Mathematical Statistics II (MAT323)

Ch6. Point Estimation

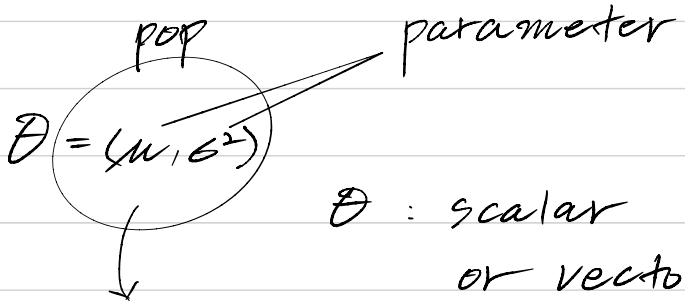
Jungsoon Choi

jungsoonchoi@hanyang.ac.kr

Table of Contents

Ch6.3 Order Statistics

parameter : (unknown) interest in pop.



θ : scalar
or vector.

$X_1, \cdots, X_n$

$\overline{X_n}$ $x_n = 3.5$ point estimation
C. I estimation.
(3.4, 3.6)

⇓

Testing.

Order statistics

Definition

Let $X_1, \ldots, X_n$ denote a random sample from a continuous distribution of $f(x)$. A random sample $Y_1, \ldots, Y_n$ is called the order statistics of $X_1, \ldots, X_n$ if

$$Y_1 < Y_2 < \cdots < Y_n$$

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

Y_i or $X_{(i)}$ is called the ith order statistic of the ramdom sample $X_1, \ldots, X_n$.

$X_1, \cdots, X_n$ random sample.

2 자체도 statistic

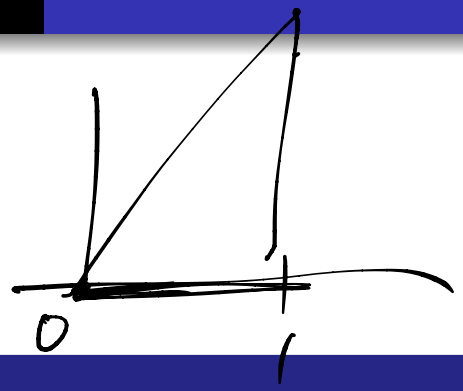
$$X_{(1)} = Y_1 < X_{(2)} = Y_2 < \cdots < X_{(n)} = Y_n$$

$X_{(1)} = \min_{i=1,\cdots,n} \{X_i\}$

$X_{(n)} = \max_{i=1,\cdots,n} \{X_i\}$

$X_{(1)} \cdots X_{(n)}$: order statistic

want to find dist^n of order statistic



Example 6.3-2

$Y_1 < Y_2 < \ldots < Y_5$ are order statistics of $X_1, \ldots, X_5$ with a pdf $f_x(x) = 2x,\ 0 < x < 1$.

- Find the probability of $P(Y_4 \leq 0.5)$.
- Find the cdf of Y_4, $F_{Y_4}(y) = P(Y_4 \leq y)$, $0 < y < 1$.
- Find the pdf of Y_4, $f_{Y_4}(y)$.

$f_X(x) = 2x \quad 0 < x < 1.$

1) $P(Y_4 \le 0.5)$.

2) $F_{Y_4}(y)$. $\quad 0 < y < 1$.

3) $f_{Y_4}(y)$.

$X_1, \cdots, X_5$.

$Y_1 < \cdots < Y_5$. $\quad Y_4 \le 0.5 \Leftrightarrow P(X \le 0.5)$ at least 4.

1) $P_5 = P(X \le 0.5) = \int_0^{0.5} 2x\,dx = [x^2]_0^{0.5} = \frac{1}{4}$.

$Z \sim Bin(5, \frac{1}{4})$.

$P(Y_4 \le 0.5) = P(Z=4) + P(Z=5)$.

$= \binom{5}{4}\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right) + \left(\frac{1}{4}\right)^5$

2). $Z \sim Bin(5, y^2)$.

$F_{Y_4}(y) = \binom{5}{4}(y^2)^4(1-y^2) + (y^2)^5$

3) $f_{Y_4}(y) = 20(y^2)^3(2y)(1-y^2) + 5(y^2)^4(2y)$.

$+ 5\cdot(y^2)^4(-2y)$

$= \binom{5}{3}(F_X(y))^3 \cdot f_X(y) \cdot \binom{2}{1}(1-F_X(y))$

$= 20(y^2)^3 \cdot 2y \cdot (1-y^2)$.

$X_i \sim f_X(x) = 2x, \quad 0 < x < 1 \qquad Y_1 < Y_2 < Y_3 < Y_4 < Y_5 < \frac{1}{2}$

$\{Y_4 \le \frac{1}{2}\}$ happens

$\Leftrightarrow$ {at least 4 r.vs among $X_1, \cdots, X_5 \le \frac{1}{2}$}

Assume $\{X_i \le \frac{1}{2}\}$ is a "success event"

Let Z be the number of success events $\sim Bin(5, p_s)$ (prob of success)

$$p_s = P(X_i \le \tfrac{1}{2}) = F_{X_i}(\tfrac{1}{2}) = \int_0^{\frac{1}{2}} 2x\,dx = [x^2]_0^{\frac{1}{2}} = \frac{1}{4}$$

$\therefore Z \sim Bin(5, \frac{1}{4})$.

1) $P(Y_4 \le \frac{1}{2}) = P(\text{at least 4 r.vs among } X_1, \cdots, X_5 \le \frac{1}{2})$

$= P(4 \text{ r.vs among } \forall X_i \le \frac{1}{2}) + P(5 \text{ r.vs among } \forall X_i \le \frac{1}{2})$

$= P(Z=4) + P(Z=5)$

$= \binom{5}{4}\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^1 + \left(\frac{1}{4}\right)^5$

2) consider $\{X_i \le y\}$ as a "success event"

$Z \sim Bin(5, P(X_i \le y) = F_{X_i}(y) = \int_0^y 2x\,dx = y^2)$

$P(Y_4 \le y) = P(Z=4) + P(Z=5)$

$= F_{Y_4}(y)$

$= \binom{5}{4}(y^2)^4(1-y^2) + (y^2)^5, \quad 0 < y < 1$

2) $f_{Y_4}(y) = \binom{5}{4} 4(y^2)^3(2y)(1-y^2) + \binom{5}{4}(y^2)^4(-2y) + 5(y^2)^4(2y)$

$= 20(y^2)^3(2y)(1-y^2)$ $\quad y^2 = F_X(y)$

Since $F_X(y) = y^2$

$$f_{Y_4}(y) = \frac{5!}{3!\,2!} \times \frac{2!}{1!} \times [F_X(y)]^3[1-F_X(y)]^1 \times f_X(y)$$

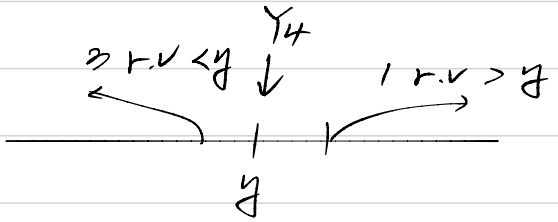
($\frac{5!}{3!2!}$: $_5C_3$, $\frac{2!}{1!}$: $_2C_1$)

$0 < y < 1$

y보다 작을 확률 3개 $\quad$ y의 확률 1개 $\quad$ y보다 클 확률 1개

(확률이 0이 아닐 때) ※개별확률은 // pdf.

$P(Y_4 = y)$



Next job : generalize the cdf & pdf of order statistic

cdf of order statistics

Suppose that $Y_1 < Y_2 < \ldots < Y_n$ are order statistics of continuous r.v $X_1, \ldots, X_n$ with a pdf $f_X(x)$ and a cdf $F_X(x)$, $a < x < b$. Since $P(X_i \leq y) = F_X(y)$,

$$F_{Y_r}(y) = P(Y_r \leq y) = \sum_{k=r}^{n} \binom{n}{k} [F_X(y)]^k [1 - F_X(y)]^{n-k}$$

$$= \sum_{k=r}^{n-1} \binom{n}{k} [F_X(y)]^k [1 - F_X(y)]^{n-k} + [F_X(y)]^n$$

pdf of order statistics

$$f_{Y_r}(y) = \sum_{k=r}^{n-1} \binom{n}{k} (k) \left[F_X(y)\right]^{k-1} f_X(y) \left[1 - F_X(y)\right]^{n-k}$$
$$+ \sum_{k=r}^{n-1} \binom{n}{k} \left[F_X(y)\right]^{k} (n-k) \left[1 - F_X(y)\right]^{n-k-1} \left[-f_X(y)\right]$$
$$+ n \left[F_X(y)\right]^{n-1} f_X(y)$$

With some calculations,

$$f_{Y_r}(y) = \frac{n!}{(r-1)!(n-r)!} \left[F_X(y)\right]^{r-1} \left[1 - F_X(y)\right]^{n-r} f_X(y)$$

$$a < y < b$$

$X_i \sim f_X(x)$, $F_X(x)$, $i = 1, \cdots, n$

$F_{Y_r}(y) = P(Y_r \le y) = P(\text{at least } "r" \text{ r.vs among } X_i \le y)$.

$= P(Z = r) + P(Z = r+1) + \cdots + P(Z = n)$.

where $Z \sim Bin(n, P(X_i \le y) = F_X(y))$

$= \sum_{k=r}^{n-1} \binom{n}{k} [F_X(y)]^k [1 - F_X(y)]^{n-k} + [F_X(y)]^n$

$\sum_{k=r}^{n} \binom{n}{k} [F_X(y)]^k [1 - F_X(y)]^{n-k}$ ← pdf 에서도.

$$f_{Y_r}(y) = \frac{\partial F_{Y_r}(y)}{\partial y} = \sum_{k=r}^{n-1} \binom{n}{k} k [F_X(y)]^{k-1} f_X(y) [1 - F_X(y)]^{n-k}$$

$$+ \sum_{k=r}^{n-1} \binom{n}{k} (n-k) [F_X(y)]^k [1 - F_X(y)]^{n-(k+1)} (-f_X(y)) + n [F_X(y)]^{n-1} f_X(y).$$

Since $\binom{n}{k} k = \frac{n!}{(k-1)!(n-k)!}$ & $\binom{n}{k}(n-k) = \frac{n!}{k!(n-(k+1))!}$

$$f_{Y_r}(y) = \frac{n!}{(r-1)!(n-r)!} [F_X(y)]^{r-1} [1 - F_X(y)]^{n-r} f_X(y)$$

$$+ \sum_{k=r+1}^{n-1} \frac{n!}{(k-1)!(n-k)!} [F_X(y)]^{k-1} [1 - F_X(y)]^{n-k} f_X(y) \quad (k' = k-1)$$

$$- \sum_{k=r}^{n-2} \frac{n!}{k!(n-(k+1))!} [F_X(y)]^k [1 - F_X(y)]^{n-(k+1)} f_X(y)$$

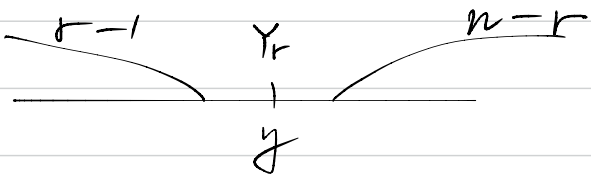
$$- \frac{n!}{(n-1)!} [F_X(y)]^{n-1} f_X(y) + n [F_X(y)]^{n-1} f_X(y) = 0.$$

$$= \frac{n!}{(r-1)!(n-r)!} [F_X(y)]^{r-1} [1 - F_X(y)]^{n-r} f_X(y)$$

$$+ \sum_{k'=r}^{n-2} \frac{n!}{k'!(n-(k'+1))!} [F_X(y)]^{k'} [1 - F_X(y)]^{n-(k'+1)} f_X(y)$$

$$- \sum_{k=r}^{n-2} \frac{n!}{k!(n-(k+1))!} [F_X(y)]^k [1 - F_X(y)]^{n-(k+1)} f_X(y).$$

$$= \frac{n!}{(r-1)!(n-r)!} [F_X(y)]^{r-1} [1 - F_X(y)]^{n-r} f_X(y). \quad a < y < b.$$



$$n \begin{cases} r-1 & \text{r.v} < y \\ n-r & \text{r.v} > y \\ 1 & \text{r.v} = y \end{cases}$$

감이 왔지만 개념적으로 이해.

$$f_{Y_r}(y) \neq P(Y_r = y).$$

$P(X<y)$ $P(X>y)$ $P(X=y)$ (≠)

$$= \binom{n}{r-1} [F_X(y)]^{r-1} \binom{n-r+1}{n-r} [1-F_X(y)]^{n-r} \cdot f_X(y).$$

$$= \frac{n!}{(r-1)!(n-r+1)!} \frac{(n-r+1)!}{(n-r)!} [F_X(y)]^{r-1} [1-F_X(y)]^{n-r} f_X(y).$$

$(n-r+1)$ 이렇게 확인하는 것.

$$f_{Y_r}(y) = \frac{n!}{(r-1)!(n-r)!} [F_X(y)]^{r-1} [1-F_X(y)]^{n-r} \cdot f_X(y).$$

pdfs of Y_1 and Y_n

$$f_{Y_1}(y) = n\left[1 - F_X(y)\right]^{n-1} f_X(y), \quad a < y < b$$

$$f_{Y_n}(y) = n\left[F_X(y)\right]^{n-1} f_X(y), \quad a < y < b$$

$X_1, \cdots, X_n$

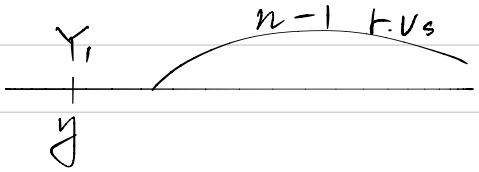
$Y_1 < Y_2 < \cdots < Y_n$

Y_1 = $\min\{X_i\}$; Y_n = $\max\{X_i\}$

$$f_{Y_1}(y) = P(Y_1 = y).$$

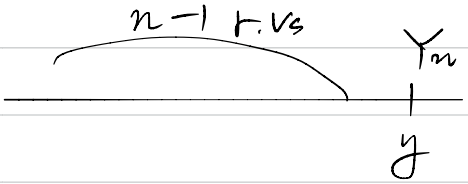
$$= \binom{n}{n-1}[1 - F_X(y)]^{n-1} f_X(y).$$

($1 - F_X(y) = P(X > y)$)



$$f_{Y_n}(y) = P(Y_n = y).$$

$$= \binom{n}{n-1}[F_X(y)]^{n-1} f_X(y).$$



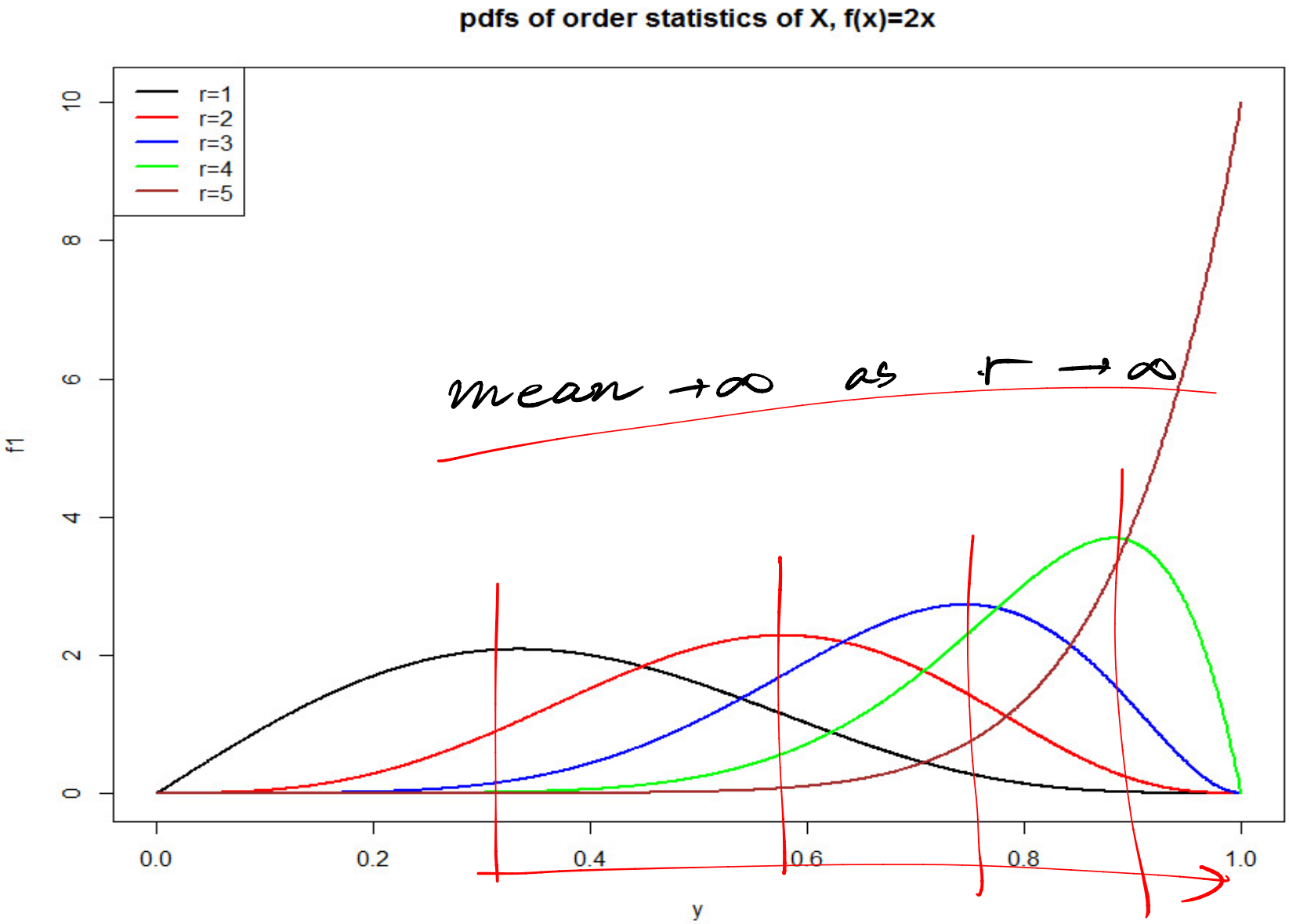
Example 6.3-3

In Example 6.3-2, find the pdfs of each $Y_1, \ldots, Y_5$.

$$f_{Y_1} = 5 \cdot [1 - F_X(y)]^4 \cdot f_X(y) = 5 \cdot (1-y^2)^4 \, 2y$$

$$f_{Y_3} = \frac{5!}{2!\,3!} [F_X(y)]^2 \cdot [1 - F_X(y)]^2 f_X(y) = 10 \cdot (y^2)^2 (1-y^2)^2 \, 2y$$

$$f_{Y_5} = 5 \cdot [F_X(y)]^4 \cdot f_X(y) = 5 \cdot (y^2)^4 \cdot 2y$$

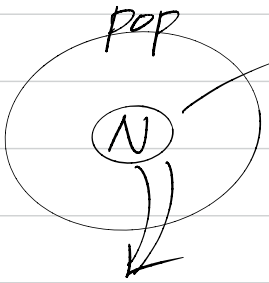

pdfs of order statistics of X, f(x)=2x
r=1
r=2
r=3
r=4
r=5
f1
y
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
mean → ∞ as r → ∞

[Review] π_p, $(100p)$th percentile

The $(100p)$th percentile is a number π_p such that the area under $f_X(x)$ to the left of π_p is p.

$$p = \int_{-\infty}^{\pi_p} f_X(x)dx = F_X(\pi_p).$$

- median=50th percentile=the second quartile, $m = \pi_{0.5}$
- the first quartile=25th percentile, $q_1 = \pi_{0.25}$
- the thrid quartile=75th percentile, $q_3 = \pi_{0.75}$



$Y_1 \quad Y_2 \quad Y_n$

$X''_{(1)} < X''_{(2)} < \cdots < X''_{(n)}$

$X_1, \cdots, X_n$: random sample.

$x_1, \cdots, x_n$: dataset.

$x_{(1)} < x_{(2)} < \cdots < x_{(n)}$

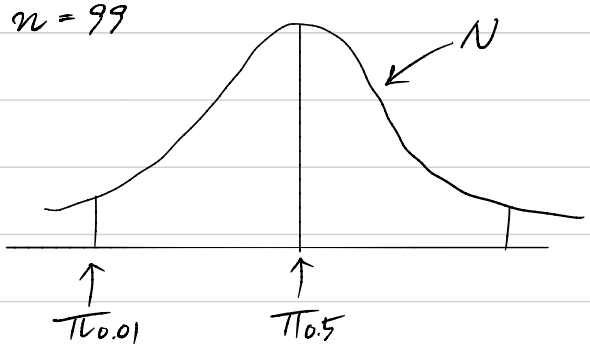
$\pi_p = (100p)^{th}$ percentile

$\pi_{0.01} = 1^{st}$ percentile.

$\vdots$

$\pi_1 = 100^{th}$.

$n = 99$



Sample percentiles (using the data)

from dataset.

Sample Percentiles ($0 < p < 1$)

The $100p^{th}$ sample percentile ($= Y_r$) is defined as

$$\begin{cases} (n+1)p^{th} \text{ order statistic} & \text{if } r = (n+1)p \text{ is integer} \\ \text{weighted average of } Y_r \text{ and } Y_{r+1} & \text{if } (n+1)p \text{ is not integer} \\ & r = [(n+1)p]\text{:the greatest integer} \end{cases}$$

we say

Notation

Q1=25th percentile=First quartile

Q2=50th percentile=Second quartile=Median

Q3=75th percentile=Third quartile

$n = 99 \qquad p = 0.01$

$(n+1)p = 100 \times 0.01 = 1$

$(100p)^{th}$ sample percentile

$= 1^{st}$ sample percentile. $= 1^{st}$ order statistic.

2^{nd} " $= 2^{nd}$ " $= x_{(2)}$.

$\vdots \qquad \vdots$

99^{th} " $= x_{(99)}$.

e.x). $n = 10. \quad x_1, \cdots, x_{10}$

$\Rightarrow x_{(1)} < \cdots < x_{(10)}$

$p = 0.5$

$50^{th} = (100p)^{th}$ sample percentile.

$r = 11 \times 0.5 = 5.5 = 5.$

$\Rightarrow \dfrac{Y_5 + Y_6}{2}.$

$(12+1)0.25 = \frac{13}{4} = 3.25$

$\frac{3}{4}Y_3 + \frac{1}{4}Y_4$

Example 6.3-4

Let X be the weight of soap in a "1000-gram" bottle. A random sample of $n = 12$ observations of X yielded the following weights, which have been ordered:

1013	1019	1021	1024	1026	1028
1033	1035	1039	1040	1043	1047

Find the first, second, and third quartiles.

25th percentile $= (12+1)0.25 = \frac{13}{4} = 3.25$

$Y_3 + (Y_4 - Y_3)0.25$

$= 0.75Y_3 + 0.25Y_4$

Q-Q plot:

To check if the assumed population distribution is reasonable based on the observed samples

Q-Q plot

Q-Q plot

A Q-Q plot is a probability plot which is a graphical method to compare two probability distributions by plotting their quantiles against each other. If the points in the Q-Q plot approximately lie on the straight line, then it is said that two distributions are similar.

sample percentile vs theoretical percentile.
↑ compare.
dataset.

To check if the observed samples follow the normal distribution, a Q-Q plot can be used.

Suppose that $x_i \sim \mathrm{N}(\mu, \sigma^2),\ i = 1, \ldots, n$.

x-axis: theoretical quantile from the normal distribution

y-axis: data quantile

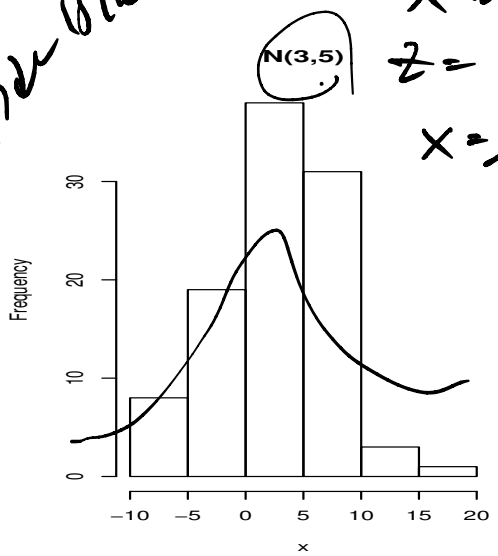
$$\text{plot of } \left(z_i = \Phi^{-1}\left(\frac{i}{n+1}\right), x_{(i)} \right), \quad i = 1, \ldots, n$$

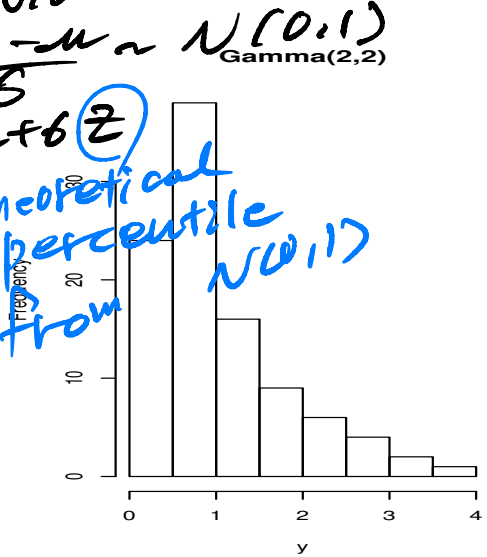
where $x_{(i)}$ is the ith ordered sample.

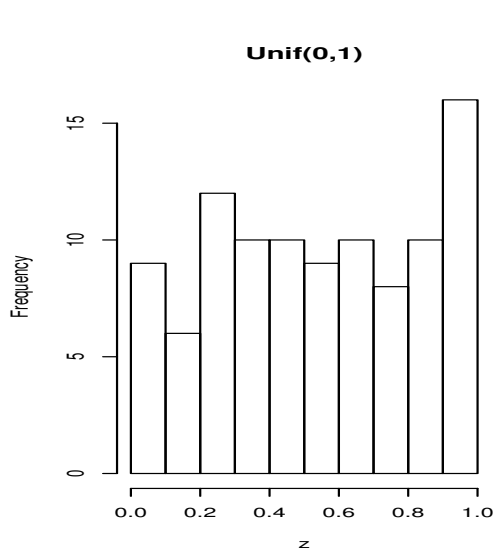
Instead of $(i/(n+1))$, $(i - 0.5)/n$ or $(i - 3/8)/(n + 1/4)$ can be used.

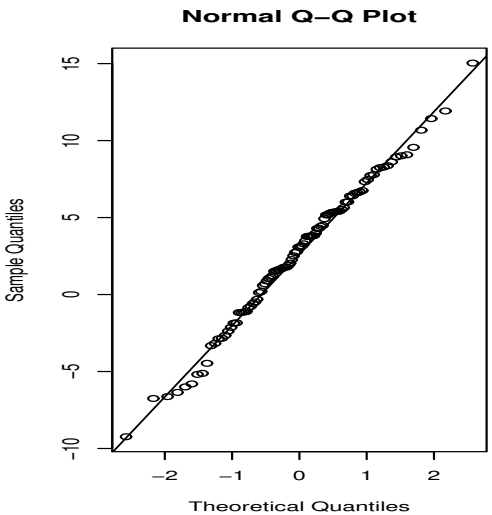
If $z_i \sim \mathrm{N}(0, 1)$ and $x_i \sim \mathrm{N}(\mu, \sigma^2)$, $x_i = \mu + \sigma z_i$.

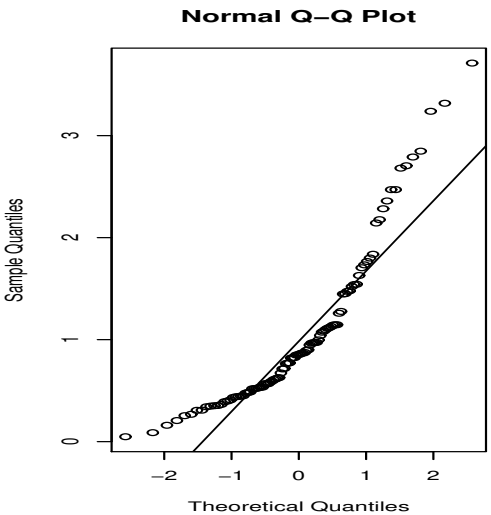
To check if the observed samples follow the normal distribution

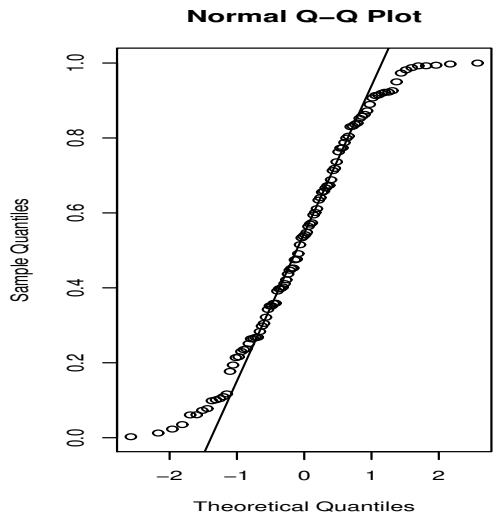


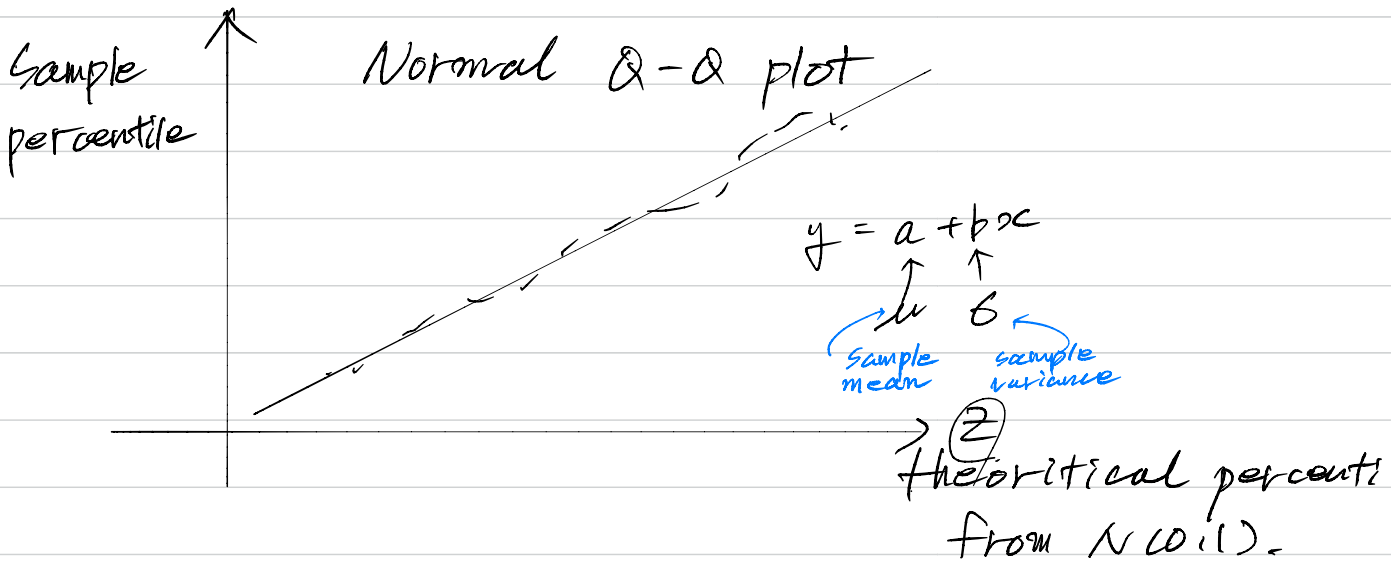











Sample percentile
Normal Q-Q plot
$y = a + bx$
μ
σ
sample mean
sample variance
z
theoritical percentile from N(0,1).

Example 6.3-5

Here are the diameters (in mm) of 30 randomly selected grains of soil:

1.24	1.36	1.28	1.31	1.35	1.20	1.39	1.35	1.41	1.31
1.28	1.26	1.37	1.49	1.32	1.40	1.33	1.28	1.25	1.39
1.38	1.34	1.40	1.27	1.33	1.36	1.43	1.33	1.29	1.34

For these data, $\overline{x} = 1.33$ and $s^2 = 0.0040$. May we assume that these are observations of a random variable X that is $N(1.33, 0.0040)$?

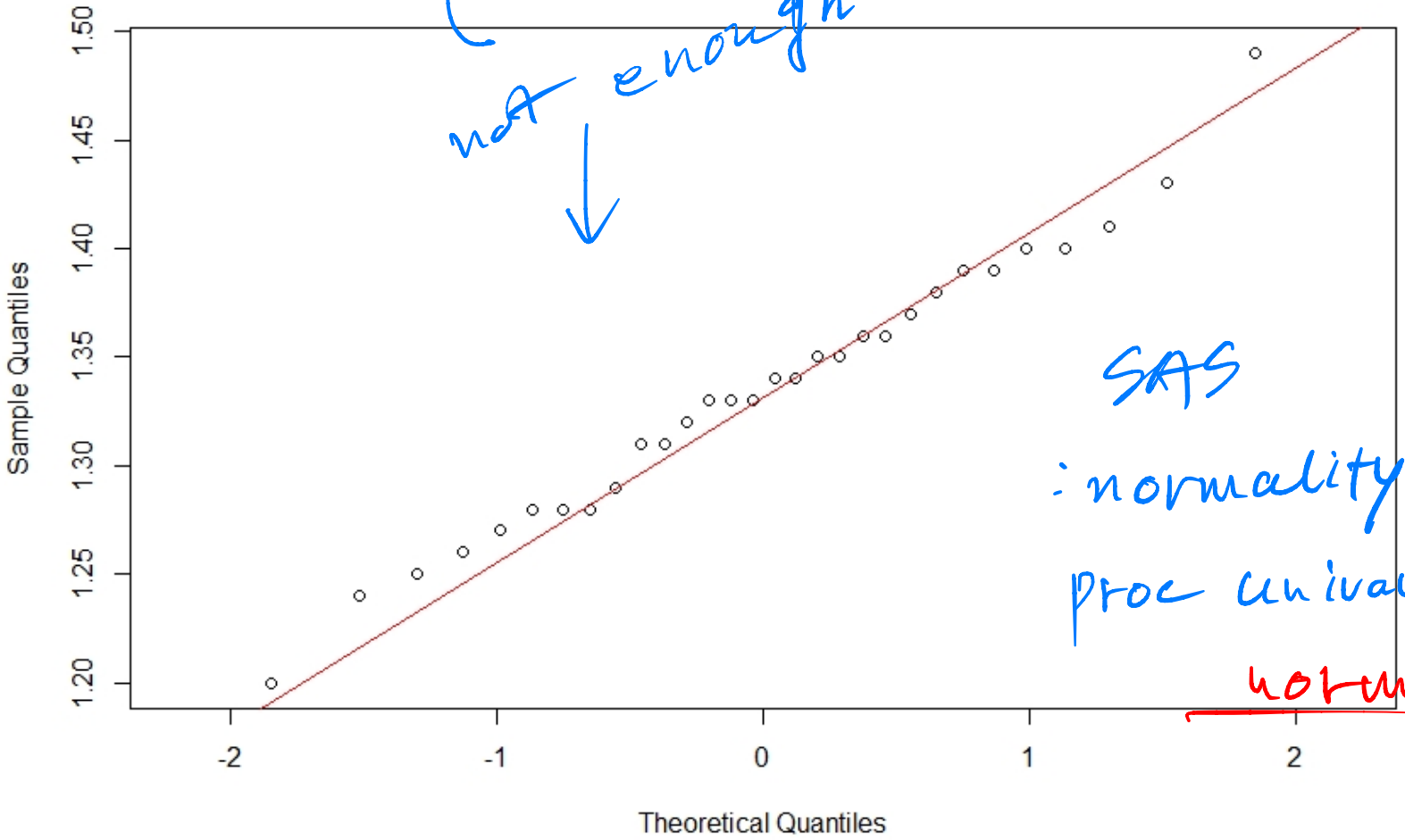
To draw a Q-Q plot, we obtain the following table:

k	y(data sorted)	$p = k/31$	z_{1-p}
1	1.20	0.0323	-1.85
2	1.24	0.0645	-1.52
3	1.25	0.0968	-1.30
	...		
29	1.43	0.9355	1.52
30	1.49	0.9677	1.85

maximum of sample but not maximum of pop

use test tool.

not enough



SAS

: normality test.

proc univariate.

normal