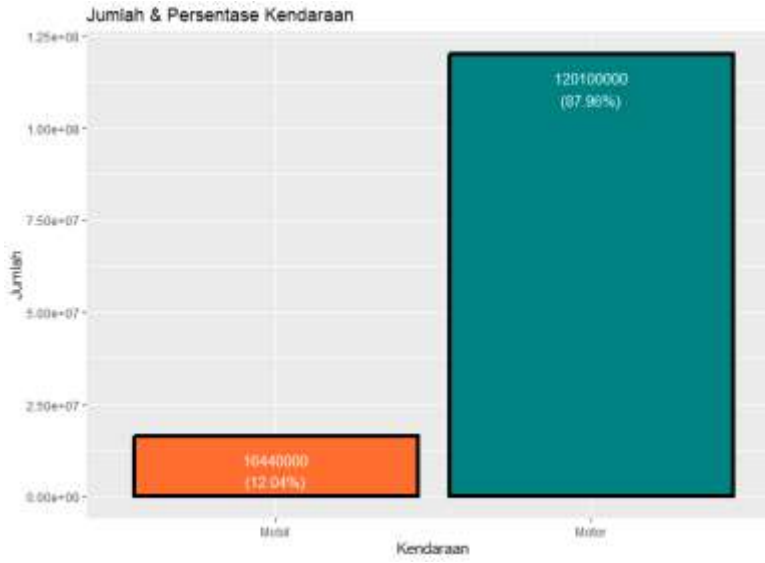
Gambar 0.3. Grafik jumlah kendaraan mobil dan motor pada tahun 2018 menurut BPS.

Implementasi dalam program R

```
Sintaks:
> rm(list = ls())     ## Untuk memastikan data tidak tertumpuk ##
> set.seed(8)         ## Setting Random                         ##
> library('ggplot2') ## package untuk plot ##
> ## data simulasi mobil sebanyak 16.44 jt##
> rd_mobil    <- round(runif(16440000 , min = 1, max = 5),0)
> ## data simulasi motor sebanyak 120.10 jt##
> rd_motor    <- round(runif(120100000 , min = 1, max = 4),0)
> ## Motor dan Mobil ##
> x_tmp             <- c('Mobil','Motor')
> y_tmp             <- c(length(rd_mobil),length(rd_motor))
> df_mm             <- data.frame(x = x_tmp, y=y_tmp)
> df_mm$fraction    <- df_mm$y / sum(df_mm$y)
> df_mm$ymax        <- cumsum(df_mm$fraction)
> df_mm$ymin        <- c(0, head(df_mm$ymax, n=-1))
> df_mm$percent      <- paste(round(df_mm$fraction*100,2), '%', sep =
'')
> df_mm$Ket          <- paste(df_mm$x,' (', df_mm$percent,')', sep =
'')
```

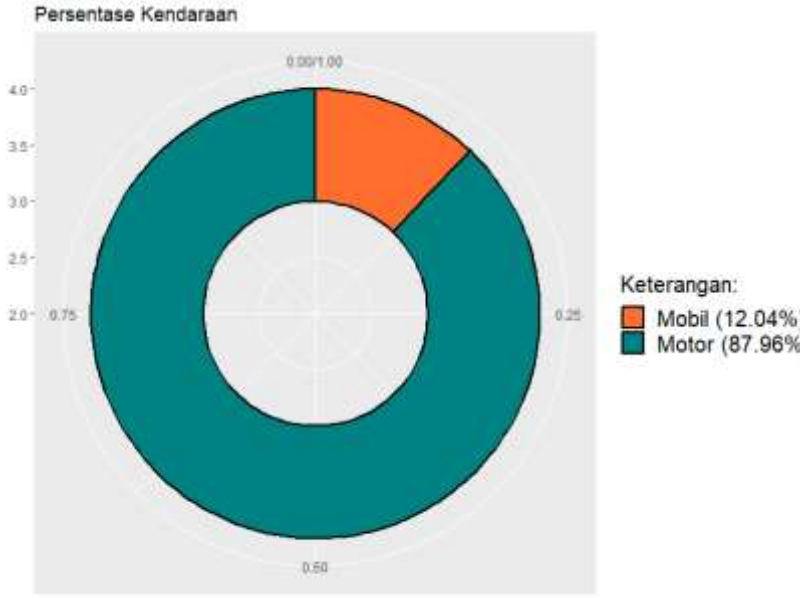
```
> ## Plot Motor dan Mobil ##
> ggplot(df_mm, aes(x=x, y=y, fill = x)) +
+   geom_bar(stat="identity", size=1.5, color='#000000') +
+   geom_text(aes(label=paste(y,'\n(',percent,')', sep = '')),
+             vjust=1.5, angle=0 , size=4, color = '#FFFFFF') +
+
+   scale_fill_manual(values = c('#FF7742',
+                                '#008080'))+
+
+   scale_color_manual(values = c('#FF7742',
+                                 '#008080'))+
+
+   labs(title = 'Jumlah & Persentase Kendaraan',
+        y='Jumlah', x='Kendaraan') +
+
+   theme(legend.position = "none")
>
```

Output:



Pada gambar 1.3 memberikan informasi kepada kita bahwa jumlah kendaraan bermotor lebih banyak dibandingkan jumlah kendaraan mobil dengan persentase

pengguna motor sebesar 87.96% sehingga dapat disimpulkan bahwa rata rata penduduk indonesia mempunyai motor. Akan tetapi penyajian grafik tidak hanya sebatas bar plot saja, terkadan jika data yang hanya memiliki dua kategori misal motor dan mobil ditampilkan dalam bentuk donut chart seperti pada gambar berikut.



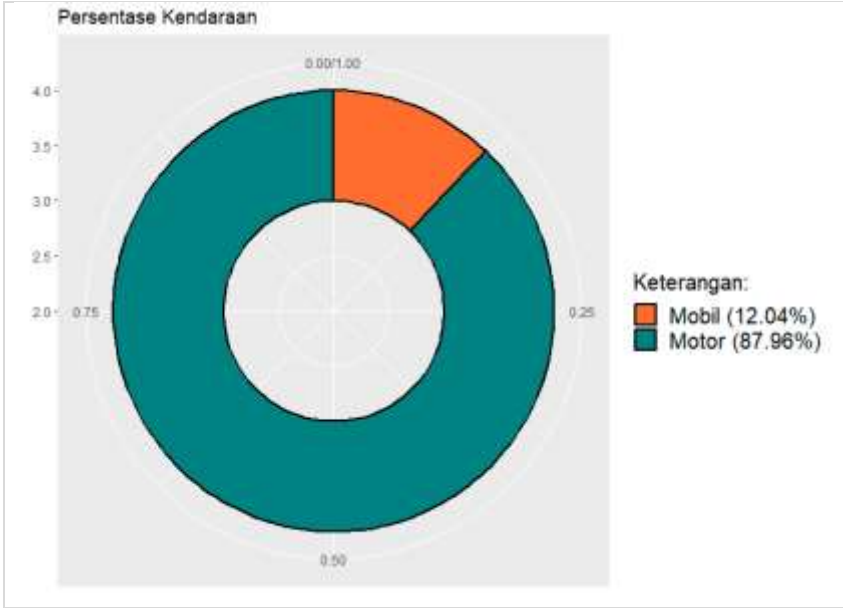
Gambar 0.4. Donut chart kendaraan mobil dan motor pada tahun 2018 menurut BPS.

Implementasi dalam program R

Sintaks:

```
> rm(list = ls())    ## Untuk memastikan data tidak tertumpuk ##
> set.seed(8)        ## Setting Random                        ##
> library('ggplot2') ## package untuk plot ##
> ## data simulasi mobil sebanyak 16.44 jt##
> rd_mobil   <- round(runif(16440000 , min = 1, max = 5),0)
> ## data simulasi motor sebanyak 120.10 jt##
> rd_motor   <- round(runif(120100000 , min = 1, max = 4),0)
> ## Motor dan Mobil ##
> x_tmp             <- c('Mobil','Motor')
> y_tmp             <- c(length(rd_mobil),length(rd_motor))
> df_mm             <- data.frame(x = x_tmp, y=y_tmp)
> df_mm$fraction    <- df_mm$y / sum(df_mm$y)
> df_mm$ymax        <- cumsum(df_mm$fraction)
> df_mm$ymin        <- c(0, head(df_mm$ymax, n=-1))
> df_mm$percent      <- paste(round(df_mm$fraction*100,2), '%', sep =
'')
> df_mm$Ket          <- paste(df_mm$x,' (', df_mm$percent,')', sep =
'')
> ## Plot Motor dan Mobil ##
> ggplot(df_mm, aes(ymax=ymax, ymin=ymin,
+                   xmax=4, xmin=3, fill=Ket)) +
+   geom_rect(color='#000000', size=1) +
+   coord_polar(theta="y") + xlim(c(2, 4)) +
+
+   scale_fill_manual(values = c('#FF7742',
+                                '#008080'))+
+
+   scale_color_manual(values = c('#FF7742',
+                                 '#008080'))+
+
+   theme(legend.text = element_text(size = 15),
+         legend.title = element_text(size = 15)) +
+
+   labs(title = 'Persentase Kendaraan') +
+
+   guides(fill = guide_legend(title = "Keterangan:"))
>
```

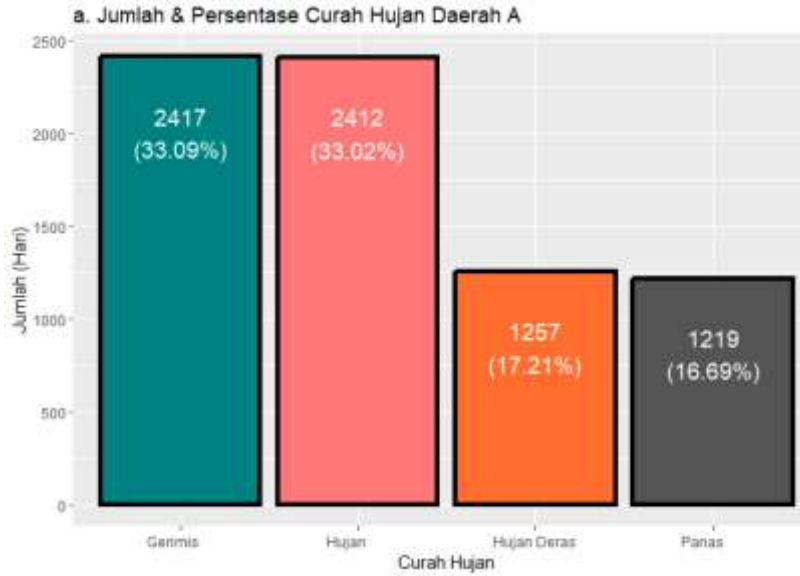
Output:



Statistik deskriptif terkait dengan organisasi dan ringkasan data. Ilmu atmosfer dipenuhi dengan data. Di seluruh dunia, pengamatan permukaan dan udara atas secara rutin dilakukan di ribuan lokasi untuk mendukung kegiatan peramalan cuaca. Proses dilengkapi dengan data pesawat, radar, profiler, dan satelit. Pengamatan atmosfer khusus untuk tujuan penelitian kurang tersebar luas, namun seringkali melibatkan pengambilan sampel yang sangat padat dalam ruang dan waktu. Selain itu, model atmosfer yang terdiri dari integrasi numerik dari persamaan yang menggambarkan dinamika atmosfer menghasilkan luaran numerik yang lebih banyak untuk tujuan operasional dan penelitian.

Sebagai konsekuensi dari aktivitas tersebut, kita sering dihadapkan pada kumpulan angka yang sangat besar yang diharapkan berisi informasi tentang fenomena alam yang

menarik. Hal ini bisa menjadi tugas yang tidak sederhana hanya untuk membuat pengertian awal dari kumpulan data tersebut. Biasanya perlu dilakukan pengaturan dan pemilihan data mentah serta mengimplementasikan representasi ringkasan yang sesuai (dalam machine learning dikenal dengan proses data preparasi). Ketika nilai data individual terlalu banyak untuk dipahami secara individual, ringkasan yang tetap menggambarkan aspek penting dari variasinya model statistic dapat sangat berguna dalam memahami data. Perlu ditekankan bahwa tujuan analisis data deskriptif bukan untuk bermain-main dengan angka. Sebaliknya, analisis ini dilakukan karena diketahui, diduga, atau diharapkan bahwa data tersebut mengandung informasi tentang fenomena alam yang menarik, yang dapat diungkap atau lebih dipahami melalui analisis statistik misal ketika ingin melihat perbedaan curah hujan dari dua daerah. Dengan mengguanakan data simulasi data curah hujan daerah A 20 tahun terakhir dapat dilihat pada gambar berikut:



Gambar 0.3. Barplot Jumlah curah hujan daerah A 20 tahun terakhir.

Implementasi dalam program R

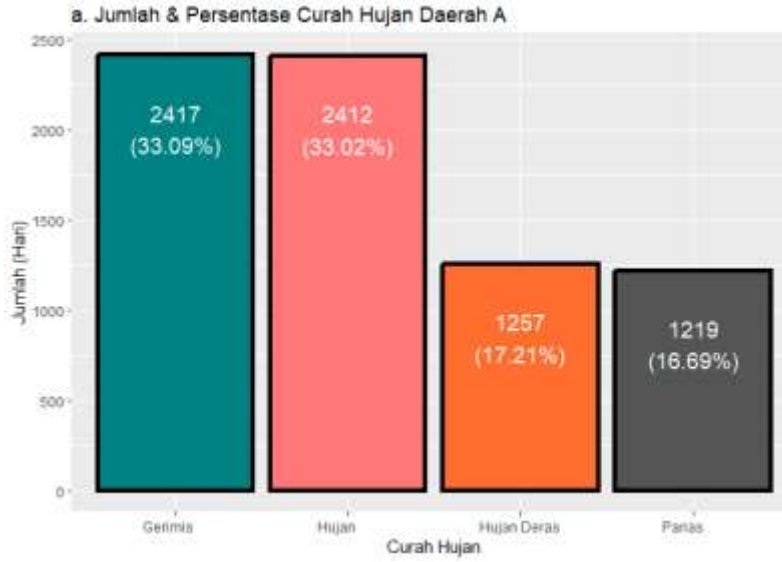
```
Sintaks:
> rm(list = ls())    ## Untuk memastikan data tidak tertumpuk ##
> set.seed(8)        ## Seting Random                          ##
> library('ggplot2') ## package untuk plot ##
> ## Function Untuk menghitung jumlah kejadian ##
> sum_of_event <- function(y){
+   y     <- as.array(y)
+   u_y   <- unique(y)
+   new_y <- c()
+   sum_y <- c()
+
+   message('========================================')
+   message('----------------------------------------')
+   for(i in 1:length(u_y)){
+     sum_y[i] <- 0
+     for (j in 1:length(y)) {
+       if(u_y[i] == y[j]){
```

```
+         sum_y[i] <- sum_y[i]+1
+         new_y[j] <- i
+       }
+     }
+     message(' Jumlah ',u_y[i],' : ', sum_y[i])
+   }
+
+   y_n       <- length(new_y)
+   y_mean    <- mean(new_y)
+   y_median  <- median(new_y)
+          y_modus          <-    new_y[which.max(tabulate(match(new_y,
unique(new_y))))]
+   y_sd      <- sd(new_y)
+   y_var     <- var(new_y)
+
+   ## Output ##
+   message('----------------------------------------')
+   message(' Jumlah sampel       : ', y_n)
+   message(' Mean                : ', round(y_mean,8))
+   message(' Median              : ', y_median)
+   message(' Mode                : ', y_modus)
+   message(' Standard Deviation  : ', round(y_sd,8))
+   message(' Variance            : ', round(y_var,8))
+   message('----------------------------------------')
+   message('========================================')
+
+   output        <- NULL
+   output$N      <- y_n
+   output$mean   <- y_mean
+   output$median <- y_median
+   output$modus  <- y_modus
+   output$sd     <- y_sd
+   output$var    <- y_var
+   output$uniq   <- u_y
+   output$data   <- data.frame(Uniq = u_y, Jumlah = sum_y)
+   output$new_y  <- new_y
+
+   return(output)
+ }
> ## data simulasi cura hujan daerah A ##
> dt_daerah_a   <- c()
> rd_daerah_a   <- round(runif(7305 , min = 1, max = 4),0)
> for (i in 1:length(rd_daerah_a)) {
+   if(rd_daerah_a[i] == 1){
```

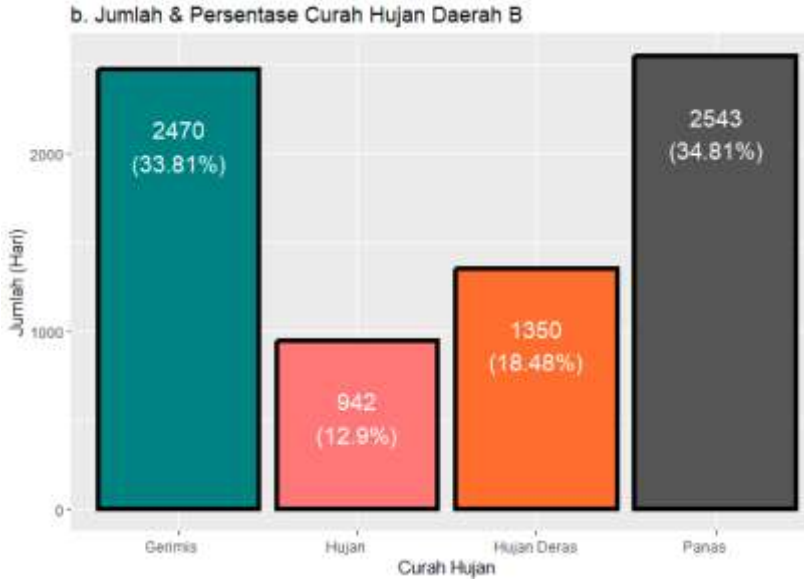
```
+     dt_daerah_a[i] <- 'Panas      '
+   } else if(rd_daerah_a[i] == 2){
+     dt_daerah_a[i] <- 'Gerimis    '
+   } else if(rd_daerah_a[i] == 3){
+     dt_daerah_a[i] <- 'Hujan      '
+   } else {
+     dt_daerah_a[i] <- 'Hujan Deras'
+   }
+ }
> summary_daerah_a     <- sum_of_event(dt_daerah_a)
==========================================
------------------------------------------
 Jumlah Gerimis     : 2417
 Jumlah Hujan       : 2412
 Jumlah Hujan Deras : 1257
 Jumlah Panas       : 1219
------------------------------------------
 Jumlah sampel       : 7305
 Mean                : 2.17494867
 Median              : 2
 Mode                : 1
 Standard Deviation  : 1.06769857
 Variance            : 1.13998023
------------------------------------------
==========================================
> x_tmp                <- summary_daerah_a$data$Uniq
> y_tmp                <- summary_daerah_a$data$Jumlah
> df_daerah_a          <- data.frame(x = x_tmp, y=y_tmp)
> df_daerah_a$fraction <- df_daerah_a$y / sum(df_daerah_a$y)
> df_daerah_a$ymax     <- cumsum(df_daerah_a$fraction)
> df_daerah_a$ymin     <- c(0, head(df_daerah_a$ymax, n=-1))
>  df_daerah_a$percent    <-  paste(round(df_daerah_a$fraction*100,2),
'%', sep = '')
>   df_daerah_a$Ket                  <-   paste(df_daerah_a$x,'   (',
df_daerah_a$percent,')', sep = '')
> ## Daerah A ##
> ggplot(df_daerah_a, aes(x=x, y=y, fill=x)) +
+   geom_bar(stat="identity", size=1.5, color='#000000')+
+
+   geom_text(aes(label=paste(y,'\n(',percent,')', sep = '')),
+             vjust=2, angle=0 , size=5, color='#FFFFFF') +
+
+   scale_fill_manual(values = c('#008080',
+                                '#FF8080',
```

```
+                                          '#FF7742',
+                                          '#555555'))+
+
+   scale_color_manual(values = c('#008080',
+                                           '#FF8080',
+                                           '#FF7742',
+                                           '#555555'))+
+
+   theme(legend.position = "none") +
+   labs(title = 'a. Jumlah & Persentase Curah Hujan Daerah A',
+        y='Jumlah (Hari)',
+        x='Curah Hujan')
>
```

Output:



Pada gambar 1.5 dapat dianalisis bahwa 20 tahun terakhir pada daerah A sering terjadi hujan, hanya 16.69% dari 20 tahun tidak terjadi hujan, dengan kata lain peluang terjadinya panas pada daerah A hanyalah 0.1669. Kemudian kita beralih ke data simulasi curah hujan daerah B yang dapat dilihat pada gambar berikut.



Gambar 0.4. Barplot Jumlah curah hujan daerah B 20 tahun terakhir.

Implementasi dalam program R

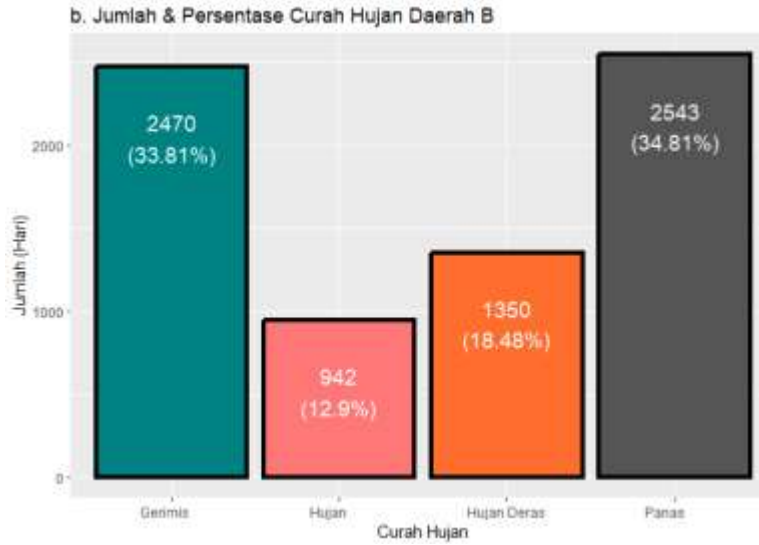
```
Sintaks:
> rm(list = ls())     ## Untuk memastikan data tidak tertumpuk ##
> set.seed(8)         ## Seting Random                          ##
> library('ggplot2') ## package untuk plot ##
> ## Function Untuk menghitung jumlah kejadian ##
> sum_of_event <- function(y){
+   y     <- as.array(y)
+   u_y   <- unique(y)
+   new_y <- c()
+   sum_y <- c()
+
+   message('========================================')
+   message('----------------------------------------')
+   for(i in 1:length(u_y)){
+     sum_y[i] <- 0
+     for (j in 1:length(y)) {
+       if(u_y[i] == y[j]){
```

```
+         sum_y[i] <- sum_y[i]+1
+         new_y[j] <- i
+       }
+     }
+     message(' Jumlah ',u_y[i],' : ', sum_y[i])
+   }
+
+   y_n       <- length(new_y)
+   y_mean    <- mean(new_y)
+   y_median  <- median(new_y)
+          y_modus          <-    new_y[which.max(tabulate(match(new_y,
unique(new_y))))]
+   y_sd      <- sd(new_y)
+   y_var     <- var(new_y)
+
+   ## Output ##
+   message('----------------------------------------')
+   message(' Jumlah sampel       : ', y_n)
+   message(' Mean                : ', round(y_mean,8))
+   message(' Median              : ', y_median)
+   message(' Mode                : ', y_modus)
+   message(' Standard Deviation  : ', round(y_sd,8))
+   message(' Variance            : ', round(y_var,8))
+   message('----------------------------------------')
+   message('========================================')
+
+   output        <- NULL
+   output$N      <- y_n
+   output$mean   <- y_mean
+   output$median <- y_median
+   output$modus  <- y_modus
+   output$sd     <- y_sd
+   output$var    <- y_var
+   output$uniq   <- u_y
+   output$data   <- data.frame(Uniq = u_y, Jumlah = sum_y)
+   output$new_y  <- new_y
+
+   return(output)
+ }
> ## data simulasi cura hujan daerah B ##
> dt_daerah_b   <- c()
> rd_daerah_b   <- round(rnorm(7305 , 1.5, 1),0)
> for (i in 1:length(rd_daerah_b)) {
+   if(rd_daerah_b[i] == 1){
```

```
+     dt_daerah_b[i] <- 'Gerimis    '
+   } else if(rd_daerah_b[i] == 2){
+     dt_daerah_b[i] <- 'Panas      '
+   } else if(rd_daerah_b[i] == 3){
+     dt_daerah_b[i] <- 'Hujan      '
+   } else {
+     dt_daerah_b[i] <- 'Hujan Deras'
+   }
+ }
> summary_daerah_b     <- sum_of_event(dt_daerah_b)
========================================
----------------------------------------
 Jumlah Panas       : 2543
 Jumlah Gerimis     : 2470
 Jumlah Hujan Deras : 1350
 Jumlah Hujan       : 942
----------------------------------------
 Jumlah sampel      : 7305
 Mean               : 2.09459274
 Median             : 2
 Mode               : 1
 Standard Deviation : 1.01976876
 Variance           : 1.03992832
----------------------------------------
========================================
> x_tmp                <- summary_daerah_b$data$Uniq
> y_tmp                <- summary_daerah_b$data$Jumlah
> df_daerah_b          <- data.frame(x = x_tmp, y=y_tmp)
> df_daerah_b$fraction <- df_daerah_b$y / sum(df_daerah_b$y)
> df_daerah_b$ymax     <- cumsum(df_daerah_b$fraction)
> df_daerah_b$ymin     <- c(0, head(df_daerah_b$ymax, n=-1))
>  df_daerah_b$percent    <-  paste(round(df_daerah_b$fraction*100,2),
'%', sep = '')
>   df_daerah_b$Ket                  <-   paste(df_daerah_b$x,'   (',
df_daerah_b$percent,')', sep = '')
> ## Daerah B ##
> ggplot(df_daerah_b, aes(x=x, y=y, fill=x)) +
+   geom_bar(stat="identity", size=1.5, color='#000000')+
+
+   geom_text(aes(label=paste(y,'\n(',percent,')', sep = '')),
+             vjust=2, angle=0 , size=5, color='#FFFFFF') +
+
+   scale_fill_manual(values = c('#008080',
+                                '#FF8080',
```

```
+                                          '#FF7742',
+                                          '#555555'))+
+
+   scale_color_manual(values = c('#008080',
+                                           '#FF8080',
+                                           '#FF7742',
+                                           '#555555'))+
+
+   theme(legend.position = "none") +
+   labs(title = 'b. Jumlah & Persentase Curah Hujan Daerah B',
+        y='Jumlah (Hari)',
+        x='Curah Hujan')
>
```

Output:



Pada gambar 1.6 dapat dianalisis pula bahwa 20 tahun terakhir pada daerah B hanya terjadi gerimis dan 34.81% dari 20 tahun tidak terjadi hujan, dengan kata lain peluang terjadinya panas pada daerah B adalah 0.3481. Maka dapat disimpulkan bahwa kedua daerah memiliki curah hujan yang berbeda, kedua daerah dapat dijadikan satu grafik untuk

memudahkan melakukan perbandingan atau dapat dilihat pada gambar berikut.

Gambar 0.5. Barplot Perbandingan Curah Hujan Daerah A dengan Daerah B 20 tahun terakhir.

Implementasi dalam program R

```
Sintaks (lanjutan dari program gambar 1.5 dan 1.6):
> ## Perbandingan 2 Daerah ##
> i_tmp <- 0
> x_tmp <- c()
> y_tmp <- c()
> z_tmp <- c()
> for(i in 1:length(summary_daerah_a$data$Uniq)){
+   for (j in 1:length(summary_daerah_b$data$Uniq)) {
+                              if(summary_daerah_a$data$Uniq[i]      ==
summary_daerah_b$data$Uniq[j]){
+       i_tmp        <- i_tmp + 1
+       x_tmp[i_tmp] <- summary_daerah_a$data$Uniq[i]
+       y_tmp[i_tmp] <- summary_daerah_a$data$Jumlah[i]
+       z_tmp[i_tmp] <- 'Daerah A'
```

```
+       i_tmp        <- i_tmp + 1
+       x_tmp[i_tmp] <- summary_daerah_b$data$Uniq[j]
+       y_tmp[i_tmp] <- summary_daerah_b$data$Jumlah[j]
+       z_tmp[i_tmp] <- 'Daerah B'
+     }
+   }
+ }
> daerah_ab <- data.frame(x=x_tmp, y=y_tmp, Ket=z_tmp)
> ## Daerah A dan B ##
> ggplot(daerah_ab, aes(x=x, y=y, fill=Ket)) +
+    geom_bar(position = 'dodge', stat='identity', color='#000000',
size=1.05) +
+
+                  annotate('text',      label=paste(daerah_ab$y[1],'
(',round(daerah_ab$y[1]/7305,2)*100,'%)', sep=''), color='#FFFFFF',
+            x = 0.8, y =2000, angle=90, size = 5) +
+                  annotate('text',      label=paste(daerah_ab$y[2],'
(',round(daerah_ab$y[2]/7305,2)*100,'%)', sep=''), color='#FFFFFF',
+            x = 1.2, y =2000, angle=90, size = 5) +
+                  annotate('text',      label=paste(daerah_ab$y[3],'
(',round(daerah_ab$y[3]/7305,2)*100,'%)', sep=''), color='#FFFFFF',
+            x = 1.75, y =2000, angle=90, size = 5)+
+                  annotate('text',      label=paste(daerah_ab$y[4],'
(',round(daerah_ab$y[4]/7305,2)*100,'%)', sep=''), color='#FFFFFF',
+            x = 2.15, y =500, angle=90, size = 5)+
+                  annotate('text',      label=paste(daerah_ab$y[5],'
(',round(daerah_ab$y[5]/7305,2)*100,'%)', sep=''), color='#FFFFFF',
+            x = 2.8, y =500, angle=90, size = 5)+
+                  annotate('text',      label=paste(daerah_ab$y[6],'
(',round(daerah_ab$y[6]/7305,2)*100,'%)', sep=''), color='#FFFFFF',
+            x = 3.2, y =500, angle=90, size = 5)+
+                  annotate('text',      label=paste(daerah_ab$y[7],'
(',round(daerah_ab$y[7]/7305,2)*100,'%)', sep=''), color='#FFFFFF',
+            x = 3.8, y =500, angle=90, size = 5)+
+                  annotate('text',      label=paste(daerah_ab$y[8],'
(',round(daerah_ab$y[8]/7305,2)*100,'%)', sep=''), color='#FFFFFF',
+            x = 4.2, y =2000, angle=90, size = 5)+
+   theme(legend.position = "none") +
+   scale_fill_manual(values = c('#008080',
+                                '#FF7742'))+
+
+   scale_color_manual(values = c('#008080',
+                                 '#FF7742'))+
+
```

```
+   ## untuk keterangan ##
+   annotate('text',label='',size=8,
+            x = 5:6, y=200) +
+
+   annotate('text',label='Keterangan:',
+            size=6, x=5.2, y=2000) +
+
+   annotate('text',label='.', color='#008080',
+            size=60, x=4.77, y=2100) +
+
+   annotate('text',label='.', color='#FF7742',
+            size=60, x=4.77, y=1900) +
+
+   annotate('text',label=': Daerah A',
+            size=4, x=5.25, y=1620) +
+
+   annotate('text',label=': Daerah B',
+            size=4, x=5.25, y=1440) +
+
+   labs(title = 'Perbandingan Curah Hujan Daerah A dengan Daerah B',
+        y='Jumlah (Hari)',
+        x='Curah Hujan')
>
Output:
```

Statistik inferensial secara tradisional dipahami terdiri dari metode dan prosedur yang digunakan untuk menarik kesimpulan mengenai proses yang mendasar yang dihasilkan data. Thiébaux dan Pedder (1987) mengungkapkan hal ini dengan menyatakan bahwa statistik adalah "seni meyakinkan dunia untuk menghasilkan

informasi tentang dirinya sendiri." Pemahaman fisik kita tentang fenomena atmosfer sebagian datang melalui manipulasi statistik dan analisis data. Dalam konteks ilmu atmosfer, mungkin masuk akal untuk menginterpretasikan statistik inferensial sedikit lebih luas juga, dan memasukkan peramalan cuaca secara statistik. Sekarang, bidang penting seperti ini memiliki tradisi panjang dan merupakan bagian penting dari peramalan cuaca operasional di pusat-pusat meteorologi di seluruh dunia.

1.3 Ketidakpastian dalam Ilmu Atmosfer

Hal mendasar tentang statistik deskriptif dan inferensial adalah gagasan tentang ketidakpastian. Jika proses atmosfer konstan, atau sangat periodik, menjelaskannya secara matematis pasti akan mudah. Perkiraan cuaca juga akan mudah, dan meteorologi akan membosankan. Tentu saja, atmosfer memperlihatkan variasi dan fluktuasi yang tidak beraturan. Ketidakpastian ini adalah kekuatan pendorong di balik pengumpulan dan analisis kumpulan data besar yang dirujuk pada bagian sebelumnya. Hal ini juga menyiratkan bahwa perkiraan cuaca tentu tidak pasti. Peramal cuaca yang memprediksi suhu tertentu pada hari berikutnya sama sekali tidak terkejut (dan mungkin bahkan senang) jika suhu yang diamati selanjutnya berbeda satu atau dua derajat. Untuk menangani ketidakpastian secara kuantitatif, perlu menggunakan probabilitas, yang merupakan bahasa matematika dari ketidakpastian.

Sebelum meninjau dasar-dasar dari probabilitas, ada baiknya mengkaji mengapa ada ketidakpastian tentang atmosfer. Lagi pula, kita memiliki model komputer yang besar dan canggih yang mewakili fisika atmosfer, dan model semacam itu digunakan secara rutin untuk meramalkan perubahannya di masa yang akan datang. Model biasa bersifat

deterministik: tidak mewakili ketidakpastian. Setelah disuplai dengan keadaan atmosfer awal tertentu (angin, suhu, kelembapan, dll., secara komprehensif melalui kedalaman atmosfer dan di sekitar planet) dan gaya batas (terutama radiasi matahari, dan kondisi permukaan laut dan daratan) masing-masing akan menghasilkan hasil tunggal tertentu. Menjalankan ulang model dengan input yang sama tidak akan mengubah hasil tersebut.

Pada prinsipnya, model atmosfer dapat memberikan perkiraan tanpa ketidakpastian, namun tidak, karena dua alasan. Pertama, meskipun model bisa sangat bagus dan memberikan perkiraan perilaku atmosfer yang cukup baik, model bukan representasi yang lengkap dan benar apabila dihadapkan pada situasi yang lebih luas. Penyebab yang tidak dapat dihindari dari masalah ini adalah bahwa beberapa proses fisik yang relevan beroperasi pada skala yang terlalu kecil untuk diwakili secara eksplisit oleh model ini, dan efeknya pada skala yang lebih besar harus didekati menggunakan informasi skala besar pula.

Bahkan jika semua ilmu fisika yang relevan entah bagaimana dapat dimasukkan dalam model atmosfer, bagaimanapun, kita masih tidak dapat menghindari ketidakpastian karena adanya unsur yang dikenal sebagai kekacauan dinamis (*dynamical chaos*). Fenomena ini ditemukan oleh seorang ilmuwan atmosfer (Lorenz, 1963), yang juga memberikan pengantar yang sangat mudah dibaca untuk subjek tersebut (Lorenz, 1993). Secara sederhana, perubahan waktu dari sistem dinamis deterministik nonlinier (misalnya, persamaan gerak atmosfer, atau atmosfer itu sendiri) sangat bergantung pada kondisi awal sistem. Jika dua realisasi dari sistem seperti itu dimulai dari dua kondisi awal yang hanya sedikit berbeda, kedua solusi tersebut pada akhirnya akan sangat berbeda. Untuk kasus simulasi

atmosfer, bayangkan salah satu dari sistem ini adalah atmosfer sebenarnya, dan yang lainnya adalah model matematis sempurna dari fisika yang mengatur atmosfer. Karena atmosfer selalu diamati secara tidak lengkap, tidak mungkin memulai model matematika dalam keadaan yang persis sama dengan sistem nyata. Jadi, meskipun modelnya sempurna, tetap tidak mungkin untuk menghitung apa yang akan dilakukan atmosfer jauh ke masa depan tanpa batas waktu. Oleh karena itu, perkiraan deterministik perilaku atmosfer masa depan akan selalu tidak pasti, dan metode probabilistik akan selalu diperlukan untuk menggambarkan perilaku tersebut.

Kesadaran bahwa atmosfer menunjukkan dinamika yang rumit telah mengakhiri harapan tentang perkiraan cuaca yang sempurna (bebas ketidakpastian) yang menjadi dasar filosofis bagi sebagian besar meteorologi abad ke-20 (penjelasan tentang sejarah dan budaya ilmiah ini disediakan oleh Friedman, 1989). Dinamika yang rumit dan kesalahan yang tidak dapat dihindari dalam representasi matematis atmosfer menyiratkan bahwa "semua masalah prediksi meteorologi, dari perkiraan cuaca hingga proyeksi perubahan iklim, pada dasarnya bersifat probabilistik" (Palmer, 2001).

Terakhir, perlu dicatat bahwa keacakan bukanlah keadaan "tidak dapat diprediksi", atau "tidak ada informasi", seperti yang dipikirkan oleh kebanyakan orang. Sebaliknya, acak berarti "tidak dapat diprediksi atau ditentukan secara tepat". Misalnya, jumlah curah hujan yang akan terjadi besok di tempat Anda tinggal adalah jumlah acak yang tidak Anda ketahui hari ini. Namun, analisis statistik sederhana dari catatan curah hujan klimatologis di lokasi Anda akan menghasilkan frekuensi relatif jumlah curah hujan yang akan memberikan lebih banyak informasi tentang curah hujan besok di lokasi Anda daripada yang saya miliki saat saya

duduk menulis kalimat ini. Gagasan hujan besok yang masih kurang pasti mungkin tersedia untuk Anda dalam bentuk ramalan cuaca. Mengurangi ketidakpastian tentang peristiwa meteorologi acak adalah tujuan dari perkiraan cuaca. Selanjutnya, metode statistik memungkinkan estimasi ketepatan prediksi, yang dengan sendirinya dapat menjadi informasi yang berharga.

BAB 2
REVIEW PROBABILITAS

2.1 Pendahuluan

Materi pada bab ini adalah ulasan singkat tentang elemen dasar probabilitas. Perlakuan yang lebih lengkap dan lebih baik dari dasar-dasar probabilitas dapat ditemukan dalam pengantar statistik, seperti Pengantar Analisis Statistik Dixon dan Massey (1983), atau Pengantar Winkler (1972) untuk Inferensi dan Keputusan Bayesian, dll.

Ketidakpastian kita tentang atmosfer, atau tentang yang lainnya, memiliki derajat yang berbeda dalam kasus yang berbeda. Misalnya, Anda tidak dapat sepenuhnya yakin bahwa besok akan terjadi hujan atau tidak, atau apakah suhu rata-rata bulan depan akan lebih besar atau lebih kecil dari suhu rata-rata bulan lalu. Tetapi Anda mungkin lebih yakin tentang salah satu dari pertanyaan tersebut.

Ketidak cukupan atau bahkan sangat informatif untuk mengatakan bahwa suatu peristiwa yang tidak pasti. Hal tersebut membuat kita dihadapkan pada masalah dalam mengungkapkan atau mengkarakterisasi derajat ketidakpastian. Pendekatan yang mungkin adalah dengan menggunakan deskriptor kualitatif seperti kemungkinan, tidak mungkin, mungkin, atau kebetulan. Bagaimanapun dalam menyampaikan ketidakpastian melalui frase tersebut hasilnya pasti ambigu namun terbuka untuk berbagai interpretasi (Beyth-Maron, 1982; Murphy dan Brown, 1983). Misalnya, tidak jelas yang mana dari “kemungkinan hujan” atau “mungkin hujan” yang menunjukkan lebih sedikit ketidakpastian tentang prospek hujan.

Umumnya lebih baik menyatakan ketidakpastian secara kuantitatif dan ini dilakukan dengan menggunakan angka yang disebut probabilitas. Dalam arti terbatas, probabilitas tidak lebih dari sistem matematika abstrak yang dapat dikembangkan secara logis dari tiga premis yang disebut Aksioma Probabilitas. Sistem ini tidak menarik bagi banyak orang, termasuk mungkin Anda sendiri, kecuali bahwa konsep abstrak yang dihasilkan relevan dengan sistem dunia nyata yang melibatkan ketidakpastian. Sebelum menyajikan aksioma probabilitas dan beberapa implikasinya untuk mendefinisikan beberapa terminologi.

2.2 Elemen Probabilitas

2.2.1 Events (Kejadian)

Suatu peristiwa (*Event*s) adalah suatu himpunan, kelas, atau kelompok dari kemungkinan hasil yang tidak pasti. Peristiwa dapat terdiri dari dua jenis yaitu: Suatu peristiwa majemuk dapat didekomposisikan menjadi dua atau lebih (sub) peristiwa dan peristiwa yang elemen-nya tidak bisa. Sebagai contoh sederhananya yaitu pelemparan suatu dadu yang memiliki enam kemungkinan. Peristiwa "munculnya angka genap" merupakan kejadian majemuk, karena akan terjadi jika muncul angka dua, empat, atau enam.

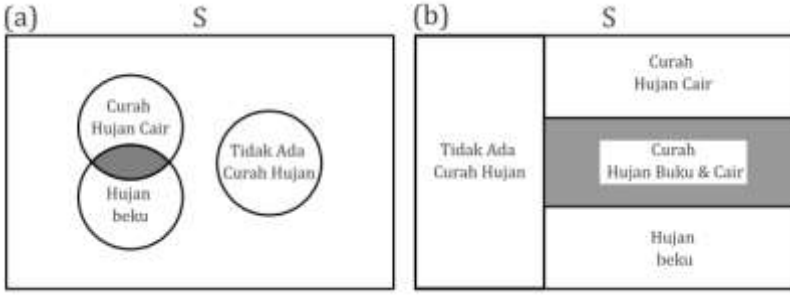
Dalam kasus sederhana seperti pada pelemparan dadu biasanya sudah jelas yang mana kejadian sederhana dan yang mana kejadian majemuk. Secara umum yang didefinisikan sebagai elemen atau kejadian majemuk yaitu kejadian yang seringkali bergantung pada masalah yang dihadapi dan tujuan dilakukannya analisis. Misalnya, peristiwa "presipitasi terjadi besok" bisa menjadi peristiwa elemen yang harus dibedakan dari peristiwa elemen "presipitasi tidak terjadi besok". Tetapi jika untuk membedakan lebih jauh antara bentuk-bentuk presipitasi (presipitasi yang terjadi) akan dianggap sebagai

peristiwa majemuk, kemungkinanya terdiri dari tiga peristiwa elemen yaitu “presipitasi cair”, “presipitasi beku”, dan “presipitasi cair dan beku”. Jika kita lebih tertarik pada berapa banyak presipitasi yang akan terjadi, ketiga peristiwa ini sendiri akan dianggap sebagai gabungan masing-masing terdiri dari setidaknya dua peristiwa elemen. Dalam kasus ini, misalnya peristiwa majemuk “presipitasi beku” akan terjadi jika salah satu peristiwa elemennya memenuhi “presipitasi beku mengandung sedikitnya 0,01-in. kadar air” atau “presipitasi cair yang mengandung kurang dari 0,01-in. kadar air”.

2.2.2 Ruang Sampel

Ruang sampel atau ruang peristiwa adalah himpunan dari semua kemungkinan kejadian. Dengan demikian ruang sampel mewakili alam semesta dari semua kemungkinan hasil atau peristiwa. Hal ini ekuivalen dengan peristiwa majemuk terbesar yang mungkin terjadi.

Hubungan antar kejadian dengan ruang sampel dapat direpresentasikan secara geometris dengan menggunakan Diagram Venn. Disini ruang sampel digambarkan sebagai persegi panjang dan kejadian digambarkan sebagai lingkaran, seperti pada Gambar 2.1a. Di sini ruang sampelnya adalah persegi panjang berlabel S, yang mungkin berisi kumpulan hasil curah hujan yang mungkin untuk besok. Empat peristiwa dasar digambarkan dalam batas tiga lingkaran. Lingkaran “Tanpa presipitasi” digambar tidak tumpang tindih dengan yang lainnya karena baik presipitasi cair maupun beku tidak dapat terjadi jika tidak terjadi presipitasi. Area yang diarsir mewakili peristiwa "presipitasi cair dan beku" yang terletak diantara "Presipitasi beku" dan "Presipitasi cair".



Gambar 2.6. Diagram Venn mewakili hubungan peristiwa curah hujan yang dipilih. Wilayah yang diaksir mewakili peristiwa "presipitasi cair dan beku." (a) Peristiwa digambarkan sebagai lingkaran dalam ruang sampel. (b) Peristiwa yang sama digambarkan sebagai persegi panjang.

Kita tidak perlu memikirkan diagram Venn menggunakan lingkaran untuk mewakili peristiwa. Gambar 2.1 b adalah diagram Venn yang menggunakan empat persegi panjang yang mengisi seluruh ruang sampel S. Hal ini jelas bahwa S terdiri dari tepat empat kejadian elemen yang mewakili yang mungkin terjadian yang mungkin terjadi. Kumpulan semua peristiwa dasar yang mungkin terjadi disebut saling eksklusif dan lengkap secara kolektif (MECE). Saling eksklusif berarti bahwa tidak lebih dari satu kejadian yang dapat terjadi. Lengkap secara kolektif berarti bahwa setidaknya satu peristiwa akan terjadi.

Contoh 0.1 Menemukan Ruang Sampel dari Pelemparan Dadu

Buatlah ruang sampel untuk percobaan yang terdiri dari melempar satu dadu. Temukan peristiwa yang sesuai setelah melakukan pelemparan dadu dengan frasa "muncul angka genap" dan "angka lebih dari dua".

Solusi:

Hasil dari semua kemungkinan yaitu himpunan S={1,2,3,4,5,6}

Hasil genap yang mungkin adalah 2,4 dan 6, jadi bilangan genap adalah himpunan {2,4,6} atau dapat dituliskan E={2,4,6}.

Demikian pula, peristiwa yang sesuai dengan ungkapan "angka yang lebih besar dari dua" adalah himpunan T={3,4,5,6} yang dilambangkan dengan T.

Contoh 0.2 Menemukan Ruang Sampel dari Pelemparan Dadu dan Koin

Misalkan kita melempar koin dan melempar dadu. Timbul pertanyaan:

- Berapa banyak hasil yang mungkin?
- Apa itu ruang probabilitas?
- Bagaimana cara mengidentifikasi peristiwa tersebut

Saat kita melempar koin, hanya ada dua hasil yang mungkin terjadi yaitu {kepala(H) atau ekor(T)}, dan saat kita melempar dadu, ada enam kemungkinan hasil yang akan terjadi yaitu {1,2,3,4,5,6}.

Maka itu berarti kita memiliki dua kejadian atau peristiwa:

Peristiwa pertama yaitu terdiri dari "kepala dan ekor".

Peristiwa lainnya terdiri dari {1,2,3,4,5,6}. Sekarang, mari kita lihat apakah kita dapat menemukan ruang sampelnya.

Jika kita melemparnya secara bersamaan, maka kita bisa mendapatkan salah satu dari skenario berikut:

Kepala dan 1	Ekor dan 1
Kepala dan 2	Ekor dan 2
Kepala dan 3	Ekor dan 3
Kepala dan 4	Ekor dan 4
Kepala dan 5	Ekor dan 5
Kepala dan 6	Ekor dan 6

Atau lebih sederhananya dinyatakan dalam ruang sampel {H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6}.

Apa yang baru saja kita lakukan adalah membuat daftar semua hasil yang mungkin dari melempar koin dan melempar dadu. Dan jika diperhatikan, kita dapat menemukan 12 kemungkinan hasil untuk eksperimen khusus ini.

2.2.3 Aksioma Probabilitas

Setelah mendefinisikan ruang sampel dan kejadian-kejadian dari berbagai kemungkinan, langkah selanjutnya adalah menghubungkan probabilitas dengan masing-masing kejadian. Aturan untuk melakukannya terdiri dari tiga Aksioma Probabilitas yaitu:

1. Probabilitas suatu peristiwa yang nilainya tidak negatif.
2. Probabilitas kejadian majemuk S adalah 1.
3. Probabilitas saling bebas merupakan kejadian dimana tidak ada elemen yang sama antara kejadian satu dengan kejadian yang lainnya.

2.3 Pengertian Probabilitas

Aksioma adalah logika dasar untuk perhitungan probabilitas. Artinya, sifat matematis dari probabilitas semua dapat disimpulkan dari aksioma. Hal ini akan dibahas pada bab ini.

Namun demikian, aksioma-aksioma tersebut tidak terlalu informatif tentang apa arti probabilitas yang sebenarnya. Ada dua pandangan yang berlaku tentang arti probabilitas yaitu dari sisi frekuensi dan Bayesian dan ada juga interpretasi lain (Gillies, 2000). Mungkin yang mengejutkan, tidak ada kontroversi kecil dalam dunia statistik tentang kebenaran dari keduanya. Karena kita lebih fokus pada subjek telah menjadi begitu besar sehingga penganut satu interpretasi atau sebagainya yang telah diketahui terhadap mereka yang memiliki pandangan yang berbeda!

Harus ditekankan bahwa matematikanya sama dalam setiap kasus, karena probabilitas Frequentist dan Bayesian mengikuti secara logis dari aksioma yang sama. Perbedaannya hanya terletak pada penafsirannya. Kedua interpretasi probabilitas yang berlaku diterima dan berguna dalam ilmu atmosfer, seperti halnya dualitas partikel/gelombang dari sifat radiasi elektromagnetik diterima dan berguna dalam fisika.

2.3.1 Interpretasi frekuensi

Interpretasi frekuensi adalah pandangan tradisional tentang probabilitas. Perkembangannya pada abad ke-18 dimotivasi oleh keinginan untuk memahami permainan peluang dan untuk mengoptimalkan taruhan yang bersangkutan. Dari perspektif ini, probabilitas suatu peristiwa adalah frekuensi relatif jangka panjangnya. Definisi ini diformalkan dalam Hukum Bilangan Besar, yang menyatakan bahwa rasio antara jumlah kejadian dari peristiwa {E} dan jumlah probabilitas {E} untuk terjadi akan konvergen ke probabilitas {E}, dinotasikan Pr{ E}, seiring dengan meningkatnya jumlah peluang. Hal ini bisa ditulis sebagai:

$$\Pr\left\{\left|\frac{a}{n} - \Pr(E)\right| \geq \varepsilon\right\} \rightarrow 0 \text{ dengan } n \rightarrow \infty \qquad (0.1)$$

dimana a adalah jumlah kejadian, n adalah jumlah peluang (dengan demikian $\frac{a}{n}$ adalah frekuensi relatif) dan ε adalah angka sangat kecil.

Interpretasi frekuensi tersebut masuk akal secara intuitif dan berdasarkan empiris. Hal ini sangat berguna dalam aplikasi seperti memperkirakan probabilitas klimatologis dengan menghitung frekuensi relatif historisnya. Misalnya, dalam 50 tahun terakhir maka terdapat $31 \times 50 = 1550$ hari pada bulan Agustus. Jika hujan turun selama 487 hari di suatu lokasi yang diinginkan, maka dpaat diperkiraan untuk probabilitas klimatologis presipitasi di lokasi tersebut pada hari terjadinya hujan di bulan Agustus adalah $\frac{487}{1550} = 0{,}314$.

2.3.2 Interpretasi Bayesian (SubJektif)

Tegasnya, menggunakan tampilan frekuensi dari probabilitas membutuhkan rangkaian panjang percobaan yang identik. Untuk memperkirakan probabilitas klimatologis dari data cuaca historis, persyaratan ini pada dasarnya tidak menimbulkan masalah. Namun, ketika seseorang berpikir tentang probabilitas untuk acara seperti {tim sepak bola perguruan tinggi dengan harapan Anda akan memenangkan setidaknya setengah dari pertandingan pada musim depan}, kami mendapatkan beberapa kesulitan dalam konteks frekuensi relatif. Meskipun secara abstrak kita dapat membayangkan serangkaian musim sepak bola dan hipotetis yang identik dengan yang akan datang, rangkaian musim sepak bola fiksi ini tidak membantu memperkirakan kemungkinan sebenarnya dari peristiwa tersebut.

Interpretasi subJektif adalah probabilitas mewakili tingkat kepercayaan atau penilaian seseorang bahwa suatu peristiwa yang tidak pasti akan terjadi. Misalnya, ada sejarah

panjang peramal cuaca secara rutin (dan sangat terampil) memperkirakan probabilitas peristiwa seperti curah hujan yang terjadi pada hari-hari dalam waktu dekat. Sama halnya dengan perguruan tinggi atau universitas Anda adalah sekolah yang cukup besar bagi para pemain profesional untuk peduli dengan hasil pertandingan sepak bola, probabilitas yang terkait dengan hasil tersebut juga dinilai secara teratur - secara subJektif.

Misalkan ada dua orang memiliki probabilitas subyektif yang berbeda dari suatu peristiwa, dan seringkali perbedaan perkiraan probabilitas tersebut disebabkan oleh informasi dan/atau pengalaman yang berbeda. Namun, hanya karena orang yang berbeda mungkin memiliki probabilitas subyektif yang berbeda untuk peristiwa yang sama, tidak berarti bahwa seseorang bebas memilih angka apa pun dan menyebutnya probabilitas. Penilaian terukur harus menjadi penilaian yang konsisten agar menjadi probabilitas subjektif yang sah. Konsistensi ini berarti, antara lain, bahwa probabilitas subyektif harus konsisten dengan aksioma probabilitas dan dengan demikian dengan sifat perhitungan probabilitas yang tersirat oleh aksioma. Monografi Epstein (1985) memberikan pengantar yang baik untuk metode Bayesian terkait dengan masalah atmosfer.

Contoh 0.3 Menemukan Probabilitas dari Pelemparan Dadu

Pelemparan dadu yang "seimbang" atau "adil". Ketika dadu tersebut dilantunkan. Tentukan probabilitas:

a. "munculnya angka genap" anggap sebagai E1.
b. "munculnya angka yang lebih besar dari dua" anggap sebagai E2.

Solusi:

Hal utama yang harus dilakukan yaitu menentukan ruang sampel dari dadu tersebut. Karena dadu memiliki 6 sisi

atau dapat dikatakan bahwa seluru ruang sampel dari dadu tersebut yaitu 1,2,3,4,5, dan 6 atau dapat dituliskan sebagai berikut:

$$S = \{1,2,3,4,5,6\}$$

a. Ruang sampel menculnya angka genap:
Karena ruang sampelnya $S = \{1,2,3,4,5,6\}$, maka angka genap yang muncul yaitu 2,4, dan 6. Sehingga dapat ditulis bahwa ruang sampel dari peristiwa pertama yaitu:

$$E_1 = \{2,4,6\}$$

Karena ruang sampel peristiwa pertama telah ditemukan, dan probabilitas setiap kejadiannya itu 1/6. Maka probabilitas munculnya angka genap:

$$\Pr(E_1) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

b. Ruang sampel menculnya angka lebih dari dua:
Karena ruang sampelnya $S = \{1,2,3,4,5,6\}$, maka angka yang lebih dari dua yang muncul yaitu 3,4,5, dan 6. Sehingga dapat ditulis bahwa ruang sampel dari peristiwa kedua yaitu:

$$E_2 = \{3,4,5,6\}$$

Sama halnya dengan bagian a, bahwa probabilitas setiap kejadian itu 1/6. Maka probabilitas munculnya angka yang lebih dari dua:

$$\Pr(E_2) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$$

Implementasi dalam program R

```
Sintaks:
> ## Membersihkan history ##
> rm(list = ls())
> S <- c(1,2,3,4,5,6)
>
> ## Bagian a ##
> E1    <- c(2,4,6)
> Pr_E1 <- length(E1)/length(S)
> cat('Probabilitas E1 =',Pr_E1)
Probabilitas E1 = 0.5
>
> ## Bagian b ##
> E2    <- c(3,4,5,6)
> Pr_E2 <- length(E2)/length(S)
> cat('Probabilitas E2 =',Pr_E2)
Probabilitas E2 = 0.6666667
>
```

2.4 Beberapa Sifat Probabilitas

Salah satu alasan diagram Venn bisa sangat berguna adalah karena memungkinkan probabilitas divisualisasikan secara geometris sebagai area. Keakraban dengan hubungan geometris di dalam dunia fisik kemudian dapat digunakan untuk lebih memahami dunia probabilitas yang lebih halus. Bayangkan luas persegi panjang pada Gambar 2.1 b adalah 1 menurut aksioma kedua. Aksioma pertama menyatakan bahwa tidak ada nilai probabilitas yang negatif. Aksioma ketiga menyatakan bahwa luas total bagian yang tidak tumpang tindih adalah jumlah luas bagian-bagian tersebut.

Bagian ini mencantumkan beberapa sifat perhitungan probabilitas yang secara logis mengikuti aksioma. Analog geometris untuk probabilitas yang disediakan oleh diagram Venn dapat digunakan untuk memvisualisasikan hal tersebut.

2.4.1 Domain, Subhimpunan, Komplemen, dan Serikat

Secara tidak lansung bahwa aksioma pertama dan kedua menyiratkan bahwa probabilitas setiap peristiwa terletak antara nol dan satu. Kendala pada rentang probabilitas ini dapat dinyatakan secara matematis sebagai:

$$0 \leq \Pr\{E\} \leq 1 \tag{0.2}$$

Jika $\Pr\{E\} = 0$, kejadian tidak terjadi. Jika $\Pr\{E\} = 1$, kejadian tersebut pasti terjadi. Jika peristiwa $\{E_2\}$ harus terjadi ketika peristiwa $\{E_1\}$ terjadi, $\{E_1\}$ dikatakan sebagai himpunan bagian dari $\{E_2\}$. Misalnya, $\{E_1\}$ dan $\{E_2\}$ masing-masing dapat menunjukkan terjadinya presipitasi beku dan terjadinya presipitasi dalam bentuk apa pun. Dalam hal ini, aksioma ketiga dapat dituliskan dalam bentuk sebagai berikut:

$$\Pr\{E_1\} \leq \Pr\{E_2\} \tag{0.3}$$

Pelengkap dari peristiwa $\{E\}$ adalah selain dari peristiwa gabungan yang (umumnya) yang tidak terjadi. Sebagai contoh, pada Gambar 2.1b, komplemen dari kejadian "presipitasi cair dan beku" adalah kejadian majemuk "tidak ada presipitasi, atau hanya presipitasi cair, atau hanya presipitasi beku". Secara tidak lansung aksioma kedua dan ketiga dapat dituliskan sebagai berikut:

$$\Pr\{E\}^C = 1 - \Pr\{E\} \tag{0.4}$$

dimana $\{E\}^C$ menunjukkan komplemen dari $\{E\}$. (Banyak penulis menggunakan simbol C sebagai notasi alternatif untuk mewakili komplemen. Penggunaan simbol ini

sangat berbeda dari arti statistik yang secara umum, yaitu untuk menunjukkan rata-rata aritmatika).

Penyatuan dua peristiwa adalah peristiwa majemuk dimana satu atau kedua peristiwa itu terjadi atau bahkan lebih dari dua. Dalam notasi himpunan, serikat dilambangkan dengan simbol ∪. Sebagai konsekuensi dari aksioma ketiga, probabilitas serikat dapat dihitung dengan rumus berikut:

$$\begin{aligned}\Pr\{E_1 \cup E_2\} &= \Pr\{E_1 \text{ atau } E_2 \text{ atau } E_n\} \\ &= \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1 \cap E_2\}\end{aligned} \quad (0.5)$$

Simbol ∩ disebut operator irisan yaitu:

$$\Pr\{E_1 \cap E_2\} = \Pr\{E_1, E_2\} = \Pr\{E_1 \text{dan } E_2\} \quad (0.6)$$

$\Pr\{E_1 \cap E_2\}$ adalah kejadian terjadinya $\{E_1\}$ dan $\{E_2\}$. Notasi $\{E_1, E_2\}$ sama dengan $\{E_1 \cap E_2\}$. Nama lain untuk $\Pr\{E_1 \cap E_2\}$ adalah probabilitas gabungan dari $\{E_1\}$ dan $\{E_2\}$. Persamaan 2.5 terkadang disebut sebagai hukum probabilitas aditif. Itu berlaku apakah $\{E_1\}$ dan $\{E_2\}$ saling eksklusif atau tidak. Akan tetapi, jika kedua peristiwa saling lepas, kemungkinan tumpang tindihnya adalah nol, karena peristiwa yang saling lepas tidak dapat terjadi keduanya.

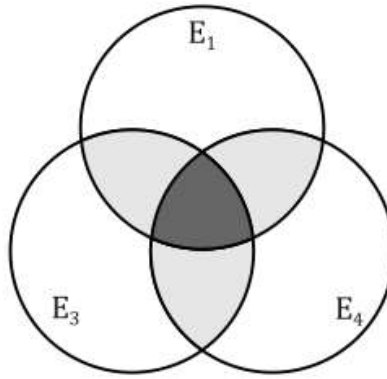
Probabilitas untuk kejadian $\Pr\{E_1, E_2\}$ yang merupakan pengurang dari dalam Persamaan 2.5 untuk mengkompensasi penghitungan dua kali ketika probabilitas untuk kejadian $\{E_1\}$ dan $\{E_2\}$ ditambahkan. Cara termudah untuk melihatnya adalah dengan memikirkan cara menemukan luas geometri total yang dikelilingi oleh dua lingkaran yang tumpang tindih pada Gambar 2.1a. Area yang diarsir pada Gambar 2.1a mewakili peristiwa persimpangan {presipitasi cair dan presipitasi beku} dan berada di dalam dua lingkaran berlabel "presipitasi cair" dan "presipitasi beku".

Hukum penjumlahan pada Persamaan 2.5, dapat diperluas menjadi gabungan dari tiga peristiwa atau lebih

dengan $\{E_1\}$ atau $\{E_2\}$ sebagai peristiwa gabungan (yaitu gabungan dari peristiwa lain) dan menerapkan Persamaan 2.5 secara rekursif. Misalnya, jika $\{E_2\} = \{E_3, E_4\}$, mensubstitusi ke persamaan 2.5, maka dapat dituliskan kembali sebagai berikut:

$$\begin{aligned}\Pr\{E_1 \cup E_3 \cup E_4\} = \Pr\{E_1\} + \Pr\{E_3\} + \Pr\{E_4\} \\ - \Pr\{E_1 \cap E_3\} - \Pr\{E_1 \cap E_4\} \\ - \Pr\{E_3 \cap E_4\} + \Pr\{E_1 \cap E_3 \cap E_4\} \quad (0.7)\end{aligned}$$

Hasil ini mungkin sulit dipahami secara aljabar, tetapi cukup mudah untuk divisualisasikan secara geometris. Gambar 2.2 mengilustrasikan situasi ini. Menjumlahkan luas dari tiga lingkaran satu per satu (baris pertama pada Persamaan 2.7) menghasilkan hitungan ganda dari luasan dengan dua pola penetasan yang tumpang tindih dan hitungan tiga kali lipat dari luas rata-rata yang terkandung dalam ketiga lingkaran. Baris kedua Persamaan 2.7 mengoreksi hitungan ganda, tetapi mengurangi luas wilayah pusat sebanyak tiga kali. Area ini ditambahkan kembali pada baris ketiga persamaan 2.7 untuk terakhir kalinya.



Gambar 2.7. Diagram Venn menggambarkan perhitungan probabilitas penyatuan tiga kejadian yang

tumpang tindih pada Persamaan 2.7. Area dengan dua pola penetasan yang tumpang tindih telah dihitung dua kali dan areanya harus dikurangi untuk mengimbanginya. Area tengah dengan tiga pola penetasan yang tumpang tindih dihitung tiga kali, tetapi kemudian dikurangi tiga kali saat mengoreksi penghitungan ganda. Lalu perlu ditambahkan kembali dengan luasnya.

2.4.2 Hukum DeMorgan

Memanipulasi pernyataan probabilitas yang melibatkan komplemen serikat atau persimpangan atau pernyataan yang melibatkan persimpangan serikat atau komplemen, difasilitasi oleh dua hubungan yang dikenal sebagai Hukum DeMorgan:

$$\Pr\{(A \cup B)^C\} = \Pr\{A^C \cap B^C\} \quad (0.8)$$

dan

$$\Pr\{(A \cap B)^C\} = \Pr\{A^C \cup B^C\} \quad (0.9)$$

Hukum pertama yaitu Persamaan 2.8 mengungkapkan fakta bahwa komplemen dari gabungan dua kejadian adalah perpotongan komplemen dari dua kejadian. Dalam istilah geometri diagram Venn, kejadian di luar gabungan $\{A\}$ dan $\{B\}$ (sisi kiri) secara bersamaan berada di luar $\{A\}$ dan $\{B\}$ (sisi kanan). Hukum kedua DeMorgan, Persamaan 2.9, menyatakan bahwa komplemen dari sebuah irisan dua kejadian adalah penyatuan komplemen dari dua kejadian individual. Di sini, dalam istilah geometris, kejadian yang tidak tumpang tindih antara $\{A\}$ dan $\{B\}$ (sisi kiri) adalah kejadian yang berada di luar $\{A\}$ atau di luar $\{B\}$ atau keduanya (sisi kanan).

2.4.3 Probabilitas Bersyarat

Kita sering tertarik pada probabilitas suatu peristiwa karena peristiwa lain telah terjadi atau akan terjadi. Misalnya, kemungkinan hujan beku saat hujan, mungkin menarik atau mungkin kita perlu mengetahui kemungkinan kecepatan angin pantai berada di atas ambang batas tertentu karena badai yang melanda di dekatnya. Ini adalah contoh probabilitas bersyarat. Kejadian ini "diberikan" atau disebut sebagai kondisi. Notasi tradisional untuk probabilitas bersyarat adalah garis vertikal, jadi kita menyebut $\{E_1\}$ sebagai peristiwa yang menarik dan $\{E_2\}$ peristiwa kodisinya, maka dapat dtuliskan sebagai berikut:

$$\Pr\{E_1|E_2\} = \Pr\{E_1 \text{ dengan } E2 \text{ telah terjadi atau akan terjadi}\} \qquad (0.10)$$

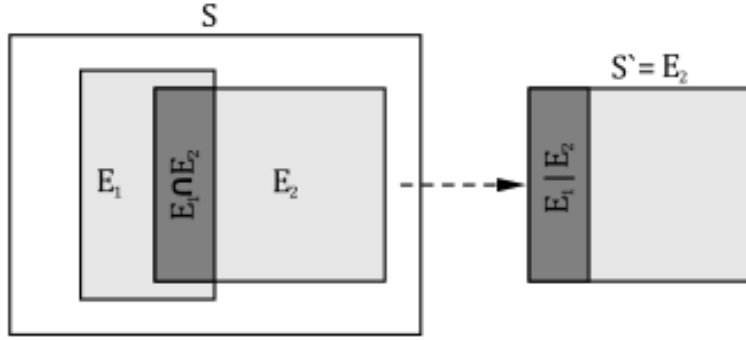
Jika peristiwa $\{E_2\}$ telah atau akan terjadi, probabilitas $\{E_1\}$ adalah probabilitas bersyarat $\Pr\{E_1|E_2\}$. Jika peristiwa pengkondisian belum atau tidak akan terjadi, ini dengan sendirinya tidak memberi tahu apa-apa tentang probabilitas $\{E_1\}$.

Secara lebih formal, probabilitas bersyarat didefinisikan dalam perpotongan peristiwa yang menarik dengan peristiwa pengkondisiannya yang dapat dituliskan dalam sabagai berikut:

$$\Pr\{E_1|E_2\} = \frac{\Pr\{E_1 \cap E_2\}}{\Pr\{E_2\}} \qquad (0.11)$$

Asalkan probabilitas peristiwa pengkondisian tidak nol. Secara intuitif, masuk akal jika probabilitas bersyarat terkait dengan probabilitas gabungan dari dua peristiwa yang dipertanyakan $\Pr\{E_1|E_2\}$. Sekali lagi, ini paling mudah dipahami dengan analogi dengan area dalam diagram Venn,

seperti yang ditunjukkan pada Gambar 2.3. Kami memahami bahwa probabilitas tak bersyarat $\{E_1\}$ diwakili oleh fraksi ruang sampel S yang ditempati oleh persegi panjang berlabel E_1. Kondisi pada $\{E_2\}$ berarti kita hanya tertarik pada hasil yang mengandung $\{E_2\}$.



Gambar 2.8. Menggambarkan definisi probabilitas bersyarat. Probabilitas tanpa syarat dari $\{E_1\}$ adalah pecahan luas S yang ditempati oleh $\{E_1\}$ di sisi kiri gambar. Pengkondisian $\{E_2\}$ bermuara pada mempertimbangkan ruang sampel baru S` yang hanya terdiri dari $\{E_2\}$, karena ini berarti bahwa kita hanya berurusan dengan kejadian dimana $\{E_2\}$ terjadi. Oleh karena itu, probabilitas bersyarat $\Pr\{E_1|E_2\}$ diberikan oleh fraksi luas ruang sampel baru S` yang ditempati oleh $\{E_1\}$ dan $\{E_2\}$. Hal ini dihitung dalam Persamaan 2.11.

Kami membuang hampir setiap bagian dari S yang tidak termasuk dalam $\{E_2\}$. Ini bermuara dengan mempertimbangkan ruang sampel baru yaitu *S`* yang bertepatan dengan $\{E_2\}$. Probabilitas bersyarat $\Pr\{E_1|E_2\}$ oleh karena itu direpresentasikan secara geometris sebagai fraksi dari area ruang sampel baru yang ditempati oleh $\{E_1\}$ dan $\{E_2\}$. Jika peristiwa pengkondisian dan peristiwa kepentingan

saling eksklusif, probabilitas bersyarat jelas harus nol karena probabilitas gabungannya akan menjadi nol.

2.4.4 Independen

Perhatikan kembali definisi probabilitas bersyarat pada Persamaan 2.11 memberikan bentuk ungkapan yang disebut hukum Multiplikatif probabilitas:

$$\Pr\{E_1 \cap E_2\} = \Pr\{E_1|E_2\}\Pr\{E_2\} = \Pr\{E_2|E_1\}\Pr\{E_1\} \quad (0.12)$$

Dua kejadian dikatakan saling bebas jika kejadian atau ketidakmunculan salah satu kejadian tidak mempengaruhi peluang kejadian yang lain. Misalnya, jika kita melempar dadu merah dan dadu putih, peluang keluarnya dadu merah tidak bergantung pada hasil dadu putih, dan sebaliknya. Hasil dari kedua dadu tidak bergantung satu sama lain. Independen antara $\{E_1\}$ dan $\{E_2\}$ menyiratkan $\Pr\{E_1|E_2\}$ $\Pr\{E1\}$ dan $\Pr\{E_2|E_1\}$ $\Pr\{E_2\}$. Peristiwa Independen membuat perhitungan probabilitas bersama sangat mudah, karena hukum multiplikatif kemudian direduksi menjadi:

$$\Pr\{E_1 \cap E_2\} = \Pr\{E_1\}\Pr\{E_2\}\text{, untuk } \{E_1\} \text{ dan } \{E_2\} \text{ independen} \quad (0.13)$$

Persamaan 2.13 dapat dengan mudah diperluas untuk menghitung probabilitas bersama untuk lebih dari dua peristiwa independen hanya dengan mengalikan semua probabilitas dari peristiwa tak bersyarat yang independen.

Contoh 0.4 Frekuensi Relatif Bersyarat

Pertimbangkan estimasi probabilitas klimatologis (yaitu jangka panjang atau frekuensi relatif) dengan menggunakan data yang tersedia di Tabel A.1 Lampiran A. Probabilitas klimatologis yang bergantung pada peristiwa lain dapat dihitung. Probabilitas seperti itu kadang-kadang

disebut sebagai probabilitas klimatologis bersyarat atau klimatologi bersyarat.

Misalkan untuk memperkirakan kemungkinan setidaknya 0,01 inci curah hujan setara cair di Ithaca pada bulan Januari, mengingat suhu minimum setidaknya 0°F. Secara umum, kedua peristiwa ini diharapkan terkait, karena suhu yang sangat dingin biasanya terjadi pada malam yang cerah dan curah hujan membutuhkan awan. Hubungan fisika ini akan membuat kita berharap bahwa kedua peristiwa ini terkait secara statistik (yaitu tidak independen) dan bahwa probabilitas curah hujan bersyarat pada kondisi suhu yang berbeda berbeda satu sama lain dan dari probabilitas tak bersyarat. Secara khusus, berdasarkan pemahaman kita tentang proses fisika yang mendasarinya, kita berharap bahwa probabilitas presipitasi pada suhu minimum 0°F atau lebih tinggi akan lebih besar daripada probabilitas bersyarat mengingat kejadian komplementer dari suhu minimum yang lebih dingin dari 0°F.

Untuk memperkirakan probabilitas ini dari kelimpahan relatif bersyarat, kami hanya tertarik pada catatan-catatan dimana suhu minimum Ithaca setidaknya 0°F. Pada Tabel A.1 ada 24 hari. Dari 24 hari ini, 14 menunjukkan presipitasi terukur (ppt), memberikan estimasi $\Pr\{\text{ppt} \geq 0.01\ inch | T_{\min} \geq 0°\text{F}\} = \frac{14}{24} \approx 0.58$. Data curah hujan selama tujuh hari dimana suhu minimum lebih dingin dari 0°F telah diabaikan. Karena presipitasi terukur tercatat hanya pada satu dari tujuh hari tersebut, kami dapat memperkirakan probabilitas presipitasi bersyarat mengingat peristiwa pengkondisian komplemen dari suhu minimum yang lebih dingin dari 0°F karena $\Pr\{\text{ppt} \geq 0.01\ inch | T_{min} < 0°F\} = 1/7 \approx 0.14$. Estimasi probabilitas presipitasi tanpa syarat yang sesuai adalah $\Pr\{\text{ppt} \geq 0.01\ inch\} = 15/31 \approx 0.48$.

Implementasi dalam program R

Sintaks:

```
> ## Membersihkan history ##
> rm(list = ls())
> ## Read data csv        ##
> data_TA1 <- read.csv('Data/Tabel A1.csv')[1:31,]
> ithaca   <- data_TA1[,2:4]
> n        <- dim(ithaca)[1]
> p        <- dim(ithaca)[2]
>
> ## PPT≥0.01 inch | T_min < 0°F##
> n_E1   <- 0
> n_tmin <- 0
> for (i in 1:n) {
+   # T_min  <0°F #
+   if(ithaca$Min.Temp.[i] < 0){
+     n_tmin <- n_tmin + 1
+     # PPT≥0.01 inch #
+     if(ithaca$Precipitation[i] >= 0.01){
```

```
+   }
+ }
>
> ## Sehingga Probabilitasnya ##
> Pr_E1 <- n_E1/n_tmin
> cat('PR{PPT ≥ 0.01 inch | T_min < 0°F} =',Pr_E1)
PR{PPT ≥ 0.01 inch | T_min < 0°F} = 0.1428571
>
> ## PPT ≥ 0.01 inch ##
> n_E2 <- 0
> for (i in 1:n) {
+   if(ithaca$Precipitation[i] >= 0.01){
+     n_E2 <- n_E2 + 1
+   }
+ }
```

Perbedaan dalam perkiraan probabilitas bersyarat yang dihitung dalam Contoh 2.4 mencerminkan ketergantungan statistik. Karena proses fisik yang

mendasarinya dipahami dengan baik, kami tidak akan terpengaruh untuk berspekulasi bahwa suhu minimum yang relatif lebih hangat entah bagaimana menyebabkan presipitasi. Sebaliknya, peristiwa suhu dan curah hujan menunjukkan hubungan statistik karena hubunganya secara fisik (berbeda) suhu dengan awan. Ketika berurusan dengan variabel dependen secara statistik yang hubungan fisiknya mungkin tidak diketahui, perlu diingat bahwa ketergantungan statistik tidak selalu menyiratkan hubungan sebab-akibat dari fisik.

Contoh 0.5 Persistensi sebagai probabilitas bersyarat

Variabel atmosfer sering menunjukkan hubungan statistik dengan nilai masa lalu atau masa depan mereka sendiri. Dalam terminologi sains atmosfer, ketergantungan sementara ini biasa disebut sebagai persistensi. Persistensi dapat didefinisikan sebagai adanya ketergantungan statistik (positif) antara nilai berurutan dari variabel yang sama atau antara kejadian berurutan dari suatu peristiwa tertentu. Ketergantungan positif artinya nilai variabel yang besar cenderung diikuti oleh nilai yang relatif besar, dan nilai variabel yang kecil cenderung diikuti oleh nilai yang relatif kecil. Sebagai aturan, ketergantungan waktu statistik dari variabel meteorologi adalah positif. Misalnya, kemungkinan besok suhu di atas rata-rata lebih tinggi jika di atas suhu rata-rata hari ini. Oleh karena itu, istilah lain untuk persistensi adalah ketergantungan serial positif. Saat ini, sifat umum ini memiliki implikasi penting untuk kesimpulan statistik yang diambil dari data atmosfer.

Pertimbangkan untuk mengkarakterisasi peristiwa persistensi {kejadian presipitasi} di Ithaca, sekali lagi dengan menggunakan kumpulan data kecil dari nilai harian pada Tabel A.1 Lampiran A. Secara fisik, ketergantungan seri dalam data ini diharapkan mengingat skala waktu yang khas untuk

Mid-latitude Sinoptik Gelombang yang diasosiasikan dengan sebagian besar presipitasi musim dingin di lokasi ini adalah beberapa hari yang lebih lama dari interval pengamatan harian. Konsekuensi statistiknya adalah bahwa hari-hari dengan curah hujan terukur yang dilaporkan akan lebih mungkin terjadi dalam jangka waktu tertentu, seperti halnya hari-hari tanpa curah hujan.

Untuk menilai ketergantungan serial kejadian curah hujan, perlu untuk memperkirakan probabilitas bersyarat jenis Pr{ppt hari ini | ppt kemarin}. Karena catatan A.1 tidak berisi catatan apa pun untuk 31 Desember 1986 atau 1 Februari 1987, ada 30 pasangan data kemarin/hari ini yang dapat digunakan. Untuk Mengestimasi Pr{ppt hari ini | ppt kemarin} kita hanya perlu menghitung jumlah hari curah hujan yang dilaporkan (sebagai peristiwa pengkondisian atau "kemarin") diikuti dengan hari berikutnya saat curah hujan dilaporkan (sebagai peristiwa yang menarik atau "hari ini"). Dalam memperkirakan probabilitas bersyarat ini, kami tidak tertarik pada apa yang terjadi setelah hari-hari ketika tidak ada curah hujan yang dilaporkan. Ada 14 hari dengan pengecualian 31 Januari ketika curah hujan dilaporkan. 10 di antaranya diikuti oleh hari lain dengan curah hujan yang dilaporkan, dan empat diikuti oleh hari kering. Estimasi frekuensi relatif bersyarat karenanya akan menjadi $\Pr\{\text{ppt } hari\ ini | \text{ppt } kemarin\} = \frac{10}{14} \approx 0.71$. Demikian pula, pengkondisian acara komplementer (tidak ada curah hujan "kemarin") memberikan $\Pr\{\text{ppt hari ini}|\text{tidak ppt kemarin}\} = \frac{5}{16} \approx 0.31$. Perbedaan antara perkiraan probabilitas bersyarat ini menunjukkan ketergantungan serial dalam data ini dan mengukur kecenderungan terjadinya hari basah dan kering dalam proses. Kedua probabilitas bersyarat ini juga membentuk "klimatologi bersyarat".

Implementasi dalam program R

```
Sintaks:
> data_TA1 <- read.csv('Data/Tabel A1.csv')[1:31,]
> ithaca   <- data_TA1[,2:4]
> n        <- dim(ithaca)[1] - 1 # Karena Pengecualian tgl 31 #
> n1_E1    <- 0
> n1_E2    <- 0
> for (i in 1:n) {
+   if(ithaca$Precipitation[i] >= 0.01){
+     n1_E1 <- n1_E1 + 1
+     if(ithaca$Min.Temp.[i] > 18){
+       n1_E2 <- n1_E2 + 1
+     }
+   }
+ }
> Pr1 = n1_E2 / n1_E1 ## Pr{ppt hari ini|ppt kemarin} #
> cat('Pr{ppt hari ini|ppt kemarin} =',Pr1)
Pr{ppt hari ini|ppt kemarin} = 0.7142857
> ## ppt hari ini||tidak ppt  kemarin ##
> n2_E1    <- 0
> n2_E2    <- 0
> for (i in 1:n) {
+   if(ithaca$Precipitation[i] < 0.01){
+     # Hari tidak hujan #
+     n2_E1 <- n2_E1 + 1
+     if(ithaca$Min.Temp.[i] > 18){
+       n2_E2 <- n2_E2 + 1
+     }
+   }
+ }
> ## Pr{ppt hari ini|tidak ppt  kemarin} #
> Pr2 = n2_E2 / n2_E1
> cat('Pr{ppt hari ini|tidak ppt  kemarin} =',Pr2)
Pr{ppt hari ini|tidak ppt  kemarin} = 0.31250
```

2.4.5 Hukum Probabilitas Total

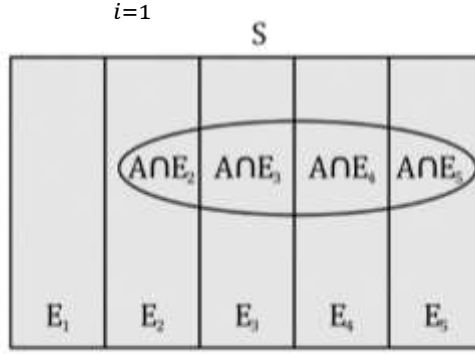
Terkadang probabilitas harus dihitung secara tidak langsung karena keterbatasan informasi. Satu hubungan yang berguna dalam situasi seperti itu adalah hukum probabilitas total. Pertimbangkan serangkaian kejadian MECE,

$\{E_i\}, i = 1 \ldots I$ pada ruang sampel yang diminati. Gambar 2.4 mengilustrasikan situasi ini untuk 15 kejadian. Jika ada kejadian $\{A\}$ yang juga ditentukan pada ruang sampel ini, probabilitasnya dapat dihitung dengan menjumlahkan probabilitas bersama

$$\Pr\{A\} = \sum_{i=1}^{I} \Pr\{A \cap E_i\} \qquad (0.14)$$

Notasi di sisi kanan, persamaan tersebut memberikan penjumlahan suku-suku yang ditentukan oleh sigma untuk semua nilai bilangan bulat indeks i antara 1 dan I. Sehingga dapat dituliskan kembali dalam persamaan sebagai berikut:

$$\Pr\{A\} = \sum_{i=1}^{I} \Pr\{A|E_i\} \Pr\{E_i\} \qquad (0.15)$$



Gambar 2.9. Ilustrasi hukum probabilitas total. Ruang sampel S berisi kejadian $\{A\}$, yang mewakili, dan lima kejadian MECE $\{Ei\}$.

Mengingat probabilitas tak bersyarat Pr{Ei} dan probabilitas bersyarat {A} mengingat peristiwa MECE {Ei}, probabilitas tak bersyarat {A} dapat dihitung. Penting untuk dicatat bahwa persamaan 2.15 benar hanya jika kejadian {Ei} mewakili partisi MECE dari ruang sampel.

Contoh 0.6 Menggabungkan probabilitas bersyarat menggunakan hukum probabilitas total

Contoh 2.5 juga dapat dilihat dari segi hukum probabilitas total. Perhatikan bahwa hanya ada $I = 2$ peristiwa MECE yang mempartisi ruang sampel: $\{E_1\}$ menunjukkan curah hujan kemarin dan $\{E_2\} = \{E_1\}^C$ menunjukkan tidak ada curah hujan kemarin. Peristiwa $\{A\}$ adalah peristiwa curah hujan hari ini. Jika data tidak tersedia, kita dapat menghitung $\Pr\{A\}$ menggunakan probabilitas bersyarat menurut hukum probabilitas total. Artinya, $\Pr\{A\} = \Pr\{A|E_1\}\Pr\{E_1\} + \Pr\{A|E_2\}\Pr\{E_2\} = \left(\frac{10}{14}\right)\left(\frac{14}{30}\right) + \left(\frac{5}{16}\right)\left(\frac{16}{30}\right) = 0.50$. Karena data tersedia pada Lampiran A, keakuratan hasil ini dapat dikonfirmasi atau di hitung.

Implementasi dalam program R

```
Sintaks (lanjutan Contoh 2.5):
> ## Pr{A|E_1 }Pr{E_1 }+Pr{A|E_2 }Pr{E_2 } ##
> PrT <- (Pr1 * (n1_E1/n)) + (Pr2 * (n1_E1/n))
> cat('Pr{A|E_1}Pr{E_1}+Pr{A|E_2}Pr{E_2} =',PrT,'atau',round(PrT,1))
Pr{A|E_1 }Pr{E_1 }+Pr{A|E_2 }Pr{E_2 } = 0.4791667 atau 0.5
```

2.4.6 Teorema Bayes

Teorema Bayes adalah kombinasi yang menarik dari hukum multiplikatif dan hukum probabilitas total. Dalam pengaturan frekuensi relatif, teorema Bayes digunakan untuk "membalikkan" probabilitas bersyarat. Artinya, jika $\Pr\{E_1|E_2\}$ diketahui, teorema Bayes dapat digunakan untuk menghitung $\Pr\{E_2|E_1\}$. Dalam kerangka Bayesian ini digunakan untuk merevisi atau memperbarui probabilitas subjektif sejalan dengan informasi baru.

Pertimbangkan lagi situasi seperti yang ditunjukkan pada Gambar 2.4, dimana ada serangkaian peristiwa MECE

$\{E_i\}$ dan peristiwa lain $\{A\}$. Hukum perkalian (Persamaan 2.11) dapat digunakan untuk menemukan dua ekspresi probabilitas bersama $\{A\}$ dan masing-masing peristiwa dari $\{E_i\}$.

$$\begin{aligned}\Pr\{A, E_i\} &= \Pr\{A|E_i\}\Pr\{E_i\} \\ &= \Pr\{E_i|A\}\Pr\{A\}\end{aligned} \qquad (0.16)$$

Menggabungkan dua sisi kanan sehingga dapat ditulis kembali menjadi:

$$\begin{aligned}\Pr\{E_i|A\} &= \frac{\Pr\{A|E_i\}\Pr\{E_i\}}{\Pr\{A\}} \\ &= \frac{\Pr\{A|E_i\}\Pr\{E_i\}}{\sum_{j=1}^{J}\Pr\{A|E_j\}\Pr\{E_j\}}\end{aligned} \qquad (0.17)$$

Hukum probabilitas total digunakan untuk menulis ulang penyebut. Persamaan 2.17 adalah ekspresi dari teorema Bayes. Ini berlaku secara terpisah untuk setiap kejadian MECE $\{E_i\}$. Perhatikan, bagaimanapun, bahwa penyebutnya sama, karena $\Pr\{A\}$ diperoleh setiap kali dengan menjumlahkan semua peristiwa yang diindeks oleh indeks *j* di penyebut.

Contoh 0.7 Teorema Bayes dari sudut pandang frekuensi relatif

Probabilitas bersyarat untuk curah hujan terjadi pada suhu minimum di atas atau di bawah 0°F diperkirakan dalam Contoh 2.4. Teorema Bayes dapat digunakan untuk menghitung probabilitas bersyarat terbalik yang terkait dengan peristiwa suhu, mengingat curah hujan telah atau belum terjadi. Misalkan $\{E_1\}$ adalah suhu minimum 0°F atau lebih dan $\{E_2\} = \{E_1\}^C$ kejadian komplemen bahwa suhu minimum lebih dingin dari0°F. Kedua peristiwa tersebut jelas membentuk partisi MECE dari ruang sampel. Perlu diingat

bahwa suhu minimum setidaknya 0°F dilaporkan pada pada tanggal 24 sampai dengan tanggal 31, sehingga perkiraan klimatologis tanpa syarat dari probabilitas kejadian suhu adalah $\Pr\{E_1\} = \frac{24}{31}$ dan $\Pr\{E_2\} = \frac{7}{31}$. Ingat juga $\Pr\{A|E_1\} = \frac{14}{24}$ dan $\Pr\{A|E_2\} = \frac{1}{7}$.

Persamaan 2.17 dapat diterapkan secara terpisah untuk masing-masing dua kejadian $\{E_i\}$. Penyebutnya adalah $\Pr\{A\} = \left(\frac{14}{24}\right)\left(\frac{24}{31}\right) + \left(\frac{1}{7}\right)\left(\frac{7}{31}\right) = \frac{15}{31}$. (Hal ini sedikit berbeda dari estimasi probabilitas presipitasi yang diperoleh pada Contoh 2.5, yang tidak dapat menjelaskan data 31 Desember). Dengan menggunakan teorema Bayes, probabilitas bersyarat untuk suhu minimal sekurang-kurangnya 0°F adalah pada kejadian Presipitasi $(14/24)(24/31)/(15/31)=14/15$. Demikian pula, probabilitas bersyarat suhu minimum di bawah 0°F dengan presipitasi bukan nol adalah $(1/7)(7/31)/(15/31)=1/15$. Karena semua data tersedia di Lampiran A, perhitungan ini dapat diperiksa secara langsung.

Implementasi dalam program R

```
Sintaks
> ## Membersihkan history ##
> rm(list = ls())
>
> ## suhu minimum setidaknya 0°F dilaporkan pada pada tanggal
24 sampai dengan  ##
> ## tanggal 31. Maka#
> E1 <- length(c(1:24))
> S  <- length(c(1:31))
> E2 <- S - E1
> Pr_E1 <- E1/S
> Pr_E2 <- E2/S
>
> ## Cari contoh sebelumbya 1 dan 14 kejadian, maka ##
> Pr_A.E1 <- 14/E1
> Pr_A.E2 <- 1/E2
>
> ## Sehingga dapat diperoleh Pr{A} ##
> Pr_A <- Pr_A.E1 * Pr_E1 + Pr_A.E2 * Pr_E2
>
> ## Output ##
> cat('Pr(E1)   =',E1,'/',S,'=',round(Pr_E1,6),
+     '\nPr(E2)   =',E2,'/',S,' =',round(Pr_E2,6),
+     '\nPr(A|E1) =',14,'/',E1,'=',round(Pr_A.E1,6),
+     '\nPr(A|E2) =',1,'/',E2,'  =',round(Pr_A.E2,6),
+     '\nPr(A)    =',14+1,'/',S,'=',round(Pr_A,6))
Pr(E1)   = 24 / 31 = 0.774194
Pr(E2)   = 7 / 31  = 0.225806
Pr(A|E1) = 14 / 24 = 0.583333
Pr(A|E2) = 1 / 7   = 0.142857
Pr(A)    = 15 / 31 = 0.483871
```

Contoh 0.8 Teorema Bayes dari sudut pandang probabilitas subyektif

Interpretasi probabilitas subyektif (Bayesian) dari Contoh 2.7 juga dapat dibuat. Misalkan prakiraan cuaca diinginkan yang memberikan probabilitas bahwa suhu minimum setidaknya 0°F. Jika informasi yang lebih canggih

tidak tersedia, wajar jika menggunakan probabilitas klimatologis tanpa syarat untuk peristiwa tersebut, $\Pr\{E_1\} = \frac{24}{31}$, untuk mewakili ketidakpastian atau tingkat kepercayaan peramal terhadap hasilnya. Dalam sistem Bayesian, keadaan dasar informasi ini dikenal sebagai probabilitas sebelumnya. Namun, misalkan peramal dapat mengetahui apakah akan ada curah hujan atau tidak pada hari itu. Informasi ini akan mempengaruhi tingkat kepastian peramal tentang hasil suhu. Seberapa percaya diri peramal tergantung pada kekuatan hubungan antara suhu dan curah hujan, yang dinyatakan dalam probabilitas bersyarat untuk terjadinya curah hujan pada dua hasil suhu minimum. Probabilitas bersyarat ini, $\Pr\{A|E_i\}$ dalam notasi contoh ini, dikenal sebagai probabilitas. Jika curah hujan terjadi, peramal lebih yakin bahwa suhu minimum setidaknya 0°F, dengan probabilitas yang direvisi diberikan oleh Persamaan 2.17 sebagai $(14/24)(24/31)/(15/31)$ $=14/15$. Penilaian probabilitas bahwa suhu minimum yang sangat dingin tidak akan terjadi (diberikan informasi tambahan mengenai terjadinya presipitasi) yang dimodifikasi atau diperbarui disebut probabilitas posterior. Di sini probabilitas posterior lebih besar dari probabilitas sebelumnya yaitu $24/31$. Demikian pula, peramal lebih yakin bahwa suhu minimum tidak akan menjadi 0°F atau lebih hangat jika tidak ada curah hujan. Perhatikan bahwa perbedaan antara contoh ini dan Contoh 2.4 adalah murni salah satu interpretasi dan hasil perhitungan numerik dan identik.

2.5 Latihan

1. Dalam catatan iklim untuk 60 musim dingin di lokasi tertentu, sembilan dari musim dingin tersebut memiliki hujan salju badai tunggal lebih besar dari 35 cm (menentukan hujan salju tersebut sebagai peristiwa "A"), dan suhu terendah di bawah 25 °C dalam 36 musim dingin (didefinisikan sebagai peristiwa "B"). Kedua peristiwa "A" dan "B" terjadi dalam tiga musim dingin.
   a. Buat sketsa diagram Venn untuk ruang sampel yang sesuai dengan data tersebut.
   b. Tulis pernyataan dalam notasi kuantitas untuk terjadinya hujan salju setebal 35 cm, suhu 25°C, atau keduanya. Perkirakan probabilitas klimatologis untuk peristiwa komposit ini.
   c. Tuliskan ekspresi dalam notasi himpunan untuk terjadinya musim dingin dengan curah salju setinggi 35 cm ketika suhu tidak turun di bawah 25°C. Perkirakan probabilitas klimatologis untuk peristiwa komposit ini.
   d. Tulis ekspresi dalam notasi himpunan untuk kejadian musim dingin tanpa suhu 25°C atau hujan salju 35 cm. Perkirakan probabilitas klimatologis lagi.
2. Menggunakan dataset Januari 1987 pada Tabel A.1, tentukan kejadian peristiwa "A" sebagai Ithaca dengan $T_{max} > 32$°F dan kejadian "B" sebagai Canandaigua $T_{max} > 32$°F.
   a. Jelaskan maksud dari $\Pr(A)$, $\Pr(B)$, $\Pr(A, B)$, $\Pr(A \cup B)$, $\Pr(A|B)$ dan $\Pr(B|A)$.
   b. Dari frekuensi relatif dalam data, perkirakan $\Pr(A)$, $\Pr(B)$ dan $\Pr(A|B)$.
   c. Dengan menggunakan hasil dari bagian (b), hitunglah $\Pr(B|A)$.
   d. Apakah peristiwa "A" dan "B" saling bebas?

3. Sekali lagi dengan menggunakan data pada Tabel A.1, perkirakan probabilitas suhu maksimum Ithaca akan berada pada atau di bawah titik beku (32°F), mengingat suhu maksimum hari sebelumnya berada pada atau di bawah titik beku.
   a. Perhitungkan kegigihan dalam data suhu.
   b. Di bawah asumsi (salah) bahwa urutan suhu harian adalah independen.
4. Tiga radar, bekerja secara independen, mencari gema "pengait" (penanda radar yang terkait dengan tornado). Misalkan ada radar yang memiliki probabilitas 0,05 untuk tidak mendeteksi tanda tangan ini saat ada tornado.
   a. Buat sketsa diagram Venn untuk contoh ruang yang cocok untuk soal ini.
   b. Berapa peluang tornado tidak terdeteksi oleh ketiga radar?
   c. Berapa probabilitas tornado akan terdeteksi oleh ketiga radar?
5. Efek penyemaian awan dalam menekan hujan es yang berbahaya dipelajari di suatu daerah dengan menyemai atau tidak menyemai calon badai dalam jumlah yang sama secara acak atau tidak. Misalkan probabilitas hujan es yang merusak dari badai dengan biji adalah 0,10 dan probabilitas hujan es yang merusak dari badai tanpa biji adalah 0,40. Jika salah satu kandidat badai baru saja menghasilkan hujan es yang merusak, seberapa besar peluangnya untuk terjadinya hujan es yang tidak merusak?

BAB 3
DISTRIBUSI EMPIRIS DAN ANALISIS DATA EKSPLORASI

3.1 Penduluan

Salah satu penerapan ide statistik yang sangat penting dalam meteorologi dan klimatologi adalah memahami kumpulan data baru. Seperti yang disebutkan dalam BAB 1, sistem pengamatan meteorologi dan model numerik, yang mendukung upaya operasional dan penelitian, menghasilkan banyak sekali data numerik. Hal ini bisa menjadi tugas yang signifikan hanya untuk merasakan sekumpulan angka baru, dan mulai memahaminya. Tujuannya adalah untuk mengekstraksi wawasan tentang proses yang mendasari pembangkitan angka.

Secara garis besar kegiatan ini dikenal dengan *Exploratory Data Analysis* (Analisis Data Eksplorasi) atau EDA. Penggunaan sistematisnya meningkat secara substansial mengikuti buku terobosan Tukey (1977). Metode EDA banyak memanfaatkan berbagai metode grafis untuk membantu pemahaman pada banyaknya angka yang dihadapi seorang analis. Grafik adalah cara yang sangat efektif untuk menggambarkan dan meringkas data, menggambarkan banyak hal dalam ruang yang kecil, dan menampilkan fitur yang tidak biasa dari kumpulan data. Fitur yang tidak biasa bisa menjadi sangat penting. Kadang-kadang poin data yang tidak biasa dihasilkan dari kesalahan dalam perekaman atau transkripsi, dan hal ini sebaiknya diketahui sedini mungkin dalam analisis. Terkadang data yang tidak biasa itu valid dan

bisa menjadi bagian yang paling menarik dan informatif dari kumpulan data.

Banyak metode EDA awalnya dirancang untuk diterapkan secara manual, dengan pensil dan kertas, pada kumpulan data kecil. Paket komputer berorientasi grafis telah muncul yang memungkinkan penggunaan metode ini dengan cepat dan mudah pada komputer desktop. Metode ini juga dapat diimplementasikan pada komputer yang lebih besar dengan jumlah pemrograman yang sederhana.

3.1.1 Robustness dan Resistensi

Banyak teknik klasik statistik bekerja dengan baik ketika asumsi yang cukup ketat tentang sifat data terpenuhi. Sebagai contoh, sering diasumsikan bahwa data akan mengikuti kurva berbentuk lonceng dari distribusi Gaussian. Prosedur klasik dapat berperilaku sangat buruk (yaitu, menghasilkan hasil yang menyesatkan) jika asumsi tidak dipenuhi.

Asumsi statistik klasik tidak dibuat karena ketidaktahuan, melainkan karena kebutuhan. Permintaan penyederhanaan asumsi dalam statistik, seperti di bidang lain, telah memungkinkan kemajuan dibuat melalui derivasi hasil analitik yang relatif sederhana namun rumus matematika yang kuat. Seperti yang telah terjadi di banyak bidang kuantitatif, munculnya daya komputasi yang relatif baru telah membebaskan peneliti data dari ketergantungan tunggal pada hasil tersebut, dengan memungkinkan alternatif yang membutuhkan asumsi yang tidak terlalu ketat menjadi praktis. Hal ini tidak berarti bahwa metode klasik tidak lagi berguna. Namun, jauh lebih mudah untuk memeriksa bahwa kumpulan data yang diberikan memenuhi asumsi tertentu sebelum prosedur klasik digunakan, dan alternatif yang baik layak secara komputasi dalam kasus dimana metode klasik mungkin tidak sesuai.

Dua sifat penting dari metode EDA adalah robust dan resisten. Robustness dan resistensi adalah dua aspek ketidakpekaan terhadap asumsi tentang sifat dari sekumpulan data. Metode yang kuat belum tentu optimal dalam keadaan tertentu, tetapi memiliki kinerja cukup baik di sebagian besar keadaan. Misalnya, rata-rata sampel adalah karakterisasi terbaik dari pusat suatu kumpulan data jika diketahui bahwa data tersebut mengikuti distribusi Gaussian. Namun, jika data tersebut jelas non-Gaussian (misalnya, jika merupakan catatan peristiwa curah hujan ekstrem), rata-rata sampel akan menghasilkan karakterisasi yang menyesatkan dari pusatnya. Sebaliknya, metode yang kuat umumnya tidak peka terhadap asumsi tertentu tentang sifat keseluruhan data. Metode yang resisten tidak terlalu dipengaruhi oleh sejumlah kecil outlier. Seperti yang ditunjukkan sebelumnya, titik-titik tersebut sering muncul dalam kumpulan data melalui kesalahan jenis satu atau lainnya. Hasil metode resisten sedikit berubah jika sebagian kecil dari nilai data diubah, bahkan jika diubah secara drastis. Selain tidak robust, rata-rata sampel juga bukan karakterisasi resisten dari pusat kumpulan data. Perhatikan himpunan kecil {11, 12, 13, 14, 15, 16, 17, 18, 19}. Rata-ratanya adalah 15. Namun, jika himpunan {11, 12, 13, 14, 15, 16, 17, 18, 91} yang dihasilkan dari kesalahan penulisan, “pusat” data (secara keliru) yang dicirikan menggunakan sampel, rata-rata malah akan menjadi 23. Ukuran resisten dari pusat kumpulan data, seperti yang akan disajikan nanti, akan diubah sedikit atau tidak sama sekali dengan mengganti “91” dengan “19” dalam contoh sederhana tadi.

3.1.2 Kuantil (Quantiles)

Banyak ukuran ringkasan umum bergantung pada penggunaan kuantil sampel yang dipilih (juga dikenal sebagai fraktil). Kuantil dan fraktil pada dasarnya setara dengan

istilah yang lebih dikenal dengan persentil. Kuantil sampel q_p adalah angka yang memiliki satuan yang sama dengan data, yang melebihi proporsi data yang diberikan oleh subskrip p, dengan $0 \leq p \leq 1$. Kuantil sampel q_p dapat ditafsirkan seperti nilai yang diharapkan melebihi anggota kumpulan data yang dipilih secara acak, dengan probabilitas p. Secara ekuivalen, sampel kuantil q_p akan dianggap sebagai persentil ke $(p \times 100)$ dari kumpulan data.

Penentuan kuantil sampel mensyaratkan bahwa kumpulan data terlebih dahulu diatur secara berurutan. Menyortir kumpulan data kecil secara manual tidak memberikan banyak masalah. Menyortir kumpulan data yang lebih besar baik dilakukan oleh komputer. Secara historis, langkah penyortiran merupakan hambatan utama dalam penerapan prosedur yang robust dan resisten terhadap kumpulan data yang besar. Penyortiran saat ini dapat dilakukan dengan mudah menggunakan spreadsheet atau program analisis data pada komputer desktop, atau salah satu dari banyak algoritma penyortiran yang tersedia dalam kumpulan rutinitas komputasi tujuan umum (Press et al., 1986).

Nilai data yang terurut, atau diberi peringkat, dari sampel tertentu disebut statistik urutan (*order statistics*) sampel. Diberikan kumpulan data $\{x_1, x_2, x_3, \dots, x_n\}$, statistik urutan untuk sampel ini akan menjadi angka yang sama, namun diurutkan dalam urutan meningkat (*ascending*). Nilai-nilai yang telah diurutkan ini secara konvensional dilambangkan menggunakan subskrip tanda kurung, yaitu, dengan himpunan $\{x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}\}$. Di sini, nilai data terkecil ke-i dilambangkan dengan $x_{(i)}$.

Kuantil sampel tertentu sering digunakan dalam ringkasan eksplorasi data. Yang paling umum digunakan adalah median, atau $q_{0.5}$, atau persentil ke-50. Nilai ini adalah

nilai tengah dari kumpulan data yang telah diurutkan, yang artinya bahwa proporsi data yang sama berada di atas dan di bawah nilainya. Jika kumpulan data berukuran ganjil, maka median hanyalah nilai tengah statistik urutan. Namun, jika data berukuran genap, maka kumpulan data memiliki dua nilai tengah. Dalam hal ini, median secara konvensional diambil sebagai rata-rata dari dua nilai tengah tersebut. Secara umum,

$$q_{0.5} = \begin{cases} x_{\left(\frac{[n+1]}{2}\right)}, & n \text{ ganjil} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\left[\frac{n}{2}\right]+1\right)}}{2}, & n \text{ genap} \end{cases} \qquad (0.1)$$

Umumnya, selain median, yang sering digunakan adalah kuartil, $q_{0.25}$ dan $q_{0.75}$. Biasanya, nilai ini masing-masing disebut sebagai kuartil bawah dan atas. Keduanya terletak di tengah-tengah antara median, $q_{0.5}$, dan nilai ekstrem, $x_{(1)}$ dan $x_{(n)}$. Tukey (1977) menyebutkan bahwa $q_{0.25}$ dan $q_{0.75}$ sebagai "engsel", dimana kumpulan data dilipat pertama kali berdasarkan median, dan kemudian dilipat berdasarkan kuartil $q_{0.25}$ dan $q_{0.75}$. Dengan demikian kuartil $q_{0.25}$ dan $q_{0.75}$ adalah dua median dari setengah kumpulan data antara $q_{0.5}$ dan nilai ekstrem. Jika n ganjil, setengah kumpulan data ini masing-masing terdiri dari $\left(\frac{n+1}{2}\right)$ titik termasuk median. Jika n genap, setengah kumpulan data ini masing-masing berisi $\frac{n}{2}$ titik dan tidak tumpang tindih.

Contoh 0.1 Perhitungan Kuantil Secara Umum

Jika terdapat $n = 9$ nilai dalam satu kumpulan data, mediannya adalah $q_{0.5} = x_{(5)}$ atau nilai terbesar ke-5 dari 9 nilai dalam data. Kuartil bawahnya adalah $q_{0.25} = x_{(3)}$, dan kuartil atasnya adalah $q_{0.75} = x_{(7)}$.

Jika $n = 10$, mediannya adalah rata-rata dari dua nilai tengah, $q_{0.25}$ dan $q_{0.75}$ adalah nilai tengah tunggal dari bagian atas dan

bawah data yang dibagi berdasarkan median. Secara matematis, $q_{0.25}$, $q_{0.5}$ dan $q_{0.75}$ adalah $x_{(3)}$, $\frac{[x_{(5)}+x_{(6)}]}{2}$, dan $x_{(8)}$ secara berurutan.

Jika $n = 11$ maka terdapat nilai tengah tunggal, namun kuartilnya ($q_{0.25}$ dan $q_{0.75}$) ditentukan dengan merata-ratakan dua nilai tengah dari bagian atas dan bawah dari data. Artinya, $q_{0.25}$, $q_{0.5}$ dan $q_{0.75}$ adalah $\frac{[x_{(3)}+x_{(4)}]}{2}$, $x_{(6)}$, dan $\frac{[x_{(8)}+x_{(9)}]}{2}$ secara berurutan.

Untuk $n = 12$, baik kuartil maupun median ditentukan dengan rata-rata pasangan nilai tengah yang artinya $q_{0.25}$, $q_{0.5}$ dan $q_{0.75}$ adalah $\frac{[x_{(3)}+x_{(4)}]}{2}$, $\frac{[x_{(6)}+x_{(7)}]}{2}$, dan $\frac{[x_{(9)}+x_{(10)}]}{2}$ secara berurutan.

3.2 Ukuran Ringkasan Numerik

Terdapat berapa ukuran ringkasan sederhana yang robust dan resisten yang dapat digunakan tanpa harus membuat plot secara manual atau memanfaatkan kemampuan grafis komputer. Hal ini seringkali menjadi ukuran awal yang dihitung dari kumpulan data. Ringkasan numerik yang tercantum dalam bagian ini dapat dibagi lagi menjadi ukuran tendensi, sebaran, dan simetri. Tendensi mengacu pada kecenderungan pusat atau besaran umum dari nilai dalam data. Sebaran menunjukkan tingkat variasi atau dispersi di sekitar nilai pusat. Simetri menggambarkan keseimbangan nilai data terdistribusi di sekitar pusatnya. Data asimetris cenderung lebih menyebar baik di sisi atas (berekor panjang di sisi kanan), maupun di sisi rendah (berekor panjang di sisi kiri). Ketiga jenis ukuran ringkasan numerik ini sesuai dengan tiga momen statistik pertama dari sampel data, namun ukuran klasik dari momen-momen ini (yaitu rata-rata sampel, varians sampel, dan koefisien

kemiringan sampel, masing-masing) tidak robust atau resisten.

3.2.1 Pusat Tendensi

Ukuran pusat tendensi yang robust dan resisten yang paling umum adalah median, $q_{0.5}$. Perhatikan kembali kumpulan data {11, 12, 13, 14, 15, 16, 17, 18, 19}. Median dan rata-rata memiliki nilai yang sama yaitu 15. Jika, seperti disebutkan sebelumnya, "19" diganti secara keliru dengan "91", dengan menggunakan rumus rata-rata

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{0.2}$$

diperoleh nilai rata-rata yaitu 23 dimana nilainya berubah drastis yang menggambarkan kurangnya resistensi terhadap outlier. Median tidak berubah karena jenis kesalahan data seperti ini.

Ukuran pusat tendensi yang sedikit lebih rumit yang memperhitungkan lebih banyak informasi tentang besaran data adalah trimean. Trimean adalah rata-rata tertimbang dari median dan kuartil, dengan median menerima dua kali bobot masing-masing kuartil:

$$\text{Trimean} = \frac{q_{0.25} + 2(q_{0.5}) + q_{0.75}}{4} \tag{0.3}$$

Rata-rata terpangkas (*trimmed mean*) adalah ukuran pusat tendensi yang resisten, yang kepekaannya terhadap outlier dikurangi dengan menghilangkan proporsi tertentu dari pengamatan terbesar dan terkecil. Jika proporsi pengamatan yang dihilangkan pada setiap ekor dalam data adalah α, maka

$$\bar{x}_\alpha = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)} \tag{0.4}$$

dimana k adalah pembulatan bilangan bulat dari $\alpha \times n$, yaitu jumlah nilai data yang "dipangkas" dari setiap ekor

dalam data. Rata-rata terpangkas direduksi menjadi rata-rata biasa (Persamaan (0.2)) jika $\alpha = 0$. Metode lain untuk mengkarakterisasi pusat tendensi dapat ditemukan di Andrews et al. (1972), Goodall (1983), Rosenberger dan Gasko (1983), dan Tukey (1977).

3.2.2 Sebaran

Ukuran sebaran yang paling umum, sederhana, robust dan resisten, yang juga dikenal sebagai dispersi atau skala, adalah Interquartile Range (IQR). Rentang Interkuartil adalah perbedaan antara kuartil atas dan bawah:

$$IQR = q_{0.75} - q_{0.25} \tag{0.5}$$

IQR adalah indeks sebaran yang baik karena IQR hanya memperhitungkan sebaran 50% data di sekitar titik pusat data. Fakta bahwa IQR mengabaikan 25% data bagian atas dan bawah membuatnya cukup resisten terhadap outlier. Penting untuk membandingkan IQR dengan simpangan baku sampel

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{0.6}$$

Nilai kuadrat simpangan baku sampel, s^2, dikenal sebagai varians sampel. Simpangan baku tidak robust atau resisten. Nilai simpangan baku hampir hanya akar kuadrat dari rata-rata perbedaan kuadrat antara titik data dan rata-rata sampelnya. Pembagian dengan $n-1$ sering dilakukan sebagai kompensasi bahwa x_i lebih dekat dengan rata-rata sampelnya dibandingkan rata-rata populasi sebenarnya: membagi dengan $n-1$ secara tepat berlawanan dengan tendensi yang dihasilkan untuk simpangan baku sampel terlalu kecil. Dengan adanya akar kuadrat dalam Persamaan (0.6), simpangan baku memiliki dimensi fisik yang sama

dengan data yang mendasarinya. Satu nilai data yang sangat besar akan membuat simpangan baku bernilai besar karena akan sangat jauh dari rata-rata. Pertimbangkan lagi himpunan {11, 12, 13, 14, 15, 16, 17, 18, 19}. Simpangan baku sampel adalah 2.74, namun berubah menjadi sangat tinggi yaitu 25.6 apabila "91" secara keliru menggantikan "19". Namun, sangat mudah untuk melihat bahwa dalam kedua kasus memiliki yang sama, $IQR = 4$.

IQR sangat mudah untuk dihitung, namun memiliki kelemahan karena tidak menggunakan sebagian besar data. Alternatif yang lebih lengkap adalah *Median Absolute Deviation* (MAD). MAD paling mudah dipahami dengan membayangkan transformasi $y_i = |x_i - q_{0.5}|$. Setiap nilai yang diubah menjadi y_i adalah nilai absolut dari selisih antara nilai data asli yang sesuai dan median, dan MAD adalah median dari kumpulan data y_i:

$$\text{MAD} = \text{median}(|x_i - q_{0.5}|) \tag{0.7}$$

Meskipun proses ini mungkin tampak agak rumit pada awalnya, hanya sedikit yang menggambarkan bahwa hal ini adalah analogi perhitungan simpangan baku, namun menggunakan operasi yang tidak melibatkan data outlier. Setiap nilai data dikurangi dengan median dan tanda negatif apa pun dihilangkan dengan operasi nilai absolut (bukan kuadrat), dan titik pusat dari selisih absolut ini terletak pada mediannya (bukan rata-rata).

Ukuran sebaran yang lebih rumit adalah varians terpangkas (*trimmed variance*). Idenya berasal dari rata-rata terpangkas (Persamaan (0.4)) yang menghilangkan nilai terbesar dan terkecil dengan proporsi tertentu dan menghitung varians sampel (nilai kuadrat dari Persamaan (0.6)).

$$s_\alpha^2 = \sqrt{\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left(x_{(i)} - \bar{x}_\alpha\right)^2} \tag{0.8}$$

Sekali lagi, k adalah bilangan bulat terdekat dengan $\alpha \times n$. Varians terpangkas terkadang dikalikan dengan faktor penyesuaian agar lebih konsisten dengan varians sampel biasa (Graedel dan Kleiner, 1985).

3.2.3 Simetri

Ukuran simetri berbasis momen konvensional dalam kumpulan data adalah koefisien kemiringan sampel,

$$\gamma = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} \tag{0.9}$$

Ukuran ini tidak robust atau resisten. Dengan demikian koefisien kemiringan sampel bahkan lebih sensitif terhadap outlier dibanding simpangan baku. Rata-rata deviasi pangkat tiga dalam pembilang dibagi dengan pangkat tiga dari simpangan baku sampel untuk membakukan (standarisasi) koefisien kemiringan sehingga perbandingan kemiringan antara kumpulan data yang berbeda dapat terlihat lebih nampak. Standarisasi juga berfungsi untuk membuat kemiringan sampel menjadi kuantitas tanpa dimensi.

Perhatikan bahwa selisih pangkat tiga antara nilai data dan rata-ratanya mempertahankan tandanya. Karena selisihnya berpangkat tiga, nilai data yang terjauh dari rata-rata akan mendominasi penjumlahan pada pembilang Persamaan (0.9). Jika terdapat beberapa nilai data yang sangat besar, maka kemiringan sampel akan cenderung positif. Oleh karena itu, kumpulan data dengan ekor panjang kanan disebut dengan miring kanan atau miring positif. Data yang dibatasi secara fisik untuk berada di atas nilai minimum

(seperti curah hujan atau kecepatan angin, dimana keduanya harus nonnegatif) seringkali miring positif (*positively skewed*). Sebaliknya, jika terdapat beberapa nilai data yang sangat kecil (atau besar negatif), maka nilainya akan jauh di bawah rata-rata. Jumlah dalam pembilang Persamaan (0.9) kemudian akan didominasi oleh beberapa nilai besar negatif sehingga koefisien kemiringan akan menjadi negatif. Data dengan ekor panjang kiri disebut miring kiri atau miring negatif. Untuk data yang pada dasarnya simetris memiliki koefisien kemiringan yang mendekati nol.

Alternatif yang robust dan resisten untuk kemiringan sampel adalah indeks Yule-Kendall

$$\gamma_{YK} = \frac{(q_{0.75} - q_{0.5}) - (q_{0.5} - q_{0.25})}{IQ} = \frac{q_{0.25} - 2(q_{05}) + q_{0.75}}{IQ} \quad (0.10)$$

Yang dihitung dengan membandingkan jarak antara median dan masing-masing kuartil. Jika data miring ke kanan, setidaknya 50% di sekitar pusat data, maka jarak median antara kuartil atas akan lebih besar dari jarak median dengan kuartil bawah. Dalam hal ini, indeks Yule-Kendall akan lebih besar dari nol, konsisten dengan konvensi biasa tentang kemiringan ke kanan (positif). Sebaliknya, data miring ke kiri akan dicirikan oleh indeks Yule-Kendall yang negatif.

3.3 Teknik Ringkasan Grafik

Pengukuran ringkasan numerik cepat dan mudah dihitung dan ditampilkan, namun hanya dapat mengungkapkan sejumlah kecil detail dari data. Selain itu, dampak visualnya terbatas. Sejumlah tampilan grafis untuk analisis data eksplorasi telah dirancang yang hanya memerlukan sedikit usaha lebih untuk diproduksi.

3.3.1 Stem-and-Leaf

Stem-and-leaf adalah alat yang sangat sederhana namun efektif untuk menghasilkan tampilan keseluruhan dari kumpulan data. Pada saat yang sama memberikan peneliti paparan awal untuk data individual. Dalam bentuk paling sederhana, tampilan *stem-and-leaf* mengelompokkan nilai data sesuai dengan digitnya. Nilai-nilai ini ditulis dalam urutan meningkat atau menurun pada sebelah kiri batang vertikal sebagai "s*tem*". Digit paling mendekati untuk setiap nilai data kemudian ditulis di sebelah kanan secara vertikal, pada baris yang sama dengan digit yang lebih mendekati. Nilai yang paling tidak mendekati merupakan "*leaf*".

Pada Gambar 3.1 (a) menunjukkan tampilan *stem-and-leaf* untuk suhu maksimum Ithaca Januari 1987 pada Tabel A.1. Nilai data dilaporkan berkisar dari 9°F hingga 53°F. Tampilan dibangun dengan menelusuri nilai data satu per satu dan menulis digit terkecilnya pada baris yang sesuai. Misalnya, suhu untuk 1 Januari adalah 33°F, maka "*stem*" pertama adalah "3" pada sisi kiri dan "*leaf*" pertama adalah "3" pada sisi kanan. Suhu untuk 2 Januari adalah 32°F, sehingga "2" ditulis pada "*leaf*" di "*stem*" yang sesuai yaitu "3".

Tampilan *stem-and-leaf* awal untuk kumpulan data ini agak ramai karena sebagian besar nilainya berada pada nilai 20-an dan 30-an. Dalam kasus seperti ini, resolusi yang lebih baik dapat diperoleh dengan membuat plot kedua, seperti pada Gambar 3.1 (b), dimana setiap batang telah dibelah untuk memuat nilai 0–4 dan 5–9.

(a)

```
0 | 9
1 | 7
2 | 2455666778899
3 | 00002223344677
4 | 5
5 | 3
```

(b)

```
0 | 9
1 |
1 | 7
2 | 24
2 | 55666778899
3 | 00002223344
3 | 677
4 |
4 | 5
5 | 3
```

Gambar 3.1. *Stem-and-leaf* untuk suhu maksimum Ithaca Januari 1987 pada Tabel A.1. Plot pada panel (a) adalah hasil setelah melewati data pertama, menggunakan 10 sebagai nilai "*stem*". Resolusi yang sedikit lebih tinggi pada panel (b) dengan membuat batang terpisah yaitu dari 0 hingga 4 dan dari 5 hingga 9.

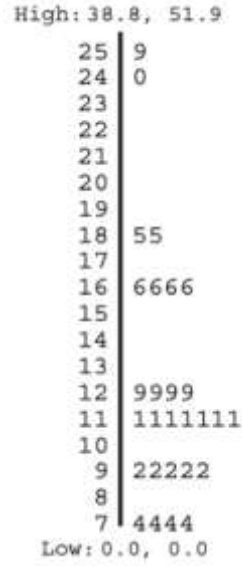
Implementasi dalam program R

Sintaks:

```
> ## Membersihkan history ##
> rm(list = ls())
> ## Read data csv        ##
> data_TA1 <- read.csv('Data/Tabel A1.csv')[1:31,]
> ithaca   <- data_TA1[,1:3]
> ## Plot a ##
> duration = ithaca$Max.Temp.
> stem(duration, scale = 0.5)

The decimal point is 1 digit(s) to the right of the |

  0 | 9
  1 | 7
  2 | 2455666778899
  3 | 00002223344677
  4 | 5
  5 | 3
>
```

Tampilan *stem-and-leaf* sangat mirip dengan histogram. Pada Gambar 3.1, misalnya, terlihat jelas bahwa data suhu cukup simetris, dengan sebagian besar nilai jatuh pada 20-an atas dan 30-an bawah. Menyortir nilai daun juga memperlihatkan kuantil yang diiginkan. Dalam kasus ini, tidak sulit untuk menentukan mediannya yaitu 30 dan kedua kuartilnya ($q_{0.25}$ dan $q_{0.75}$) adalah 26 dan 33.

```
> ## plot b ##

> duration = ithaca$Max.Temp.
> stem(duration)
  The decimal point is 1 digit(s) to the right of the |

  0 | 9
  1 |
  1 | 7
  2 | 24
  2 | 55666778899
  3 | 00002223344
  3 | 677
  4 |
  4 | 5
  5 | 3

>
```



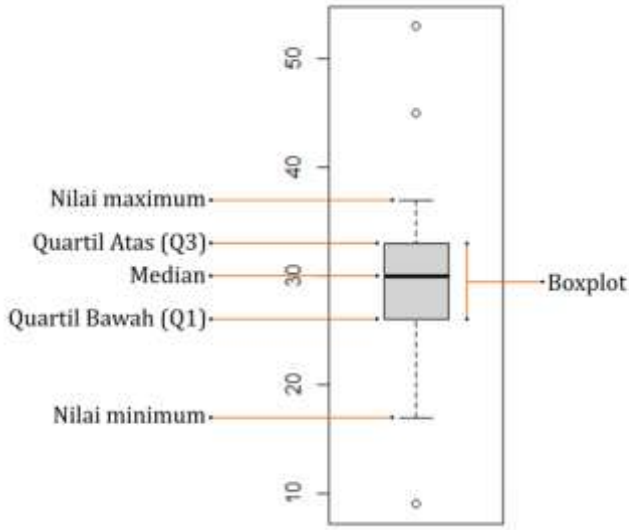
Gambar 3.2. *Stem-and-leaf* dari kecepatam angin (km/jam) pada jam 01:00 AM di Newark, New Jersey, Airport selama Desember 1974. Nilai yang sangat tinggi dan rendah dituliskan pada bagian luar plot untuk menghindari banyaknya batang yang kosong.

Selain itu, memungkinkan bahwa ada satu atau lebih titik data yang terletak jauh dari bagian utamanya. Daripada memplot banyak batang kosong, biasanya lebih mudah untuk mencantumkan nilai ekstrem ini secara terpisah di ujung atas dan bawah plot, seperti pada Gambar 3.2, berupa data yang diambil dari Graedel dan Kleiner (1985), kecepatan angin dalam kilometer per jam (km/jam). Dengan menuliskan dua nilai yang sangat besar dan dua nilai angin tenang di bagian atas dan bawah plot akan mengurangi panjang tampilan hingga lebih dari setengahnya. Dalam plot terlihat bahwa data sangat miring ke kanan.

Tampilan *stem-and-leaf* pada Gambar 3.2 juga memperlihatkan sesuatu yang mungkin terlewatkan dalam daftar tabel. Semua nilai daun pada setiap batang adalah sama. Sehingga, timbul kecurigaan bahwa proses pembulatan telah diterapkan pada data.

3.3.2 Boxplot

Boxplot atau *box-and-whisker* plot, adalah grafis yang sering digunakan yang diperkenalkan oleh Tukey (1977). Plot ini adalah plot sederhana karena hanya memperhitungkan lima kuantil sampel yaitu nilai minimum $x_{(1)}$, kuartil bawah $q_{0.25}$, median $q_{0.5}$, kuartil atas $q_{0.75}$, dan maksimum $x_{(n)}$. Dengan menggunakan lima nilai ini, boxplot pada dasarnya menyajikan sketsa distribusi data.



Gambar 3.3. Boxplot sederhana atau *box-and-whiskers* untuk data temperatur Ithaca maksimum pada bulan Januari 1987. Ujung atas dan bawah dari kotak (*box*) digambar pada titik kuartil dan batang dalam *box* digambar pada titik median. Kumis

(*whiskers*) digambar memanjang dari kuartil sampai nilai data maksimum dan minimum.

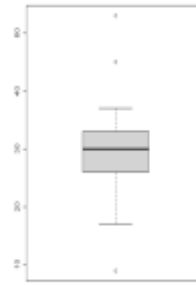
Pada Gambar 3.3 menunjukkan boxplot untuk data suhu maksimum Ithaca Januari 1987 pada Tabel A.1. Kotak di tengah diagram dibatasi oleh kuartil atas dan bawah yang merepresentasikan 50% data. Garis dalam kotak merepresentasikan median dan kumis (*whiskers*) memanjang dari kotak ke dua nilai ekstrim.

Implementasi dalam program R

Sintaks:

```
> ## Membersihkan history ##
> rm(list = ls())

> ## Read data csv ##
> data_TA1 <- read.csv('Data/Tabel A1.csv')
> ithaca   <- data_TA1[1:31, 1:3]
> ## boxplot ##
> boxplot(ithaca$Max.Temp.)
```

Output:



Boxplots dapat menyampaikan informasi dalam jumlah yang sangat besar secara sekilas. Pada Gambar 3.3, terlihat bahwa data terkonsentrasi cukup dekat dengan nilai

$30°F$. Karena hanya didasarkan pada median dan kuartil, bagian boxplot sangat resisten terhadap outlier yang mungkin ada. Kisaran lengkap data juga terlihat sekilas. Terakhir, terlihat bahwa data ini hampir simetris, karena mediannya berada dekat dengan bagian tengah kotak dan kumisnya memiliki panjang yang sebanding.

3.3.3 Plot Skematik

Kelemahan dari boxplot adalah informasi tentang ekor data sangat digeneralisasi. Kumis (whiskers) meluas ke nilai tertinggi dan terendah, namun tidak ada informasi tentang distribusi titik data dalam kuartil atas dan bawah data. Sebagai contoh, meskipun Gambar 3.3 menunjukkan bahwa suhu maksimum tertinggi adalah $53°F$, tidak memberikan informasi apakah nilai tersebut merupakan titik terisolasi.

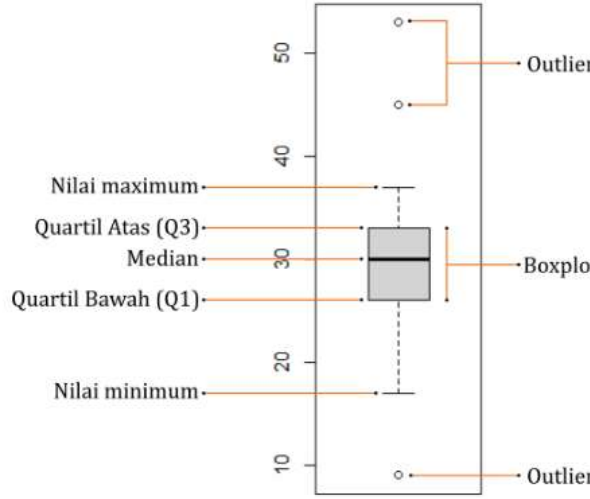
Seringkali gagasan tentang tingkat keanehan (anomali) dari nilai-nilai ekstrim sangat dibutuhkan dalam proses analisis pada data. Penyempurnaan dari boxplot yang menyajikan detail lebih lanjut pada bagian ekor adalah plot skematik, yang juga dipernalkan oleh Tukey (1977). Plot skematik identik dengan boxplot, kecuali titik ekstrim yang dianggap cukup tidak biasa diplot secara individual. Tingkat ekstremnya dinilai tidak biasa tergantung pada variabilitas data dalam cakupan median sampel, sebagaimana tercermin dari IQR. Nilai ekstrim yang diberikan dianggap kurang biasa jika kedua kuartil berjauhan (yaitu, jika IQR besar), dan lebih tidak biasa jika kedua kuartil sangat dekat satu sama lain (IQR kecil).

Garis pemisah antara titik yang kurang atau tidak biasa dikenal dalam terminologi istimewa Tukey sebagai batas atau "$fences$". Terdapat empat batas yang digunakan yaitu:

$$\text{Upper outer fence} = q_{0.75} + 3(IQR)$$
$$\text{Upper inner fence} = q_{0.75} + \frac{3(IQR)}{2} \qquad (0.11)$$

$$\text{Lower inner fence} = q_{0.25} - \frac{3(IQR)}{2}$$
$$\text{Lower outer fence} = q_{0.25} - 3(IQR)$$

Dengan demikian dua batas terluar (*outer*) terletak tiga kali jarak rentang interkuartil (IQR) di atas dan di bawah kedua kuartil. Batas bagian dalam (*iiner*) berada di tengah-tengah antara batas luar dan kuartil, menjadi setengah jarak rentang interkuartil dari kuartil.



Gambar 3.4. Plot skematik untuk data temperatur maksimum Ithaca pada Januari 1987. Porsi pusat kotak identik dengan boxplot pada Gambar 3.3. Tiga nilai diluar *inner fence* diplot secara terpisah. Tidak ada nilai diluar *outer fence*. Perhatikan bahwa kumis (*whiskers*) memanjang ke nilai ekstrim bagian dalam (*inside*), bukan ke *fence*.

Implementasi dalam program R

Sintaks:

```
> boxplot_ithaca <- boxplot(ithaca$Max.Temp.)
> IQR <- boxplot_ithaca$stats[4] - boxplot_ithaca$stats[2]
> inner_fence_1 <- boxplot_ithaca$stats[2] - (3/2)*IQR
> inner_fence_2 <- boxplot_ithaca$stats[4] + (3/2)*IQR
> outer_fence_1 <- boxplot_ithaca$stats[2] - 3*IQR
> outer_fence_2 <- boxplot_ithaca$stats[4] + 3*IQR
> lv_inside <- max(boxplot_ithaca$stats)
> sv_inside <- min(boxplot_ithaca$stats)
> out_lower <- boxplot_ithaca$out[boxplot_ithaca$out < sv_inside]
> out_upper <- boxplot_ithaca$out[boxplot_ithaca$out > lv_inside]

> ## output ##

> cat(
+   'Kuartil data  :',boxplot_ithaca$stats[2], 'dan',
boxplot_ithaca$stats[4],
+   '\nBatas dalam   :',inner_fence_1,'dan',inner_fence_2,
+   '\nBatas luar    :',outer_fence_1,'dan',outer_fence_2,
+   '\nOutlier atas  :',out_upper,
+   '\nOutlier bawah :',out_lower
+ )
Kuartil data  : 26 dan 33
Batas dalam   : 15.5 dan 43.5
Batas luar    : 5 dan 54
Outlier atas  : 53 45
Outlier bawah : 9
>
```

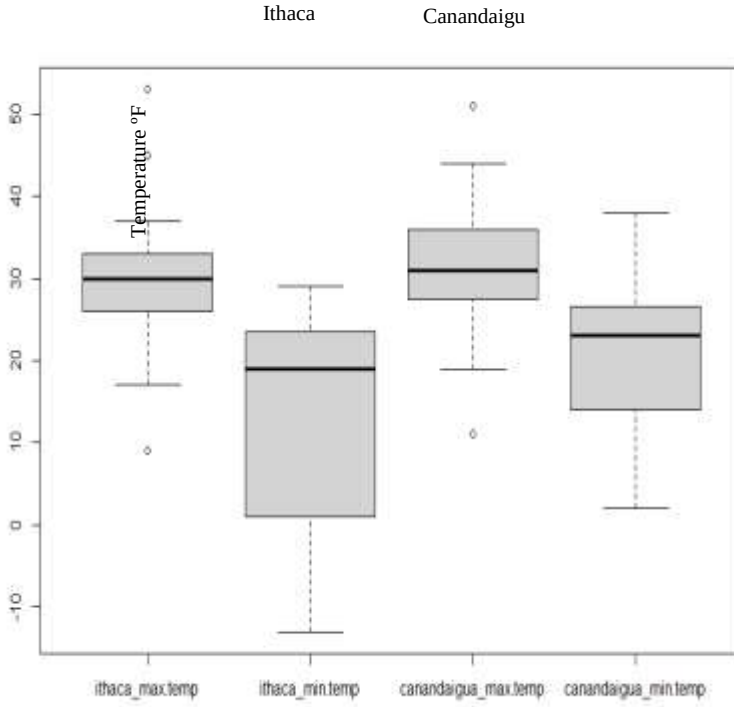
Dalam plot skematik, titik-titik dalam batas bagian dalam disebut "*inside*". Rentang titik bagian dalam direpresentasikan oleh panjang kumis (*whiskers*). Titik data antara batas dalam dan luar disebut sebagai "*outside*" dan diplot secara individual dalam plot skematik. Titik-titik di atas batas luar atas atau di bawah batas luar bawah disebut "*far out*" dan diplot secara secara terpisah dengan simbol yang berbeda. Perbedaan ini diilustrasikan pada Gambar 3.4. Sama

dengan boxplot, kotak dalam plot skematik menunjukkan lokasi kuartil dan median.

Contoh 0.2 Konstruksi Plot Skematik

Gambar 3.4 adalah plot skematik untuk data suhu maksimum Ithaca Januari 1987. Seperti pada Gambar 3.1, kuartil datanya adalah pada $33°F$ dan $26°F$, dan $IQR = 33 - 26 = 7°F$. Dari informasi tersebut, mudah untuk menghitung lokasi batas bagian dalam (*inner fence*) yaitu $33 + \left(\frac{3}{2}\right)(7) = 43.5°F$ dan $26 - \left(\frac{3}{2}\right)(7) = 15.5°F$. Demikian pula, batas luar berada di $33 + (3)(7) = 54°F$ dan $26 - (3)(7) = 5°F$.

Dua suhu terpanas, yaitu $53°F$ dan $45°F$, lebih besar dari batas bagian dalam atas dan disimbolkan secara terpisah dengan lingkaran. Suhu terdingin, yaitu $9°F$, kurang dari batas bagian dalam bawah, dan juga diplot secara terpisah. Kumis (*whiskers*) memanjang pada suhu paling ekstrem dalam batas, yaitu $37°F$ dan $17°F$. Jika suhu terpanas adalah $55°F$ bukan $53°F$, maka nilai tersebut akan jatuh di luar batas luar (*far out*) dan akan diplot secara terpisah dengan simbol yang berbeda. Simbol terpisah untuk titik jauh ini sering digambarkan dengan tanda bintang.



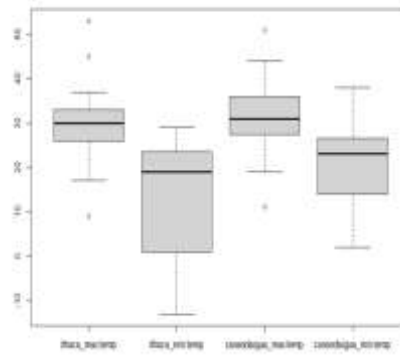
Gambar 3.5. Plot skematik berdampingan untuk data temperatur pada Januari 1987 dalam Tabel A.1. Data temperatur minimum untuk dua lokasi semuanya berada di bagian dalam batas luar sehingga plot skematik identik dengan boxplot biasa.

Implementasi dalam program R

Sintaks:

```
> ## Membersihkan history ##
> rm(list = ls())
> ## Read data csv ##
> data_TA1             <- read.csv('Data/Tabel A1.csv')
> ithaca_max.temp      <- data_TA1$Max.Temp.[1:31]
> ithaca_min.temp      <- data_TA1$Min.Temp.[1:31]
> canandaigua_max.temp <- data_TA1$Max.Temp..1[1:31]
> canandaigua_min.temp <- data_TA1$Min.Temp..1[1:31]
> ithaca_n_canand      <- cbind(ithaca_max.temp,
                                ithaca_min.temp,
                                canandaigua_max.temp,
                                canandaigua_min.temp)

> ## show boxplot ##
> boxplot(ithaca_n_canand)
```

Output:

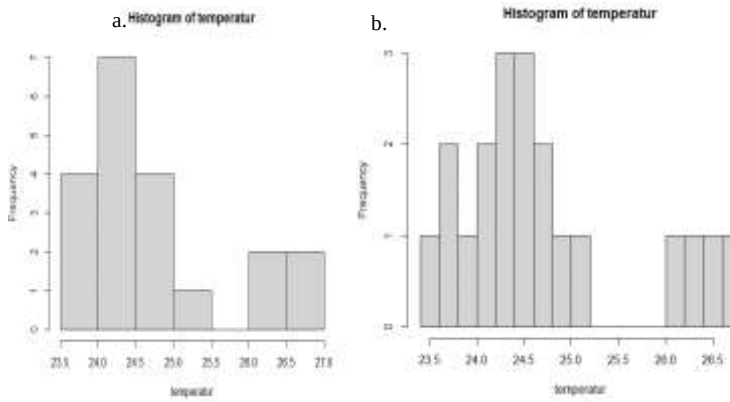


Salah satu penggunaan penting dari plot skematik atau boxplot adalah perbandingan grafik simultan dari beberapa kumpulan data. Penggunaan plot skematik ini diilustrasikan

pada Gambar 3.5, yang menunjukkan plot berdampingan untuk keempat kumpulan data suhu pada Tabel A.1. Sebelumnya diketahui bahwa suhu maksimum lebih hangat daripada suhu minimum dan membandingkan plot skematiknya menunjukkan perbedaan yang cukup besar. Namun, dari Gambar 3.5 juga diketahui bahwa Canandaigua sedikit lebih hangat daripada Ithaca pada bulan tersebut dan terlihat jelas pada suhu minimum. Suhu minimum Ithaca ternyata lebih bervariasi daripada suhu minimum Canandaigua. Untuk kedua lokasi, temperatur minimum lebih bervariasi daripada temperatur maksimum, khususnya di bagian tengah distribusi yang diwakili oleh kotak. Lokasi median di ujung atas kotak plot skematik suhu minimum yang menunjukkan kecenderungan kemiringan negatif, seperti halnya ketidaksebandingan panjang kumis (*whiskers*) untuk data suhu minimum Ithaca. Suhu maksimum tampaknya cukup simetris untuk kedua lokasi. Perhatikan bahwa tidak ada data suhu minimum yang berada di luar batas bagian dalam sehingga boxplot dengan data yang sama akan identik.

3.3.4 Histogram

Histogram adalah perangkat tampilan grafik yang sangat familiar untuk satu kumpulan data. Rentang data dibagi menjadi interval kelas (*bin*) dan jumlah nilai yang jatuh ke setiap interval dihitung. Histogram kemudian terdiri dari serangkaian persegi panjang yang lebarnya ditentukan oleh batas kelas dan tingginya bergantung pada jumlah nilai yang berada di setiap interval kelas. Contoh histogram ditunjukkan pada Gambar 3.6. Histogram dengan jelas mengungkapkan atribut data seperti tendensi pusat, sebaran, dan simetri dari data.



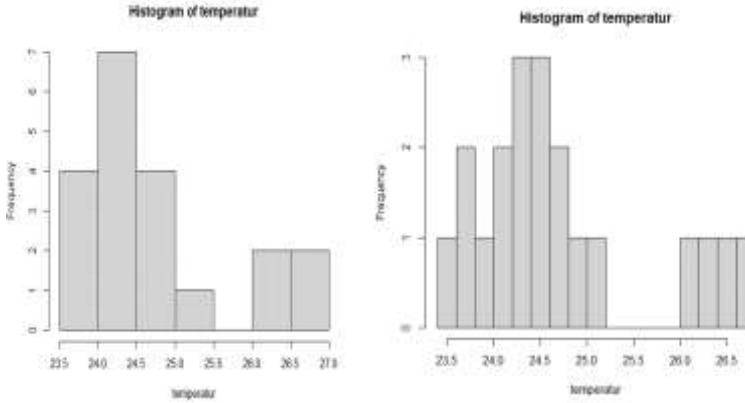
Gambar 3.6. Histogram dari data temperatur June Guayaquil dalam Tabel A.3, mengilustrasikan perbedaan yang dapat timbul karena pergeseran interval kelas. Gambar ini juga mengilustrasikan bahwa setiap batang histogram dapat dipandang seperti blok yang ditumpuk sesuai dengan jumlah nilai yang berada dalam interval kelas. Plot titik dibawah setiap histogram menunjukkan lokasi data sebenarnya.

Implementasi dalam program R

Sintaks:

```
> rm(list = ls())
>
> ## Read Data Tabel A.3 ##
> data_tA3 <- read.csv('Data/Tabel A3.csv')
>
> ## Histogram ##
> ## Histogram a ##
> temperatur <- data_tA3$x2
> hist(temperatur, breaks = 11)

> ## Histogram b ##
> temperatur <- data_tA3$x2
> hist(temperatur, breaks = 12)
```

Output:



Biasanya, lebar interval kelas yang dipilih adalah sama. Dalam hal ini, tinggi batang histogram kemudian proporsional dengan jumlah hitungannya. Sumbu vertikal dapat diberi label untuk memperlihatkan jumlah hitungan yang diwakili oleh setiap batang (frekuensi absolut) atau proporsi seluruh sampel yang diwakili oleh setiap batang (frekuensi relatif). Oleh karena itu, mungkin perlu untuk menskalakan histogram

sehingga luas total yang terkandung dalam batang histogram berjumlah 1.

Masalah utama yang sering dihadapi saat membuat histogram adalah pilihan lebar bin. Interval yang terlalu lebar akan menghasilkan detail penting dari data akan tersamarkan (histogram terlalu halus). Interval yang terlalu sempit akan menghasilkan plot yang tidak beraturan dan sulit diinterpretasikan (histogram terlalu kasar). Secara umum, bin histogram yang lebih sempit dibenarkan untuk sampel data yang besar, namun sifat data juga memengaruhi pemilihan lebar bin. Salah satu pendekatan untuk memilih lebar bin (disimbolkan h) adalah,

$$h \cong \frac{c\ IQR}{n^{\frac{1}{3}}} \tag{0.12}$$

dimana c adalah konstanta dalam rentang antara 2.0 sampai 2.6. Hasil yang diberikan dalam Scott (1992) mengindikasikan bahwa $c = 2.6$ optimal untuk data Gaussian (*bentuk lonceng*) dan nilai yang lebih kecil akan lebih sesuai untuk data miring (*skewed*).

Lebar bin awal yang dihitung menggunakan Persamaan (0.12) atau yang diperoleh menurut aturan lain hanya dianggap sebagai pedoman atau aturan praktis. Pertimbangan lain juga akan masuk ke dalam pilihan lebar bin agar batas kelas jatuh pada nilai yang wajar sehubungan dengan data yang ada. Program komputer yang menampilkan histogram menggunakan aturan seperti pada Persamaan (0.12) dan salah satu indikasi kehati-hatian dalam penulisan perangkat lunak adalah apakah histogram yang dihasilkan memiliki batas bin alami. Sebagai contoh, maksimum Ithaca Januari 1987 data suhu memiliki $IQR = 7^{\circ}F$ dan $n = 31$. Lebar bin sebesar $5.7^{\circ}F$ akan disarankan pada awalnya dengan Persamaan (0.12) menggunakan $c = 2.6$ karena data terlihat setidaknya mendekati Gaussian. Biasanya, mungkin

dengan memilih 10 bin dengan lebar $5°F$ menghasilkan histogram yang mirip dengan tampilan *stem-and-leaf* pada Gambar 3.1 (b).

3.3.5 Pemulusan Densitas Kernel

Salah satu interpretasi histogram adalah sebagai estimator nonparametrik dari distribusi probabilitas yang mendasari data yang telah dikumpulkan. Namun, penjajaran bin histogram pada garis sebenarnya dapat dipilih secara sebarang dan konstruksi histogram pada dasarnya mengharuskan setiap nilai data dibulatkan ke tengah bin tempat ia jatuh. Misalnya, pada Gambar 3.6 (a) bin telah diselaraskan sehingga terpusat pada nilai suhu bilangan bulat ± 0.25℃, sedangkan histogram pada Gambar 3.6 (b) telah digeser sebesar 0.25℃. Kedua histogram pada Gambar 3.6 menampilkan gambaran data yang agak berbeda, meskipun keduanya berasal dari data yang sama. Kesulitan lain yang mungkin terjadi pada histogram adalah sifat persegi panjang dari batang histogram menampilkan tampilan kasar dan seperti menyiratkan bahwa nilai apa pun yang ada pada bin memiliki kemungkinan yang sama.

Sebuah alternatif untuk histogram yang tidak memerlukan pembulatan acak ke pusat bin dan memberikan hasil yang mulus adalah pemulusan densitas kernel. Penerapan pemulusan kernel pada distribusi frekuensi kumpulan data menghasilkan perkiraan densitas kernel yang merupakan alternatif nonparametrik untuk pemasangan fungsi densitas probabilitas parametrik. Tepatnya pemulusan densitas kernel dapat dilihat sebagai perpanjangan dari histogram. Seperti yang diilustrasikan pada Gambar 3.6, setelah membulatkan setiap nilai data ke pusat bin, histogram dapat dilihat sebagai tumpukan blok di atas setiap pusat bin, dengan jumlah blok sama dengan jumlah titik data di setiap

bin. Pada Gambar 3.6 distribusi data ditunjukkan di bawah setiap histogram dalam bentuk dotplot yang menempatkan setiap nilai data dengan sebuah titik dan menunjukkan contoh data berulang dengan tumpukan titik.

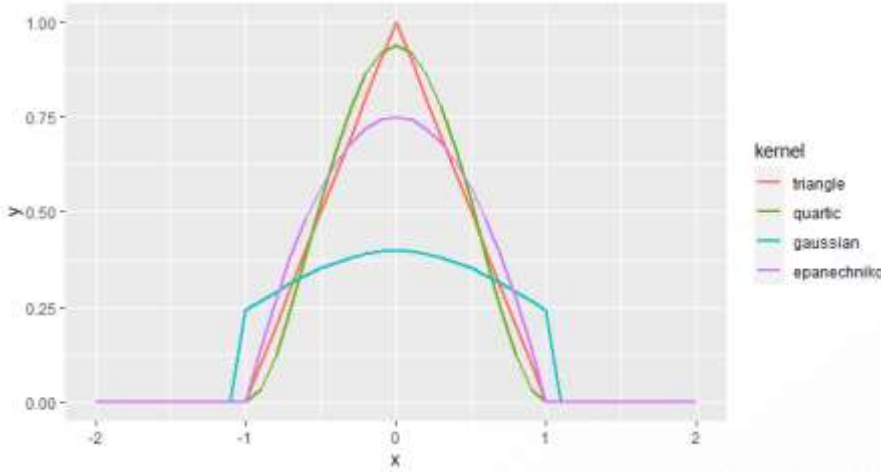
Blok persegi panjang pada Gambar 3.6 masing-masing memiliki luas sama dengan lebar bin 0.5°F, karena sumbu vertikal hanyalah jumlah mentah hitungan di setiap bin. Sebaliknya, jika sumbu vertikal telah dipilih sehingga luas setiap blok penyusun adalah $\frac{1}{n}$, histogram yang dihasilkan akan menjadi penaksir kuantitatif dari distribusi probabilitas karena total luas histogram akan menjadi 1 dalam setiap kasus dan probabilitas total harus berjumlah 1.

Tabel 3.1. Beberapa kernel pemulusan yang umum digunakan.

Nama Pemulusan	$\boldsymbol{K(t)}$	$\boldsymbol{t}$ **untuk setiap** $\boldsymbol{K(t) > 0}$	$\boldsymbol{\frac{1}{\sigma_k}}$
Triangular	$1 - \lvert t \rvert$	$-1 < t < 1$	$\sqrt{6}$
Quadratic (Epanechnikov)	$\left(\frac{3}{4}\right)(1 - t^2)$	$-1 < t < 1$	$\sqrt{5}$
Quaric (Biweight)	$\left(\frac{15}{16}\right)(1 - t^2)^2$	$-1 < t < 1$	$\sqrt{7}$
Gaussian	$(2\pi)^{-\frac{1}{2}} \exp\left[-\frac{t^2}{2}\right]$	$-\infty < t < \infty$	1

Pemulusan densitas kernel berlangsung dengan cara yang analog, menggunakan bentuk karakteristik yang disebut kernel, yang umumnya lebih halus daripada persegi panjang. Tabel 3.1 mencantumkan empat kernel pemulusan yang umum digunakan dan Gambar 3.7 menunjukkan bentuknya secara grafis. Ini semua adalah fungsi non-negetif dengan luas satu, yaitu $\int K(t)\,\mathrm{d}t = 1$, dalam setiap kasus, sehingga masing-masing adalah fungsi kepadatan peluang yang tepat

(dibahas lebih detail di Bab 4). Selain itu, semua berpusat pada nol. Kernel yang tercantum dalam Tabel 3.1 sesuai untuk digunakan dengan data kontinu (mengambil nilai pada semua atau sebagian dari garis sebenarnya). Beberapa kernel yang sesuai untuk data diskrit (hanya mampu mengambil sejumlah nilai terbatas) disajikan dalam Rajagopalan et al. (1997).



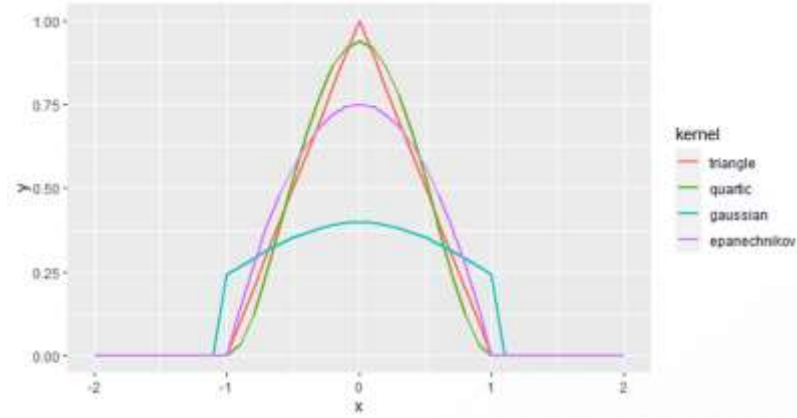
Gambar 3.7. Empat kernel pemulisan yang umum digunakan dalam Tabel 3.1.

Implementasi dalam program R

Sintaks:

```
> ## Package R ##
> library('ggplot2')
> library('reshape2')
> library('kableExtra')
> library('spNetwork')
> library('RColorBrewer')
> x <- seq(-2,2,by = 0.1)
> df <- data.frame(
+   x = x,
+   triangle = triangle_kernel(x,1),
+   quartic = quartic_kernel(x,1),
+   gaussian = gaussian_kernel(x,1),
+   epanechnikov = epanechnikov_kernel(x,1))
> df2 <- melt(df, id.vars = "x")
> names(df2) <- c("x","kernel","y")
> ## Plot ##
> ggplot(df2) + geom_line(aes(x=x, y=y, color=kernel), size=1)
```

Output:



Alih-alih menumpuk kernel persegi panjang yang berpusat pada titik tengah bin (yang merupakan salah satu cara untuk melihat konstruksi histogram), pemulusan densitas kernel dibentuk dengan menumpuk bentuk kernel, jumlahnya sama dengan jumlah nilai data, dengan setiap elemen yang ditumpuk dan dipusatkan pada nilai yang diwakilinya. Tentu saja secara umum bentuk kernel tidak cocok satu sama lain seperti blok bangunan, tetapi pemulusan densitas kernel dicapai melalui ekuivalen matematis dari penumpukan dengan menambahkan ketinggian semua fungsi kernel yang berkontribusi pada perkiraan pemulusan pada nilai tertentu, x_0,

$$\hat{f}(x_0) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right) \qquad (0.13)$$

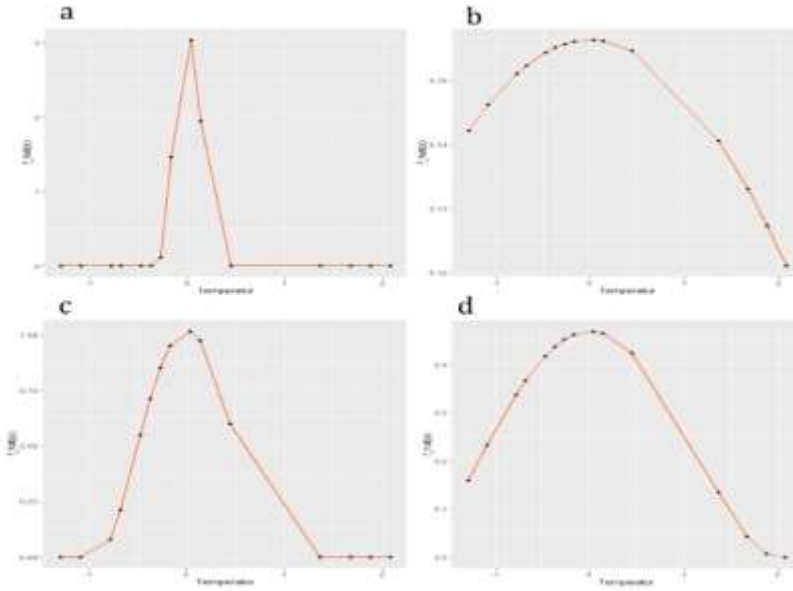
Argumen dalam fungsi kernel menunjukkan bahwa setiap kernel yang digunakan dalam pemulusan dipusatkan pada nilai data masing-masing x_i dan diskalakan dalam lebar relatif terhadap bentuk seperti yang diplot pada Gambar 3.7 dengan parameter pemulusan h. Pertimbangkan misalnya, kernel segitiga pada Tabel 3.1, dengan $t = \frac{x_0 - x_i}{h}$. Fungsi $K\left[\frac{x_0 - x_i}{h}\right] = 1 - \left|\frac{x_0 - x_i}{h}\right|$ adalah segitiga sama kaki dengan pendukung (misalnya, tinggi tidak nol) untuk $x_i - h < x_0 < x_i + h$ dan luas dalam segitiga ini adalah h karena luas dalam $1 - |t|$ adalah 1 dan alasnya telah diperluas (atau dikerutkan) dengan faktor h. Oleh karena itu, dalam Persamaan 3.13 tinggi kernel yang ditumpuk pada nilai x_0 akan sesuai dengan salah satu dari x_i pada jarak yang lebih dekat dengan x_0 daripada h. Agar luas di bawah seluruh fungsi dalam Persamaan 3.13 berintegrasi menjadi 1, yang diinginkan jika hasilnya dimaksudkan untuk memperkirakan fungsi kerapatan probabilitas, masing-masing dari n kernel yang akan

ditumpangkan harus memiliki luas $\frac{1}{n}$. Hal ini dicapai dengan membagi setiap $K\left[\frac{x_0-x_i}{h}\right]$, atau dengan cara yang sama membagi jumlahnya dengan hasil kali nh.

Pilihan jenis kernel tampaknya kurang menarik dibandingkan dengan pilihan parameter pemulusan. Kernel Gaussian menarik secara intuitif, tetapi komputasinya lebih lambat karena pemanggilan fungsi eksponensial dan pendukung tak terbatasnya menyebabkan semua nilai data berkontribusi pada estimasi yang dihaluskan pada setiap x_0 (tidak ada n suku dalam Persamaan 3.13 yang akan nol). Di sisi lain, semua turunan dari fungsi yang dihasilkan akan ada dan probabilitas bukan nol diestimasi dimana pun pada garis real, sedangkan ini bukan karakteristik dari kernel lain yang tercantum dalam Tabel 3.1.

Contoh 0.3 Perkiraan Kepadatan Kernel Data Suhu Guayaquil

Gambar 3.8 menunjukkan estimasi densitas kernel untuk data suhu Guayaquil bulan Juni pada Tabel A.3, sesuai dengan histogram pada Gambar 3.6.



Gambar 3.8. Estimasi densitas kernel untuk data suhu Guayaquil bulan Juni pada Tabel A.3 dibangun menggunakan kernel quartic dan (a) $h = 0.3$, (b) $h = 6$, (c) $h = 0.92$, dan (d) $h = 2.0$. Juga ditunjukkan pada panel (b) adalah masing-masing kernel yang telah ditambahkan bersama untuk membuat perkiraan. Data yang sama ini ditampilkan sebagai histogram pada Gambar 3.6.

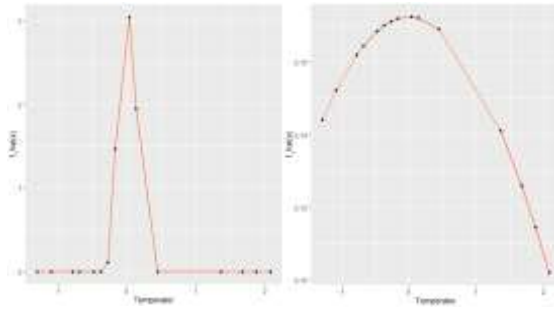
Implementasi dalam program R

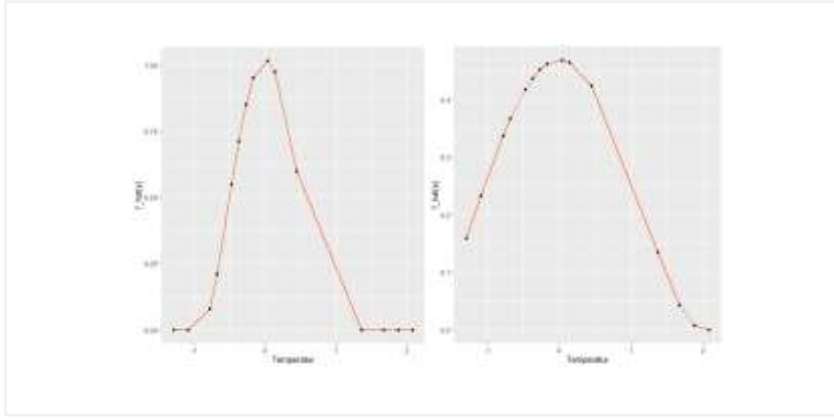
Sintaks:

```
> rm(list = ls())
> ## Package R ##
> library('ggplot2')
> library('reshape2')
> library('kableExtra')
> library('spNetwork')
> library('RColorBrewer')
>
> ## Read Data Tabel A.3 ##
> data_tA3   <- read.csv('Data/Tabel A3.csv')
> temperatur <- data_tA3$x2
> x          <- array((temperatur -
+                        mean(temperatur))/sd(temperatur))
>
> ## plot a ##
> h   <- 0.3
> df  <- data.frame(x = x,quartic = quartic_kernel(x, h))
> df2 <- melt(df, id.vars = "x");names(df2) <- c("x","kernel","y")
>
> ggplot(df2) +
+   geom_line(aes(x=x, y=y), size=1, color='coral') +
+   geom_point(aes(x=x, y=y)) +
+   ylab('f_hat(x)') +
+   xlab('Temperatur')
>
> ## plot b ##
> h   <- 6
> df  <- data.frame(x = x,quartic = quartic_kernel(x, h))
> df2 <- melt(df, id.vars = "x");names(df2) <- c("x","kernel","y")
> ggplot(df2) +
+   geom_line(aes(x=x, y=y), size=1, color='coral') +
+   geom_point(aes(x=x, y=y)) +
+   ylab('f_hat(x)') +
+   xlab('Temperatur')
>
```

```
> ## plot b ##
> h   <- 6
> df  <- data.frame(x = x,quartic = quartic_kernel(x, h))
> df2 <- melt(df, id.vars = "x");names(df2) <- c("x","kernel","y")
> ggplot(df2) +
+   geom_line(aes(x=x, y=y), size=1, color='coral') +
+   geom_point(aes(x=x, y=y)) +
+   ylab('f_hat(x)') +
+   xlab('Temperatur')
>
> ## plot c ##
> h   <- 0.92
> df  <- data.frame(x = x,quartic = quartic_kernel(x, h))
> df2 <- melt(df, id.vars = "x");names(df2) <- c("x","kernel","y")
> ggplot(df2) +
+   geom_line(aes(x=x, y=y), size=1, color='coral') +
+   geom_point(aes(x=x, y=y)) +
+   ylab('f_hat(x)') +
+   xlab('Temperatur')
>
> ## plot c ##
> h   <- 2
> df  <- data.frame(x = x,quartic = quartic_kernel(x, h))
> df2 <- melt(df, id.vars = "x");names(df2) <- c("x","kernel","y")
> ggplot(df2) +
+   geom_line(aes(x=x, y=y), size=1, color='coral') +
+   geom_point(aes(x=x, y=y)) +
+   ylab('f_hat(x)') +
+   xlab('Temperatur')
>
```

Output:

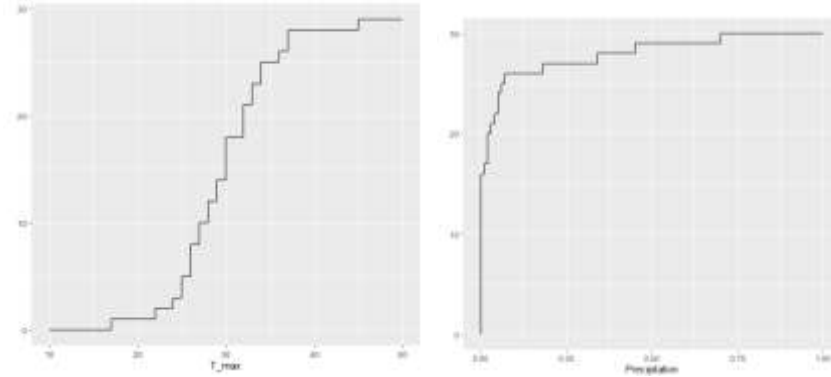




Empat perkiraan kerapatan telah dibangun menggunakan kernel kuartik (*quartic*) dan empat pilihan untuk parameter pemulusan h, yang meningkat dari panel (a) hingga (d). Peran parameter pemulusan analog dengan lebar bin histogram, juga disebut h, dimana nilai yang lebih besar menghasilkan bentuk yang lebih halus yang semakin menekan detail informasi dalam data. Nilai yang lebih kecil menghasilkan bentuk yang lebih tidak beraturan yang menampilkan lebih banyak detail, termasuk variabilitas sampel. Gambar 3.8 (b) diplot dengan menggunakan $h = 0.6$ menunjukkan masing-masing kernel yang telah dijumlahkan untuk menghasilkan pendugaan densitas yang dihaluskan. Karena $h = 0.6$ dan pendukung dari kernel kuartik adalah $-1 < t < 1$ (lihat Tabel 3.1) lebar masing-masing kernel pada Gambar 3.8 (b) adalah 1.2. Lima nilai data berulang 23.7, 24.1, 24.3, 24.5 dan 24.8 (lihat dotplot di bagian bawah Gambar 3.6) diwakili oleh lima kernel yang lebih tinggi dengan luas masing-masing $\frac{2}{n}$. 10 nilai data yang tersisa adalah unik dan kernelnya masing-masing memiliki luas $\frac{1}{n}$.

3.3.6 Distribusi Frekuensi Kumulatif

Distribusi frekuensi kumulatif adalah tampilan yang terkait dengan histogram. Ini juga dikenal sebagai fungsi distribusi kumulatif empiris. Distribusi frekuensi kumulatif adalah plot dua dimensi dimana sumbu vertikal menunjukkan estimasi probabilitas kumulatif yang terkait dengan nilai data pada sumbu horizontal. Artinya, plot mewakili perkiraan frekuensi relatif untuk probabilitas bahwa datum acak kedepannya tidak akan melebihi nilai yang sesuai pada sumbu horizontal. Dengan demikian, distribusi frekuensi kumulatif seperti integral dari histogram dengan lebar bin sempit yang berubah-ubah. Gambar 3.9 menunjukkan dua fungsi distribusi kumulatif empiris yang mengilustrasikan bahwa keduanya adalah fungsi langkah (*step function*) dengan lompatan probabilitas yang terjadi pada nilai data. Sama seperti histogram yang dapat dihaluskan menggunakan estimator kerapatan kernel, versi yang dihaluskan dari fungsi distribusi kumulatif empiris dapat diperoleh dengan mengintegrasikan hasil pemulusan kernel.



Gambar 3.9. Fungsi distribusi frekuensi kumulatif empiris untuk data suhu maksimum Ithaca Januari 1987 (a) dan data curah hujan (b). Bentuk S yang diperlihatkan oleh data temperatur merupakan

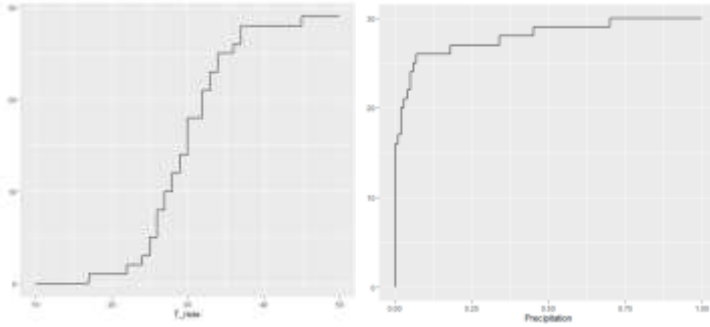
karakteristik dari data yang cukup simetris, dan tampilan cekung ke bawah yang ditunjukkan oleh data curah hujan merupakan karakteristik dari data yang miring ke kanan.

Implementasi dalam program R

Sintaks:

```
> ## Package R ##
> library('ggplot2')
> rm(list = ls())
> ## Read data csv ##
> data_TA1 <- read.csv('Data/Tabel A1.csv')
> ithaca   <- data_TA1[1:31, 1:3]
> ## Plot a ##
> duration <- ithaca$Max.Temp.
> breaks   <- seq(10, 50, by=0.01)
> duration.cut  <- cut(duration, breaks, right=FALSE)
> duration.freq <- table(duration.cut)
> cumfreq0      <- c(0, cumsum(duration.freq))
> data_plot     <- data.frame(breaks, cumfreq0)
> ggplot(data_plot, aes(x = breaks, y =cumfreq0)) +
+   geom_line(lwd = 1, color = '#555555') +
+   xlab('T_max') +
+   ylab(' ')
> ## Plot b ##
> duration <- ithaca$Precipitation
> breaks   <- seq(0, 1, by=0.001)
> duration.cut  <- cut(duration, breaks, right=FALSE)
> duration.freq <- table(duration.cut)
> cumfreq0      <- c(0, cumsum(duration.freq))
> data_plot     <- data.frame(breaks, cumfreq0)
> ggplot(data_plot, aes(x = breaks, y =cumfreq0)) +
+   geom_line(lwd = 1, color = '#555555') +
+   xlab('Precipitation') +
+   ylab(' ')
>
```

Output:



Sumbu vertikal pada Gambar 3.9 menunjukkan fungsi distribusi kumulatif empiris, $p(x)$, yang dapat dinyatakan sebagai:

$$p(x) \approx P\{X \leq x\} \quad (0.14)$$

Notasi di sisi kanan persamaan diatas mungkin agak membingungkan pada awalnya, namun persamaan tersebut merupakan standar dalam pekerjaan statistik. Huruf besar X mewakili variabel acak umum, atau nilai "acak kedepannya" yang dijelaskan pada paragraf sebelumnya. Huruf kecil x, pada kedua sisi Persamaan 3.14, mewakili nilai tertentu dari besaran acak. Dalam distribusi frekuensi kumulatif, nilai spesifik ini diplot pada sumbu horizontal.

Untuk membangun distribusi frekuensi kumulatif, perlu untuk memperkirakan $p(x)$ menggunakan peringkat, i, dari urutan statistik, $x_{(i)}$. Dalam literatur hidrologi perkiraan ini dikenal sebagai *plotting position* (e.g., Harter, 1984) yang mencerminkan penggunaan historisnya dalam membandingkan secara grafis distribusi empiris kumpulan data dengan kandidat fungsi parametrik yang mungkin digunakan untuk mewakilinya. Ada literatur substansial yang

dikhususkan untuk persamaan yang dapat digunakan untuk menghitung *plotting position* dan memperkirakan probabilitas kumulatif. Sebagian besar adalah kasus khusus dari rumus

$$p\left(x_{(i)}\right) = \frac{i - a}{n + 1 - 2a}, \qquad 0 \leq a \leq 1 \qquad (0.15)$$

Dimana nilai yang berbeda untuk konstanta menghasilkan estimator posisi plotting yang berbeda, beberapa di antaranya ditunjukkan pada Tabel 3.2. Nama-nama dalam tabel ini berhubungan dengan penulis yang mengusulkan berbagai estimator dan bukan distribusi probabilitas tertentu yang mungkin dinamai untuk penulis yang sama. Empat posisi plot pertama pada Tabel 3.2 dimotivasi oleh karakteristik distribusi sampling probabilitas kumulatif yang terkait dengan statistik urutan. Gagasan tentang distribusi sampling dipertimbangkan secara lebih rinci di Bab 5, tetapi secara singkat, pikirkan tentang secara hipotetis mendapatkan sejumlah besar sampel data berukuran n dari beberapa distribusi yang tidak diketahui. Statistik urutan ke-i dari sampel-sampel ini akan agak berbeda satu sama lain, namun masing-masing akan sesuai dengan probabilitas kumulatif tertentu dalam distribusi dari mana data diambil. Secara agregat atas sejumlah besar sampel, hipotetis akan memiliki distribusi probabilitas kumulatif yang sesuai dengan statistik urutan ke-i. Salah satu cara untuk membayangkan distribusi pengambilan sampel ini adalah sebagai histogram probabilitas kumulatif yang terkecil (atau statistik urutan lainnya) dari nilai n di setiap kumpulan data. Gagasan distribusi sampling untuk probabilitas kumulatif ini diperluas lebih lengkap dalam konteks klimatologis oleh Folland dan Anderson (2002).

Tabel 3.2. Beberapa estimator *plotting position* yang umum untuk probabilitas kumulatif yang sesuai dengan statistik urutan ke-i, $x_{(i)}$, dan nilai yang sesuai dari parameter a dalam Persamaan 3.15.

Nama	**Rumus**	a	**Interpretasi**
Weibull	$\frac{i}{n+1}$	0	rata-rata distribusi sampel
Benard & Bos-Levenbach	$\frac{i-0.3}{n+0.4}$	0.3	Perkiraan median distribusi sampel
Tukey	$\frac{i-\frac{1}{3}}{n+\frac{1}{3}}$	1/3	Perkiraan median distribusi sampel
Gumbel	$\frac{i-1}{n-1}$	1	Modus distribusi sampel
Hazen	$\frac{i-\frac{1}{2}}{n}$	1/2	Titik tengah n interval yang sama dalam [0, 1]
Cunnane	$\frac{i-\frac{2}{5}}{n+\frac{1}{5}}$	2/5	Pilihan subyektif, biasa digunakan dalam hidrologi

Bentuk matematis dari distribusi sampel probabilitas kumulatif yang berhubungan dengan statistik urutan ke-i dikenal sebagai distribusi Beta dengan parameter $p = i$ dan $q = n - i + 1$, terlepas dari distribusi dari x telah ditarik secara independen. Dengan demikian estimator *plotting position* weibull ($a = 0$) adalah rata-rata dari probabilitas kumulatif yang sesuai dengan $x_{(i)}$ tertentu, dirata-ratakan

pada banyak sampel hipotetis ukuran n. Demikian pula, penaksir Benard & Bos-Levenbach ($a = 0.3$) dan Tukey $a = \frac{1}{3}$, penaksir mendekati median dari distribusi ini. *Plotting position* Gumbel ($a = 1$) menempatkan probabilitas kumulatif modal meskipun itu menganggap nol dan unit probabilitas kumulatif masing-masing untuk $x_{(1)}$ dan $x_{(n)}$. Dimungkinkan juga untuk menurunkan formula *plotting position* menggunakan perspektif terbalik, berpikir tentang distribusi sampling kuantil data x_i sesuai dengan probabilitas kumulatif tetap tertentu (misalnya, Cunnane, 1978; Stedinger et al., 1993). Tidak seperti empat posisi plot pertama pada Tabel 3.2, *plotting position* yang dihasilkan dari pendekatan ini bergantung pada distribusi dari mana data telah diambil, meskipun *plotting position* Cunnane ($a = \frac{2}{5}$) adalah pendekatan kompromi. Dalam prakteknya sebagian besar dari berbagai formula *plotting position* menghasilkan hasil yang sangat mirip, terutama ketika dinilai dalam kaitannya dengan variabilitas intrinsik dari distribusi sampling dari probabilitas kumulatif, yang jauh lebih besar daripada perbedaan di antara berbagai *plotting position* pada Tabel 3.2. Umumnya hasil yang sangat masuk akal diperoleh dengan menggunakan *plotting position* moderat (dalam hal parameter a) seperti Tukey atau Cunnane. Gambar 3.9 (a) menunjukkan distribusi frekuensi kumulatif untuk data suhu maksimum Ithaca Januari 1987, menggunakan *plotting position* Tukey untuk memperkirakan probabilitas kumulatif. Gambar 3.9 (b) menunjukkan data curah hujan Ithaca yang ditampilkan dengan cara yang sama. Sebagai contoh, untuk suhu terdingin dari 31 suhu pada Gambar 3.10 (a) adalah $x_{(1)} = 9°\text{F}$ dan $p(x_{(1)})$ diplot pada $\frac{1-0.333}{31+0.333} = 0.0213$. Kecuraman di tengah plot mencerminkan konsentrasi nilai data di tengah distribusi dan kerataan pada suhu tinggi dan rendah dihasilkan dari

kelangkaannya. Karakter berbentuk S dari distribusi kumulatif ini merupakan indikasi dari distribusi yang cukup simetris, dengan jumlah pengamatan yang sebanding di kedua sisi median pada jarak tertentu dari median. Fungsi distribusi kumulatif untuk data curah hujan (lihat Gambar 3.9 (b)) naik dengan cepat di sebelah kiri karena tingginya konsentrasi nilai data di sana, dan kemudian naik lebih lambat di tengah dan kanan gambar karena pengamatan yang relatif lebih sedikit. Tampilan cekung ke bawah ke fungsi distribusi kumulatif ini menunjukkan data yang miring secara positif. Plot untuk kumpulan data miring negatif hanya akan menunjukkan karakteristik kebalikannya: kemiringan yang sangat dangkal di kiri dan tengah diagram, naik tajam ke kanan, menghasilkan fungsi yang akan cekung ke atas.

3.4 Penskalaan Ulang (*Rescalling*)

Ada kemungkinan bahwa skala pengukuran asli dapat mengaburkan fitur-fitur penting dalam sekumpulan data. Jika demikian, analisis dapat difasilitasi, atau dapat menghasilkan hasil yang lebih terbuka jika data terlebih dahulu mengalami transformasi matematis. Transformasi tersebut juga dapat sangat berguna untuk membantu data atmosfer sesuai dengan asumsi analisis regresi atau memungkinkan penerapan metode statistik multivariat yang dapat mengasumsikan distribusi Gaussian. Dalam terminologi analisis data eksplorasi, transformasi data tersebut dikenal sebagai ekspresi (penskalaan) ulang pada data.

3.4.1 Transformasi

Seringkali transformasi data dilakukan untuk membuat distribusi nilai lebih mendekati simetris dan hasilnya memungkinkan penggunaan teknik statistik yang lebih familiar dan tradisional. Terkadang transformasi penghasil simetri dapat membuat analisis eksplorasi, seperti

yang dijelaskan dalam bab ini, menjadi lebih terbuka. Transformasi ini juga dapat membantu dalam membandingkan kumpulan data yang berbeda, misalnya dengan meluruskan hubungan antara dua variabel. Kegunaan penting lainnya dari transformasi adalah membuat variasi atau dispersi (yaitu, penyebaran) dari satu variabel kurang bergantung pada nilai variabel lain, dalam hal ini transformasi disebut stabilisasi varians (*variance stabilizing*).

Tidak diragukan lagi, transformasi penghasil simetri yang paling umum digunakan (walaupun bukan satu-satunya yang mungkin) adalah transformasi pangkat yang didefinisikan oleh dua fungsi yang berkaitan erat.

$$T_1(x) = \begin{cases} x^{\lambda}, & \lambda > 0 \\ \ln(x), & \lambda = 0 \\ -(x^{\lambda}), & \lambda < 0 \end{cases} \qquad (0.16)$$

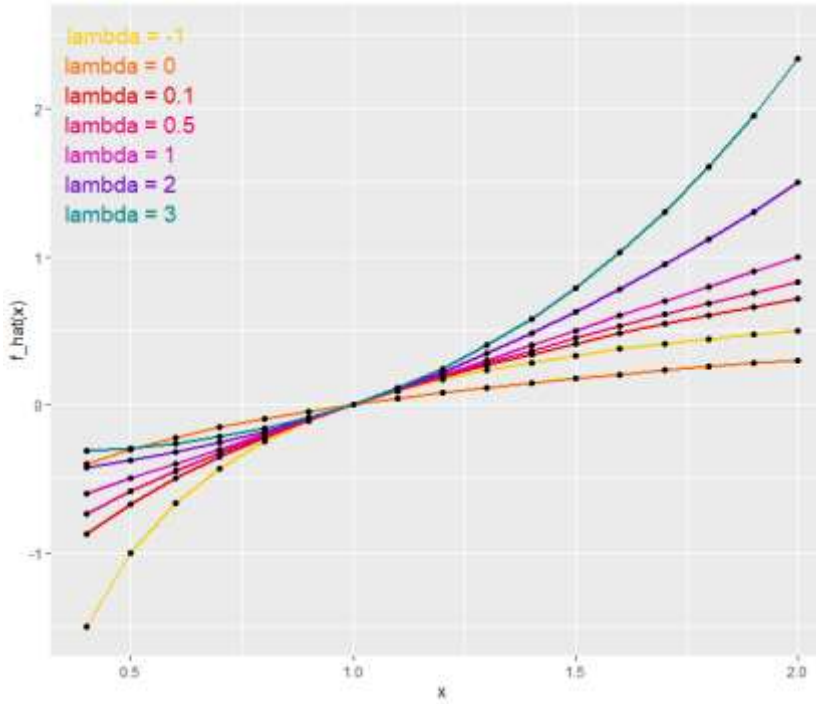
dan

$$T_2(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x), & \lambda = 0 \end{cases} \qquad (0.17)$$

Transformasi ini berguna ketika berhadapan dengan distribusi unimodal (*single-humped*) dari variabel data positif. Masing-masing fungsi ini mendefinisikan keluarga transformasi yang diindeks oleh parameter tunggal λ. Nama transformasi pangkat berasal dari fakta bahwa pekerjaan penting dari transformasi ini diselesaikan dengan eksponensial, atau menaikkan nilai data menjadi pangkat λ. Dengan demikian himpunan transformasi dalam Persamaan 0.16 dan 0.17 sebenarnya cukup sebanding dan nilai tertentu λ menghasilkan efek yang sama pada keseluruhan bentuk data dalam kedua kasus tersebut. Transformasi dalam Persamaan 3.16 memiliki bentuk yang sedikit lebih sederhana dan sering digunakan karena lebih mudah. Transformasi

dalam Persamaan 0.17, juga dikenal sebagai transformasi Box-Cox adalah versi sederhana dari Persamaan 0.16 yang digeser dan diskalakan, dan terkadang lebih berguna saat membandingkan berbagai transformasi. Juga, Persamaan 0.17 secara matematis lebih baik karena batas transformasi $\lambda \to 0$ sebenarnya adalah fungsi $\ln(x)$.

Untuk transformasi data yang menyertakan beberapa nilai nol atau negatif, rekomendasi dari Box dan Cox (1964) adalah memodifikasi transformasi dengan menambahkan konstanta positif ke setiap nilai data dengan besaran konstanta yang cukup besar untuk semua data yang akan digeser ke setengah positif dari garis nyata. Pendekatan yang mudah ini seringkali cukup, namun cukup acak. Yeo dan Johnson (2000) telah mengusulkan perluasan terpadu dari transformasi Box-Cox yang mengakomodasi data dimanapun pada garis nyata.



Gambar 3.10. Grafik transformasi pangkat pada Persamaan 0.17 untuk nilai parameter transformasi λ. Untuk $\lambda = 1$ transformasinya linier dan tidak menghasilkan perubahan bentuk data. Untuk $\lambda < 1$ transformasi mengurangi semua nilai data dan nilai yang lebih besar akan lebih terpengaruh. Efek sebaliknya dihasilkan oleh transformasi dengan $\lambda > 1$.

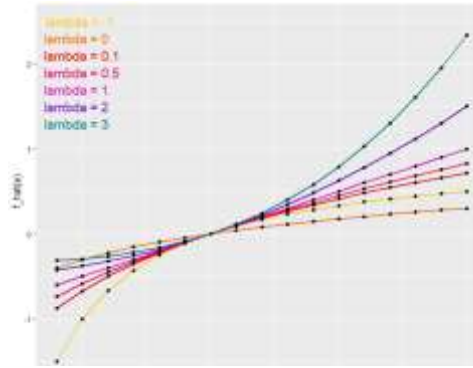
Implementasi dalam program R

Sintaks:

```
## Package R ##
## packages ##
library('ggplot2')

## lambda tidak sama dengan 0 ##
x <- seq(0.4,2, by = 0.1)
fx_1  <- ((x^(-1))  - 1)/(-1)
fx_0.1 <- ((x^(0.1)) - 1)/(0.1)
fx_0.5 <- ((x^(0.5)) - 1)/(0.5)
fx_1  <- ((x^(1))   - 1)/(1)
fx.2   <- ((x^(2))   - 1)/(2)
fx.3   <- ((x^(3))   - 1)/(3)

## lambda sama dengan 0 ##
fx_0   <- log10(x)
df     <- data.frame(fx._1,
                     fx.0,
                     fx_0.1,
                     fx.0.5,
                     fx.1,
                     fx.2,
                     fx.3,
                     x)
ggplot(df) +
  geom_line(aes(x=x, y=fx._1), size=1, color='#F1CD08') +
  geom_point(aes(x=x, y=fx._1)) +
  annotate(
    "text",
    label = "lambda = -1 ",
    x = 0.5, y = 2.5, size = 5, colour = "#F1CD08"
  ) +
  geom_line(aes(x=x, y=fx.0), size=1, color='#E67200') +
  geom_point(aes(x=x, y=fx.0)) +
  annotate(
    "text",
    label = "lambda = 0   ",
    x = 0.5, y = 2.3, size = 5, colour = "#E67200"
  ) +

  geom_line(aes(x=x, y=fx_0.1), size=1, color='#CC1300') +
  geom_point(aes(x=x, y=fx_0.1)) +
  annotate(
    "text",
    label = "lambda = 0.1 ",
    x = 0.5, y = 2.1, size = 5, colour = "#CC1300"
  ) +

  geom_line(aes(x=x, y=fx.0.5), size=1, color='#CC0070') +
  geom_point(aes(x=x, y=fx.0.5)) +
  annotate(
    "text",
    label = "lambda = 0.5",
    x = 0.5, y = 1.9, size = 5, colour = "#CC0070"
  ) +

  geom_line(aes(x=x, y=fx.1), size=1, color='#C407C5') +
  geom_point(aes(x=x, y=fx.1)) +
  annotate(
    "text",
    label = "lambda = 1   ",
    x = 0.5, y = 1.7, size = 5, colour = "#C407C5"
  ) +

  geom_line(aes(x=x, y=fx.2), size=1, color='#6B07C5') +
  geom_point(aes(x=x, y=fx.2)) +
  annotate(
    "text",
    label = "lambda = 2   ",
    x = 0.5, y = 1.5, size = 5, colour = "#6B07C5"
  ) +

  geom_line(aes(x=x, y=fx.3), size=1, color='#008080') +
  geom_point(aes(x=x, y=fx.3
```

```
+   annotate(
+     "text",
+     label = "lambda = 3   ",
+     x = 0.5, y = 1.3, size = 5, colour = "#008080"
+   ) +
+
+   ylab('f_hat(x)') +
+   xlab('x')
>
```

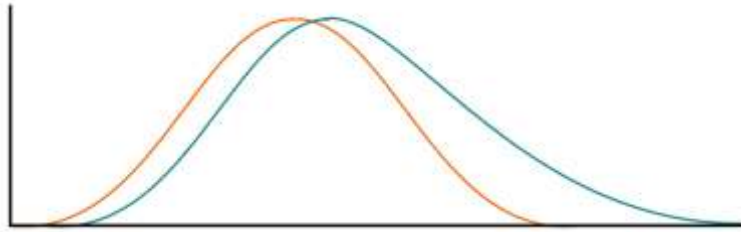
Output:



Dalam Persamaan 0.16 dan 0.17, menyesuaikan nilai parameter λ menghasilkan anggota spesifik dari serangkaian transformasi halus yang pada dasarnya terus bervariasi. Transformasi ini terkadang disebut sebagai "tangga pangkat". Beberapa dari fungsi transformasi ini diplot pada Gambar 3.10. Kurva pada gambar ini adalah fungsi yang ditentukan oleh Persamaan 0.17, meskipun kurva yang sesuai dari Persamaan 0.16 memiliki bentuk yang sama. Gambar 3.10 memperjelas bahwa penggunaan transformasi logaritmik untuk $\lambda = 0$ sangat cocok dengan spektrum transformasi pangkat. Gambar ini juga mengilustrasikan sifat lain dari transformasi pangkat, yaitu bahwa semuanya adalah fungsi naik dari variabel awal, x. Sifat ini dicapai dalam Persamaan 3.16 dengan tanda negatif pada transformasi dengan $\lambda < 1$.

Untuk transformasi pada Persamaan 0.16 pembalikan tanda ini dicapai dengan membaginya dengan λ. Properti transformasi pangkat yang meningkat secara ketat ini menyiratkan bahwa mereka mempertahankan urutan, sehingga nilai terkecil dalam kumpulan data asli akan sesuai dengan nilai terkecil dalam kumpulan data yang diubah, begitupun dengan nilai terbesar. Faktanya, akan ada korespondensi satu sama lain antara semua statistik urutan dari distribusi asli dan yang diubah. Jadi median, kuartil, dan seterusnya, dari data asli akan menjadi kuantil yang sesuai dari data yang diubah.

Jelas untuk $\lambda = 1$ data pada dasarnya tetap tidak berubah. Untuk $\lambda > 1$ nilai data dinaikkan (kecuali untuk pengurangan $\frac{1}{\lambda}$ dan pembagian dengan λ, jika Persamaan 0.17 digunakan), dengan nilai yang lebih besar dinaikkan lebih banyak daripada yang lebih kecil. Oleh karena itu transformasi pangkat dengan $\lambda > 1$ akan membantu menghasilkan kesimterisan bila diterapkan pada data yang miring negatif. Kebalikannya berlaku untuk $\lambda < 1$, dimana nilai data yang lebih besar berkurang lebih banyak daripada nilai yang lebih kecil. Oleh karena itu, transformasi pangkat dengan $\lambda < 1$ umumnya diterapkan pada data yang awalnya miring secara positif, untuk menghasilkan data yang hampir simetris.



Gambar 3.11. Pengaruh transformasi pangkat dengan $\lambda < 1$ pada kumpulan data dengan kemiringan positif (hijau). Panah menunjukkan bahwa transformasi memindahkan semua titik ke kiri, dengan nilai yang lebih besar dipindahkan lebih banyak. Distribusi yang dihasilkan (orange) cukup simetris.

Gambar 3.11 mengilustrasikan mekanisme dari proses ini untuk distribusi yang awalnya miring secara positif (hijau). Menerapkan transformasi pangkat dengan $\lambda < 1$ mengurangi semua nilai data, tetapi memengaruhi nilai yang lebih besar dengan lebih kuat. Pilihan yang tepat dari λ seringkali dapat menghasilkan setidaknya perkiraan simetri melalui proses ini. Memilih nilai yang terlalu kecil atau negatif untuk λ akan menghasilkan koreksi berlebih, yang mengakibatkan distribusi yang ditransformasi menjadi miring secara negatif.

Inspeksi awal plot data eksplorasi seperti diagram skematik dapat menunjukkan dengan cepat arah dan perkiraan besarnya kemiringan dalam kumpulan data. Oleh karena itu biasanya jelas apakah transformasi pangkat dengan $\lambda > 1$ atau $\lambda < 1$ sesuai, tetapi nilai spesifik untuk eksponen tidak akan begitu jelas. Sejumlah pendekatan untuk memilih parameter transformasi yang tepat telah disarankan.

Yang paling sederhana dari ini adalah statistik d_λ (Hinkley 1977),

$$d_\lambda = \frac{|\text{mean}(\lambda) - \text{median}(\lambda)|}{\text{spread}(\lambda)} \tag{0.18}$$

Di sini, sebaran (*spread*) adalah ukuran dispersi yang resisten, seperti IQR atau MAD. Setiap nilai λ akan menghasilkan rata-rata, median, dan sebaran yang berbeda dalam kumpulan data tertentu, dan ketergantungan pada λ ini ditunjukkan dalam persamaan. Hinkley d_λ digunakan untuk menentukan di antara transformasi pangkat yang dicobakan, dengan menghitung nilainya untuk masing-masing dari sejumlah pilihan yang berbeda untuk λ. Biasanya nilai-nilai percobaan λ diberi jarak dengan interval 1/2 atau 1/4. Pemilihan λ ditentukan dengan d_λ terkecil kemudian diadopsi untuk mengubah data (transformasi data). Salah satu cara yang sangat mudah untuk melakukan perhitungan adalah dengan menggunakan bantuan komputasi.

Dasar dari statistik d_λ adalah untuk data yang terdistribusi secara simetris, rata-rata dan median akan sangat dekat. Oleh karena itu, seiring transformasi pangkat yang lebih kuat secara berturut-turut (nilai λ semakin jauh dari 1) memindahkan data ke arah simetri, pembilang dalam Persamaan 0.18 akan bergerak menuju nol. Karena transformasi menjadi terlalu kuat, pembilang akan mulai meningkat relatif terhadap ukuran sebaran, menghasilkan peningkatan lagi.

Persamaan 0.18 adalah pendekatan sederhana untuk menemukan transformasi pangkat yang menghasilkan simetris atau mendekati simetris dalam data yang ditransformasikan. Pendekatan yang lebih canggih disarankan dalam makalah Box and Cox (1964), yang sangat sesuai ketika data yang diubah harus memiliki distribusi sedekat mungkin dengan Gaussian berbentuk lonceng,

misalnya ketika hasil dari beberapa transformasi akan dirangkum secara bersamaan melalui multivariat Gaussian, atau distribusi normal multivariat. Secara khusus, Box dan Cox menyarankan memilih eksponen transformasi pangkat untuk memaksimalkan fungsi log-likelihood untuk distribusi Gaussian

$$L(\lambda) = -\frac{n}{2}\ln[s^2(\lambda)] + (\lambda - 1)\sum_{i=1}^{n}\ln[x_i] \qquad (0.19)$$

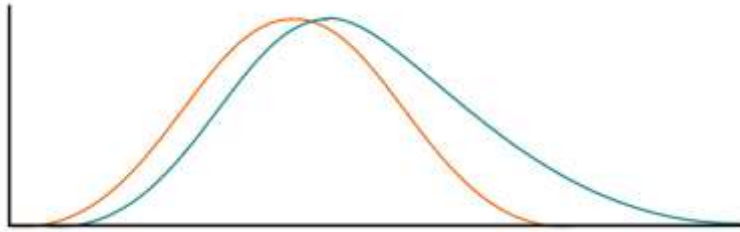
Di sini n adalah ukuran sampel dan s^2 adalah varians sampel (dihitung dengan pembagi n, bukan $n-1$) dari data setelah transformasi dengan eksponen λ. Seperti kasus untuk menggunakan statistik Hinkley (Persamaan 0.18), nilai yang berbeda dari λ dicoba dan yang menghasilkan nilai $L(\lambda)$ terbesar dipilih sebagai yang paling tepat. Ada kemungkinan bahwa dua kriteria akan menghasilkan pilihan yang berbeda karena Persamaan 0.18 hanya membahas simetris data yang diubah, sedangkan Persamaan 0.19 mencoba mengakomodasi semua aspek distribusi Gaussian, termasuk namun tidak terbatas pada simetrinya.

Contoh 0.4 Memilih Transformasi Pangkat yang Tepat

Tabel 3.3 menunjukkan data curah hujan Ithaca Januari 1933–1982 dari Tabel A.2 di Lampiran A, diurutkan dalam urutan menaik dan mengalami transformasi pangkat menggunakan Persamaan 1.2, untuk $\lambda = 1$, $\lambda = 0.5$, $\lambda = 0$, dan $\lambda = -0.5$. Untuk $\lambda = 1$ transformasi hanya berarti mengurangkan 1 dari setiap nilai data. Perhatikan bahwa bahkan untuk eksponen negatif $\lambda = -0.5$ urutan data asli dipertahankan dalam semua transformasi, sehingga mudah untuk menentukan median dan kuartil dari data asli dan data yang diubah.

Tabel 3.3. Curah hujan Ithaca Januari 1933–1982 dari Tabel A.2 ($\lambda = 1$). Data telah disortir dengan transformasi pangkat pada Persamaan 3.17 diterapkan untuk $\lambda = 1$, $\lambda = 0.5$, $\lambda = 0$, dan $\lambda = -0.5$. Plot skematis dari data ini ditunjukkan pada Gambar 3.11.

Year	$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0$	$\lambda = -0.5$	Year	$\lambda = 1$	$\lambda = 0.5$	$\lambda = 0$	$\lambda = -0.$
1933	-0.56	-0.67	-0.82	-1.02	1948	0.72	0.62	0.54	0.48
1980	-0.48	-0.56	-0.65	-0.77	1960	0.75	0.65	0.56	0.49
1944	-0.46	-0.53	-0.62	-0.72	1964	0.76	0.65	0.57	0.49
1940	-0.28	-0.30	-0.33	-0.36	1974	0.84	0.71	0.61	0.53
1981	-0.13	-0.13	-0.14	-0.14	1962	0.88	0.74	0.63	0.54
1970	0.03	0.03	0.03	0.03	1951	0.98	0.81	0.68	0.58
1971	0.11	0.11	0.10	0.10	1954	1.00	0.83	0.69	0.59
1955	0.12	0.12	0.11	0.11	1936	1.08	0.88	0.73	0.61
1946	0.13	0.13	0.12	0.12	1956	1.13	0.92	0.76	0.63
1967	0.16	0.15	0.15	0.14	1965	1.17	0.95	0.77	0.64
1934	0.18	0.17	0.17	0.16	1949	1.27	1.01	0.82	0.67
1942	0.30	0.28	0.26	0.25	1966	1.38	1.09	0.87	0.70
1963	0.31	0.29	0.27	0.25	1952	1.44	1.12	0.89	0.72
1943	0.35	0.32	0.30	0.28	1947	1.50	1.16	0.92	0.74
1972	0.35	0.32	0.30	0.28	1953	1.53	1.18	0.93	0.74
1957	0.36	0.33	0.31	0.29	1935	1.69	1.28	0.99	0.78
1969	0.36	0.33	0.31	0.29	1945	1.74	1.31	1.01	0.79
1977	0.36	0.33	0.31	0.29	1939	1.82	1.36	1.04	0.81
1968	0.39	0.36	0.33	0.30	1950	1.82	1.36	1.04	0.81
1973	0.44	0.40	0.36	0.33	1959	1.94	1.43	1.08	0.83
1941	0.46	0.42	0.38	0.34	1976	2.00	1.46	1.10	0.85
1982	0.51	0.46	0.41	0.37	1937	2.66	1.83	1.30	0.95
1961	0.69	0.60	0.52	0.46	1979	3.55	2.27	1.52	1.06
1975	0.69	0.60	0.52	0.46	1958	3.90	2.43	1.59	1.10
1938	0.72	0.62	0.54	0.48	1978	5.37	3.05	1.85	1.21



Gambar 3.12. Pengaruh transformasi pangkat dalam Persamaan 0.17 pada data curah hujan total bulan Januari untuk Ithaca, 1933–1982 (Tabel A.2). Data asli ($\lambda = 1$) sangat miring ke kanan, dengan nilai terbesar berada jauh di luar. Transformasi akar kuadrat ($\lambda = 0.5$) sedikit meningkatkan simetri. Transformasi logaritmik ($\lambda = 0$) menghasilkan distribusi yang cukup simetris. Ketika mengalami transformasi akar kuadrat terbalik yang lebih ekstrem ($\lambda = -0.5$) data mulai menunjukkan kemiringan negatif. Transformasi logaritmik akan dipilih sebagai yang terbaik berdasarkan statistik Hinkley d_λ (Persamaan 0.18), dan log-likelihood Gaussian (Persamaan 0.19).

Gambar 3.12 menunjukkan plot skematik untuk data pada Tabel 3.3. Data yang tidak diubah (plot paling kiri) jelas miring positif, yang biasa dijumpai pada distribusi jumlah curah hujan. Ketiga nilai di luar pagar adalah jumlah yang besar, dengan yang terbesar berada jauh di luar. Tiga plot skematik lainnya menunjukkan hasil transformasi pangkat yang semakin kuat dengan $\lambda < 1$. Transformasi logaritmik ($\lambda = 0$) keduanya meminimalkan statistik Hinkley d_λ (Persamaan 0.18) dengan IQR sebagai ukuran penyebaran,

dan memaksimalkan log-likelihood Gaussian (Persamaan 0.19). Kesimetrisan dekat yang ditunjukkan oleh plot skematik untuk data yang diubah secara logaritmik mendukung kesimpulan bahwa ini adalah yang terbaik di antara kemungkinan yang dipertimbangkan menurut kedua kriteria. Transformasi akar-akar terbalik yang lebih ekstrem ($\lambda = -0.5$) jelas telah dikoreksi berlebihan untuk kemiringan positif, karena tiga jumlah terkecil sekarang berada di luar pagar bawah.

3.4.2 Standarisasi Anomali (Normalisasi)

Transformasi juga dapat bermanfaat saat tertarik untuk bekerja secara bersamaan dengan kumpulan data yang terkait, tetapi tidak dapat dibandingkan secara ketat. Salah satu contoh dari situasi ini terjadi ketika data bergantung pada variasi musiman. Misalnya, perbandingan langsung suhu mentah bulanan biasanya akan menunjukkan pengaruh yang mendominasi dalam siklus musiman. Rekor Januari yang hangat masih akan jauh lebih dingin daripada rekor Juli yang sejuk. Dalam situasi semacam ini, penskalaan ulang data dalam hal ini standarisasi anomali bisa sangat membantu.
Standarisasi anomali, z, dihitung hanya dengan mengurangkan rata-rata sampel pada data mentah x, dan membaginya dengan standar deviasi sampel yang sesuai:

$$z = \frac{x - \bar{x}}{s_x} = \frac{x'}{s_x} \tag{0.20}$$

Dalam ilmu atmosfer, anomali x' dipahami sebagai pengurangan dari nilai data rata-rata yang relevan, seperti pada pembilang Persamaan 0.20. Istilah anomali tidak berkonotasi dengan nilai data atau peristiwa yang tidak normal atau bahkan tidak biasa. Standarisasi anomali pada Persamaan 0.20 dihasilkan dengan membagi anomali pada pembilang dengan standar deviasi yang sesuai. Transformasi

ini terkadang juga disebut sebagai normalisasi. Dimungkinkan juga untuk membangun Standarisasi anomali menggunakan ukuran lokasi dan penyebaran yang resisten, misalnya, mengurangkan median dan membaginya dengan IQR, namun ini jarang dilakukan. Penggunaan standarisasi anomali dimotivasi oleh ide-ide yang diturunkan dari distribusi Gaussian berbentuk lonceng. Namun, tidak perlu berasumsi bahwa kumpulan data mengikuti distribusi tertentu untuk mengekspresikannya kembali dalam bentuk standarisasi anomali dan mengubah data non-Gaussian menurut Persamaan 3.20 tidak akan membuatnya menjadi Gaussian kembali.

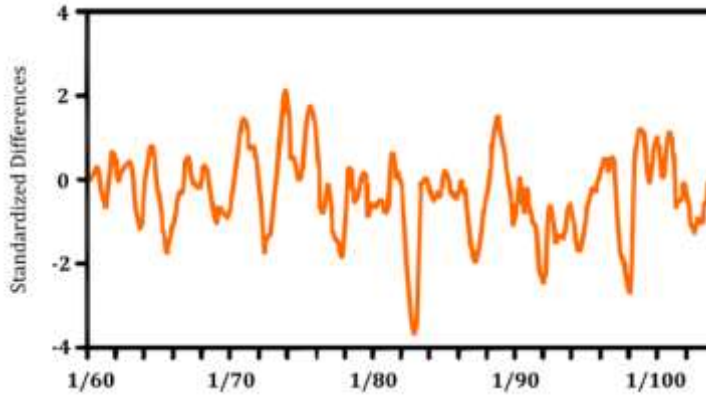
Ide di balik Standarisasi anomali adalah mencoba menghilangkan pengaruh lokasi dan penyebaran dari sekumpulan data. Unit fisik dari data asli dibatalkan, sehingga standarisasi anomali selalu merupakan kuantitas tanpa dimensi. Pengurangan rata-rata menghasilkan serangkaian anomali, x', yang terletak di suatu tempat mendekati nol. Pembagian dengan standar deviasi menempatkan ekskursi dari rata-rata dalam kumpulan data yang berbeda pada pijakan yang sama. Secara kolektif, kumpulan data yang telah diubah menjadi sekumpulan standarisasi anomali akan menunjukkan rata-rata nol dan standar deviasi 1.

Misalnya, sering ditemukan bahwa suhu musim panas kurang bervariasi dibandingkan suhu musim dingin. Mungkin ditemukan bahwa standar deviasi untuk suhu rata-rata bulan Januari di beberapa lokasi adalah sekitar 3°C, tetapi standar deviasi untuk suhu rata-rata bulan Juli di lokasi yang sama mendekati 1°C. Suhu rata-rata bulan Juli 3°C lebih dingin daripada rata-rata jangka panjang untuk bulan Juli akan sangat tidak biasa, sesuai dengan standarisasi anomali standar -3. Suhu rata-rata Januari 3°C lebih hangat daripada suhu rata-rata Januari jangka panjang di lokasi yang sama

akan menjadi kejadian yang cukup biasa, sesuai dengan standarisasi anomali +1. Cara lain untuk melihat standarisasi anomali adalah sebagai ukuran jarak dalam satuan standar deviasi antara nilai data dan rata-ratanya.

Contoh 0.5 Mengekspresikan Data Iklim dalam Standarisasi Anomali

Gambar 3.12 mengilustrasikan penggunaan standarisasi anomali dalam konteks operasional. Titik yang diplot adalah nilai indeks Indeks Osilasi Selatan yang merupakan indeks fenomena El-Niño-Osilasi Selatan (ENSO) yang digunakan oleh Pusat Prediksi Iklim dari Pusat Prediksi Lingkungan Nasional AS (Ropelewski dan Jones, 1987). Nilai indeks ini dalam gambar diperoleh dari perbedaan bulan demi bulan dalam standarisasi anomali tekanan permukaan laut di dua lokasi tropis: Tahiti, di tengah Samudra Pasifik; dan Darwin, di Australia utara. Dalam hal Persamaan 0.20, langkah pertama untuk menghasilkan Gambar 3.12 adalah menghitung perbedaan $\Delta z = z_{Tahiti} - z_{Darwin}$ untuk setiap bulan selama tahun-tahun yang diplot. Standarisasi anomali z_{Tahiti} untuk Januari 1960, misalnya, dihitung dengan mengurangkan tekanan rata-rata untuk semua Januari di Tahiti dari tekanan bulanan yang diamati untuk Januari 1960. Perbedaan ini kemudian dibagi dengan standar deviasi yang mencirikan variasi tahun-ke-tahun dari Tekanan atmosfer Januari di Tahiti.



Gambar 3.13. Standarisasi selisih antar standarisasi anomali tekanan permukaan laut bulanan di Tahiti dan Darwin (Indeks Osilasi Selatan), 1960–2002. Nilai bulanan individu telah dihaluskan dalam waktu.

Sebenarnya, kurva pada Gambar 3.13 didasarkan pada nilai bulanan yang merupakan standarisasi anomali dari perbedaan standarisasi anomali Δz, sehingga Persamaan 0.20 telah diterapkan dua kali pada data asli. Yang pertama dari dua standarisasi dilakukan untuk meminimalkan pengaruh perubahan musiman pada tekanan bulanan rata-rata dan variabilitas tekanan bulanan dari tahun ke tahun. Standardisasi kedua, menghitung standarisasi anomali dari perbedaan Δz, memastikan bahwa indeks yang dihasilkan akan memiliki satuan standar deviasi.

Secara fisik, selama peristiwa El Niño, pusat aktivitas curah hujan Pasifik tropis bergeser ke arah timur dari Pasifik barat (dekat Darwin) ke Pasifik tengah (dekat Tahiti). Pergeseran ini dikaitkan dengan tekanan permukaan yang lebih tinggi dari rata-rata di Darwin dan tekanan permukaan yang lebih rendah dari rata-rata di Tahiti, yang bersama-sama

menghasilkan nilai negatif untuk indeks yang diplot pada Gambar 3.12. Peristiwa El-Niño yang sangat kuat pada tahun 1982–1983 sangat menonjol dalam gambar ini.

3.5 Latihan

1. Bandingkan median, trimester, dan rata-rata data curah hujan pada Tabel A.3.
2. Hitung MAD, IQR dan simpangan baku dari data pada Tabel A.3.
3. Gambarlah plot histogram dan boxplot untuk data suhu pada Tabel A.3.
4. Hitung indeks Yule-Kendall dan koefisien skewness dengan menggunakan data temperatur pada Tabel A.3.
5. Plot distribusi frekuensi kumulatif empiris untuk data pada Tabel A.3. Bandingkan dengan histogram dari data yang sama.
6. Bandingkan boxplot dan plot skematis yang mewakili data curah hujan pada Tabel A.3.

DAFTAR PUSTAKA

Abramowitz, M., and I.A. Stegun, eds., 1984. *Pocketbook of Mathematical Functions. Frankfurt*, Verlag Harri Deutsch, 468 pp.

Agresti, A., 1996. *An Introduction to Categorical Data Analysis*. Wiley, 290 pp. Agresti, A., and B.A. Coull, 1998. Approximate is better than "exact" for interval estimation of binomial proportions. American Statistician, 52, 119–126.

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

Allen, M.R., and L.A. Smith, 1996. Monte Carlo SSA: Detecting irregular oscillations in the presence of colored noise. *Journal of Climate*, 9, 3373–3404.

Anderson, J.L., 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9, 1518–1530.

Anderson, J.L., 1997. The impact of dynamical constraints on the selection of initial conditions for ensemble predictions: low-order perfect model results. *Monthly Weather Review*, 125, 2969–2983.

Buizza, R., M. Miller, and T.N. Palmer, 1999b. Stochastic representation of model uncer- tainties in the ECMWF Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908.

Burman, P., E. Chow, and D. Nolan, 1994. A cross-validatory method for dependent data. *Biometrika*, 81, 351–358.

Carter, G.M., J.P. Dallavalle, and H.R. Glahn, 1989. Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Weather and Forecasting*, 4, 401–412.

Cunnane, C., 1978. Unbiased plotting positionsa review. *Journal of Hydrology*, 37, 205–222.

Daan, H., 1985. Sensitivity of verification scores to the classification of the predictand. *Monthly Weather Review*, 113, 1384–1392.

Ehrendorfer, M., 1997. *Predicting the uncertainty of numerical weather forecasts: a review*. Meteorol. Zeitschrift, 6, 147–183.

Gleeson, T.A., 1970. Statistical-dynamical predictions. *Journal of Applied Meteorology*, 9, 333–344.

Goldsmith, B.S., 1990. *NWS verification of precipitation type and snow amount forecasts during the AFOS era*. NOAA Technical Memorandum NWS FCST 33. National Weather Service, 28 pp.

Golub, G.H., and C.F. van Loan, 1996. *Matrix Computations. Johns Hopkins Press*, 694 pp.

Houtekamer, P.L., L. Lefaivre, J. Derome, H. Ritchie, and H.L. Mitchell, 1996. A system simulation approach to ensemble prediction. *Monthly Weather Review*, 124, 1225–1242.

Houtekamer, P.L., and H.L. Mitchell, 1998. Data assimilation using an ensemble Kalman filtering technique. *Monthly Weather Review*, 126, 796–811.

Hsu, W.-R., and A.H. Murphy, 1986. The attributes diagram: a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, 2, 285–293.

Hu, Q., 1997. On the uniqueness of the singular value decomposition in meteorological applications. *Journal of Climate*, 10, 1762–1766.

Klein, W.H., B.M. Lewis, and I. Enger, 1959. Objective prediction of five-day mean temperature during winter. *Journal of Meteorology*, 16, 672–682.

Knaff, J.A., and C.W. Landsea, 1997. An El Niño-southern oscillation climatology and persistence (CLIPER) forecasting scheme. *Weather and Forecasting*, 12, 633–647.

Krzysztofowicz, R., 1983. *Why should a forecaster and a decision maker use Bayes' theorem? Water Resources Research*, 19, 327–336.

Krzysztofowicz, R., 2001. The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, 249, 2–9.

Lin, J. W.-B., and J.D. Neelin, 2002. *Considerations for stochastic convective parameterization. Journal of the Atmospheric Sciences, 59, 959–975. Lindgren, B.W., 1976*. Statistical Theory. MacMillan, 614 pp.

Lipschutz, S., 1968. *Schaum's Outline of Theory and Problems of Linear Algebra*. McGraw-Hill, 334 pp.

Livezey, R.E., 1985. Statistical analysis of general circulation model climate simula- tion: sensitivity and prediction experiments. *Journal of the Atmospheric Sciences*, 42, 1139–1149.

Mylne, K.R., C. Woolcock, J.C.W. Denholm-Price, and R.J. Darvell, 2002b. Operational calibrated probability forecasts from the ECMWF ensemble prediction system: imple- mentation and verification. Preprints, Symposium on Observations, Data Analysis, and Probabilistic Prediction, (Orlando, Florida), *American Meteorological Society*, 113–118.

Namias, J., 1952. The annual course of month-to-month persistence in climatic anomalies. *Bulletin of the American Meteorological Society*, 33, 279–285.

National Bureau of Standards, 1959. *Tables of the Bivariate Normal Distribution Function and Related Functions*. Applied Mathematics Series, 50, U.S. Government Printing Office, 258 pp.

Nehrkorn, T., R.N. Hoffman, C. Grassotti, and J.-F. Louis, 2003: Feature calibration and alignment to represent model forecast errors: Empirical regularization. *Quarterly Journal of the Royal Meteorological Society*, 129, 195–218.

Penland, C., and P.D. Sardeshmukh, 1995. The optimal growth of tropical sea surface temperatures anomalies. *Journal of Climate*, 8, 1999–2024.

Peterson, C.R., K.J. Snapper, and A.H. Murphy 1972. Credible interval temperature forecasts. Bulletin of the American Meteorological Society, 53, 966–970.

Pitcher, E.J., 1977. Application of stochastic dynamic prediction to real data. *Journal of the Atmospheric Sciences*, 34, 3–21.

Pitman, E.J.G., 1937. Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society*, B4, 119–130.

Plaut, G., and R. Vautard, 1994. Spells of low-frequency oscillations and weather regimes in the Northern Hemisphere. *Journal of the Atmospheric Sciences*, 51, 210–236.

Tukey, J.W., 1977. *Exploratory Data Analysis*. Reading, Mass., Addison-Wesley, 688 pp.

Unger, D.A., 1985*. A method to estimate the continuous ranked probability score*. Preprints, 9th Conference on Probability and Statistics in the Atmospheric Sciences. American Meteorological Society, 206–213.

Zwiers, F.W., and H.J. Thiébaux, 1987. Statistical considerations for climate experiments. Part I: scalar tests. *Journal of Climate and Applied Meteorology*, 26, 465–476.

Zwiers, F.W., and H. von Storch, 1995. Taking serial correlation into account in tests of the mean. *Journal of Climate*, 8, 336–351.

3.6 Lampiran

Tabel A

Kumpulan Contoh Data

Dalam aplikasi nyata dari analisis data klimatologi, kami berharap dapat menggunakan lebih banyak data (misalnya semua data harian Januari yang tersedia, bukan hanya data untuk satu tahun) dan membiarkan komputer melakukan perhitungan. Kumpulan data kecil ini digunakan dalam beberapa contoh di seluruh buku ini sehingga perhitungan dapat dilakukan dengan tangan dan pemahaman yang lebih jelas tentang prosedur dapat dicapai.

Tabel A.1 Pengamatan curah hujan harian (inci) dan suhu (°F) di Ithaca dan Canandaigua, New York, untuk Januari 1987.

Date	Ithaca			Canandaigua		
	Precipitation	Max Temp.	Min Temp.	Precipitation	Max Temp.	Min Temp.
1	0.00	33	19	0.00	34	28
2	0.07	32	25	0.04	36	28
3	1.11	30	22	0.84	30	26
4	0.00	29	−1	0.00	29	19
5	0.00	25	4	0.00	30	16
6	0.00	30	14	0.00	35	24
7	0.00	37	21	0.02	44	26
8	0.04	37	22	0.05	38	24
9	0.02	29	23	0.01	31	24
10	0.05	30	27	0.09	33	29
11	0.34	36	29	0.18	39	29
12	0.06	32	25	0.04	33	27
13	0.18	33	29	0.04	34	31
14	0.02	34	15	0.00	39	26

Date	Ithaca			Canandaigua		
	Precipitation	Max Temp.	Min Temp.	Precipitation	Max Temp.	Min Temp.
15	0.02	53	29	0.06	51	38
16	0.00	45	24	0.03	44	23
17	0.00	25	0	0.04	25	13
18	0.00	28	2	0.00	34	14
19	0.00	32	26	0.00	36	28
20	0.45	27	17	0.35	29	19
21	0.00	26	19	0.02	27	19
22	0.00	28	9	0.01	29	17
23	0.70	24	20	0.35	27	22
24	0.00	26	−6	0.08	24	2
25	0.00	9	−13	0.00	11	4
26	0.00	22	−13	0.00	21	5
27	0.00	17	−11	0.00	19	7
28	0.00	26	−4	0.00	26	8
29	0.01	27	−4	0.01	28	14
30	0.03	30	11	0.01	31	14
31	0.05	34	23	0.13	38	23
sum/avg.	3.15	29.87	13	2.4	31.77	20.23
std. dev.	0.243	7.71	13.62	0.168	7.86	8.81

Tabel A.2 Curah hujan Januari di Ithaca, New York, 1933–1982, inci.

1933	0.44	1945	2.74	1958	4.9	1970	1.03
1934	1.18	1946	1.13	1959	2.94	1971	1.11
1935	2.69	1947	2.5	1960	1.75	1972	1.35
1936	2.08	1948	1.72	1961	1.69	1973	1.44
1937	3.66	1949	2.27	1962	1.88	1974	1.84
1938	1.72	1950	2.82	1963	1.31	1975	1.69
1939	2.82	1951	1.98	1964	1.76	1976	3
1940	0.72	1952	2.44	1965	2.17	1977	1.36
1941	1.46	1953	2.53	1966	2.38	1978	6.37
1942	1.3	1954	2	1967	1.16	1979	4.55
1943	1.35	1955	1.12	1968	1.39	1980	0.52
1944	0.54	1956	2.13	1969	1.36	1981	0.87
		1957	1.36			1982	1.51

Tabel A.3 Data Iklim Juni untuk Guayaquil, Ekuador, 1951-1970. Tanda bintang menunjukkan tahun El Niño.

1951∗	26.1	43	1009.5
1952	24.5	10	1010.9
1953∗	24.8	4	1010.7
1954	24.5	0	1011.2
1955	24.1	2	1011.9
1956	24.3	Missing	1011.2
1957∗	26.4	31	1009.3
1958	24.9	0	1011.1
1959	23.7	0	1012
1960	23.5	0	1011.4
1961	24	2	1010.9
1962	24.1	3	1011.5
1963	23.7	0	1011
1964	24.3	4	1011.2
1965∗	26.6	15	1009.9
1966	24.6	2	1012.5
1967	24.8	0	1011.1
1968	24.4	1	1011.8
1969∗	26.8	127	1009.3
1970	25.2	2	1010.6

Tabel B

Tabel Probabilitas

Tabel probabilitas untuk distribusi probabilitas bersama terpilih yang tidak mengikuti ekspresi tertutup untuk fungsi distribusi kumulatif.

Tabel B.1. Probabilitas kumulatif arah kiri untuk distribusi Gaussian standar, $\phi(z) = \Pr\{Z \leq z\}$. Nilai-nilai variabel Gaussian standar z dicantumkan hingga sepersepuluh terdekat di kolom paling kanan dan paling kiri. Judul kolom yang tersisa menunjukkan tempat keseratus mis. Probabilitas sisi kanan diperoleh dengan $\Pr\{Z > z\} = 1 - \Pr\{Z \leq z\}$. Probabilitas untuk $Z > 0$ diperoleh dengan menggunakan simetri distribusi Gaussian, $\Pr\{Z \leq z\} = 1 - \Pr\{Z \leq -z\}$.

Z	0.0900	0.0800	0.0700	0.0600	0.0500	0.0400	0.0300	0.0200	0.0100	0.0000	Z
−4.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	−4.0
−3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	−3.9
−3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	−3.8
−3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	−3.7
−3.6	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002	−3.6
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	−3.5
−3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
−3.3	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3

−3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
−3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
−3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0014	−3.0
−2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
−2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
−2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
−2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0042	0.0043	0.0044	0.0045	0.0047	−2.6
−2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
−2.4	0.0064	0.0066	0.0068	0.0070	0.0071	0.0073	0.0076	0.0078	0.0080	0.0082	−2.4
−2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
−2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0126	0.0129	0.0132	0.0136	0.0139	−2.2
−2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
−2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
−1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
−1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0352	0.0359	−1.8
−1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
−1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
−1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5
−1.4	0.0681	0.0694	0.0708	0.0722	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	−1.4

−1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	−1.3
−1.2	0.0985	0.1003	0.1020	0.1038	0.1057	0.1075	0.1094	0.1112	0.1131	0.1151	−1.2
−1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	−1.1
−1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1563	0.1587	−1.0
−0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	−0.9
−0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	−0.8
−0.7	0.2148	0.2177	0.2207	0.2236	0.2266	0.2297	0.2327	0.2358	0.2389	0.2420	−0.7
−0.6	0.2451	0.2483	0.2514	0.2546	0.2579	0.2611	0.2644	0.2676	0.2709	0.2743	−0.6
−0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	−0.5
−0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	−0.4
−0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	−0.3
−0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4091	0.4129	0.4168	0.4207	−0.2
−0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	−0.1
−0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4841	0.4880	0.4920	0.4960	0.5000	0

Tabel B.2. Kuantil distribusi gamma standar ($\boldsymbol{\beta = 1}$). Item yang ditabulasikan adalah nilai dari variabel acak standar ξ yang sesuai dengan probabilitas kumulatif $F(\xi)$ yang diberikan dalam judul kolom untuk nilai parameter bentuk (α) yang diberikan pada kolom pertama. Untuk menemukan kuantil untuk distribusi dengan parameter skala lainnya, masukkan tabel pada baris yang sesuai, baca nilai standar pada kolom yang sesuai, dan kalikan nilai tabel dengan parameter skala. Untuk mengekstrak probabilitas kumulatif yang sesuai dengan nilai tertentu dari variabel acak, bagi nilai dengan parameter skala, masukkan tabel di baris yang sesuai dengan parameter bentuk, dan interpolasi hasil dari judul kolom

Cumulative Probability															
a	0.001	0.01	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99	0.999
0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.077	0.262	1.057	2.423
0.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.018	0.070	0.264	0.575	1.554	3.035
0.2	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.006	0.021	0.062	0.164	0.442	0.820	1.894	3.439
0.2	0.000	0.000	0.000	0.000	0.000	0.002	0.007	0.021	0.053	0.122	0.265	0.602	1.024	2.164	3.756
0.3	0.000	0.000	0.000	0.000	0.001	0.006	0.018	0.044	0.095	0.188	0.364	0.747	1.203	2.395	4.024
0.3	0.000	0.000	0.000	0.000	0.003	0.013	0.034	0.073	0.142	0.257	0.461	0.882	1.365	2.599	4.262
0.4	0.000	0.000	0.000	0.001	0.007	0.024	0.055	0.108	0.192	0.328	0.556	1.007	1.515	2.785	4.477

Cumulative Probability															
0.4	0.000	0.000	0.000	0.002	0.013	0.038	0.080	0.145	0.245	0.398	0.644	1.126	1.654	2.958	4.677
0.5	0.000	0.000	0.001	0.005	0.022	0.055	0.107	0.186	0.300	0.468	0.733	1.240	1.786	3.121	4.863
0.5	0.000	0.000	0.002	0.008	0.032	0.074	0.138	0.228	0.355	0.538	0.819	1.349	1.913	3.274	5.040
0.6	0.000	0.000	0.004	0.012	0.045	0.096	0.170	0.272	0.411	0.607	0.904	1.454	2.034	3.421	5.208
0.6	0.000	0.000	0.006	0.018	0.059	0.120	0.204	0.316	0.467	0.676	0.987	1.556	2.150	3.562	5.370
0.7	0.000	0.001	0.009	0.025	0.075	0.146	0.240	0.362	0.523	0.744	1.068	1.656	2.264	3.698	5.526
0.7	0.000	0.001	0.012	0.033	0.093	0.173	0.276	0.408	0.579	0.811	1.149	1.753	2.374	3.830	5.676
0.8	0.000	0.002	0.017	0.043	0.112	0.201	0.314	0.455	0.636	0.878	1.227	1.848	2.481	3.958	5.822
0.8	0.000	0.003	0.022	0.053	0.132	0.231	0.352	0.502	0.692	0.945	1.305	1.941	2.586	4.083	5.964
0.9	0.000	0.004	0.028	0.065	0.153	0.261	0.391	0.550	0.749	1.010	1.382	2.032	2.689	4.205	6.103
0.9	0.001	0.006	0.035	0.078	0.176	0.292	0.431	0.598	0.805	1.076	1.458	2.122	2.790	4.325	6.239
1	0.001	0.008	0.043	0.091	0.199	0.324	0.471	0.646	0.861	1.141	1.533	2.211	2.888	4.441	6.373
1	0.001	0.011	0.052	0.106	0.224	0.357	0.512	0.694	0.918	1.206	1.607	2.298	2.986	4.556	6.503
1.1	0.002	0.013	0.061	0.121	0.249	0.391	0.553	0.742	0.974	1.270	1.681	2.384	3.082	4.669	6.631
1.1	0.002	0.017	0.071	0.138	0.275	0.425	0.594	0.791	1.030	1.334	1.759	2.469	3.177	4.781	6.757
1.2	0.002	0.020	0.082	0.155	0.301	0.459	0.636	0.840	1.086	1.397	1.831	2.553	3.270	4.890	6.881
1.2	0.002	0.024	0.094	0.173	0.329	0.494	0.678	0.889	1.141	1.460	1.903	2.636	3.362	4.998	7.003
1.3	0.003	0.027	0.106	0.191	0.357	0.530	0.720	0.938	1.197	1.523	1.974	2.719	3.453	5.105	7.124

Cumulative Probability															
1.3	0.004	0.032	0.119	0.210	0.385	0.566	0.763	0.987	1.253	1.586	2.045	2.800	3.544	5.211	7.242
1.4	0.004	0.037	0.133	0.230	0.414	0.602	0.806	1.036	1.308	1.649	2.115	2.881	3.633	5.314	7.360
1.4	0.005	0.043	0.145	0.250	0.443	0.639	0.849	1.085	1.364	1.711	2.185	2.961	3.722	5.418	7.476
1.5	0.007	0.049	0.160	0.272	0.473	0.676	0.892	1.135	1.419	1.773	2.255	3.041	3.809	5.519	7.590
1.5	0.008	0.056	0.175	0.293	0.504	0.713	0.935	1.184	1.474	1.834	2.324	3.120	3.897	5.620	7.704
1.6	0.011	0.063	0.191	0.313	0.534	0.750	0.979	1.234	1.530	1.896	2.392	3.199	3.983	5.720	7.816
1.6	0.014	0.071	0.207	0.336	0.565	0.788	1.023	1.283	1.585	1.957	2.461	3.276	4.068	5.818	7.928
1.7	0.018	0.078	0.224	0.359	0.597	0.826	1.067	1.333	1.640	2.018	2.529	3.354	4.153	5.917	8.038
1.7	0.023	0.087	0.241	0.382	0.628	0.865	1.111	1.382	1.695	2.079	2.597	3.431	4.237	6.014	8.147
1.8	0.031	0.096	0.259	0.406	0.661	0.903	1.155	1.432	1.750	2.140	2.664	3.507	4.321	6.110	8.255
1.8	0.036	0.104	0.277	0.430	0.693	0.942	1.199	1.481	1.805	2.200	2.731	3.584	4.405	6.207	8.362
1.9	0.041	0.115	0.296	0.454	0.726	0.980	1.244	1.531	1.860	2.261	2.798	3.659	4.487	6.301	8.469
1.9	0.045	0.124	0.314	0.479	0.759	1.020	1.288	1.580	1.915	2.321	2.865	3.735	4.569	6.396	8.575
2	0.049	0.136	0.334	0.505	0.790	1.059	1.333	1.630	1.969	2.381	2.931	3.809	4.651	6.490	8.679
2	0.053	0.151	0.354	0.530	0.823	1.099	1.378	1.680	2.024	2.442	2.997	3.883	4.732	6.582	8.783
2.1	0.057	0.164	0.374	0.556	0.857	1.138	1.422	1.729	2.079	2.501	3.063	3.958	4.813	6.675	8.887
2.1	0.066	0.175	0.395	0.583	0.891	1.178	1.467	1.779	2.133	2.561	3.129	4.032	4.894	6.767	8.989
2.2	0.070	0.186	0.415	0.610	0.925	1.218	1.512	1.829	2.188	2.620	3.195	4.105	4.973	6.858	9.091

Cumulative Probability															
2.2	0.074	0.200	0.437	0.637	0.959	1.258	1.557	1.879	2.242	2.680	3.260	4.179	5.053	6.949	9.193
2.3	0.085	0.212	0.458	0.664	0.994	1.298	1.603	1.928	2.297	2.739	3.325	4.252	5.132	7.039	9.294
2.3	0.090	0.226	0.481	0.691	1.029	1.338	1.648	1.978	2.351	2.799	3.390	4.324	5.211	7.129	9.394
2.4	0.095	0.238	0.502	0.718	1.064	1.379	1.693	2.028	2.405	2.858	3.455	4.396	5.289	7.219	9.493
2.4	0.100	0.253	0.524	0.747	1.099	1.420	1.738	2.078	2.459	2.917	3.519	4.468	5.367	7.308	9.592
2.5	0.113	0.268	0.548	0.775	1.134	1.460	1.784	2.127	2.514	2.976	3.584	4.540	5.445	7.397	9.691
2.5	0.118	0.280	0.575	0.804	1.170	1.500	1.829	2.178	2.568	3.035	3.648	4.612	5.522	7.484	9.789
2.6	0.124	0.296	0.598	0.833	1.205	1.539	1.875	2.227	2.622	3.093	3.712	4.683	5.600	7.572	9.886
2.6	0.130	0.313	0.621	0.862	1.241	1.581	1.920	2.277	2.676	3.152	3.776	4.754	5.677	7.660	9.983
2.7	0.147	0.326	0.646	0.890	1.277	1.622	1.966	2.327	2.730	3.210	3.840	4.825	5.753	7.746	10.079
2.7	0.152	0.343	0.671	0.920	1.314	1.663	2.011	2.376	2.784	3.269	3.903	4.896	5.830	7.833	10.176
2.8	0.158	0.356	0.694	0.950	1.350	1.704	2.058	2.427	2.838	3.328	3.967	4.966	5.906	7.919	10.272
2.8	0.165	0.374	0.719	0.980	1.386	1.746	2.103	2.476	2.892	3.386	4.030	5.040	5.982	8.004	10.367
2.9	0.186	0.392	0.744	1.009	1.423	1.787	2.149	2.526	2.946	3.444	4.093	5.120	6.058	8.090	10.461
2.9	0.192	0.406	0.770	1.040	1.460	1.829	2.195	2.576	2.999	3.502	4.156	5.190	6.133	8.175	10.556
3	0.198	0.424	0.794	1.070	1.497	1.871	2.241	2.626	3.054	3.560	4.220	5.260	6.208	8.260	10.649
3	0.205	0.439	0.819	1.101	1.534	1.913	2.287	2.676	3.108	3.618	4.283	5.329	6.283	8.345	10.743
3.1	0.212	0.458	0.845	1.134	1.571	1.954	2.333	2.726	3.161	3.676	4.346	5.398	6.357	8.429	10.837

Cumulative Probability															
3.1	0.239	0.478	0.872	1.165	1.607	1.996	2.378	2.776	3.215	3.734	4.408	5.468	6.432	8.513	10.930
3.2	0.245	0.492	0.898	1.197	1.645	2.038	2.425	2.825	3.268	3.792	4.471	5.537	6.506	8.596	11.023
3.2	0.251	0.513	0.925	1.227	1.682	2.080	2.471	2.875	3.322	3.850	4.533	5.605	6.580	8.680	11.113
3.3	0.259	0.528	0.950	1.259	1.720	2.123	2.517	2.925	3.376	3.907	4.595	5.675	6.654	8.763	11.205
3.3	0.267	0.548	0.977	1.291	1.758	2.165	2.563	2.975	3.430	3.965	4.658	5.743	6.727	8.845	11.298
3.4	0.300	0.570	1.004	1.323	1.796	2.207	2.610	3.025	3.483	4.022	4.720	5.811	6.801	8.928	11.389
3.4	0.306	0.585	1.031	1.354	1.834	2.250	2.656	3.075	3.537	4.079	4.782	5.879	6.874	9.010	11.480
3.5	0.313	0.607	1.059	1.386	1.872	2.292	2.702	3.125	3.590	4.137	4.843	5.948	6.947	9.093	11.570
3.5	0.320	0.623	1.087	1.418	1.910	2.334	2.748	3.175	3.644	4.194	4.905	6.015	7.020	9.174	11.660
3.6	0.328	0.645	1.115	1.451	1.948	2.377	2.795	3.225	3.697	4.252	4.967	6.084	7.092	9.255	11.749
3.6	0.337	0.661	1.141	1.483	1.985	2.420	2.841	3.274	3.750	4.309	5.028	6.152	7.165	9.337	11.840
3.7	0.377	0.684	1.169	1.516	2.024	2.462	2.887	3.324	3.804	4.366	5.091	6.219	7.237	9.418	11.929
3.7	0.383	0.708	1.197	1.549	2.062	2.505	2.934	3.374	3.858	4.423	5.152	6.286	7.310	9.499	12.017
3.8	0.390	0.723	1.226	1.582	2.101	2.547	2.980	3.425	3.911	4.480	5.214	6.354	7.381	9.579	12.107
3.8	0.398	0.748	1.255	1.613	2.140	2.590	3.027	3.474	3.964	4.537	5.275	6.420	7.454	9.659	12.195
3.9	0.406	0.764	1.284	1.646	2.179	2.633	3.073	3.524	4.018	4.594	5.336	6.488	7.525	9.740	12.284
3.9	0.416	0.788	1.310	1.680	2.218	2.676	3.120	3.574	4.071	4.651	5.397	6.555	7.596	9.820	12.371
4	0.426	0.805	1.339	1.713	2.257	2.719	3.163	3.624	4.124	4.708	5.458	6.622	7.668	9.900	12.459

Cumulative Probability															
4	0.471	0.829	1.369	1.746	2.295	2.762	3.209	3.674	4.177	4.765	5.519	6.689	7.739	9.980	12.546
4.1	0.478	0.847	1.398	1.780	2.334	2.805	3.256	3.724	4.231	4.822	5.580	6.755	7.811	10.059	12.634
4.1	0.485	0.871	1.429	1.814	2.373	2.848	3.302	3.774	4.284	4.879	5.641	6.821	7.882	10.137	12.721
4.2	0.494	0.900	1.455	1.848	2.413	2.891	3.350	3.823	4.337	4.936	5.701	6.888	7.952	10.216	12.807
4.2	0.503	0.914	1.485	1.882	2.451	2.935	3.396	3.874	4.390	4.992	5.762	6.954	8.023	10.295	12.894
4.3	0.513	0.942	1.515	1.916	2.491	2.978	3.443	3.924	4.444	5.049	5.823	7.020	8.093	10.374	12.981
4.3	0.524	0.958	1.545	1.950	2.531	3.021	3.489	3.974	4.497	5.105	5.883	7.086	8.170	10.453	13.066
4.4	0.578	0.986	1.576	1.985	2.572	3.065	3.537	4.024	4.550	5.162	5.944	7.153	8.264	10.531	13.152
4.4	0.584	1.002	1.603	2.017	2.612	3.108	3.584	4.074	4.603	5.218	6.005	7.219	8.334	10.609	13.238
4.5	0.592	1.029	1.634	2.051	2.653	3.152	3.630	4.123	4.656	5.274	6.065	7.284	8.405	10.687	13.324
4.5	0.600	1.046	1.665	2.085	2.691	3.195	3.677	4.173	4.709	5.331	6.126	7.350	8.475	10.765	13.410
4.6	0.610	1.074	1.696	2.120	2.731	3.239	3.724	4.223	4.762	5.387	6.186	7.415	8.544	10.843	13.495
4.6	0.620	1.092	1.727	2.155	2.771	3.283	3.771	4.273	4.815	5.443	6.246	7.480	8.615	10.920	13.578
4.7	0.632	1.119	1.755	2.190	2.812	3.326	3.817	4.323	4.868	5.501	6.306	7.546	8.684	10.998	13.663
4.7	0.698	1.138	1.786	2.225	2.852	3.369	3.864	4.373	4.921	5.557	6.366	7.611	8.754	11.075	13.748
4.8	0.703	1.165	1.817	2.260	2.890	3.412	3.911	4.423	4.974	5.613	6.426	7.676	8.823	11.152	13.832
4.8	0.710	1.184	1.849	2.295	2.930	3.456	3.958	4.474	5.027	5.669	6.486	7.742	8.892	11.229	13.916
4.9	0.717	1.211	1.881	2.330	2.970	3.500	4.005	4.524	5.081	5.725	6.546	7.807	8.962	11.306	14.000

Cumulative Probability															
4.9	0.726	1.247	1.909	2.366	3.011	3.544	4.052	4.573	5.134	5.781	6.606	7.872	9.031	11.382	14.084
5	0.737	1.258	1.940	2.398	3.051	3.588	4.099	4.623	5.186	5.837	6.665	7.937	9.100	11.457	14.168
5	0.748	1.293	1.972	2.434	3.091	3.632	4.146	4.673	5.239	5.893	6.725	8.002	9.169	11.534	14.251

Tabel B.3. Kuantil kanan dari distribusi Chi-kuadrat. Untuk v besar, distribusi chi-kuadrat kira-kira Gaussian, dengan rata-rata v dan varians 2v.

Cumulative Probability						
V	0.5	0.9	0.95	0.99	0.999	0.9999
1	0.455	2.706	3.841	6.635	10.828	15.137
2	1.386	4.605	5.991	9.21	13.816	18.421
3	2.366	6.251	7.815	11.345	16.266	21.108
4	3.357	7.779	9.488	13.277	18.467	23.512
5	4.351	9.236	11.07	15.086	20.515	25.745
6	5.348	10.645	12.592	16.812	22.458	27.855
7	6.346	12.017	14.067	18.475	24.322	29.878
8	7.344	13.362	15.507	20.09	26.124	31.827
9	8.343	14.684	16.919	21.666	27.877	33.719
10	9.342	15.987	18.307	23.209	29.588	35.563

Cumulative Probability						
11	10.341	17.275	19.675	24.725	31.264	37.366
12	11.34	18.549	21.026	26.217	32.91	39.134
13	12.34	19.812	22.362	27.688	34.528	40.871
14	13.339	21.064	23.685	29.141	36.123	42.578
15	14.339	22.307	24.996	30.578	37.697	44.262
16	15.338	23.542	26.296	32	39.252	45.925
17	16.338	24.769	27.587	33.409	40.79	47.566
18	17.338	25.989	28.869	34.805	42.312	49.19
19	18.338	27.204	30.144	36.191	43.82	50.794
20	19.337	28.412	31.41	37.566	45.315	52.385
21	20.337	29.615	32.671	38.932	46.797	53.961
22	21.337	30.813	33.924	40.289	48.268	55.523
23	22.337	32.007	35.172	41.638	49.728	57.074
24	23.337	33.196	36.415	42.98	51.179	58.613
25	24.337	34.382	37.652	44.314	52.62	60.14
26	25.336	35.563	38.885	45.642	54.052	61.656
27	26.336	36.741	40.113	46.963	55.476	63.164
28	27.336	37.916	41.337	48.278	56.892	64.661

Cumulative Probability						
29	28.336	39.087	42.557	49.588	58.301	66.152
30	29.336	40.256	43.773	50.892	59.703	67.632
31	30.336	41.422	44.985	52.191	61.098	69.104
32	31.336	42.585	46.194	53.486	62.487	70.57
33	32.336	43.745	47.4	54.776	63.87	72.03
34	33.336	44.903	48.602	56.061	65.247	73.481
35	34.336	46.059	49.802	57.342	66.619	74.926
36	35.336	47.212	50.998	58.619	67.985	76.365
37	36.336	48.363	52.192	59.892	69.347	77.798
38	37.335	49.513	53.384	61.162	70.703	79.224
39	38.335	50.66	54.572	62.428	72.055	80.645
40	39.335	51.805	55.758	63.691	73.402	82.061
41	40.335	52.949	56.942	64.95	74.745	83.474
42	41.335	54.09	58.124	66.206	76.084	84.88
43	42.335	55.23	59.304	67.459	77.419	86.28
44	43.335	56.369	60.481	68.71	78.75	87.678
45	44.335	57.505	61.656	69.957	80.077	89.07
46	45.335	58.641	62.83	71.201	81.4	90.456

Cumulative Probability						
47	46.335	59.774	64.001	72.443	82.721	91.842
48	47.335	60.907	65.171	73.683	84.037	93.221
49	48.335	62.038	66.339	74.919	85.351	94.597
50	49.335	63.167	67.505	76.154	86.661	95.968
55	54.335	68.796	73.311	82.292	93.168	102.776
60	59.335	74.397	79.082	88.379	99.607	109.501
65	64.335	79.973	84.821	94.422	105.988	116.16
70	69.334	85.527	90.531	100.425	112.317	122.754
75	74.334	91.061	96.217	106.393	118.599	129.294
80	79.334	96.578	101.879	112.329	124.839	135.783
85	84.334	102.079	107.522	118.236	131.041	142.226
90	89.334	107.565	113.145	124.116	137.208	148.626
95	94.334	113.038	118.752	129.973	143.344	154.989
100	99.334	118.498	124.342	135.807	149.449	161.318

LATAR BELAKANG PENULIS

Nurtiti Sunusi, Lahir di Siwa Kabupaten Wajo Provinsi Sulawesi Selatan. Sekolah Dasar dan Sekolah Menengah Pertama diselesaikan di Siwa. Tahun 1987 melanjutkan sekolah ke SMA Negeri 2 Makassar. Tahun 1996 menyelesaikan Studi Sarjana dari Jurusan Matematika FMIPA Unhas dengan gelar S.Si. Selama kuliah di Universitas Hasanuddin, penulis menjalani Ikatan Dinas dari pemerintah dan mendapatkan beasiswa dari pemeritah Canada melalui project *Eastern Indonesia University Development Project* (EIUDP) yang merupakan project persiapan untuk pembukaan Fakultas MIPA di Indonesia Bagian Timur. Tahun 1997 penulis mendapat SK CPNS sebagai dosen yang ditempatkan di Fakultas MIPA Universitas Haluoleo. Tahun 1998 penulis mendapatkan beasiswa BPPS untuk melanjutkan studi ke jenjang Magister di Institut Teknologi Bandung (ITB) Jurusan Matematika dan berhasil menyelesaikan studi pada tahun 2010 dengan gelar M.Si. Tahun 2015 penulis melanjutkan Pendidikan Doktor di ITB dengan beasiswa BPPS dari pemerintah Indonesia. Selama di ITB penulis bergabung dengan Kelompok Keahlian Statistika sesuai dengan minat yang akan ditekuni. Setelah menyelesaikan Pendidikan Doctor pada Januari 2010, penulis kembali mengabdi di Universitas Haluoleo Kendari. Pada Tahun 2012, penulis pindah ke Universitas Hasanuddin Jurusan Matematika mengikuti suami yang bekerja sebagai dosen di Unhas. Tahun 2019 – 2023 dimanahi tugas sebagai Ketua Departemen Statistika FMIPA Unhas. Berbagai karya ilmiah telah dihasilkan oleh penulis yang telah dipublikasikan, baik pada jurnal Nasional terakreditasi maupun Jurnal Internasional bereputasi. Karya-karya ilmiah tersebut telah diseminarkan baik di tingkat Nasional maupun di Tingkat Internasional. Penulis tergabung dalam organisasi Profesi yaitu ISI (*International Statistical Intitute*), Forstat (Forum Pendidikan Tinggi Statistika), dan Indonesian Mathematics Society (IndoMS). Saat ini penulis diamanahi sebagai Wakil Gubernur bidang Akademik dan Pengajaran di IndoMS wilayah Sulawesi Gorontalo.