STATISTIKA TERAPAN
dengan Menggunakan R

Siana Halim

STATISTIKA TERAPAN
dengan Menggunakan R

Siana Halim

Penerbit:
Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM)
Universitas Kristen Petra

Statistika Terapan dengan Menggunakan R
Siana Halim
Surabaya, Bagian Penerbitan Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM), Universitas Kristen Petra, 2025
e-ISBN: 978-623-5457-19-2 (PDF)

Kutipan Pasal 44
1. Barang siapa dengan sengaja dan tanpa hak mengumumkan atau memperbanyak suatu ciptaan atau memberi ijin untuk itu, dipidana paling lama 7 (tujuh) tahun dan/atau denda paling banyak Rp 100.000.000,- (seratus juta rupiah)
2. Barang siapa dengan sengaja menyiarkan, memamerkan, mengedarkan, atau menjual kepada umum dalam ayat (1) dipidana dengan pidana penjara paling lama 5 (lima) tahun dan/atau denda paling banyak Rp 50.000.000,- (lima puluh juta rupiah)

Statistika Terapan dengan Menggunakan R
Cetakan Pertama, Januari 2025

Penulis:
Siana Halim

@Hak cipta ada pada penulis
Hak penerbit pada penerbit

Tidak boleh diproduksi sebagian atau seluruhnya dalam bentuk apapun tanpa seijin tertulis dari pengarang dan/atau penerbit

Penerbit:

Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM)
Universitas Kristen Petra
Jl. Siwalankerto 121-131, Surabaya 60236, Indonesia
Telp. 031-2983147

Statistika Terapan dengan Menggunakan R

Siana Halim

Table of contents

Kata Pengantar

Buku ini dituliskan bagi para mahasiswa dan umum yang ingin mempelajari Statistika terapan dengan menggunakan software R sebagai piranti untuk membantu menyelesaikan perhitungan statistik. Uraian-uraian diberikan secara ringkas dengan menambahkan R-script sebagai bagian dari materi.

Referensi pelengkap disertakan pada akhir Bab. Diharapkan para pembaca dapat memperlengkapinya dengan membaca uraian detail pada referensi-referensi yang dianjurkan. Pembaca diharapkan pula untuk mencoba R-script yang telah dituliskan, serta mengembangkan kemampuan untuk dapat menuliskan R-script secara mandiri untuk menyelesaikan masalah industri ataupun keseharian yang terkait dengan statistika.

Satu quotation dari Florence Nightingale mengatakan:

To understand God's thoughts we must study statistics, for these are the measure of His purpose.
– Florence Nightingale

Selamat belajar Statistik... kiranya buku ini membantu pembaca untuk mencintainya dan menemukan "*the measure of God purpose*".

Surabaya, Januari 2025

Siana Halim

Buku dapat diakses secera online pada link berikut:

https://sianahalim.quarto.pub/statistika-terapan-dengan-menggunakan-r/

Menggunakan R

R adalah *open source* yang dapat diunduh di http://www.r-project.org. R terdiri dari program utama (base) yang harus diunduh bila R ini di *install* untuk pertama kalinya (https://cran.project.org/bin/windows/base/).Selain itu base, R memiliki ribuan packages yang merupakan *statistics library* (https://cran.r-project.org/web/packages/available_packages_by_name.html). Packages yang tersedia di R merupakan statistical tools yang telah teruji dan selalu *update*.

Alternative lain untuk menampilkan R adalah dengan menggunakan R-studio, yang dapat diunduh di https://www.rstudio.com/, dan R dapat pula ditampilkan secara *interactive* dengan menggunakan shiny (http://shiny.rstudio.com/).

R menghasilkan graphics dengan kualitas bagus dan dapat disimpan dalam berbagai format misalnya, jpeg, postscript, eps, pdf dan bmp.

R sangat berguna untuk menangani *project* berskala kecil maupun besar.

Memulai R

Cara termudah untuk mempelajari R adalah dengan menirukan perintah-perintah R (R *command*) yang dapat ditemui dengan mudah di internet. Untuk mendapatkan bantuan (help) tentang penggunaan R *command* ada tiga sumber bantuan yaitu: file help, R-help archive, dan R-help. File R dapat diakses dengan menggunakan perintah help(), sebagai contoh:

```
? mean              # Berhasil! mean merupakan R command
```

```
starting httpd help server ... done
```

```
help(mean)          # Berhasil! mean merupakan R command
```

Namun demikian, kita tidak akan mendapatkan informasi yang kita butuhkan, bila perintah tersebut tidak tercantum pada sistem help.

Kita dapat memberikan catatan pada R-script dengan memberikan tanda # di awal kalimat.

```
help(regression)        # Gagal! regression bukan merupakan R command
```

```
No documentation for 'regression' in specified packages and libraries:
you could try '??regression'
```

Perintah ini gagal, karena "regression" bukanlah R command. Bila kita ingin mencari command yang harus digunakan untuk menyelesaiakan masalah regresi, maka kita dapat menggunakan perintah.

```
help.search("regression")
```

Perintah ini akan memberitahu memberi kita petunjuk perintah-perintah di R yang dapat digunakan untuk menyelesaikan masalah regresi.

Installing Packages

Seringkali kita membutuhkan R-packages untuk dapat menjalankan perintah-perintah statistics yang terdapat di R. Perintah yang digunakan adalah install.packages("nama packages"), misalkan:

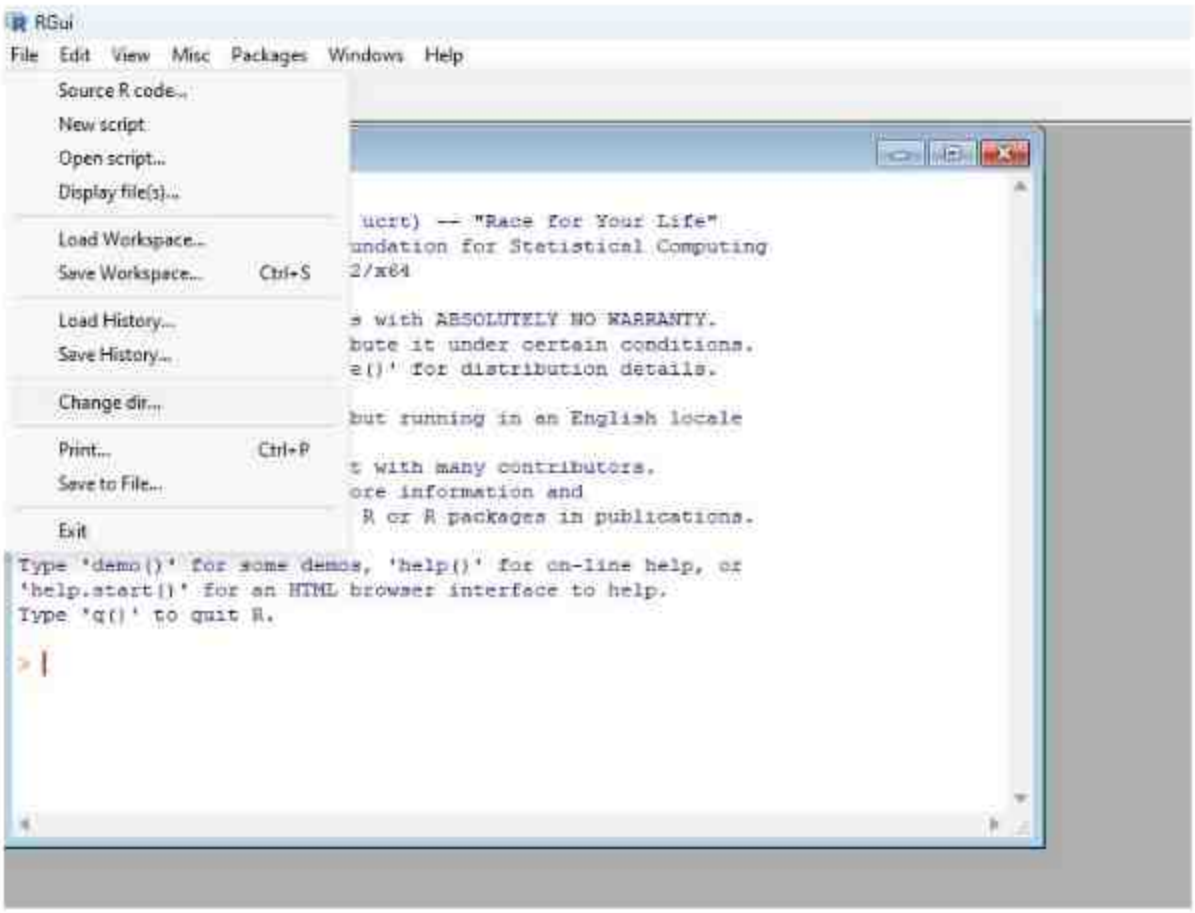
```
#Catatan delete tanda # ketika anda menjalankan intall.packages di bawah ini
#install.packages("car")
```

Menyiapkan Direktori Kerja

Untuk memulai suatu project dengan menggunakan R, sebaiknya kita mempersiapkan direktori yang akan digunakan untuk menyimpan data, hasil ataupun R-command yang telah disimpan dalam bentuk script.

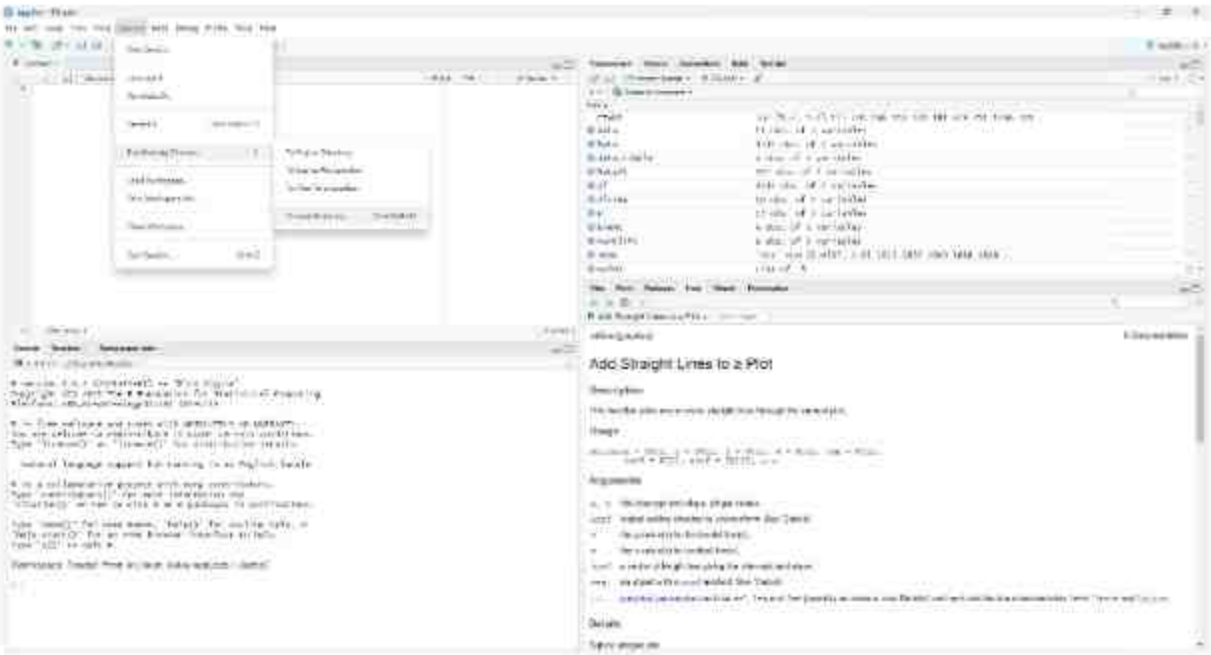
Agar R bekerja di direktori yang telah disiapkan, maka kita harus melakukan *change directory* dengan cara:

KLIK Console-> KLIK File -> KLIK Change Directory tentukan di directory mana anda akan bekerja.



Jika anda menggunakan R-Studio maka untuk mengubah direktori dapat dilakukan dengan cara:

KLIK Session -> Setting Working Direktory -> Choose Directory (ctrl+shift+H)



Selain itu kita dapat menyiapkan direktori dengan cara sebagai berikut:

Pada console ketik

```
#untuk mengetahui direktori yang digunakan saat ini
DIR = getwd()

#create direktori
DIR1 = paste0(DIR,"/testdir")
dir.create(file.path(DIR1), recursive = TRUE)
```

```
Warning in dir.create(file.path(DIR1), recursive = TRUE): 'D:\Buat
Buku\AppStat\testdir' already exists
```

```
#Mengubah setting direktori
setwd(DIR1)

#Mengembalikan setting direktori ke direktori awal
setwd(DIR)
```

Work Space

R menggunakan konsep *work space* agar *objects* yang digunakan dalam R *command* terorganisasi dengan baik. Seluruh *objects* yang dibuat baik secara langsung maupun tidak langsung dapat disimpan dalam sebuah *workspace*. *Workspace* yang digunakan dapat di save, load, share ataupun archive. Isi dari workspace dapat dilihat dengan menggunakan perintah ls(), selain itu seluruh objects yang tersimpan dapat dihapus dengan menggunakan perintah rm().

Contoh:

```
test <- 1:10    # Membuat sebuat object
ls()              # List seluruh objects yang ada di workspace
```

```
[1] "DIR"  "DIR1" "test"
```

```
# Menyimpan seluruh objects dalam sebuah binary file
save.image(file="labstat.RData")

# Menyimpan objects  bernama test
save(test, file="labstat.RData")

rm(test)            # Menghapus object dengan nama "test"
ls()              # List seluruh objects yang ada
```

```
[1] "DIR"  "DIR1"
```

```
rm(list=ls())     # Menghapus seluruh object yang ada
load(file="labstat.RData")  # Load object
ls()                        # List seluruh objects yang ada
```

```
[1] "test"
```

Menulis Script

R command dapat dituliskan dalam sebuah notepad ataupun text editor yang tersedia pada R, ataupun R-Studio. Untuk menampilkan script:

KLIK File -> KLIK New Script.

Script ini biasanya disimpan dengan extension .R, misalnya file.R. Perintah-perintah yang telah tersimpan pada script ini dapat dijalankan pada R console dengan cara copy-paste perintah tersebut dari script ke console. Bila seluruh script akan dijalankan pada R, maka kita dapat menggunakan perintah:

source(file="C://path/to/filename/file.R", echo=T)

source(file="../directory/file.R", echo=T)

source(file="file.R", echo=T) # Jika file.R berada pada wd yang sedang digunakan.

Membuat Objects

Object dapat dibuat dan diberi nilai dengan menggunakan tanda "<-" atau "=", misal:

```
# Membuat object "a" dan memberi nilai 1
a <- 1
# Menggantikan nilai 1 dengan 1.5
a <- 1.5
# Menggantikan nilai 1.5 dengan karakter "labstat"
a <- "labstat"
# Membuat object "b" dan memberi nilai sama dengan a
b <- a

# Mengganti "labstat" dengan sebuah vektor dengan nilai 1,2 dan 3
a <- c(1,2,3)

#  c adalah sebuah R command untuk menyatakan vector
#  Jangan menggunakan c sebagai nama sebuah object

a <- 1:3    # Untuk membuat deret 1,2,3 dapat dituliskan dengan 1:3
b <-mean(a) # Assign mean dari object a ke dalam object b
```

Tipe Object

Ada setidaknya enam tipe object yang dikenali oleh R.

Konstan

Tipe data konstan ini dapat didefinisikan secara langsung, misalnya

```
a <- 1
a <- 1.5
a <- "R Datacamp"
b <- a
```

Array atau vektor

Array atau vektor dapat dinyatakan dengan menggunakan *'reserved word'* , c(), dan mendaftarkan anggota vektor di antara tanda (), misalnya

```
a <- c(1,2,3)
b <- c(4,6,2,8,1)
```

Pada contoh di atas *a* adalah vektor dengan anggota 1,2,3 dan *b* adalah vektor dengan anggota 4,6,2,8,1. Cara lain untuk membuat vektor yang berurutan dengan menggunakan tanda : di antara nilai awal dan nilai akhir yang ditentukan. Misalnya

```
a <- 1:3
b <- 1:10
```

Pada contoh di atas *a* adalah vektor dengan anggota 1,2,3 dan *b* adalah vektor dengan anggota 1 sampai dengan 10. Bila vektor dapat pula dinyatakan sebagai deret (*sequence*), dari nilai awal dan nilai akhir dengan nilai beda yang sebarang.

```
a <- seq(1,50,2)
b <- seq(100,1,-2)
```

Pada contoh di atas a adalah deret atau vektor dengan anggota 1 hingga 50 dengan beda 2 dan *b* adalah deret dengan anggota 100 hingga 1 dengan beda -2. Untuk mengetahui panjang sebuah vektor kita dapat menggunakan perintah **length**, misalnya

```
L <- length(a) # panjang array
```

Untuk mengakses nilai dari sebuah vektor pada indeks tertentu digunakan perintah NamaVektor[indeks]. Misalnya

```
a <- seq(1,50,2)
a[1]
```

```
[1] 1
```

```
a[1:5]
```

```
[1] 1 3 5 7 9
```

```
a[c(1,3,5)]
```

```
[1] 1 5 9
```

Untuk mendapatkan sintak lengkap dari sebuah fungsi atau perintah di R kita dapat menggunakan help dengan mengetik ?NamaFungsi misalnya

```
?seq
```

Matrix

Bila kita ingin menyatakan data dalam dua dimensi, maka kita dapat menggunakan Matrix untuk menyimpan data tersebut. Matrix dapat dinyatakan dengan syntax sebagai berikut

A <- matrix(vektor, n,m, byrow = FALSE)

Pada perintah di atas kita akan mendifinisikan A sebagai matrix, vektor adalah nilai yang akan kita simpan, nilai tersebut dinyatakan dalam vektor, n adalah jumlah baris, m adalah jumlah kolom dan nilai tersebut akan diatur secara kolom terlebih dahulu. Contoh

```
Mat1 <- matrix(1:100,10,10)
```

Pada contoh ini vektor yang beranggotakan angka 1 hingga 100 akan disimpan dalam matriks berukuran 10 x 10 secara kolom (Mat1) dan secara baris (Mat2)

```
Mat2 <- matrix(1:100,10,10, byrow = TRUE)
```

Untuk dapat mengakses nilai dari sebuah matriks, kita menggunakan perintah

NamaVar[Posisi Baris, Posisi kolom]

Pada contoh ini Mat1[3,5] akan memberikan nilai yang tersimpan pada variable Mat1 di baris ketiga dan kolom kelima, yaitu 43. Kita juga dapat menyatakan posisi baris dan posisi kolom yang tersimpan pada sebuah matrix sebagai vektor.

```
Mat1[3,5]
```

```
[1] 43
```

Pada contoh ini Mat1[1:4, 2:5] akan menghasilkan submatrix yang tersimpan di baris 1 hingga 4, dan kolom 2 hingga 5.

```
Mat1[1:4,2:5]
```

```
     [,1] [,2] [,3] [,4]
[1,]   11   21   31   41
[2,]   12   22   32   42
[3,]   13   23   33   43
[4,]   14   24   34   44
```

Dataframe

Dataframe adalah struktur data dalam bentuk vektor dua dimensi. Dataframe merepresentasikan multidimensional data set secara optimal. Pada dataframe, setiap kolom merepresentasikan sebuah variabel dan setiap baris merepresentasikan sebuah observasi.

Sebuah dataframe dapat menyimpan vektor dalam berbagai basic class. Bisa saja dalam sebuah dataframe, kolom pertama berupa string, sedangkan kolom kedua berupa faktor dan kolom ketiga berupa numerik. Berikut adalah cara untuk menampilkan object dalam sebuah dataframe.

```
Alphabet <- c("a","b","c","d","e")
Numerik  <- 1:5
Konstan  <- 1
Persamaan <- Numerik^2 + Konstan
#mendefinisikan dataframe
Data <- data.frame(Alphabet,Numerik,Persamaan)
```

```
is.data.frame(Data)
```

```
[1] TRUE
```

Dataframe dapat diakses dengan menggunakan

NamaDataFrame$NamaVariabel[posisi baris].

Misalkan

```
Alpha <- Data$Alphabet
Alpha
```

```
[1] "a" "b" "c" "d" "e"
```

Kelemahan dari dataframe adalah seringkali dataframe ini akan mengubah data numerik menjadi data text. Bila hal itu terjadi, maka kita dapat mengkonversikan kembali data tersebut ke numerik dengan menggunakan perintah as.numeric.

```
Num <- Data$Numerik
Num <- as.numeric(Num)
Num
```

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

Tibble

Tibble juga merupakan dataframe. Namun tibble memperbaiki kekurangan yang ada pada dataframe. Tibble tidak akan mengubah tipe data. Data bertipe numerik akan tetap ditampilkan sebagai numerik, walaupun dataframe tersebut merupakan memiliki data yang bertipe string dan numerik. Tibble juga hanya menampilkan 10 baris teratas saja beserta dengan tipe datanya. Untuk dapat menggunakan tipe data tibble, maka kita harus install packages tidyverse terlebih dahulu. dan selanjutkan mengaksesnya dengan menggunakan library(tidyverse). Berikut adalah contoh mendefinisikan tipe data tibble pada R

```r
#Diberi tanda # karena packages ini telah terinstall di PC, sehingga tidak perlu diinstall u
#install.packages("tibble")
#install.packages("tidyverse")
#install.packages("conflicted")

#Peringatan yang muncul pada output tidak akan ditampilkan
options(warn = -1)

#Bila terjadi conflict antara dua package pada R, maka package "dplyr"
#akan digunakan sebagai pendukung library tidyverse daripada package "filter"
#ataupun "lag"

library(conflicted)
suppressMessages(conflict_prefer("filter", "dplyr"))
suppressMessages(conflict_prefer("lag", "dplyr"))

#perintah suppressMessanges agar message yang terdapat pada library tidyverse tidak muncul
#saat library(tidyverse) diaktifkan.
options(warn = -1)
suppressMessages(library(tidyverse))

data_tibble = tibble(x = 1:5, y = 1, z = x^2+y)
data_tibble
```

```
# A tibble: 5 x 3
      x     y     z
  <int> <dbl> <dbl>
1     1     1     2
2     2     1     5
3     3     1    10
4     4     1    17
5     5     1    26
```

```r
data_tibble = tibble(Alpha,data_tibble)
data_tibble
```

```
# A tibble: 5 x 4
  Alpha     x     y     z
  <chr> <int> <dbl> <dbl>
1 a         1     1     2
2 b         2     1     5
```

```
3 c          3     1     10
4 d          4     1     17
5 e          5     1     26
```

Tibble adalah dataframe, cara mengakses tibble sama seperti cara mengakses dataframe. Tibble dapat diakses berdasarkan nama variabel

```
data_tibble$z
```

```
[1]  2  5 10 17 26
```

```
data_tibble[["z"]]
```

```
[1]  2  5 10 17 26
```

atau berdasarkan indeks

```
data_tibble[4]
```

```
# A tibble: 5 x 1
      z
  <dbl>
1     2
2     5
3    10
4    17
5    26
```

```
data_tibble$z[1]
```

```
[1] 2
```

Pipelines

Kita dapat menggunakan pipes untuk menuliskan coding secara ringkas dan mudah dibaca dan dimengerti. Pada tibble pipes dituliskan dengan mengguanakn %>%. Contoh

Kita dapat menuliskan summary data_tibble di atas dengan cara

```
summary(data_tibble)
```

```
   Alpha                x           y           z
Length:5           Min.   :1   Min.   :1   Min.   : 2
Class :character   1st Qu.:2   1st Qu.:1   1st Qu.: 5
Mode  :character   Median :3   Median :1   Median :10
                   Mean   :3   Mean   :1   Mean   :12
                   3rd Qu.:4   3rd Qu.:1   3rd Qu.:17
                   Max.   :5   Max.   :1   Max.   :26
```

```
summary(data_tibble[-1])
```

```
      x           y           z
Min.   :1   Min.   :1   Min.   : 2
1st Qu.:2   1st Qu.:1   1st Qu.: 5
Median :3   Median :1   Median :10
Mean   :3   Mean   :1   Mean   :12
3rd Qu.:4   3rd Qu.:1   3rd Qu.:17
Max.   :5   Max.   :1   Max.   :26
```

Kita dapat menuliskan perintah tersebut dalam satu kesatuan dengan menggunakan pipes sebagai berikut

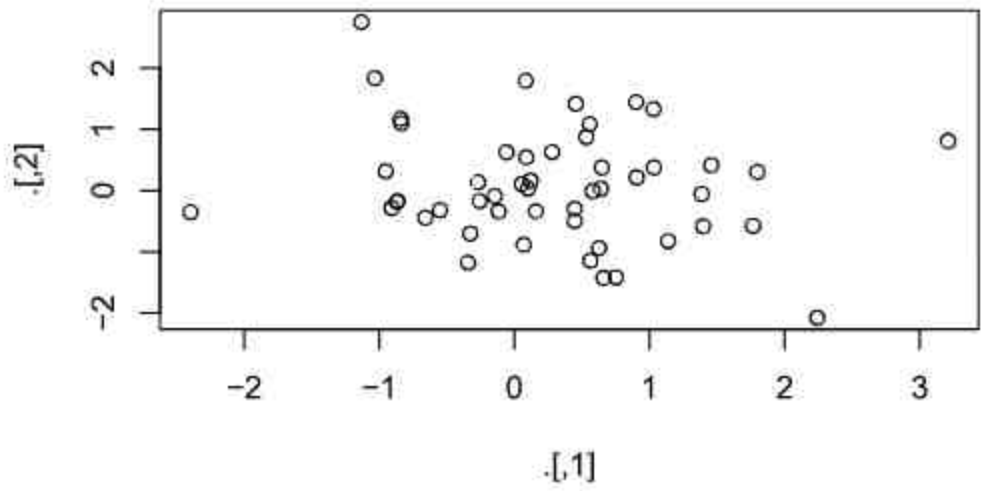
```
data_tibble %>%
    summary()
```

```
   Alpha                x           y           z
Length:5           Min.   :1   Min.   :1   Min.   : 2
Class :character   1st Qu.:2   1st Qu.:1   1st Qu.: 5
Mode  :character   Median :3   Median :1   Median :10
                   Mean   :3   Mean   :1   Mean   :12
                   3rd Qu.:4   3rd Qu.:1   3rd Qu.:17
                   Max.   :5   Max.   :1   Max.   :26
```

```
    #summary(data_tibble[-1])
```

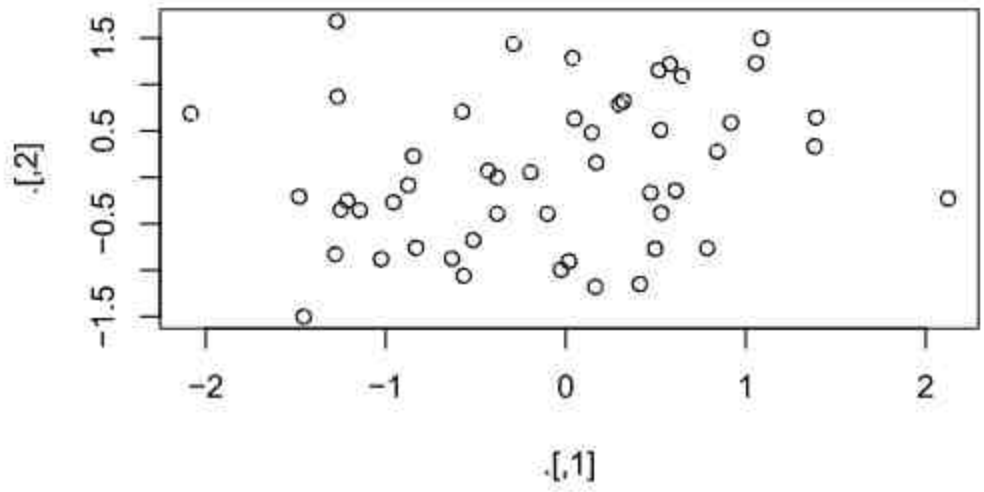
Terkadang %>% bermasalah, yaitu tidak menghasilkan apapun (NULL), dan menghentikan pipe. Untuk mengatasi hal ini gunakan tee pipes tee pipes %T>%

```
rnorm(100) %>%
    matrix(ncol = 2) %>%
    plot() %>%
     str()
```




```
NULL
```

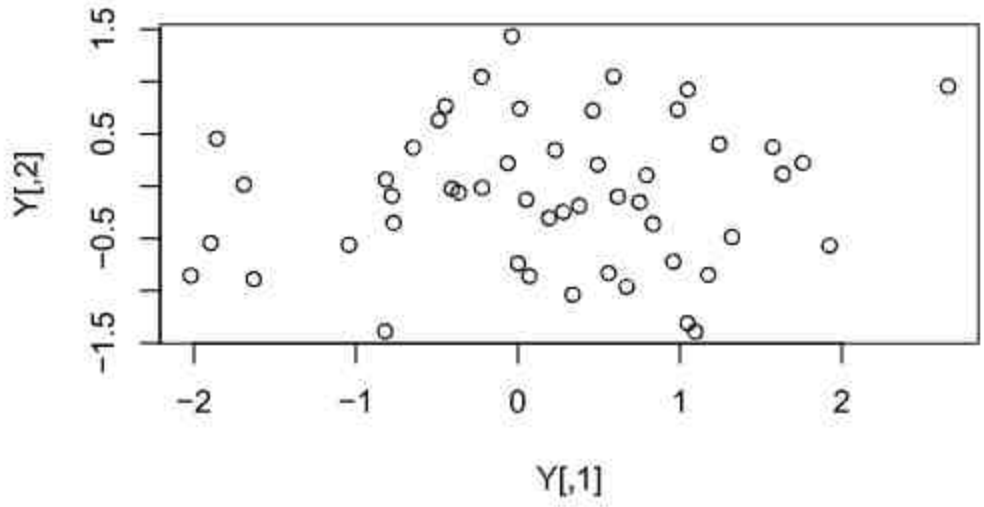
```
rnorm(100) %>%
    matrix(ncol = 2) %T>%
     plot() %>%
    str()
```




```
 num [1:50, 1:2] 0.471 0.516 -0.957 1.381 -1.146 ...
```

Pada penulisan coding tanpa pipes, perintah di atas dapat dituliskan sebagai berikut

```
X = rnorm(100)
Y = matrix(X, ncol=2)
plot(Y)
```




```
str(Y)
```

```
 num [1:50, 1:2] 0.461 -0.449 0.833 -0.364 0.985 ...
```

Perintah ini akan menghasilkan random plot dari 50 titik yang Y1 dan Y2 yang di generate dari 100 titik X dengan distribusi standard normal.

List

List adalah R objects yang terdiri dari berbagai tipe tipe elemen yang berbeda, seperti, numerik, string, vektor, matrix, dataframe atau list lainnya. Berikut adalah contoh dari membuat list dengan menggunakan R

```
list_data = list(Matrix = Mat1, Dataset = Data)
list_data
```

```
$Matrix
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1   11   21   31   41   51   61   71   81    91
[2,]    2   12   22   32   42   52   62   72   82    92
[3,]    3   13   23   33   43   53   63   73   83    93
```

```
 [4,]    4   14   24   34   44   54   64   74   84    94
 [5,]    5   15   25   35   45   55   65   75   85    95
 [6,]    6   16   26   36   46   56   66   76   86    96
 [7,]    7   17   27   37   47   57   67   77   87    97
 [8,]    8   18   28   38   48   58   68   78   88    98
 [9,]    9   19   29   39   49   59   69   79   89    99
[10,]   10   20   30   40   50   60   70   80   90   100

$Dataset
  Alphabet Numerik Persamaan
1        a       1         2
2        b       2         5
3        c       3        10
4        d       4        17
5        e       5        26
```

Untuk mengakses list kita dapat menggunakan dua cara sebagai berikut

```
list_data$Matrix
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    1   11   21   31   41   51   61   71   81    91
 [2,]    2   12   22   32   42   52   62   72   82    92
 [3,]    3   13   23   33   43   53   63   73   83    93
 [4,]    4   14   24   34   44   54   64   74   84    94
 [5,]    5   15   25   35   45   55   65   75   85    95
 [6,]    6   16   26   36   46   56   66   76   86    96
 [7,]    7   17   27   37   47   57   67   77   87    97
 [8,]    8   18   28   38   48   58   68   78   88    98
 [9,]    9   19   29   39   49   59   69   79   89    99
[10,]   10   20   30   40   50   60   70   80   90   100
```

```
list_data[[1]]
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    1   11   21   31   41   51   61   71   81    91
 [2,]    2   12   22   32   42   52   62   72   82    92
 [3,]    3   13   23   33   43   53   63   73   83    93
 [4,]    4   14   24   34   44   54   64   74   84    94
 [5,]    5   15   25   35   45   55   65   75   85    95
 [6,]    6   16   26   36   46   56   66   76   86    96
```

```
 [7,]    7   17   27   37   47   57   67   77   87    97
 [8,]    8   18   28   38   48   58   68   78   88    98
 [9,]    9   19   29   39   49   59   69   79   89    99
[10,]   10   20   30   40   50   60   70   80   90   100
```

```
A = list_data$Matrix
```

List dapat diuraikan lagi menjadi vektor dengan melakukan unlist pada list data.

```
B = unlist(list_data)
```

Exporting dan Importing Data

Import Data

Ada beberapa hal yang harus diperhatikan untuk import data dalam R. Pertama, kita harus tahu dimana data tersebut berada dan akan disimpan dimana, serta adakah karakter khusus yang tersimpan dalam data tersebut. Bila data yang kita miliki tersimpan dalam bentuk spreadsheet, Excel misalnya, hal-hal berikut perlu diperhatikan:

1. Apakah baris dan kolomnya konsisten? Seringkali terjadi masalah ketika kita import data dari Excel spreadsheets.
2. Apakah kolom atau barisnya memiliki label? Apakah label tersebut mudah diolah? (i.e. hindari penamaan yang menggunakan spasi, tanda persen, underscore, dsb)
3. Apakah ada data yang hilang? Apakah data hilang ini tercantum dalam data? Jika iya, simbol apa yang digunakan untuk merepresentasikan data hilang ini?
4. Adakah simbol yang sulit untuk diinterpretasikan?

Untuk membaca file dalam bentuk csv, gunakan command read.csv, contoh:

Bukalah file DataLatihan1.xlsx lalu save as CSV(Comma delimited)(*.csv) menjadi

DataLatihan1.csv

1 Diskripsi Data

1.1 Data

Data adalah catatan yang kita kumpulkan dari berbagai kejadian yang kita amati. Data dapat dikumpulkan melalui survey, eksperimen ataupun pengamatan (observasi). Data dapat kita jumpai dimanapun. Di pasar ataupun toko tradisional, setiap pemilik toko akan ingat akan semua pelangannya. Mereka mengingat kebiasaan ataupun barang-barang yang biasanya para pelanggan itu beli, bahkan bila salah seorang dari pelanggan tersebut lama tidak muncul di toko mereka, mereka akan menanyakan kabar dan kesehatan pelanggan tersebut. Tak jarang pemilik toko ini tahu akan keluarga dari para pelanggan mereka. Catatan transaksi dan kebiasaan dari masing-masing pelanggan pun tak luput dari ingatan mereka. Pemilik toko tradisional ini mengingat ketersediaan barang yang ada di tokonya, mereka hafal harga dari setiap barang yang dijualnya. Data persediaan barang yang ada di toko terekam dengan jelas dalam benak mereka. Tentu saja hal ini bisa mereka lakukan, karena pada umumnya toko kelontong adalah toko kecil dengan jumlah pelanggan yang terbatas.

Lain halnya dengan "toko modern" ataupun "toko online", pada kedua macam toko ini semua data transaksi tercatat secara automatis. Tentu saja "pemilik" toko ini tidak bersinggungan secara langsung dengan para pelanggannya. Terkadang, pelanggan hanyalan "nomor identitas", namun melalui semua data tercatat yang mereka miliki, pemilik toko ini "mengenal" kebiasan dari para pelanggan mereka. Data pelanggan, persedian barang, siapa saja penjual yang ada pada platform toko online tercatat secara lengkap dalam sebuah sistem yang kita sebut sebagai *data base*. Data inilah yang akan diolah untuk menjadi informasi berguna bagi pemilik toko untuk mengembangkan bisnisnya.

Bila kita disodori data seperti terlihat pada Tabel 1.1, maka data di atas tidak akan memiliki arti apapun bagi kita. Ia hanyalah sekedar angka, simbol ataupun kumpulan alfabet tanpa makna. Kita bahkan tidak dapat menduga, apa yang tercatat pada kumpulan data ini. Data, apapun nilainya, bila tidak disertai dengan konteks tidaklah berguna.

Tabel 1.1 Data tercatat

10129	W	12-12-2020	AQ1134	Sambarose	Putih	W	38,67	1,8	T	0,0
10023	W	14-12-2020	DH3152	Freak2	Coklat	K	31,5	1,4	Y	0,4
20129	P	17-12-2020	SKE5171	Monster	Hitam	M	44	1,2	T	0,0
10150	W	19-12-2020	SKE1295	D'lite3	Hitam	W	39	1,0	Y	0,1

Bila data pada Tabel 1.1 diatur seperti yang telihat pada Tabel 1.2, maka data ini akan memiliki makna. Data ini merupakan catatan pembelian sepatu pada sebuah toko. Baris pada data ini menunjukaan *who* – siapa yang membeli sepatu di toko tersebut. *What*: menyatakan karakteristik dari tiap individu yang tercatat pada data di atas. Karakteristik ini biasa disebut sebagai variabel (dapat dilihat pada kolom pada Tabel 1.2).

Tabel 1.2 Data penjualan pada sebuah toko sepatu

ID	Sex	Tanggal	Kode barang	Nama barang	Warna	Jenis	Ukuran	Harga*	Diskon	Besar diskon
10129	W	12-12-2020	AQ1134	Sambarose	Putih	W	38.67	1,8	T	0,0
10023	W	14-12-2020	DH3152	Freak2	Coklat	K	31,5	1,4	Y	0,4
20129	P	17-12-2020	SKE5171	Monster	Hitam	M	44	1,2	T	0,0
10150	W	19-12-2020	SKE1295	D'lite3	Hitam	W	39	1,0	Y	0,1

Pada Tabel 1.2, perhatikan kolom **Id**, angka-angka yang tercantum pada Id ini hanyalah label yang menunjukkan pelanggan dari toko ini dan nilainya adalah sebarang. Mereka mewakili kategori dari sebuah variabel, dan biasanya disebut sebagai **variabel kategori** (kualitatif). Pada data kategori ini, kita tidak bisa membandingkan bahwa angka 10023 lebih kecil dari 10129. Kedua angka ini hanyalah merujuk pada pelangan yang diwakilinya, dan kedua angka tersebut tidak menunjukkan suatu ukuran tertentu.

Id ataupun Kode Barang yang terlihat pada Tabel 1.2, biasanya akan terhubung dengan tabel lain yang menyimpan tentang informasi tentang pelanggan ataupun barang secara detail.

Perhatikan juga kolom **Harga**. Variabel yang tercatat pada kolom ini merupakan bilangan yang disebut sebagai **variabel kuantitatif**. Variabel kuantitatif merupakan variabel yang

terukur. Angka 1,8 pada variabel ini dapat dibandingkan dengan angka 1. Kita dapat mengatakan bahwa sepatu Sambarose lebih mahal dari sepatu D'lite 3.

Secara umum tipe data dapat diklasifikasikan menjadi dua (lihat Gambar 1.1.), yaitu: data kuantitatif (data numerik) dan data kualitatif (data kategori).

Data kuantitatif merupakan data yang menyatakan ukuran, misalnya umur, tinggi badan, berat badan, jumlah anak, jumlah pekerja, dsb. Ukuran pada data kuantitatif ini dapat berupa bilangan real atau kontinu, misalnya, suhu ruangan, tinggi badan, jarak, dsb; ataupun berupa bilangan integer atau diskrit, misalnya, jumlah anak, jumlah pekerja dsb. Data kuantitatif dapat digolongkan menjadi dua yaitu data interval dan data rasio.

Data interval adalah data kuantitatif yang diukur dalam suatu skala. Setiap nilai pada data interval memiliki arti dan jarak yang konsisten. Data interval dapat dijumlahkan dan dikurangkan, tetapi tidak dapat dibagi ataupun dikalikan. Tidak terdapat nilai nol mutlak (true zero). Nol pada data interval diartikan sebagai sebuah titik sebarang, bukan diartikan sebagai kenihilan dari sebuah variabel yang diukur. Contoh: Temperature, setiap perbedaan 1 derajat akan memiliki makna yang konsisten di seluruh skala. Temperatur dapat ditambahkan ataupun dikurangi, namun tidak dapat dibagi, karena tidak akan memiliki makna. Temperature memiliki nilai 0 derajat. Tetapi nilai 0 ini menunjukkan titik referensi yang disepakati secara ilmiah, ketika es membeku maka temperaturenya adalah nol. Bukan berarti tidak ada temperatur.

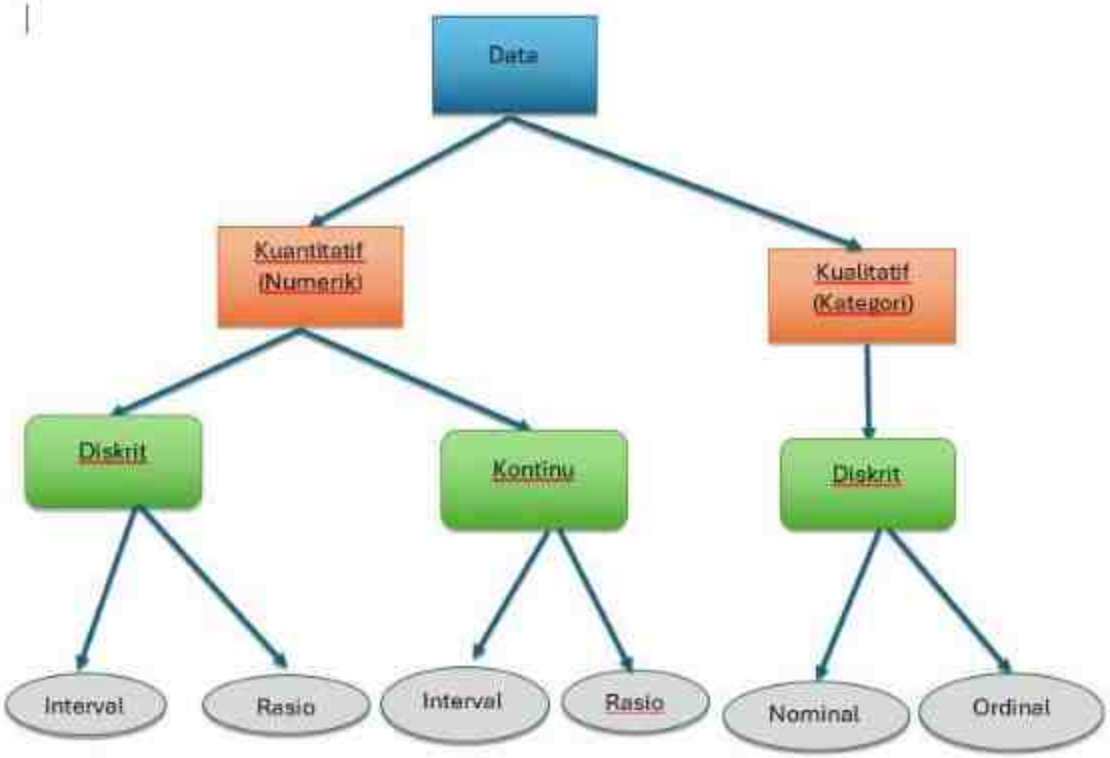
Data ratio adalah data kuantitatif yang diukur dalam suatu skala, memiliki nilai nol mutlak, dan semua operator aritmatika (tambah, kurang, kali, bagi) dapat dioperasikan pada data ini. Contoh data ratio adalah harga, umur, tinggi badan, berat badan, jarak, waktu tempuh. Nilai nol pada sebuah harga barang menandakan bahwa barang tersebut diberikan secara gratis.

Data kualitatif (data kategori), merupakan data yang mengkategorikan atau mengklasifikasikan obyek-obyek yang diamati ke dalam suatu kategori ataupun kelas tertentu. Misalnya, pelajar dapat dikategorikan menjadi anak SD, SMP, SMU dan Universitas. Pegawai dapat dikelompokkan berdasarkan kelompok A, kelompok B dsb. Untuk mempermudah pengolahan data, biasanya tipe data ini dilabelkan ke dalam angka integer (diskrit). Misalnya: 1. SD, 2. SMP, 3. SMU dan 4. Universitas; ataupun 1. Kelompok A, 2. Kelompok B, Kelompok C, dst. Angka diskrit pada data kualitatif, tidak dapat diperbandingkan.

Bila dicermati lagi, maka tipe data kualitatif ini dapat dibagi menjadi dua, yaitu data nominal dan data ordinal. Data nominal adalah data kualitatif yang tidak memperhatikan adanya urutan ataupun strata. Misalnya, pada pengelompokan pegawai berdasarkan 1. Kelompok A, 2. Kelompok B dan 3. Kelompok C. Di sini angka 1, 2 dan 3 benar-benar hanya menyatakan label saja dan tidak bisa diperbandingkan bahwa 1 lebih kecil dari 2, atau Kelompok A lebih buruk dari Kelompok B. Pelabelan inipun dapat diganti-ganti, misalnya 1. Kelompok C, 2. Kelompok B, dan 3. Kelompok A. Lain halnya pada pengelompokan pelajar berdasarkan jenjang pendidikan di sini, anak SMP, memiliki jenjang pendidikan lebih tinggi bila dibandingkan dengan anak SD, demikian pula anak SMU memiliki jenjang pendidikan lebih tinggi bila dibandingakn dengan anak SD maupun anak SMP. Data kategori yang memperhatikan

urutan ataupun jenjang seperti ini dinamakan data ordinal. Pada data ordinal ini label 1,2,3 pada 1. SD, 2. SMP dan 3. SMU, memiliki makna urutan.

Tujuan dari mengetahui tipe data adalah data kuantitatif dan data kualitatif memiliki karakteristik yang berbeda, keduanya akan dianalisa dengan cara yang berbeda pula.



Gambar 1.1. Tipe data

1.2 Mengetahui Jenis Data di R

R mengenali jenis data: Integer, numerik, complex, character dan logical. Untuk mengetahui jenis data yang terdapat pada R dapat digunakan fungsi class().

```
DataInteger = -10:10    #Bilangan bulat (integer)
class(DataInteger)
```

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

```
DataNumerik = runif(100) #Bilangan real
class(DataNumerik)
```

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

```
DataChar  = c("A", "B", "C") #Character
class(DataChar)
```

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

```
Benar = TRUE    #logical
Salah = FALSE
class(Benar)
```

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

```
class(Salah)
```

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

```
#Bilangan kompleks
Kompleks = complex(real = 2, imaginary = 1)
class(Kompleks)
```

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

```
Mat = as.matrix(DataInteger,10,10)  #Matrix atau array
class(Mat)
```

```
[1] "matrix" "array"
```

```
list_data = list(Matrix = Mat, Num = DataInteger)
class(list_data)
```

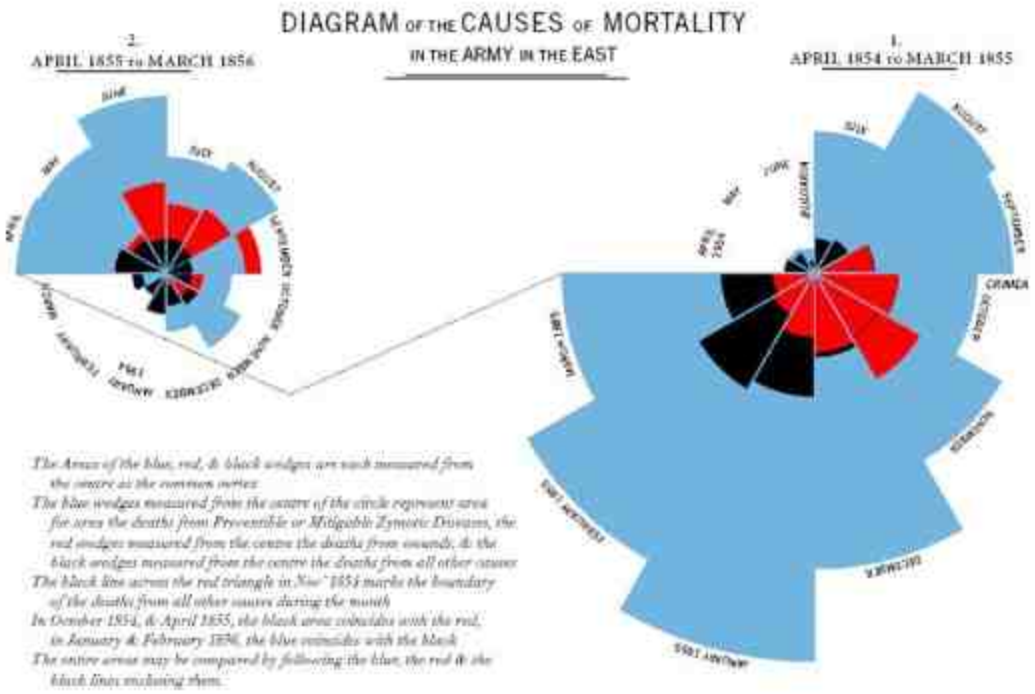
```
[1] "list"
```

1.3 Menampilkan dan Mendeskripsikan Data Kategori

Menampilkan dan mendiskripsikan data, sangatlah penting. Kita tidak dapat melihat hal-hal yang tersembunyi pada tumpukan data dalam tabel. Menampilkan data dengan desain yang baik akan membantu kita untuk melihat pola (*pattern*), hubungan (*relationship*) yang tersembunyi antara satu variabel dengan variabel yang lain, dan kecenderungan (*trend*).

Berikut adalah salah satu contoh display yang sangat terkenal yaitu diagram penyebab kematian pada perang dunia 1 (Gambar1.2.). Florence Nightangle, adalah seorang perawat dan statistikawan. Dia adalah pelopor dunia keperawatan modern. Pada perang Krimea (1854), Florence Nightangle menemukan bahwa, sebagian besar para tentara tersebut mati bukan karena tertembak musuh, namun karena terluka. Perawatan terhadap para tentara yang terluka, melalui kebersihan rumah sakit dan penanganan terhadap luka yang baik akan menyelamatkan banyak jiwa. Florence menganalisa data tersebut dan menampilkannya dalam bentuk diagaram yang disebut sebagai diagram penyebab kematian.

(https://en.wikipedia.org/wiki/Florence_Nightingale).



Gambar 1.2. Diagram penyebab kematian yang terjadi pada perang dunia pertama

https://commons.wikimedia.org/wiki/File:Nightingale-mortality.jpg

Contoh data yang akan diolah pada bab ini adalah data tentang penumpang kapal RMS Titanic. RMS Titanic adalah kapal penumpang Inggris yang dioperasikan oleh White Star Line. Titanic, di produksi di Belfast – Irlandia. Pada tanggal 2 April 1912, Titanic memulai pelayaran percobaan dari Belfast-Irlandia, menuju Southampton (Inggris). Setelah percobaan pelayaran ini berhasil Titanic mulai pelayaran perdananya di Southampton pada tanggal 10 April 1912. Titanic diperkirakan mengangkut 2224 penumpang termasuk awak kapal, lebih dari 1500 penumpang meninggal pada kejadian ini. Penumpang kapal ini sebagian besar, 920 orang, berangkat dari Southampton (Inggris), 179 orang adalah penumpang kelas satu; 247 penumpang kelas dua dan 494 penumpang kelas tiga. Penumpang yang lain berangkat dari Cherbourg (Perancis) dan Queenstown (Irlandia). Sebagian data penumpang Kapal Titanic ditampilkan pada Tabel 1.3. Namun sayangnya, Kapal ini tenggelam di Samudra Atlantic Utara pada tanggal 15 April 1912, setelah menabrak gunung es pada pelayaran perdananya dari Southampton, Inggris menuju New York, Amerika Serikat.

Tabel 1.3. Sebagian data penumpang Kapal Titanic (dicuplik dari: https://www.encyclopedia-titanica.org/)

Gender	Age	Class	Embarked	Fare	Sibsp	Parc	Survived
Male	19	3rd	Southampton	8.01	0	0	No
Male	30	2nd	Cherbourg	24.00	1	0	No
Female	45	1st	Southmpton	164.17	1	1	Yes
Male	33	crew	Southmpton				No
Male	46	crew	Southmpton				Yes
Male	46	crew	Southmpton				No
Male	32	crew	Belfast				No
Male	65	1st	Cherbourg	26.11	0	0	No
Male	21	crew	Southmpton				Yes
Male	37	1st	Southmpton	26.11	0	0	No

Terdapat delapan variabel yang dicantumkan pada Tabel 1.3 yaitu

- Gender: data kategori yang menyatakan jenis kelamin dari penumpang, pria (Male) atau wanita (Female)
- Age: data numerik yang merepresentasikan umur penumpang
- Class: data kategori yang menyatakan kelas penumpang yang menaiki Kapal Titanic dan awak kapal. Terdapat tiga kelas: 1st, 2nd, 3rd dan crew
- Embarked: data kategori yang menyatakan pelabuhah embarkasi
- Fare: data numerik yang merepresentasikan biaya tiket yang dibayarkan oleh penumpang
- Sibsp: data numerik yang merepresentasikan jumlah saudara/pasangan yang ikut serta dalam pelayaran
- Parch: data numerik yang merepresentasikan jumlah orang tua/anak yang ikut serta dalam pelayaran
- Survived: data kategori yang menyatakan apakah penumpang selamat (Yes) atau meninggal (No)

Catatan: Crew terdiri dari Deck Crew, Engineering Crew, Restaurant Staff and Victualling Crew.

Data Titanic terdapat pada package stablelearner, dan dapat diakses dengan cara

```
options(warns = 0)
library(stablelearner)
```

```
Warning: package 'stablelearner' was built under R version 4.1.3
```

```
data("titanic")
head(titanic)
```

```
                          name gender age class embarked       country
1            Abbing, Mr. Anthony   male  42   3rd        S United States
2      Abbott, Mr. Eugene Joseph   male  13   3rd        S United States
3    Abbott, Mr. Rossmore Edward   male  16   3rd        S United States
4 Abbott, Mrs. Rhoda Mary 'Rosa' female  39   3rd        S       England
5    Abelseth, Miss. Karen Marie female  16   3rd        S        Norway
6  Abelseth, Mr. Olaus Jørgensen   male  25   3rd        S United States
  ticketno  fare sibsp parch survived
1     5547  7.11     0     0       no
2     2673 20.05     0     2       no
3     2673 20.05     1     1       no
4     2673 20.05     1     1      yes
5   348125  7.13     0     0      yes
6   348122  7.13     0     0      yes
```

```
dim(titanic)
```

```
[1] 2207   11
```

Untuk mengetahui variable yang digunakan pada data titanic dapat digunakan perintah

```
names(titanic)
```

```
 [1] "name"     "gender"   "age"      "class"    "embarked" "country"
 [7] "ticketno" "fare"     "sibsp"    "parch"    "survived"
```

1.3.1 Tabel Frekuensi dan Tabel Frekuensi Relatif

Untuk mendiskripsikan data tersebut, pertama-tama kita dapat membuat tabel frekuensi dan tabel frekuensi relatif. Untuk itu kita perlu mendefinisikan variabel Kelas dan Selamat sebagai data kategori dari class dan survived. Hal ini dapat dilakukan dengan menggunakan R sebagai berikut:

```
Kelas   = factor(titanic$class)
Selamat = factor(titanic$survived)

levels(Kelas)
```

```
[1] "1st"              "2nd"              "3rd"              "deck crew"
[5] "engineering crew" "restaurant staff" "victualling crew"
```

```
levels(Selamat)
```

```
[1] "no"  "yes"
```

Pada kedua data kategori ini kita ketahui bahwa Kelas memiliki 7 levels dan Selamat memiliki 2 levels. Untuk menyederhanakan deskripsi data, maka crew akan dijadikan menjadi satu level saja.

```
Kelas = as.character(Kelas)
Kelas[Kelas == "deck crew"] = "crew"
Kelas[Kelas == "engineering crew"] = "crew"
Kelas[Kelas == "restaurant staff"] = "crew"
Kelas[Kelas == "victualling crew"] = "crew"

Kelas = factor(Kelas)
levels(Kelas)
```

```
[1] "1st"  "2nd"  "3rd"  "crew"
```

```
titanic$class = Kelas
```

Saat ini variabel Kelas hanya memiliki 4 levels saja. Frekuensi dari masing-masing level pada Kelas penumpang dapat dihitung dengan menggunakan perintah table sebagai berikut

```
L = length(Kelas)       #Jumlah data

Freq.Kelas = table(Kelas)   #Tabel Frekuensi
Freq.Kelas
```

```
Kelas
 1st  2nd  3rd crew
 324  284  709  890
```

Sedangkan nilai frekuensi relatif dapat diperoleh dengan membagi Freq.Kelas dengan jumlah data, yaitu L (length dari Kelas)

```
Freq.Relatif = Freq.Kelas/L #Tabel Frekuensi Relatif
Freq.Relatif
```

```
Kelas
      1st       2nd       3rd      crew
0.1468056 0.1286815 0.3212506 0.4032623
```

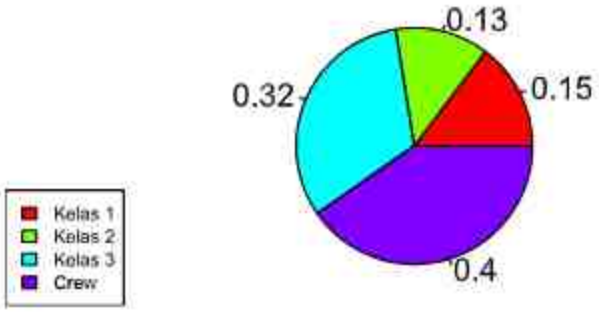
Dari sini dapat kita ketahui bahwa jumlah penumpang Kelas 1 yang tercatat pada data ini adalah 324 (14.68 %) orang, 284 (12.87 %) orang penumpang Kelas 2, 709 (32.13 %) orang penumpang Kelas 3 dan 890 (40.33 %) adalah crew dari kapal Titanic.

1.3.2 Bar Chart dan Pie Chart

Bar chart dan Pie chart adalah chart sederhana yang biasanya digunakan untuk menggambarkan data kuantitatif. Tabel frekuensi di atas dapat digambarkan dengan menggunakan barchart, sedangkan frekuensi relatif dapat digambarkan dengan menggunakan pie chart (lihat Gambar 1.3). Dari sini kita dapat melihat bahwa lebih dari 40.33 persen penumpang Kapal Titanic, adalah penumpang awak kapal (*crew*).

```
barplot(Freq.Kelas, xlab = "Kelas Penumpang", ylab = "Frekuensi",
   main = "Frekuensi Penumpang Titanic berdasarkan Kelas", col = rainbow(4))
```


```
pie(Freq.Relatif,col = rainbow(4),labels = round(Freq.Relatif,2),
      main = "Pie Chart Penumpang Titanic berdasarkan Kelas")
legend("bottomleft",1,legend = c("Kelas 1", "Kelas 2", "Kelas 3","Crew"),
       cex = 0.6, fill = rainbow(4))
```



1.3.3 Tabel Kontingensi (*Contingency Table*)

Selanjutnya kita ingin mengetahui bagaimana proporsi penumpang yang selamat di tiap kelas ini. Apakah ketiga kelas ini memiliki kesempatan yang sama untuk selamat (*survived*), atau tingkat keselamatan ini timpang antara satu kelas dengan yang lain.

Untuk membandingkan dua buah variabel secara bersama-sama kita bisa membuat tabel kontingensi (*Contingency Table*) atau krosstabulasi (*Cross Tabulation*), atau sering juga disebut sebagai pivoting. Ada dua cara untuk membuat pivot table dengan menggunakan R yang akan kita pelajari. Pertama kita dapat membuat pivot table secara manual atau menggunakan **package pivottabler**.

Untuk membuat pivot table secara manual, maka pertama kita akan membuat cross tabulasi antara variabel Selamat dan variabel Kelas.

```
Tab.Sel.Kel = table(Selamat,Kelas)      #CrossTabulasi Selamat&Kelas
Tab.Sel.Kel
```

```
       Kelas
Selamat 1st 2nd 3rd crew
    no  123 166 528  679
    yes 201 118 181  211
```

Setelah itu kita akan menghitung jumlah dari penumpang yang tidak selamat (No) dan penumpang yang selamat (Yes). Tabel di atas dapat dianggap sebagai matrix, dan kita dapat menggunakan perintah *apply* untuk menghitung jumlahan per baris ataupun per kolom dari sebuah matrix. Untuk tabel di atas dapat dituliskan perintah

```
TotSel = apply(Tab.Sel.Kel,1,sum)                #Total per variabel Kelas
TotSel
```

```
  no  yes
1496  711
```

Pada variabel TotSel di atas, perintah *apply* akan menjumlah (sum) tiap baris (1) dari matrix Tab.Sel.Kel. Bila kita ingin menjumlah tiap kolom maka nilai 1 di atas diganti dengan 2. Secara umum fungsi sum dapat diubah sesuai dengan kebutuhan. Kita juga dapat mendifinisikan sebuah fungsi sendiri untuk digunakan pada perintah *apply*.

Selanjutanya kita akan menggabungkan secara kolom cross tabulasi dan GrandKelas dengan menggunakan perintah **cbind** (*column binding*)

```
CTab1 = cbind(Tab.Sel.Kel, TotSel)       #Menggabungkan secara kolom
CTab1
```

```
    1st 2nd 3rd crew TotSel
no  123 166 528  679   1496
yes 201 118 181  211    711
```

Demikian selanjutnya, kita akan menjumlah tiap kolom dari CTab1 dan menggabungkan hasilnya secara baris (**rbind** -*row binding*) dengan tabel yang telah kita buat di atas.

```
TotKelas      = apply(CTab1,2,sum)  #Total per variable Selamat
Pivot.Sel.Kel = rbind(CTab1,TotKelas)   #Menggabungkan secara baris
Pivot.Sel.Kel
```

```
         1st 2nd 3rd crew TotSel
no       123 166 528  679   1496
yes      201 118 181  211    711
TotKelas 324 284 709  890   2207
```

Cara kedua untuk membuat pivot table adalah dengan menggunakan *library(pivottabler)*. Tentu saja kita harus install.packages("pivottabler"), sebelum menggunakan library ini.

```
library(pivottabler)
```

```
Warning: package 'pivottabler' was built under R version 4.1.3
```

```
qhpvt(titanic,"class","survived","n()")
```

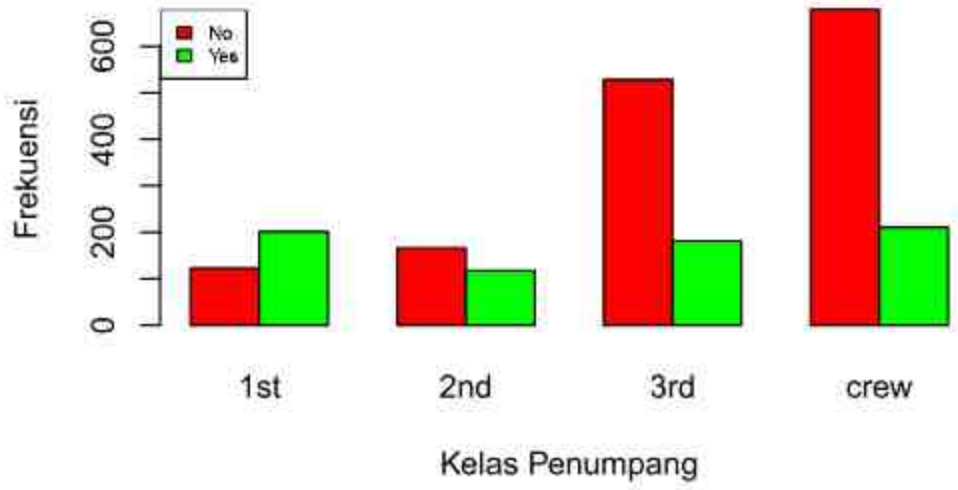
```
pt1  = PivotTable$new()
pt1$addData(titanic)
pt1$addColumnDataGroups("class")
pt1$addRowDataGroups("survived")
pt1$defineCalculation(calculationName = "Total", summariseExpression ="n()")
pt1$renderPivot()
```

```
pt1$evaluatePivot()
pt1
```

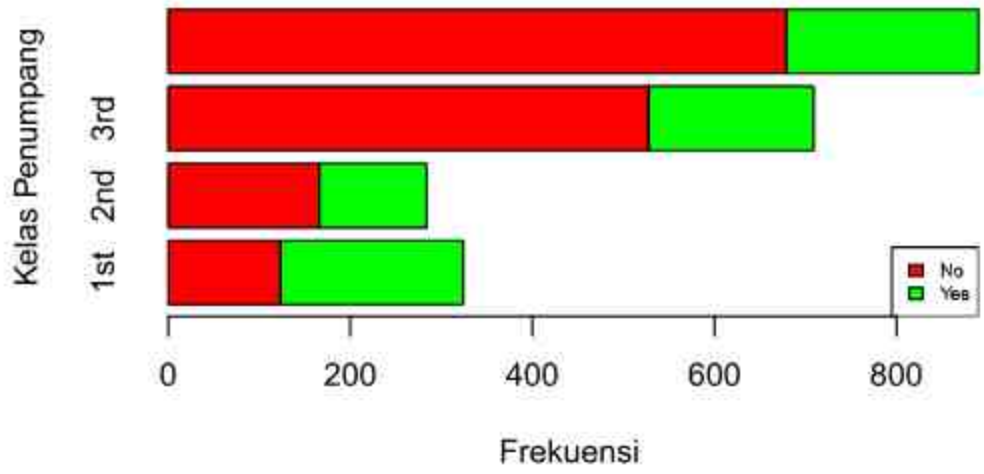
```
       1st  2nd  3rd  crew  Total
no     123  166  528   679   1496
yes    201  118  181   211    711
Total  324  284  709   890   2207
```

Perbandingan distribusi data pada setiap kelas secara bersama-sama ini, seringkali ditambilkan dalam bentuk *grouped barchart* ataupun *stacked barchart*. Bila kedua distribusi ini disandingkan terlihat bahwa variable Selamat tidak tersebar secara merata untuk seluruh penumpang Titanic. Terlihat tingkat kematian pada penumpanng kelas tiga dan crew sangat tinggi bila dibandingkan dengan tingkat kematian pada penumpang kelas satu ataupun kelas dua (Gambar 1.4).

```
barplot(Tab.Sel.Kel,xlab = "Kelas Penumpang", ylab = "Frekuensi",beside = TRUE,
      col = c("red","green"))
legend("topleft",1,legend = c("No", "Yes"),
cex = 0.6,fill = c("red","green"))
```




```
barplot(Tab.Sel.Kel, ylab = "Kelas Penumpang", xlab="Frekuensi",col = c("red","green"),
        horiz = TRUE)
legend("bottomright",1,legend = c("No", "Yes"),
cex = 0.6,fill = c("red","green"))
```



1.3.4 Distribusi Marginal

Pada tabel kontingensi kita memperhatikan penyebaran/distribusi dari dua variabel kuantitatif secara bersama-sama. Pada contoh di atas kita memperhatikan distribusi dari variabel Selamat dan Kelas secara bersama-sama. Namun bila kita hanya ingin memperhatikan salah satu dari keduanya, misalnya hanya distribusi dari variabel Selamat saja (kolom TotSel) atau distribusi dari variabel Kelas saja (baris Tot Kelas) maka distribusi ini disebut sebagai **Distribusi Marginal** atau distribusi pinggiran.

Terkadang melakukan melihat distribusi secara marginal akan memberikan informasi yang lebih dalam bila dibandingkan dengan melihat secara keseluruhan. Hal ini dapat kita gali dari tingkat keselamatan penumpang terhadap kelas penumpang tersebut.

Bila kita membandingkan terhadap total data yang tercatat, terlihat bahwa **secara keseluruhan** 9,1% penumpang Kelas 1 selamat. Hal ini tidak terlihat terlalu berbeda dengan *crew*, **secara keseluruhan** 9.6% *crew* selamat.

```
options(digits = 2)
Persen_Total = Pivot.Sel.Kel/L
Persen_Total
```

```
          1st   2nd   3rd  crew TotSel
no      0.056 0.075 0.239 0.308   0.68
```

```
yes       0.091 0.053 0.082 0.096   0.32
TotKelas 0.147 0.129 0.321 0.403   1.00
```

Demikian juga hal nya kita bisa membandingkan distribusi dari penumpang yang selamat ini terhadap salah satu variabel secara marginal. Pada contoh di bawah ini, kita membandingkan keselamatan penumpang Titanic, terhadap variabel Selamat secara marginal. Pada pembandingan ini kita dapat menyatakan **dari seluruh penumpang yang selamat**, 28% berasal dari Kelas 1 dan 30% adalah crew. Namun **dari seluruh penumpang yang tidak selamat**, hanya 9% merupakan penumpang Kelas 1, dan 45% adalah *crew*.

```
Persen_Selamat = apply(Pivot.Sel.Kel,2,"/",c(TotSel,L))
Persen_Selamat
```

```
            1st  2nd  3rd crew TotSel
no        0.082 0.11 0.35 0.45      1
yes       0.283 0.17 0.25 0.30      1
TotKelas 0.147 0.13 0.32 0.40      1
```

Bila kita melakukan perbandingan distribusi keselamatan penumpang Titanic, untuk setiap kelas secara marginal, maka dapat kita katakan 62% **dari seluruh penumpang Kelas 1**, selamat, namun hanya 24% **dari seluruh *crew*** yang selamat pada peristiwa tenggelamnya kapal Titanic.

```
Persen_Kelas =  t(apply(Pivot.Sel.Kel,1,"/",TotKelas))
Persen_Kelas
```

```
            1st  2nd  3rd crew TotSel
no        0.38 0.58 0.74 0.76   0.68
yes       0.62 0.42 0.26 0.24   0.32
TotKelas 1.00 1.00 1.00 1.00   1.00
```

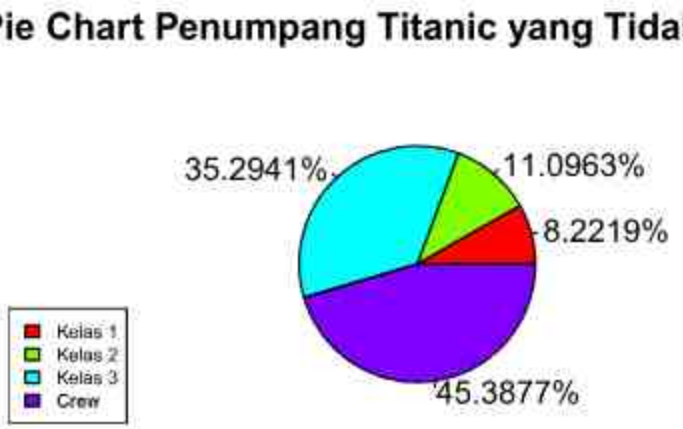
Kita perlu berhati-hati dalam menyampaikan dan membaca laporan statistik, memilih perbandingan terhadap variabel yang berbeda akan memberikan efek yang berbeda ketika seseorang membacanya. Bila kita melihat distribusi yang tergambar pada Gambar 1.4 maka perbandingan kondisi keselamatan penumpang yang didasarkan pada tiap kelas adalah perbandingan yang sesuai (*fair*). Pada peristiwa tenggelamnya kapal Titanic, 76% dari seluruh *crew* tidak selamat, sedangkan 62% dari seluruh penumpang Kelas 1 selamat.

Melakukan perbandingan secara tidak fair, disebut sebagai **Simpson Paradox**. Pada contoh ini membandingkan keselamatan penumpang terhadap seluruh penumpang yang tercatat tanpa mengklasifikasikan penumpang berdasarkan Kelasnya adalah Simpson Paradox. Hal ini karena kita menyamaratakan prosentase keselamatan penumpang Titanic, tanpa memandang bahwa jumlah penumpang Titanic per Kelas berbeda.

1.3.5 Distribusi Bersyarat (*Conditional Distribution*)

Sebuah distribusi dikatakan bersyarat, apabila distribusi dari data tersebut dibatasi (diberi syarat) pada salah satu pilihan yang terdapat pada salah satu variabel yang ada pada data kita. Misalnya pada contoh di atas, distribusi bersyarat dari penumpang yang Tidak Selamat adalah

```
Persen_Selamat1 <- as.vector(Persen_Selamat[1,1:4])

label_percent = function(X)
{ paste0(round(X*100,4),"%")
}
Persen.Selamat1 <- label_percent(Persen_Selamat1)

pie(Persen_Selamat1,col = rainbow(4),labels = Persen.Selamat1,
      main = "Pie Chart Penumpang Titanic yang Tidak Selamat")
legend("bottomleft",1,legend = c("Kelas 1", "Kelas 2", "Kelas 3", "Crew"),
cex = 0.6,fill = rainbow(4))
```



Sedangkan distribusi bersyarat dari penumpang yang Selamat adalah

```
Persen_Selamat1 <- as.vector(Persen_Selamat[2,1:4])
Persen.Selamat1 <- label_percent(Persen_Selamat1)
```

```
pie(Persen_Selamat1,col = rainbow(4),labels = Persen.Selamat1,
    main = "Pie Chart Penumpang Titanic yang Selamat")
legend("bottomleft",1,legend = c("Kelas 1", "Kelas 2", "Kelas 3", "Crew"),
    cex = 0.6,fill = rainbow(4))
```

Dalam tabel kontingensi, apabila distribusi dari satu variabel sama untuk semua kategori dari variabel yang lain, maka kita dapat katakan kedua variabel tersebut saling independen.

```
library(lattice)

trellis.device()
barchart(t(Tab.Sel.Kel), horizontal = FALSE,
 groups = FALSE, xlab = "Kelas Penumpang", col = "blue")
```

Referensi buku untuk mempelajari graph di R dapat digunakan link berikut:

https://r-graphics.org/

1.3.6 Simpson Paradox

Simpson's Paradox adalah fenomena dalam statistik, dimana asosiasi yang terjadi diantara dua variable menjadi hilang atau berubah arah ketika populasi dari data tersebut dibagi menjadi

beberapa subpopulasi. (https://plato.stanford.edu/entries/paradox-simpson/)

Salah satu contoh yang terkenal dari Simpson Paradox adalah UC Berkeley gender bias (Freedman *et al.*, 2007, Bickel *et al.*, 1975). Bila dilihat data penerimaan mahasiswa di UC Berkeley pada tahun 1973 (Tabel 1.4) terlihat bahwa pelamar pria memiliki kesempatan lebih besar untuk diterima di UC Berkeley di bandingkan wanita.

Tabel 1.4. Pelamar yang diterima di UC Berkeley 1973 secara keseluruhan

	Semua		Pria		Wanita	
	Pelamar	Diterima	Pelamar	Diterima	Pelamar	Diterima
Total	12763	41%	8442	44%	4321	35%

Namun bila populasi pendaftar di UC Berkeley ini dibagi-bagi lagi menjadi subpopulasi, yaitu per departemen (Tabel 1.5), maka akan terlihat bahwa asosiasi yang dinyatakan pada Tabel 1.4 tidak seluruhnya benar. Pada departemen dengan persentase penerimaan rendah (Departemen F), wanita lebih banyak diterima daripada pria. Demikian juga dengan dua departemen teratas di UC Berkeley (Departemen A dan B), wanita juga lebih banyak diterima untuk bersekolah di sana bila dibandingkan pria. Melakukan perbandingan secara "tidak fair" seperti ini disebut sebagai Simpson Paradox.

Tabel 1.5. Pelamar yang diterima di UC Berkeley 1973 per departemen

Departemen	Semua		Pria		Wanita	
	Pelamar	Diterima	Pelamar	Diterima	Pelamar	Diterima
A	933	64%	825	62%	108	82%
B	585	63%	560	63%	25	68%
C	918	35%	325	37%	593	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	393	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

1.4 Menampilkan dan Mendiskripsikan Data Numerik

Data numerik merepresentasikan suatu ukuran, misalnya panjang, berat badan, biaya, volume, usia, dan waktu. Data numerik dapat bersifat diskrit bila subyek yang diukur tidak dapat dapat direpresentasikan sebagai bilangan pecahan atau desimal; misalnya, jumlah telur, jumlah siswa, jumlah kendaraan yang lewat di pintu tol, dan sebagainya. Namun bila subyek yang diukur tersebut dapat direpresentasikan sebagai bilangan pecahan, maka data numerik

tersebut bersifat kontinu; misalnya berat telur, tinggi seorang siswa. tonase kendaraan yang lewat di pintu tol.

Studi Kasus: Jakarta Stock Exchange

Data berikut merupakan sebagian dari index saham gabungan dari Jakarta stock exchange (JKSE). Pada data ini ditampilkan tanggal, index saham saat pembukaan pasar saham (Open), index gabungan tertinggi, terendah, dan saat pasar saham tersebut tutut (close) dan harga penutupan yang disesuaikan (Adj close), serta volume pembelian yang terjadi di hari itu.

```
warnings(-1)
suppressMessages(library(quantmod))
```

```
Warning: package 'xts' was built under R version 4.1.3

Warning: package 'zoo' was built under R version 4.1.3

Warning: package 'TTR' was built under R version 4.1.3
```

```
warnings(-1)
getSymbols("^JKSE",src='yahoo')
```

```
Warning: ^JKSE contains missing values. Some functions will not work if objects contain miss
values in the middle of the series. Consider using na.omit(), na.approx(), na.fill(), etc to
or replace them.

[1] "JKSE"
```

```
df = data.frame(Date=index(JKSE),coredata(JKSE))
df = na.omit(df)
options(width = 100)
tail(df,5)
```

	Date	JKSE.Open	JKSE.High	JKSE.Low	JKSE.Close	JKSE.Volume	JKSE.Adjusted
4397	2024-11-07	7374	7382	7244	7244	2.0e+08	7244
4398	2024-11-08	7264	7350	7264	7287	1.5e+08	7287
4399	2024-11-11	7259	7282	7182	7266	2.2e+08	7266
4400	2024-11-12	7271	7344	7269	7322	2.9e+08	7322
4401	2024-11-13	7344	7370	7305	7309	2.3e+08	7309

Bila kita hanya melihat angka-angka ini maka kita tidak mendapatkan pola, hubungan antar data ataupun trend yang terdapat pada data di atas. Untuk itu kita perlu mendeskripsikan data tersebut melalui *chart* ataupun membuat ringkasan (*summary*) statistik.

1.4.1 Histogram

Salah satu *chart* yang sering digunakan untuk mendiskripsikan data numerik adalah histogram. Histogram ini mirip dengan *bar chart*, namun tidak memiliki jeda di antara batang-batang nya. Histogram dibuat dengan cara membagi-bagi interval data yang tercatat ke dalam sejumlah kelas yang disebut sebagai bin. Kita akan menghitung frekuensi data yang berada di dalam setiap bin yang sesuai. Histogram menggambarkan distribusi frekuensi data pada kelas-kelas yang terbagi di sepanjang interval dari data tersebut.

Bila jumlah bin ini banyak, maka lebar tiap bin ini akan sempit. Akibatnya jumlah data yang masuk ke dalam tiap bin tersebut sedikit, sehingga distribusi dari data itu sangat bergerigi.

Bila jumlah bin sedikit, maka lebar tiap bin akan sangat besar. Akibatnya jumlah data yang masuk ke dalam bin tersebut banyak, sehingga distribusi dari data itu sangat mulus (*over smooth*).

Formula yang sering digunakan untuk menghitung jumlah bin adalah Sturges

$$K = 1 + 3,222Log(N)$$

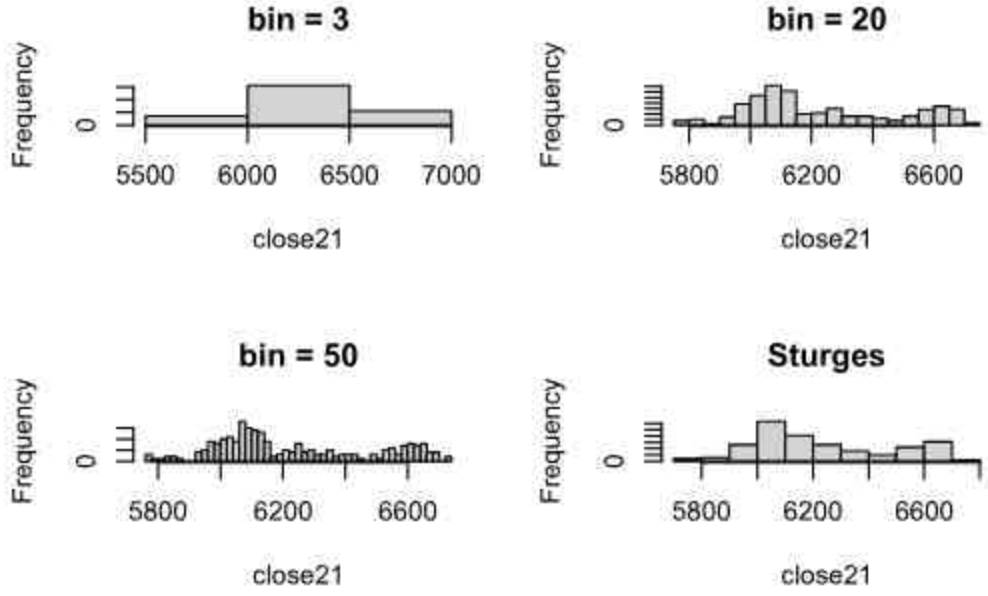
dimana K adalah jumlah bin dan N adalah jumlah data

Gambar di bawah menunjukkan histogram dari data Jakarta Stock Exchange (JKSE) tahun 2021, dengan nilai bin yang berubah-ubah. Dapat dilihat bahwa jumlah bin sangat mempengaruhi bentuk dan sebaran dari data JKSE ini. Pada kasus JKSE ini, bila bin = 3, terlihat bahwa index saham gabungan (IHSG) paling sering muncul di kisaran 6000 hingga 6500. Range IHSG sebesar 500 tentulah sangat besar, dan kita tidak bisa melihat sebenarnya pada kisaran harga berapakah IHSG itu sering terjadi bila histogram ini terlalu mulus (*over smooth*). Sebaliknya bila jumlah bin = 50, besaran interval antar bin adalah 19 point. Harga IHSG yang sering muncul berada di kisaran 6069-6088, 6088-6107, 6107-6126 dan pada masing-masing interval frekuensi kemunculanya hanya 16 kali. Jumlah data perdagangan setahun 245 hari, maka probabilitas kemunculan nilai IHSG di range tersebut hanyalah 16/245 = 0.065. Tentu saja ini tidak memberikan informasi apapun.

```
#install.packages("lubridate")
warnings(-1)
suppressMessages(library(lubridate))
```

```
Warning: package 'lubridate' was built under R version 4.1.3
```

```
close = df$JKSE.Close
Year  = year(df$Date)
Data  = data.frame(close,Year)
Data21 = Data[which(Data$Year==2021),]
close21 = Data21$close
op = par(mfrow = c(2,2))
hist(close21,nclass = 3, main = "bin = 3")
hist(close21,nclass = 20, main = "bin = 20")
hist(close21,nclass = 50, main = "bin = 50")
hist(close21, main = "Sturges")
```




```
par(op)
```

1.4.1.1 Bentuk, Pusat dan Sebaran

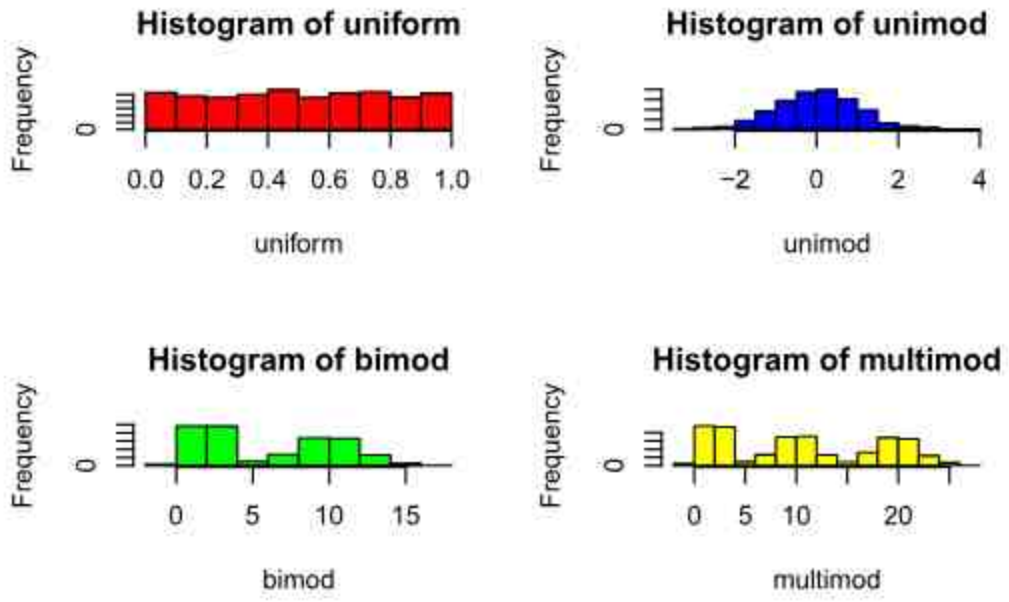
1.4.1.1.1 Bentuk Distribusi

Bentuk histogram merepresentasikan darimana data kita berasal. Bila data berasal dari pengukuran yang seragam, misalnya data tinggi tentara, maka data tersebut tidak memiliki puncak. Histogram dari data yang tidak memiliki puncak ini disebut **uniform** (seragam). Histogram yang memiliki satu puncak disebut dengan **unimodal**, disebut **bimodal** bila memiliki dua puncak, dan **multimodal** bila memiliki lebih dari dua puncak (Gambar 1.8).

Data yang berasal dari dua populasi yang berbeda dengan jarak antar puncak yang berjauhan akan memiliki bimodal. Contoh, ukuran panjang rambut pria dan wanita akan memiliki distribusi binodal. Seperti kita ketahui, pria cenderung memiliki rambut pendek dan wanita cenderung memiliki rambut panjang. Tentu saja terdapat pria berambut panjang dan wanita berambut pendek. Namun jumlah pria berambut panjang tidaklah sebanyak pria berambut pendek demikian juga dengan jumlah wanita berambut pendek tidak sebanyak wanita berambut panjang.

Melihat bentuk histogram kita dapat menduga, asal populasi dari pengukuran data ini, apakah dari satu populasi dengan batasan tertentu (misalkan tinggi tentara), satu populasi homogen, atau populasi yang bercampur (heterogen).

```
uniform   = runif(1000)
unimod    = rnorm(1000)
bimod     = c(rnorm(1000, 2,1), rnorm(1000, 10,2))
multimod  = c(rnorm(1000, 2,1), rnorm(1000,10,2), rnorm(1000, 20,2))

op = par(mfrow=c(2,2))
hist(uniform, col = "red")
hist(unimod, col = "blue")
hist(bimod, col = "green")
hist(multimod, col = "yellow")
```

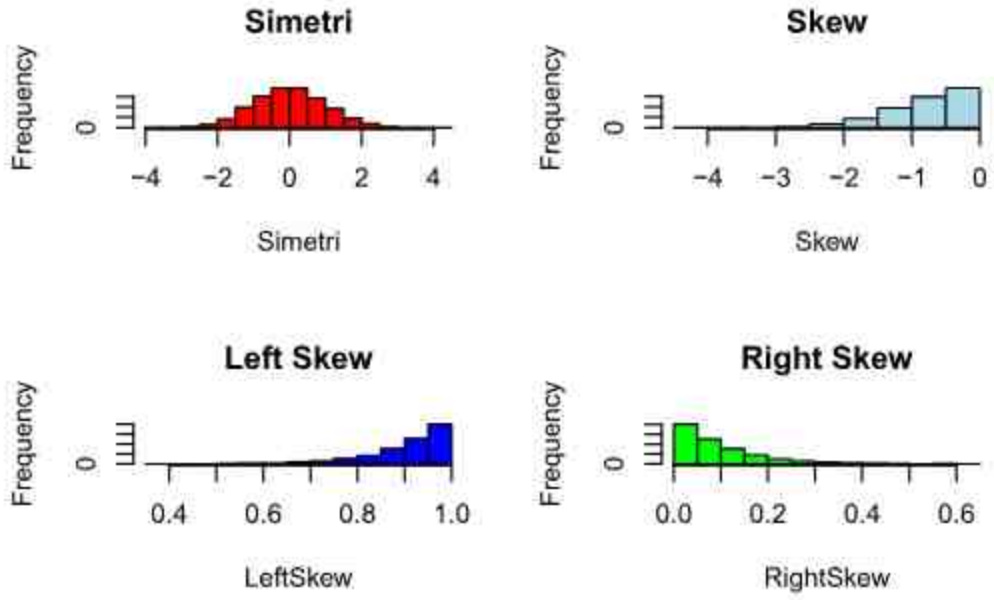


```
par(op)
```

1.4.1.1.2 Simetri

Bentuk histogram dapat pula dilihat dari kesimetriannya. Bila histogram itu dilipat di tengah dan kedua ujungnya dapat berhimpitan cukup dekat, maka histogram itu dikatakan simetri. Namun bila hal di atas tidak terpenuhi, maka histrogram tersebut tidak simetri (*skew*). Gambar di bawah menunjukkan histogram yang simetri dan histogram yang miring.

```
LeftSkew = rbeta(10000,10,1)
RightSkew= rbeta(10000,1,10)
#Simetri = rbeta(10000,10,10)
Simetri  = rnorm(10000)
Skew     = -abs(Simetri)

op = par(mfrow=c(2,2))
hist(Simetri, col = "red", main = 'Simetri')
hist(Skew, col = "light blue", main = 'Skew')
hist(LeftSkew, col = "blue", main = 'Left Skew')
hist(RightSkew, col = "green", main = 'Right Skew')
```



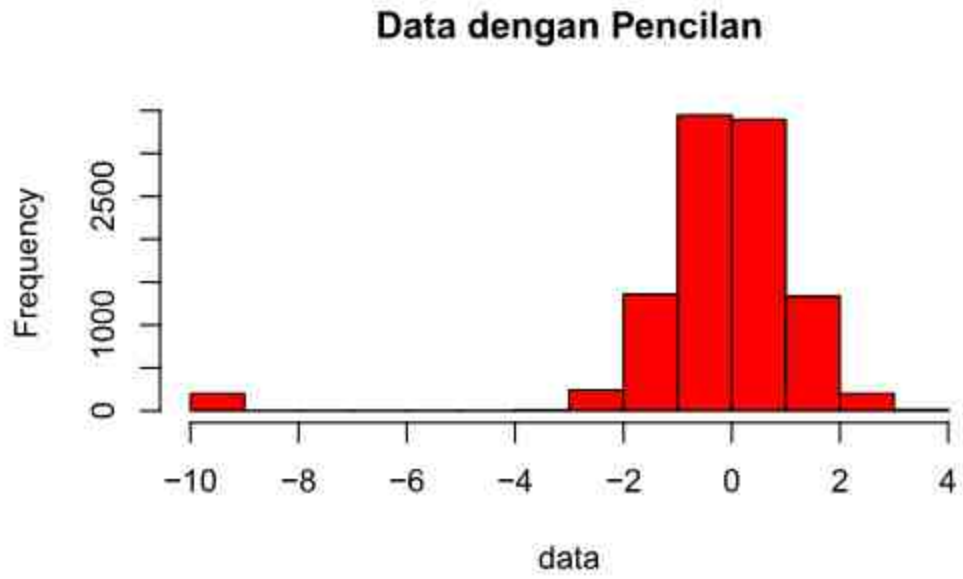
```
par(op)
```

Pada histogram yang miring (*skew*), biasanya ujung yang lebih tipis disebut sebagai ekor (*tail*). Bila ekor itu berada di ujung kiri, maka histrogram itu disebut miring ke kiri (*left skew*), sebaliknya bila ekornya berada di ujung kanan, maka histogram itu disebut miring ke kanan (*right skew*).

1.4.1.1.3 Pencilan (*Outlier*)

Seringkali kita menjumpai data yang keluar jauh dari distribusinya. Data seperti ini disebut sebagai pencilan atau *outlier*. Pencilan mungkin saja merupakan bagian data yang paling informatif, atau mungkin juga pencilan tersebut merupakan kesalahan pencatatan, kesalahan yang lain yang timbul dalam sistem. Untuk itu perlu dilakukan penyelidikan yang seksama terhadap pencilan ini sebelum membuangnya. Bila pencilan ini terjadi karena kesalahan pencatatan maka pencilan ini dapat diperbaiki bila memungkinkan, namun bila tidak data seperti ini dapat dibuang. Bila pencilan yang muncul bukan karena kesalahan pencatatan, maka perlu diperhatikan sistem kerja yang diwakili oleh data tersebut. Bisa saja terjadi perubahan dalam sistem tersebut, dan pencilan seperti ini tidak boleh dibuang karena ia mengindikasikan suatu pergeseran di dalam sistem.

```
data = c(rnorm(10000),rep(-10,200))
hist(data, col = 'red', main = 'Data dengan Pencilan')
```



Secara umum, untuk mendeteksi adanya pencilan di dalam data, kita dapat melihat hitogram utama dari data yang kita miliki. Bila bagian utama dari histogram itu tampak simetri, kemudian muncul satu punukan data entah di ujung kiri atau kanan, maka punukan tersebut dapat kita duga sebagai pencilan (Gambar 1.10). Namun bila histogram yang kita miliki miring (*skew*), maka ekor dari histogram itu tidak dapat kita anggap sebagai pencilan.

1.4.1.1.4 Pusat dan Sebaran

Pusat data biasanya digunakan untuk mendiskripsikan ringkasan yang mewakili seluruh data. Bila histogram tersebut unimodal dan simetrik, maka pusatnya adalah nilai yang muncul di tengah. Namun bila histrogram tersebut miring (*skew*) untuk menentukan pusat data tentulah tidak mudah. Bila histogram tersebut memiliki lebih dari satu modus (multimodal) maka konsep dari pusat data tidaklah dapat digunakan.

Sebaran data mengukur seberapa besar variasi data terhadap pusatnya. Apakah data tersebut berkumpul secara rapat disekitar pusatnya, ataukah tersebar menjauh dari pusatnya.

1.4.2 Diagram Batang dan Daun (Stem and Leafplot, Tukey 1977)

Diagram batang dan daun menampilkan seluruh informasi yang terdapat pada histogram, beserta dengan nilai dari setiap data. Secara prosedur, diagram batang dan daun dapat dituliskan sebagai berikut:

1. Setiap data observasi akan diubah ke dalam batang dan daun. Batang merupakan digit dari data, sedangkan daun terdiri dari angka tunggal
2. Urutkan nilai-nilai yang akan diletakan sebagai batang dari atas ke bawah berdasarkan nilai terkecil ke nilai terbesar, gambarlah garis lurus di kanan batang dan tambahkan daun di kanan garis.
3. Tuliskan nilai-nilai yang ada pada daun dari kiri ke kanan terurut berdasarkan nilai terkecil ke nilai terbesar.

Data berikut dapat digambarkan sebagai diagram batang dan daun:

54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22

Diagram batang dan daun:

2:2 5

3:4 5

4:1 1 6 6 7 9

5:4 4 9

6: 0

Stem and leafplot menggunakan R

Pada contoh ini digunakan data faithful yang terdapat pada R. Data ini terdiri dari kumpulan pengamatan yang dilakukan di Old Faitful geyser di Taman Nasional Yellowstone di Amerika. Data yang dicatat adalah waktu saat geyser ini menyemburkan uap panas. Data eruptions mencatat durasi geysey tersebut mengalami erupsi, sedangkan data waiting mencatat lama waktu

menunggu hingga erupsi berikutnya terjadi. (https://www.frommers.com/slideshows/848448-beyond-old-faithful-a-geyser-gazing-guide-to-yellowstone-national-park)

Data ini dapat diakses di R dengan mengetik faithful di R console. Ada dua observasi yang dicatat, *eruptions* dan *waiting*. Untuk mendeskripsikan data durasi dengan diagram batang dan daun (stem and leafplot), tuliskan.

```
duration = faithful$eruptions
stem(duration)
```

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

 16 | 070355555588
 18 | 000022233333335577777777888822335777888
 20 | 00002223378800035778
 22 | 0002335578023578
 24 | 00228
 26 | 23
 28 | 080
 30 | 7
 32 | 2337
 34 | 250077
 36 | 0000823577
 38 | 2333335582225577
 40 | 000000335778888800233555577778
 42 | 033355557788002333333555577778
 44 | 022223355577800000000233333577788888
 46 | 000023335770000002357 8
 48 | 00000022335800333
 50 | 0370
```

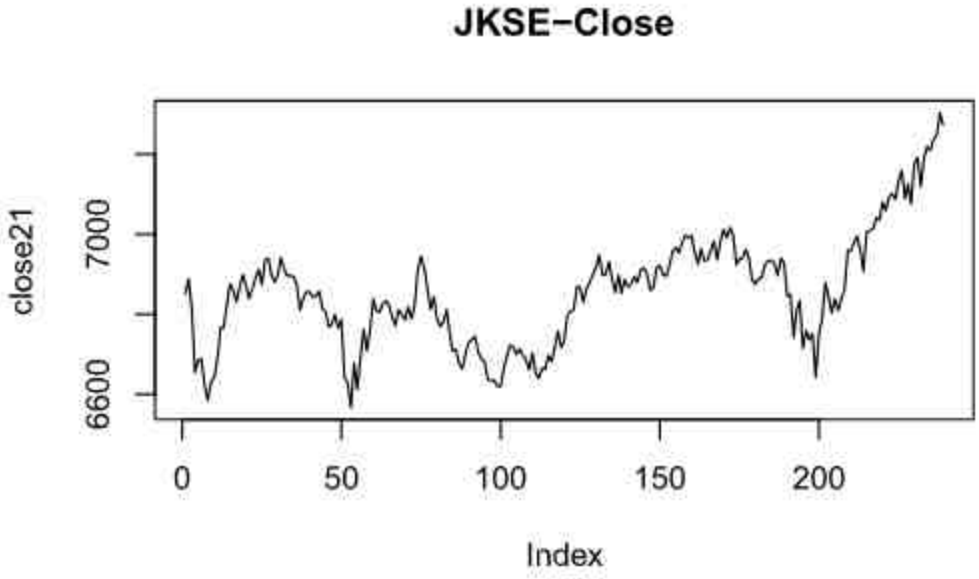
Dari diagram batang dan daun di atas terlihat pada durasi erupsi pada Old Faitful geyser memiki dua puncak (bimodal), yaitu 16 menit dan 44 menit.

Keuntungan dari diagram batang dan daun adalah distribusi dari data terlihat secara langsung: simetri, unimodal, bimodal, dsb. Gap antar data dapat terlihat secara langsung. Kerugiannya adalah diagram ini tidak dapat digunakan apabila jangkauan data cukup besar, misalnya data memuat nilai dari satuan hingga ratusan ribu.

1.4.3 Line Plot

Bila data dicatat secara terurut berdasarkan waktu, seperti pada data JKSE, maka line plot dapat membantu kita melihat *trend* data dari waktu ke waktu. Terlihat bahwa data JKSE memiliki nilai trend yang meningkat dari tahun ke tahun.

```
#install.packages("lubridate")
warnings(-1)
suppressMessages(library(lubridate))
close = df$JKSE.Close
Year  = year(df$Date)
Data  = data.frame(close,Year)
Data21 = Data[which(Data$Year==2023),]
close21 = Data21$close
plot(close21, type = 'l', main = 'JKSE-Close')
```



1.4.4 Ringkasan Data Numerik

Ringkasan data numerik biasanya dinyatakan dalam beberapa statistik, di antaranya adalah:

1.4.4.1 Minimum dan Maximum (Min, Max):

Statistik ini menyatakan nilai terkecil dan nilai terbesar yang terdapat pada data. Pada R dapat dituliskan dengan singkat:

```
min(close)
```

```
[1] 1111
```

```
max(close)
```

```
[1] 7905
```

1.4.4.2 Quantile

Secara umum data dapat dibagi menjadi tiga quantile.

Quantile 1 (Q1), merupakan data dengan urutan ke 25% dari total jumlah data.

Quantile 2 (Q2), biasa juga disebut sebagai Median, merupakan data dengan urutan ke 50% dari total jumlah data. Median merupakan nilai tengah dari seluruh data. Median dapat menyatakan pusat data, terutama bila data yang dimiliki miring (*skew*).

Quantile 3 (Q3), merupakan data dengan urutan ke 75% dari total jumlah data.

Secara prosedur, quantile data dapat dicari dengan cara berikut:

Urutkan data dari kecil ke besar.

Bila jumlah data genap, maka Q2 dihitung dengan membagi 2 nilai yang ada di tengah

Pada contoh ini

3, 3, 5, 9, **12, 15,** 17, 21, 22, 24

Nilai Q1 adalah 5, nilai Q2 adalah (12+15)/2 = 13.5 dan nilai Q3 adalah 21

Bila jumlah data ganjil, maka Q2 merupakan nilai tengah dari data, Q1 diperoleh dengan membagi dua nilai yang berada di 25% dari jumlah data, demikian juga dengan Q3 diperoleh dengan membagi dua nilai yang berada di 75% dari jumlah data.

Pada contoh ini

3, 3, 5, 9, 12, **15,** 17, 21, 22, 24, 25

Nilai Q1 adalah (5+9)/2 =7, nilai Q2 adalah 15, dan nilai Q3 adalah (21+22)/2 = 21,5

R menggunakan nilai interpolasi untuk menghitung quantile (Hyndman and Fan, 1996). Quantile ke- i didefinisikan sebagai berikut:

$$Q[i](p) = (1-\gamma)x[j] + \gamma x[j+1]$$

dimana:

$1 \leq i \leq g$; $\frac{j-m}{n} \leq p \leq \frac{j-m+1}{n}$; $x[j]$ adalah data pada urutan ke- j , n adalah jumlah data, m adalah konstan yang ditentukan berdasarkan tipe quantile, dan γ adalah fungsi dari $j = \lfloor np + m \rfloor$; $g = np + m - j$

Pada contoh di atas, hasil yang didapatkan dari R sedikit berbeda.

```
contoh = c(3,3,5,9,12,15,17,21,22,24)
quantile(contoh)
```

```
  0%  25%  50%  75% 100%
   3    6   14   20   24
```

```
contoh = c(3,3,5,9,12,15,17,21,22,24,25)
quantile(contoh)
```

```
  0%  25%  50%  75% 100%
   3    7   15   22   25
```

1.4.4.3 Median

Median biasa digunakan untuk menyatakan nilai tengah (Q2) atau pusat data, terutama bila data tersebut miring (*skew*). Median juga tangguh (*robust*) terhadap adanya nilai pencilan (*outlier*).

1.4.4.4 Mean (Rerata)

Mean juga digunakan untuk menyatakan nilai tengah atau pusat data. Mean dirumuskan sebagai

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

Jumlah seluruh data dibagi dengan jumlah data (n). Mean tidak tangguh terhadap adanya pencilan.

Contoh: Andaikan kita ingin mengetahui rata-rata gaji dari pegawai di sebuah perusahaan yang gaji seluruh karyawan dalam (jutaan Rp) di tuliskan di bawah ini

4, 4, 5, 5, 7, 8, 10, 10, 12, 100

Pada data di atas, gaji karyawan dimulai dari 1 juta hingga 100 juta. Bila kita mengambil nilai rata-rata dari seluruh gaji kita dapatkan bahwa rata-rata gaji 16.5 juta Rupiah. Tentu nilai ini sangatlah besar. Seolah-olah perusahaan tersebut mengaji karyawan nya sangat tinggi. Padalah nilai rata-2 ini terlihat besar karena ada 1 gaji katakanlah pemilik yang nilainya jauh di atas para karyawan yang lain. Bila kita menghitung nilai rata-rata ini dengan menggunakan Median, maka median dari gaji ini adalah 7.5 juta Rupiah. Nilai median ini lebih mencerminkan rata-rata gaji karyawan dibandingkan dengan nilai mean nya.

1.4.4.5 Trimmed Mean

Untuk menghindari adanya data pencilan, pusat data dapat dihitung dengan menggunakan trimmed mean (mean yang dipotong). Nilai $100\alpha\%$ trimmed mean dapat dihitung dengan cara sebagai berikut:

$$\bar{x_\alpha} = \frac{x_{([n\alpha]+1)} + \ldots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

$[n\alpha]$ nilai integer terbesar yang kurang dari atau sama dengan $n\alpha$

Perintah R untuk menghitung median, mean dan trimmed mean adalah

```
contoh = c(4,4,5,5,7,8,10,10,12,100)
mean(contoh)
```

```
[1] 16
```

```
median(contoh)
```

```
[1] 7.5
```

```
mean(contoh, trim = 0.1)
```

```
[1] 7.6
```

Median, mean, trimmed mean adalah statistik yang biasa digunakan untuk mengukur pusat data. Untuk mengukur sebaran atau variasi data terhadap pusat dapat digunakan statistik berikut.

1.4.4.6 Range

Range data adalah besarnya perbedaan antara nilai tertinggi terhadap nilai terendah

$$Range = Max - Min$$

1.4.4.7 Interquartile Range (IQR)

Interquartile range adalah besarnya perbedaan antar quartile. IQR biasa digunakan untuk mengukur sebaran data, bila distribusi data miring (*skew*)

$$IQR = Quartile\ atas(Q3) - Quartile\ bawah(Q1)$$

1.4.4.8 Variance

Varians biasa digunakan untuk mengukur sebaran data terhadap mean bila distribusi data simetri.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

1.4.4.9 Standard Deviasi

Standard deviasi adalah akar dari variance, S

1.4.5 Summary (Ringkasan)

Summary statistik yang biasa ditampilkan oleh software statistik meliputi:

Five – Number Summary (lima ringkasan statistik): Max, Q3, Median, Q1, Min

Seven – Number Summary (7 Ringkasan Statistik): Max, Q3, Median, Q1, Min, Mean, Standard Deviasi

Perintah R untuk menghitung summary statistik adalah

```r
Q       =  quantile(close)
Q
```

```
  0%  25%  50%  75% 100%
1111 3754 5017 6189 7905
```

```r
Range   =  max(close) - min(close)
Range
```

```
[1] 6794
```

```r
Range   =  Q[5]-Q[1]
Range
```

```
100%
6794
```

```r
IQR     =  Q[4]-Q[2]
IQR
```

```
 75%
2434
```

```r
Mean    =  mean(close)
Mean
```

```
[1] 4833
```

```r
Med     =  median(close)
Med
```

```
[1] 5017
```

```r
Var     =  var(close)
Var
```

```
[1] 2793684
```

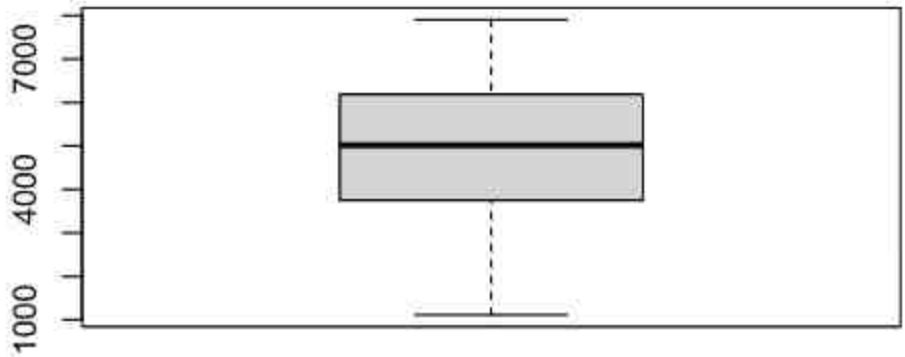
```
summary(close)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1111    3754    5017    4833    6189    7905
```

1.4.6 Box Plot

Boxplot merupakan grafik yang merepresentasikan lima ringkasan statistik (Tukey). Sebagai contoh kita akan menggambarkan boxplot dari JKSE-Close. Kotak (*box*) pada grafik ini merepresentasikan quartile 1 (Q1), median (Q2) dan quartile 3 (Q3), sedangkan tinggi kotak akan merepresentasikan IQR. Pada boxplot terdapat dua 'pagar' (*fence*), pagar atas dan pagar bawah.

```
boxplot(close)
```



Pagar atas dihitung dengan membandingkan antara nilai max yang terdapat pada data, dengan Q3 + 1,5 IQR.

Bila nilai max tersebut lebih kecil dari nilai Q3+1,5 IQR; maka panjang pagar atas akan sama dengan nilai max yang terdapat pada data. Nilai Q3 + 1,5 IQR biasanya digunakan sebagai indikator terjadinya pencilan. Bila terdapat nilai dalam data yang melebihi Q3 + 1,5 IQR

maka nilai tersebut dapat diindikasikan sebagai nilai pencilan. Pada contoh di atas tidak terdapat nilai pencilan.

Bila nilai max tersebut lebih besar dari nilai Q3+1,5 IQR; maka panjang pagar atas akan sama dengan nilai dari Q3 + 1,5 IQR.

$$P_a(x) = min(Q3 + 1,5 * IQR(x);\ max(x))$$

Pagar bawah dihitung dengan membandingkan antara nilai min yang terdapat pada data, dengan Q1 - 1,5 IQR.

Bila nilai min tersebut lebih besar dari nilai Q1-1,5 IQR; maka panjang pagar atas akan sama dengan nilai min yang terdapat pada data. Nilai Q1 - 1,5 IQR biasanya digunakan sebagai indikator terjadinya pencilan. Bila terdapat nilai dalam data yang kurang dari Q1 - 1,5 IQR maka nilai tersebut dapat diindikasikan sebagai nilai pencilan. Pada contoh di atas tidak terdapat nilai pencilan.

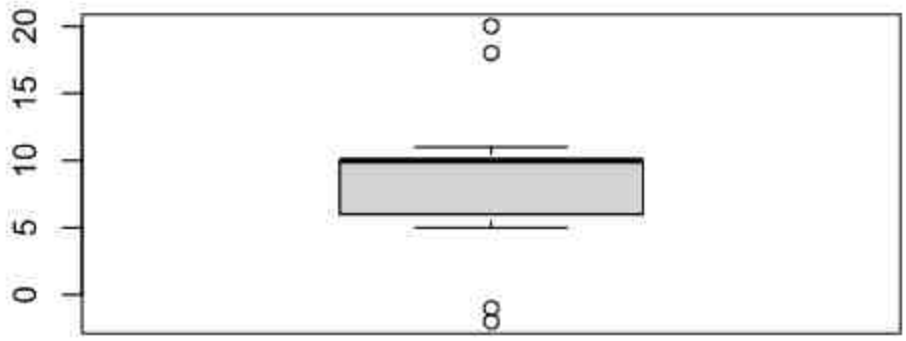
Bila nilai min tersebut lebih kecil dari nilai Q1-1,5 IQR; maka panjang pagar atas akan sama dengan nilai dari Q1 - 1,5 IQR.

$$P_b(x) = min(Q1 - 1,5 * IQR(x);\ min(x))$$

```
data_out = c(-2,-1,9,10,10,10,6,11,5,10,9,10,18,20)
summary(data_out)
```

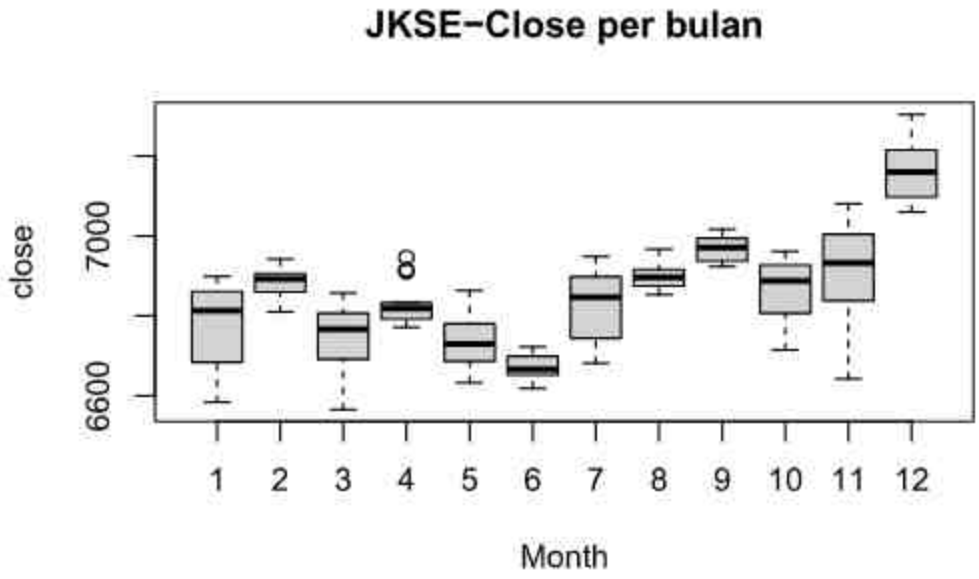
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -2.0     6.8    10.0     8.9    10.0    20.0
```

```
boxplot(data_out)
```



Pada boxplot kita dapat membandingkan perubahan data yang terjadi dari waktu ke waktu atau dari satu variabel ke variabel lain yang diukur. Sebagai contoh, kita akan melihat perubahan index penutupan pada Jakarta Stock Exchange.

```
library(lubridate)
date = df$Date
Month = month(date)
Year  = year(date)
NewData = data.frame(close,Month,Year)
NewData21 = NewData[which(NewData$Year == 2023),]
boxplot(close~Month, data = NewData21, main = 'JKSE-Close per bulan')
```



Bila dilihat dari mediannya maka terlihat nilai penutupan JKSE terendah terjadi di bulan Mei 2021. Antara bulan April hingga September 2021 harga penutupan saham bergerak di kisaran Rp 6000 – Rp 6100 an. Nilai penutupan saham ini mulai bergerak naik di bulan Oktober dan mencapai puncaknya di bulan November.

Selain itu, bila dilihat dari IQR nya (tinggi box), terlihat bahwa pada bulan Januari variasi dari nilai penutupan saham ini paling besar. Variasi nilai penutupan harga saham terkecil terlihat di bulan September. Namun di bulan itu terdapat nilai pencilan atas ataupun pencilan bawah. Grafik ini akan menceritakan lebih banyak hal lagi, bila kita mampu menghubungkannya dengan kondisi ekonomi dan sosial yang terjadi di Indonesia saat itu.

Pada bab ini kita telah mempelajari tentang data, tipe data dan cara mendiskripsikannya. Tipe data yang berbeda akan didiskripsikan dengan cara yang berbeda pula. Diskripsi data yang tepat akan membantu kita untuk menyingkapkan informasi yang tersembunyi di dalam data yang kita miliki.

1.5 Latihan

1. Selidikilah distribusi keselamatan penumpang terhadap Gender dan pelabuhan Embarkasi. Tariklah kesimpulan dari penyelelidikan yang anda lakukan.
2. Selidikilah distribusi dari JKSE open, high, low, Adj.close dan volume dengan menggunakan histogram, diagram batang dan daun, dan line plot. Tariklah kesimpulan dari penyelidikan yang anda lakukan.

3. Buatlah boxplot dari JKSE open, high, low, Adj.close dan volume per bulan. Tariklah kesimpulan dari penyelidikan yang anda lakukan.

2 Probabilitas

2.1 Memahami Keacakan (*Randomness*)

Menurut kamus besar Bahasa Indonesia, kata acak (random) diartikan sebagai tanpa pola. Keacakan ini sering kita jumpai dalam berbagai permainan, misalkan lembar dadu, permainan kartu, lotere, semua aplikasi permainan yang ada di smartphone. Permainan-permainan ini dianggap acak bila permainan ini adil (*fair*), yaitu, semua pemain tidak dapat menebak hasil luaran dari permainan ini hingga luaran itu terjadi.

Sebagai gambaran, misalkan kita melempar mata uang, gambaran adil di sini adalah kita tidak bisa menebak apakah yang akan jatuh di permukaan lantai adalah Gambar (*Head*) atau Angka (*Tail*), hingga mata uang tersebut benar-benar menunjukkan Gambar atau di Angka di atas permukaan lantai.

Hal lain yang perlu diperhatikan tentang keacakan adalah jika kita melakukan berulang-ulang, maka luaran yang diperoleh akan konsisten dan dapat ditebak kemungkinan terjadinya suatu kejadian (De Veaux *et al.* 2016).

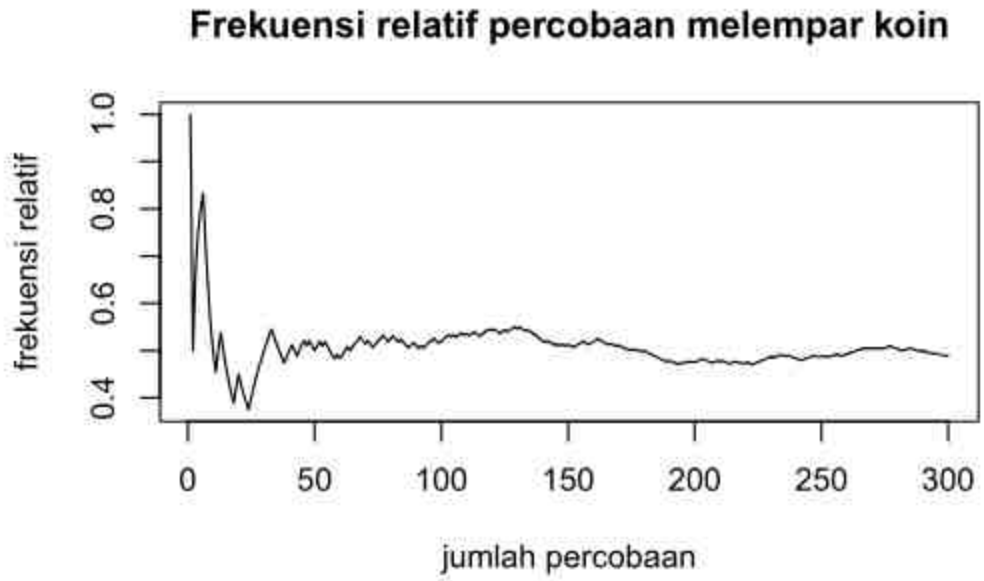
Misalkan kita melemparkan mata uang koin berulang-ulang, maka kita dapat memprediksi bahwa kemungkinan munculnya Gambar. Pada percobaan pertama, mungkin saja yang muncul adalah Gambar, sehingga kita mendapatkan frekuensi relative munculnya Gambar adalah 1. Kita ulang lagi percobaan ini. Percobaan kedua, kebetulan muncul lagi Gambar, maka frekuensi relative kemunculan Gambar adalah 2/2. Kita ulang lagi, pada percobaan ketiga muncul Angka, maka frekuensi relative kemunculan Gambar adalah $2/3 = 0{,}66$ (Tabel 2.1). Demikian seterusnya, bila kita melakukan percobaan ini berulang-ulang, hingga N yang cukup besar maka frekuensi relative dari munculnya Gambar adalah ½. Pada R-script di bawah ini menyimulasikan percobaan melempar koin sebesar 300 kali, terlihat setelah percobaan ini diulang berkali-kali maka frekuensi relative dari munculnya Gambar pada koin akan mendekati ½. Dari sini dapat kita simpulkan, percobaan melempar koin ini adalah peristiwa acak. Kita tidak dapat menebak munculnya Gambar atau Angka hingga dia terjadi, bila percobaan ini diulang-ulang kita mendapatkan pola bahwa frekuensi relative munculnya Gambar (atau Angka) adalah sama yaitu ½.

Tabel 2.1. Frekuensi relatif percobaan melempar koin

No	Kejadian yang muncul	Frekuensi relative
1	Gambar	$1/1 = 1$
2	Gambar	$2/2 = 1$
3	Angka	$2/3 = 0{,}666$
4	Angka	$2/4 = 0{,}5$
5	Gambar	$3/5 = 0{,}6$
6	Gambar	$4/6 = 0{,}666$
...		
N	Gambar	$144/300 = 0{,}48$

Probabilitas adalah ukuran yang digunakan untuk mengukur kemungkinan munculnya suatu kejadian. Dalam kasus melempar koin di atas, kita ingin mengetahui berapa kemungkinan muncul Gambar atau Angka. Berdasarkan percobaan di atas (secara empiris), nilai probabilitas dapat dinyatakan sebagai frekuensi relative munculnya suatu kejadian yang diharapkan terhadap semua kejadian yang mungkin terjadi. Pada kasus melempar koin, kejadian yang diharapkan adalah munculnya Gambar; sedangkan semua kejadian yang mungkin terjadi adalah {Gambar, Angka}. Oleh sebab itu secara empiris probabilitas munculnya Gambar adalah total munculnya Gambar dibagi dengan total percobaan.

```
N = 300
X = sample(2,N, replace = TRUE)-1

Freq = c()
for(i in 1:N)
  Freq[i] = sum(X[1:i])/i
plot(Freq, type = 'l', main = "Frekuensi relatif percobaan melempar koin",
     xlab = 'jumlah percobaan', ylab = 'frekuensi relatif')
```



Bila percobaan ini diulang-ulang sebanyak mungkin nilai probabilitasnya akan sama dengan jumlah anggota kejadian yang diinginkan {Gambar}, dibagi dengan jumlah anggota seluruh kejadian yang mungkin terjadi {Gambar, Angka}:

$$prob(Gambar) = \frac{|Gambar|}{|\{Gambar, Angka\}|} = \frac{1}{2}$$

Seluruh kejadian yang mungkin terjadi biasanya disebut sebagai ruang sample, sedangkan kejadian yang diharapan untuk muncul disebut sebagai *event*.

2.2 Ruang Sample (*Sample Space*) dan Kejadian (*Event*)

Dalam suatu percobaan, eksperimen akan menghasilkan luaran (*outcome*). Pada kejadian melempar koin maka luarannya adalah Gambar atau Angka. Pada kejadian melempar dadu maka luaran yang mungkin terjadi adalah mata dadu 1, 2,3,4,5 dan 6.

Semua luaran yang mungkin terjadi dan berkaitan degan suatu eksperimen atau kejadian disebut sebagai ruang sampel, atau ruang cuplikan (*sample space*). Ruang sampel biasa dinotasikan sebagai Ω, sedangkan anggota atau elemen dari ruang sampel dinotasikan dengan ω.

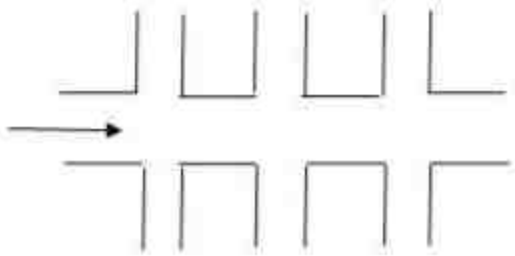
Ruang sampel seringkali dinyatakan dalam bentuk himpunan dengan anggota seluruh luaran (*outcome*) yang mungkin terjadi.

Pada kejadian melempar koin, ruang sampelnya: $\Omega = \{Gambar, Angka\}$

Pada kejadian melempar dadu, ruang sampelnya: $\Omega = \{1, 2, 3, 4, 5, 6\}$

Contoh 2.1.

Pada sebuah gudang terdapat tiga lorong. Sebuah forklift akan berjalan melewati tiga buah perempatan yang terdapat pada gudang tersebut. Tentukan ruang sampel bahwa dia akan berhenti (H) atau berjalan (J) disetiap perempatan. Lihat (Gambar 2.1)



Gambar 2.1. Ilustrasi lorong pada sebuah gudang

$\Omega = \{JJJ, JJH, JHJ, HJJ, HHJ, HJH, JHH, HHH\}$

Contoh 2.2

Jumlah barang dalam antrian yang terdapat pada jalur produksi dapat dimodelkan sebagai fenomena acak. Ruang sampel nya adalah $\Omega = \{0.1, 2, 3, ...\}$

yaitu semua bilangan bulat tak-negative. Pada contoh ini, ruang sampelnya tidak berhingga (*infinite*). Ruang sampel seperti ini disebut sebagai ruang sampel diskrit tak berhingga. Jika dalam antrian tersebut terdapat kapasitas maksimum, misalnya N, maka ruang sampelnya berubah menjadi: $\Omega = \{0, 1, 2, 3, ..., N\}$

Ruang sampel ini adalah ruang sampel yang berhingga (*finite*). Ruang sampel seperti ini disebut sebagai ruang sampel diskrit berhingga.

Contoh 2.3.

Kebisingan dalam suatu pabrik sering juga dimodelkan sebagai fenomena acak. Sebuah kondisi dikatakan sebagai bising, apabila suara yang terjadi pada tempat tersebut melebihi 85 DB (Keputusan Menteri Tenaga Kerja No. 13 Tahun 2011). Nilai kebisingan dapat diukur secara kontinu, ruang sampel dari kebisingan ini dapat dinyatakan sebagai $\Omega = \{t|t > 85\}$

Ruang sampel ini adalah ruang sample yang kontinu tak berhingga.

Contoh 2.4.

Ruang sampel untuk sebuah koin yang dilempar 1 kali: $\Omega = \{G, A\}$; G- Gambar, A- Angka

```
S = c("G","A")
S
```

```
[1] "G" "A"
```

Ruang sampel untuk sebuah koin yang dilempar 3 kali:

$$\Omega = \{GGG, GGA, GAG, AGG, GAA, GAG, AGA, AAA\}$$

```
s  = c(rep("G",3), rep("A",3))
samplespace = unique(replicate(100, (paste(sample(s,3), collapse=""))))
samplespace
```

```
[1] "GGA" "AGG" "AAG" "AGA" "GGG" "GAA" "AAA" "GAG"
```

Ruang sampel untuk dadu yang dilembar 1 kali: $\Omega = \{1, 2. 3, 4, 5. 6\}$

```
S = 1:6
S
```

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

Ruang sampel untuk dadu yang dilembar 2 kali:

$$\Omega = \{(1,1), (1,2), ..., (1,6), (2,1), (2,2), ..., (2,6), ..., (6,6)\}$$

Terdapat 36 luaran yang mungkin terjadi pada kasus pelemparan dadu 2 kali.

```
s = expand.grid(1:6,1:6)
t(s)  #transpose(s)
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
Var1    1    2    3    4    5    6    1    2    3     4     5     6     1     2
Var2    1    1    1    1    1    1    2    2    2     2     2     2     3     3
     [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26]
Var1     3     4     5     6     1     2     3     4     5     6     1     2
Var2     3     3     3     3     4     4     4     4     4     4     5     5
     [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36]
Var1     3     4     5     6     1     2     3     4     5     6
Var2     5     5     5     5     6     6     6     6     6     6
```

2.2.1 Mengambil Cuplikan secara Acak (*Random Sampling*)

Mencuplik secara acak (*random sampling*) adalah mengambil sebagian kecil bagian dari ruang sampel tanpa memperhatikan komposisi dari ruang sampel tesebut. Contoh yang seringkali digunakan untuk menggambarkan *random sampling* ini adalah mengambil bola yang berada dalam jambangan. Andaikan terdapat tiga bola bernomor di dalam jambangan tersebut, mengambil dua bola dari jambangan tersebut secara acak dapat dilakukan dengan dua cara.

- Cara pertama adalah setelah diambil bola tersebut dikembalikan (*sampling with replacement*).

  Sampling seperti ini disebut sebagai sampling yang *independent*. Hal ini karena,hasil sampling berikutnya tidak dipengaruhi oleh sampling yang dilakukan saat ini. Jumlah bola dalam jambangan selalu sama. kemungkinan untuk mendapatkan bola yang sama dengan sampling pertama tetap dimungkinkan.

- Cara kedua adalah setelah diambil bola tersebut tidak dikembalikan (*sampling without replacement*)

  Sampling ini adalah sampling yang *dependent*. Hasil sampling berikutnya tergantung pada bola yang diambil pada sampling pertama, karena tidak dikembalikan, berarti bola tersebut sudah tidak berada di dalam ruang sampel yang baru. Hal ini berarti pula, bola-bola pada sampling pertama tidak akan pernah muncul di sampling kedua.

Selain cara pengambilan bola dikembalikan atau tidak, hal lain yang biasa diperhatikan adalah urutan dari pengambilan cuplikan. Terdapat dua kemungkinan, yaitu

- Jika urutan diperhatikan, maka bola1 bernilai 1. dan bola kedua bernilai 2 memiliki arti yang berbeda jika bola 1 bernilai 2 dan bola kedua bernilai 1.
- Jika urutan tidak diperhatikan, maka pada contoh di atas yang dilihat hanyalah kedua bola tersebut bernilai 1 dan 2, urutan luaran dari kedua bola tersebut diabaikan.

Contoh 2.5.

Berikut adalah luaran dari semua kemungkinan bila 2 bola dicuplik dari jambangan yang berisi 3 bola.

Bila bola dikembalikan, dan urutan diperhatikan

$$\Omega = \{(1,1)(2,1)(3,1)(1,2)(2,2)(3,2)(1,3)(2,3)(3,3)\}$$

```
S = expand.grid(1:3,1:3)
t(S)    #transpose(s)
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
Var1    1    2    3    1    2    3    1    2    3
Var2    1    1    1    2    2    2    3    3    3
```

Bila bola tidak dikembalikan, dan urutan diperhatikan

$$\Omega = \{(1,2)(1,3)(2,1)(2,3)(2,2)(3,1)(3,2)\}$$

```
#install.packages("gtools")
library(gtools)
```

```
Warning: package 'gtools' was built under R version 4.1.3
```

```
S= permutations(3,2)
S
```

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

Bila bola tidak dikembalikan, dan urutan tidak diperhatikan

$$\Omega = \{(1,2)(1,3)(2,3)\}$$

```
S=combn(1:3,2)
S
```

```
     [,1] [,2] [,3]
[1,]    1    1    2
[2,]    2    3    3
```

Bila bola dikembalikan, dan urutan tidak diperhatikan

$$\Omega = \{(1,1)(1,2)(1,3)(2,2)(2,3)(3,3)\}$$

```
S = expand.grid(1:3,1:3)
S[S[,2]>= S[,1],]
```

```
  Var1 Var2
1    1    1
4    1    2
5    2    2
7    1    3
8    2    3
9    3    3
```

2.2.2 Kejadian (Event)

Dalam suatu peristiwa, tidak semua anggota dari ruang sampel akan muncul secara bersama-sama, bisa saja hanya sebagian dari anggota ruang sampel itu yang terjadi. Peristiwa yang muncul atau terjadi pada suatu ruang sampel ini disebut sebagai Kejadian (*Event*). Event seringkali dinyatakan sebagai himpunan bagian dari ruang sampel, setiap anggota dari Event ada di ruang sampel.

Misalnya dalam Contoh 2.1, kejadian bahwa forklift tersebut berhenti pada percempatan pertama, adalah himpunan bagian dari Ω yang dinyatakan sebagai:

$$E = \{HHH, HHJ, HJJ, HJH\}$$

Dalam Contoh 2.2, peristiwa bahwa orang yang berada dalam antrian tidak lebih dari 10 orang merupakan suatu Kejadian (*Event*) yang dapat dinyatakan sebagai

$$E = \{0.1, 2, 3, ..., 10\}$$

Contoh 2.6.

Peristiwa Forklift pada Contoh 2.1 dapat dianggap sebagai peristiwa melempar koin 3 kali, karena keduanya memiliki *outcome* yang sama yaitu 0 -"Berhenti" atau 1- "Berjalan" pada Contoh 2.1 dan 0- "Head" atau 1- "Tail" pada peristiwa melempar koin. Untuk itu ruang sampel pada peristiwa Forklift dan Event bahwa Forklift tersebut berhenti di Lorong pertama dapat dituliskan dengan R sebagai berikut

```
s = c("H","J")
#Ruang sampel dari forklift berjalan di gudang dengan 3 lorong
Forklift = expand.grid(s,s,s)

#Event forklift tersebut berhenti di Lorong pertama
Event = subset(Forklift, Forklift[,1]=="H")
Event
```

```
  Var1 Var2 Var3
1    H    H    H
3    H    J    H
5    H    H    J
7    H    J    J
```

Contoh 2.7

Beberapa perintah dalam R untuk menentukan event dari ruang sample.

```
#Ruang sample: melempar koin 2 kali, Event: Koin pertama yang muncul adalah Tail
s = c("H","T")
S = expand.grid(s,s)
E = subset(S, S[,1]=="T")
E
```

```
  Var1 Var2
2    T    H
4    T    T
```

```
#Ruang sample: melempar 3 dadu, Event: jumlah ketiga dadu lebih dari 16
s = 1:6
S = expand.grid(s,s,s)
SA = apply(S,1,sum)
E  = subset(S,SA>16)
E
```

```
    Var1 Var2 Var3
180    6    6    5
210    6    5    6
215    5    6    6
216    6    6    6
```

2.3 Teori Himpunan

Beberapa Kejadian (*Event*) dapat terjadi secara bersama-sama, beririsan, terjadi di Event1 tetapi tidak di Event2, dan sebagainya. Untuk menyatakan peristiwa seperti ini maka diperlukan teori himpunan sebagai berikut.

2.3.1 Himpunan Bagian (*Subset*)

Himpunan bagian dari sebuah Kejadian A adalah B, apabila semua Kejadian yang terjadi di B, terjadi di A, dinotasikan sebagai $B \subset A$

Contoh 2.8.

A = Kejadian bahwa forklift berhenti di perempatan pertama

B = Kejadian bahwa forklift selalu berhenti di ketiga Lorong

$A = \{HHH, HHJ, HJJ, HJH\}$

$B = \{HHH\}$

Pada R, fungsi **%in%** berguna untuk menguji apakah sebuah himpunan atau vektor merupakan himpunan bagian dari himpunan yang lain. Contoh ini dapat dituliskan dalam R sebagai berikut:

```
A = c("HHH", "HHJ", "HJJ", "HJH")
B = c("HHH")
B %in% A
```

```
[1] TRUE
```

Selain itu dapat pula digunakan fungsi **all** dan **subset**, untuk menguji apakah seluruh anggota himpunan (vector) B berada di A.

```
all(B%in%A)
```

```
[1] TRUE
```

2.3.2 Gabungan (Union)

Gabungan (Union) dari dua Kejadian A dan B adalah C. Kejadian C adalah Kejadian yang terjadi karena A terjadi atau B terjadi atau kedua-duanya terjadi.

$$C = A \cup B$$

Contoh 2.9.

A = Kejadian bahwa forklift berhenti pada perempatan pertama

B = Kejadian bahwa forklift berhenti pada perempatan ketiga

$A = \{HHH, HHJ, HJJ, HJH\}$

$B = \{HHH, HJH, JJH, JHH\}$

$C = A \cup B = \{HHH, HHJ, HJJ, HJH, JJH, JHH\}$

Pada R fungsi **union** digunakan untuk menyatakan gabungan antara dua himpunan A, dan B

```
A = c("HHH", "HHJ", "HJJ", "HJH")
B = c("HHH", "HJH", "JJH", "JHH")
C = union(A,B)
C
```

```
[1] "HHH" "HHJ" "HJJ" "HJH" "JJH" "JHH"
```

2.3.3 Irisan (*Intersection*)

Irisan (Intersection) dari dua Kejadian A dan B adalah D. Kejadian D adalah Kejadian yang terjadi baik di A maupun di B. $D = A \cap B$

Contoh 2.10.

Dari Contoh 2.9 di atas, kejadian bahwa forklift tersebut berhenti di perempatan pertama dan juga di perempatan ketiga adalah irisan dari Kejadian A dan Kejadian B, dan dapat dinyatakan sebagai berikut:

$D = A \cap B = \{HHH, HJH\}$

Pada R fungsi **intersect** digunakan untuk menyakatan irisan antara dua himpunan A, dan B

```
D = intersect(A,B)
D
```

```
[1] "HHH" "HJH"
```

2.3.4 Perbedaan (*Difference*)

Perbedaan (*Difference*) antara himpunan A dan B, dinotasikan sebagai $A \backslash B$, adalah himpunan yang merupakan anggota A tetapi bukan anggota B.

Contoh 2.11

Dari Contoh 2.9 di atas, kejadian forklift berhenti dilorong pertama dan kedua, namun berjalan di lorong ketiga (HHJ), dan kejadian forklift berjalan di lorong kedua dan ketiga, namun berhenti di lorong pertama(HJJ) adalah pembeda antara himpunan A dan B.

Pada R fungsi **setdiff** digunakan untuk menyakatan *difference* antara dua himpunan A, dan B

```
E = setdiff(A,B)
E
```

```
[1] "HHJ" "HJJ"
```

```
F = setdiff(B,A)
F
```

```
[1] "JJH" "JHH"
```

2.3.5 Komplemen

Komplemen suatu Kejadian A disebut dengan A^c adalah semua anggota ruang sampel yang tidak terjadi di A . Dari Contoh 4, A^c adalah kejadian dimana forklift tidak berhenti di perempatan pertama.

$A^c = \{JJJ, JJH, JHH, JHJ\}$

Pada R komplemen dari suatau himpunan dapat dinyatakan sebagai setdiff himpuanan tersebut terhadap ruang sampelnya.

```
S = c("JJJ", "JJH", "JHJ", "HJJ","HHJ", "HJH", "JHH", "HHH")
G = setdiff(S,A)
G
```

```
[1] "JJJ" "JJH" "JHJ" "JHH"
```

2.3.6 Himpunan Kosong

Himpunan kosong dinotasikan dengan himpunan tanpa anggota {} atau ϕ. Himpunan kosong menyatakan suatu kejadian yang tidak memiliki luaran (*outcome*).

Contoh 2.12.

Misalkan E adalah kejadian dimana forklift berjalan tanpa hambatan, A adalah kejadian dimana forklift selalu berhenti di perempatan pertama. Kedua kejadian ini tidak akan pernah terjadi bersama-sama. Irisan kedua kejadian ini adalah himpunan kosong. Kejadian seperti ini disebut sebagai saling asing (*disjoint*).

$$A \cap E = \phi$$

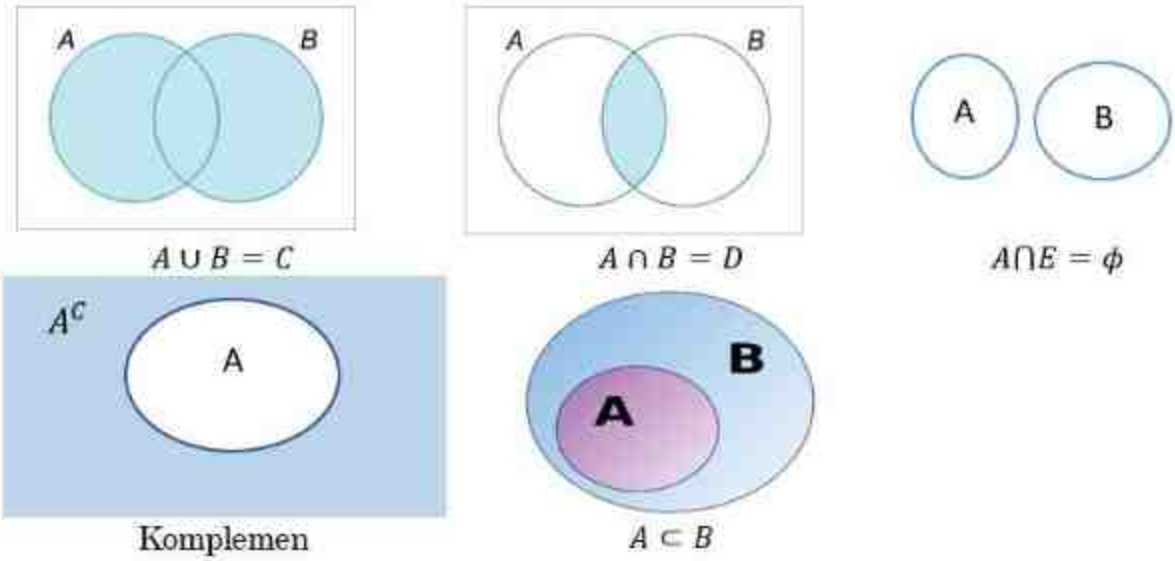
cobalah

```
setdiff(A,S) #anda akan mendapatkan himpunan kosong.
```

```
character(0)
```

2.3.7 Diagram Venn

Diagram Venn adalah gambar yang biasa digunakan untuk menyatakan hubungan antara satu Kejadian dengan Kejadian yang lain. Biasanya suatu Kejadian dinyatakan dengan lingkaran dan ruang sampel digambarkan dengan kotak yang melingkupi lingkaran tersebut. Beberapa contoh Diagram Venn yang menyatakan hubungan Union, Intersection, Komplemen dan himpunan kosong (lihat Gambar 2.2)



Gambar 2.2. Diagram Venn

2.3.8 Hukum dalam Teori Himpunan

1. Hukum Komutatif (Cummutating Law)

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

2. Hukum Asosiatif (Associative Law)

$$(A \cup B) \cap C = A \cup (B \cap C)$$

$$(A \cap B) \cup C = A \cap (B \cup C)$$

3. Hukum Distributif (Distributive Law)

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

2.4 Ukuran Probabilitas (*Probability Measure*)

Ukuran probabilitas (*probability measure*) pada Ω adalah fungsi p pada himpunan bagian (Event) Ω ke bilangan real:

$$p : A \rightarrow R, A \subseteq \Omega$$

dan memenuhi aksioma sebagai berikut:

1. $p(\Omega) = 1$
2. Jika $A \subset \Omega$, maka $p(A) \geq 0$

Jika A_1 dan A_2 saling asing (*disjoint*) maka

$$p(A_1 \cup A_2) = p(A_1) + p(A_2)$$

Secara umum jika $A_1, A_2, ..., A_n$ satu dengan yang lain saling asing, maka

$$p(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} p(A_i)$$

Catatan: Aksioma adalah anggapan yang dianggap logis dan benar.

Sifat A: $p(A^c) = 1 - p(A)$

Bukti:

$A \cup A^c = \Omega \Rightarrow p(A \cup A^c) = p(\Omega) = 1$ (Ak.1)

$A \cap A^c = \phi \Rightarrow p(A) + p(A^c) = 1$
$\Rightarrow p(A^c) = 1 - p(A)$■

Sifat B: $p(\phi) = 0$

Bukti:

$\Omega^c = \phi$ (karena Sifat A)

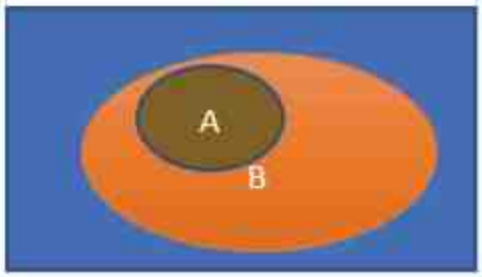
$p(\Omega^c) = 1 - p(\Omega) = 1 - 1 = 0$■

Probabilitas bahwa tidak ada *outcome* sama sekali adalah nol.

Sifat C: Jika $A \subseteq B$, maka $p(A) \leq p(B)$

Bukti:

Perhatikan gambar di bawah ini



$B = A \cup (B \cap A^c)$ merupakan gabungan dari dua himpunan saling asing.

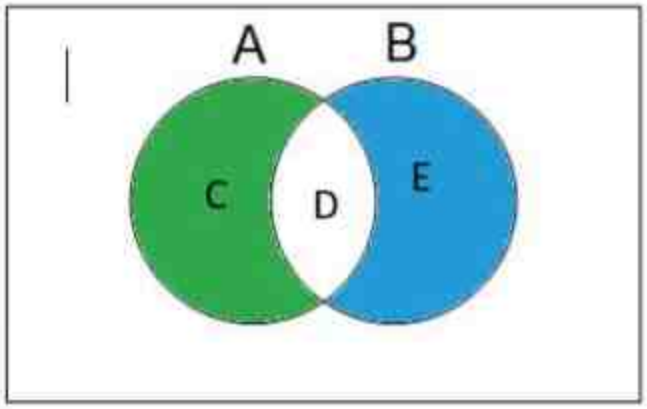
$p(B) = p(A) + p(B \cap A^c)$ (Ak.3)

$p(A) = p(B) - p(B \cap A^c) \leq p(B)$ (Ak. 3) ■

Sifat D: (Hukum Penjumlaha)

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Bukti: Perhatikan gambar berikut



Perhatikan bahwa

$C = A \cap B^c$; $D = A \cap B$; $E = A^c \cap B$.

Ketiganya saling asing, karena itu berdasarkan hukum penjumlahan maka:

$p(A \cup B) = p(C) + p(D) + p(E)$

Perhatikan:

$A = C \cup D$; C dan D saling asing maka, $p(A) = p(C) + p(D)$

$B = D \cup E$; D dan E saling asing maka, $p(B) = p(D) + p(E)$

$$p(A) + p(B) = p(C) + p(D) + p(D) + p(E)$$

$$p(A) + p(B) = p(A \cup B) + p(D)$$

$$\Rightarrow p(A \cup B) = p(A) + p(B) - p(D)$$

$$\Rightarrow p(A \cup B) = p(A) + p(B) - p(A \cap B) \blacksquare$$

2.4.1 Menentukan Nilai Probabilitas secara Empiris

Ada dua pendekatan dalam menentukan nilai probabilitas, yaitu secara empiris dan secara model probabilitas. Pada bagian ini, akan dibahas penentuan nilai probabilitas secara empiris. Seperti telah dijelaskan di atas, bahwa dalam pendekatan empiris, nilai probabilitas merupakan frekuensi relative suatu Event terjadi. Untuk itu, dalam pendekatan ini, bagaimana suatu Event terjadi sangatlah penting. Selain itu, karena Event adalah himpunan bagian dari ruang sampel, maka jumlah anggota dari suatu Event jugalah penting untuk diketahui.

Untuk menghitung nilai probabilitas secara empiris dapat dilakukan melalui metode pencacahan (*counting methods*)

Andaikan ruang sampel yang terdiri dari n anggota, dan masing-masing anggota memiliki kemungkinan untuk terjadi yang sama yaitu $\frac{1}{n}$, maka kemungkinan sebuah Kejadian (*Event*) A terjadi adalah

$$p(A) = \frac{\text{jumlah cara } A \text{ dapat terjadi}}{n}$$

Cara penghitungan nilai probabilitas seperti pada persamaan di atas, disebut sebagai *equally likely model* (ELM)

Contoh 2.13:

Pada peristiwa mengocok sebuah dadu, probabilitas munculnya dadu genap dapat dijabarkan sebagai berikut. Ruang sampel dari peristiwa ini adalah $\Omega = \{1, 2, 3, 4, 5, 6\}$, jumlah anggota dari ruang sampel ini adalah $n = 6$, dan masing-masing diasumsikan memiliki kemungkinan yang sama untuk terjadi. Kejadian munculnya mata dadu genap $A = \{2, 4, 6\}$, jumlah cara A dapat terjadi adalah 3 , maka

$$p(A) = \frac{3}{6}$$

```
S=1:6
A=subset(S,S%%2==0)
ProbA = length(A)/length(S)
ProbA
```

```
[1] 0.5
```

Contoh 2.14:

Dalam sebuah permainan dengan dua buah dadu, seorang pemain akan memenangkan hadiah bila jumlah dadu yang muncul di lebih besar atau sama dengan 8. Berapa probabilitas bahwa pemain tersebut akan mendapatkan hadiah?

Ruang sampel dari peristiwa ini adalah

Ω = { (1,1), (1,2), (1,3), (1,4), (1,5), (1,6),

(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),

(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),

(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),

(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),

(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)}

Jumlah mata dadu yang muncul di atas 8 terjadi pada

A = {(2,6), (3,5), (3,6), (4,4),(4,5), (4,6), (5,3), (5,4), (5,5), (5,6), (6,2),(6,3), (6,4), (6,5), (6,6)}

$$p(A) = \frac{15}{36}$$

```
s = 1:6
S = expand.grid(s,s)
E = subset(S, apply(S,1,sum) >= 8)
Prob_E = nrow(E)/nrow(S)
Prob_E
```

```
[1] 0.4166667
```

Metode pencacahan ini sangat tergantung pada penghitungan jumlah luaran (*outcome*) dari suatu ruang sampel beserta dengan jumlah luaran dari Kejadian (*Event*) yang akan dihitung. Terdapat beberapa cara untuk mengetahui jumlah luaran tersebut, di antaranya adalah:

2.4.2 Asas Pelipatan (*Multiplication Principle*)

Jika suatu percobaan memiliki M luaran, dan percobaan lain memiliki N luaran, maka akan ada $M \times N$ luaran yang mungkin terjadi bila kedua percobaan tersebut dilaksanakan bersama-sama.

Percobaan melempar dua dadu pada Contoh 8, merupakan contoh dari asas pelipatan. Pelemparan dadu 1 akan memiliki 6 luaran, dan pelemparan dadu 2 juga memiliki 6 luaran. Ruang sampel pelemparan kedua dadu ini akan memiliki 6 x 6 = 36 luaran.

Secara umum, bila terdapat K percobaan, percobaan pertama memiliki M_1 luaran, percobaan kedua memiliki M_2 luaran, ..., percobaan ke - K memiliki M_K luaran. Secara keseluruhan akan terdapat $M_1 \times M_2 \times ... \times M_K$ luaran yang mungkin terjadi untuk K percobaan tersebut.

2.4.3 Permutasi dan Kombinasi

Permutasi adalah suatu susunan yang terbentuk dari obyek-obyek yang dicuplik sebagian atau seluruhnya dari suatu percobaan dengan memperhatikan urutannya.

Andaikan $A = \{a_1, a_2, ..., a_n\}$ adalah himpunan dari suatu bola bernomor yang berada di dalam keranjang. Pilih r buah anggota A dan urutan dari pilihan ini diperhatikan. Misalkan $r = 2$, maka $\{a_1, a_2\}$ dan $\{a_2, a_1\}$ dianggap sebagai pilihan yang berbeda. Ada berapa cara hal ini dapat dilakukan?

Pada masalah pengambilan bola bernomor di dalam keranjang, ada dua hal kemungkinan hal ini dapat dilakukan:

- Menarik cuplikan **tanpa dikembalikan**: Kasus ini merupakan kasus ***dependent***. Hal ini dikarenakan hasil dari pengambilan Bola II tergantung dari Bola I yang terpilih, demikian juga dengan hasil dari Bola III tergantung dari luaran pengambilan Bola II dan Bola I, demikian dan seterusnya. Ruang sampel dari kasus ini berkurang setiap kali pengambilan cuplikan.

  Bola I dapat dipilih dengan n cara

  Bola II dapat dipilih dengan $n - 1$ cara

  Bola III dapat dipilh dengan $n - 2$ cara, dan seterusnya hingga

  Bola ke - r hanya dapat dipilih dengan $n - r + 1$ cara

  Secara keseluruhan akan terdapat $n \times (n-1) \times (n-2) \times ..., \times (n - r_1)$ cara.

- Menarik cuplikan **dengan dikembalikan**: Kasus ini merupakan kasus ***independent***. Bola yang selalu dikembalikan setelah diambil, tidak akan mempengaruhi luaran dari pengambilan Bola selanjutnya. Ruang sample dari kasus ini selalu sama.

  Bola I dapat dipilih dengan n cara

  Bola II dapat dipilih dengan n cara

  Bola III dapat dipilh dengan n cara, dan seterusnya hingga

  Bola ke - r hanya dapat dipilih dengan n cara

  Secara keseluruhan akan terdapat $n \times n \times n \times, ..., \times n = n^r$ cara.

Pada kasus ini permutasi adalah peristiwa pengembalian bola tanpa dikembalikan. Jika terdapat n pilihan, dan hanya boleh r pilihan saja yang boleh dipilih, maka permutasi pemilihan sebanyak r dan n jumlah pilihan yang tersedia dapat dirumuskan sebagai

$$P_r^n = n \times (n-1) \times (n-2) \times ... \times (n-r-1)$$

Jika $n! = n \times (n-1) \times (n-2) \times ... \times 1$, maka permutasi dapat dirumuskan sebagai berikut:

$$P_r^n = \frac{n!}{(n-r)!} = \frac{n \times (n-1) \times (n-2) \times ... \times 1}{(n-r) \times (n-r-1) \times ... \times 1}$$

$$= n \times (n-1) \times (n-2) \times ... \times (n-r+1)$$

Contoh 2.15:

Anda ingin membentuk team untuk mengikuti lomba yang terdiri dari 3 orang. Ketiga orang itu akan berperan sebagai ketua, wakil dan anggota. Ada 10 siswa yang potensial untuk terpilih. Ada berapa cara team lomba ini akan terbentuk?

Dalam hal ini, karena peran dari sangat diperhatikan, maka urutan dari penyusunan team ini akan penting. Misalkan di antara 10 siswa tersebut terdapat siswa bernama Ani, Budi dan Cahyo, maka team dengan urutuan Ani, Budi dan Cahyo sebagai ketua, wakil dan anggota, akan berbeda bila team itu terdiri dari Budi, Ani dan Cahyo sebagai ketua, wakil dan anggota.

Untuk itu, dalam kasus ini jumlah cara pembentukan team lomba dapat dihitung dengan permutasi

$$P_3^{10} = \frac{10!}{7!} = 10 \times 9 \times 8 = 720$$

```
#install.packages(gtools)
library(gtools)
Per3_10 = nrow(permutations(10,3))
Per3_10
```

```
[1] 720
```

Kombinasi adalah cara penyusunan sebagian atau seluruh obyek tanpa memperhatikan urutanya. Pada contoh pengambilan bola di atas, pengambilan bola dengan urutan $\{a_1, a_2\}$ dianggap sama dengan pengambilan bola dengan urutan $\{a_2, a_1\}$.

Jika r obyek diciuplik dari suatu himpunan n obyek tanpa pengembalian dan tanpa memperhatikan urutan, maka cara obyek ini dapat disusun merupakan kombinasi yang dirumuskan sebagai berikut:

$$C_r^n - \frac{n!}{(n-r)!r!}$$

Contoh 2.16

Pada pembentukan team di Contoh 2.15. bila dalam lomba ini tidak ditentukan peran sebagai ketua, wakil dan anggota. maka sebuah team yang beranggotakan Ani, Budi dan Cahyo, akan dianggap sama bila anggota team ditulis sebagai Budi, Ani dan Cahyo.

Perhatikan: Pembentukan nama anggota bila urutan diperhatikan ada 6 ($3 \times 2 \times 1 = 3!$) cara, yaitu: Ani, Budi, Cahyo; Ani, Cahyo, Budi; Budi, Ani, Cahyo; Budi, Cahyo, Ani; Cahyo, Budi, Ani; Cahyo, Ani, Budi.

Secara keseluruhan ada 720 cara untuk membentuk team beranggotakan 3 orang dari 10 orang dengan urutan diperhatikan (Contoh 2.15). Jika urutan ini sekarang tidak diperhatikan, maka dari 720 cara tadi akan terdapat 6 cara yang dianggap sama (berulang). Untuk itu jumlah cara pembentukan team bila urutan tidak diperhatikan akan sama dengan $720/6 = 120$ cara saja. Perhitungan ini dapat dituliskan secara kombinasi:

$$C_3^{10} = \frac{10!}{7!3!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120$$

Catatan:

Bilangan C_r^n disebut sebagai koefisien binomial yang terjadi pada ekspansi

$$(a+b)^n = \sum_{k=0}^{n} C_k^n a^k b^{n-k}$$

```
Comb3_10 = nrow(combinations(10,3))
Comb3_10
```

```
[1] 120
```

Contoh 2.17: Masalah ulang tahun (*The Birthday Problem*)

Andaikan terdapat n orang dalam sebuah ruangan. Setiap orang mengumumkan hari ulang tahunnya secara bergiliran. Berapakah probabilitas sekurang-kurangnya ada satu orang yang berulangtahun sama?

Jawab:

A : Event sekurang-kurangnya ada satu pasang orang yang berulang tahun sama

A^c: Event tidak ada satupun yang berulang tahun sama

S: Semua kemungkinan hari ulang tahun dari n orang yang ada di ruangan itu.

Jika kita mengabaikan tahun kabisat, maka orang pertama memiliki kemungkinan 365 hari ulang tahun, demikian juga dengan orang kedua hingga orang ke- n. Total jumlah hari urutan ulang tahun dari n orang adalah $\#(S) = 365^n$

Jika kita menghitung langsung A, akan sangat sulit, lebih mudah bila kita menghitung A^c

Bila tidak ada satupun orang dalam ruangan itu yang memiliki ulang tahun yang sama maka orang pertama akan memiliki kemungkinan 365 hari ulang tahun, orang kedua akan memiliki kemungkinan 364 hari ulang tahun, ... Orang ke- n akan memiliki kemungkinan $(365-n+1)$ hari ulang tahun. Berdasarkan prinsip multiplikasi maka $\#(A^c) = 365 \times 364 \times ... \times (365-n+1)$

Dengan demikian probabilitas A terjadi dapat dihitung dengan menggunakan sifat

$$p(A) = 1 - p(A^c) = 1 - \frac{\#(A^c)}{\#(S)}$$

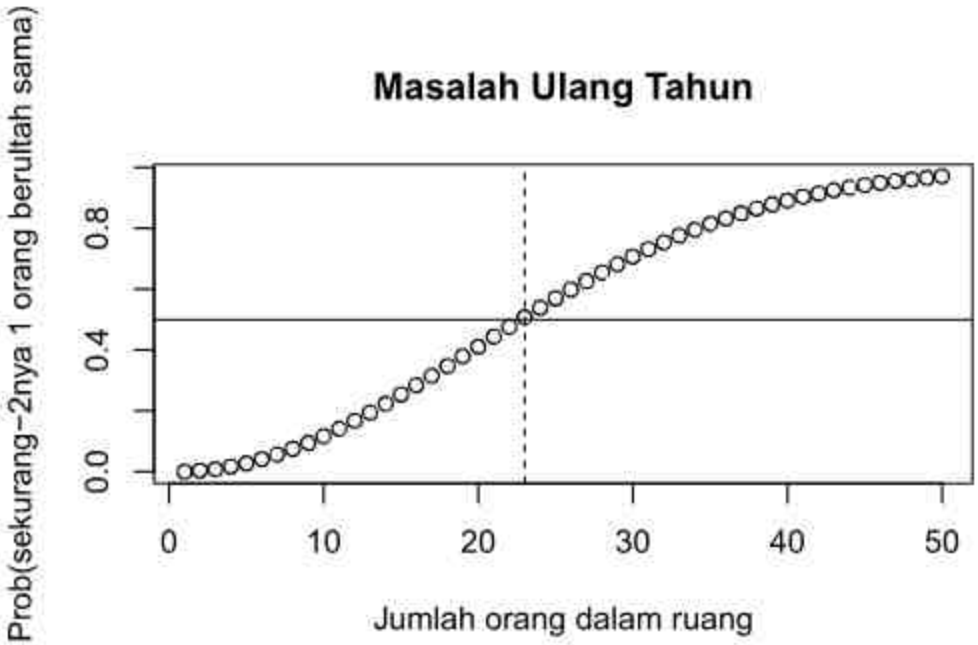
$$p(A) = 1 - \frac{365 \times 364 \times ... \times (365-n+1)}{365^n}$$

$$p(A) = 1 - \frac{364}{365} \times \frac{363}{365} \times ... \times \frac{365-n+1}{365}$$

Bertanyaan berikut dari masalah ini adalah berapa orang harus berada dalam satu ruangan sedemikian hingga probabilitas bahwa sekurang-kurangnya sepasang orang memiliki hari ulang tahun yang sama.

Bila jumlah orang dalam ruangan itu hanya 1 saja, tentu saja tidak ada sepasang orang yang memiliki ulang tahun yang sama, karena itu jika $n = 1$, maka $p(A) = 0$. Dengan mengganti-ganti nilai n, akan didapat $p(A)$ mendekati 0,5 bila $n = 23$. Perhitungan ini dapat dilakukan dengan menggunakan R sebagai berikut:

```
#fungsi untuk menghitung probabilitas ulang tahun di hari yang sama
p_ultah = function(n)
{ if(n == 1) pac = 1
  if(n >= 2)
  { pac = 1
    for(i in 2:n)
      pac = pac*(365-i+1)/365
  }
  p_ultah = 1-pac
  return(p_ultah)
}

#Menghitung semua probabilitas ulang tahun
n = 50  #jumlah orang
g = c()
for(i in 1:n)
  g= append(g, p_ultah(i))

#Plot
plot(1:50, g,
xlab = "Jumlah orang dalam ruang",
ylab = "Prob(sekurang-2nya 1 orang berultah sama)",
main = "Masalah Ulang Tahun")
abline (h = 0.5)
abline (v = 23, lty = 2) # dashed line
```



2.5 Kemungkinan Maksimum (*Maximum Likelihood*)

Kemungkinan maksimum adalah cara/aturan untuk memilih nilai dari n obyek keseluruhan yang membuat keluaran (*outcome*) yang diamati paling mungkin terjadi.

Contoh 2.18

Andaikan terdapat 10 ikan yang berada di sebuah sungai, ditangkap, diberi tanda lalu dilepaskan. Pada suatu waktu tertentu 20 ikan ditangkap, dan ditemukan bahwa 4 dari mereka adalah yang pernah diberi tanda. Berapa besar populasi dari ikan yang berada di sungai tersebut?

Jawab:

Asumsikan terdapat n ekor ikan dalam populasi, darinya 10 ekor diberi tanda. Jika keduapuluh ikan yang tertangkap tersebut, secara keseluruhan dapat ditangkap dengan C_{20}^{n} cara, tiap cara memiliki kemungkinan yang sama untuk terjadi maka probabilitas 4 dari mereka adalah ikan yang bertanda adalah

$$\frac{C_4^{10} C_{20-4}^{n-10}}{C_{20}^{n}}$$

Pada contoh ini jelas n tidak dapat dengan tepat ditentukan dari informasi yang ada, namun nilai ini dapat diperkirakan dengan menggunakan pendekatan ***maximum likelihood***.

Untuk menemukan estimasi yang memiliki kemungkinan maksimum, kita mengubah contoh di atas menjadi lebih umum. Andaikan pada umumnya t ikan diberi tanda, kemudian dari seluruh populasi ikan n ekor ikan tertangkap, diantaranya r ikan yang bertanda tertangkap kembali.

Estimasi nilai n, dengan menggunakan maximum likelihood adalah

$$L_n = \frac{C_r^t C_{m-r}^{n-t}}{C_m^n}$$

Untuk menemukan nilai n yang memaksimumkan L_n, perhatikan perbandingan dari suku-suku berurutan. Setelah beberapa langkah penyederhanaan secara aljabar, didapat:

$$\frac{L_n}{L_{n-1}} = \frac{(n-t)(n-m)}{n(n-t-m+r)}$$

Nilai perbandingan ini lebih besar dari satu, yaitu nilai L_n akan bertambah jika

$$(n-t)(n-m) > n(n-t-m+r)$$

$$n^2 - nt - nm + mt > n^2 - nt - nm + nr$$

$$mt > nr$$

$$n < \frac{mt}{r}$$

Jadi nilai L_n akan bertambah jika $n < \frac{mt}{r}$ dan nilai L_n akan berkurang jika $n > \frac{mt}{r}$. Hal ini berarti nilai n yang akan memaksimumkan L_n adalah bilangan bulat terbesar yang tidak melebihi $\frac{mt}{r}$

Pada kasus ikan di atas, maka jumlah populasi ikan di sungai tersebut $n = \frac{20 \times 10}{4} = 50$

(Dari jumlah populasi ikan, 10 diberitanda, diambil 20 ikan diantaranya ditemukan 4 ikan bertanda)

$$p(A) = \frac{10}{50} = \frac{4}{20} = \frac{1}{5}$$

2.6 Probabilitas Bersyarat (*Conditional Probability*)

Probabilitas bersyarat adalah nilai kemungkinan sebuah event akan terjadi, jika event yang lain sudah terjadi.

Contoh 2.19

Andaikan 2 buah dadu dilemparkan dan andaikan pula ke-36 luaran (*outcome*) yang mungkin terjadi memiliki kemungkinan terjadi yang sama yaitu 1/36.

Andaikan diamati bahwa dadu pertama yang muncul adalah 3; dengan diketahui informasi ini, berapakah probabilitas bahwa jumlah kedua dadu tersebut sama dengan 8.

Jawab:

Jika diberikan syarat bahwa luaran pertama dari dadu tersebut adalah 3,

Luaran yang mungkin muncul adalah (3,1) (3,2) (3,3) (3,4) (3,5) dan (3,6)

Dari sample space ini luaran yang jumlahnya 8 adalah (3,5)

maka probabilitas bahwa jumlah luaran tersebut 8 dengan syarat luaran pertamanya 3, adalah 1/6.

Catatan

Pada probabilitas bersyarat ini kita membatasi ruang sampel. Bila tanpa syarat, jumlah angora dari ruang sampel dari percobaan melempar dua dadu ini adalah 36. Dengan adanya syarat bahwa lemparan pertama adalah 3, ruang sampel ini menjadi lebih sempit, dan anggota ruang sampel ini menjadi 6 saja.

Notasi:

E : Kejadian bahwa kedua dadu adalah 8

F: Kejadian bahwa luaran dari lemparan dadu pertama adalah 3

Maka probabilitas bersyarat dinotasikan sebagai

$$p(E|F)$$

Secara umum:

Jika kejadian F terjadi, maka agar kejadian E terjadi, maka kedua kejadian tersebut harus terjadi secara Bersama-sama. Hal ini berarti irisan E dan F bukanlah himpunan kosong $E \cap F = EF \neq \phi$

Setelah kejadian F terjadi, berarti ruang sampel dari percobaan ini tereduksi. Hal ini berarti probabilitas bahwa kejadian EF terjadi akan sama dengan probabilitas EF relative terhadap probabiltias F terjadi

Definisi:

Jika $p(F) > 0$, maka $p(E|F) = \frac{p(EF)}{p(F)}$

Contoh 2.20

Terapi menggunakan obat herbal untuk penderita penyumbatan jantung, tetapi terdapat risiko keracunan dari pengobatan herbal tersebut. Risiko keracunan ini merupakan suatu akibat sampingan yang serius dan sulit untuk didiagnosis. Untuk memperbaiki peluang suatu diagnosis yang benar, konsentrasi daun-daunan dalam darah akan diukur. Beller dan kawan-kawan (1971) melakukan studi hubungan antara konsentrasi daun-daunan dalam darah dan keracunan daun-daunan dalam 135 penderita. Di dapat tabel sebagai berikut

	D+	D-	Total
T+	25	14	39
T-	18	78	96
Total	43	92	135

Notasi

T+: Konsentrasi darah yang tinggi (Tes positif)

T-: Konsentrasi darah yang rendah (Tes negatif)

D+: Keracunan

D-: Tidak keracunan

Jika seorang dokter tahu bahwa tes tersebut positf, berapakah probabilitas bahwa pasien tersebut akan keracunan bila konsentrasi darahnya tinggi.

Jawab:

Asumsikan bahwa pasien diambil secara acak (ada 135 pasien). Tabel diubah dalam bentuk frekuensi relative (dalam perbandingan dengan ukuran sampel 135), adalah

Tabel 2.3. Probabilitas pasien akan keracunan bila konsentrasi darahnya tinggi.

	D+	D-	Total
T+	0,185	0,104	0,289
T-	0,133	0,578	0,711
Total	0,318	0,682	1

$$p(T^+) = 0,289; p(D^+) = 0,318$$

$$p(D^+|T^+) = \frac{D^+ \cap T^+}{p(T^+)} = \frac{0,185}{0,289} = 0,640$$

Probabilitas pasien tidak keracunan bila konsentrasi darahnya rendah.

$$p(D^-|T^-) = \frac{D^- \cap T^-}{p(T^-)} = \frac{0,578}{0,711} = 0,848$$

Probabilitas pasien tidak keracunan bila konsentrasi darahnya tinggi.

$$p(D^-|T^+) = \frac{D^- \cap T^+}{p(T^+)} = \frac{0,104}{0,289} = 0,360$$

Probabilitas pasien keracunan bila konsentrasi darahnya rendah

$$p(D^+|T^-) = \frac{D^+ \cap T^-}{p(T^-)} = \frac{0,133}{0,711} = 0,187$$

2.6.1 Hukum Kelipatan (*Multiplication Law*)

Andaikan A dan B adalah dua kejadian dan asumsikan $p(B) \neq 0$, maka

$$p(A \cap B) = p(A|B)p(B)$$

$$p(B \cap A) = p(B|A)p(A)$$

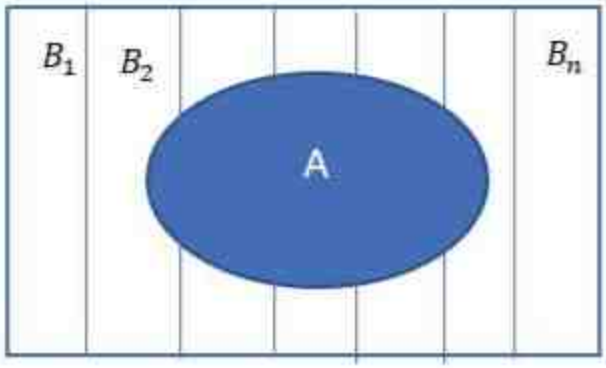
2.6.2 Hukum Total Probabilitas (*Law of Total Probability*)

Andaikan $B_1, B_2, ..., B_n$ adalah event sedemikian hingga $\bigcup_{i=1}^{n} B_i = \Omega$ dan $B_i \cap B_j = \phi$ untuk $i \neq j$ dengan $p(B_i) > 0$ untuk setiap i . Maka untuk sebarang kejadian A, berlaku

$$p(A) = \sum_{i=1}^{n} p(A|B_i)p(B_i)$$

Bukti:

Perhatikan gambar berikut



$$p(A) = p(A \cap \Omega)$$

$$= p(A \cap (\bigcup_{i=1}^{n} B_i))$$

$$= p(\bigcup_{i=1}^{n} (A \cap B_i))$$

Karena kejadian A dan B_i saling asing maka

$$p(\bigcup_{i=1}^{n} (A \cap B_i)) = \sum_{i=1}^{n} p(A \cap B_i)$$

Dengan menggunakan hukum asas pelipatan didapat

$$p(\bigcup_{i=1}^{n} (A \cap B_i)) = \sum_{i=1}^{n} p(A|B_i)p(B_i) \blacksquare$$

2.7 Tree

Seringkali permasalahan dalam probabilitas bersyarat dapat digambarkan dalam bentuk pohon (*tree*).

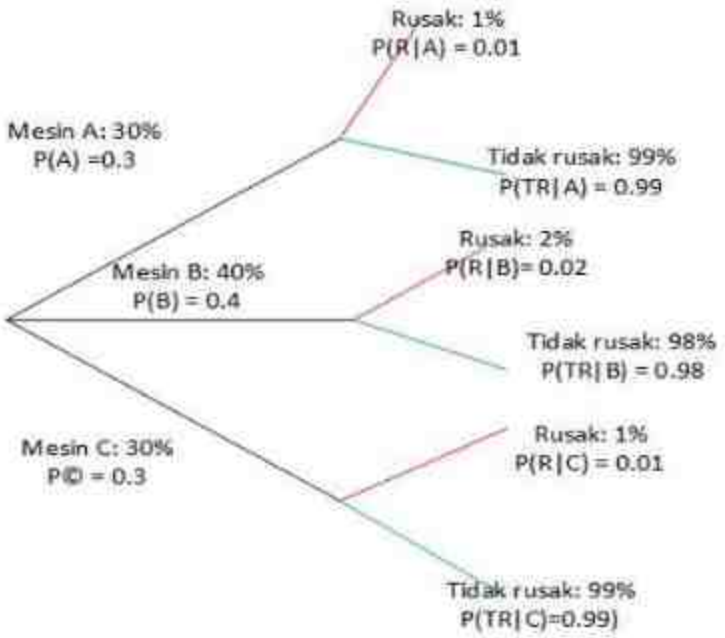
Contoh 2.21

Sebuah line produksi terdiri dari tiga buah mesin identik, Mesin A, Mesin B dan Mesin C. Mesin A mampu memproduksi 30% dari total produksi harian di line produksi tersebut, Mesin

B mampu memproduksi 40% dan Mesin C mampu memproduksi 30%. Masing-masing mesin memiliki tingkat kecacatan produk yang berbeda. Produk cacat yang dihasilkan oleh Mesin A adalah 1% dari total produksi yang dihasilkannya, Mesin B menghasilkan 2% produk cacat dan Mesin C menghasilkan 1% produk cacat. Berapakah total probabilitas bahwa barang yang diambil dari Mesin A rusak, barang yang diambil dari Mesin A tidak rusak? Demikian juga untuk Mesin B dan Mesin C.

Jawab:

Permasalahan ini dapat digambarkan sebagai tree. Dengan menggunakan hukum total probability, maka kita dapat menghitung total probabilitas bahwa barang yang diambil dari Mesin A rusak ataupun tidak rusak. Demikian juga dengan Mesin B dan Mesin C sebagai berikut.



$$
\begin{aligned}
p(A \cap R) &= p(R|A)p(A) \\
&= 0{,}01 * 0{,}3 = 0{,}003 \\
p(A \cap TR) &= p(TR|A)p(A) \\
&= 0{,}99 * 0{,}3 = 0{,}297 \\
p(B \cap R) &= p(R|B)p(B) \\
&= 0{,}02 * 0{,}4 = 0{,}008 \\
p(B \cap TR) &= p(TR|B)p(B) \\
&= 0{,}98 * 0{,}4 = 0{,}392 \\
p(C \cap R) &= p(R|C)p(C) \\
&= 0{,}01 * 0{,}3 = 0{,}003 \\
p(C \cap TR) &= p(TR|C)p(C) \\
&= 0{,}99 * 0{,}3 = 0{,}297
\end{aligned}
$$

2.8 Aturan Bayes (*Bayes Rule*)

Andaikan A dan $B_1, B_2, ..., B_n$ adalah suatu kejadian, B_i saling asing, i.e, $B_i \cap B_j = \phi, i \neq j$ dan $\bigcup_{i=1}^{n} B_i = \Omega$, $p(B_i) > 0$ untuk setiap i. maka

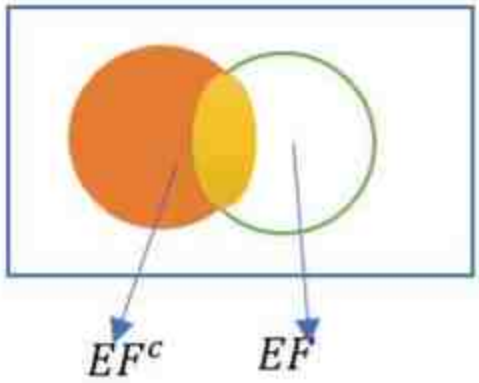
$$p(B_j|A) = \frac{p(A|B_j)p(B_j)}{\sum_{i=1}^{n} p(A|B_i)p(B_i)}$$

jika $j = 1$

$$p(B_1|A) = \frac{p(A|B_1)p(B_1)}{\sum_{i=1}^{n} p(A|B_i)p(B_i)}$$

Ilustrasi 1

Misal E dan F adalah suatu kejadian, E dapat dituliskan sebagai berikut



$$E = EF \cup EF^e; EF \cap EF^e = \phi$$

E dan F disebut sebagai mutually disjoint

$$p(E) = p(EF) + p(EF^e)$$

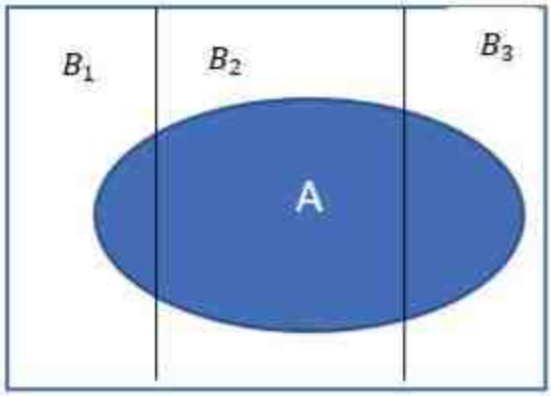
$$= p(E|F)p(E) + p(E|F^e)p(F^e)$$

$$= p(E|F)p(E) + p(E|F^e)[1 - p(F)]$$

$$p(F|E) = \frac{p(EF)}{p(E)} = \frac{p(E|F)p(F)}{p(E|F)p(E) + p(E|F^e)[1 - p(F)]}$$

Ilustrasi 2

Misalkan A dan B_1, B_2, B_3 adalah suatu kejadian, B_1, B_2, B_3 *mutually disjoint*. Perhatikan gambar berikut. $A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3)$



$$p(B_1|A) = \frac{p(B_1 \cap A)}{p(A)} = \frac{p(A \cap B_1)}{p(A)} = \frac{p(A|B_1)p(B_1)}{p(A)}$$

$$p(A) = p(A|B_1)p(B_1) + p(A|B_2)p(B_2) + p(A|B_3)p(B_3)$$

Contoh 2.22

Dari Contoh 2.21, diketahui kejadian hasil produksi rusak. Berapakah probabilitas bahwa produksi rusak itu dihasilkan oleh Mesin A?

Jawab

Pada kasus ini yang kita ketahui dulu adalah hasil produksinya rusak, dan yang ingin diketahui bahwa produksi rusak tersebut berasal dari Mesin A. Hal ini sama seperti kita diminta untuk menghitung $p(A|R)$. Untuk dapat menghitungnya, kita akan menggunakan aturan Bayes, dengan pertama-tama menghitung probabilitas produk tersebut rusak.

$$\begin{aligned} p(R) &= p(R|A)p(A) + p(R|B)p(B) + p(R|C)p(C) \\ &= 0,03 + 0,08 + 0,03 = 0,14 \\ p(A|R) &= \frac{p(A \cap R)}{p(R)} = \frac{0,03}{0,14} = 0,214 \end{aligned}$$

Latihan

Dengan cara yang sama, hitunglah bahwa produk rusak tersebut berasalah dari Mesin B, dan Mesin C.

2.9 Independensi (Ketidaktergantungan)

A dan B dikatakan sebagai dua kejadian yang independent jika terjadinya A tidak mempengaruhi apapakah B terjadi atau tidak.

$$p(A|B) = p(A); p(B|A) = p(B)$$

$$p(A) = p(A|B) = \frac{p(A \cap B)}{p(B)}$$

$$p(A \cap B) = p(A)p(B)$$

Contoh 2.23

Sebuah koin dilempar dua kali. Ruang sampelnya adalah $S = \{HH, HT, TH, TT\}$. Misalkan H_1 adalah luaran lemparan pertama adalah H. dan H_2 adalah luaran lemparan kedua adalah H, dan $2H$ adalah kedua lemparan adalah H, maka $p(H_1) = \frac{2}{4}$; $p(H_2) = \frac{2}{4}$; $p(2H) = \frac{1}{4}$. Dengan demikian

$$p(2H|H_1) = \frac{p(2H)}{p(H_1)} = \frac{1\backslash 4}{2\backslash 4} = p(2H)$$

Maka dapat disimpulkan bahwa peristiwa koin dilempar dua kali ini adalah peristiwa independent. Hasil lemparan pertama tidak mempengaruhi hasil lemparan kedua dan sebaliknya.

Mutually Independent

Kumpulan kejadian $A_1, A_2, ..., A_n$ dikatakan saling indepedent jika untuk sebarang sub kumpulan $A_{i_1}, A_{i_2}, ..., A_{i_m}$ maka

$$p(A_{i_1} \cap A_{i_2} \cap ... \cap Ai_m) = p(A_{i_1}) \cap p(A_{i_2}) \cap ... \cap p(A_{i_m})$$

2.9.1 Peubah Acak (*Random Variable*)

Pada subbab terdahulu kita telah mempelajari teutang percobaan acak, ruang sampel dan events. Semua yang kita pelajari ini menghasilkan *outcome* yang dapat kita nyatakan sebagai anggota himpunan. Pada subbab ini, kita ingin mengasosiasikan *outcome* yang didapatkan dari suatu percobaan acak, ke dalam sebuah angka. Untuk setiap outcome ω yang berada dalam ruang sampel, akan diasosiasikan dengan sebuah angka $X(\omega) = x$.

Definisi 2.1

Random Variable X adalah fungsi $X : S \rightarrow R$, yaitu fungsi yang mengasosiasikan setiap outcome $\omega \in S$ tepat ke sebuah angka $X(\omega) = x$

Contoh 2.24

Misalkan E adalah percobaan melempar koin dua kali. Ruang sampel dari percobaan ini adalah $S = \{HH, HT, TH, TT\}$. Didefinisikan sebuah random variabel X sebagai jumlah head yang muncul. Pada percobaan ini $X(HH) = 2, X(HT) = 1$. Seluruh kemungkinan yang ada dapat ditabelkan sebagai berikut

$\omega \in S$	HH	HT	TH	TT
$X(\omega) = x$	2	1	1	0

Baris kedua dari tabel di atas disebut sebagai *support* dari X, yaitu himpunan dari seluruh angka yang terjadi pada X. *Support* ini dituliskan sebagai $S_X = 0, 1, 2$

Contoh 2.25

Misalkan E adalah percobaan melempar koin berkali-kali sehingga sebuah Head muncul. Ruang sampel dari percobaan ini adalah $S = \{H, TH, TTH, TTTH, ...\}$. Didefinisikan random variabel Y sebagai jumlah Tail yang terjadi sebelum Head pertama muncul. Maka *support* dari Y adalah $S_Y = \{0, 1, 2, ...\}$

Contoh 2.26

Misalkan E adalah percobaan melempar koin diudara, didefinisikan random variable Z sebagai waktu yang dibutuhkan koin tersebut jatuh ke tanah (dalam detik). Pada kasus ini ruang sampel dari percobaan ini adalah ruang sampel yang kontinu, pada setiap saat koin tersebut mungkin jatuh ke tanah, karena itu *support* dari Z adalah $S_Z = (0, \infty)$.

Catatan:

Dari ketiga contoh ini, *support* dari X adalah diskrit terhingga (*finite*), *support* dari Y, diskrit tak berhingga (*infinite*), dan *support* dari Z adalah kontinu tak berhingga (*infinite*).

Berdasarkan dari support yang mungkin terjadi ini maka random variable dapat dibedakan sebagai random variable diskrit (*discrete random variable*) dan random variable kontinu (*continuous random variable*).

Referensi

De Veaux, R., Velleman, P., and Bock D., (2016), *Stats: Data and Models*, 5th Eds. Pearson

J.A. Rice, *Mathematical Statistics & Data Analysis*, (2006), Duxbury Press

Kerns, G. J., (2010), *Introduction to Probability and Statistics using R*, GNU Free documentation Licence.

3 Distribusi Diskrit

3.1 Diskrit Random Variable

Pada bab ini kita akan mempelajari random variabel diskrit, yaitu random variabel yang memiliki support berhingga (*finite*) atau tercacah tak berhingga (*coutably infinite*):

$$S_X = \{u_1, u_2, ..., u_k\}; \text{atau } S_X = \{u_1, u_2, u_3, ...\}$$

3.1.1 Probability Mass Function (PMF)

Setiap random variable X memiliki asosiasi dengan sebuah probability mass function (PMF): $f_X : S_X \to [0, 1]$, dan didefinisikan sebagai

$$f_X(x) = p(X = x), x \in S_X$$

PMF adalah fungsi yang memetakan setiap anggota, x, yang berada di dalam *support* X ke nilai probabilitas x tersebut terjadi.

PMF adalah nilai probabilitas, hal ini berarti PMF memenuhi sifat-sifat probabilitas sebagai berikut:

1. $f_X(x) > 0$ untuk $x \in S$; nilai probabilitas kejadian x yang berada di dalam ruang sampel S lebih dari nol.
2. $\sum_{x \in S} f_X(x) = 1$; total probabilitas untuk kejadian x yang berada di dalam ruang sampel S adalah satu.
3. $p(X \in A) = \sum_{x \in A} f_X(x)$ untuk setiap event $A \subset S$

Contoh 3.1

Sebuah koin dilempar 3 kali, memiliki ruang sampel sebagai berikut:

$$S = \{HHH, HTH, THH, TTH, HHT, HTT, THT, TTT\}$$

Misalkan X adalah random variable yang merepresentasikan jumlah *Head* yang muncul, maka X akan memiliki *support* $S_X = \{0, 1, 2, 3\}$. Asumsikan koin tersebut fair, maka setiap luaran akan memiliki probabilitas $\frac{1}{8}$, dengan demikian PMF dari X dapat dituliskan sebagai berikut:

$x \in S_X$	0	1	2	3	Total
$f_X(x) = p(X = x)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	1

```
#install.packages("stringr")

library(stringr)
Koin    = c("H","T")
S = c()
for(i in 1:100)
{ Sample3 = sample(Koin,3,replace = TRUE, prob = c(0.5,0.5))
  Sample3 = paste0(Sample3[1],Sample3[2],Sample3[3])
  S = c(S,Sample3)
}
S = unique(S)
S
```

```
[1] "HTH" "HHH" "THT" "HTT" "TTT" "HHT" "TTH" "THH"
```

```
LS = length(S)
Koin3 = str_split(S,"")
Sx0 = c()
for(i in 1:LS)
{ count = sum((Koin3[[i]]=="H")*1)
  Sx0 = c(Sx0,count)
}
Sx = sort(unique(Sx0))
L  = length(Sx)

fx = c()
for(i in 1:L)
 fx[i] = sum(Sx0==Sx[i])
fx = fx/LS
fx
```

```
[1] 0.125 0.375 0.375 0.125
```

```
barplot(fx, col = "blue",xlab = "#Head yang muncul", ylab = "probabilitas",
        main = "PMF #Head pelemparan koin 3 kali")
```

3.1.2 Nilai Ekspektasi (Mean), Varians dan Standar Deviasi

Nilai ekspektasi (nilai harapan) atau sering kali disebut sebagai mean, dari sebuah random variable sebenarnya sama dengan nilai rata-rata berbobot. Bobot dari nilai random variable yang muncul ini adalah nilai probabilitasnya.

$$\mu = EX = \sum_{x \in S} x f_X(x)$$

Varians dari random variable ini mengukur penyimpangan nilai random variable tersebut terhadap mean-nya.

$$\sigma^2 = VAR(X) = E(X - \mu)^2 = \sum_{x \in S} (x - \mu)^2 f_X(x)$$

Persamaan di atas dapat disederhanakan menjadi

$$\sigma^2 = EX^2 - (EX)^2$$

Akar dari varians adalah standar deviasi (simpangan baku): $\sigma = \sqrt{\sigma^2}$

Contoh 3.2

Mean, varians dan standar deviasi dari Contoh 3.1 adalah

$\mu = \sum_{x=0}^{3} x f_X(x) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 3,5$

$\sigma^2 = \sum_{x=0}^{3} (x - 3,5)^2 f_X(x)$

$= (0 - 3,5)^2 \times \frac{1}{8} + (1 - 3,5)^2 \times \frac{3}{8} + (2 - 3,5)^2 \times \frac{3}{8} + (3 - 3,5)^2 \times \frac{1}{8}$

$= \frac{12,25}{8} + \frac{18,75}{8} + \frac{6,75}{8} + \frac{0,25}{8} = \frac{38}{8} = 4,75$

$\sigma = \sqrt{4,75} = 2,18$

Interpretasi dari hasil di atas adalah apabila kita melakukan percobaan ini berulang kali secara independent, maka nilai harapan bahwa head yang akan muncul pada percobaan ini sebesar 3,5 kali dengan simpangan sebesar 2,18.

Contoh 3.3.

Sebuah perusahaan asuransi kecelakaan akan membayar klien nya sebesar Rp 100 juta jika klien nya mengalami kecelakaan dan meninggal, Rp 50 juga jika klien tersebut menjadi cacat karena kecelakaan itu atau Rp 0 jika klien tidak mengalami kecelakaan. Andaikan dari pengalaman masa lalu diketahui bahwa dari 1000 klien yang dimiliki perusahaan tersebut 1 orang mengalami kecelakaan dan meninggal, 2 orang mengalami kecelakaan dan menjadi cacat, sisanya tidak pernah mengalami kecelakaan berapakah nilai rata-rata asuransi yang dibayarkan oleh perusahaan tersebut terhadap kliennya? Berapa standar deviasinya?

Jawab

Peristiwa di atas merupakan peristiwa acak. dengan ruang sampel S = {Meninggal, Cacat, Tidak terjadi kecelak
Random variable yang berasosiasi dengan ruang sampel ini adalah nilai asuransi yang dibayarkan, X = jumlah pembayaran klaim asuransi, dengan support dari X, S_X = {100 juta, 50 juta, 0} Rupiah, dengan probabilitas terjadi $f_X(x) = \{\frac{1}{1000}, \frac{2}{1000}, \frac{997}{1000}\}$ dalam juta Rupiah.

$\mu = E(X) = 100 \times \frac{1}{1000} + 50 \times \frac{2}{1000} + 0 \times \frac{997}{1000} = 0,2$

Ruang sampel (S)	Pembayaran (S_X)	Probabilitas ($f_X(x)$)	Deviasi ($X - \mu$)
Meninggal	100	1/1000	(100-0,2) = 99,8
Cacat	50	2/1000	(50-0,2) = 49,8
Tidak terjadi kecelakaan	0	997/1000	(0-0.2) = -0,2

$\sigma^2 = 99,8^2 \frac{1}{1000} + 49,8^2 \frac{2}{1000} + (-0,2)^2 \frac{997}{1000} = 14,96$ juta Rupiah

$\sigma = \sqrt{\sigma^2} = 3,87$ juta Rupiah

Dalam hal ini, secara average perusahaan membayar klaim asuransi sebesar Rp 200 ribu dengan standar deviasi sebesar Rp 3, 87 juta.

Sifat-sifat Mean dan Varians

Jika c - konstan maka

$E(X \pm c) = E(X) \pm c$	$Var(X \pm c) = Var(X)$
$E(cX) = cE(X)$	$Var(cX) = c^2 Var(X)$
$E(X \pm Y) - E(X) \pm E(Y)$	

Bila X dan Y dua random variable yang saling independen, maka

$Var(X \pm Y) = Var(X) + Var(Y)$

Perhatikan, karena random variable adalah fungsi yang memiliki domain ruang sampel yang berasal dari peristiwa random, maka

$$X + X + X \neq 3X$$

Nilai X tidak akan pernah diketahui, hingga dia terjadi (terealisasi). Dalam notasi matematika seringkali dituliskan bila X adalah sebuah random variable, maka x adalah realisasi dari X.

Pada Contoh 3.1, X adalah jumlah Head yang muncul, setelah koin itu dilempar tiga kali. Misalkan kita telah melempar koin tiga kali, ternyata yang muncul adalah HHH, maka realisasi dari X adalah 3. Apabila koin ini sekali lagi dilempar 3 kali, kemudian luaran yang muncul adalah HHT, maka realisasi dari X adalah 2. Karena realisasi dari X juga merupakan variable, maka seringkali realisasi itu dituliskan sebagai x.

Selain itu perhatikan bahwa pada pelomparan koin tiga kali yang kedua belum tentu menghasilkan luaran yang sama. Oleh sebab itu $X + X \neq 2X$.

3.1.3 Cummulative Distribution Function (CDF)

Fungsi distribusi kumulatif (*Cummulative Distribution Funcion -CDF*) adalah nilai probabilitas dari sebuah random variable secara kumulatif.

$$F_X(t) = p(X \leq t), -\infty < t < \infty$$

CDF, F_X memenuhi sifat

1. F_X adalah fungsi non-decreasing, jika $t_1 < t_2$ maka, $F_X(t_1) < F_X(t_2)$
2. F_X adalah fungsi yang kontinu dari kanan; $\lim_{t \to a^+} F_X(t) = F_X(a)$ untuk semua $a \in \mathbb{R}$
3. $\lim_{t \to -\infty} F_X(t) = 0$ dan $\lim_{t \to \infty} F_X(t) = 1$

Dari definisi di atas kita bisa menyebutkan random variable X berdistribusi F_X dan memenuliskannya sebagai $X \sim F_X$

Contoh 3.4

CDF dari Contoh 3.1 adalah

$x \in S_X$	0	1	2	3	Total
$f_X(x) = p(X = x)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	1
$F_X(x) = p(X \leq x)$	$\frac{1}{8}$	$\frac{4}{8}$	$\frac{7}{8}$	1	

```
X  = Sx
mu = sum(X*fx)
mu
```

```
[1] 1.5
```

```
sigma2 = sum((X-mu)^2*fx)
sigma2
```

```
[1] 0.75
```

```
sigma  = sqrt(sigma2)
sigma
```

```
[1] 0.8660254
```

```
FX = cumsum(fx)
FX
```

```
[1] 0.125 0.500 0.875 1.000
```

```
barplot(FX, col = "blue",xlab = "#Head yang muncul", ylab = "probabilitas",
        main = "CDF #Head pelemparan koin 3 kali")
```


```
#install.packages("distrEx)
options(warn=-1)
suppressMessages(library(distrEx))
X = DiscreteDistribution(supp=0:3, prob = c(1,3,3,1)/8)
E(X)
```

```
[1] 1.5
```

```
var(X)
```

```
[1] 0.75
```

```
sd(X)
```

```
[1] 0.8660254
```

3.2 Distribusi Uniform Diskrit

Random variable X yang memiliki support pada bilangan bulat $1, 2, ..., m$, dikatakan berdistribusi diskrit bila, masing-masing luaran pada support memiliki probabilitas yang sama (*equally likely*) untuk terjadi. Dapat pula dikatakan X memiliki PMF

$$f_X(x) = \frac{1}{m}, x = 1, 2, ..., m$$

Notasi: $X \sim disunif(m)$

Contoh 3.5

Sebuah perusahaan lottery, menerbitkan 1 juta kupon bernomor. Tiap kupon memiliki nomor yang berbeda. Misalkan X adalah nomor yang muncul pada lottery tersebut, maka $X \sim disunif(m = 1000000)$. Hal ini dikarenakan setiap nomor yang muncul akan memiliki kemungkinan yang sama untuk terjadi yaitu 1/1.000.000.

$f_X(x) = \frac{1}{1000000}, x = 1, 2, ..., 1000000$

Bila $X \sim diunif(m)$ maka

Mean dari X adalah

$$\mu = \sum_{i=1}^{m} x f_X(x) = \sum_{i=1}^{m} x \frac{1}{m} = \frac{1}{m}(1 + 2 + ... + m) = \frac{m+1}{2}$$

Varians dari X adalah

$$EX^2 = \frac{1}{m}\sum_{i=1}^{m} x^2 = \frac{1}{m}\frac{m(m+1)(2m+3)}{6} = \frac{(m+1)(2m+1)}{6}$$

$$\sigma^2 = EX^2 - (EX)^2 = \frac{(m+1)(2m+1)}{6} - \left(\frac{m+1}{2}\right)^2 = ... = \frac{m^2-1}{12}$$

Contoh 3.6

Pada percobaan melempar sebuah dadu, X adalah mata dadu yang muncul, maka $m = 6$

$\mu = \frac{7}{2} = 3,5$ dan $\sigma^2 = \frac{6^2-1}{12} = \frac{35}{12}$

Dalam R, untuk membuat random variable dengan diskrit uniform dapat dilakukan dengan perintah sample(x, size, replace =TRUE). Misalkan

```r
#Melempar dadu 30 kali
sample(6, size = 30, replace  = TRUE)
```

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

```r
#Memilih angka dari 10 hingga 20, sebanyak 6 kali
sample(10:20, size = 6, replace = TRUE)
```

```
[1] 10 16 12 17 15 17
```

```r
#Melempar koin 10 kali
sample(c("H","T"), size = 10, replace = TRUE)
```

```
 [1] "T" "T" "T" "T" "H" "H" "T" "T" "H" "H"
```

3.3 Percobaan Bernoulli

Percobaan Bernoulli ditemukan pertama kali oleh Daniel Bernoulli. Percobaan Bernoulli, adalah percobaan yang memiliki *outcome* yang terdiri dari "ya" atau "tidak"; "benar" atau "salah", "satu" atau "nol", "cacat" atau "tidak cacat" dan sebagainya. Oleh karena itu, random variable Bernoulli hanya akan memiliki dua nilai yaitu 1 dan 0 dengan probabilitas p dan q secara berturutan. Dengan demikian, random variable Bernoulli akan memiliki PMF

$$f(x) = \begin{cases} p^x(1-p)^{1-x}, & \text{jika } x = 0 \text{ atau } x = 1 \\ 0, & \text{lainnya} \end{cases}$$

Percobaan Bernoulli, seringkali disebut sebagai percobaan dengan *outcome* nol-satu, atau percobaan "sukses-gagal". Pada percobaan ini, p – lebih umum disebut sebagai proporsi sukses dan q – proporsi gagal.

$$p = \frac{\#\text{sukses terjadi}}{\#\text{seluruh percobaan}}$$

$$q = \frac{\#\text{gagal terjadi}}{\#\text{seluruh percobaan}}$$

Karena outcome dari percobaan ini hanya sukses dan gagal, maka $p + q = 1$; atau seringkali dituliskan $q = 1 - p$.

Contoh 3.7

Dalam sebuah lantai produksi, diketahui bahwa proporsi produk cacat adalah 0,02. Berarti dari seratus produk yang diperiksa didapati dua produk cacat. Pada kasus ini dapat dituliskan "sukses" di sini adalah menemukan produk cacat, $p = 0,02$; dan "gagal" adalah tidak menemukan produk cacat, $q = 0,98$

Bila dilakukan inspeksi dengan mengambil produk satu persatu, maka probabilitas ditemukan produk cacat pada pengambilan produk pertama adalah

$$p(X = 1) = 0,02$$

Random variable X di sini adalah jumlah produk yang diambil hingga ditemukan produk tersebut cacat.

Bila inspeksi ini dilakukan secara independen. *Outcome* dari pengambilan produk pertama tidak mempengaruhi *outcome* dari pengambilan kedua dan seterusnya. Berapakah probabilitas bahwa ditemukan produk cacat, pada pengambilan produk kedua?

Pada kasus ini, produk pertama yang diambil tidak cacat; pada pengambilan kedua ditemukan produk tersebut cacat. Oleh karena kasus ini independent, maka probabiltas didapatkan produk cacat, pada pengambilan produk kedua adalah

$$p(X = 2) = (0,98)(0,02)$$

```
prob = dgeom(1, p = 0.02)
prob
```

```
[1] 0.0196
```

Bila inspeksi ini dilakukan berulang kali, probabilitas akan ditemukan produk cacat pada pengambilan produk kelima adalah

$$p(X = 5) = (0,98)^4(0.02)$$

```
prob = dgeom(4,p=0.02)
prob
```

```
[1] 0.01844736
```

Catatan:

R menggunakan rumus: $p(X = x) = p(1-p)^x$

Untuk menemukan sebuah produk cacat, berapa banyak produk ini harus diinspeksi secara rata-rata? Karena proporsi produk cacat ini adalah 2%, atau 1 per 50 produk, maka secara rata-rata kita berharap pada pengambilan produk ke-50 kita akan menemukan sebuah produk cacat, $\mu = \frac{1}{0.02} = 50$ produk.

3.4 Distribusi Geometri

Sebuah random variable X akan berdistribusi Geometric, $X \sim Geom(p)$; apabila random variable tersebut muncul dari percobaan Bernoulli yang berulang-ulang, dengan X adalah jumlah percobaan hingga sukses pertama terjadi. PMF dari Xadalah

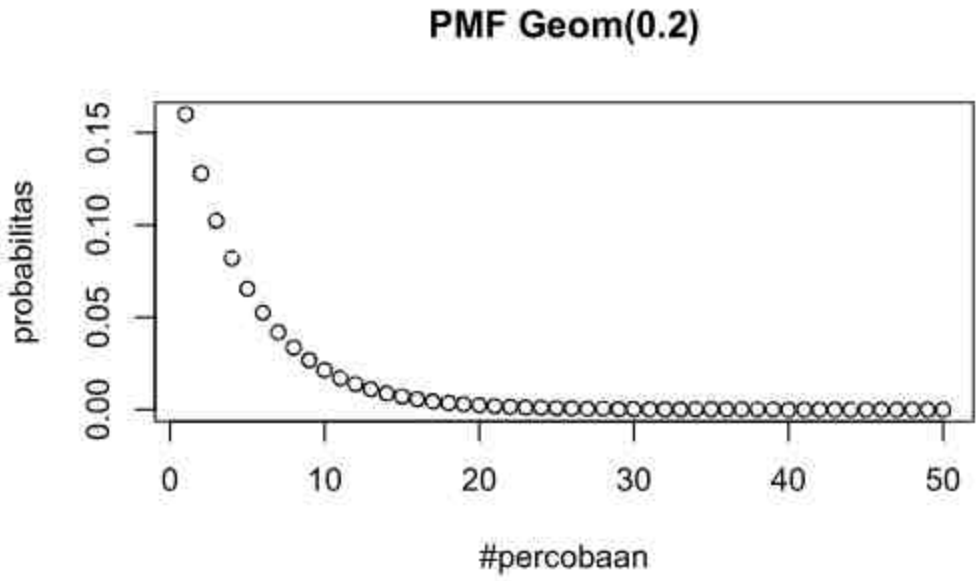
$$f_X(x) = p(X = x) = q^{x-1}p$$

dimana p - proporsi sukses, q - proporsi gagal. Mean dan variance dari distribusi Geometric adalah

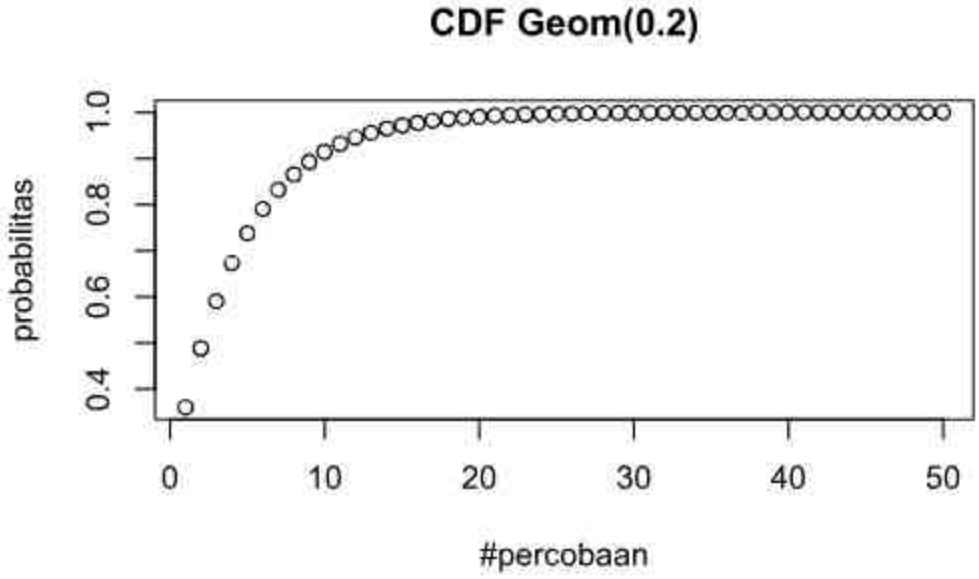
$$\mu = \frac{1}{p};\ \sigma^2 = \frac{q}{p^2}$$

PMF dan CDF dari distribusi Geometrik dapat dilihat pada Gambar berikut secara berturutan.

```
x  = 1:50
p  = 0.2
fx = dgeom(x,p)
plot(fx, main = "PMF Geom(0.2)", xlab = "#percobaan", ylab = "probabilitas")
```




```
Fx = pgeom(x,p)
plot(Fx, main = "CDF Geom(0.2)",xlab = "#percobaan", ylab = "probabilitas")
```



Pada distribusi Geometrik, p disebut sebagai distribusi parameter. Nilai p inilah yang menentukan bentuk distribusi.

Catatan: Kondisi 10%

Percobaan Bernoulli yang dilakukan harus independen. Jika asumsi ini tidak terpenuhi, distribusi Geometric ini masih tetap dapat diterapkan asalkan sampel yang diambil lebih kecil dari 10% dari total seluruh data (populasi data).

Contoh 3.7 menggambarkan tentang distribusi Geometric ini. Untuk melihat perubahan bentuk distribusi ubahlah nilai p dan perhatikan bentuk dari distribusi Geometrik.

Contoh 3.8

Andaikan saat ini inspeksi produk cacat pada Contoh 3.7 tidak dilakukan satu persatu, namun dilakukan per batch, katakanlah tiap batch terdiri dari lima produk, berapakah probabilitas ditemukan dua produk cacat pada batch tersebut?

Kasus ini merupakan percobaan Bernoulli yang berulang, namun saat ini kita tertarik untuk mengetahui probabilitas mendapatkan 2 sukses dari 5 kali percobaan.

Jika p – proporsi sukses, dan q – proporsi gagal, maka probabilitas untuk mendapatkan satu kombinasi didapatkan 2 produk cacat dari 5 produk yang diperiksa adalah p^2q^3. Secara keseluruhan terdapat $C_2^5 = \frac{5!}{3!2!}$ kombinasi yang mungkin dari penemuan produk cacat di antara 5 produk yang diperiksa ini; yaitu

SSGGG, SGSGG, SGGSG, SGGGS, GSSGG, GSGSG, GSGGS, GGSSG, GGSGS, GGGSS

S – Sukses (ditemukan produk cacat), G – Gagal (tidak ditemukan produk cacat).

Hal ini berarti bahwa probabilitas ditemukan 2 produk cacat dari 5 produk yang diperiksa adalah

$$p(X = 2) = C_2^5 p^2 q^3$$

3.5 Distribusi Binomial

Sebuah random variable X akan berdistribusi Binomial, $X \sim binom(n, p)$; apabila random variable tersebut muncul dari percobaan Bernoulli yang berulang-ulang, dengan X- jumlah sukses yang diperoleh dari n percobaan, dimana n, jumlah produk dalam satu batch ditentukan sebelum percobaan dilakukan. PMF dari X adalah

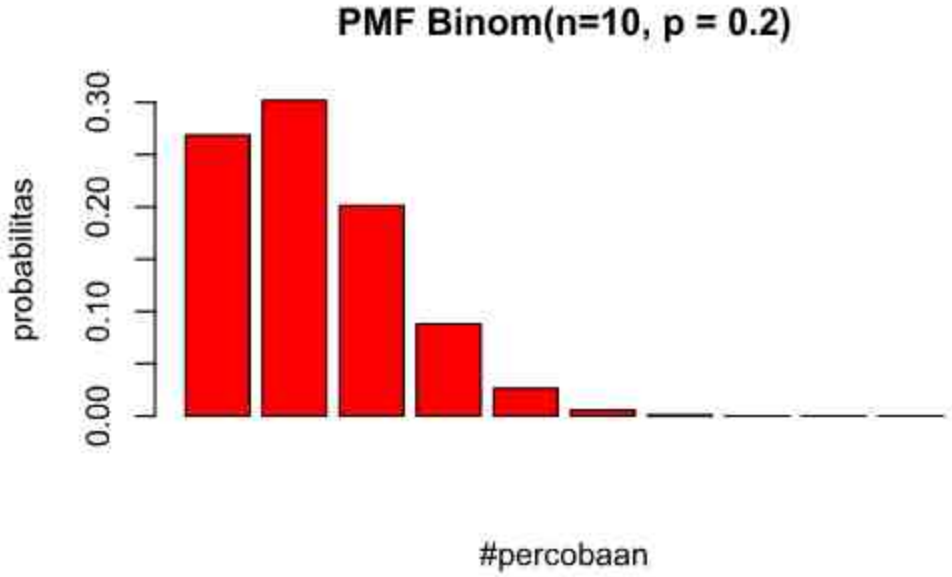
$$f_X(x) = p(X = x) = C_x^n p^x q^{n-x}$$

dimana p - proporsi sukses, q - proporsi gagal. Mean dan varians dari distribusi Binomial adalah

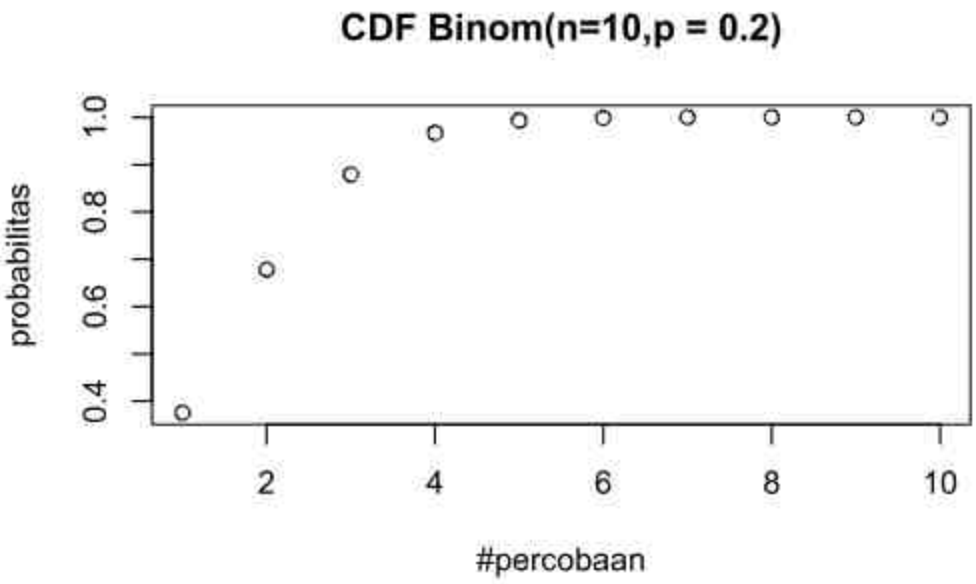
$$\mu = np; \sigma^2 = npq$$

Pada distribusi Binomial, n, p disebut sebagai distribusi parameter. Nilai n dan p inilah yang menentukan bentuk distribusi, dan kedua parameter ini harus ditentukan/diketahui di awal. Terlihat bahwa ketika kita mengubah nilai n, p, maka bentuk distribusi juga berubah

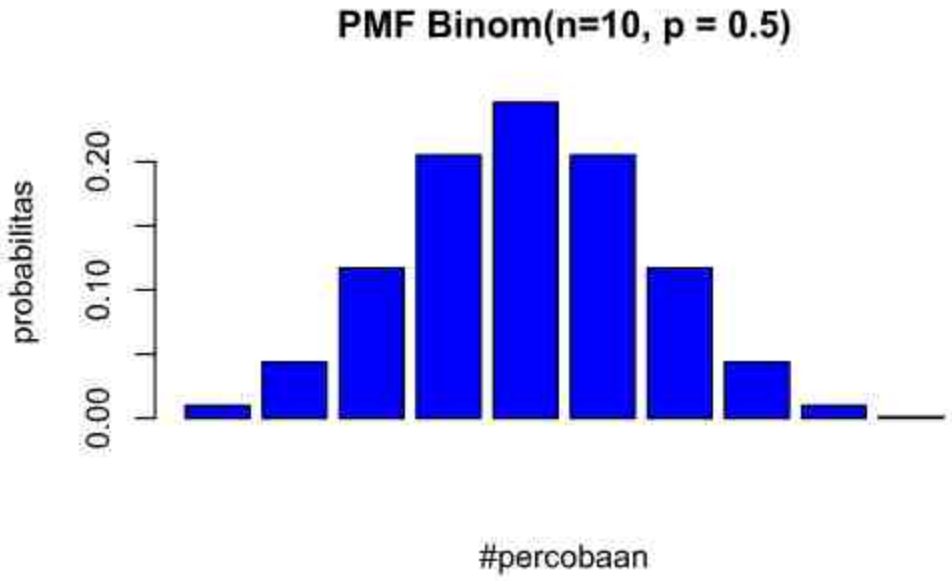
```
x  = 1:10
p  = 0.2
n  = 10
fx = dbinom(x,n,p)
barplot(fx, col = 'red', main = "PMF Binom(n=10, p = 0.2)",
              xlab = "#percobaan", ylab = "probabilitas")
```




```
Fx = pbinom(x,n,p)
plot(Fx, main = "CDF Binom(n=10,p = 0.2)",
     xlab = "#percobaan", ylab = "probabilitas")
```




```
p  = 0.5
fx = dbinom(x,n,p)
barplot(fx, col = 'blue',main = "PMF Binom(n=10, p = 0.5)",
        xlab = "#percobaan", ylab = "probabilitas")
```



Andaikan pada Contoh 3.8, nilai $p = 0.98$ dan $q = 0,02$

$$p(X = 2) = C_2^5(0,98)^2(0,03)^3$$

```
prob = dbinom(2,5,0.98)
prob
```

```
[1] 7.6832e-05
```

Catatan

R menggunakan rumus: $p(X = x) = C_x^n p^x (1-p)^{n-x}$

3.6 Distribusi Poisson

Andaikan saat ini inspeksi produk cacat, tidak dilakukan per batch, namun inspektor hanya tertarik pada jumlah cacat yang dijumpainya. Kasus ini seperti pada Contoh 3.8, hanya saja saat ini n tidak dibatasi, nilai n dianggap sangat besar, yaitu seluruh total produksi, sedangkan proporsi kecacatan yang terjadi sangat kecil.

Simeon Denis Poisson, seorang matematikawan berkebangsaan Perancis tertarik untuk meneliti suatu kejadian dengan probabilitas sangat kecil. Poisson menurunkan model

matematikanya dengan menggunakan pendekatan model Binomial, dengan proporsi sukses, p sangat kecil dan jumlah percobaan n, sangat besar.

Sebuah random variable X, dikatakan berdistribusi Poisson, $X \sim Poisson(\lambda)$, apabila X memiliki PMF

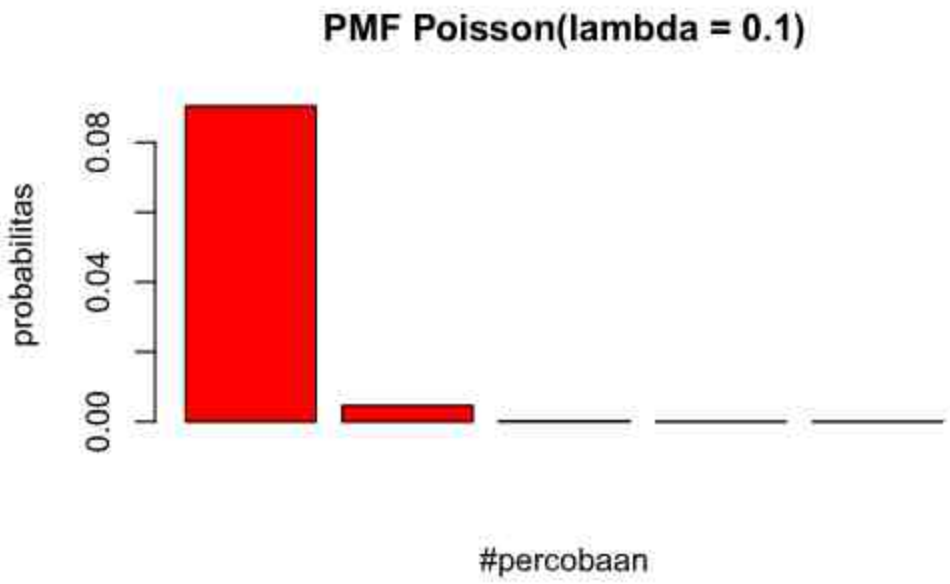
$$f_X(x) = p(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

λ - mean dari jumlah kesuksesan

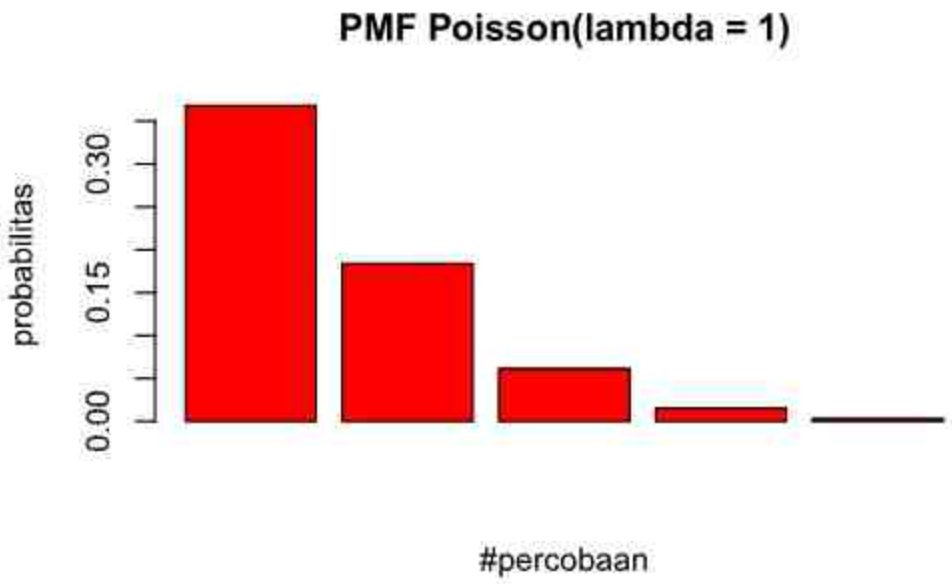
X- jumlah kesuksesan yang terjadi.

Mean dan varians dari distribusi Poisson adalah $\mu = \lambda; \sigma^2 = \lambda$.Parameter pada distribusi Poisson adalah λ. Nilai λ inilah yang menentukan bentuk distribusi. Gambar berikut menunjukkan perubahan bentuk distribusi Poisson, bila nilai λ diubah.

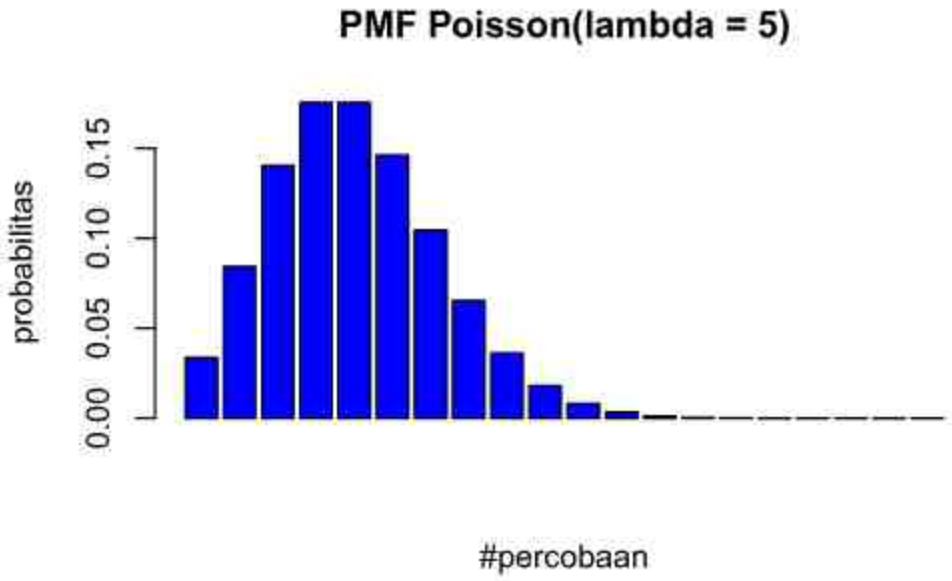
```
x = 1:5
lambda = 0.1
fx = dpois(x,lambda)
barplot(fx, col = 'red', main = "PMF Poisson(lambda = 0.1)",
        xlab = "#percobaan", ylab = "probabilitas")
```



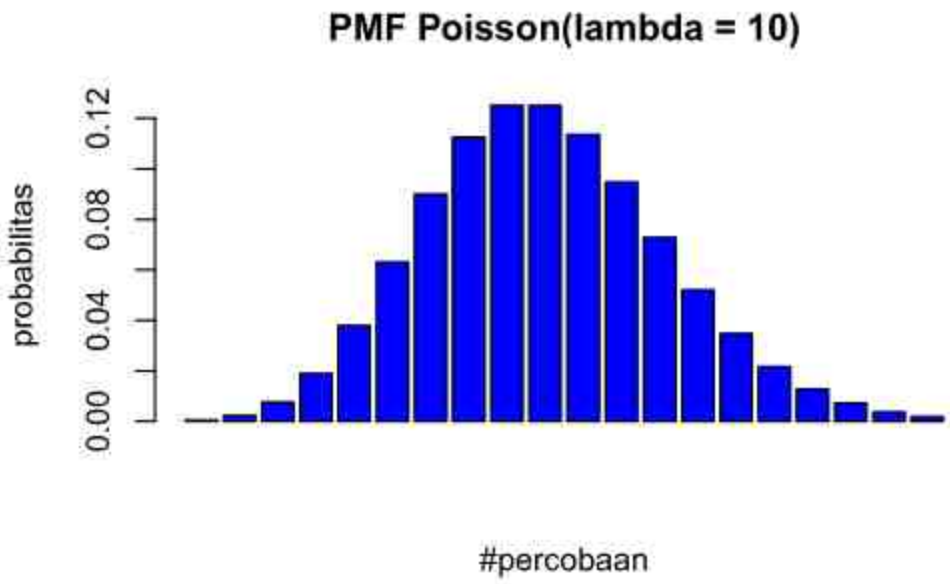
```r
lambda = 1
fx = dpois(x,lambda)
barplot(fx, col = 'red', main = "PMF Poisson(lambda = 1)",
        xlab = "#percobaan", ylab = "probabilitas")
```




```r
x = 1:20
lambda = 5
fx = dpois(x,lambda)
barplot(fx, col = 'blue', main = "PMF Poisson(lambda = 5)",
        xlab = "#percobaan", ylab = "probabilitas")
```




```
lambda = 10
fx = dpois(x,lambda)
barplot(fx, col = 'blue', main = "PMF Poisson(lambda = 10)",
        xlab = "#percobaan", ylab = "probabilitas")
```



Distribusi Poisson, seringkali digunakan untuk memodelkan jumlah kecelakaan, jumlah orang dalam antrian, jumlah klaim asuransi, dan sebagainya.

Contoh 3.9

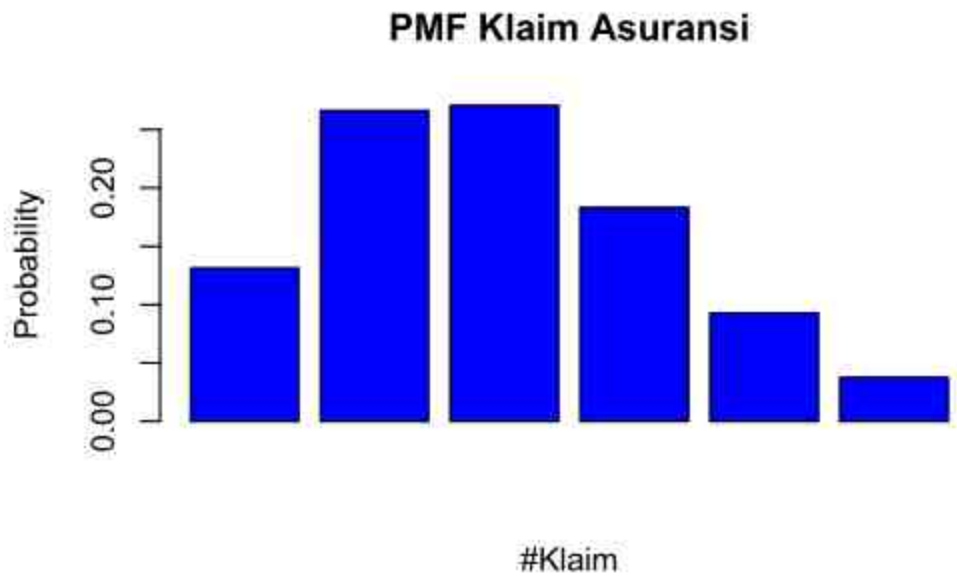
Pada sebuah agen asuransi didapat jumlah klaim asuransi sebagai berikut

#Klaim	0	1	2	3	4	≥ 5	Total
#Observasi	22	53	58	39	20	8	200

Rata-rata jumlah klaim $= \frac{(0\times 22)+(1\times 53)+(2\times 58)+(3\times 39)+(4\times 20)+(5\times 8)}{200} = 2.03$

Distribusi pengajuan klaim dapat dituliskan sebagai berikut:

```
#RScript
Klaim = 0:5
Obs = c(22,53,58,39,20,8)
Total = sum(Obs)
Lambda = sum(Klaim*Obs)/Total
Prob = dpois(Klaim,Lambda)
barplot(Prob, col = 'blue', main = "PMF Klaim Asuransi",
        xlab = "#Klaim", ylab = "Probability")
```



Contoh 3.10

Andaikan inspektor pada Contoh 3.7 tidak hanya tertarik mendapatkan produk cacat sekali saja, namun dia tertarik untuk mendapatkan produk cacat tiga kali. Berapakah jumlah produk tidak cacat yang harus diperiksanya hingga didapatkan produk cacat ketiga?

Pada kasus ini random variable X adalah jumlah produk tidak cacat yang diperiksa sebelum ditemukan produk cacat sebanyak tiga buah. Percobaan ini seperti pada percobaan Binomial, hanya saja n, jumlah sampel yang harus diperiksa pada kasus Binomial, sangat tergantung pada nilai x.

Jika $x = 0$, menunjukkan bahwa pada saat inspeksi dilakukan, ditemukan tiga produk cacat berturut-turut. $p(X = 0) = p^3q^0$

```
#Rscript, misalkan p = 0.98
prob = dnbinom(0,3,0.98)
prob
```

```
[1] 0.941192
```

Jika $x = 1$, menunjukkan ditemukan satu produk tidak cacat, sebelum ditemukan tidak produk cacat pada inspeksi tersebut. Kombinasi sukses (menemukan produk cacat) atau gagal (tidak menemukan produk cacat), pada kasus ini adalah GSSS, SGSS, SSGS. Berarti $p(X = 1) = 3p^3q^1$

```
#Rscript, misalkan p = 0.98
prob = dnbinom(1,3,0.98)
prob
```

```
[1] 0.05647152
```

Jika $x = 2$, menunjukkan ditemukan dua produk tidak cacat, sebelum ditemukan tidak produk cacat pada inspeksi tersebut. Kombinasi sukses (menemukan produk cacat) atau gagal (tidak menemukan produk cacat), pada kasus ini adalah GGSSS, GSGSS, GSSGS, SGGSS, SGSGS, SSGGS. Berarti $p(X = 2) = 6p^3q^2$.

```
#Rscript, misalkan p = 0.98
prob = dnbinom(2,3,0.98)
prob
```

```
[1] 0.002258861
```

Perhatikan bahwa sepintas, perhitungan yang kita lakukan mirip dengan kasus Binomial. Hanya saja kombinasi pada kasus ini harus berakhir dengan Sukses, sehingga bila jumlah sukses adalah r, dan jumlah gagal adalah x, maka jumlah sampel yang bisa dikombinasikan adalah $r+x-1$ terhadap $r-1$. Kita lihat pada saat $x = 2, r = 3$, kombinasi Sukses dan Gagal, hanya terjadi pada empat $(x + r - 1)$ inspeksi saja, karena inspeksi terakhir yang ditemukan adalah produk cacat ketiga dan inspeksi berhenti, sehingga tinggal $2(r - 1)$ kejadian produk cacat yang ditemukan yang dapat dikombinasikan. Oleh karena itu kombinasi yang mungkin terjadi adalah C_{r-1}^{x+r-1}, dan

$$P(X = x) = C_{r-1}^{r+x-1} p^r q^x$$

Percobaan ini merupakan percobaan negative binomial.

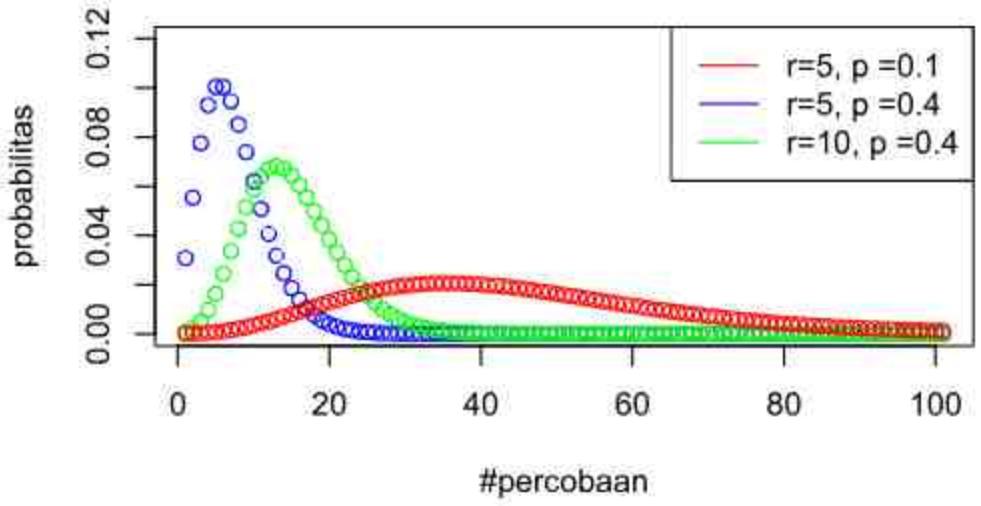
3.7 Distribusi Negative Binomial

Distribusi Negative Binomial adalah percobaan Bernoulli yang berulang, dengan random variable X – jumlah kegagalan yang terjadi sebelum diperoleh r sukses.

Distribusi ini merupakan generalisasi dari distribusi Geometrik. Pada distribusi Geometrik, kesuksesan yang diinginkan hanya satu kali saja, sedangkan pada distribusi Negative Binomial, ada r sukses yang diinginkan terjadi. PMF dari distribusi Negative Binomial adalah

$$f_X(x) = P(X = x) = C_{r-1}^{r+x-1} p^r q^x, x = 0, 1, 2, ...$$

```r
#R Script
x = 1:101
y = seq(0,0.12,0.0012)

r = 5; p = 0.1
fx1 = dnbinom(x,r,p)

p = 0.4
fx2 = dnbinom(x,r,p)

r = 10;p = 0.4
fx3 = dnbinom(x,r,p)

plot(x,y,type = 'n', xlab = "#percobaan", ylab = "probabilitas")
points(fx2, col = 'blue', main = "PMF Neg Binom")
points(fx3, col = 'green')
points(fx1,col = 'red')
legend("topright",c("r=5, p =0.1", "r=5, p =0.4", "r=10, p =0.4"),
  col = c("red","blue","green"),lty = 1)
```



Contoh 3.11

Andaikan produk barang yang dihasilkan pada Contoh 3.7 berasal dari dua Mesin. Mesin

A menghasilkan 100 produk, dan Mesin B menghasilkan 200 produk. Produk tersebut dikumpulkan pada satu line, dan dilakukan inspeksi dengan mengambil lima produk secara acak tanpa pengembalian. Berapakah kemungkin terdapat 3 produk berasal dari Mesin A?

Bila tanpa memperhatikan produk tersebut berasl dari Mesin A atau Mesin B, pengambilan sambel sebanyak 5 produk akan memiliki ruang sampel sebanyak C_5^{300} kombinasi

Produk yang berada dalam sampel tersebut 3 berasal dari Mesin A, berarti 2 sisanya berasalah dari Mesin B. Pemilihan sampel ini akan memiliki $C_3^{100}C_2^{200}$ kombinasi. Probabilitas bahwa tiga produk berasal dari Mesin A dapat dituliskan sebagai

$$P(X=3) = \frac{C_3^{100}C_2^{200}}{C_5^{300}}$$

```
prob = dhyper(3,100,200,5)
prob
```

```
[1] 0.1643189
```

atau

```
prob = choose(100,3)*choose(200,2)/choose(300,5)
prob
```

```
[1] 0.1643189
```

Catatan: R menggunakan rumus $p(x) = choose(m.x)choose(n, k-x)/choose(m+n, k)$

3.8 Distribusi Hypergeometric

Contoh 3.10 di atas merupakan contoh dari distribusi Hypergeometric.

Secara umum, bila terdapat sebuah himpunan terdiri dari

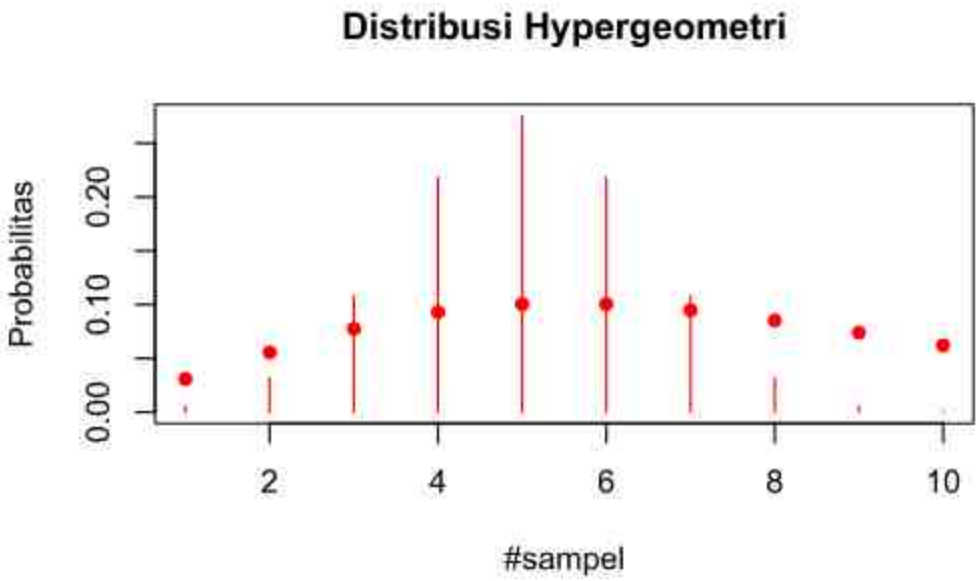
- M obyek diklasifikasikan sebagai Sukses (produk yang berasal dari Mesin A)
- N obyek yang diklasifikasikan sebagai Gagal (produk berasal dari Mesin B)

Sebuah sampel berukuran K dipilih secara acak (tanpa pengembalian)

Random variable X adalah jumlah sukses yang terjadi di dalam sampel. X berdistribusi hypergeometric, $X \sim hyper(m = M, n = N, k = K)$ dan memiliki PMF

$$f_X(x) = p(X = x) = \frac{C_x^M C_{K-x}^N}{C_K^{M+N}}$$

```
#R-script
x = 1:10
m = 25; n = 25; k = 10
fx = dhyper(x,m,n,k)
plot(fx, col = "red", type = "h",
main="Distribusi Hypergeometri",
ylab = "Probabilitas", xlab = "#sampel")
points(fx2,col = "red",pch=16)
```



References

De Veaux, R., Velleman, P., and Bock D., (2016), *Stats: Data and Models*, 5th Eds. Pearson

J.A. Rice, *Mathematical Statistics & Data Analysis*, (2006), Duxbury Press

Kerns, G. J., (2010), *Introduction to Probability and Statistics using R*, GNU Free documentation Licence.

Montgomery, D.C., and Runger, G.C., (2018) *Applied Statistics and Probability for Engineers*, 7th Eds., Wiley, USA.

4 Distribusi Kontinu

4.1 Random Variabel Kontinu

Random variabel kontinu akan memiliki support berupa interval atau gabunngan dari beberapa interval:

$$S_X = [a, b] \text{atau } (a, b)$$

Contoh random variabel kontinu: Tinggi, berat, panjang, lebar, durasi waktu, dan sebagainya.

4.1.1 Probability Density Function (PDF)

Setiap random variabel kontinu X memiliki sebuah probability density function yang dinotasikan sebagai f_x, PDF ini memenuhi tiga sifat dasar yaitu:

- Fungsi PDF adalah fungsi non-negatif, $f_X > 0$ untuk $x \in S_X$
- Total probability dari sebuah random variabel $X di seluruh supportnya adalah satu $\int_{x \in S_X} f_X(x)dx = 1$

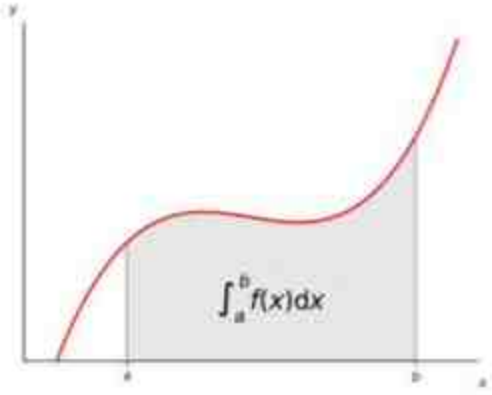
Probabiltias bahwa sebuah even terjadi dapat di definisikan sebagai integral terhadap PDF dengan support A, dimana support A merupakan himpunan bagian dari support X. $p(X \in A) = \int_{X \in A} f_X(x)dx$, untuk sebuah even $A \subset S_X$

Catatan

- Biasanya himpunan A pada sifat (3) berupa interval, misalkan $A = [a, b]$, maka

$$p(X \in A) = \int_a^b f_X(x)dx$$

- Hal ini berarti nilai probabilitas random variabel X terjadi sebenarnya merupakan luasan di bawah kurva PDF f_X yang dibatasi interval $[a, b]$ (lihat Gambar 4.1)



Gambar 4.1. Kurva PDF

- Karena luasan sebuah garis $X = a$ pada sebuah bidang adalah nol, maka $p(X = a) = 0$ untuk sebarang nilai a. Hal ini berarti

$$p(a \leq X \leq b) = p(a < X \leq b) = p(a \leq X < b) = p(a < X < b)$$

- PDF f_X dapat saja bernilai lebih dari satu, hal ini tentu saja berbeda dengan PMF. Pada random variabel diskrit, setiap nilai PMF merupakan nilai probabilitas yang didefinisikan pada interval $[0, 1]$.

4.1.2 Cumulative Distribution Function (CDF)

Fungsi distribusi kumulatif didefinisikan sebagai

$$F_X(t) = p(X \leq t) = \int_{-\infty}^{t} f_X(x)dx, -\infty < t < \infty$$

Untuk sebarang CDF F_X kontinu berlaku:

- F_X adalah fungsi non-decreasing, hal ini berarti untuk sebarang $t_1 < t_2$ maka $F_X(t_1) < F_X(t_2)$
- F_X adalah fungsi kontinu
- Limit kiri:$\lim_{t \to -\infty} F_X(t) = 0$ dan

  limit kanan: $\lim_{t \to \infty} F_X(t) = 1$

4.1.3 Ekspektasi dan Variance dari Random variabel Kontinu

Bila X adalah random variabel kontinu maka nilai ekspektasi dari X didefinisikan sebagai

$$\mu = EX = \int_{x \in S} x f_X(x) dx$$

Variance dari X didefinisikan sebagai

$$\sigma^2 = E(X - \mu)^2 = \int_{x \in S} (X - \mu)^2 f_X(x) dx$$

Sifat-sifat mean dan variance:

Bila X, Y adalah random variabel; a dan c adalah konstanta maka

$E(X \pm c) = E(X) \pm c;$ $\qquad Var(X \pm c) = Var(X)$

$E(aX) = aE(X);$ $\qquad Var(aX) = a^2 Var(X)$

Secara umum mean dari jumlahan dua random variabel adalah jumlahan dari mean. Mean dari selisih dari dua random variabel adalah selisih dari mean.

Jika dua atau lebih random variabel saling independent, maka variance dari jumlahan atau selisih dari kedua random variabel tersebut selalu merupkan jumlahan dari variance.

$E(X \pm Y) = E(X) + E(Y);$ $\qquad Var(X \pm Y) = Var(X) + Var(Y)$

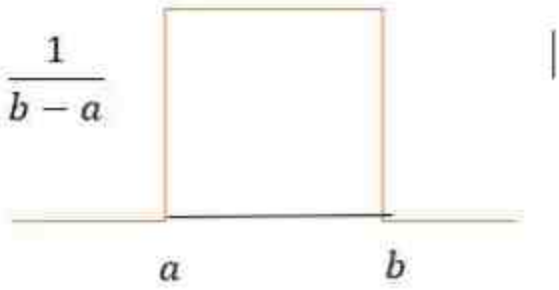
Hati-hati untuk random variabel

$X + X + X \neq 3X$

4.2 Distribusi Uniform Kontinu

Sebuah random variabel X dikatakan berdistribusi uniform kontinu pada interval (a, b), bila random variabel tersebut memiliki PDF (Gambar 4.2)

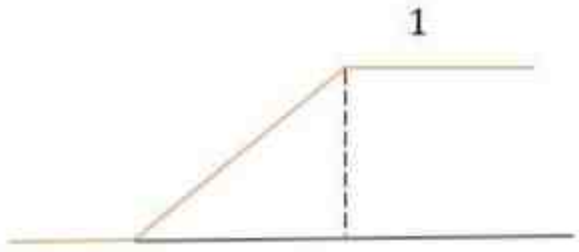
$$f_X(x) = \frac{1}{b - a}, a < X < b$$



Gambar 4.2. PDF Uniform Kontinu

dan CDF (Gambar 4.3) sebagai berikut

$$F_X(t) = \begin{cases} 0, & t < 0 \\ \frac{t-a}{b-a}, & a \leq t \leq b \\ 1, & t \geq b \end{cases}$$



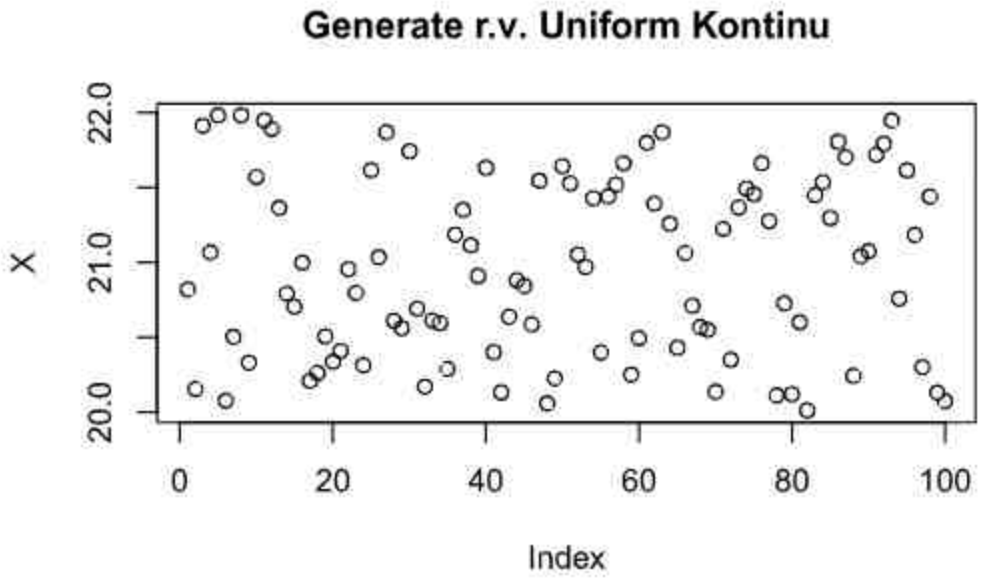
Gambar 4.3. CDF Uniform Kontinu

Distribusi uniform kontinu biasa digunakan untuk memodelkan eksperimen yang memiliki *outcome* yang berada dalam sebuah interval tertentu, (a, b) dan setiap outcome memiliki kemungkingkan yang sama untuk terjadi (*equally likely*).

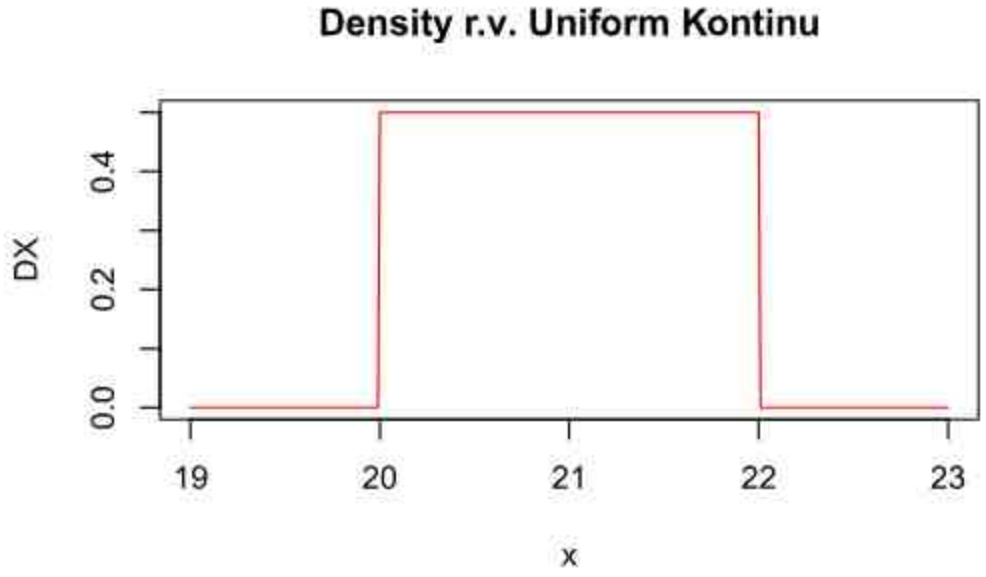
Misalkan kita memodelkan debit air, atau arus listrik keduanya dapat dimodelkan berdistribusi uniform kontinu pada interval tertentu. Katakanlah debit air Bengawan Sore di musim kemarau adalah unif(20 m^3/detik, 22 m^3/detik), maka kita dapat memodelkan aliran air di Bengawan tersebut dengan membangkitkan (*generate*) bilangan random yang mengikuti distribusi uniform kontinu di interval (20, 22) dengan menggunakan komputer. Bilangan random

ini dapat kita gunakan untuk mensimulasikan kondisi debit air di Bengawan tersebut di saat kemarau.

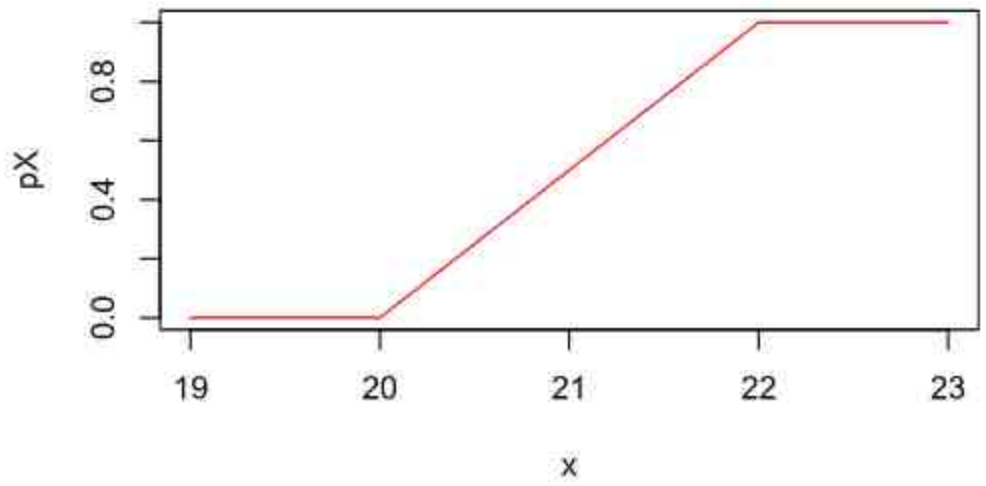
```
X   = runif(100,20,22)     #Generate random uniform kontinu
plot(X, main = "Generate r.v. Uniform Kontinu")
```



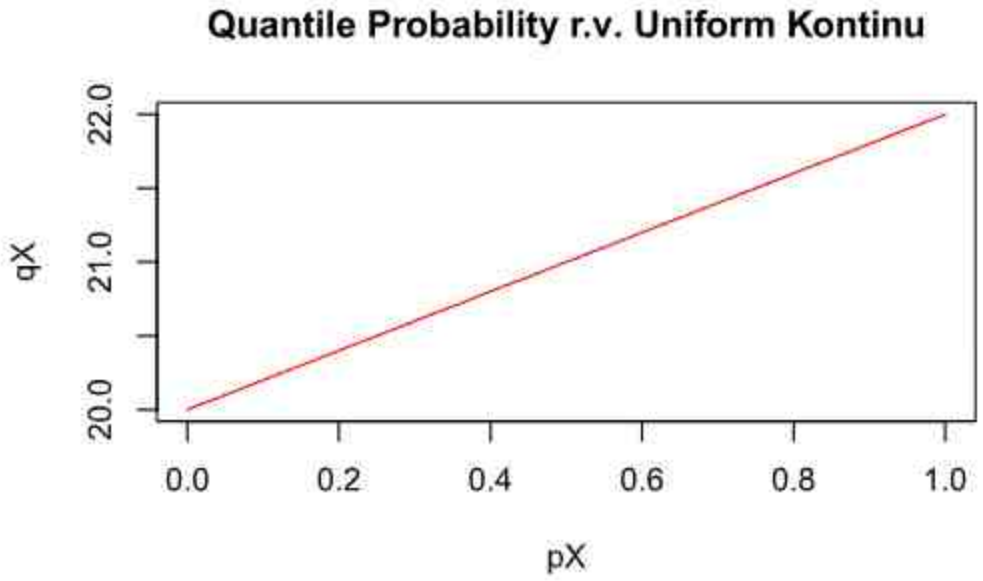

```
x   = seq(19,23, 0.01)
DX  = dunif(x,20,22)       #Density random uniform kontinu
plot(x,type = "l",col = "red", DX,
     main = "Density r.v. Uniform Kontinu")
```




```
pX  = punif(x,20,22)       #Probability random uniform kontinu
plot(x,pX, type ="l", col = "red",
     main = "Probability r.v. Uniform Kontinu")
```



```r
qX  = qunif(pX,20,22)     #Quantile random uniform kontinu
plot(pX,qX, type ="l", col = "red",
     main = "Quantile Probability r.v. Uniform Kontinu")
```



Gambar 4.3a menunjukkan plot seratus data yang dibangkitkan (*generate*) dengan menggunakan distribusi uniform kontinu $X \sim unif(20, 22)$. Terlihat bahwa seratus data yang muncul ini tersebar secara seragam di antara dua nilai yaitu 20 hingga 22 dan memiliki PDF $f_X(x) = \frac{1}{22-20} = \frac{1}{2}$ (Gambar 4.3b). Plot CDF dari random variabel ini $F_X(x) = \frac{x-20}{22-20}, 20 < x < 22$ terlihat pada Gambar 4.3c., sedangkan Gambar 4.3d menunjukkan nilai quantile (invers probability) dari distribusi uniform kontinu. Pada quantile distribusi (Gambar 4.3d) sumbu-x merepresentasikan nilai probabilitas sedangkan sumbu-y merepresentasikan nilai random variabel X yang muncul. Quantile distribusi uniform diskrit bisa dituliskan sebagai

$$Q_X(t) = \begin{cases} -\infty, & p = 0 \\ bp + a(1-p), & p > 0 \end{cases}$$

Secara umum quantile distribusi didefinisikan sebagai nilai x terkecil sedemikian hingga nilai kumulatif distribusi pada x sama dengan nilai CDF dicari p; $F_X(x) = p$

$$Q_X(x) = min\{x \in R | F_X(x) = p\}$$

Jika $X \sim unif(20, 22)$, maka nilai quantile pada $F_X(x) = 0$ adalah 20 dan $F_X(x) = 1$ adalah 22.

Mean dan Variance

Jika $X \sim unif(a, b)$ maka mean dan variance dari X adalah

$$\mu = EX = \frac{a+b}{2}$$

$$\sigma^2 = Var(X) = \frac{a^2 + ab + b^2}{3}$$

4.3 Distribusi Normal

Sebuah random variabel X dikatakan berdistribusi normal, bila X memiliki PDF dengan dua parameter yaitu $\mu; -\infty < \mu < \infty$ dan $\sigma > 0$.

$$f_X(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

dan CDF:

$$F_X(X) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

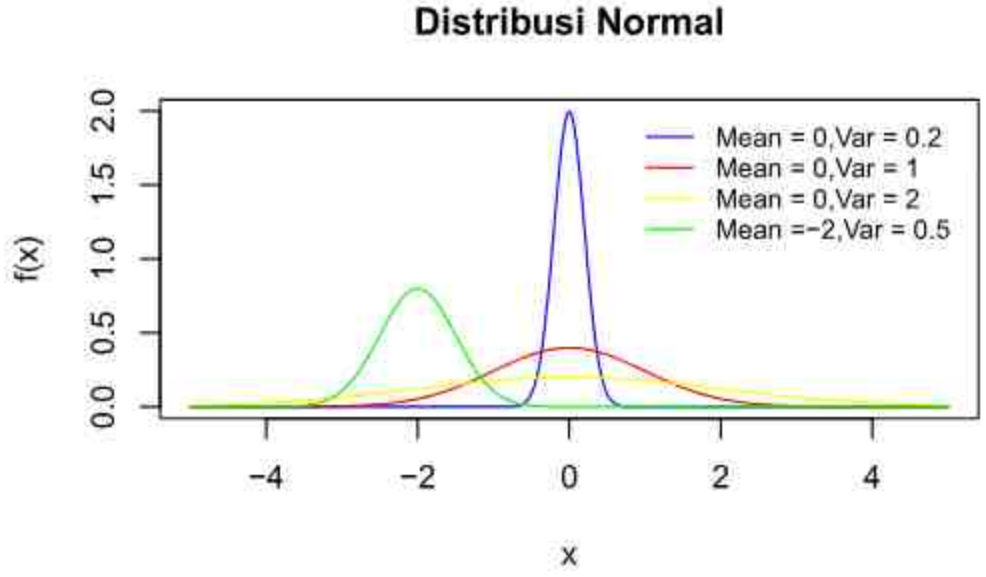
Mean dan variance dari random variabel ini adalah

$$EX = \mu \text{ dan } Var(X) = \sigma^2$$

Bila X berdistribusi normal, X dapat dituliskan sebagai $X \sim N(\mu, \sigma^2)$. Bila $\mu = 0$ dan $\sigma^2 = 1$, maka dikatakan berdistribusi normal standar atau normal baku, $X \sim N(0, 1)$.

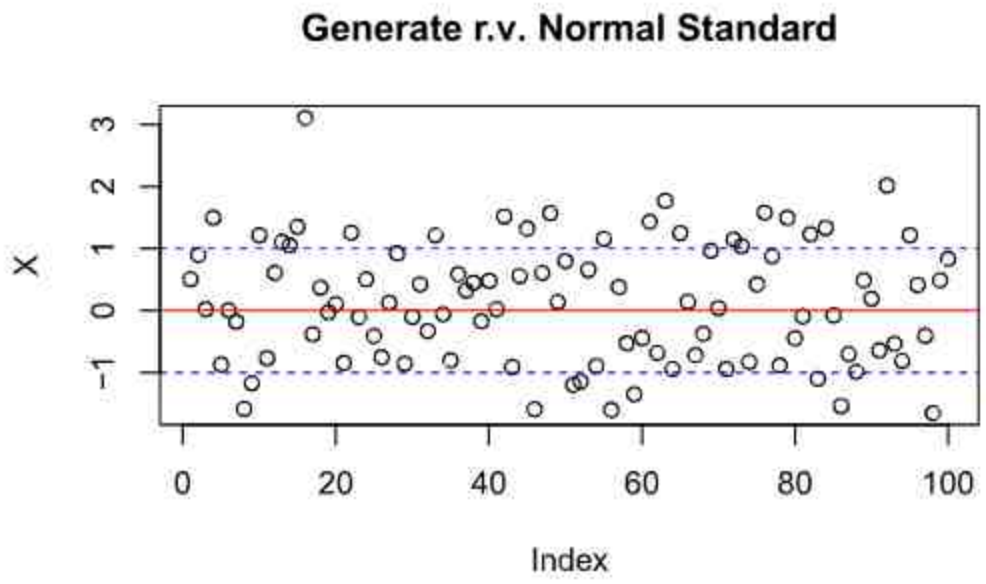
Bentuk distribusi normal ditentukan oleh kedua parameter ini. Gambar di bawah ini menunjukkan perubahan bentuk distribusi normal bila kita mengubah-ubah parameter mean (μ) dan var (σ^2). Nilai mean (μ) menunjukkan pusat dari distribusi, bila nilai mean (μ) diubah, maka pusat distribusinya agak bergeser. Namun bila nilai var (σ^2) yang diubah maka sebaran dari data akan berubah. Variance yang besar akan membuat distribusi normal ini terlihat landai, sedangkan nilai variance kecil akan membuat distribusi normal terlihat runcing.

```
x    = seq(-5,5,0.01)
DX0  = dnorm(x,0,0.2)
DX1  = dnorm(x,0,1)
DX2  = dnorm(x,0,2)
DX3  = dnorm(x,-2,0.5)

plot(x,DX0, type ="n",ylab = "f(x)")
lines(x,col = "blue", DX0)
lines(x,col = "red", DX1)
lines(x,col = "yellow",DX2)
lines(x,col = "green",DX3)

title("Distribusi Normal")
legend ("topright",inset = 0.02,legend =
c("Mean = 0,Var = 0.2","Mean = 0,Var = 1",
   "Mean = 0,Var = 2","Mean =-2,Var = 0.5"),
    col = c("blue","red","yellow","green"),
    lty=1,cex=0.8, box.col = "white")
```



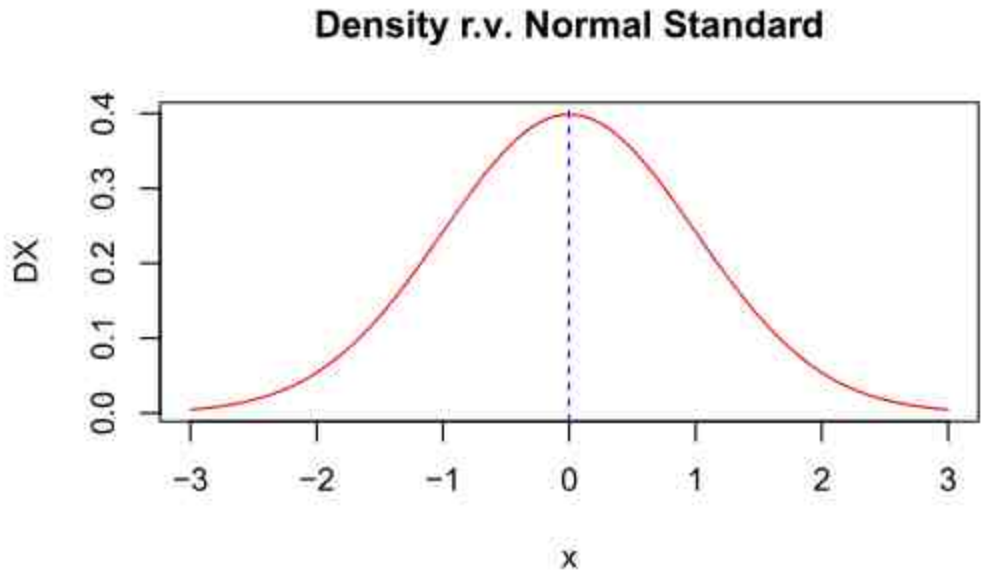
Gambar di bawah ini menunjukkan plot seratus data yang dibangkitkan dari distribusi normal standar. Terlihat bahwa sebagian besar data berada di sekitar mean $\mu = 0$, dan $\sigma^2 = 1$. Beberapa data berada di luar $\sigma^2 = 1$ (garis biru).

```
options(warn=-1)
X   = rnorm(100,0,1)    #Generate random normal standard
plot(X, main = "Generate r.v. Normal Standard")
abline(h=0, col = "red")
abline(h=1, type = "l",lty = 2,col = "blue")
abline(h=-1, type ="l",lty = 2,col = "blue")
```

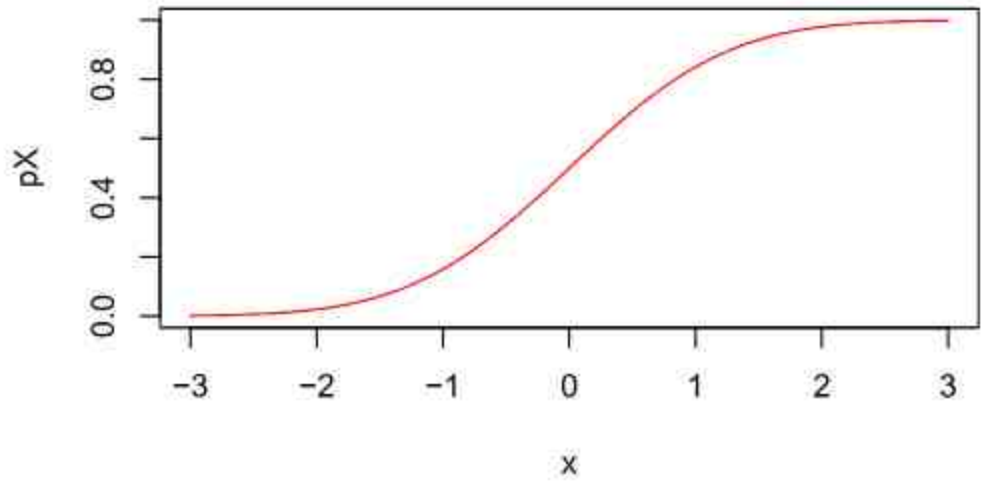


Gambar-gambar berikut menunjukkan plot distribusi normal baku, cummulatif distribusi dan quantile plot distribusi normal secara berturutan. Terlihat bahwa quantile plot (qq-plot) dari distribusi normal adalah garis diagonal yang membentang dari nilai $p(x) = 0$ hingga $p(x) = 1$. Normal qq-plot ini sering digunakan sebagai rujukan untuk menunjukkan normalitas dari suatu data.

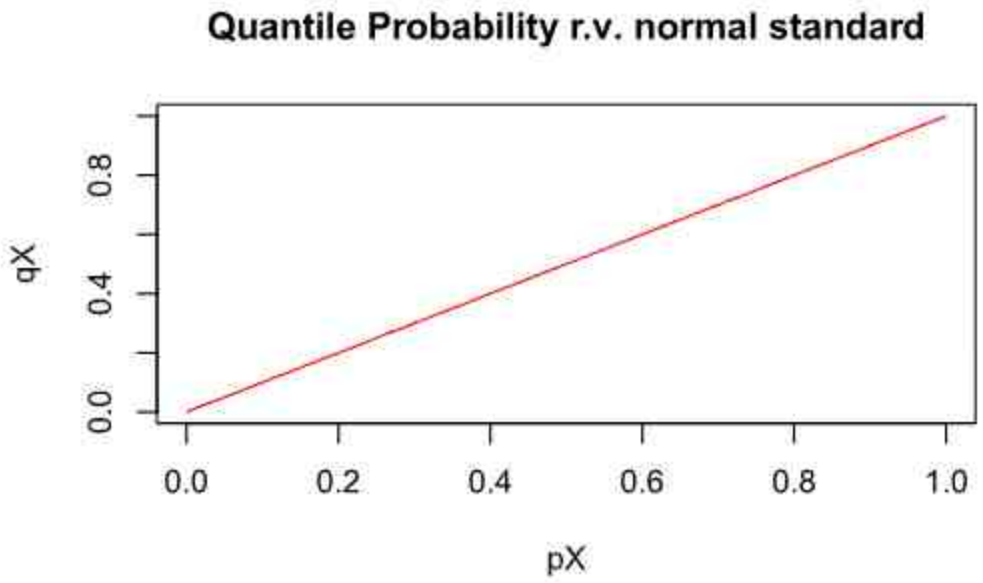
```
options(warn=-1)
x   = seq(-3,3,0.01)
DX  = dnorm(x,0,1)      #Density random normal standard
plot(x,type = "l",col = "red", DX,
     main = "Density r.v. Normal Standard")
abline(v=0, type = "l",lty = 2,col = "blue")
```




```
pX  = pnorm(x,0,1)     #Probability random normal standard
plot(x,pX, type ="l", col = "red",
     main = "Probability r.v. normal standard")
```



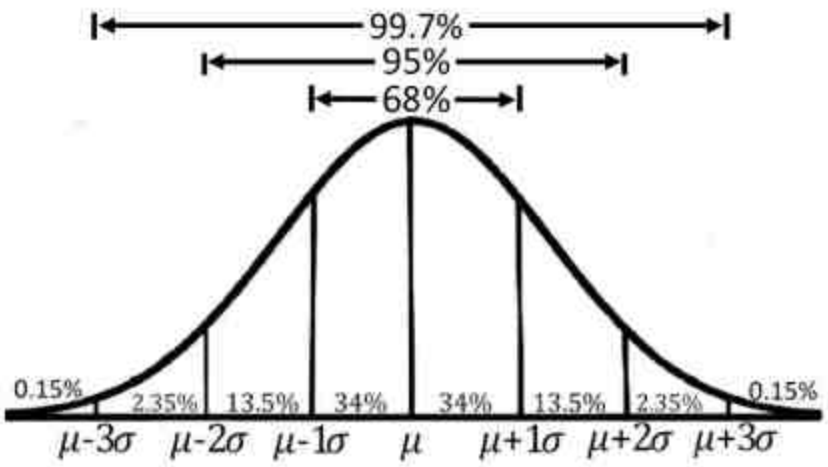
```r
qX  = qunif(pX,0,1)      #Quantile random normal standard
plot(pX,qX, type ="l", col = "red",
     main = "Quantile Probability r.v. normal standard")
```



4.3.1 Aturan Normal Empiris (*Empirical Rule Normal Distribution*)

Bila sebuah random variabel X berdistribusi normal dengan mean $= \mu$ dan standar deviasi $= \sigma^2$, aturan normal empiris menyebutkan bahwa

- probabilitas X terjadi di antara $\mu - \sigma$ hingga $\mu + \sigma$ adalah 68%
- probabilitas X terjadi di antara $\mu - 2\sigma$ hingga $\mu + 2\sigma$ adalah 95%
- probabilitas X terjadi di antara $\mu - 3\sigma$ hingga $\mu + 3\sigma$ adalah 99.7%

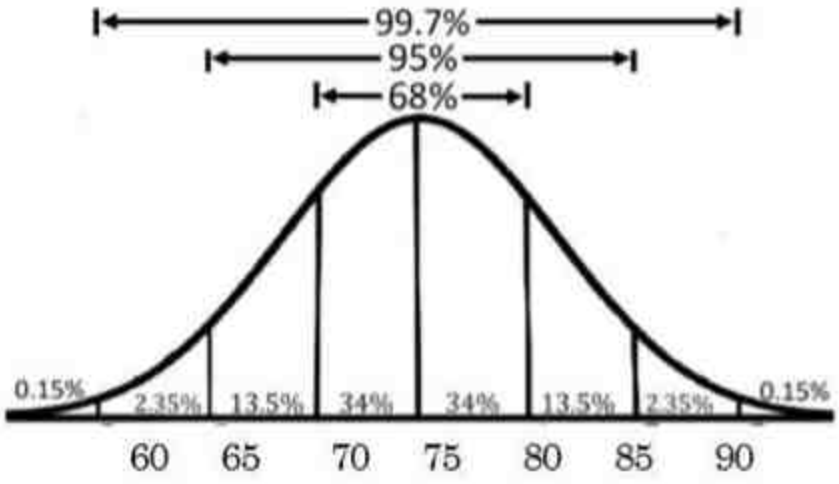


Gambar 4.4. Gambar 4.3 Aturan normal empiris

Aturan ini sering juga disebut sebagai aturan tiga-sigma atau aturan 68-95-99,7.

Contoh 4.1

Rata-rata nilai kelas statistik diasumsikan berdistribusi normal dengan mean 75 dan standar deviasi 5. Hal ini dapat diartikan, di kelas tersebut:



- 68% mahasiswa memiliki nilai statistik antara 70 hingga 80.

- Hanya 0,15% saja mahasiswa yang memiliki nilai di atas 90 atau di bawah 60
- 16% mahasiswa memiliki nilai di atas 80
- 13,5% mahasiswa memiliki nilai antara 65 hingga 70, atau 80 hingga 85.

4.3.2 Sifat-sifat Distribusi Normal

Sifat 1:

Jika $X \sim N(\mu, \sigma^2)$, maka $Z = \frac{X-\mu}{\sigma}$ akan berdistribusi Normal standar. $N(0,1)$. Sehingga

$$p(X \leq b) = p\left(\frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = p\left(Z \leq \frac{b-\mu}{\sigma}\right)$$

$$p(a \leq X \leq b) = p\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = p\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)$$

Sifat 2:

Jika $X \sim N(\mu, \sigma)$. maka $Y = a + bX \sim N(a + b\mu, |b|\sigma)$

Mengkalikan dengan sebuah konstanta a dan menambahkan dengan sebuah konstanta b pada sebuah random variabel yang berdistribusi Normal, hanya akan mengubah mean dan variance dari distribusi normal itu sendiri.

Sifat 3: Distribusi dari jumlahan dari dua random variabel independen yang berdistribusi Normal adalah Normal.

Jika $X \sim N(\mu_1, \sigma_1)$ dan $Y \sim N(\mu_2, \sigma_2)$, X dan Y saling independen, maka

$$X + Y \sim N(\mu, \sigma)$$

$\mu = \mu_1 + \mu_2$; $\sigma^2 = \sigma_1^2 + \sigma_2^2$; $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$

Sifat 4: Pendekatan Normal

Jika $X \sim bin(n, p)$ dengan nilai n sangat besar dan p dekat dengan 0 atau 1, maka variabel yang distandarkan; $Z = (X - np)/\sqrt{npq}$ akan mendekati distribusi normal standar $N(0,1)$

Tanpa koreksi kontinuitas

$$p(a \leq X \leq b) \approx p\left(\frac{a-np}{\sqrt{np(1-p)}} \leq Z \leq \frac{n-np}{\sqrt{np(1-p)}}\right)$$

Dengan koreksi kontinuitas

$$p(a \leq X \leq b) \approx p\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{n + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

4.3.3 Menghitung probabilitas dengan menggunakan R

Perintah dalam R yang dapat digunakan untuk menghitung nilai probabilitas sebuah random variabel yang berdistribusi normal dengan mean = 0, dan sd = 1 adalah

F(X=q) = pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

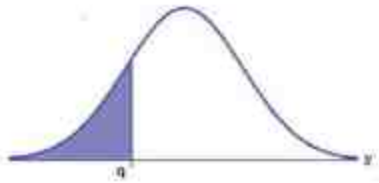
$F(X - q) = \int_{-\infty}^{q} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$

Secara default pnorm menghitung nilai integral dari kiri (lower.tail = TRUE). Bila nilai log.p =TRUE, maka kita menghitung distribusi dari log.normal.

Contoh 4.2

Lihat Contoh 4.1: X - nilai kelas statistik, mean = 75, sd = 5.

- Hitunglah probabilitas bahwa seorang mahasiswa di kelas tersebut mendapatkan nilai kurang dari 70.



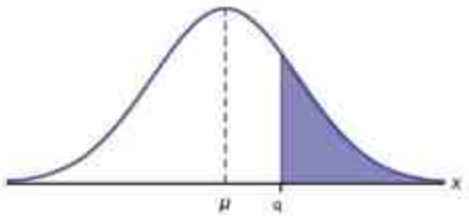
$P_1 = p(X \leq 70)$

```
P1 = pnorm(70, 75,5)
P1
```

```
[1] 0.1586553
```

Probabilitas bahwa seorang mahasiswa di kelas tersebut mendapatkan nilai kurang dari 70 adalah 0.1587

- Hitunglah probabilitas bahwa seorang mahasiswa di kelas tersebut mendapatkan nilai lebih dari 90.



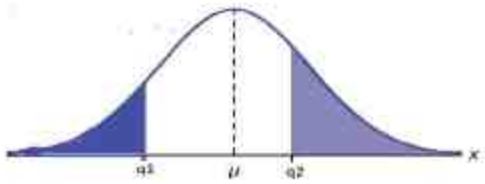
$P_2 = p(X \geq 90)$

```
P2 = 1-pnorm(90, mean = 75, sd = 5, lower.tail = TRUE)
P2
```

```
[1] 0.001349898
```

Probabilitas bahwa seorang mahasiswa di kelas tersebut mendapat nilai lebih dari 90 adalah 0.13 %. Di kelas ini, mahasiswa sulit untuk mendapatkan nilai di atas 90.

- Hitunglah probabilitas bagi seorang mahasiswa untuk mendapatkan kurang dari 70 atau nilai lebih dari 90.

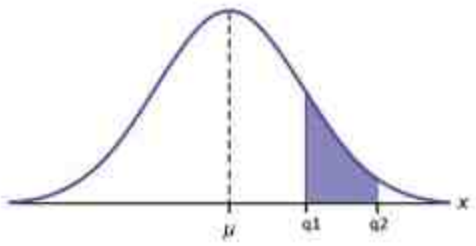


```
P3 = P1 + P2
P3
```

```
[1] 0.1600052
```

Probabilitas bahwa seorang mahasiswa di kelas tersebut mendapat nilai kurang dari 70 atau lebih dari 90 adalah 16 %.

- Hitunglah probabilitas bahwa seorang mahasiswa di kelas tersebut mendapatkan nilai antara 80 hingga 90.




```
P4 = pnorm(90,75,5)-pnorm(80,75,5)
P4
```

```
[1] 0.1573054
```

Probabilitas bahwa seorang mahasiswa di kelas tersebut mendapat nilai antara 80 hingga 90 adalah 15.73%.

Contoh 4.3 (de Veaux, et al., 2017; p.458)

Sebuah perusahaan memproduksi mini stereo sistem. Pada akhir proses produksi, stereo ini akan dikemas dan dikirim. Proses 1 yang dilakukan adalah *packing*. Pada proses ini pekerja mengumpulkan seluruh komponen (unit utama, dua speakers, kabel listik, antenna dan beberapa kabel yang lain) menjadi satu, membungkus masing-masing komponen itu dengan plastic dan kemudian meletakkannya ke dalam styrofoam form. Kemasan ini kemudian dipindahkan ke Proses 2, yaitu "boxing". Para pekerja akan membungkus kemasan dari Proses 1 kedalam kardus, memberikan kartu instruksi instalasi dan garansi, menutupnya dengan seloptip dan memberi label untuk pengiriman. Berdasarkan data yang dikumpulkan, waktu untuk Proses 1 memiliki distribusi Normal dengan mean 9 menit dan standard deviasi 1,5 menit. Proses 2 juga memiliki distribusi Normal dengan mean 6 menit dan standard deviasi 1 menit.

Pertanyaan:

a. Berapakah kemungkinan *packing* dua sistem secara berturut-turut memerlukan waktu lebih dari 20 menit?

b. Berapa percent waktu yang dibutuhkan untuk *packing* lebih lama dari waktu untuk membungkus (*boxing*)?

Jawab:

Misalkan

P_1= waktu yang dibutuhkan untuk *packing* sistem 1

P_2= waktu yang dibutuhkan untuk *packing* sistem 2

T = total waktu yang dibutukan untuk mengemas keduanya.

$T = P_1 + P_2$

Menggunakan Sifat 3 distribusi Normal,

$E(T) = E(P_1 + P_2) = E(P_1) + E(P_2) = 9 + 9 = 18$

$Var(T) = Var(P_1 + P_2)$, keduanya independent

$Var(T) = Var(P_1) + Var(P_2) = (1,5)^2 + (1.5)^2 = 4,5$ $Sd(T) = \sqrt{4.5} \approx 2.12$ menit $T \sim N(18; 4,5)$ $p(T > 20) = p\left(Z > \frac{20-18}{2,12}\right) = p(Z > 0,94) = 0,1736$

```
P5a = 1-pnorm(20,18,2.12)
P5a
```

```
[1] 0.1727391
```

atau

```
P5b = pnorm(20,18,2.12, lower.tail = FALSE)
P5b
```

```
[1] 0.1727391
```

Kemungkinan *packing* dua sistem secara berturut-turut memerlukan waktu lebih dari 20 menit adalah 17.27%

b. Misalkan

P = waktu yang dibutuhkan untuk mengemas (*packing*)

B = waktu yang dibutuhkan untuk memasukan ke dalam kardus (*boxing*)

D = selisih antara waktu untuk packing dan *boxing*

$D = P - B$

Menggunakan Sifat 3 distribusi Normal. $E(D) - E(P - B) - E(P) - E(B) - 9 - 6 - 3$

$Var(D) = Var(P - B)$, keduanya independent

$Var(D) = Var(P) + Var(B) = (1,5)^2 + (1)^2 = 3,25$ $Sd(T) = \sqrt{3,25} \approx 1,80$ menit $D \sim N(3; 3,25)$ $p(D > 0) = p\left(Z > \frac{0-3}{1,8}\right) = p(Z > -1,67) = 0,9525$

```
P6a = 1-pnorm(0,3,1.8)
P6a
```

```
[1] 0.9522096
```

atau

```
P6b = pnorm(0,3,1.8, lower.tail = FALSE)
P6b
```

```
[1] 0.9522096
```

Kemungkinan waktu yang dibutuhkan untuk *packing* lebih lama dari waktu untuk membungkus (*boxing*) adalah 95.22%

Contoh 4.4

Pada saat pandemic, kebutuhan plasma convalesce sangatlah mendesak dan menjadi harapan bagi orang yant terpapar Covid-19. Misalkan, saat itu Palang Merah Indonesia (PMI) cabang Surabaya mengantisipasi atas kebutuhan plasma ini sekurang-kurang nya untuk 1850 pasien. Sulitnya mendapatkan plasma ini, membuat PMI Surabaya untuk mengumpulkan sebanyak mungkin pendonor. Andaikan saat itu terkumpul 32.000 pendonor. Berapa kemungkinan PMI Surabaya tidak mampu memenuhi kebutuhan plasma tersebut?

Jawab.

Bila masalah ini diselesaikan model Binomial dengan $n = 32.000$ dan $p = \frac{1850}{32000} = 0,06$; maka perhitungan nilai probabilitasnya akan sulit. Untuk itu kita dapat menggunakan pendekatan model Normal dengan mean $np = 1920$ dan standard deviasi $\sqrt{npq} \approx 42,48$.

$$p(X < 1850) = p\left(Z < \frac{1850 - 1920}{42,48}\right) \approx p(Z < -1,65) \approx 0,05$$

```
P7 = pnorm(1850,1920,42.48)
P7
```

```
[1] 0.04969334
```

Kemungkinan PMI Surabaya tidak mampu memenuhi kebutuhan plasma tersebut adalah 4.9 %

References

De Veaux, R., Velleman, P., and Bock D., (2016), *Stats: Data and Models*, 5th Eds. Pearson

J.A. Rice, *Mathematical Statistics & Data Analysis*, (2006), Duxbury Press

Kerns. G. J.. (2010), *Introduction to Probability and Statistics using R*, GNU Free documentation Licence.

Montgomery, D.C., and Runger, G.C., (2018) *Applied Statistics and Probability for Engineers*, 7th Eds., Wiley, USA.

5 Inferensia

5.1 Populasi dan sampel

Populasi adalah suatu kumpulan subyek, variabel, konsep atau fenomena (Glosary of statistical term, 2023). Menurut Howel (2012: 7), populasi adalah kumpulan dan peristiwa dimana kita tertarik untuk meneliti peristiwa atau fenomena tersebut. Misalkan kita ingin meneliti tentang demografi dari penduduk Surabaya, maka populasi dari penelitian ini adalah seluruh penduduk Surabaya. Bila kita hanya ingin meneliti demografi dari sebuah kelurahan. maka populasi yang akan kita teliti adalah seluruh penduduk di kelurahan tersebut.

Seringkali kita ingin mengetahui karakteristik yang terdapat dalam suatu populasi. Namun demikian mengumpulkan data dari seluruh populasi terkadang tidak praktis atau bahkan tidak mungkin. Populasi selalu berubah dari waktu ke waktu. Misalkan kita mencatat demografi dari penduduk di Surabaya, bisa saja saat dicatat penduduk tersebut ada, namun beberapa saat kemudian penduduk tersebut meninggal dunia atau pindah keluar kota. Untuk itu seringkali kita akan mengambil data dari sebagian kecil dari populasi yang kita sebut sebagai sampel atau cuplikan.

Sampel yang kita cuplik dari populasi haruslah mewakili karakateristik dari populasi tersebut. Jika kita misalkan populasi kita adalah sepanci sup, kita tidak perlu makan sepanci sup untuk dapat merasakan sup tersebut, kita hanya perlu mencicipi semangkuk sup saja. Namun demikian, sebelum kita mengambil sup tersebut. kita perlu mengaduknya, agar semua bahan yang digunakan untuk membuat sup tersebut berada dalam mangkok tersebut (de Veaux, 2016).

Demikian juga dalam mengambilan sampel dari suatu populasi maka kita perlu mengambil sampel tersebut secara acak/random. Tergantung dari sifat dari suatu populasi, terdapat beberapa teknik sampling yang dapat kita lakukan, agar data sampel yang kita peroleh dapat mewakili karakteristik dari populasi yang ingin kita teliti.

5.1.1 Simple Random Sampling (SRS)

Apabila populasi yang akan kita sampling cukup homogen maka kita dapat mengambil sampel dengan cara mengacaknya secara sederhana (Simple Random Sampling). Misalkan saja kita ingin mengetahui tinggi badan siswa di sebuah kelas. Tentu saja, kita dapat mengukur semua tinggi badan siswa tersebut. Namun demikian bila kita hanya ingin mengambil sampel dari

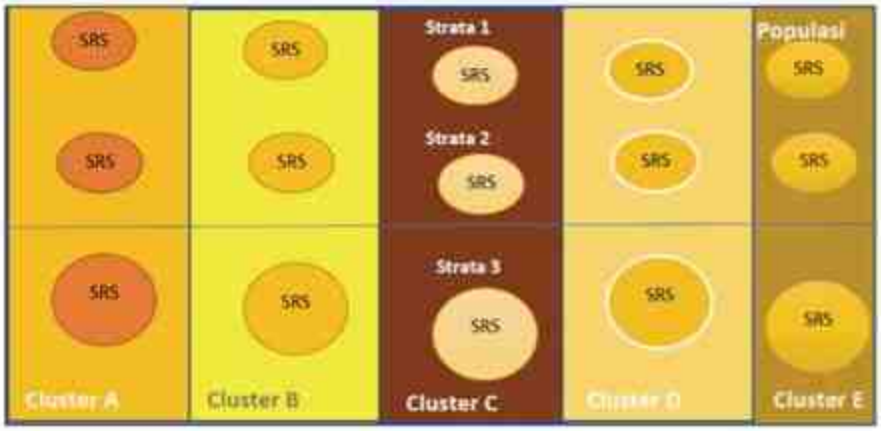
10 anak saja, kita dapat mengacak nomor induk yang terdaftar di kelas itu, dan mengukur tinggi siswa sesuai dengan nomor induk yang terpilih. Pada pengambilan sampel secara acak sederhana ini setiap anggota dari populasi memiliki kesempatan yang sama untuk terpilih.

5.1.2 Stratified Sampling

Terkadang populasi yang akan kita teliti secara keseluruhan tidak homogen dan akan memiliki sifat homogen bila populasi tersebut kita pisahkan (*slice*) menjadi kelompok-kelompok yang homogen, yang kita sebut sebagai strata. Misalkan kita ingin mengukur pendapatan sebuah kota, bila kita mengambil sampel secara acak sederhana, maka akan terdapat kemungkinan sampel yang kita peroleh tidak mewakili populasi yang kita inginkan. Bisa saja, sampel itu hanya mewakili sebagian masyarakat yang kurang mampu saja, ataupun komposisi status ekonomi yang terambil dalam sampel tidak sesuai dengan komposisi status ekonomi yang terdapat pada populasi. Untuk itu, kita dapat membagi dulu masyarakat di kota tersebut berdasarkan strata ekonomi mereka, kemudian berdasarkan strata yang sudah terbentuk kita mengambil sambel secara acak sederhana di setiap strata tersebut. Dengan demikian populasi dari pendapatan dari kota tersebut dapat terwakili dari sampel yang terpilih.

5.1.3 Cluster dan Multistage Sampling

Apabila populasi yang akan kita teliti sangatlah luas dan heterogen, seringkali kita membaginya menjadi beberapa kelas (*cluster*) untuk mempermudah pengambilan sampelnya. Katakanlah kita ingin mengukur tingkat pendapatan penduduk kota Surabaya. Pertama-tama, kita akan membagi populasi Kota Surabaya ini menjadi lima wilayah Surabaya yaitu Utara, Selatan, Timur, Barat dan Pusat. Di setiap kelas ini, kita dapat melakukan stratified sampling berdasarkan tingkat pendapatan, setelah itu melakukan simple random sampling untuk menentukan sampel yang akan kita ambil untuk mewakili nilai populasi yang kita teliti. Bila mengambilan cluster ini masih dirasa cukup besar, kita masih dapat membaginya lagi menjadi beberapa stage, misalnya per Kecamatan, kemudian per Kelurahan, sebelum melakukan stratified random sampling dan simple random sampling. Pengambilan sampel seperti ini biasa kita sebut sebagai Multistage Sampling (lihat Gambar 5.1).



Gambar 5.1. Ilustrasi dari multistage random sampling

5.2 Populasi dan sampel Parameter

Data yang dikumpulkan melalui sampling akan diwakilkan dengan nilai parameter sampel. Katakanlah dalam sebuah survei kependudukan, terdapat beberapa hal yang ingin kita ketahui tentang nilai parameter sampelnya, seperti:

- Proporsi ($\bar{p}$) antara penduduk pria dan wanita
- Rata-rata ($\bar{x}$) pendapatan per keluarga
- Standar deviasi (s) dari nilai pendapatan per keluarga
- Korelasi (r) antara jumlah tahun pendidikan yang diselesaikan dengan gaji yang diperolah

Hubungan antara gaji yang diperoleh (Y) dengan jumlah tahun pendidikan (X_1) , lama bekerja (X_2), gender (X_3) yang dinyatakan sebagi persamaan linear. Dalam hal ini kita ingin mengetahui nilai koefisien regresi (a, b_1, b_2, b_3) dari persamaan berikut:

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3$$

Dalam statistik kita membedakan antara nilai parameter sampel dan nilai parameter populasi. Nilai parameter populasi biasanya diturunkan dari model matematika, atau nilai yang diperoleh dari seluruh populasi. Nilai parameter sampel adalah nilai estimasi parameter populasi yang diperolah dari sampel data. Nilai proporsi, rata-rata, standar deviasi yang dihitung/diestimasi dari data yang diperoleh, disebut sebagai nilai parameter sampel. yaitu, proporsi sampel, rata-rata sampel, standar deviasi sampel.

Berikut adalah perbedaan antara nilai parameter populasi dan parameter sampel.

5.2.1 Proporsi Populasi dan Proporsi Sampel

Proporsi populasi adalah fraksi (bagian) dari populasi yang memiliki karakteristik tertentu. Misalkan dalam sebuah populasi terdapat 1000 orang, 457 di antaranya adalah wanita, maka fraksi dari populasi tersebut yang merupakan Wanita adalah 457 dari 1000 orang atau 0.457.

$$p = \frac{X}{N}$$

Dimana p adalah proporsi populasi, X adalah fraksi karakteristik dari populasi tersebut, dan N adalah jumlah anggota dari populasi. Bila populasi tersebut digantikan oleh sampel maka proporsi sampel dapat diestimasi sebagai

$$\hat{p} = \frac{x}{n}$$

dimana $\hat{p}$ adalah proporsi sampel, x adalah jumlah karakteristik tertentu dari sampel tersebut, dan n adalah jumlah anggota dari sampel.

5.2.2 Mean Populasi dan Mean Sampel

Mean Populasi

Mean populasi adalah nilai ekspektasi yang dihitung berdasarkan PMF jika sebuah random variabel X berdistribusi diskrit dan PDF jika X berdistribusi kontinu. Mean populasi didefinisikan sebagai:

$$\mu = EX = \sum_{x \in S_X} x f_X(x)$$

bila X adalah random variabel diskrit, S_X adalah support dari X dan $f_X(X)$ adalah PMF dari X.

$$\mu = EX = \int_{x \in S_X} x f_X(x) dx$$

bila X adalah random variabel kontinu, $f_X(x)$ adalah PDF dari X.

Mean Sampel

Nilai mean populasi ini diestimasi dari data sampel sebagai nilai rata-rata sampel

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

5.2.3 varians Populasi dan varians sampel

varians Populasi dapat dituliskan sebagai:

$$\sigma^2 = Var(X) = E(X-\mu)^2 = \sum_{x \in S_X} (x-\mu)^2 f_X(x)$$

bila X adalah random variabel diskrit, dan

$$\sigma^2 = Var(X) = E(X-\mu)^2 = \int_{x \in S_X} (x-\mu)^2 f_X(x) dx$$

bila X adalah random variabel kontinu, $\sigma = \sqrt{\sigma^2}$

varians sampel merupakan nilai estimasi dari varians populasi, dan dapat dituliskan sebagai:

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standar deviasi sampel dapat dituliskan sebagai $S = \sqrt{S^2}$

5.2.4 Kovarians dan Korelasi Populasi

Bila X dan Y adakah dua random variabel diskrit, untuk mengukur dependensi antara kedua random variabel tersebut dapat diukur dengan menggunakan

$$\gamma = Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y)$$

$$= \sum_{x \in S_X} (X-\mu_X)(Y-\mu_y) f_{XY}(x,y)$$

dimana, μ_X adalah mean dari random variabel X, μ_Y adalah mean dari random variabel Y, $f_{XY}(X,Y)$ adalah PMF gabungan antara random variabel X dan Y.

Bila X dan Y adakah dua random variabel kontinu, maka

$$\gamma = Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y)$$

$$= \int_{x \in S_X} (X - \mu_X)(Y - \mu_y) f_{XY}(x, y) dx$$

dimana $f_{XY}(x, y)$ adalah PDF gabungan antara random variabel X dan Y. Covarians populasi yang diestimasi dari data sampel disebut sebagai covarians sampel.

$$\hat{\gamma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Kedua random variabel dikatakan independen bila nilai $Cov(X, Y) = 0$. Namun demikian nilai covarians masih mengandung unit pengukuran. Misalkan X adalah tinggi badan (m), dan Y adalah berat badan (Kg), maka nilai $Cov(X, Y)$ masih mengandung unit pengukuran Kg. m. Selain itu $Cov(X, Y)$ memiliki range nilai yang besar: $-\infty < Cov(X, Y) < \infty$. Untuk menyederhanakan kedua hal ini maka kita dapat menggunakan nilai korelasi antara dua random variabel X dan Y untuk mengukur independensi antara keduanya.

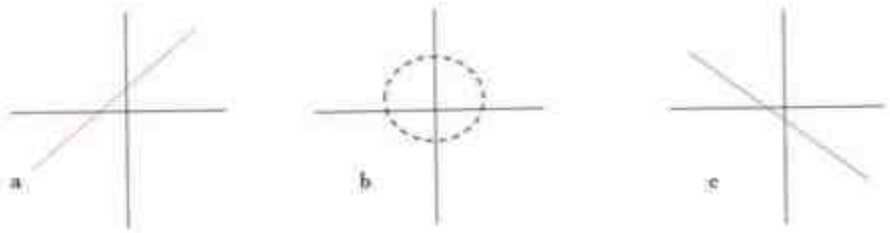
Korelasi populasi antara X dan Y didefinisikan sebagai:

$$\rho = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Korelasi sampel antara X dan Y didefinisikan sebagai

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Korelasi tidak memiliki unit pengukuran lagi dan nilainya berkisar antara -1 dan 1, $-1 < r < 1$. Bila nilai $r = 0$ maka random variabel X dan Y dikatakan independen (Gambar 5.2b). Nilai negative dan positive pada nilai korelasi r menandakan arah hubungan antara X dan Y. Bila nilai $r > 0$, maka X dan Y memiliki hubungan yang searah, artinya bila nilai X semakin meningkat, maka nilai Y juga meningkat (Gambar 5.2a). Namun, bila $r < 0$, maka X dan Y memiliki hubungan yang berlawanan arah, semakin tinggi nilai Xmaka nilai Y akan semakin rendah dan sebaliknya (Gambar 5.2 c). Besaran nilai korelasi, $|r|$menandakan kuat lemahnya hubungan tersebut. Secara kasar, bila $|r| \geq 0.5$ maka korelasi tersebut dikatakan kuat, dan sebaliknya.



Gambar 5.2. Korelasi positif (a), tidak berkorelasi (b), korelasi negative (c)

Semua nilai parameter yang diestimasi dari data kita sebut sebagai statistik. Beberapa contoh statistik diberikan pada Tabel 5.1.

Tabel 5.1. Table 5.1 Beberapa statistik

	Statistik	Parameter
Proporsi	$\hat{p}$	p
Mean	$\bar{X}$	μ (mu)
standar deviasi	S	σ (sigma)
Korelasi	r	ρ (rho)
Koefisien regresi	b	β (beta)

Contoh 5.1:

Cortez dan Silva (2008), menggunakan data mining untuk memprediksi nilai dari siswa SMP di Portugal. Data terdiri dari 33 variabel yang diukur. Diantara variabel tersebut, kita akan menggunakannya untuk menghitung nilai parameter populasi dari gender (sex), nilai parameter mean, standar deviasi dan varians dari umur (age) dan nilai korelasi dari jumlah ketidakhadiran (absences) dan nilai yang diperoleh (G1, G2, G3). Berikut adalah r-script dari contoh ini:

```
Dir = getwd()
setwd(Dir)
data      = read.csv("student-mat.csv") #membaca data
gender    = data$sex            #membaca data gender
age       = data$age            #membaca data umur
absences  = data$absences   #membaca data jumlah ketidakhadiran
G1          = data$G1             #Nilai G1
G2          = data$G2             #Nilai G2
G3          = data$G3             #Nilai G3
```

Untuk mengestimasi proporsi dari jumlah siswa ("Male") dan jumlah siswi ("Female"), kita dapat menggunakan persamaan (5.2), dimana karakteristik x di sini adalah jumlah siswa jika kita ingin menghitung proporsi siswa, atau x adalah jumlah siswi jika kita ingin menghitung proporsi sampel dari siswi.

```
#Anggota dari himpunan gender yang merupakan siswi
Female   = which(gender == "Female")
#Anggota dari himpunan gender yang merupakan siswa
Male       = which(gender == "Male")

PropFemale = length(Female)/length(gender)  #Proporsi sampel dari siswi
PropMale   = length(Male)/length(gender)    #Proporsi sampel dari siswa
```

Pada data di atas nilai proporsi siswi adalah 0.5265823 dan siswa adalah 0.4734177

R telah menyediakan fungsi untuk mengestimasi nilai sampel mean, sampel varians dan sampel standar deviasi. Untuk memperoleh nilai sampel mean, varians dan standar deviasi dari umur seluruh siswa kita dapat menggunakan perintah sebagai berikut:

```
MeanAge = mean(age)      #Rata-rata umur
VarAge  = var(age)       #varians umur
SdAge   = sd(age)            #Standar deviasi umur
```

Nilai sampel mean = 16.6962025, sampel varians = 1.628285. dan sampel standar deviasi = 1.2760427

Demikian juga nilai nilai sampel korelasi. Kita dapat menggunakan fungsi $cor(X, Y)$ dalam R untuk mendapatkan nilai estimasi dari sampel korelasi. Nilai sampel korelasi antara jumlah ketidakhadiran terhadap Grade 1 adalah

```
cor(absences,G1)
```

```
[1] -0.0310029
```

Pada contoh ini didapat nilai korelasi antara jumlah ketidakhadiran dan Grade 1 adalah negative. Hal ini menandakan bahwa semakin besar jumlah ketidakhadiran, maka semakin rendah nilai G1 yang diperoleh. Namun, nilai korelasi ini hanya 0,03. Hal ini menunjukkan bahwa korelasi antara ketidakhadiran dan nilai G1 lemah.

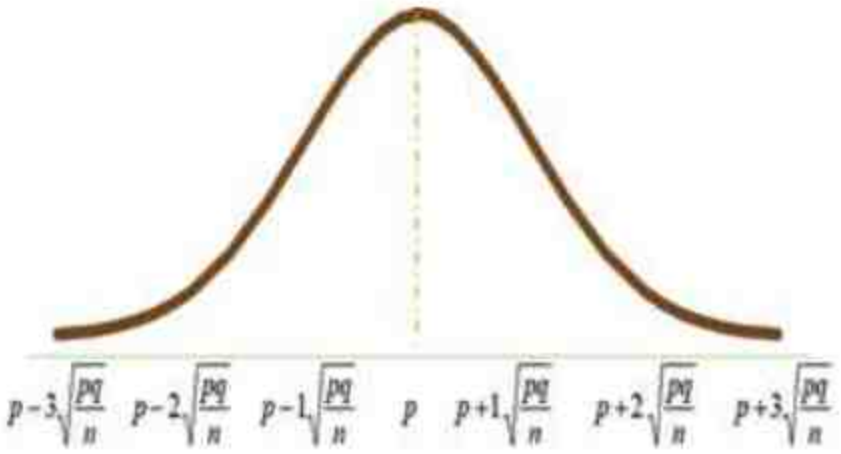
5.3 Model Sampling Distribusi

5.3.1 Model Sampling Distribusi dari Proporsi

Ketika kita melakukan sebuah survei, walaupun survei itu dilakukan pada saat yang sama, oleh sebuah lembaga yang sama, dan dengan pertanyaan-pertanyaan yang sama, hasil yang kita peroleh selalu bervariasi. Hal ini adalah wajar, karena setiap survei selalu dilakukan pada sampel yang berbeda.

Katakanlah hasil survei yang kita peroleh, kita menghitung nilai proporsi. Nilai proporsi ini bervariasi dari sampel ke sampel. Hal ini dikarenakan setiap sampel terdiri dari orang-orang yang berbeda. Bila pengambilan sampel ini dilakukan berulang-kali, maka nilai proporsi yang kita peroleh dari sampel yang berbeda-beda ini akan membentuk distribusi.

Bila pengambilan sampel ini dilakukan secara independen dan ukuran sampel cukup besar, maka distribusi sampling dari proporsi sampel $\hat{p}$ dapat dimodelkan sebagai normal model dengan mean $\mu(\hat{p})$ dan standar deviasi $Sd(\hat{p}) = \sqrt{\frac{pq}{n}}$ (lihat Gambar 5.3)



Gambar 5.3. Distribusi sampling proporsi

$\hat{p} = \frac{y}{n}$; $\hat{q} = 1 - \hat{p}$; y adalah jumlah sukses dan n adalah jumlah sampel.

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

p adalah proporsi sukses populasi

q adalah proporsi gagal populasi

Contoh 5.2

Katakanlah kita akan mengadakan survei pada sebuah populasi yang jumlahnya sangat besar. Setiap kali kita melakukan survei ini kita hanya mengambil sampel sebanyak 100 saja. Dari sampel tersebut kita ingin mengetahui proporsi sampel pria (=1) terhadap wanita (=0). Andaikan kita mengetahui proposi populasi pria adalah 0,6 dan kita melakukan survei terhadap 100 orang ini diulang-2 hingga 1000 kali, maka simulasi dari survei ini dapat kita tuliskan dalam R-script sebagai berikut:

```
set.seed(12345)
L  = 1000       #Jumlah pengulangan
n  = 100            #Jumlah sampel
pr = 0.6            #Proporsi populasi

#Simulasi 1000 kali dari sampel data dengan output (1 = pria, 0 = wanita)
p = c()
for(i in 1:L)
{ X = sample(c(0,1), replace = TRUE, size = n, prob = c(1-pr,pr))   #sampling
  p = append(p,sum(X)/n)    #proporsi sampel
}

mean_p = mean(p)
sd_p   = sd(p)
sd_p0  = sqrt(pr*(1-pr)/n)
mean_p; sd_p0;sd_p
```

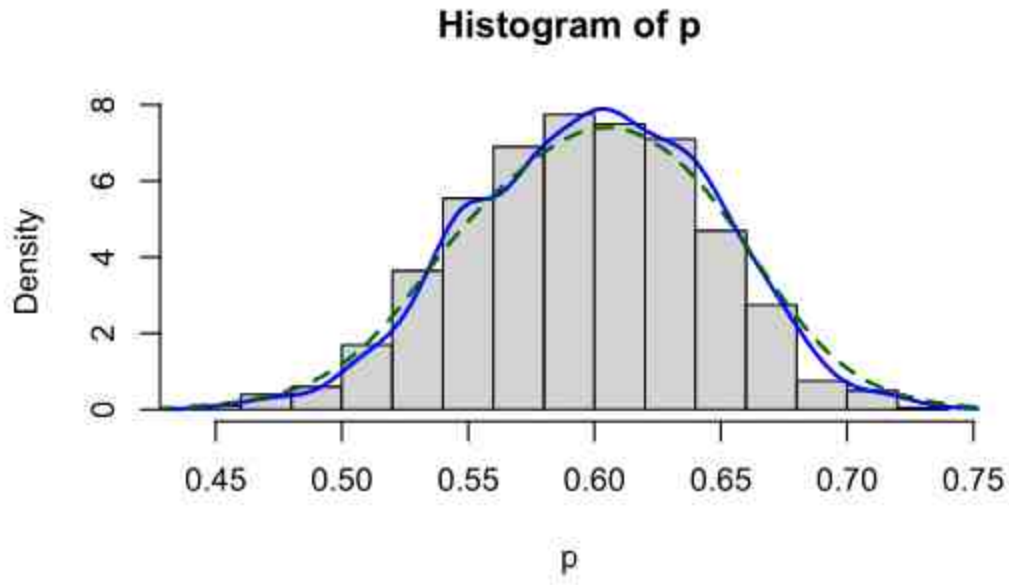
```
[1] 0.59979
```

```
[1] 0.04898979
```

```
[1] 0.04742262
```

Hitogram dari simulasi ini menunjukkan bahwa proporsi sampel bila diulang-ulang, akan memiliki distribusi normal dengan nilai mean 0.59979. Nilai proporsi mean ini mendekati nilai mean dari proporsi populasi yaitu 0.6 dan nilai standar deviasi dari proporsi sampelnya, 0.0474226 mendekati nilai proporsi populasi yaitu 0.0489898

```
hist(p, prob = TRUE)
lines(density(p), col= "blue",lwd = 2)
lines(density(p, adjust=2), lty = "dotted", col = "darkgreen", lwd = 2)
```



Terlihat histogram dari proporsi sampel ini mengikuti distribusi normal dengan mean = 0.6 dan standar deviasi = 0.0489898 .

5.3.2 Central Limit Theorem (CLT) - Teorema Limit Pusat

Teorema limit pusat dicetuskan oleh De Moivre (1811), teorema ini mengatakan:

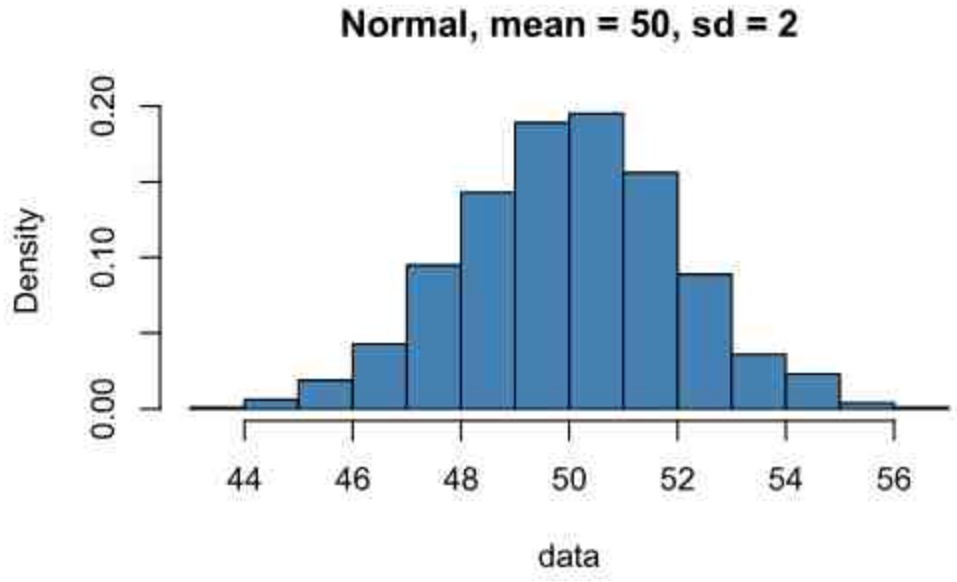
Jika ukuran sampel n, meningkat, maka rata-rata n independen random variabel, $X_1, X_2, ..., X_n$ akan berdistribusi normal dengan mean $\mu(\bar{x})$ dan standar deviasi $\sigma(\bar{x}) = SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

CLT memiliki dua asumsi yaitu sampling dilakukan secara acak dan independen.

Contoh 5.3

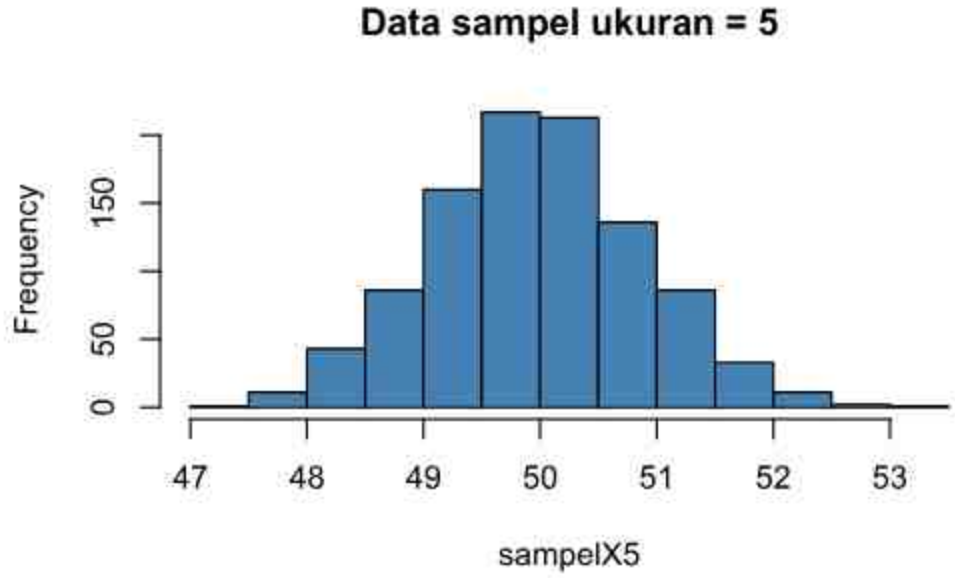
Sebuah mesin pemotong plastik yang sudah memiliki presisi yang kurang baik. Panjang rata-rata plastik yang dihasil oleh mesin ini adalah 50mm dengan standar deviasi 2 mm. Panjang plastik ini diasumsikan mengikuti distribusi normal. Seorang mahasiswa yang sedang kerja praktek di perusahaan itu melakukan sampling produksi. Sampling pertama berukuran 5 plastik, sampling kedua berukuran 30 plastik. Pada setiap sampling dia mengukur mean sampel dan mengulang pengambilan sampel ini masing-masing sebanyak 1000 kali. Kemudian dia menggambarkan distribusi dari populasi, mean distribusi mean dari sampling pertama, dan mean distribusi dari sampling kedua. Berikut adalah R dari hasil kerja praktek tersebut.

```r
# Random variabel dengan distribusi normal yang merepresentasikan
# populasi dari panjang plastic dari hasil mesin
# pemotongan plastik yang sudah tua.

N    = 1000
data = rnorm(N, mean= 50, sd = 2)

#Histogram dari populasi
hist (data, col = 'steelblue', freq = FALSE, main = 'Normal, mean = 50, sd = 2')
```




```r
#Fungsi sampling
sampling = function(n,N=1000,Data=data)
{ sampel_n = c()
  for(i in 1:N)
    sampel_n[i] = mean(sample(Data,n,replace = TRUE))
  return(sampel_n)
}

#Data sampling dengan ukuran n = 5
n = 5
sampelX5 = sampling(n)
```

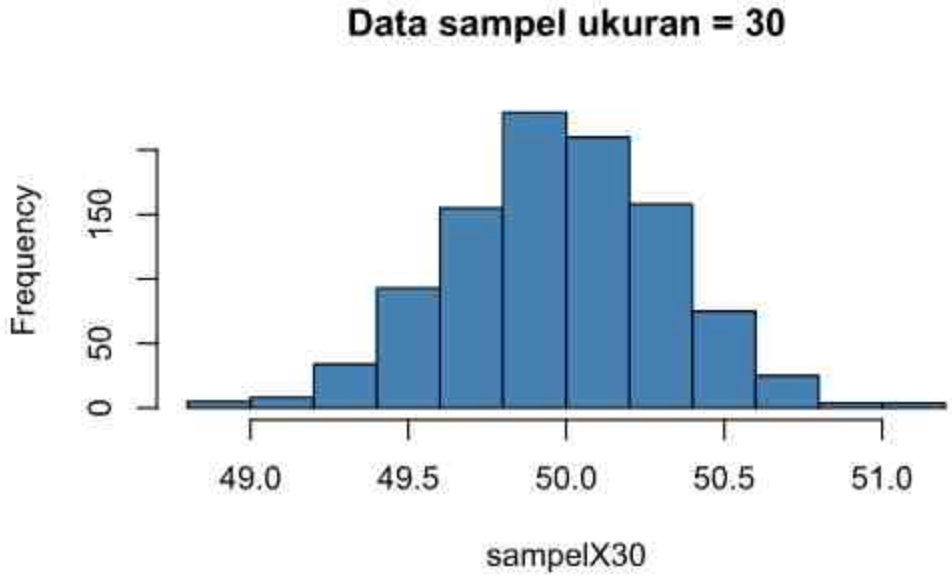
```r
#Menghitung mean dan standar deviasi dari sampel means
Mean5 = mean(sampelX5)
SD5 = sd(sampelX5)

hist(sampelX5, col = 'steelblue', main = 'Data sampel ukuran = 5')
```



Pada sampling dengan ukuran sampel 5 yang diulang sebanyak 1000 kali ini, histogram dari sampel mean mengikuti distribusi normal juga. Mean dari sampel mean (49.9770998) mendekati mean dari populasi yaitu 50, sedangkan standar deviasinya (0.8988228)mendekati nilai CLT de Moivre yaitu $\frac{2}{\sqrt{5}} = 0,894$. Selanjutkan akan dihitung sampel mean dengan ukuran sampel 30, sebagai berikut:

```r
#Data sampling dengan ukuran n = 30
n = 30

#Fungsi sampling
sampling = function(n,N=1000,Data=data)
{ sampel_n = c()
  for(i in 1:N)
    sampel_n[i] = mean(sample(Data,n,replace = TRUE))
  return(sampel_n)
}
```

```
sampelX30 = sampling(n)

#Menghitung mean dan standar deviasi dari sampel means
Mean30 = mean(sampelX30)
SD30 = sd(sampelX30)
hist(sampelX30, col = 'steelblue', main = 'Data sampel ukuran = 30')
```



Terlihat bahwa mean dari sampel mean (49.9782544) mendekati nilai mean populasi yaitu 50, dan standar deviasi dari sampel mean mendekati nilai CLT de Moivre yaitu $\frac{2}{\sqrt{30}} = 0,365$. Histogram dari sampel mean mengikuti distribusi normal.

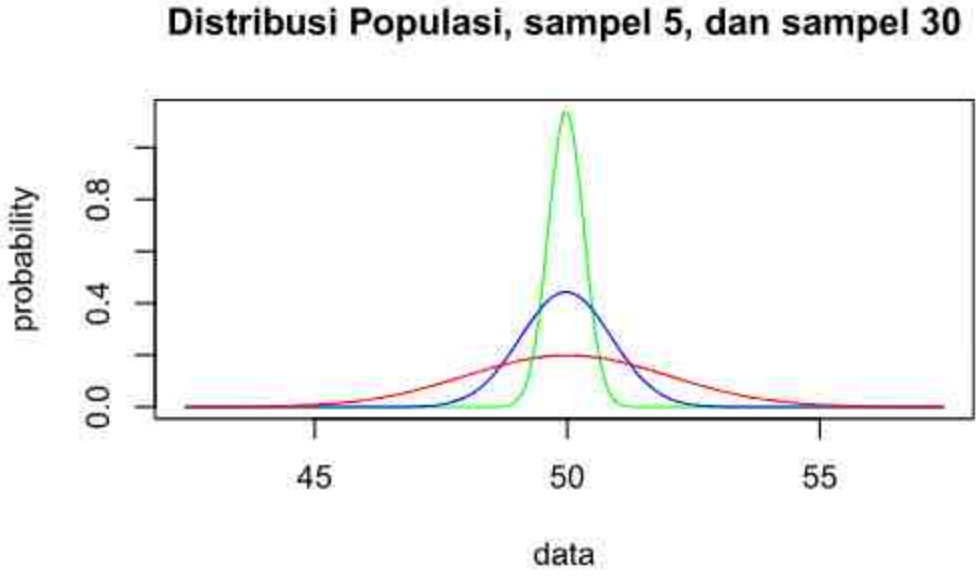
Perbandingan dari ketiga histogram di atas dapat dituliskan dalam R-script sebagai berikut. Terlihat bahwa mean dari ketiga distribusi mendekati nilai 50 (mean populasi), sedangkan standar deviasinya berbeda. Semakin besar ukuran sampel, maka standar deviasi dari sampel mean akan semakin kecil dan distribusinya semakin runcing.

```
#Menggambarkan ketiga distribusi dalam satu grafik.
i     = seq(min(data)-1, max(data)+1,by = 0.1)
di1   = dunif(i,min = 2, max = 6)

di1   = dnorm(i, 50,2)
di2   = dnorm(i,mean(sampelX5), sd(sampelX5))
di3   = dnorm(i,mean(sampelX30), sd(sampelX30))
```

```
plot(i,di3, type = "n", xlab = "data", ylab = "probability")
lines(i,di3, col = 'green')
lines(i,di2, col = 'blue')
lines(i,di1, col = 'red', freq = FALSE)
```

```
Warning in plot.xy(xy.coords(x, y), type = type, ...): "freq" is not a graphical
parameter
```

```
legend(1,c("Populasi", "sampel 5", "sampel 30"),pch = '-',
       col = c('red','green','blue'))
title('Distribusi Populasi, sampel 5, dan sampel 30')
```

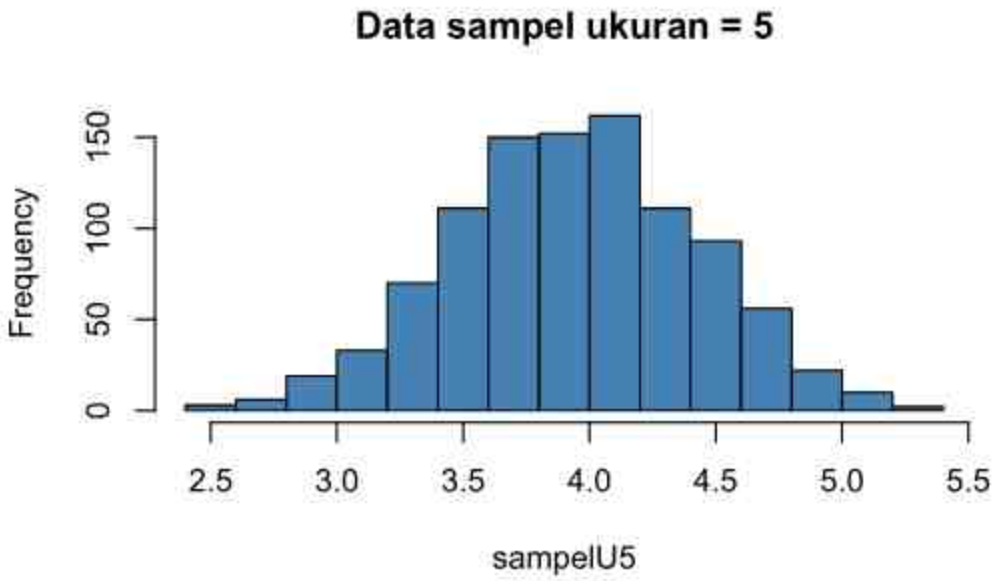


Contoh 5.4 Sebuah tanaman memiliki tinggi yang tidak rata. Tinggi tanaman tersebut mengikuti distribusi Uniform dengan min = 2 cm dan max = 6 cm. Dilakukan sampling pengukuran pada tanaman tersebut di sebuah ladang dengan ukuran sampel 5 dan 30. Seperti pada contoh di atas sampling ini diulang sebanyak 1000 kali, dan setiap kali dihitung sampel mean nya. Dengan menggunakan R-script yang mirip, dan mengubah data menjadi:

```
N = 1000
data = runif(N, min = 2, max = 6)
MeanUnif = mean(data)
Sdunif   = sd(data)
```

Mean dan standar deviasi dari data tersebut adalah Mean = 3.9336271; standar deviasi = 1.1383097

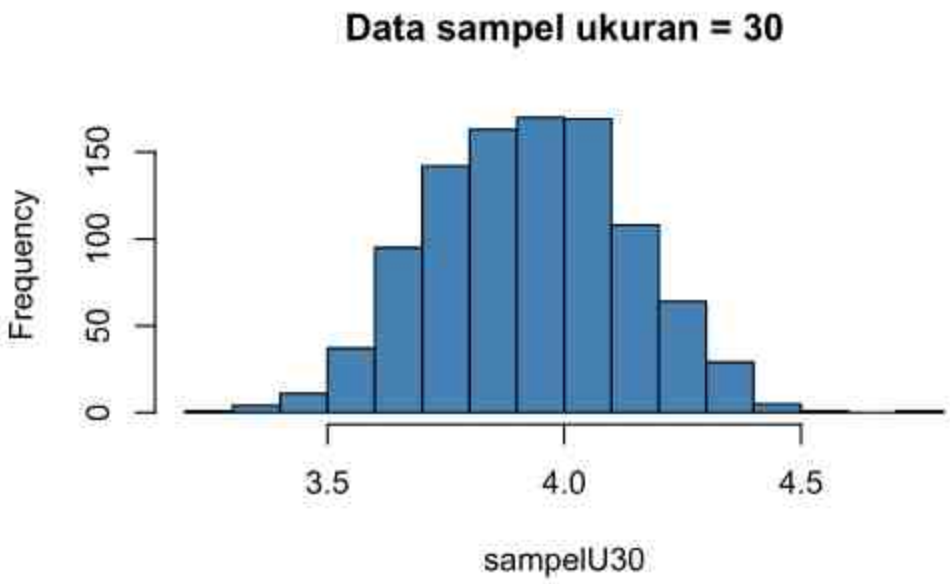
```
n = 5
sampelU5 = sampling(n)
MeanU5 = mean(sampelU5)
sdU5 = sd(sampelU5)
hist(sampelU5, col = 'steelblue', main = 'Data sampel ukuran = 5')
```



Apabila kita melalukan sampling dengan ukuran sampel $n = 5$, maka mean dan standar deviasi dari sampel mean dengan ukuran sampel 5 = 3.938423, dengan standar deviasi = 0.4840718. Nilai mean ini hampir sama seperti pada data populasi yaitu 3.9336271 dan standar deviasinya mengikuti nilai CLT dari de Moivre yaitu: $\frac{\sigma}{\sqrt{n}} = 0.5090676$

Demikian juga dengan mean dan standar devisi dari sampel mean dengan ukuran sampel 30.

```
n = 30
sampelU30 = sampling(n)
MeanU30 = mean(sampelU30)
sdU30 = sd(sampelU30)
hist(sampelU30, col = 'steelblue', main = 'Data sampel ukuran = 30')
```



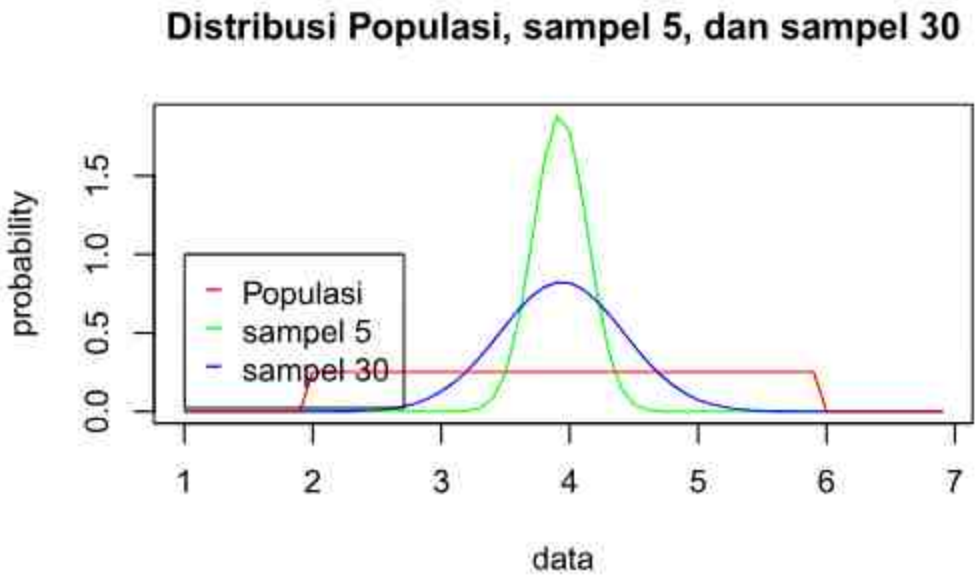
Mean dan standar deviasi dari sampel mean dengan ukuran sampel 30 = 3.9269211, dengan standar deviasi = 0.210494. Nilai mean ini hampir sama seperti pada data populasi yaitu 3.9336271 dan standar deviasinya mengikuti nilai CLT dari de Moivre yaitu: $\frac{\sigma}{\sqrt{n}} = 0.207826$

Sesuai dengan hukum Central Limit Theorem, apapun distribusi dari populasi, sampel mean akan mengikuti distribusi normal. Perbandingan dari ketiga distribusi ini dapat di lihat pada Gambar 5.4

```
i    = seq(min(data)-1, max(data)+1,by = 0.1)
di1  = dunif(i,min = 2, max = 6)

#di1  = runif(i, 50,2)
di2   = dnorm(i,MeanU5, sdU5)
di3   = dnorm(i,MeanU30,sdU30)

plot(i,di3, type = "n", xlab = "data", ylab = "probability")
lines(i,di3, col = 'green')
lines(i,di2, col = 'blue')
lines(i,di1, col = 'red', freq = FALSE)
```

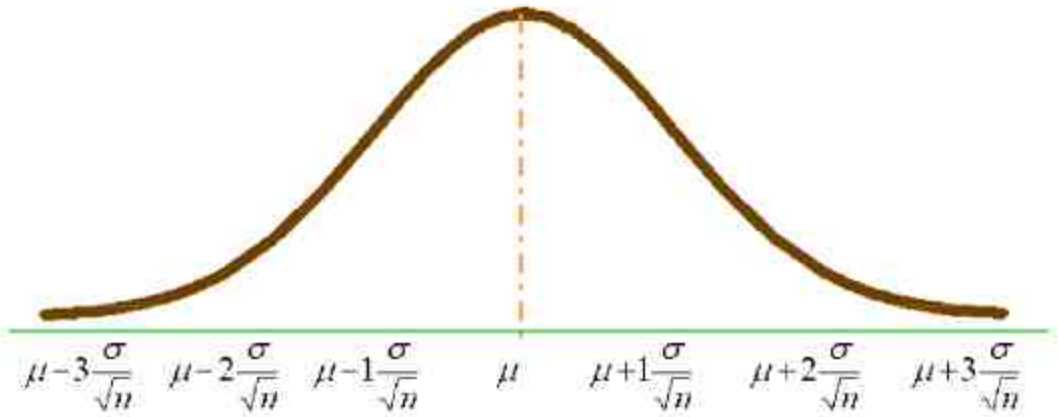
```
Warning in plot.xy(xy.coords(x, y), type = type, ...): "freq" is not a graphical
parameter
```

```
legend(1,c("Populasi", "sampel 5", "sampel 30"),pch = '-',
       col = c('red','green','blue'))
title('Distribusi Populasi, sampel 5, dan sampel 30')
```



5.3.3 Model Sampling Distribusi dari Rata-rata (sampel Mean)

Jika asumsi independen, dan mengambilan sampel dilakukan secara acak terpenuhi serta ukuran sampel n cukup besar, maka nilai rata-rata dapat dimodelkan mengikuti distribusi normal dengan mean sama dengan mean dari populasi, μ dan standar deviasi sama dengan $\frac{\sigma}{\sqrt{n}}$



Gambar 5.4. Distribusi sampling rata-rata

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

$$\hat{\sigma} = s = \frac{1}{n-1}\sum_{i=1} n(x_i - \bar{x})^2$$

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Contoh 5.4:

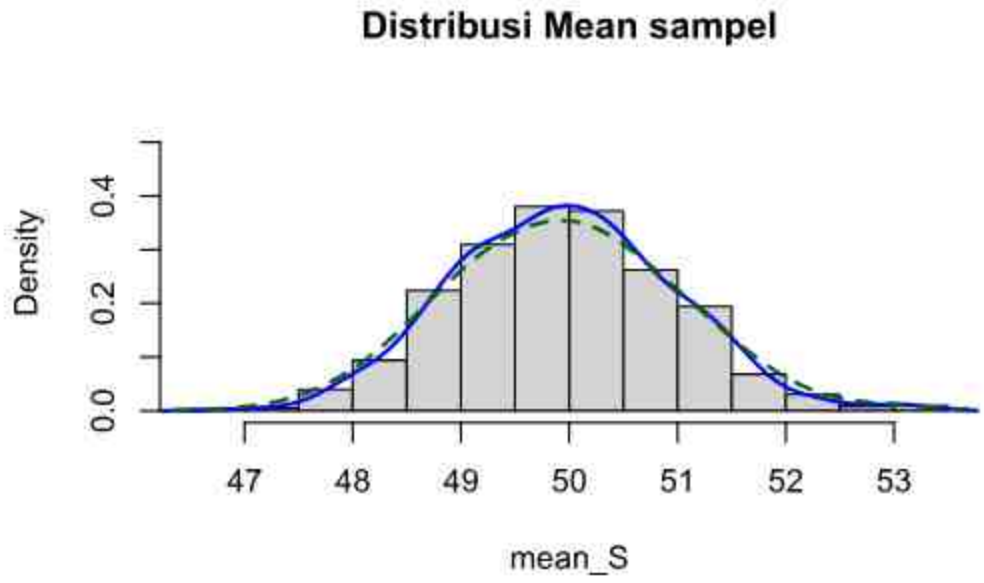
Pada survei yang sama seperti pada Contoh 5.2, tetapi kali ini kita ingin mensimulasikan distribusi dari rata-rata pengeluaran pembelian paket data internet. Andaikan diketahui, bahwa pembelian paket data memiliki distribusi normal dengan rata-rata populasi pembelian paket data internet adalah 50 ribu Rupiah, dengan standar deviasi 10 ribu Rupiah, jumlah populasi dari kota yang sedang kita survei adalah 100 ribu orang. Seperti pada survei di Contoh 5.2, setiap kali akan diambil sampel sebesar 100 orang, dan kegiatan ini diulang sebanyak 1000 kali. Simulasi ini dapat dituliskan dengan R-script sebagai berikut:

```
Pop     = 100000    # Ukuran populasi
n       = 100       # Ukuran sampel
L     = 1000        # Jumlah pengulangan pengambilan sampel
Mean    = 50          # Mean populasi pembelian paket data
SD      = 10          # Standar deviasi populasi pembelian paket data

#Simulasi pembelian paket data yang berdistribusi normal dengan
#Mean dan SD populasi
```

```
BeliPaketData = rnorm(Pop,Mean,SD)

#Simulasi pengambilan sampel, dengan ukuran sampel = 100 dan
#diulang sebanyak 1000 kali
mean_S = c()
for(i in 1:L)
{ Idx = sample(1:Pop, replace = TRUE, size = n)
  Data= BeliPaketData [Idx]
  mean_S = append(mean_S, mean(Data))
}
```

Simulasi ini menunjukkan bahwa bila perhitungan sampel mean dari suatu populasi, bila kita hitung berulang kali dari sampel yang berbeda, akan memiliki distribusi normal dengan nilai mean sampel mendekati nilai mean populasi dan standar deviasi dari mean sampel akan sama dengan standar deviasi populasi dibagi dengan akar dari ukuran sampel.

```
#Distribusi dari sampel mean
hist(mean_S, freq =FALSE, ylim = c(0,0.55),
main = "Distribusi Mean sampel")
lines(density(mean_S), col= "blue",lwd = 2)
lines(density(mean_S, adjust=2), lty = "dotted",
col = "darkgreen", lwd = 2)
```



```
#Menghitung nilai mean dan standar deviasi
#dari sampel mean
Mean_S = mean(mean_S)
SD_S   = sd(mean_S)
SD_S0  = SD/sqrt(n)
Mean_S;SD_S0;SD_S
```

```
[1] 49.94765
```

```
[1] 1
```

```
[1] 1.011826
```

5.3.4 standar Error

Apabila nilai standar deviasi dari sebuah distribusi sampling diestimasi dari data, maka nilai estimasi ini disebut sebagai standar error

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}$$

dimana

$\hat{p}$ adalah sampel proporsi sukses yang dihitung dari jumlah kesuksesan dibagi dengan jumlah sampel dan $\hat{q}$ adalah proporsi gagal

n adalah ukuran sampel

$\bar{y}$ adalah sampel mean

s adalah sampel standar deviasi

5.3.5 Distribusi t

Distribusi student-t ditemukan oleh William Sealy Gosset, distribusi ini memiliki bentuk lonceng seperti distribusi normal, namun bentuk distribusinya berubah bila ukuran sampelnya berubah. Distribusi ini memiliki satu parameter yang disebut sebagai derajat kebebasan (*degree of freedom*) yang nilainya sama dengan $n-1$ (n - jumlah sampel). $df = n-1$

Bila n semakin besar, distribusi ini akan menyerupai distribusi normal

```
#curve(dt(x, df=5), from=-4, to=4, add = TRUE, col = 'red')
#curve(dt(x, df=30), from=-4, to=4, add = TRUE, col = 'blue')
#curve(dnorm(x), from=-4, to=4, add = TRUE, col = 'green')
#legend("topleft",pch='____', c("t, df=5", "t, df = 30", "normal"),
#       col = c('red', 'blue', 'green'))
```

5.4 Confidence Interval

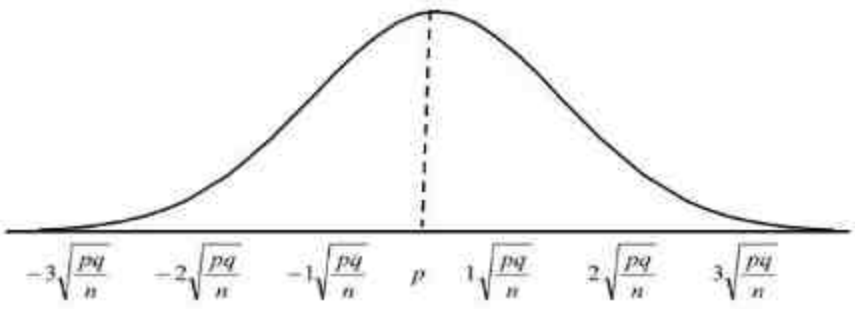
Telah kita ketahui bahwa mendapatkan seringkali parameter populasi sulit untuk didapatkan atau bahkan mustahil untuk didapatkan. Sebagai gambarannya, setiap lima tahun sekali Indonesia mengadakan Pemilihan Umum (Pemilu) Presiden. Pada saat Pemilu dilaksanakan, sering kali lembaga-lembaga survei meluncurkan hitung cepat (*quick on*) untuk meramalkan siapa pemenang dari Pemilu. Hasil hitung cepat ini diperolah dari data sampel, yang diperoleh dari beberapa tempat pemilihan suara (TPS) di berbagai wilayah di Indonesia. Pada saat hitung cepat dilaksanakan, perhitungan suara nasional tentu saja belum selelai. Hitung cepat ini adalah gambaran dari parameter sampel, sedangkan perhitungan suara nasional adalah gambaran dari parameter populasi yang baru akan diperoleh nilainya beberapa bulan setelah Pemilu usai.

Andaikan, tim *quick count* mencatat hasil perolehan suara di suatu kecamatan. Didapati di kecamatan tersebut 54 dari 104 TPS dimenangkan oleh Calon A. Pada sampel ini proporsi kemenangan Calon A adalah 51,9%. Kita tahu bahwa model distribusi sampling (lihat Gambar 5.5) untuk proporsi adalah normal dengan pusat di proporsi populasi dan standar deviasi distribusinya adalah

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}}$$

Namun demikian pada saat *quick on* ini dijalankan p , yaitu nilai proporsi populasi kemenangan Calon A maupun q, yaitu nilai proporsi kemenangan Calon B, belum diketahui. Hal yang dapat diketahui adalah standar error dari proporsi sampel $\hat{p} = 0,519$ yaitu sebesar

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0,519)(0,418)}{104}} = 0,049$$



Gambar 5.5. Model distribusi sampling proporsi

Dari aturan 68-95-99,9 pada distribusi normal dapat diperkirakan 68% dari seluruh sampel 104 TPU, akan memiliki proporsi sampel $\hat{p}$ yang berada pada range

$$\hat{p} \pm 1\ SE = 0,519 \pm 0,049$$

yaitu interval $[0,47; 0.568]$. Pada range

$$\hat{p} + 2\ SE = 0,519 + 0,098$$

dapat diperkirakan 95% dari seluruh sampel TPU sebanyak 104, akan memiliki proporsi sampel pada interval $[0.421; 0,617]$.

Dari nilai interval kita dapat menyatakan bahwa 95% kita menyakini (*confidence*) bahwa proporsi populasi p berada di dalam interval tersebut. Tentu saja ada kemungkinan sebesar 5% bahwa proporsi populasi p tidak berada di dalam interval tersebut. Interval ini kita sebut sebagai interval keyakinan (*Confidence interval*).

Selang Kepercayaan (*Confidence Interval*) Proporsi dapat didefinisikan sebagai

$$\hat{p} \pm z_{\alpha/2} SE(\hat{p})$$

Selang Kepercayaan (*Confidence Interval*) Mean dapat didefinisikan sebagai

$$\bar{x} \pm z_{\alpha/2} SE(\bar{x})$$

dimana $z_{\alpha/2}$ adalah quantile distribusi normal, α biasa disebut sebagai signifikan level (akan dijelaskan lebih lanjut di sub-bab berikutnya). $z_{\alpha/2}$ sering pula disebut sebagai nilai kritis (*critical value*)

Definisi quantile distribusi

Bila X adalah random variabel, dan $F_X : \mathbb{R} \rightarrow [0,1]$ adalah fungsi distribusi kumulatif yang bersifat kontinu dan monotonik, maka $Q : [0,1] \rightarrow \mathbb{R}$ akan memetakan nilai probabilitas p pada nilai x sedemikian hingga probabilitas random variabel $X \leq x$ akan sama dengan p.

$$F_X(x) := Pr(X \leq x) = p$$
$$Q(p) = F_X^{-1}(p) = x$$

Sebagai ilustrasi:

Andaikan sebuah random variabel X berdistribusi normal baku dengan mean nol dan varians satu; $X \sim N(0,1)$. Ingin dicari nilai x_1 dan x_2 sedemikian hingga

$$Pr(x_1 \leq x_2) = 0,95$$

Daerah yang diarsir pada Gambar 5.6 menunjukkan nilai $Pr(x_1 \leq x_2) = 0,95$; karena nilai total probabilitas adalah satu, maka nilai seluruh daerah yang tidak diarsir adalah 0,05. Daerah yang tidak diarsir ada di kiri dan di kanan, masing-masing akan memiliki luasan (probabilitas) sebesar $0,05/2 = 0,025$.

Nilai x_1 dapat dicari sebagai $Pr(X \leq x_1) = 0,025$ atau $Q(0,025) = F_X^{-1}(0,025) = x_1$ Nilai x_2 dapat dicari sebagai $Pr(X \leq x_2) = 1 - 0,025 = 0,975$ atau $Q(0,975) = F_X^{-1}(0,975) = x_2$

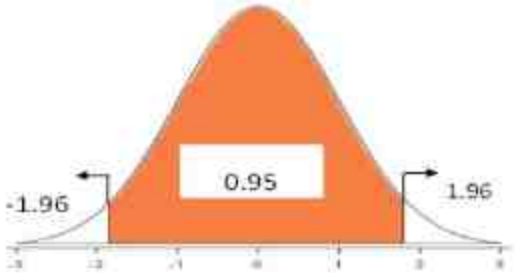
Kedua nilai tersebut dapat dihitung dengan menggunakan R sebagai berikut

```
qnorm(0.025)
```

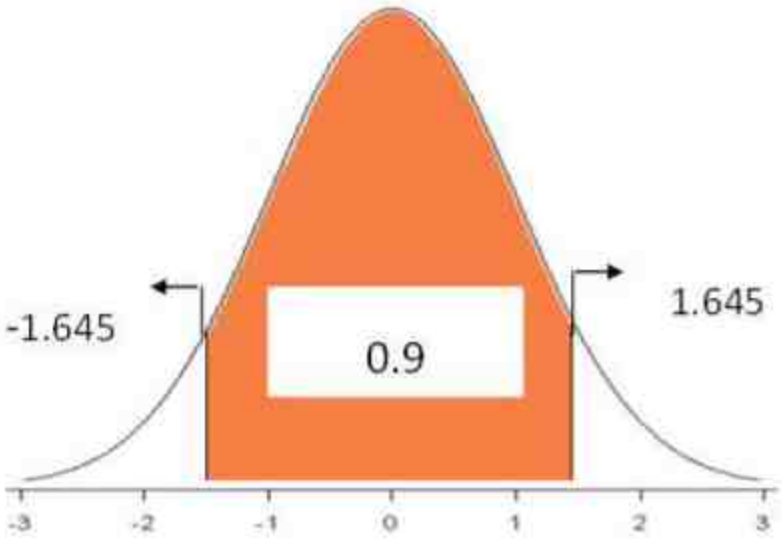
```
[1] -1.959964
```

```
qnorm(0.975)
```

```
[1] 1.959964
```



Gambar 5.6. Quantile distribusi normal dengan alpha = 5%



Gambar 5.7. Quantile distribusi normal dengan alpha = 5%

Dengan cara yang sama dapat dihitung nilai pada Gambar 5.7

$$Pr(x_1 \leq X \leq x_2) = 0,90$$

dengan menggunakan R

```
qnorm(0.05)
```

```
[1] -1.644854
```

```
qnorm(0.95)
```

```
[1] 1.644854
```

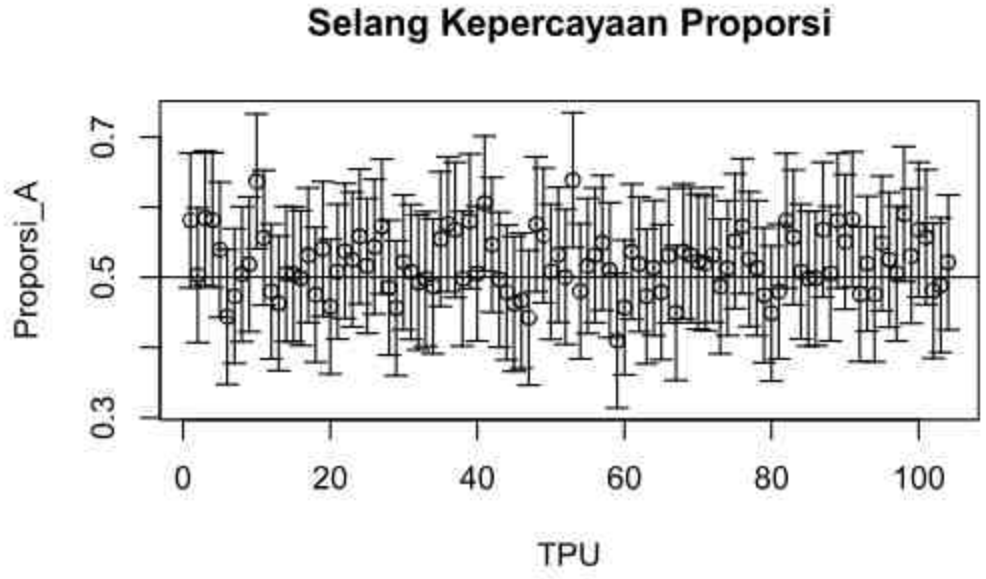
**Simulasi*

Pada kasus *Quickcount* di atas, andaikan proporsi populasi kemenangan calon A adalah $p_A = 0,519$; maka berdasarkan model sampling proporsi yang telah dibaha di subbab 5.3.1, proporsi kemenangan calom A ini akan berdistribusi normal, dengan mean $= p_A$ dan standar deviasi $= \sqrt{\frac{p_A q_A}{n}}$

Andaikan kita mensimulasikan Confidence Interval dari ke $n = 104$ TPU tersebut dengan menggunakan distribusi $N\left(p_A; \sqrt{\frac{p_A q_A}{n}}\right)$ dan signifikan level $\alpha = 0,05$; maka akan dapat dilihat bahwa tidak semua nilai proporsi populasi berada di dalam confidence interval ini. Ada kemungkinan 5% diantaranya tidak confidence interval yang terbentuk berada di luar nilai proporsi populasi. R-script berikut mensimulasikan kasus ini.

```
library("plotrix")
n    = 104
pA   = 0.519            #Probabilitas A terpilih
qA   = 1- pA            #Probabilitas A tidak terpilih
seA = sqrt(pA*qA/n)     #standar Error
alpha = 0.05            #Signifikan level

set.seed(0)
# Generate n random normal dengan mean pA, dan
# standar deviasi seA

p   = rnorm(n,pA,seA)

#Generate confidence interval
LowP = p + qnorm(alpha/2)* seA
UppP = p + qnorm(1-alpha/2)*seA

#Construct data.frame
data = data.frame(x = 1:n, y = p, lower = LowP, upper = UppP)
TPU = data$x
Proporsi_A = data$y

#Plot Confidence Internval
plotCI(x = TPU,              # plotrix plot with confidence intervals
```

```
       y = Proporsi_A,
       li = data$lower,
       ui = data$upper)
abline(h=0.5)
title("Selang Kepercayaan Proporsi")
```




```
Count = sum((LowP > 0.5 | UppP < 0.5)*1)
Count
```

```
[1] 3
```

5.5 Selang Kepercayaan (*Confidence Interval*) untuk beda (*difference*) antara dua proporsi

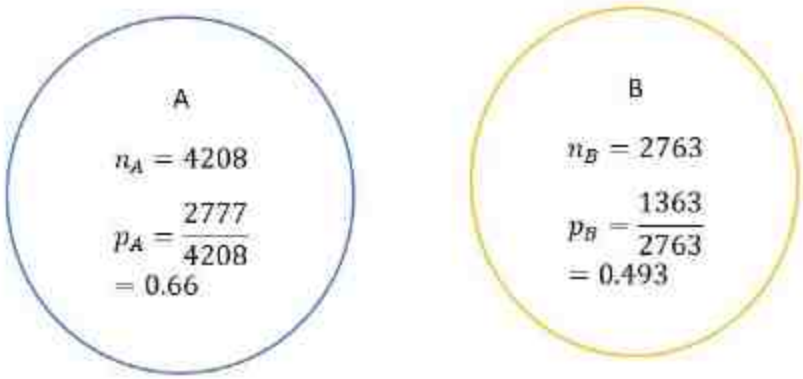
Seringkali kita membandingkan proporsi antara dua sampel dan ingin mengetahui perbedaan proporsi antara kedua sampel tersebut.

Contoh 5.5:

Sebagai contoh: Dalam sebuah studi ditemukan bahwa dari 4208 pengendara pria dengan penumpang wanita berada di mobilnya: 2777 (66%) di antaranya akan menggunakan seat

belts. Namun, di antara 2763 pengendara pria dengan penumpang pria di mobilnya, hanya 1363 (49.3%) saja yang menggunakan seat belts. Apakah yang dapat disimpulkan dari kedua perbedaan situasi ini?

Pada contoh ini kita memiliki dua sampel



A : Sampel pengendara pria dengan penumpang wanita berada di dalam mobilnya

B : Sampel pengendara pria dengan penumpang pria berada di dalam mobilnya.

5.5.1 Model sampling distribusi untuk beda antara dua proporsi

Apabila asumsi independen antara kedua sampel ini terpenuhi, maka $\hat{p}_1 - \hat{p}_2$ akan berdistribusi normal dengan standar deviasi:

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Pada sampling distribusi apabila proporsi populasi tidak diketahui maka standar deviasi bisa didekati dengan menggunakan standar error:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Selang kepercayaan dari dua proporsi ini dapat dirumuskan sebagai berikut:

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \times SE(\hat{p}_1 - \hat{p}_2)$$

Pada Contoh 5.5: $\hat{p}_1 = 0,660; \hat{q}_1 = 1 - \hat{p}_1 = 0,340$ $\hat{p}_2 = 0,493; \hat{q}_2 = 1 - p_2 = 0,507$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{(0,660)(0,340)}{4208} + \frac{(0,493)(0,507)}{2763}} = 0,012$$

Selang kepercayaan dari: $p_1 - p_2$ dengan $\alpha = 5$ adalah

$$0,164 \pm (1,96)(0,012) = 0,164 \pm 0,0236 = [0,1404; 0,1875]$$

Perbedaan ketaatan dalam menggunakan seat belt bagi Pria bila terdapat wanita dalam kendaraannya dan bila pria yang lain berada di dalam kendaraannya berada dalam selang 14,04% hingga 18,75%.

```
p1 = 0.660; q1 = 1-p1; n1 = 4208
p2 = 0.493; q2 = 1-p2; n2 = 2763

p = p1-p2
SE_p = sqrt((p1*q1)/n1 + (p2*q2)/n2)
CI_L = p + qnorm(0.025)*SE_p
CI_U = p + qnorm(0.975)*SE_p

CI_L
```

```
[1] 0.1434975
```

```
CI_U
```

```
[1] 0.1905025
```

5.6 Selang Kepercayaan (*Confidence Interval*) untuk beda (*difference*) antara dua mean

Apabila sampel data yang kita miliki adalah data numerik, maka seringkali kita juga membandingkan rerata (mean) dari dua sampel yang kita miliki.

Contoh 5.6

Dalam sebuah penelitian pasar. Seorang peniliti tertarik untuk membandingkan harga barang bekas apabila barang bekas tersebut dibeli dari teman atau dari orang asing yang sama sekali tidak dikenal. Berikut adalah table perbandingan harga yang dia peroleh

Beli dari teman	Beli dari orang asing
275	260
300	250
260	175
300	130
255	200
275	225
290	240
300	

Peneliti ini ingin mengetahui rerata perbedaan dari kedua situasi ini.

5.6.1 Model sampling distribusi untuk beda antara dua mean

Apabila asumsi independen terpenuhi $\bar{y}_1 - \bar{y}_2$ akan berdistribusi normal dengan standar deviasi

$$SD(\bar{y}_1 - \bar{y}_2) - \sqrt{(Var(\bar{y}_1) + Var(\bar{y}_2)} - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Pada sampling distribusi apabila varians populasi tidak diketahui maka standar deviasi bisa didekati dengan menggunakan standar error:

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Selang kepercayaan dari dua mean dengan distribusi Z dapat dirumuskan sebagai:

$$(\bar{y}_1 - \bar{y}_2) \pm Z_{\alpha/2} \times SE(\bar{y}_1 - \bar{y}_2)$$

Pada Contoh 5.6, diperoleh Rerata membeli barang bekas dari teman $\bar{y}_1 = 281,88$ sedangkan rerata membeli barang bekas dari orang yang tak dikenal $\bar{y}_2 = 211,43$. Perbedaan antara kedua mean: $\bar{y}_1 - \bar{y}_2 = 70,45$

Sampel varians membeli barang bekas dari teman $s_1^2 = 335,268$; Sampel varians membeli barang bekas dari asing $s_2^2 = 2155,952$;

standar error dari beda mean:

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{335,268}{8} + \frac{2155,952}{7}} = 18,71$$

Selang kepercayaan dari dua mean:

$$70,45 \pm 1,96 \times 18,71 = 70,45 \pm 36,67 = [33,78; 107,12]$$

Dari selang kepercayaan ini dapat disimpulkan bahwa membeli dari teman lebih mahal bila dibandingkan membeli dari orang asing dengan selang kepercayaan 95% dari rerata perbedaan di antara 33,78 hingga 107,12. Hal ini dikarenakan nilai nol berada di luar selang kepercayaan. Sehingga kemungkinan bahwa membeli barang bekas dari teman ataupun asing adalah sama tidak terdapat pada selang kepercayaan tersebut.

```
BeliDariTeman   = c(275,300,260,300,255,275,290,300)
BeliDariAsing   = c(260,250,175,130,200,225,240)

MeanTeman   = mean(BeliDariTeman)
MeanAsing   = mean(BeliDariAsing)
VarTeman    = var(BeliDariTeman)
VarAsing    = var(BeliDariAsing)
nTeman      = length(BeliDariTeman)
nAsing      = length(BeliDariAsing)
y           = MeanTeman - MeanAsing
SE_y        = sqrt(VarTeman/nTeman + VarAsing/nAsing)
CIy_L       = y + qnorm(0.025)*SE_y
CIy_U       = y + qnorm(0.975)*SE_y
CIy_L
```

```
[1] 33.78401
```

```
CIy_U
```

```
[1] 107.1088
```

5.6.2 Selang kepercayaan dari dua mean dengan distribusi t dapat dirumuskan sebagai:

Bila ukuran sampel kecil maka distribusi t lebih sesuai untuk digunakan dalam membangung selang kepercayaan.

$$(\bar{y}_1 - \bar{y}_2) \pm t* \times SE(\bar{y}_1 - \bar{y}_2)$$

dimana distribusi t tersebut memiliki derajat kebebasan (*degree of freedom*) sebagai berikut:

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

```
Warning: package 'knitr' was built under R version 4.1.3
```

```
    Welch Two Sample t-test

data:  Harga by Ket
t = -3.766, df = 7.6229, p-value = 0.006003
alternative hypothesis: true difference in means between group Asing and group Teman is not
95 percent confidence interval:
 -113.95597  -26.93688
sample estimates:
mean in group Asing mean in group Teman
          211.4286            281.8750
```

Dengan menggunakan distribusi t, selang kepercayaan perbedaan rerata harga barang beli dibeli dari teman dan dari asing adalah [26, 94; 113, 96]. Selang kepercayaan ini sedikit berbeda bila dihitung dengan menggunakan distribusi Z, yaitu [33, 78: 107, 11].

5.6.3 Pooling

Apabila varians dari kedua sampel ini diasumsikan sama, maka standar error difference dapat digabungkan dan disebut sebagai pooled.

standar error pada beda proporsi, apabila varians dari kedua grup ini diasumsikan sama adalah:

$$S^2_{pooled} = \frac{(n_1-1)^2 s_1^2 + (n_2-1)^2 s_2^2}{n_1+n_2}$$

$$\hat{q}_{pooled} = 1 - \hat{p}_{pooled}$$

$$SE_{pooled}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{pooled} \times \hat{q}_{pooled}}{n_1} + \frac{\hat{p}_{pooled} \times \hat{q}_{pooled}}{n_2}}$$

standar error pada beda means, apabila varians dari kedua grup ini diasumsikan sama adalah:

$$S^2_{pooled} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

$$SE_{pooled}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{S^2_{pooled}}{n_1} + \frac{S^2_{pooled}}{n_2}} = S_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Pada Contoh 5.6 di atas apabila varians diasumsikan sama maka selang kepercayaan dari beda mean yang dihitung dengan menggunakan distribusi t akan diperoleh hasil yang mendekati perhitungan yang dilakukan dengan menggunakan pendekatan distribusi Z dengan asumsi varians beda yaitu [32,11; 108, 78]. Catatan: Pooling ini akan memberikan hasil yang sesuai apabila asumsi varians sama benar-benar terpenuhi.

```
options(width = 20)
t.test(Harga~Ket, var.equal = TRUE)
```

```
    Two Sample
    t-test

data:  Harga by Ket
t = -3.9699, df =
13, p-value =
0.0016
alternative hypothesis: true difference in means between group Asing and group Teman is not
95 percent confidence interval:
 -108.78238  -32.11047
sample estimates:
mean in group Asing
           211.4286
mean in group Teman
           281.8750
```

References

Glossary of statistical terms: Population. Statistics.com. Diunduh 11 November 2023

D. Howell, *Statistical Methods for Psychology*, Wadsworth Publishing Company: International ed of edition, 2012

De Veaux, R., Velleman, P., and Bock D., (2016), *Stats: Data and Models*, 5th Eds. Pearson

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. Available at: https://www.kaggle.com/ishandutta/student-performance-data-set

6 Uji Hipotesa

6.1 Pendahuluan

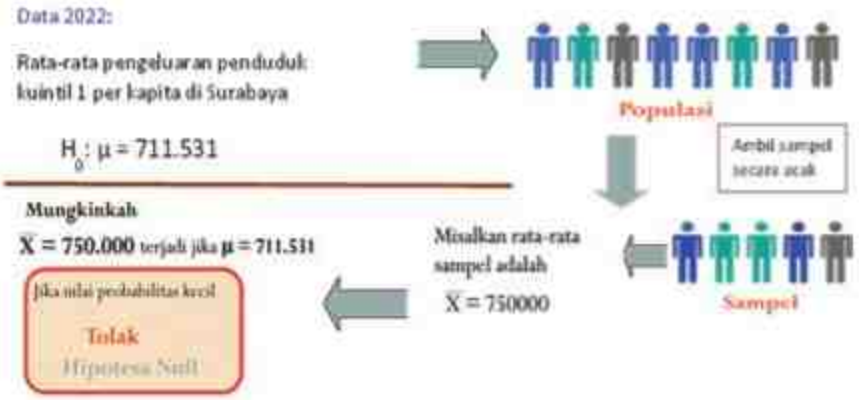
Tujuan dari menguji hipotesa adalah untuk menentukan apakah suatu dugaan terhadap parameter populasi didukung secara kuat oleh informasi yang diperoleh dari data sampel. Sebagai contoh, dari data masa lalu diketahui bahwa rata-rata pengeluaran per kapita sebulan penduduk kuintil 1 di Surabaya pada tahun 2022 adalah Rp 711.531 (BPS, 2023). Pada tahun 2023 ini dilakukan pengukuran secara acak dengan mengambil sampel penduduk pada kuintil 1 di Surabaya, ternyata didapati bahwa rata-rata pengeluaran per kapita sebulan penduduk kuintil 1 pada sampel tersebut adalah Rp 750.000. Pernyataan yang ingin dibuktikan pada uji hipotesa adalah:

Apakah nilai rata-rata pengeluaran per kapita sebesar $\bar{X} = 750.000$ dapat terjadi jika rata-rata populasi pengeluaran per kapita adalah sebesar $\mu = 711.531$ (lihat Gambar 6.1)

- Jika probabilitas bahwa $\bar{X} = 750.000$ terjadi bila $\mu = 711.531$ masih berlaku adalah **kecil**, maka kita **menolak** dugaan awal (Hipotesa Null). Hal ini menandakan bahwa $\mu = 711.531$ sudah tidak didukung oleh data.
- Namun demikian jika probabilitas bahwa $\bar{X} = 750.000$ terjadi bila $\mu = 711.531$ masih berlaku adalah **besar**, maka kita **gagal menolak** dugaan awal (Hipotesa Null). Hal ini menandakan bahwa $\mu = 711.531$ masih berlaku, tidak cukup data untuk menolak bahwa rata-rata pengeluaran penduduk pada kuintil 1 masih mengikuti nilai yang lama yaitu $\mu = 711.531$.

Pada ilustrasi ini terdapat beberapa langkah yang harus dilakukan agar uji hipotesa ini dapat disimpulkan. Langkah-langkah tersebut adalah

1. Memformulasikan hipotesa
2. Menentukan nilai probabilitas dari parameter sampel terjadi jika parameter populasi masih berlaku.
3. Menentukan suatu ambang batas agar kita dapat memutuskan apakah nilai probabilitas ini besar atau kecil. Sehingga kita dapat menolak ataupun gagal menolak hipotesa awal (Null Hypothesis)



Gambar 6.1. Ilustrasi menguji hipotesa

Contoh 6.1

Dalam sebuah industri bola lampu, diketahui bahwa tingkat kecacatan dari produk ini adalah 20%. Seorang insinyur di perusahaan tersebut menemukan metode baru. dan mengujicobakan pada 400 buah bola lampu, didapatkan hanya 17% saja yang cacat. Apakah insinyur tersebut dapat mengatakan bahwa metode temuannya tersebut mampu menurunkan kecacatan produk?

Jawab:

1. Memformulasikan hipotesa

Dalam ilmu statistika, hipotesa merupakan sebuah usulan model. Usulan ini kemudian diuji dengan data yang tersedia.

- Jika data tersebut konsisten dengan usulan model yang diberikan. maka tidak ada alasan untuk tidak mempercayai hipotesa yang diusulkan.
- Namun jika data yang tersedia tersebut tidak konsisten dengan model yang diusulkan maka ada dua hal yang harus diperhatikan, yaitu
  - Jika perbedaan antara data dengan model tersebut tidak terlalu besar, maka model usulan tersebut masih dapat diterima
  - Jika kenyataan data yang ada bertentangan dengan model yang diusulkan, maka terdapat bukti kuat bahwa model tersebut tidak sesuai dengan kenyataan yang ada di lapangan.

Hipotesa awal, atau disebut juga sebagai hipotesa null (H_0)menyatakan model populasi parameter yang saat ini merupakan fakta yang berlaku. Pada contoh di atas,H_0 dituliskan sebagai:

$$H_o : p = 0,2$$

Dimana p adalah proporsi tingkat kecacatan produk saat itu, yaitu 20%

Hipotesa alternatif (H_1), menyatakan sampel parameter yang hendak dibandingkan dengan populasi parameter. Pada kasus di atas, pihak manajemen mengklaim bahwa metode baru mereka dapat menurunkan tingkat kecacatan sebesar 17%.

$$H_1 : p < 0,2$$

Setelah memformulasikan hipotesa ini, kita perlu mengukur kemungkinan bahwa model usulan didukung oleh data atau tidak didukung oleh data.

2. Menentukan nilai kemungkinan bahwa nilai populasi parameter masih didukung oleh data atau sudah tidak didukung oleh data sampel yang baru dilakukan.

   Untuk itu kita perlu menghitung kemungkinan berapa nilai proporsi kecacatan yang baru ini terjadi jika proporsi populasi masih berlaku.

$$Pr(\hat{p} < 0,17|p = 0,2)$$

Kita tahu bahwa model proporsi mengikuti distribusi normal. Nilai probabilitas di atas dapat dihitung dengan menggunakan distribusi normal dengan mean $(\hat{p}) = p$ dan standar deviasi, $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$

$$SD(\hat{p}) = \sqrt{\frac{(0,2)(0,8)}{400}} = 0,02$$

$$Pr(\hat{p} < 0,17|p = 0,2) = Pr\left(Z < \frac{0,17 - 0,20}{0,02}\right)$$

$$= Pr(Z < -1,5) = 0.067$$

Kemungkinan bahwa tingkat kecacatan produk kurang dari 17% bila tingkat kecacatan populasi, 20% dalah 0,067.

3. Menentukan ambang batas

Untuk dapat membandingkan apakah nilai probabilitas sebesar 0,067 dapat dikatakan besar atau kecil, maka nilai probabilitas ini perlu dibandingkan dengan ambang batas (*threshold*).

Andaikan nilai ambang yang diambil adalah 10%, maka nilai probabilitas sebesar 0,067 ini bila dibandingkan dengan nilai ambang, masih di bawah 10%. Dengan demikian dapat dikatakan bahwa perbedaan antara nilai parameter sampel dengan nilai parameter populasi, **kecil**. Hal ini dapat disimpulkan bahwa hipotesa awal bahwa nilai populasi sampel = 0,2 kecil kemungkinannya untuk terjadi atau sudah tidak didukung oleh data lagi (tolak H_o).

Namun demikian apabila nilai ambang yang diambil adalah 5%, maka nilai probabilitas sebesar 0,067 ini bila dibandingkan dengan nilai ambang, lebih besar 5%. Dengan demikian dapat dikatakan bahwa perbedaan antara nilai parameter sampel dengan nilai parameter populasi, **Besar**. Hal ini dapat disimpulkan bahwa hipotesa awal bahwa nilai populasi sampel = 0,2 masih mungkin terjadi, atau masih didukung oleh data (gagal tolak H_o).

Ambang batas ini biasa disebut sebagai signifikan level. Topik ini akan dibicarakan secara khusus pada Subbab 6.2.3

Pada kasus ini, bila ambang batas adalah 5%, pihak manajemen tidak dapat mengklaim bahwa metode baru menghasilkan tingkat kecacatan lebih rendah bila dibandingkan dengan metode lama. Perbedaan proporsi kecacatan yang dihasilkan oleh metode baru tidak berbeda secara signifikan bila dibandingkan dengan proporsi kecacatan yang dihasilkan oleh metode lama.

Namun demikian bila ambang batas yang diambil adalah 10%, maka pihak manajemen dapat mengatakan bahwa metode baru tersebut menghasilkan tingkat kecacatan yang lebih rendah dan merupakan inovasi yang berhasil.

6.2 Konsep Uji Hipotesa

6.2.1 Formulasi Hipotesa

Secara konsep uji hipotesa mirip dengan konsep pengadilan pada umumnya. Pengadilan selalu mengedepankan **asal praduga tak besalah (*innocent defendant*)**. Orang yang sedang diadili adalah orang yang diduga tidak bersalah.

H_o : "terdakwa tidak bersalah" (innocent defendant)

Jika bukti-bukti yang dikumpulkan selama persidangan tidak mendukung asumsi praduga tak bersalah ini, maka hakim di pengadilan dapat **menolak hipotesa awal** dan menyatakan terdakwa bersalah (*guilty*).

Namun, jika tidak ditemukan bukti yang cukup untuk menghukum terdakwa, hakim tidak dapat menyatakan bahwa Ho benar, yaitu terdakwa *innocent*. Hakim hanya dapat menyatakan **gagal menolak** H_o dan menyatakan terdakwa tidak bersalah (*not guilty*).

Dalam hal pengadilan, bila suatu saat didapatkan bukti yang cukup untuk menghukum terdakwa, maka putusan pengadilan ini dapat berubah.

Demikian halnya dalam konsep uji hipotesa statistik. Hipotesa awal (null hypothesis) menyatakan nilai parameter yang akan digunakan dalam model. Nilai parameter yang digunakan adalah nilai populasi yang saat diuji sedang terjadi dan diasumsikan masih berlaku (*innocent defendant*). Pengujian hipotesa statistik hampir selalu mengenai model parameter.

$$H_0 : \text{parameter} = \text{nilai yang dihipotesakan}$$

Untuk melakukan uji hipotesa, kita harus menentukan hipotesa alternative. Hipotesa alternative ini biasanya merupakan range dari selain nilai yang dihipotesakan yang mungkin terjadi.

H_1 terdiri dari nilai parameter yang akan kita terima jika kita menolak hipotesa awal.

Perlu diingat bahwa kita tidak akan pernah bisa membuktikan hipotesis nol, kita hanya bisa menolaknya atau gagal menolaknya. Jika kami menolaknya, maka kami menerima alternatifnya.

Contoh 6.2

Andaikan kita ingin menguji preferensi mahasiswa dalam menggunakan dua aplikasi Grab Food ataupun Go Food untuk memesan makanan secara online. Untuk memformulasikan masalah ini kedalam uji hipotesa kita butuh untuk menentukan parameter dan nilai yang akan diuji. Dalam keseharian masalah di atas dapat dihipotesakan sebagai

H_o: Kedua aplikasi tersebut memiliki preferensi yang sama untuk digunakan para mahasiswa untuk memesan makanan.

H_1: Salah kedua aplikasi tersebut tidak disukai

Agar dapat diuji, rumusan hipotesa di atas harus diubah ke dalam bentuk model parameter. Pada masalah di atas terdapat dua pilihan saja yaitu suka dan tidak suka, maka tingkat preferensi dapat dinyatakan dalam bentuk parameter proporsi.

Misalkan p adalah proporsi mahasiswa yang lebih suka menggunakan Grab Food daripada Go Food. Rumusan hipotesa di atas dapat dituliskan sebagai

$$H_o : p = 0,5$$

$$H_1 : p \neq 0,5$$

Jika dilakukan survei dan ternyata tidak cukup bukti untuk mendukung Ho, maka kita akan menolak hipotesa null dan menerima hipotesa alternative.

Contoh 6.3

Pengalaman menunjukkan bahwa tingkat penyembuhan penyakit tertentu dengan menggunakan pengobatan yang saat ini sering dijalankan di masyarakat adalah 60%. Diketemukan suatu obat baru yang dinyatakan lebih manjur dari pengobatan cara lama tersebut. Andaikan obat baru ini diujicobakan pada 20 pasien, dan jumlah orang yang sembuh telah dicatat. Apakah cukup bukti untuk menyatakan cara pengobatan baru ini memiliki tingkat kemanjuran lebih tinggi dari cara pengobatan lama?

Uji hipotesa dari masalah ini dapat dituliskan sebagai:

H_o: Tingkat pengobatan lama masih tidak berubah

H_1: Tingkat pengobatan baru lebih baik dari tingkat pengobatan lama

Misalkan p adalah proporsi tingkat penyembuhan, maka hipotesa di atas dapat dituliskan ke dalam bentuk parameter sebagai berikut

$$H_o : p = 0,6$$

$$H_1 : p > 0,6$$

6.2.2 P-value

Dalam pengadilan di Indonesia, diperlukan Kitab Undang-undang Hukum Pidana (KUHP) untuk kasus pidana dan Kitab Undang-undang Hukum Perdata untuk kasus perdata. Kitab undang-undang ini digunakan oleh para Hakim sebagai pijakan dalam membuat putusan apakah terdakwa bersalah atau tidak bersalah.

Demikian pula dalam uji hipotesa diperlukan sebuah alat ukur untuk menimbang apakah kemungkinan H_o masih didukung oleh data atau sudah tidak lagi didukung oleh data. Alat ukur ini disebut sebagai P-Value (nilai probabilitas)

P-value didefinisikan sebagai probabilitas bersyarat bahwa nilai parameter statistik yang kita amati (H_1) terjadi jika hipotesa null benar (masih berlaku)

$$P_{Value} = Pr(\text{nilai parameter statistik yang kita amati}|H_o)$$

Jika P-value ini nilainya besar, maka H_o masih mungkin terjadi, kondisi awal masih didukung oleh data observasi. Kondisi ini disebut sebagai gagal tolak H_o (*failed to reject the hypotheses*).

Namun jika P-value ini nilainya kecil, maka H_o sudah tidak didukung oleh data observasi. Kondisi ini disebut sebagai tolak H_o (*reject the hypotheses*).

6.2.3 Tingkat Signifikansi (*Significant level*)

Untuk menentukan apakah P-value dianggap besar atau kecil, dibutuhkan sebuah nilai ambang batas (*threshold*). Bila P-value melebihi nilai ambang batas, maka P-value dianggap besar, sebaliknya bila P-value di bawah nilai ambang batas, maka P-value dianggap kecil.

Ambang batas ini disebut sebagai tingkat signifikansi (*significant level*), dan biasa disimbolkan sebagai nilai α (alpha).

Paradigma Neyman – Pearson

Jerzy Neyman dan Egon Pearson (1933) menyatakan bahwa dalam uji hipotesa terdapat empat kemungkinan yang terjadi yang dikenal sebagai paradigma Neyman-Pearson. Keempat hal tersebut dapat dinyatakan sebagai berikut (lihat Gambar 6.2):

- Bila dalam kenyataannya Ho benar dan keputusan uji hipotesa gagal menolak H_o maka uji hipotesa ini sudah benar (OK).
- Bila dalam kenyataannya Ho salah dan keputusan uji hipotesa adalah tolak H_o, kita melakukan hal yang benar. Kemampuan untuk mendeteksi hipotesa yang salah disebut sebagai kekuatan uji (*power of the test*).

Selain kedua hal yang benar ini, maka terdapat dua kesalahan yang mungkin terjadi saat kita melakukan uji hipotesa.

- Kesalahan tipe I (*Type I error*): Hipotesa Null benar, tetapi kita melakukan kesalahan dengan menolaknya (*False Positive*)
- Kesalahan tipe II (*Type II error*): Hipotesa Null salah, tetapi kita melakukan kesalahan dengan gagal menolaknya (*False Negative*)

Type I error dapat terjadi ketika hipotesa null benar, namun kita menarik data sampel yang kurang sesuai.

Keputusan	Yang sesungguhnya terjadi	
	Ho benar	Ho salah
Tolak Ho	Error tipe I	Power
Gagal tolak Ho	OK	Error Tipe II

Gambar 6.2. Paradigma Neyman-Pearson

Untuk menolak H_o, maka P-value harus kurang dari α. Pada kenyataannya bilai nilai H_o benar, maka terjadilah kesalahan tipe I. Dengan demikian probabilitas terjadinya kesalahan tipe I ini adalah sebesar α. Oleh karena itu nilai αi ni dapat didefinisikan sebagai probabilitas kesalahan tipe I terjadi.

$$\alpha = Pr(\text{Type I error}) = Pr(\text{Tolak } H_0 \text{ ketika } H_0 \text{ benar})$$

Untuk itu nilai α biasanya ditentukan sedemikian hingga probabilitas kesalahan tipe I ini kecil terjadinya. Umumnya nilai α yang sering digunakan adalah 5%. Namun demikian untuk kasus-kasus yang sangat sensitive, seperti misalnya dalam uji coba obat baru, atau metode baru dalam kesehatan yang menyangkut nyawa orang, nilai α yang diambil kurang dari 1%.

Apabila H_o salah dan kita gagal untuk menolaknya, maka kita membuat kesalahan Tipe II. Probabilitas kesalahan tipe II terjadi ini disebut sebagai β.

$$\beta = Pr(\text{Type II error}) = Pr(\text{Gagal tolak } H_0 \text{ ketika } H_0 \text{ salah})$$

Illustrasi Paradigma Neyman-Pearson

Mary Jane Veloso yang merupakan terpidana mati asal Filipina karena tertangkap membawa 2,6-kilogram heroin in bandara Yogyakarta pada bulan April 2010. May Jane mengklaim narkoba tersebut dijahitkan dalam kopor miliknya tanpa sepengetahuan dirinya (detiknews, 16 Jan 2024).

Kita akan menganalisa kisah di atas dengan menggunakan Paradigma Neyman-Pearson (Gambar 6.3).

Keputusan	Yang sesungguhnya terjadi	
	Mary Jane benar	Mary Jane salah
Tolak Ho	Error tipe I	Power
Gagal Tolak Ho	OK	Error Tipe II

Gambar 6.3. Paradigma Neyman-Pearson pada kisah Mary Jane

Uji hipotesa, seperti halnya pada pengadilan yang berlaku di seluruh dunia menganut asas praduga tak bersalah (*innocent*).

- H_o: Mary Jane tidak bersalah
- H_1: Mary Jane bersalah

Terdapat empat kemungkinan yang dapat terjadi pada kasus pengadilan Mary Jane ini. Dua hal yang pertama bila keputusan pengadilan sesuai dengan kenyataan yang sesungguhnya terjadi:

- Pada kenyataannya Mary Jane benar dan pengadilan memutuskan dia tidak bersalah (gagal tolak H_o). Pengadilan akan membebaskan Mary Jane
- Pada kenyataannya Mary Jane salah dan pengadilan memutuskan dia bersalah (tolak H_o). Pengadilan akan menghukum mati Mary Jane

Kesalahan yang mungkin terjadi pada kasus pengadilan ini adalah:

- Kesalahan Tipe I: Pada kenyataannya Mary Jane benar, tetapi pengadilan memutuskan dia bersalah (menolak H_o). Dalam hal ini Pengadilan akan menghukum mati orang yang tidak bersalah.
- Kesalahan Tipe II: Pada kenyataannya Mary Jane salah, tetapi pengadilan memutuskan dia tidak bersalah (gagal H_o). Dalam hal ini Pengadilan akan membebaskan orang yang bersalah.

Dari dua tipe kesalahan ini terlihat bahwa kesalahan Tipe I lebih berat daripada kesalahan Tipe II. Pada ilustrasi ini bila terjadi kesalahan Tipe I maka Mary Jane di hukum mati. Apabila disuatu saat nanti dapat dibuktikan ternyata Mary Jane ini tidak bersalah, maka pengadilan tidak dapat menghidupkan Mary Jane kembali dari hukuman mati yang sudah dijalaninya. Namun demikian apabila kesalahan Tipe II yang terjadi, pengadilan masih dapat menangkap kembali Mary Jane, bila ditemukan cukup bukti bahwa Mary Jane bersalah.

Untuk itu α nilai probabilitas kesalahan tipe I terjadi, harus diambil dengan hati-hati, agar kesalahan tipe I ini terjadi sekecil mungkin.

6.3 Uji Hipotesa sampel Tunggal

Uji hipotesa sampel tunggal digunakan untuk menguji apakah nilai parameter dari sebuah populasi sudah berubah atau tidak terjadi perubahan nilai yang signifikan. Untuk itu akan diambil sampel dari populasi tersebut dan dihitung nilai probabilitas nilai parameter yang baru tersebut bila nilai parameter yang lama masih berlaku (p-value).

6.3.1 Uji Proporsi untuk sampel Tunggal (*One-sample proportion test*)

Andaikan proporsi populasi adalah p_0 maka hipotesa null dari uji proporsi untuk sampel tunggal, maka terdapat dua kemungkinan uji hipotesa yang dapat kita tuliskan.

1. Uji satu arah (*one-sided test*), bila kita tertarik untuk menguji apakah nilai proporsi populasi yang lama lebih besar atau lebih kecil dari nilai proporsi baru/usulan atau proporsi sampel.

$$H_0 : p = p_o$$

$$H_1 : p > p_o \text{ atau}$$

$$H_1 : p < p_o$$

Uji dua arah (*two-sided test*), bila kita tertarik untuk menguji apakah nilai proporsi populasi tidak sama dengan nilai proporsi baru/usulan atau proporsi sampel. Uji ini disebut sebagai uji dua arah, karena proporsi sampel tersebut bisa lebih besar ataupun lebih kecil, tetapi hal tersebut tidak diperhatikan. Perhatian utama adalah asalkan nilai proporsi populasi tidak sama dengan nilai proporsi sampel.

$$H_1 : p \neq p_o$$

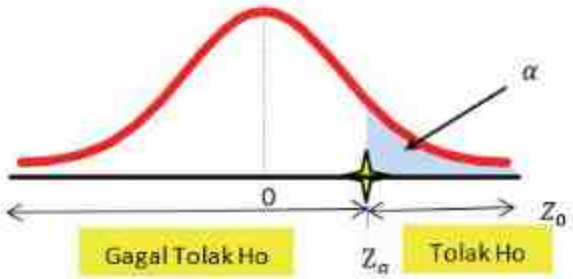
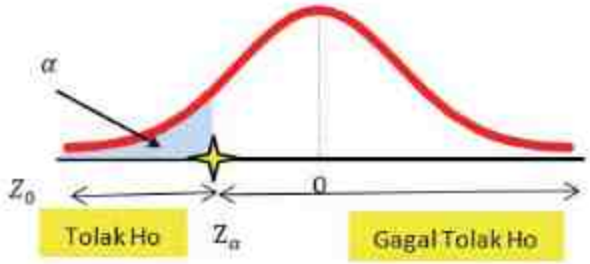
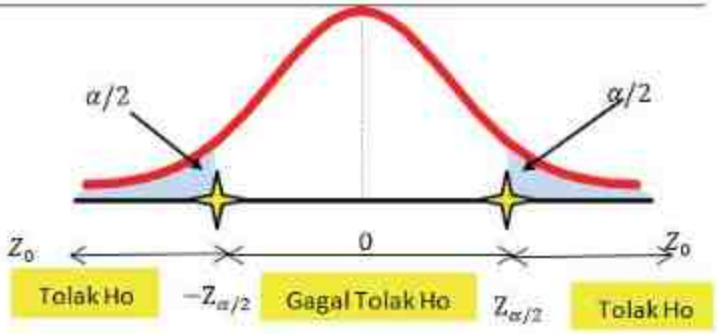
Test-Statistics

Pada uji proporsi data diasumsikan berdistribusi binomial. Bila nilai p tidak terlalu dekat nilainya ke nol atau ke satu dan ukuran sampelnya cukup besar, maka distribusi binomial dapat didekati dengan distribusi normal. Oleh karena itu test-statistics untuk uji proporsi di dasarkan pada pendekatan distribusi normal. Ringkasan penentuan nilai p-value dapat dilihat pada Gambar 6.4

Test statistics: $p \sim N[np_o, np_o(1-p_o)]$

$$Z_o = \frac{p - p_o}{\sqrt{p_o(1-p_o)/n}}$$

Syarat: $np_o \geq 5$ dan $n(1-p_o) \geq 5$. Sampel diambil secara independen.

Hipotesa	P-value	
$H_o: p = p_0$ $H_1: p > p_0$	P-value = Pr $(Z \geq Z_0)$ $= 1 - \Phi(Z_0)$ $\Phi(.)$ – CDF dari Normal α – Signifikan level Z_α – nilai kritis (*critical value*)	
$H_o: p = p_0$ $H_1: p < p_0$	P-value = Pr $(Z \leq Z_0)$ $= \Phi(Z_0)$	
$H_o: p = p_0$ $H_1: p \neq p_0$	P-value = Pr $(\|Z\| \leq Z_0)$ $= 2[1 - \Phi(Z_0)]$	

Gambar 6.4. P-Value

Daerah Kritis (*Critical Value*) - lihat Tabel 6.1

Pada gambar di atas Z_α didefinisikan sebagai nilai kritis (*critical value*). Beberapa contoh penentuan nilai kritis dapat dilihat pada Tabel 6.1. Nilai ini adalah

$$Pr(Z < Z_\alpha) = \alpha$$

atau

$$Z_\alpha = \Phi^{-1}(\alpha)$$

Tabel 6.1. Nilai Z_α

α	1-sided	2-sided
0,05	1,645	1,96
0,01	2,28	2,575
0,001	3,09	3,29

- Bila P-value $< \alpha$ maka kemungkinan bahwa H_1 terjadi bila H_o masih dianggap berlaku kecil kemungkinannya terjadi. Sehingga dapat disimpulkan bahwa proporsi populasi p_o sudah tidak didukung oleh data sampel, dan kita menolak H_o
- Sebaliknya P-value $> \alpha$ maka kemungkinan bahwa H_1 terjadi bila H_o masih sangat mungkin terjadi. Dalam hal ini proporsi populasi p_o masih mungkin terjadi, maka kita gagal menolak (Tabel 6.2).
- Selain membandingkan P-value dengan signifikan level (α), kita dapat juga melihat daerah kritis. Pada gambar illustrasi pada Tabel Pvalue, adalah daerah yang diarsir.

Tabel 6.2. Daerah Kritis

Hipotesa	Daerah Kritis	Keputusan
$H_o : p = p_o$ $H_1 : p > p_o$	P-value $< \alpha$ $Z_o > Z_\alpha$	Tolak H_o
$H_o : p = p_o$ $H_1 : p < p_o$	P-value $< \alpha$ $Z_o < -Z_\alpha$	Tolak H_o
$H_o : p = p_o$ $H_o : p \neq p_o$	P-value $< \alpha$ $Z_o > Z_\alpha$ atau $Z_o < -Z_\alpha$	Tolak H_o

Contoh 6.4

Diperkirakan lebih dari 70% gangguan pada saluran transmisi disebabkan oleh petir. Dalam sampel acak yang terdiri dari 200 kesalahan dari basis data yang besar, 151 disebabkan oleh petir. Apakah data memberikan bukti kuat yang mendukung anggapan ini? Tes pada tingkat signifikansi = 0,01; dan laporkan nilai p-value.

Penyelesaian:

a. Uji satu sisi (One-sided test)

Hipotesa dari permasalahan ini adalah satu sisi:

$H_o : p = 0,7$

$H_1 : p > 0,7$

$n = 200 : \hat{p} = 151/200 = 0,755$

$$Z_o = \frac{\hat{p} - p_o}{\sqrt{p_o(1-p_o)/n}} = \frac{0,755 - 0,7}{\sqrt{(0,7)(0,3)/200}} = 1,697$$

$$\text{P-value} = 1 - \Phi(1,697) = 1 - 0,9554 = 0,0446$$

Andaikan pihak manajemen mengambil nilai signifikan level $\alpha = 5\%$, maka p-value = 0,0446 $< \alpha$; nilai kritis $Z_{\alpha} = 1,645$. Dapat dilihat bahwa $Z_o > Z_{\alpha}$. Dapat disimpulkan bahwa hipotesa null ditolak dan manajemen dapat mengklaim bahwa lebih dari 70% gangguan pada saluran transmisi disebabkan oleh petir.

Namun demikian, bila nilai signifikan level $\alpha = 1\%$, maka p-value = 0,0446 $> \alpha$; nilai kritis $Z_o = 2,28$. Dapat dilihat bahwa $Z_o < Z\alpha$. Dapat disimpulkan uji ini gagal menolak hipotesa null ditolak. Gangguan transmisi disebabkan oleh petir tidak lebih dari 70%.

b. Uji dua sisi (*two-sided test*)

Hipotesa dari permasalahan ini adalah dua sisi:

$H_o : p = 0,7$

$H_1 : p \neq 0,7$

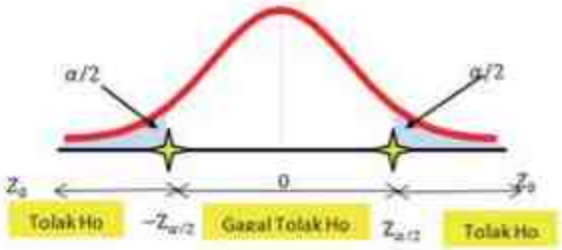
Bila nilai signifikan level yang diambil adalah $\alpha = 5\%$ ($\frac{\alpha}{2}$=2,5%), maka nilai p-value = 0,0446 $> \frac{\alpha}{2}$; nilai kritis $Z_{\alpha/2} = 1,96$. Dapat dilihat bahwa $Z_o < Z_{\alpha/2}$ dan kita gagal menolak H_o. Pada two-sided test ini disimpulkan bahwa gangguan transmisi yang disebabkan oleh petir masih sama dengan 70%.

```
p = 0.7; n = 200; p_hat = 151/n
sd_p = sqrt(p*(1-p)/n)
Z0 = (p_hat-p)/sd_p
pvalue = 1-pnorm(Z0)
pvalue
```

```
[1] 0.0448165
```

6.3.2 Hubungan antara Confidence Interval dan Uji Hipotesa

Selang kepercayaan (*Confidence Interval*) dan uji hipotesa dihitung dengan menggunakan pendekatan yang sama. Keduanya memiliki asumsi dan syarat yang sama. Secara alamiah, selang kepercayaan memiliki batas bawah dan batas atas (dua sisi, *two-sided*), maka selang kepercayaan memiliki hubungan dengan uji hipotesa dua sisi (*two-sided test*).



Selang kepercayaan dari $\hat{p}$ berada pada daerah gagal tolak H_o, yaitu

$$[\hat{p} - Z\_\alpha/2\ SE; \hat{p} + Z\_\alpha/2\ SE]$$

dengan demikian dapat disimpulkan apabila p_o berada di antara confidence interval maka kita gagal menolak H_o. Sebaliknya apabila p_o berada di luar interval kepercayaan, maka kita menolak H_o.

Contoh 6.5

Selang kepercayaan dari masalah di Contoh 6.4 di atas dapat dihitung sebagai berikut:

$\hat{p} = 151/200 = 0,755$

$SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{(0,755)(0,245)/200} = 0,0304$

Jika $\alpha = 5\%$ ($\alpha/2$ =2,5%), maka $Z_{\alpha/2} = 1.96$

Selang kepercayaan $= [0,755 - (1,96) * (0,0304); 0,755 + (1,96) * (0.0304)] = [0,695; 0.814]$

Pada uji hipotesa dua sisi (*two-sided test*)

$H_o : p = 0,7$

$H_1 : p \neq 0,7$

Dapat kita lihat bahwa $p_o = 0,7$ berada di dalam selang kepercayaan tersebut.

$$\hat{p} - Z_{\alpha/2}\ SE \leq p_o \leq \hat{p} + Z_{\alpha/2}\ SE$$

$$0,695 \leq 0,7 \leq 0,814$$

maka dapat disimpulkan bahwa kita gagal menolak Ho. Kesimpulan yang sama seperti yang kita ambil pada Contoh 6.4.

```
p = 0.7
n = 200
alpha = 0.05
p_hat = 151/n
se_p = sqrt(p_hat*(1-p_hat)/n)
Batas_bawah = p_hat+qnorm(alpha/2)*se_p
Batas_atas  = p_hat-qnorm(alpha/2)*se_p
Batas_bawah
```

```
[1] 0.6953941
```

```
Batas_atas
```

```
[1] 0.8146059
```

Perintah R yang dapat digunakan untuk menguji proporsi adalah:

```
Uji eksak dengan menggunakan distribusi binomial:
```

```
binom.test(x, n, p = 0.5,alternative = c("two.sided", "less", "greater"),
conf.level = 0.95)
```

atau dengan menggunakan pendekatan distribusi normal:

prop.test(x, n, p = NULL,alternative = c("two.sided", "less", "greater"),conf.level = 0.95, correct = TRUE)

dimana:

x – jumlah sukses

n – jumlah sampel

p – proporsi populasi

Contoh 6.4 dan 6.5 dapat diselesaikan dengan menggunakan R-script sebagai berikut:

```
BinTest = binom.test(151, n=200, p = 0.7, alternative = "two.sided")
BinTest
```

```
    Exact binomial test

data:  151 and 200
number of successes = 151, number of trials = 200, p-value = 0.1048
alternative hypothesis: true probability of success is not equal to 0.7
95 percent confidence interval:
 0.6893601 0.8129159
sample estimates:
probability of success
                 0.755
```

Pada uji hipotesa dengan menggunakan perintah binom.test ini diperoleh pvalue = 0.1047992 dan confidence interval berada pada selang 0.6893601, 0.8129159. Dari uji ini Kita gagal menolak H_o. Nilai proporsi populasi tidak berubah, $p = 0,7$.

```
PropTest = prop.test(151, n=200, p = 0.7, alternative = "two.sided",
                     correct = TRUE)
PropTest
```

```
    1-sample proportions test with continuity correction

data:  151 out of 200, null probability 0.7
X-squared = 2.625, df = 1, p-value = 0.1052
alternative hypothesis: true p is not equal to 0.7
95 percent confidence interval:
 0.6883070 0.8116839
sample estimates:
    p
0.755
```

Pada uji hipotesa dengan menggunakan perintah prop.test ini diperoleh pvalue = 0.1051925 dan confidence interval berada pada selang 0.688307, 0.8116839 Dari uji ini Kita gagal menolak H_o. Nilai proporsi populasi tidak berubah, $p = 0,7$.

6.3.3 Uji Mean untuk sampel Tunggal (One-sample mean test)

Pada prinsipnya uji mean untuk sampel tunggal memiliki cara yang mirip seperti pada uji proporsi untuk sampel tunggal. Hanya saja, pada uji mean untuk sampel tunggal ini perlu diperhatikan apakah varians dari sampel diketahui atau tidak diketahui.

Andaikan kita akan menguji apakah mean populasi masih didukung oleh data ataupun tidak, maka hipotesa null dari permasalahan ini dapat dituliskan sebagai

$$H_o : \mu = \mu_o$$

dimana μ_o adalah mean populasi, atau nilai mean yang sudah diketahui. Untuk menguji hipotesa null di atas, maka diambilah sampel secara acak dari populasi yang berdistribusi normal, $X_1, X_2, ..., X_n$ dan dihitung nilai rata-ratanya, $\bar{X}$.

A. Bila varians populasi diketahui

Diketahui berdasarkan *central limit theory*: $\bar{X} \sim N(\mu_o, \sigma/\sqrt{n})$ dengan σ adalah standar deviasi yang nilainya diketahui dan n adalah jumlah sampel yang diambil. Tabel 6.3 memberikan ringkasan nilai P-value dan daerah kritis dari uji mean bila varians populasi diketahui.

Uji mean dengan varians diketahui, biasa disebut juga sebagai Z-test dapat dituliskan sebagai berikut:

Test statistics:

$$Z_o = \frac{\bar{X} - \mu_o}{\sigma/\sqrt{n}}$$

Tabel 6.3. P-value - Daerah kritis dari uji mean bila varians populasi diketahui

Hipotesa	P-value	Daerah kritis	Keputusan
$H_o : \mu = \mu_o$	P-value $= Pr(Z \geq Z_o)$	P-value $< \alpha$	Tolak H_o
$H_a : \mu > \mu_o$	$= 1 - \Phi(Z_o)$	$Z_o > Z_\alpha$	
$H_o : \mu = \mu_o$	P-value $= Pr(Z \leq Z_o)$	P-value $< \alpha$	Tolak H_o
$H_a : \mu < \mu_o$	$= \Phi(Z_o)$	$Z_o < -Z_\alpha$	
$H_o : \mu = \mu_o$	P-value $= 2[1 - \Phi(Z_o)]$	P-value $< \alpha$	Tolak H_o
$H_a : \mu \neq \mu_o$		$Z_o > Z_{\alpha/2}$ atau $Z_o < -Z_{\alpha/2}$	

Contoh 6.6

Sebuah UMKM memproduksi kripik pisang yang dibungkus per 50 gram secara rata-rata, dengan simpangan baku $\sigma = 2$ gram. Seorang mahasiswa sedang mengerjakan kerja praktek

di UMKM tersebut dan tertarik untuk menguji apakah kripik pisang ini memiliki rata-rata berat sesuai yang diutarakan pemiliknya. Untuk itu mahasiswa tersebut mengambil secara acak 25 bungkus kripik pisang dan diperoleh rata-rata sampel sebesar $\bar{X} = 51,2$ gram. Bila signifikansi $\alpha = 0,05$; kesimpulan apa yang dapat ditarik oleh mahasiswa tersebut?

Jawab:

Hipotesa dari masalah ini dapat dituliskan sebagai berikut:

$$H_o : \mu = 50$$

$$H_1 : \mu \neq 50$$

Test statistics:

$$Z_o = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{51,3 - 50}{2/\sqrt{25}} = 3,25$$

$$\text{P-value} = 2[1 - \Phi(3,25)] = 2[1 - 0,999423] = 0,0012$$

dengan $\alpha = 5\%$; $Z_{\alpha/2} = 1,96$; terlihat bahwa p-value $< \alpha$, $Z_o > Z_{\alpha/2}$ maka mahasiswa tersebut akan menolak H_o. Dapat disimpulkan bahwa rata-rata berat kripik pisang tidak sama dengan 50 gram lagi per bungkus.

```
miu = 50; n = 25; sigma = 2
X_bar = 51.3
Z0 = (X_bar-miu)/(sigma/sqrt(n))
pvalue = 1 - pnorm(Z0)
pvalue
```

```
[1] 0.000577025
```

Jika hipotesa yang dilakukan adalah satu sisi:

$$H_1 : \mu > 50$$

Test statistics: $Z_o = 3,25$

$$\text{P-value} = [1 - \Phi(3,25)] = [1 - 0,999423] = 0,000577$$

dengan $\alpha = 5\%$; $Z_\alpha = 1,65$; terlihat bahwa p-value $< \alpha$, $Z_o > Z_{\alpha/2}$ maka mahasiswa tersebut akan menolak H_o. Dapat disimpulkan bahwa rata-rata berat kripik pisang sudah lebih dari 50 gram lagi per bungkus.

B. Bila varians populasi tidak diketahui

Bila varians populasi tidak diketahui (lihat Tabel 6.4), maka nilai varians yang digunakan adalah nilai varians sampel, sehingga pada uji hipotesa

$$H_0 : \mu = \mu_o$$

$$H_1 : \mu \neq \mu_o$$

Statistik uji yang digunakan adalah:

$$T_o = \frac{\bar{X} - \mu_o}{S/\sqrt{n}}$$

T_o akan berdistribusi t dengan derajat kebebasan $n - 1$

n adalah jumlah sampel

$\bar{X}$ adalah rata-rata sampel

S adalah standar deviasi sampel

Tabel 6.4. P-value - Daerah kritis dari uji mean bila varians populasi tidak diketahui

Hipotesa	P-value	Daerah kritis	Keputusan
$H_o : \mu = \mu_o$	P-value $= Pr(T \geq T_o)$	P-value $< \alpha$	Tolak
$H_1 : \mu > \mu_o$		$T_o > t_{\alpha,n-1}$	H_o
$H_o : \mu = \mu_o$	P-value $= Pr(T \leq T_o)$	P-value $< \alpha$	Tolak
$H_1 : \mu < \mu_o$		$T_o < -t_{\alpha,n-1}$	H_o
$H_o : \mu = \mu_o$	P-value $= Pr(T >$	P-value $< \alpha$	Tolak
$H_1 : \mu \neq \mu_o$	$\lvert T_o \rvert) + Pr(T \leq -\lvert T_o \rvert)$	$T_o > t_{\alpha/2,n-1}$ atau $T_o < -t_{\alpha/2,n-1}$	H_o

Contoh 6.7

Batas kecepatan untuk melaju di Jalan Tol Surabaya-Malang adalah 80 Km/jam (Detiknews, 8 Juli 2019). Untuk melihat apakah kendaraan yang melaju di jalan tol Surabaya-Malang mentaati aturan tersebut dilakukan uji kecepatan secara acak terhadap 25 kendaraan dari 100 yang lewat di jalan tol tersebut, dan diperoleh data sebagai berikut

89	89	84	124	94	94	94	92	96	88
91	81	120	87	104	89	97	96	98	89
81	91	86	95	100					

Apa yang dapat anda simpulkan dari sampel data ini? Apakah para pengemudi taat pada batas kecepatan yang ditetapkan oleh pemerintah untuk melaju di jalan tol Surabaya-Malang?

Jawab:

Hipotesa dari permasalahan di atas adalah:

$$H_{,,} : \mu = 80$$

$$H_1 : \mu \geq 80$$

Dari data di atas diperoleh jumlah sampel $n = 25$; dengan derajat kebebasan $df = n - 1 = 24$. Rata-rata sampel, $\bar{X} = 93,96$; dengan standar deviasi $S = 10,11$

$$T_0 = \frac{93,96 - 80}{10,11/\sqrt{25}} = 6,9014$$

Probabilitas bahwa kecepatan rata-rata pengendara yang melaju di jalan tol Surabaya - Malang lebih dari 80 Km/jam, jika Ho masih berlaku adalah

$$\text{P-value} = Pr\ (T > 6,9014) = 4,574e - 7 \sim 0$$

Nilai p-value yang diperoleh sangat kecil, mendekati nol, dan nilai ini kurang dari signifikan level α=5%. Dengan demikian Ho sudah tidak didukung kuat oleh sampel data dan kita menolak Ho. Para pengendara yang melanju di jalan tol Surabaya-Malang sudah melebihi batas kecepatan yang ditetapkan oleh pemerintah.

```
#Cara1: Dihitung secara manual
DataKecepatan   = c(89,89,84,124,94,94,94,92,96,88,91,81,120,87,104,
                    89,97,96,98,89,81,91,86,95,100)
alpha       = 0.05
n           = length(DataKecepatan)
X_bar       = mean(DataKecepatan)
Mu            = 80
S           = sd(DataKecepatan)
To            = (X_bar - Mu)/(S/sqrt(n))
Pvalue      = dt(To,n-1)
```

```
T           = qt(1-alpha,n-1)
Pvalue
```

```
[1] 4.574729e-07
```

```
T
```

```
[1] 1.710882
```

Pada cara perhitungan manual diperoleh nilai pvalue = 4.5747286×10^{-7} dan T-statistik = rT . Perhitungan ini dapat diringkas dengan menggunakan perintah t.test pada R sebagai berikut

```
#Cara2: Dihitung dengan menggunakan perintah t.test
t.test(DataKecepatan, mu = 80,alternative = "greater")
```

```
    One Sample t-test

data:  DataKecepatan
t = 6.9014, df = 24, p-value = 1.942e-07
alternative hypothesis: true mean is greater than 80
95 percent confidence interval:
 90.49928      Inf
sample estimates:
mean of x
    93.96
```

Bila digunakan two-sided test maka cara2 di atas dapat diubah menjadi:

```
#Cara3: Bila diuji dengan menggunakan two.sided test
t.test(DataKecepatan, mu = 80)
```

```
    One Sample t-test

data:  DataKecepatan
t = 6.9014, df = 24, p-value = 3.884e-07
alternative hypothesis: true mean is not equal to 80
```

```
95 percent confidence interval:
 89.78521 98.13479
sample estimates:
mean of x
    93.96
```

Pada two-sided test ini terlihat bahwa $\mu = 80$ berada di luar selang kepercayaan $[89,78521; 98,1347]$. Dapat disimpulkan bahwa ketetapan batas berkendara 80Km/jam sudah tidak dipatuhi oleh para pengendara yang melaju di tol Surabaya-Malang.

6.4 Uji Hipotesa untuk Dua sampel (Two-sample Test)

Pada sub-bab ini kita akan menguji apakah parameter statistik yang diambil dari dua sampel yang berbeda adalah sama ataukah memiliki perbedaan.

6.4.1 Uji Proporsi untuk Dua sampel (*Two-sample proportion test*)

Andaikan dua independent random sampel yang berukuran n_1 dan n_2 diambil dari dua populasi: X_1 dan X_2 menunjukkan jumlah "kesuksesan" yang berasal dari sampel 1 dan sampel 2. Andaikan pendekatan distribusi normal diaplikasikan pada distribusi binomial dari masing-masing populasi ini, maka estimasi dari proporsi "sukses" untuk tiap sampel adalah $p_1 = \frac{X_1}{n_1}$ dan $p_2 = \frac{X_2}{n_2}$ Uji hipotesa: $H_o : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$

Statistik uji:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Akan berdistribusi normal $N(0,1)$.

Pooled estimator untuk proporsi p dapat dituliskan sebagai berikut

$$\widehat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Statistik uji untuk $H_o : p_1 = p_2$ akan berubah menjadi

$$Z_0 = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Contoh 6.8 Dalam sebuah wawancara terhadap 593 mahasiswa, didapatkan 62 dari 325 mahasiswi ingin melanjutkan studi S2 setelah lulus dari S1 dan 75 dari 268 mahasiswa ingin

melanjutkan studi S2. Adakah perbedaan yang signifikan antara proporsi mahasiswi dan mahasiswa yang ingin melanjutkan studi S2?

Jawab $p_1 = \frac{62}{325} = 0,19$; $p_2 = \frac{75}{268} = 0,28$;

$\hat{p} = \frac{137}{593} = 0,23$

Uji Hipotesa: $H_o : p_1 = p_2$; vs $H_1 : p_1 < p_2$ Statistik uji:

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = -2,56$$

P-value = 0,0052

Kesimpulan: Nilai P-value $< \alpha$ (5%), maka kita menolak hipotesa null. Proporsi mahasiswa yang ingin melanjutkan S2 lebih besar dari proporsi mahasiswi yang ingin melanjutkan studi S2 secara signifikan.

```
X_Mhsi = 62
n_Mhsi = 325
X_Mhsa = 75
n_Mhsa = 268

p_Mhsi = X_Mhsi/n_Mhsi
p_Mhsa = X_Mhsa/n_Mhsa
p  = (X_Mhsi + X_Mhsa)/(n_Mhsi + n_Mhsa)
q  = 1-p
Zhit  = (p_Mhsi - p_Mhsa)/sqrt(p*q*(1/n_Mhsi + 1/n_Mhsa))
Pvalue = pnorm(Zhit)
Pvalue
```

```
[1] 0.005212216
```

Pada perhitungan manual di atas diperoleh nilai pvalue = 0.0052122. Nilai pvalue ini lebih kecil dari $\alpha = 5\%$, sehingga kita menolak H_o. Dapat disimpulkan bahwa proporsi mahasiswi yang ingin melanjutkan studi ke S2 lebih kecil daripada proporsi mahasiswa yang ingin melanjutkan ke S2 secara signifikan.

Uji dua proporsi ini dapat dilakukan dengan menggunakan perintah prop.test pada R sebagai berikut:

```
prop.test(x = c(X_Mhsi, X_Mhsa), n = c(n_Mhsi, n_Mhsa),
          alternative = 'less')
```

```
    2-sample test for equality of proportions with continuity correction

data:  c(X_Mhsi, X_Mhsa) out of c(n_Mhsi, n_Mhsa)
X-squared = 6.069, df = 1, p-value = 0.006879
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.02806058
sample estimates:
   prop 1    prop 2
0.1907692 0.2798507
```

6.4.2 Uji Mean untuk Dua sampel (*Two-sample t-test*)

Apabila sampel data berasal dari dua populasi yang berbeda, dan kita tertarik untuk menguji apakah perbedaan mean dari kedua sampel tersebut sama dengan Δ_0. Dimana Δ_0 adalah sebuah konstanta yang ditetapkan. Apabila kita ingin menguji apakah kedua sampel mean tersebut sama, maka nilai $\Delta_0 = 0$.

Andaikan μ_1 adalah mean dari populasi pertama dan μ_2 adalah mean dari populasi kedua. Kita tertarik untuk menguji apakah perbedaan dari kedua mean tersebut, yaitu $\mu_1 - \mu_2 = \Delta_0$.

Seperti telah kita bahas pada Bab 5, apabila kedua populasi tersebut berdistribusi normal, maka perbedaan sampel mean dari populasi tersebut juga akan normal dengan varians

$$SD(\bar{X}_1 - \bar{X}_2) = \sqrt{(Var(\bar{X}_1) + Var(\bar{X}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

dimana $\bar{X}_1$ adalah sampel mean dari populasi pertama $\bar{X}_2$ adalah sampel mean dari populasi kedua. n_1 adalah ukuran sampel dari populasi pertama n_2 adalah ukuran sampel dari populasi kedua. σ_1^2 adalah varians dari populasi pertama σ_2^2 adalah varians dari populasi kedua.

Dengan demikian uji hipotesa dari perbedaan mean antara dua populasi dapat diformulasikan sebagai berikut (ringkasan dapat dilihat pada Tabel 6.6).

A. Bila varians dari kedua populasi tersebut diketahui

Hipotesa Null: $H_o : \mu_1 - \mu_2 = \Delta_0$ Statistik uji yang digunakan adalah:

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Tabel 6.6. P-value - Daerah kritis dari uji mean dua sampel

Hipotesa	P-value	Daerah kritis	Keputusan
$H_o : \mu_1 - \mu_2 = \Delta_o$ $H_1 : \mu_1 - \mu_2 > \Delta_o$	P-value $= Pr(Z \geq Z_o)$ $= 1 - \Phi(Z_o)$	P-value $< \alpha$ $Z_o > Z_\alpha$	Tolak H_o
$H_o : \mu_1 - \mu_2 = \Delta_o$ $H_1 : \mu < \mu_o$	P-value $= Pr(Z \leq Z_o)$ $= \Phi(Z_o)$	P-value $< \alpha$ $Z_o < -Z_\alpha$	Tolak H_o
$H_o : \mu_1 - \mu_2 = \Delta_o$ $H_1 : \mu_1 - \mu_2 \neq \Delta_o$	P-value $= Pr(Z \geq$ $\|z_o\|) + Pr(Z \leq -\|z_o\|)$ $= 2[1 - \Phi(\|Z_o\|)]$	P-value $< \alpha$ $Z_o > Z_{\alpha/2}$ atau $Z_o < -Z_{\alpha/2}$	Tolak H_o

Contoh 6.9

Seorang pengembang produk tertarik untuk mengurangi waktu pengeringan cat primer. Dua formulasi cat diuji; formulasi 1 adalah kimia standar, dan formulasi 2 memiliki bahan pengeringan baru yang seharusnya mengurangi waktu pengeringan. Dari pengalaman diketahui bahwa standar deviasi waktu pengeringan adalah 8 menit, dan variabilitas bawaan ini tidak terpengaruh oleh penambahan bahan baru. Sepuluh benda uji dicat dengan formulasi 1, dan 10 benda uji lainnya dicat dengan formulasi 2; 20 spesimen dicat secara acak. Rata-rata waktu pengeringan kedua sampel berturut-turut adalah $\bar{x}_1 = 121$ menit dan $\bar{x}_2 = 112$ menit. Kesimpulan apa yang dapat diambil oleh pengembang produk tentang efektivitas bahan baru tersebut, dengan menggunakan $\alpha = 0,05$?

Jawaban. Terdapat tujuh langkah yang dapat digunakan untuk menyelesaikan masalah ini.

1. Menentukan parameter yang akan diuji. Pada contoh ini parameter yang akan diuji adalah perbedaan mean dari waktu pengeringan cat primer, $\mu_1 - \mu_2$, dan $\Delta_0 = 0$
2. Menentukan hipotesa null. $H_o : \mu_1 - \mu_2 = 0$ atau $H_o : \mu_1 = \mu_2$
3. Menentukan hipotesa alternative: $H_1 : \mu_1 > \mu_2$
4. Menentukan statistik uji yang akan digunakan

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ dimana} \sigma_1^2 = \sigma_2^2 = 8^2 = 64 \text{ dan } n_1 = n_2 = 10$$

5. Menentukan daerah tolak H_o. Tolak $H_o : \mu_1 = \mu_2$ jika P-value kurang dari 0.05
6. Menghitung nilai Z_0. Diketahui $x_1 = 121$ dan $x_2 = 112$, maka

$$z_0 = \frac{121 - 112}{\sqrt{\frac{64}{10} + \frac{64}{10}}} = 2,52$$

Kesimpulan: diperoleh nilai $z_0 = 2,52$. P-value $= Pr\ (Z \geq 2,5) = 1 - \Phi(2.52) = 0,0059$. Nilai P-value kurang dari 0,05 maka kita menolak hipotesa null. Secara praktis dapat disimpulkan bahwa menambahkan bahan baru ke dalam cat akan mengurangi waktu pengeringan secara signifikan.

B. Bila varians dari kedua populasi tersebut tidak diketahui Kasus 1: varians dari kedua populasi sama: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ Andaikan kita memiliki dua independen populasi yang saling independen, dengan nilai mean μ_1 dan μ_2 tidak diketahui; dan varians dari kedua populasi tersebut sama $\sigma_1^2 = \sigma_2^2 = \sigma^2$ serta tidak diketahui nilainya. Untuk kasus ini misalkan $\bar{x}_1, \bar{x}_2$ adalah sampel mean dan s_1^2, s_2^2 adalah sampel varians, maka kita dapat menggunakan standar error pooled seperti yang telah dibahas pada Bab 5.

$$S_{pooled}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$$SE_{pooled}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}} = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Hipotesa Null:

$$H_o : \mu_1 - \mu_2 = \Delta_0$$
$$H_1 : \mu_1 - \mu_2 \neq \Delta_0$$

Statistik uji yang digunakan adalah:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Z akan berdistribusi $N(0, 1)$. Dengan mensubstitusi σ dengan s_{pooled} akan mengubah statistik uji menjadi

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

T akan berdistribusi t dengan derajat kebebasan $n_1 + n_2 - 2$.

Contoh 6.10

Dua katalis sedang dianalisis untuk menentukan pengaruhnya terhadap mean hasil dari suatu proses kimia. Secara khusus, katalis 1 saat ini digunakan; tetapi katalis 2 dapat diterima. Karena katalis 2 lebih murah, katalis ini sebaiknya digunakan, jika tidak mengubah hasil proses. Sebuah tes dijalankan di pabrik percontohan dan menghasilkan data yang ditunjukkan pada Tabel 6.7. Gambar YY menyajikan plot probabilitas normal dan plot kotak komparatif data dari kedua sampel tersebut. Apakah ada perbedaan dalam hasil rata-rata? Gunakan $\alpha = 0,05$ dan asumsikan variansi yang sama.

Tabel 6.7. Data hasil observasi (Montgomery, p.385)

Observasi	Katalis 1	Katalis 2
1	91,50	89,19
2	94,18	90,95
3	92,18	90,46
4	95,39	93,21
5	91,79	97,19
6	89,07	97,04
7	94,72	91,07
8	89,21	92,75
	$\bar{x}_1 - 92,255$	$\bar{x}_2 - 92,733$
	$s_1 = 2,39$	$s_2 = 2,98$

Jawab

$$s^2_{pooled} = \frac{(8-1)(2,39)^2 + (8-1)(2,98)^2}{(8+8-2} = 7,30$$

$$s_{pooled} = \sqrt{7,30} = 2,70$$

$$t_0 = \frac{92,255 - 92.733}{2,70\sqrt{\frac{1}{8} + \frac{1}{8}}} = -0,35$$

Dari table student t diperoleh $t_{0,40;14} = 0,258$ dan $t_{0,25;14} = 0,692$ maka $0,258 < 0,35 < 0,692$. P-value berada di antara $0,50 < P - value < 0,80$. P-value $> 5\%$, maka kita gagal menolak hipotesa. Dapat disimpulkan bahwa rata-rata yield yang dihasilkan oleh kedua katalis ini tidak berbeda secara signifikan.

Bila diselesaikan secara manual. R-script dari masalah ini adalah sebagai berikut:

```
Katalis1 = c(91.50,94.18,92.18,95.39,91.79,89.07,94.72,89.21)
Katalis2 = c(89.19,90.95,90.46,93.21,97.19,97.04,91.07,92.75)

MKat1 = mean(Katalis1)
MKat2 = mean(Katalis2)
SKat1 = sd(Katalis1)
SKat2 = sd(Katalis2)
NKat1 = length(Katalis1)
NKat2 = length(Katalis2)
Sp2   = ((NKat1-1)*SKat1^2 + (NKat2-1)*SKat2^2)/(NKat1+NKat2-2)
Sp    = sqrt(Sp2)
```

```
SEp   = Sp*sqrt(1/NKat1 + 1/NKat2)

Thit  = (MKat1-MKat2)/SEp
Df    = NKat1+NKat2-2
pvalue= 2* pt(Thit,Df)
pvalue
```

```
[1] 0.7289136
```

Selain itu R- menyediakan fungsi t.test yang dapat digunakan secara langsung untuk menyelesaikan masalah ini:

```
options(width = 60)
t.test(Katalis1,Katalis2, alternative = "two.sided",var.equal = TRUE)
```

```
    Two Sample t-test

data:  Katalis1 and Katalis2
t = -0.35359, df = 14, p-value = 0.7289
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.373886  2.418886
sample estimates:
mean of x mean of y
  92.2550   92.7325
```

Dari hasil t.test diperoleh nilai p.value = 0.7289 dengan CI dari nilai beda (*difference*) berada pada selang [-3.373; 2.4188]. Terlihat bahwa nilai Nol berada di dalam selang kepercayaan tersebut. Hal ini menandakan bahwa nilai beda kedua katalis tersebut tidak berbeda secara signifikan.

Kasus 2: varians dari kedua populasi tidak sama; $\sigma_1^2 \neq \sigma_2^2$ Hipotesa Null:

$$H_o : \mu_1 - \mu_2 = \triangle_0$$

$$H_1 : \mu_1 - \mu_2 \neq \triangle_0$$

Statistik uji yang digunakan adalah:

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

akan berdistribusi t dengan derajat kebebasan sebagai berikut

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

Bulatkan nilai ν ke bilangan bulat terdekat, bila ν bukan bilangan bulat.

Contoh 6.11: Cokelat dan Kesehatan Kardiovaskular (Montgomery, p. 389)

Serafini et.al. (2003) menjelaskan sebuah percobaan di mana subjek mengonsumsi berbagai jenis coklat untuk mengetahui efek makan coklat untuk mengukur kesehatan jantung. Percobaan tersebut mempertimbangkan hasil hanya untuk coklat hitam dan coklat susu. Dalam percobaannya, 12 subjek mengonsumsi 100 gram coklat hitam dan 200 gram coklat susu, salah satu jenis coklat.per hari, dan setelah satu jam, total kapasitas antioksidan plasma darah mereka diukur dalam suatu pengujian. Subjek terdiri dari tujuh perempuan dan lima laki-laki dengan rentang usia rata-rata $32,2 \pm 1$ tahun, berat badan rata-rata $65,8 \pm 3,1$ kg, dan rata-rata indeks massa tubuh $21,9 \pm 0,4$ kg/m2. Data serupa dilaporkan pada artikel berikut.

Coklat hitam (Dark Chocolate)	118,8	122,6	115,6	113,6	119,5	115,9
	115,8	115,1	116,9	115,4	115,6	107,9
Coklat susu(Milk Chocolate)	102,1	105,8	99,6	102,7	98,8	100,9
	102,8	98,7	94,7	97,8	99,7	98.6

Apakah ada bukti yang mendukung klaim bahwa mengonsumsi coklat hitam menghasilkan rata-rata kadar darah total yang lebih tinggi kapasitas antioksidan plasma dibandingkan mengkonsumsi coklat susu? Misalkan μ_1 adalah rata-rata kapasitas antioksidan plasma darah yang dihasilkan dari mengonsumsi coklat hitam dan μ_2 adalah rata-rata kapasitas antioksidan plasma darah yang dihasilkan dari mengonsumsi susu cokelat.

Hipotesis yang ingin diuji:

$$H_o : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Rata-rata sampel yang diperoleh adalah: $\bar{x}_1 = 116,06$ dan $\bar{x}_2 = 100,19$ standar deviasi sampel: $s_1 = 3,53$ dan $s_2 = 2,89$ Perbedaan nilai rata-rata: $\bar{x}_1 - \bar{x}_2 = 15.87$

Statistik uji:

$$T_0^* = \frac{15,87}{\sqrt{\frac{12,46}{12} + \frac{8,35}{12}}} = 12,05$$

Derajat kebebasan:

$$\nu = \frac{\left(\frac{12,46}{12} + \frac{8,35}{12}\right)^2}{\frac{\left(\frac{12,46}{12}\right)^2}{11} + \frac{\left(\frac{8,35}{12}\right)^2}{11}} = 21,17$$

Derajat kebebasan $\nu = 21$.

```
#R-script
Dark = c(118.8,122.6,115.6,113.6,119.5,115.9,115.8,115.1,116.9,115.4,
         115.6,107.9)
Milk = c(102.1,105.8,99.6,102.7,98.8,100.9,102.8,98.7,94.7,97.8,99.7,
         98.6)

MDark = mean(Dark)
MMilk = mean(Milk)
VDark= var(Dark)
VMilk = var(Milk)
NDark= length(Dark)
NMilk = length(Milk)
Thit   = (MDark-MMilk)/sqrt(VDark/NDark + VMilk/NMilk)
Df1    = (VDark/NDark + VMilk/NMilk)^2
Df2    = ((VDark/NDark)^2/(NDark-1)) + ((VMilk/NMilk)^2/(NMilk-1))
Df     = round(Df1/Df2)
pvalue= pt(Thit,Df, lower.tail = FALSE)
pvalue
```

```
[1] 3.383454e-11
```

Diperoleh nilai P-value < 0.001. Kesimpulan tolak hipotesa null, rata-rata kapasitas antioksidan plasma darah yang dihasilkan dari coklat hitam lebih tinggi daripada yang dihasilkan oleh coklat susu.

Uji dengan menggunakan perintah t.test diperoleh hasil sebagai berikut:

```
t.test(Dark,Milk,alternative = "greater")
```

```
    Welch Two Sample t-test

data:  Dark and Milk
t = 12.048, df = 21.167, p-value = 3.053e-11
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 13.60845      Inf
sample estimates:
mean of x mean of y
 116.0583  100.1833
```

6.4.3 Uji Mean Berpasangan (*Paired t-test*)

Apabila observasi yang dilakukan pada dua populasi tersebut berpasangan maka data tersebut tidak lagi saling independent. Untuk data berpasangan, pengujian hipotesa dilakukan terhadap nilai beda (*difference*) dari pasangan-pasangan data tersebut.

Observasi	X_1	X_2	Beda	
1	X_{11}	X_{21}	$X_{11} - X_{21}$	D_1
2	X_{12}	X_{22}	$X_{12} - X_{22}$	D_2
3	X_{13}	X_{23}	$X_{13} - X_{23}$	D_3
...				
n	X_{1n}	X_{2n}	$X_{1n} - X_{2n}$	D_n

Null hypothesis:

$$H_0 : \mu_D = \triangle_0$$

Uji statistik:

$$T_0 = \frac{\bar{D} - \triangle_0}{\frac{S_D}{\sqrt{n}}}$$

Contoh 6.12: Parkir mobil parallel (Montgomery, p. 403)

Olson and Wachsler (1962) melaporkan sebuah penelitian di mana $n = 14$ orang diminta untuk memarkir paralel dua mobil yang jarak sumbu roda dan putarannya sangat berbeda jari-jari. Waktu dalam detik untuk setiap waktu yang diperlukan untuk memarkir mobil secara paralel dicatat dan diberikan pada table berikut.

Observasi	Mobil 1	Mobil 2	Beda
1	37,0	17,8	19,2
2	25,8	20,2	5,6
3	16,2	16,8	-0,6
4	24,2	41,4	-17,2
5	22,0	21,4	0,6
6	33,4	38,4	-5,0
7	23,8	16,8	7,0
8	58,2	32,2	26,0
9	33,6	27,8	5,8
10	24,4	23,2	1,2
11	23,4	29,6	-6,2
12	21,2	20,6	0,6
13	36,2	32,3	4,0
14	29,8	53,8	-24,0

Hipotesa yang ingin diuji:

$$H_0 : \mu_D = 0$$

Rata-rata beda: $\bar{D} = 1,21$
standar deviasi beda: $s_D = 12,68$

Statistik uji:

$$T_0 = \frac{1,21}{(12,68/\sqrt{14})} = 0,36$$

P-value = 0,64 Simpulan: Gagal tolak Ho. Tidak ada perbedaan waktu yang signifikan dalam memarkir secara parallel baik Mobil1 ataupun Mobil2.

```
Mobil1 = c(37,25.8,16.2,24.2,22,33.4,23.8,58.2,33.6,24.4,23.4,21.2,36.2,
           29.8)
Mobil2 = c(17.8,20.2,16.8,41.4,21.4,38.4,16.8,32.2,27.8,23.2,29.6,20.6,32.3,
           53.8)
Diff   = Mobil1 - Mobil2
MDiff  = mean(Diff)
SDiff  = sd(Diff)
NDiff  = length(Diff)
Thit   = MDiff/(SDiff/sqrt(NDiff))
Df     = NDiff-1
pvalue = pt(Thit,Df)
pvalue
```

```
[1] 0.6362668
```

```
t.test(Mobil1,Mobil2, paired = TRUE)
```

```
    Paired t-test

data:  Mobil1 and Mobil2
t = 0.35612, df = 13, p-value = 0.7275
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.115959  8.530245
sample estimates:
mean of the differences
               1.207143
```

Referensi

BPS, 2023, "Surabaya dalam angka 2023", https://surabayakota.bps.go.id/id/publication/2023/02/28/219438c973b16c7c80f11868/kota-surabaya-dalam-angka-2023.html

Detik.com: 16 Januari 2024: Perjalanan kasus Mary Jane Terpidana Mati hingga ibunya memohon Jokowi bebaskan: https://www.detik.com/jogja/berita/d-7145156/perjalanan-kasus-mary-jane-terpidana-mati-hingga-ibunya-mohon-jokowi-bebaskan

Montgomery, D.C, and Runger, G.C., Applied Statistics and Probability for Engineeris, (2011), 6th Edition, Wiley,

Olson, P.L and Wachsler, R.A., "Relative controlling of dissimilar cars", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 4(6), pp. 375-380, (1962).

Serafini, M., Bugianesi, R., Maiani, G. *et al.* "Plasma antioxidants from chocolate". *Nature* **424**, 1013 (2003). https://doi.org/10.1038/4241013a

Referensi

Bickel, P.J., Hammel, E. A., and O'Connell, J.W., (1975), 'Sex bias in graduate admissions: Data from Berkeley', *Science*, **187** (4175): 398-404.

BPS, 2023, "Surabaya dalam angka 2023", https://surabayakota.bps.go.id/id/publication/2023/02/28/219438e973b16c7c80f11868/kota-surabaya-dalam-angka-2023.html

Cortez, P. and Silva, A., Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. Available at: https://www.kaggle.com/ishandutta/student-performance-data-set

De Veaux, R., Velleman, P., and Bock D., (2016), *Stats: Data and Models*, 5th Eds. Pearson

Detik.com: 16 Januari 2024: Perjalanan kasus Mary Jane Terpidana Mati hingga ibunya memohon Jokowi bebaskan: https://www.detik.com/jogja/berita/d-7145156/perjalanankasus-mary-jane-terpidana-mati-hingga-ibunya-mohon-jokowi-bebaskan

Encyclopedia-Titanica, Wikipedia: https://www.encyclopedia-titanica.org/

Florence Nightingale, Wikipedia: https://en.wikipedia.org/wiki/Florence_Nightingale,

Freedman D., Pisani R., and Purves, R., (2007) *Statistics*, 4th Eds, W.W. Norton

Hyndman, R. J. and Fan, Y. (1996) 'Sample quantiles in statistical packages', *American Statistician* **50**: 361–365.

Jakarta Stock Exchange Index, Finance yahoo, https://finance.yahoo.com/quote/%5EJKSE

Kerns, G. J., (2010), *Introduction to Probability and Statistics using R*, GNU Free documentation Licence.

Montgomery, D.C., and Runger, G.C., (2018) *Applied Statistics and Probability for Engineers*, 7th Eds., Wiley, USA.

Old faithful Geyser: https://www.frommers.com/slideshows/848448-beyond-old-faithful-a-geyser-gazing-guide-to-yellowstone-national-park

Olson, P.L and Wachsler, R.A., "Relative controlling of dissimilar cars", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 4(6), pp. 375-380, (1962).

R-Graphics, open source: https://r-graphics.org/

Rice, J.A., (2006), *Mathematical Statistics & Data Analysis*, Duxbury Press

Serafini, M., Bugianesi, R., Maiani, G. *et al.* "Plasma antioxidants from chocolate". *Nature* **424**, 1013 (2003). https://doi.org/10.1038/4241013a213

Simpson Paradox: https://plato.stanford.edu/entries/paradox-simpson/

Penerbit:
Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM)
Universitas Kristen Petra
Jl. Siwalankerto 121-131, Surabaya 60236, Indonesia

ISBN 978-623-5457-19-2 (PDF)
9 786235 457192