Modul IS240 Lab 4: Distribusi Peluang

Prodi Sistem Informasi

1 Pendahuluan

1.1 Tujuan Pembelajaran

- Mahasiswa bisa menghitung distribusi peluang Binomial dan Gaussian (Normal).
- Mahasiswa bisa membuat grafik distribusi peluang Binomial dan Gaussian.
- Mahasiswa bisa melakukan uji normalitas.

1.2 Topik Praktikum

- Distribusi peluang Binomial
- Distribusi peluang Gaussian (Normal)
- Uji normalitas

1.3 Petunjuk Umum

Gunakan file tutorial/ cheat sheet sebagai bahan referensi anda. File-file tersebut dapat diunduh dari Dropbox folder TutorialManual atau dari elearning.
- **2RTutorialTTH2018DescriptiveStatistics.pdf**, - **2RTutorialTTHHypothesisTesting.pdf**

2 Materi Praktikum (Baca sebelum mengerjakan soal.)

2.1 Distribusi Binomial

Distribusi Binomial adalah sebuah distribusi diskrit dengan ciri-ciri sebagai berikut.
* n = banyaknya trial dalam sebuah eksperimen. * X = variabel acak yang menyatakan banyaknya sukses x dalam n trial. * p = peluang sukses (peluang memperoleh sebuah sukses). Peluang ini harus konstan dalam sebuah trial.

Rumus Distribusi Binomial

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$E(X = x) = np$$

$$Var(X) = np(1-p)$$

2.1.1 $P(X = x)$

Contoh:
Sebuah kuis terdiri dari 10 soal pilihan ganda. Setiap pertanyaan bersifat independen (tidak bergantung pada pertanyaan lain) dan memiliki 4 pilihan jawaban. Hitung peluang seorang siswa menjawab 8 pertanyaan dengan benar.

Jawaban:
Soal ini merupakan soal distribusi Binomial $P(X = 8)$ di mana:

- X = banyaknya pertanyaan yang dijawab dengan benar.
- $n = 10$
- $x = 8$
- $p = 0,25$

$P(X = 8) = \binom{10}{8}(0,25)^8(0,75)^{10-2} = 0,000386$

```
dbinom(x = 8, size = 10, prob = 0.25)
```

```
## [1] 0.0003862381
```

2.1.2 $P(X \leq x)$

Contoh:
Hitung peluang seorang siswa menjawab maksimal 8 pertanyaan dengan benar bila suatu kuis pilihan ganda terdiri dari 10 pertanyaan dan tiap pertanyaan memiliki 4 pilihan jawaban.

Jawaban: Soal ini merupakan soal distribusi Binomial $P(X \leq 8)$ di mana:

- X = banyaknya pertanyaan yang dijawab dengan benar.
- $n = 10$
- $x = 8$
- $p = 0,25$

$P(X \leq 8) = P(X = 0) + P(X = 1) + \cdots + P(X = 8)$.
Masing-masing suku $P(X = 0)$ dan seterusnya dihitung dengan rumus $P(X = x)$.

```
#peluang P(X = 0) sampai dengan P(X = 8)
dbinom(x = 0:8, size = 10, prob = 0.25)
```

```
## [1] 0.0563135147 0.1877117157 0.2815675735 0.2502822876 0.1459980011
## [6] 0.0583992004 0.0162220001 0.0030899048 0.0003862381
```

```
#P(X <= 8)
sum(dbinom(x = 0:8, size = 10, prob = 0.25))
```

```
## [1] 0.9999704
```

```
#dapat juga dihitung dengan menggunakan fungsi pbinom
pbinom(q = 8, size = 10, prob = 0.25, lower.tail = TRUE)
```

```
## [1] 0.9999704
```

```r
#lower.tail = TRUE merupakan default value pada fungsi pbinom
pbinom(q = 8, size = 10, prob = 0.25)
```

```
## [1] 0.9999704
```

Opsi *lower.tail* pada pbinom akan melakukan perhitungan untuk $P(X \leq 8)$.

2.1.3 $P(X > x)$

$P(X > x) = 1 - P(X \leq x)$

Contoh:
Hitung peluang seorang siswa menjawab lebih dari 8 pertanyaan dengan benar bila suatu kuis pilihan ganda terdiri dari 10 pertanyaan dan tiap pertanyaan memiliki 4 pilihan jawaban.

```r
pbinom(q = 8, size = 10, prob = 0.25, lower.tail = FALSE)
```

```
## [1] 2.95639e-05
```

```r
#atau
1 - pbinom(q = 8, size = 10, prob = 0.25, lower.tail = TRUE)
```

```
## [1] 2.95639e-05
```

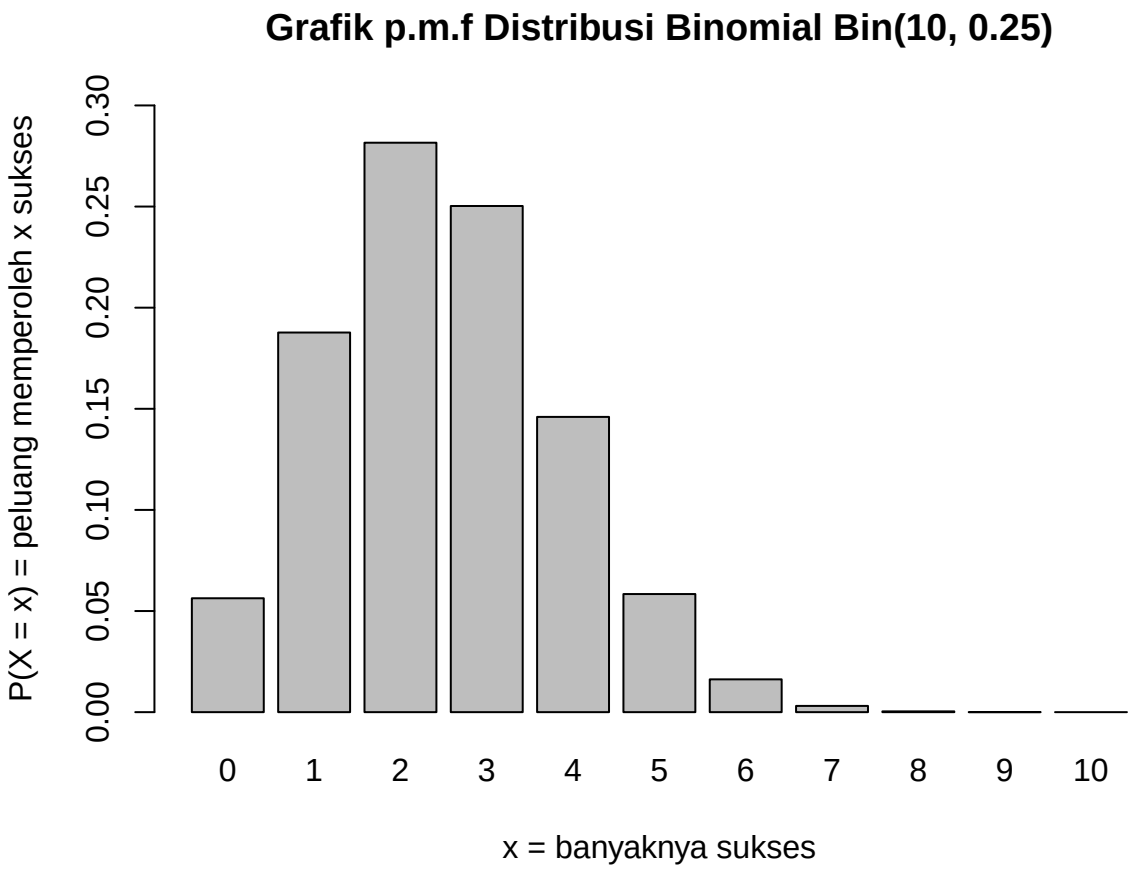
```r
#atau
1 - pbinom(q = 8, size = 10, prob = 0.25)
```

```
## [1] 2.95639e-05
```

2.1.4 Grafik fungsi p.m.f. (probability mass function) Distribusi Binomial

Grafik fungsi p.m.f. Binomial menunjukkan distribusi variabel acak Binomial. Dari grafik p.m.f. dapat dilihat bahwa peluang $P(X = x)$ paling tinggi adalah $P(X = 2)$ dan $P(X = 3)$. Nilai x dengan peluang paling tinggi berada di sekitar mean.
Untuk soal di atas, mean $= E(X) = np = (10)(0,25) = 2,5$.

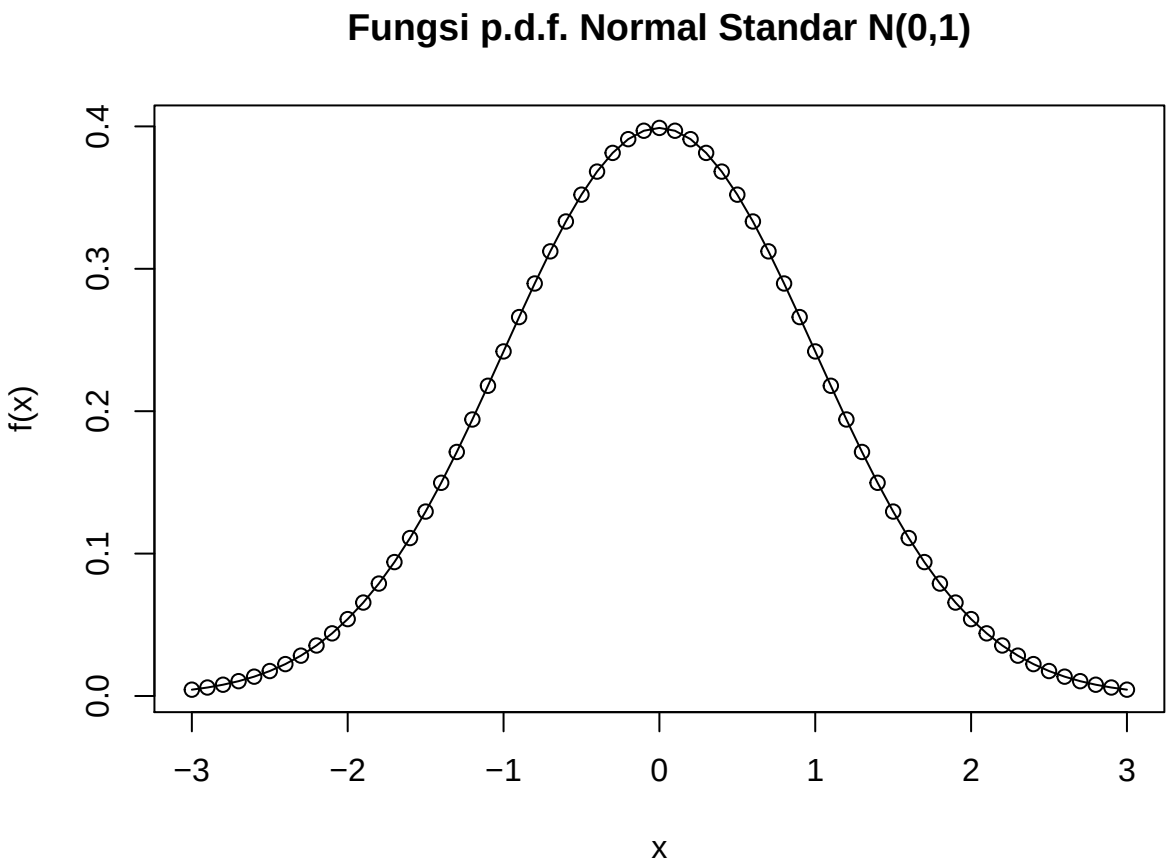
```r
successcount <- 0:10
probx <- dbinom(x = successcount, size = 10, prob = 0.25)
#menyatukan x dan prob dalam sebuah data frame
dat <- data.frame(successcount, probx)

#persiapan membuat bar chart
labelx <- "x = banyaknya sukses"
labely <- "P(X = x) = peluang memperoleh x sukses"
judul <- "Grafik p.m.f Distribusi Binomial Bin(10, 0.25)"
batasy <- c(0, 0.3)
barplot(dat$probx, names.arg = successcount, xlab = labelx, ylab = labely,
  ylim = batasy, main = judul)
```



2.2 Distribusi Gaussian (Distribusi Normal)

Distribusi Gaussian atau distribusi Normal memiliki mean μ dan varians σ^2.

```
x <- seq(from = -3, to = 3, by = 0.1)
mu <- 0
sigma <- 1
y <- dnorm(x, mean = mu, sd = sigma)
dat <- data.frame(x, y)
labelx <- "x"
labely <- "f(x)"
judul <- "Fungsi p.d.f. Normal Standar N(0,1)"
plot(dat$x, dat$y, type = "p",  main = judul, xlab = labelx, ylab = labely) #plot point
lines(dat$x, dat$y) #plot line
```



2.2.1 $P(Z \leq z)$ di mana $Z \sim N(0, 1^2)$

Diketahui sebuah variabel acak (random variabel) Z mengikuti distribusi Gaussian dengan $\mu = 0$ dan simpangan baku $\sigma = 1$. Hitung peluang $P(Z \leq 0)$ untuk $Z \sim N(0, 1^2)$.

```
pnorm(q = 0, mean = 0, sd = 1, lower.tail = TRUE)
```

```
## [1] 0.5
```

```
#atau
pnorm(0)
```

```
## [1] 0.5
```

2.2.2 $P(Z > z)$ di mana $Z \sim N(0, 1^2)$

Diketahui sebuah variabel acak (random variabel) Z mengikuti distribusi Gaussian dengan $\mu = 0$ dan simpangan baku $\sigma = 1$. Hitung peluang $P(Z > 2)$.

```r
pnorm(q = 2, mean = 0, sd = 1, lower.tail = FALSE)
```

```
## [1] 0.02275013
```

```r
pnorm(q = 2, lower.tail = FALSE) #atau
```

```
## [1] 0.02275013
```

2.2.3 $P(X \leq x)$ di mana $X \sim N(\mu, sigma^2)$

Diketahui sebuah variabel acak (random variabel) X mengikuti distribusi Gaussian dengan $\mu = 80$ dan simpangan baku $\sigma = 10$. Hitung peluang $P(X \leq 90)$.

```r
pnorm(q = 90, mean = 80, sd = 10, lower.tail = TRUE)
```

```
## [1] 0.8413447
```

2.2.4 $P(x_1 \leq X \leq x_2)$

Diketahui sebuah variabel acak (random variabel) X mengikuti distribusi Gaussian dengan $\mu = 80$ dan simpangan baku $\sigma = 10$. Hitung peluang $P(60 \leq X \leq 90)$.

Jawaban:
$P(60 \leq X \leq 90) = P(X \leq 90) - P(X \leq 60)$

```r
mu <- 80
sigma <- 10
pnorm(q = 90, mean = mu, sd = sigma)
```

```
## [1] 0.8413447
```

```r
pnorm(q = 60, mean = mu, sd = sigma)
```

```
## [1] 0.02275013
```

```r
(ans <- pnorm(q = 90, mean = mu, sd = sigma) - pnorm(q = 60, mean = mu, sd = sigma))
```

```
## [1] 0.8185946
```

2.3 Pengecekan Normalitas

Pengecekan normalitas untuk sebuah variabel numerik dapat dilakukan dengan cara: 1. menggambarkan grafik distribusi a. histogram b. density plot c. Q-Q plot 2. melakukan uji hipotesa normalitas

Terdapat beberapa uji normalitas.

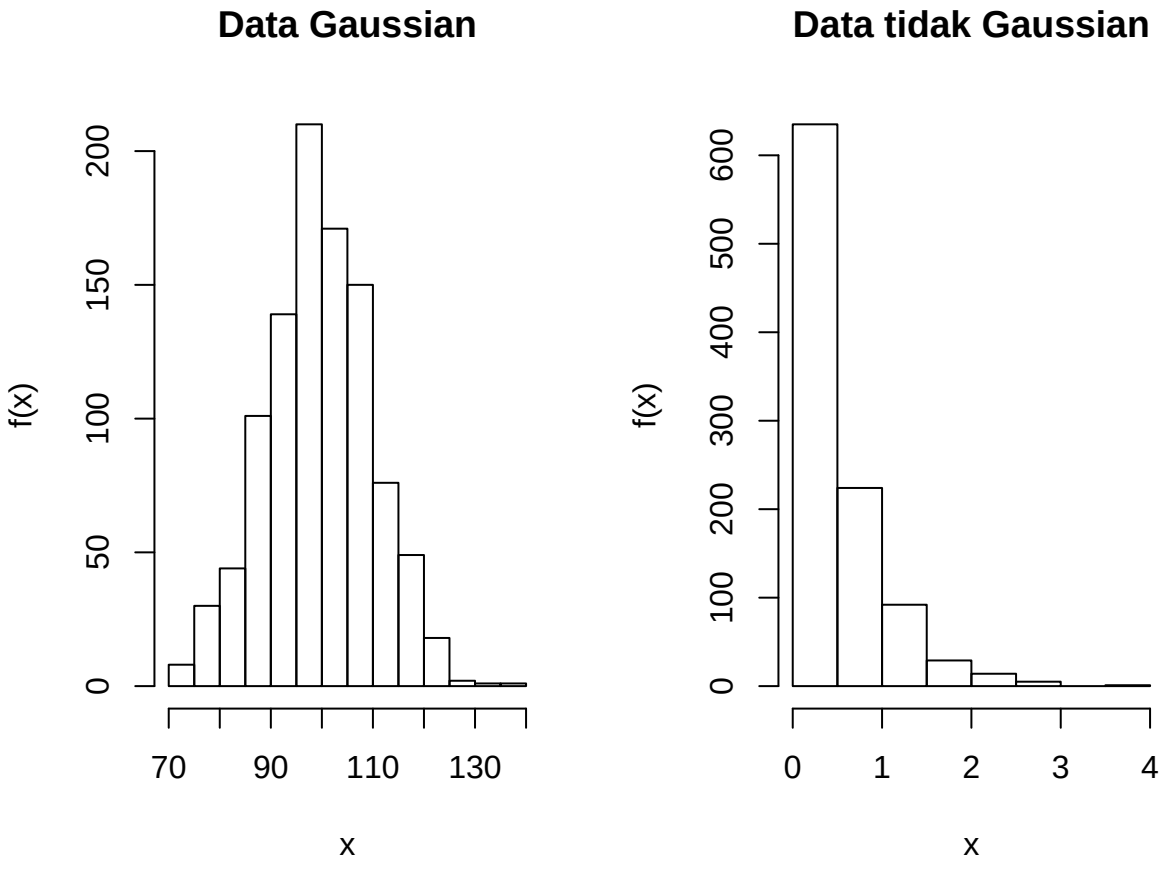
- Shapiro-Wilk
- Anderson-Darling

- Shapiro-Francia

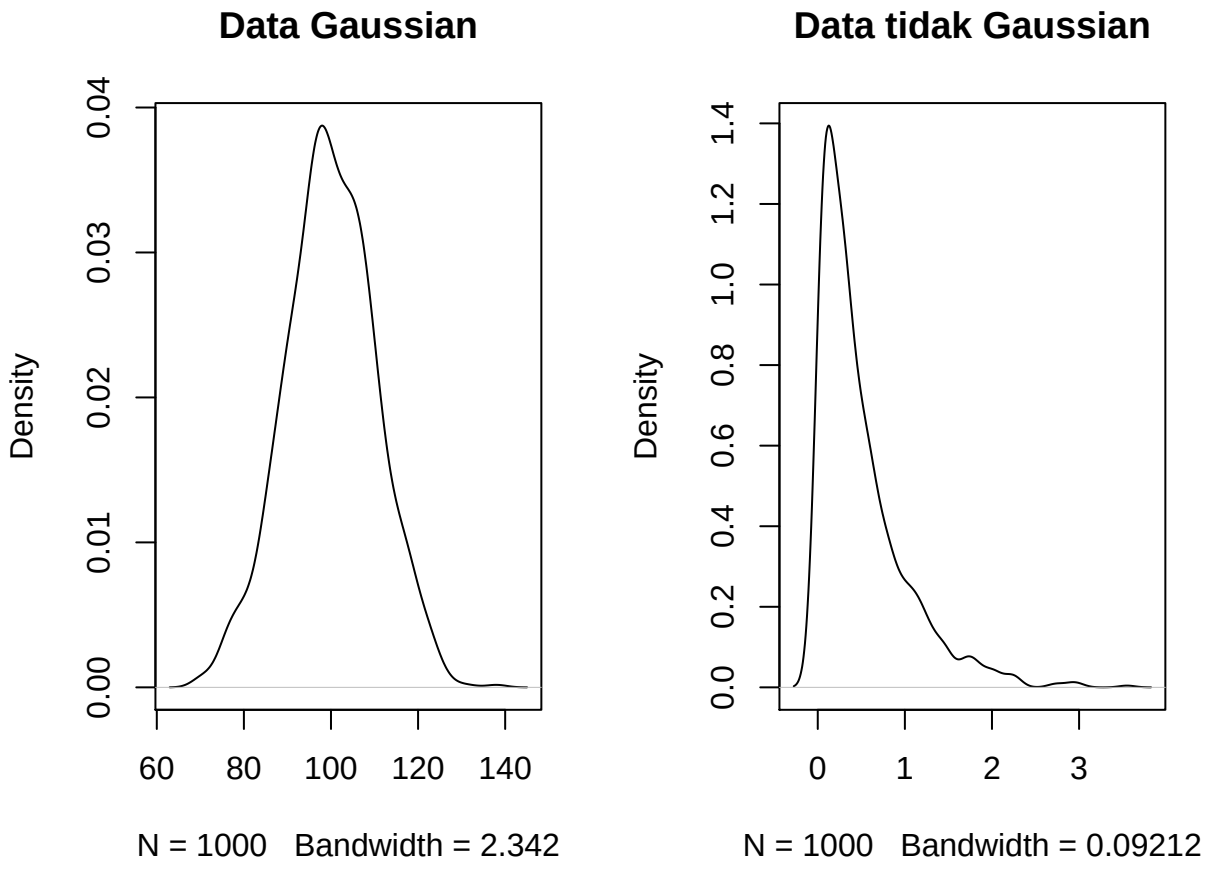
H_0: data mengikuti distribusi Gaussian (distribusi Normal).
H_a: data tidak mengikuti distribusi Gaussian.

Hipotesa null H_0 akan ditolak pada tingkat signifikansi α bila p-value $< \alpha$.
Bila p-value $< \alpha$, kita simpulkan data tidak mengikuti distribusi Gaussian.
Bila p-value $>= \alpha$, kita simpulkan data mengikuti distribusi Gaussian.

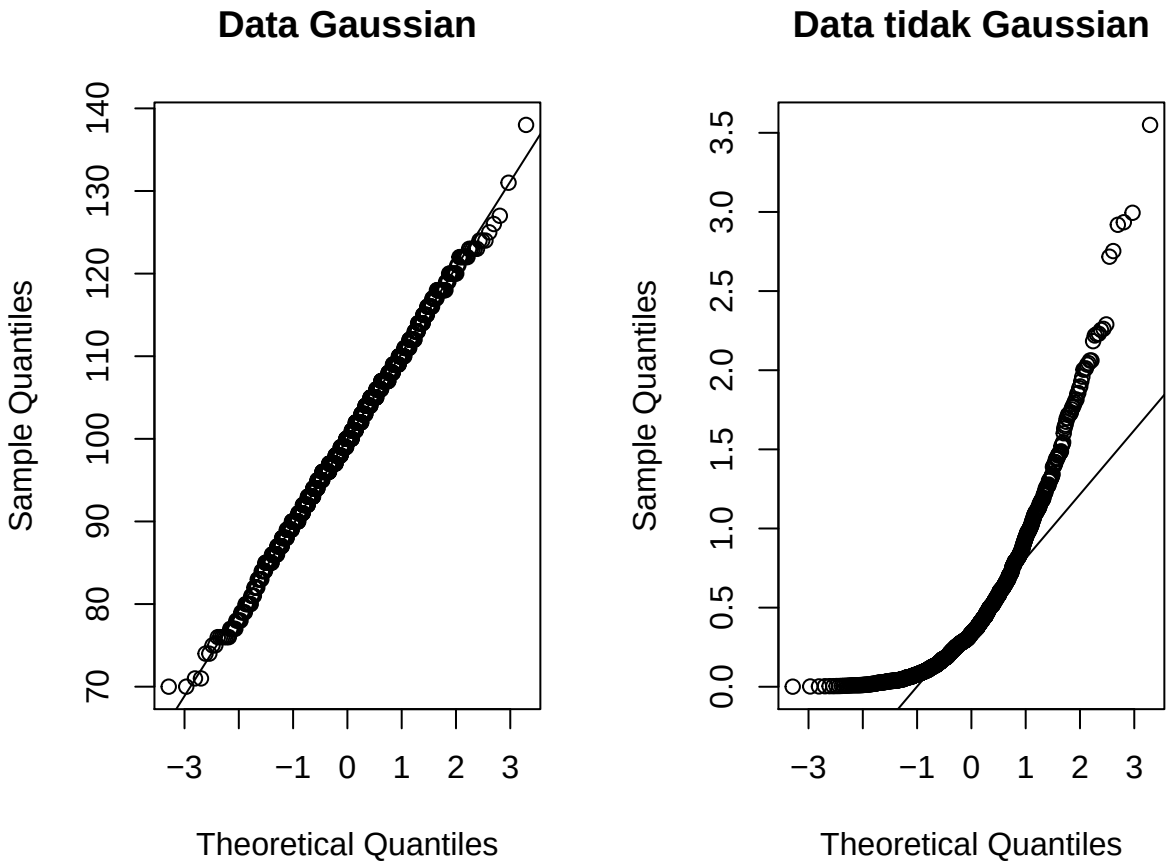
```
set.seed(1)
normaldata <- round(rnorm(n = 1000, mean = 100, sd = 10))
exponentialdata <- rexp(n = 1000, rate = 2)

judulnormal <- "Data Gaussian"
judultdknormal <- "Data tidak Gaussian"
labelx <- "x"
labely <- "f(x)"

par(mfrow = c(1,2)) #set up agar 2 plot per baris
hist(normaldata, main = judulnormal, xlab = labelx, ylab = labely)
hist(exponentialdata, main = judultdknormal, xlab = labelx, ylab = labely)
```



```
plot(density(normaldata), main = judulnormal)
plot(density(exponentialdata), main = judultdknormal)
```




```
qqnorm(normaldata, main = judulnormal)
qqline(normaldata)
qqnorm(exponentialdata, main = judultdknormal)
qqline(exponentialdata)
```



- Histogram, density plot dan Q-Q plot data pada kolom sebelah kiri menunjukkan data mengikuti distribusi Gaussian (distribusi Normal).
- Histogram, density plot dan Q-Q plot data distribusi Eksponensial pada kolom sebelah kanan menunjukkan data tidak mengikuti distribusi Gaussian.

Kita dapat melakukan uji hipotesa normalitas agar lebih yakin akan normalitas data.

```
(test1 <- stats::shapiro.test(normaldata))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  normaldata
## W = 0.99797, p-value = 0.2701
```

```
(test2 <- nortest::ad.test(normaldata))
```

```
##
##  Anderson-Darling normality test
##
## data:  normaldata
## A = 0.58755, p-value = 0.1255
```

```
(test3 <- nortest::sf.test(normaldata))
```

```
##
##  Shapiro-Francia normality test
##
## data:  normaldata
## W = 0.99802, p-value = 0.2591
```

- p-value untuk uji hipotesa Shapiro-Wilk adalah 0.2701
- p-value untuk uji hipotesa Anderson-Darling adalah 0.1255
- p-value untuk uji hipotesa Shapiro-Francia adalah 0.2591

Kesimpulan:
p-value pada ketiga test di atas $> \alpha = 0,05$ sehingga kita tidak cukup bukti untuk menolak hipotesa null normalitas data.
Kita simpulkan data normaldata mengikuti distribusi Gaussian.

```
(test1 <- stats::shapiro.test(exponentialdata))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  exponentialdata
## W = 0.8209, p-value < 2.2e-16
```

```
(test2 <- nortest::ad.test(exponentialdata))
```

```
##
##  Anderson-Darling normality test
##
## data:  exponentialdata
## A = 47.983, p-value < 2.2e-16
```

```
(test3 <- nortest::sf.test(exponentialdata))
```

```
##
##  Shapiro-Francia normality test
##
## data:  exponentialdata
## W = 0.82044, p-value < 2.2e-16
```

- p-value untuk uji hipotesa Shapiro-Wilk adalah $5.9312488 \times 10^{-32}$.
- p-value untuk uji hipotesa Anderson-Darling adalah 3.7×10^{-24}.
- p-value untuk uji hipotesa Shapiro-Francia adalah $2.3719438 \times 10^{-28}$.

Kesimpulan:
p-value pada ketiga test di atas sangat kecil $< \alpha = 0,05$ sehingga kita memiliki cukup bukti untuk menolak hipotesa null normalitas data.
Kita simpulkan data exponentialdata **tidak** mengikuti distribusi Gaussian.

3 TIPS R

3.1 Ubah Tipe Variabel

Untuk analisa dengan menggunakan software R, variabel kategorikal yang bukan merupakan nama biasanya diubah menjadi tipe **factor** agar perintah R seperti plot dapat membedakan tipe variabel.

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
head(mtcars, 3)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

Dalam data mtcars di atas, variabel *cyl* merupakan variabel kategorikal yang diberi kode angka dan dimasukkan sebagai variabel numerik. Sebaiknya variabel cyl, vs, am, gear, dan carb diubah menjadi tipe **factor**. Pastikan anda save hasil konversi ke dalam data frame anda.

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$am <- as.factor(mtcars$am)
# dan seterusnya
```

3.2 Mengakses obyek hasil test

```
(test1 <- stats::shapiro.test(mtcars$mpg))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.94756, p-value = 0.1229
```

```r
#menambahkan tanda kurung di pada awal dan akhir perintah R
#akan menampilkan output ke console.

#tampilkan obyek test1
names(test1)
```

```
## [1] "statistic" "p.value"   "method"    "data.name"
```

```r
#akses p.value
test1$p.value
```

```
## [1] 0.1228814
```

```r
#akses test statistic
test1$statistic
```

```
##         W
## 0.9475647
```