STK335 Analisis Eksplorasi Data

Pertemuan 03

Pemeriksaan Sebaran Data

Bagus Sartono

Outline

- **Quantile-Quantile Plot**
  - Apa itu kuantil?
  - Plot kuantil
  - QQplot
  - QQplot Normal
  - QQplot selain normal

- **Goodness of Fit Test**
  - Chi-Square Test
  - Kolmogorov-Smirnov Test

Persentil dan Kuantil

- Persentile ke-k dari sebuah dataset adalah sebuah nilai yang membagi sedemikian rupa sehingga terdapat k% amatan yang kurang dari nilai tersebut dan (100-k)% amatan bernilai lebih besar dari nilai persentil tersebut
  - Persentil ke-25 disebut juga sebagai lower quartile atau Q1
  - Persentil ke-50 disebut juga sebagai median
  - Persentil ke-75 disebut juga sebagai upper quartile atau Q3

- Dalam analisis statistik, istilah kuantil lebih umum digunakan dibandingkan persentil, meskipun maknanya sama. Hanya saja sering digunakan indeks yang berbeda.
  - P25 → Q(0.25)
  - P50 → Q(0.5)
  - P75 → Q(0.75)

Kuantil

- Misalkan ada dataset berikut

  3.7 2.7 3.3 1.3 2.2 3.1

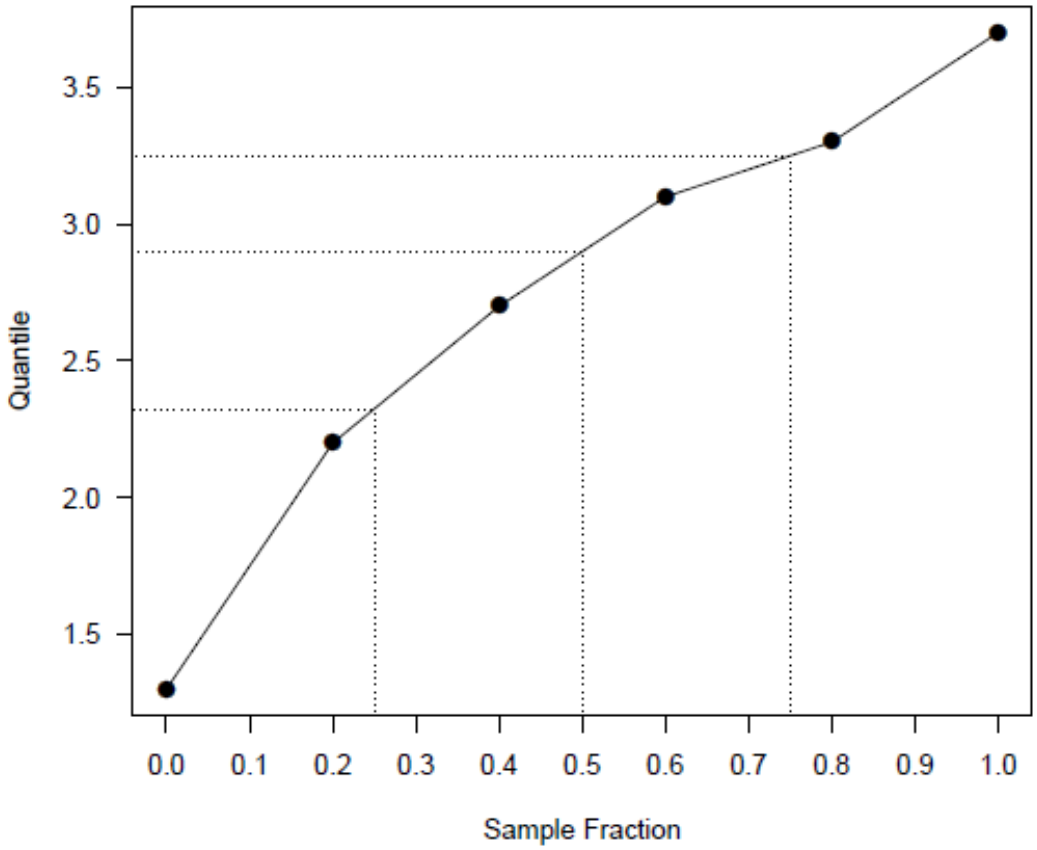
- Pertama urutkan datanya

  1.3 2.2 2.7 3.1 3.3 3.7

- Padankan setiap nilai yang terurut dengan bilangan fraksi antara 0 dan 1 dengan jarak yang sama

Sample fraction	0	.2	.4	.6	.8	1
Quantile	1.3	2.2	2.7	3.1	3.3	3.7

Kuantil



Kuantil yang lain diperoleh menggunakan interpolasi linear

Kuantil

- Andaikan terdapat suatu gugus data $x_1, x_2, \ldots, x_n$. Kuantil dengan fraksi tertentu diperoleh dengan cara sebagai berikut:
  - Urutkan datanya $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
  - Setiap data yang terurut merupakan kuantil yang bersesuaian dengan fraksi

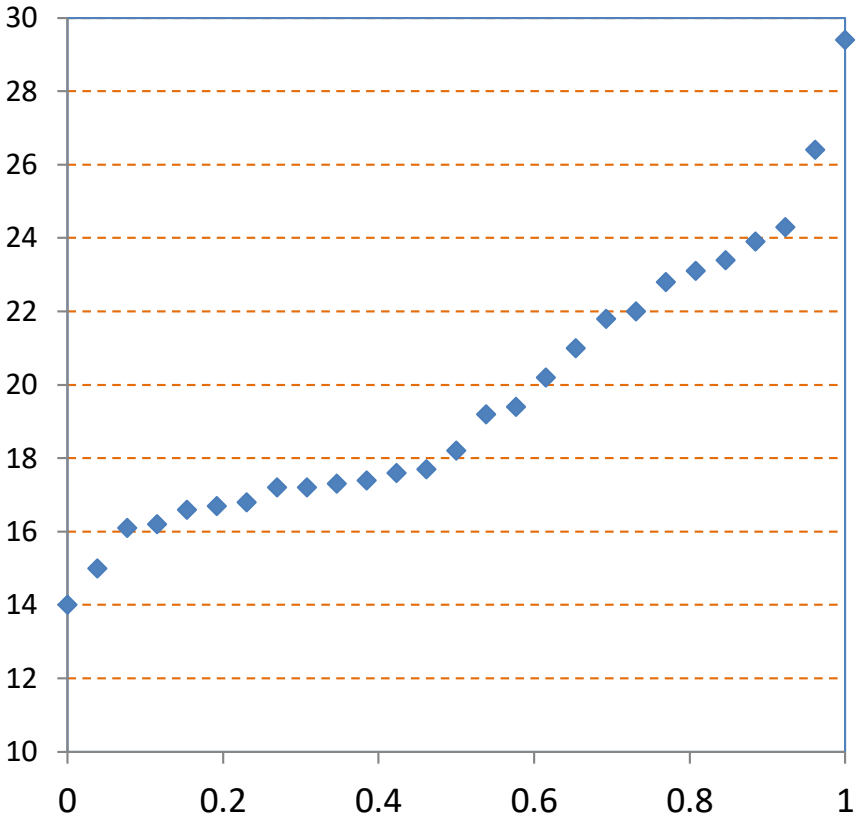
$$p_i = \frac{i-1}{n-1}$$

    untuk i = 1, . . . , n
  - Kuantil untuk fraksi lain diperoleh dengan melakukan interpolasi linear

Plot Kuantil

- Merupakan plot antar nilai kuantil dan fraksinya
- Serupa dengan plot dari fungsi sebaran kumulatif empirik (menukar sumbu)

24.3	14.0	1	14.0	0.0000
17.7	15.0	2	15.0	0.0385
23.4	16.1	3	16.1	0.0769
20.2	16.2	4	16.2	0.1154
22.8	16.6	5	16.6	0.1538
16.1	16.7	6	16.7	0.1923
22.0	16.8	7	16.8	0.2308
21.8	17.2	8	17.2	0.2692
17.6	17.2	9	17.2	0.3077
16.7	17.3	10	17.3	0.3462
18.2	17.4	11	17.4	0.3846
14.0	17.6	12	17.6	0.4231
29.4	17.7	13	17.7	0.4615
19.4	18.2	14	18.2	0.5000
16.2	19.2	15	19.2	0.5385
16.6	19.4	16	19.4	0.5769
17.4	20.2	17	20.2	0.6154
23.9	21.0	18	21.0	0.6538
19.2	21.8	19	21.8	0.6923
17.3	22.0	20	22.0	0.7308
16.8	22.8	21	22.8	0.7692
21.0	23.1	22	23.1	0.8077
15.0	23.4	23	23.4	0.8462
17.2	23.9	24	23.9	0.8846
26.4	24.3	25	24.3	0.9231
23.1	26.4	26	26.4	0.9615
17.2	29.4	27	29.4	1.0000



Plot QQ

- Plot Kuantil-Kuantil
- Theoretical QQ Plot
- Scatter plot antara quantil data dengan quantil berdasarkan sebaran hipotetik tertentu
- Digunakan untuk mengidentifikasi apakah sebaran data mengikuti sebaran hipotetik yang digambarkan
- Pola garis lurus mengindikasikan hal tersebut

Plot QQ

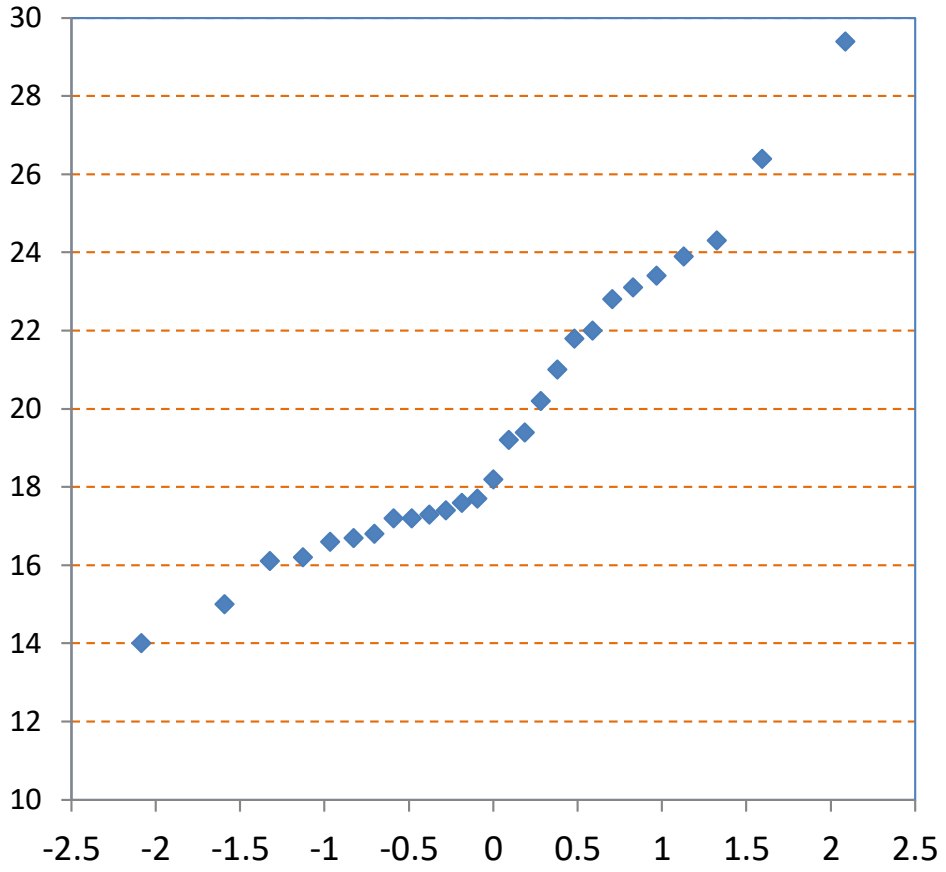
- Tahapan pembuatan
  - Urutkan data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
  - Hitung pi = (i – 0.5)/n
  - Untuk sebaran hipotetik tertentu, hitung $Q_i = F^{-1}(p_i)$ dengan F adalah fungsi sebaran kumulatif, dengan kata lain Qi adalah sebuah nilai sehingga $P(Y \leq Q_i) = p_i$
  - Plot $x_{(i)}$ vs Q_i

Plot QQ Normal

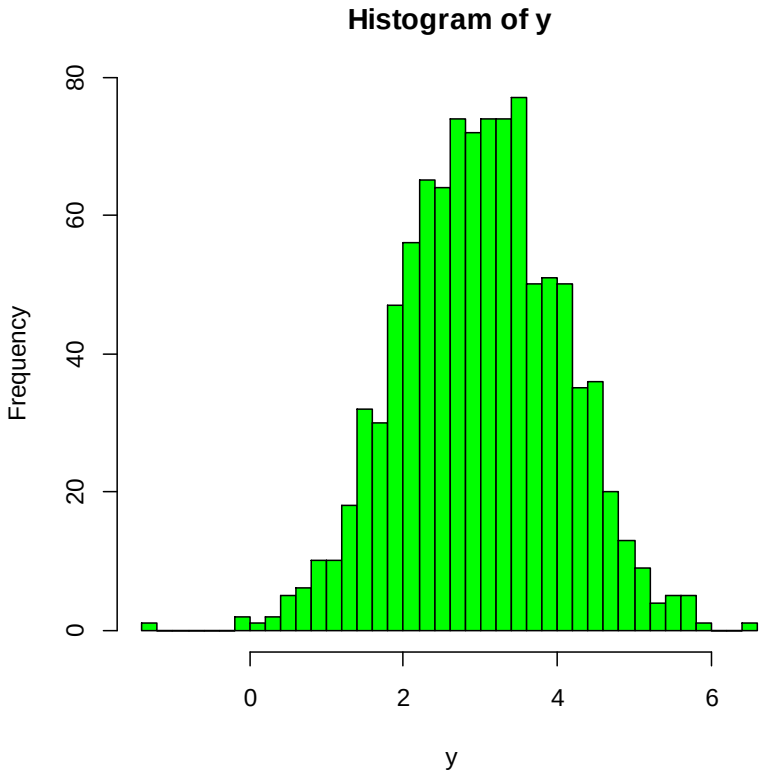
- Tahapan pembuatan
  - Urutkan data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
  - Hitung pi = (i – 0.5)/n
  - Tentukan skor normal Z, untuk setiap p_i
  - Plot $x_{(i)}$ vs Z_i

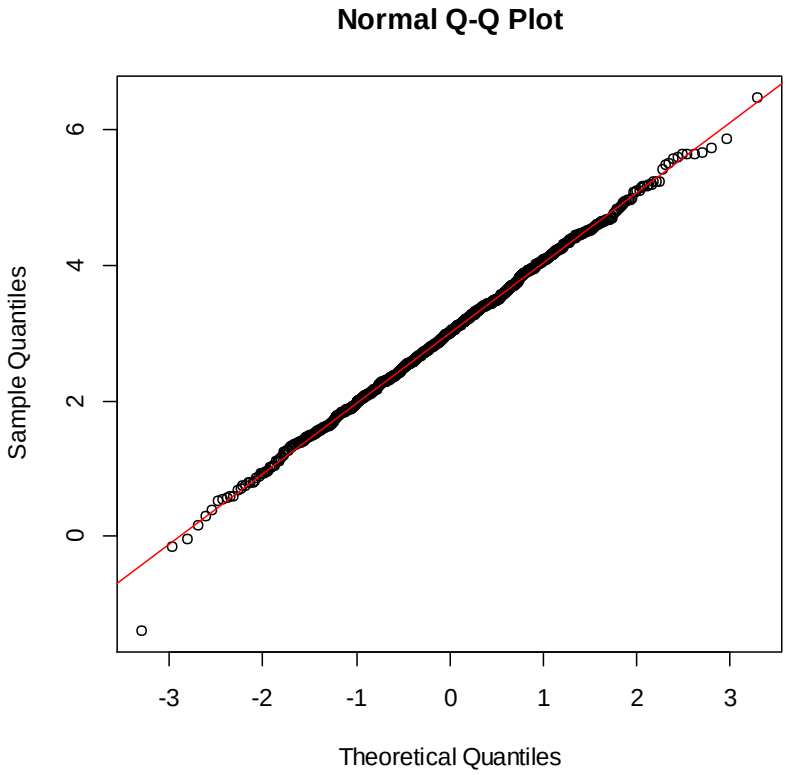
- Digunakan untuk melihat apakah distribusi data mengikuti sebaran normal

1	14.0	0.0185	-2.08536
2	15.0	0.0556	-1.59322
3	16.1	0.0926	-1.32496
4	16.2	0.1296	-1.12814
5	16.6	0.1667	-0.96742
6	16.7	0.2037	-0.82846
7	16.8	0.2407	-0.70392
8	17.2	0.2778	-0.58946
9	17.2	0.3148	-0.48225
10	17.3	0.3519	-0.38033
11	17.4	0.3889	-0.28222
12	17.6	0.4259	-0.18676
13	17.7	0.4630	-0.09297
14	18.2	0.5000	-1.4E-16
15	19.2	0.5370	0.092972
16	19.4	0.5741	0.186756
17	20.2	0.6111	0.282216
18	21.0	0.6481	0.380326
19	21.8	0.6852	0.482248
20	22.0	0.7222	0.589456
21	22.8	0.7593	0.703922
22	23.1	0.7963	0.828465
23	23.4	0.8333	0.967422
24	23.9	0.8704	1.128144
25	24.3	0.9074	1.324958
26	26.4	0.9444	1.593219
27	29.4	0.9815	2.085356



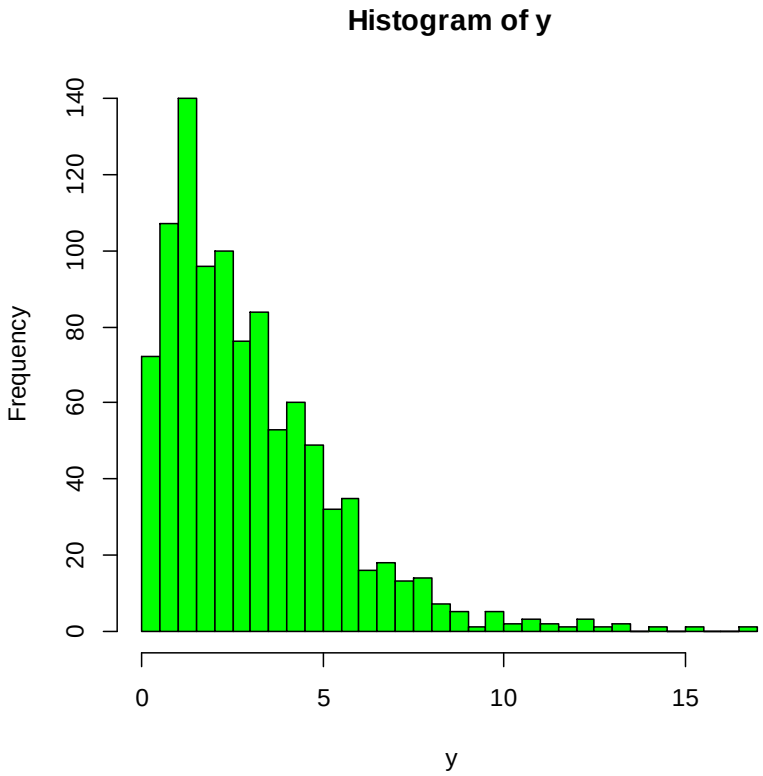
QQPlot Normal untuk Data yang Mengikuti Sebaran Normal

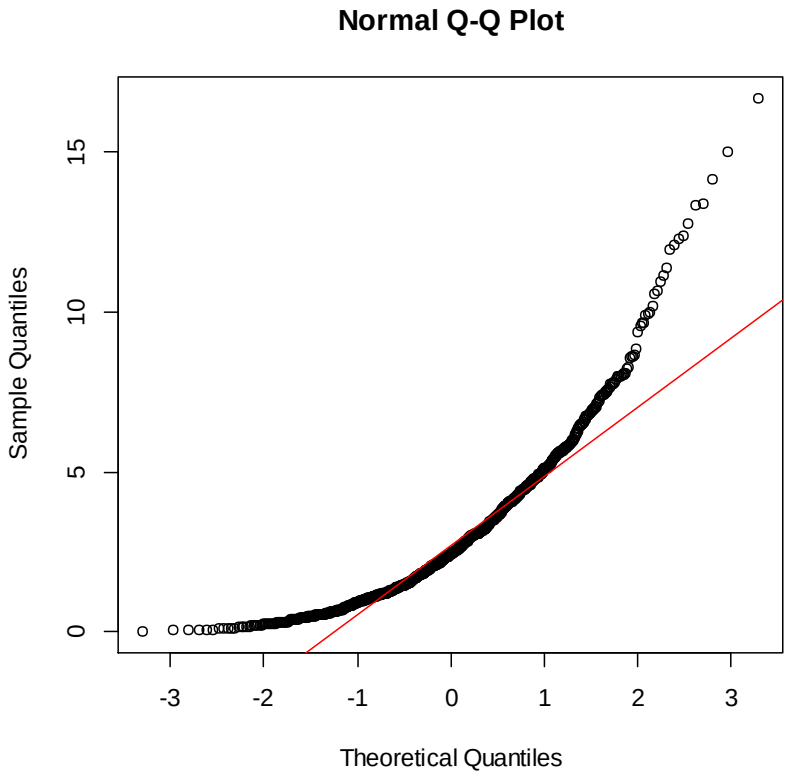





```
y <- rnorm(1000, 3,1)
hist(y, breaks=30, col="green")
qqnorm(y)
qqline(y, col = "red")
```

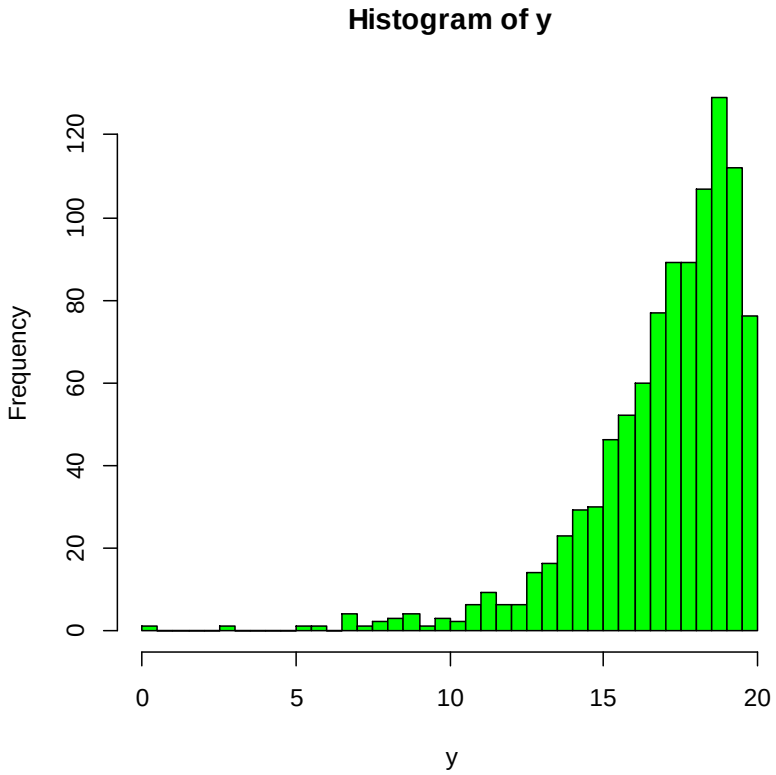
QQPlot Normal untuk Data yang Sebarannya Menjulur ke Kanan

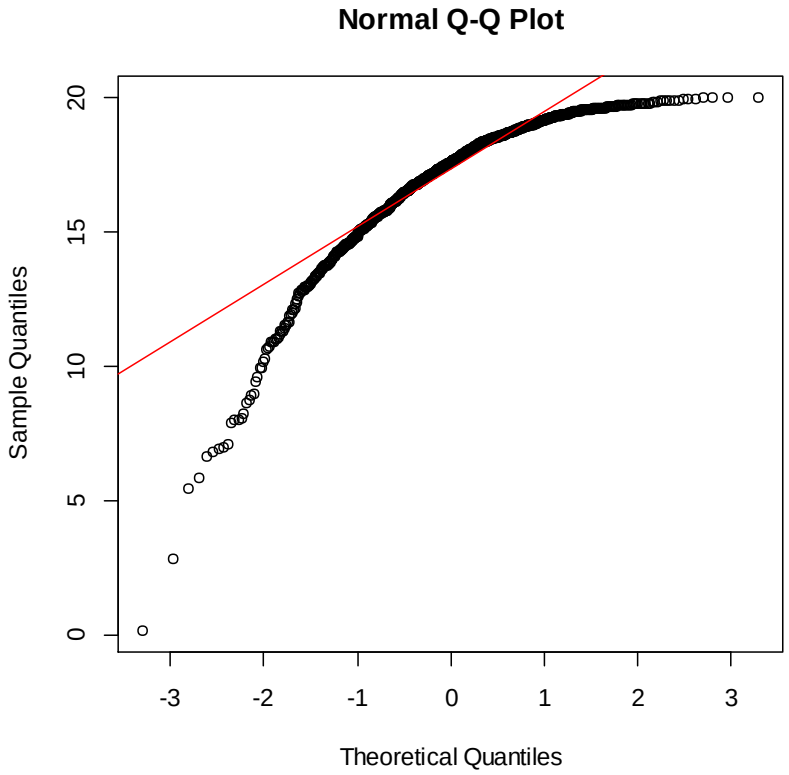




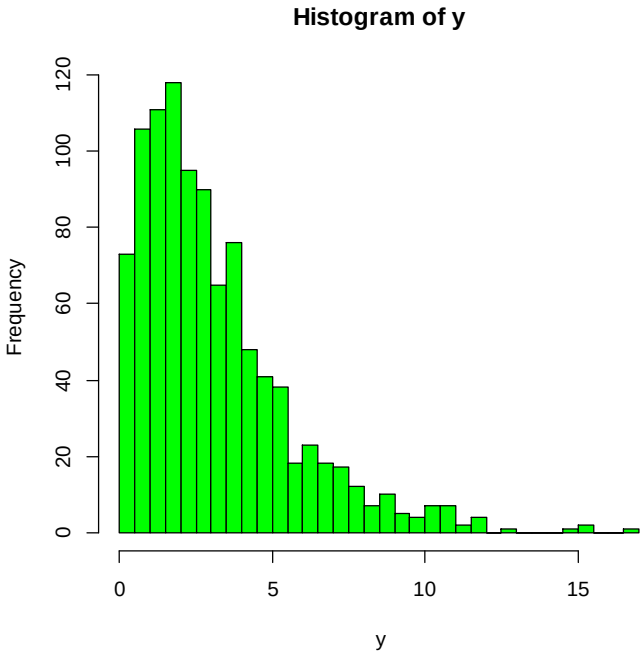

```
y <- rchisq(1000, 3)
hist(y, breaks=30, col="green")
qqnorm(y)
qqline(y, col = "red")
```

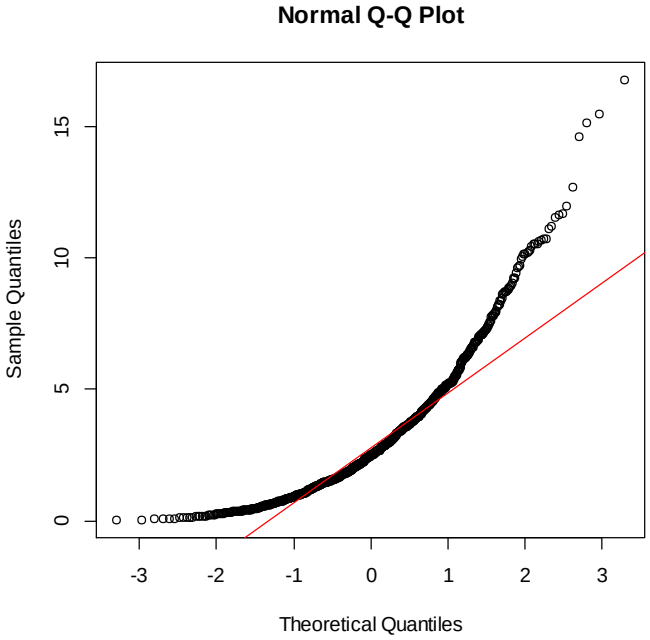
QQPlot Normal untuk Data yang Sebarannya Menjulur ke Kiri

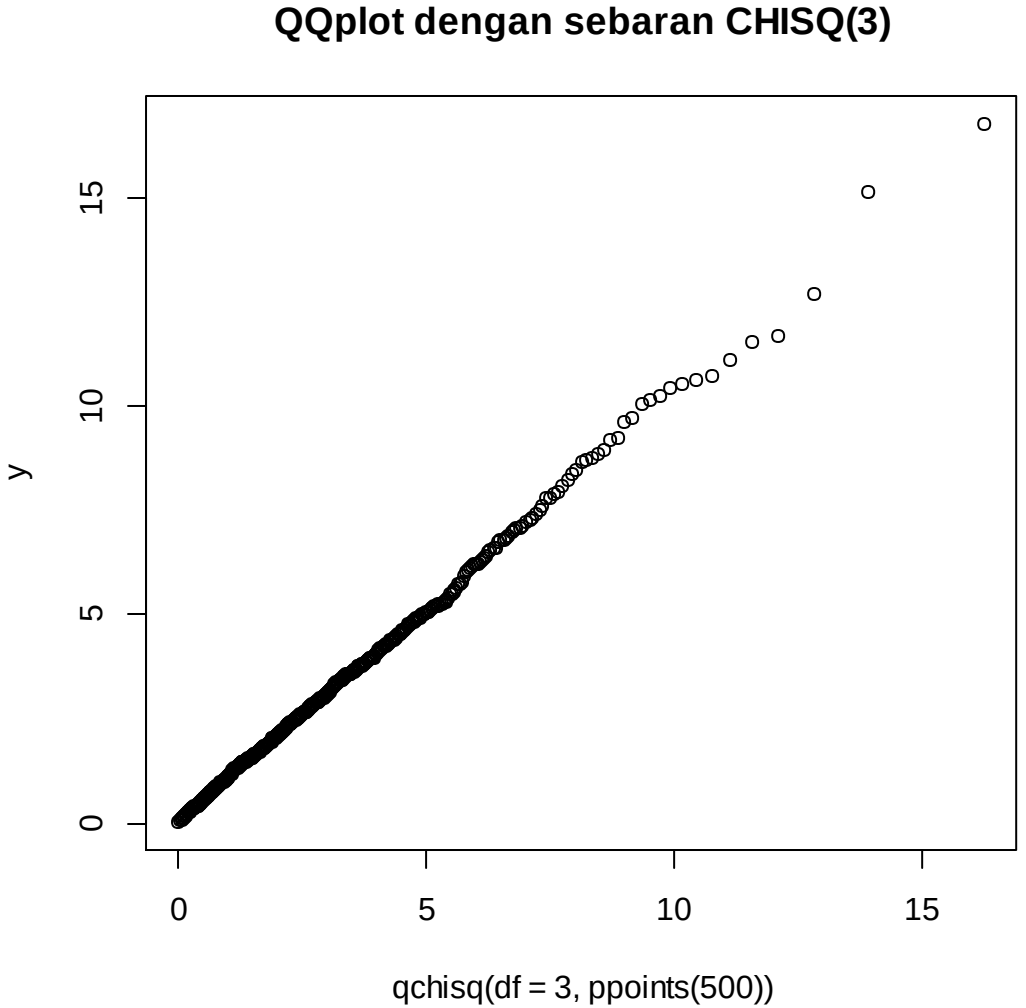





```
y <- 20 - rchisq(1000, 3)
hist(y, breaks=30, col="green")
qqnorm(y)
qqline(y, col = "red")
```








```
y <- rchisq(1000, 3)
hist(y, breaks=30, col="green")
qqplot(qchisq(df=3, ppoints(500)), y, main =
"QQplot dengan sebaran CHISQ(3)")
```

Goodness of Fit Test

- Uji formal untuk apakah suatu gugus data mengikuti sebaran hipotetik tertentu

- H0: data mengikuti sebaran hipotetik
- H1: data tidak mengikuti sebaran hipotetik

- Chi-Square test, didasarkan pada perbandingan frekuensi amatan antara data empirik dengan kondisi jika sebarannya mengikuti fungsi kepekatan/massa peluang tertentu
- Kolmogorov-Smirnov test, didasarkan pada perbandingan antara fungsi sebaran kumulatif empirik dan fungsi sebaran kumulatif hipotetik

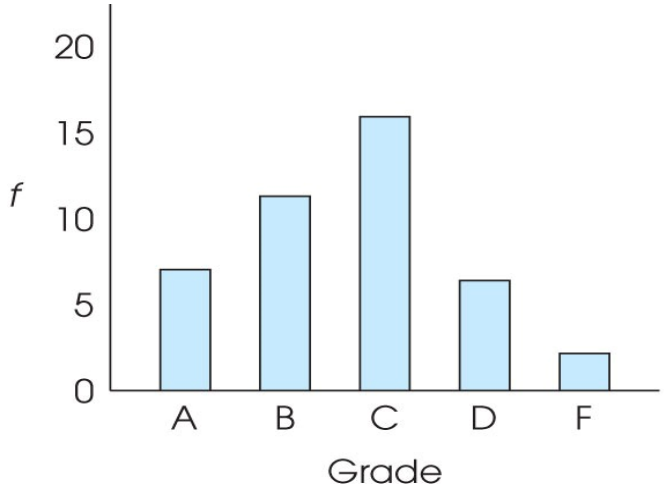
Chi-Square Test

- Membandingkan frekuensi amatan (observed, O) dengan frekuensi harapan (expected, E) berdasarkan sebaran tertentu
- Statistika Uji $\chi^2_{hitung} = \sum_{i=1}^{p} \frac{(O_i - E_i)^2}{E_i}$
- χ^2_{hitung} mengikuti sebaran χ^2 dengan derajat bebas (p – 1)
- Ingat! Ada beberapa batasan kevalidan uji ini…

  (pelajari di berbagai sumber bacaan terkait hal ini)

Chi-Square Test

- **Ilustrasi: Apakah data berikut mengikuti sebaran seragam?**



X	f
A	5
B	11
C	16
D	6
F	2

A	B	C	D	E
5	11	16	6	2

n = 40

H0 : P(A) = P(B) = P(C) = P(D) = P(F) = 0.2
H1: selainnya

Chi-Square Test

H0 : P(A) = P(B) = P(C) = P(D) = P(F) = 0.2

H1: selainnya

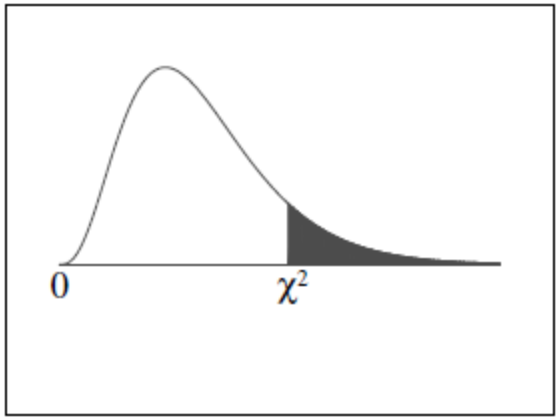
Nilai	Observed	Expected
A	5	8
B	11	8
C	16	8
D	6	8
F	2	8

$$\chi^2_{hitung} = \sum_{i=1}^{p} \frac{(O_i - E_i)^2}{E_i}$$

$$= 15.25$$

Terima H0 atau Tolak H0?

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

Chi-Square Test

- **Ilustrasi: Apakah data berikut mengikuti sebaran Normal?**

18.5	21.3	20.3	20.0	20.6	21.1	18.0	20.5	20.3	20.3	19.3	20.9	21.3
19.3	20.2	20.7	20.4	20.5	20.2	20.6	18.2	20.4	20.4	19.3	20.9	22.5
19.1	20.1	19.9	19.2	19.3	19.4	18.4	22.9	20.8	20.5	19.3	19.7	20.8
20.1	18.6	21.2	20.2	19.5	19.9	20.9	20.6	19.9	20.9	20.7	20.8	19.2

H0 : data menyebar normal
H1: data tidak menyebar normal

H0 : data menyebar Normal(?, ?)
H1: data tidak menyebar Normal(?, ?)

H0 : data menyebar Normal(mu=20.2, sigma=0.972)
H1: data tidak menyebar Normal(mu=20.2, sigma=0.972)

Chi-Square Test

H0 : data menyebar Normal(mu=20.2, sigma=0.972)
H1: data tidak menyebar Normal(mu=20.2, sigma=0.972)

Selang Nilai	Frekuensi	Peluang Normal sesuai H0	Ekspektasi	χ^2_{hitung}
18-19	5	0.11761	6.115714	0.203544
19-20	14	0.319512	16.61465	0.411466
20-21	27	0.370859	19.28467	3.086720
21-22	4	0.163252	8.48909	2.373862
22-23	2	0.027061	1.40718	0.249745

$$\chi^2_{hitung} = \sum_{i=1}^{p} \frac{(O_i - E_i)^2}{E_i}$$

$$= 6.33$$

Kolmogorov-Smirnov Test

- **Introduction**
- *A test for goodness of fit usually involves examining a random sample* from some unknown distribution in order to test the null hypothesis that the unknown distribution function is in fact a known, specified function.

- *A random sample X1,X2, . . . , Xn is drawn from some population* and is compared with $F^*(x)$ *in some way to see if it is reasonable to* say that $F^*(x)$ *is the true distribution function of the random sample.*

- *One logical way of comparing the random sample with* $F^*(x)$ *is by* means of the **empirical distribution function *S(x)***

Kolmogorov-Smirnov Test

- Definition
- Let X1,X2, . . . , Xn be a random sample. The empirical distribution function S(x) is a function of x, which equals the fraction of Xis that are less than or equal to x for each x, $-\infty<x<\infty$, i.e

$$S(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \leq x\}}$$

Kolmogorov-Smirnov Test

- The data consist of a random sample X1,X2, . . . , Xn of size n associated with some unknown distribution function,denoted by F(x)
- The sample is a random sample

- Let S(x) be the empirical distribution function based on the random sample X1,X2, . . . , Xn. Let $F^*(x)$ be a completely specified hypothesized distribution function

- Let the test statistic T be the greatest (denoted by ”sup” for supremum) vertical distance between S(x) and $F*(x)$. In symbols we say

$$T = \sup_x \mid F^*(x) - S(x) \mid$$

Critical values for the Kolmogorov-Smirnov Test for goodness of fit

For completely specified continuous distributions:

n \ $1-\alpha$	0.9	0.95	0.99
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.636	0.708	0.829
4	0.565	0.624	0.734
5	0.510	0.563	0.669
6	0.468	0.520	0.617
7	0.436	0.483	0.576
8	0.410	0.454	0.542
9	0.387	0.430	0.513
10	0.369	0.409	0.489
11	0.352	0.391	0.468
12	0.338	0.375	0.450
13	0.325	0.361	0.432
14	0.314	0.349	0.418
15	0.304	0.338	0.404
16	0.295	0.327	0.392
17	0.286	0.318	0.381
18	0.279	0.309	0.371
19	0.271	0.301	0.361
20	0.265	0.294	0.352

n \ $1-\alpha$	0.9	0.95	0.99
21	0.259	0.287	0.344
22	0.253	0.281	0.337
23	0.247	0.275	0.330
24	0.242	0.269	0.323
25	0.238	0.264	0.317
26	0.233	0.259	0.311
27	0.229	0.254	0.305
28	0.225	0.250	0.300
29	0.221	0.246	0.295
30	0.218	0.242	0.290
31	0.214	0.238	0.285
32	0.211	0.234	0.281
33	0.208	0.231	0.277
34	0.205	0.227	0.273
35	0.202	0.224	0.269
> 35	$\frac{1.224}{\sqrt{n}}$	$\frac{1.358}{\sqrt{n}}$	$\frac{1.628}{\sqrt{n}}$

Kolmogorov-Smirnov Test

- **Ilustrasi: Apakah data berikut mengikuti sebaran Normal**(mu=20.2, sigma=0.972)**?**

18.5	21.3	20.3	20.0	20.6	21.1	18.0	20.5	20.3	20.3	19.3	20.9	21.3
19.3	20.2	20.7	20.4	20.5	20.2	20.6	18.2	20.4	20.4	19.3	20.9	22.5
19.1	20.1	19.9	19.2	19.3	19.4	18.4	22.9	20.8	20.5	19.3	19.7	20.8
20.1	18.6	21.2	20.2	19.5	19.9	20.9	20.6	19.9	20.9	20.7	20.8	19.2

H0 : data menyebar Normal(mu=20.2, sigma=0.972)
H1: data tidak menyebar Normal(mu=20.2, sigma=0.972)

Kolmogorov-Smirnov Test

- **Ilustrasi: Apakah data berikut mengikuti sebaran Normal**(mu=20.2, sigma=0.972)**?**

18.5	21.3	20.3	20.0	20.6	21.1	18.0	20.5	20.3	20.3	19.3	20.9	21.3
19.3	20.2	20.7	20.4	20.5	20.2	20.6	18.2	20.4	20.4	19.3	20.9	22.5
19.1	20.1	19.9	19.2	19.3	19.4	18.4	22.9	20.8	20.5	19.3	19.7	20.8
20.1	18.6	21.2	20.2	19.5	19.9	20.9	20.6	19.9	20.9	20.7	20.8	19.2

H0 : data menyebar Normal(mu=20.2, sigma=0.972)
H1: data tidak menyebar Normal(mu=20.2, sigma=0.972)

i	x		S(x)	F(x)	abs(S-F)
1	18	1	0.019231	0.011806	0.007424
2	18.2	2	0.038462	0.019814	0.018648
3	18.4	3	0.057692	0.032024	0.025669
4	18.5	4	0.076923	0.040148	0.036775
5	18.6	5	0.096154	0.049873	0.046281
6	19.1	6	0.115385	0.128883	0.013498
7	19.2	8	0.153846	0.151785	0.002061
8	19.2	8	0.153846	0.151785	0.002061
9	19.3	13	0.25	0.177242	0.072758
10	19.3	13	0.25	0.177242	0.072758
11	19.3	13	0.25	0.177242	0.072758
12	19.3	13	0.25	0.177242	0.072758
13	19.3	13	0.25	0.177242	0.072758
14	19.4	14	0.269231	0.205241	0.06399
15	19.5	15	0.288462	0.235712	0.05275
16	19.7	16	0.307692	0.303485	0.004207
17	19.9	19	0.365385	0.378797	0.013412
18	19.9	19	0.365385	0.378797	0.013412

Dst....

T = 0.1203

T kritis = 0.1883

Terima H0

ks.test()

```
> ks.test(data, "pnorm", 20.2, 0.972)

        One-sample Kolmogorov-Smirnov test

data:  data
D = 0.12033, p-value = 0.4388
alternative hypothesis: two-sided
```

```
require(graphics)

y <- rt(200, df = 5)
qqnorm(y); qqline(y, col = 2)
qqplot(y, rt(300, df = 5))

qqnorm(precip, ylab = "Precipitation [in/yr] for 70 US cities")

## "QQ-Chisquare" : ------------------------
y <- rchisq(500, df = 3)
## Q-Q plot for Chi^2 data against true theoretical distribution:
qqplot(qchisq(ppoints(500), df = 3), y,
       main = expression("Q-Q plot for" ~~ {chi^2}[nu == 3]))
qqline(y, distribution = function(p) qchisq(p, df = 3),
       prob = c(0.1, 0.6), col = 2)
mtext("qqline(*, dist = qchisq(., df=3), prob = c(0.1, 0.6))")
```

```
require(graphics)

y <- rt(200, df = 5)
qqnorm(y); qqline(y, col = 2)
qqplot(y, rt(300, df = 5))

qqnorm(precip, ylab = "Precipitation [in/yr] for 70 US cities")

## "QQ-Chisquare" : ----------------------------
y <- rchisq(500, df = 3)
## Q-Q plot for Chi^2 data against true theoretical distribution:
qqplot(qchisq(ppoints(500), df = 3), y,
     main = expression("Q-Q plot for" ~~ {chi^2}[nu == 3]))
qqline(y, distribution = function(p) qchisq(p, df = 3),
     prob = c(0.1, 0.6), col = 2)
mtext("qqline(*, dist = qchisq(., df=3), prob = c(0.1, 0.6))")


library(moments)
library(nortest)
library(e1071)

set.seed(777)
x <- rnorm(250,10,1)

# skewness and kurtosis, they should be around (0,3)
skewness(x)
kurtosis(x)

# Shapiro-Wilks test
shapiro.test(x)

# Kolmogorov-Smirnov test
ks.test(x,"pnorm",mean(x),sqrt(var(x)))

# Anderson-Darling test
ad.test(x)

# qq-plot: you should observe a good fit of the straight line
qqnorm(x)
qqline(x)

# p-plot: you should observe a good fit of the straight line
```