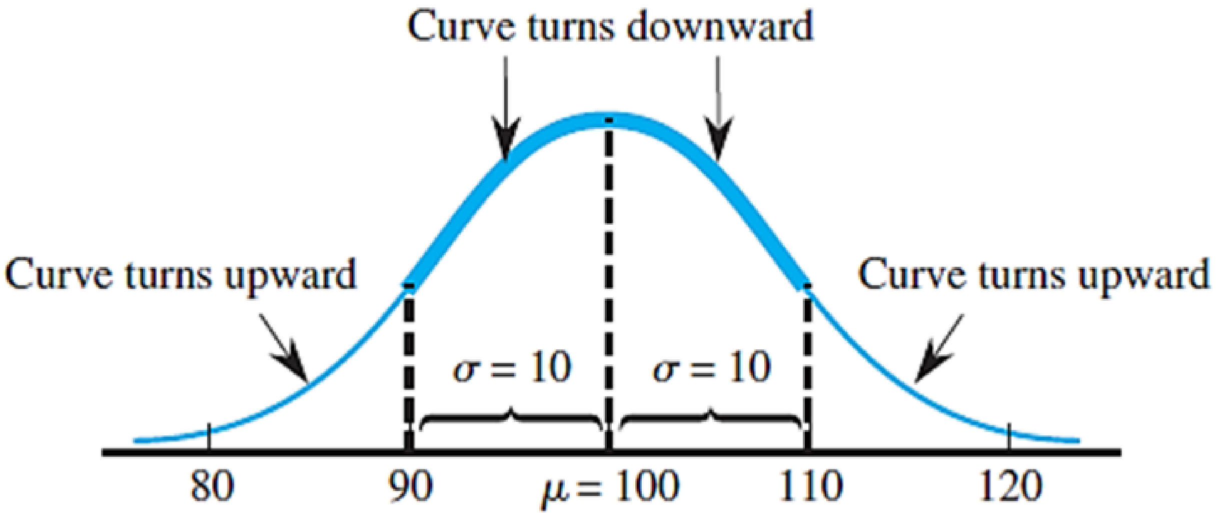
Engineering Statistics

Distribution



Dr. Vvn Weian Chao (趙韋安)
https://ce.nctu.edu.tw/member/teachers/23
Department of Civil Engineering, National Yang Ming Chiao Tung University, Taiwan

Distribution

Choice
of
an appropriate
distribution

Distribution

Decide which family is reasonable

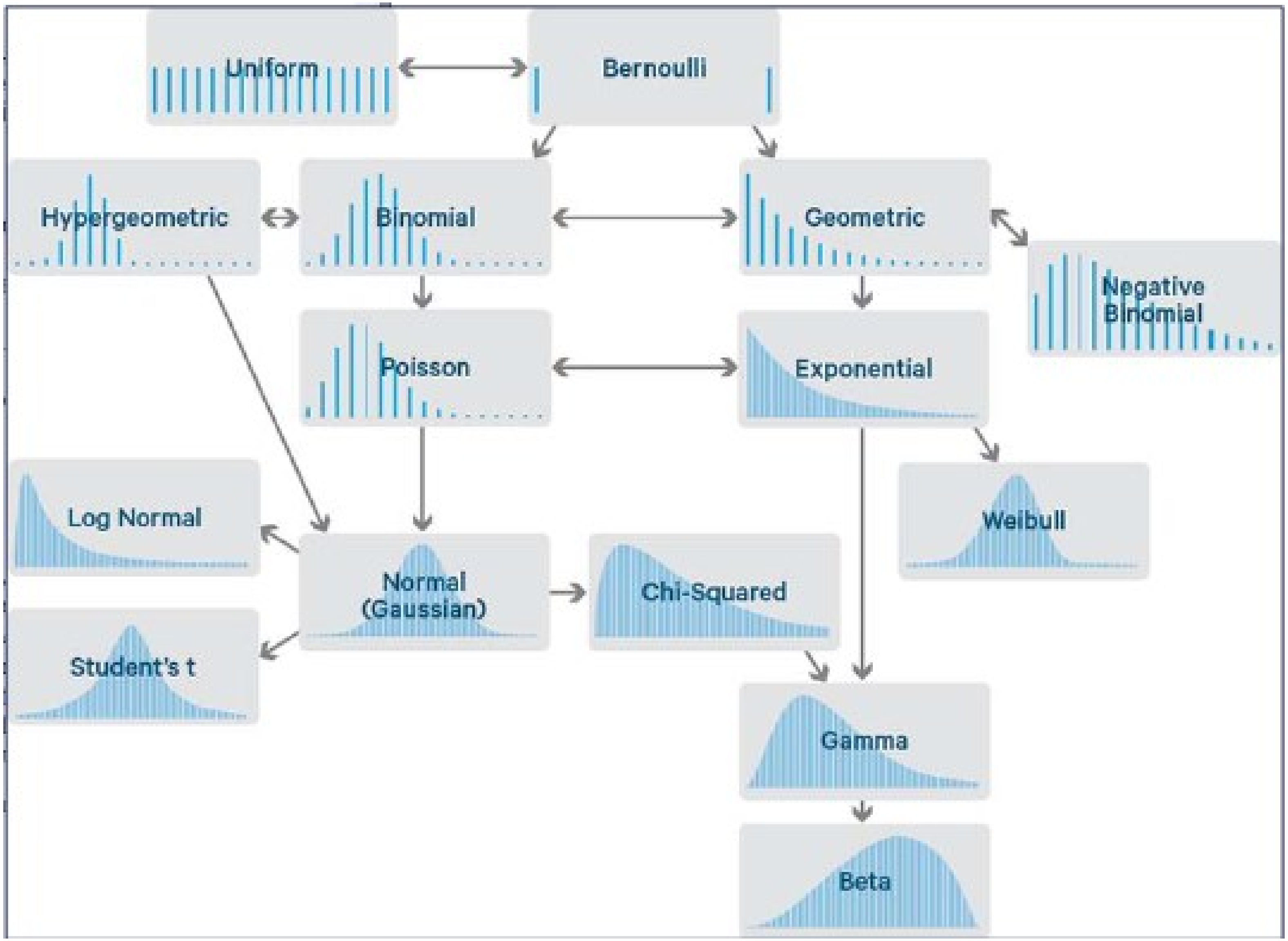
Distribution (統計模型)

若能決定所觀察現象的機率分布之參數，就可以了解所觀察現象的本質

Distribution

Continuous	Discrete

Distribution



Distribution in R

normal
lognormal
exponential
weibull
poisson
gamma
chi-squared
beta

Distribution	R name	additional arguments
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df, ncp
exponential	exp	rate
F	f	df1, df1, ncp
gamma	gamma	shape, scale
geometric	geom	prob
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logistic	logis	location, scale
negative binomial	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
Student's	t	t df, ncp
uniform	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

Normal distribution in R

dnorm() –常態機率密度函數
pnorm() –常態累積機率函數
qnorm() –常態機率函數之分位數
rnorm() –常態隨機亂數

Distribution, d
Probability, p
Quantile, q
Random, r

How to change legend title for ggplot

scale_color_discrete (name =, labels =).

TRY
it
in
R

R: Distribution

R_distribution_a.R

R: Distribution Quantile Plot

An effective way to check a distributional assumption (also called a probability plot).

R: Distribution Quantile Plot

If the plot is based on the correct distribution, the points in the plot will fall close to a **straight line**.

Quantile Plot

判斷資料是否服從某機率分佈(如常態分佈)的描述性方法之一，就是畫出資料次數分佈的直方圖或莖葉圖，若資料近似服從分佈函數，則圖形的形狀與函數應該相似

Quantile Plot

但是，實際上更常用的是繪製樣本資料的分位數-分位數圖(Q-Q plot)，Q-Q plot是根據觀測值的實際分位數與理論分佈的分位數的符合程度繪製

Quantile Plots: Sample Quantiles (樣本分位數)

Definition:

- Let $x_{(1)}$ denote the smallest sample observation, $x_{(2)}$ the second smallest sample observation, . . . , $x_{(n)}$ the largest sample observation.
- Take $x_{(1)}$ to be the (.5/n)th sample quantile, $x_{(2)}$ to be the (1.5/n)th sample quantile, . . . , $x_{(n)}$ to be the [(i – .5)/n]th sample quantile.
- For i = 1, . . . , n,

$x_{(i)}$ is the **[(i – .5)/n]th sample quantile.**

A Normal Quantile Plot

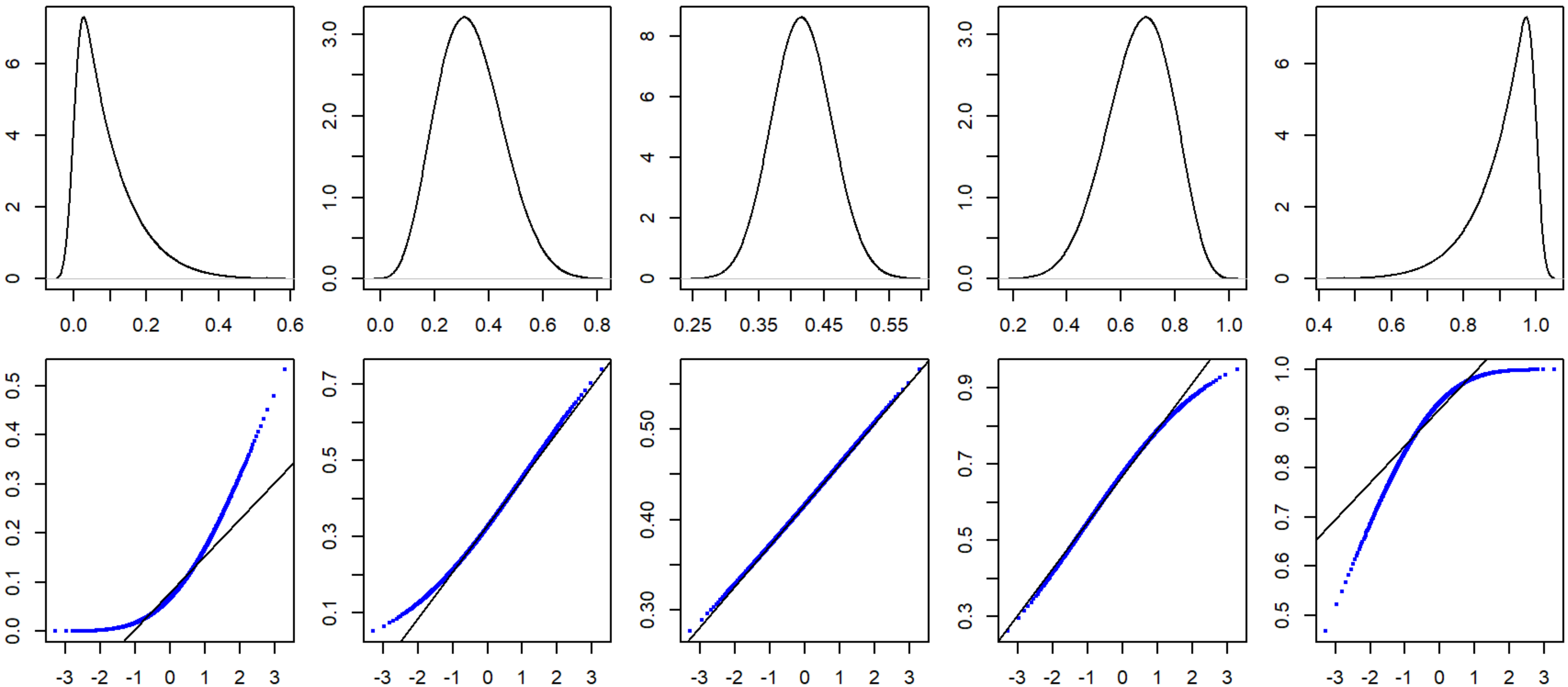
Definition:

- A **normal quantile plot** is a plot of the (z quantile, observation) pairs.
- The linear relation between normal (μ, σ) quantiles and z quantiles implies that if the sample has come from a normal distribution with particular values of μ and σ, the points in the plot should **fall close to a straight line with slope σ and vertical intercept μ.**
- Thus a plot for which the points fall close to *some* straight line suggests that the assumption of a normal population or process distribution is plausible.

R: Distribution Sample Quantile quantile()

quantile produces sample quantiles corresponding to the given probability.

R: Distribution QQPlot



Source :https://mgimond.github.io/ES218/Week06a.html

R: Distribution QQPlot in ggplot2

stat_qq(distribution). : produce quantile-quantile data points

Source :https://mgimond.github.io/ES218/Week06a.html

R: Distribution QQPlot in ggplot2

stat_qq_line(
line.p =
).
: produce line

Source :https://mgimond.github.io/ES218/Week06a.html

TRY
it
in
R

R: Distribution QQPlot in ggplot2

R_distribution_b.R

Continuous Distributions

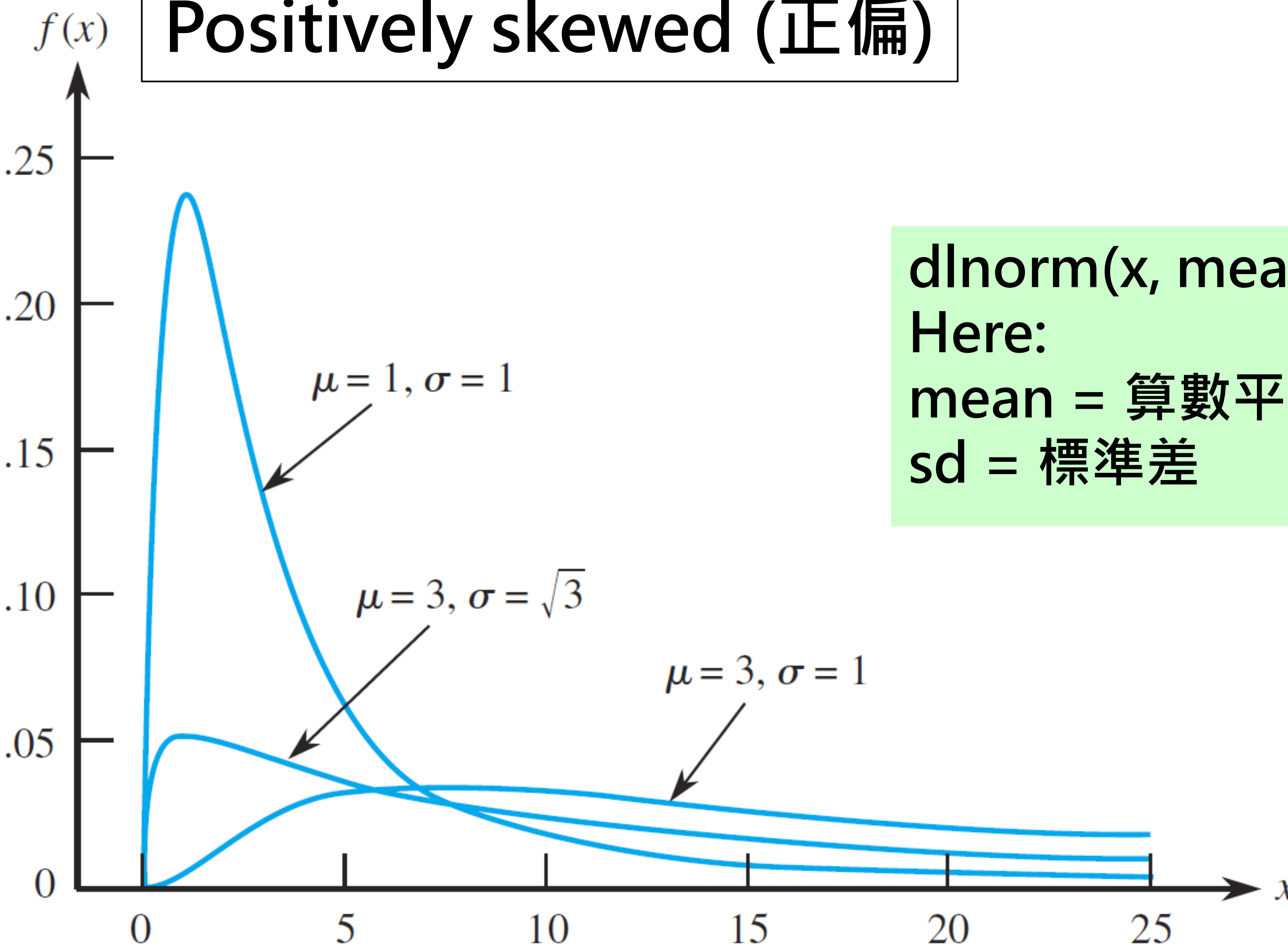
Definition:

- A nonnegative variable x is said to have a **lognormal distribution** if ln(x) has a normal distribution with parameters μ and σ.
- The density function of x:

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma x} e^{-[\ln(x)-\mu]^2/(2\sigma^2)} & x > 0 \\ 0 & \text{for } x \le 0 \end{cases}$$

Continuous Distributions

Positively skewed (正偏)



dlnorm(x, mean =, sd =)
Here:
mean = 算數平均數
sd = 標準差

Continuous Distributions

油管最大可埋藏的深度可安全運作的機率

Example 1.19

According to the article "Predictive Model for Pitting Corrosion in Buried Oil and Gas Pipelines" (*Corrosion*, 2009: 332–342), the lognormal distribution has been reported as the best option for describing the distribution of maximum pit depth data from cast iron pipes in soil. [redacted] is appropriate for maximum pit depth (mm) of buried pipelines.

Since $x < 2$ is equivalent to $\ln(x) < \ln(2) = .693$,

$$\begin{aligned}\text{proportion of pipelines with } x < 2 &= \text{proportion of pipelines with } \ln(x) < .693 \\ &= \text{area under normal } (.353, .754) \text{ curve to the left of } .693 \\ &= \text{area under } z \text{ curve to the left of } (.693 - .353)/.754 \\ &= \text{area under } z \text{ curve to the left of } .45 \\ &= .6736\end{aligned}$$

Similarly, since $\ln(1) = 0$ and $(0 - .353)/.754 = -0.47$,

$$\begin{aligned}\text{proportion of pipelines with } 1 < x < 2 &= \text{area under } z \text{ curve between } -0.47 \text{ and } 0.45 \\ &= .6736 - .3192 \\ &= .3544\end{aligned}$$

TRY it in R

R: Distribution
Log-normal distribution

R_distribution_c.R

Continuous Distributions

A lognormal variable x is one for which ln(x) has a normal distribution with mean value μ. That is, $\mu_{\ln(x)}=\mu$. Therefore, it might seem that $\mu_x=e^{\mu}$, but this not the case. It can be shown that:

$$\boxed{\mu_x = e^{\mu+\sigma^2/2}}$$

$$\boxed{V(x) = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)}$$

Continuous Distributions

Definition:

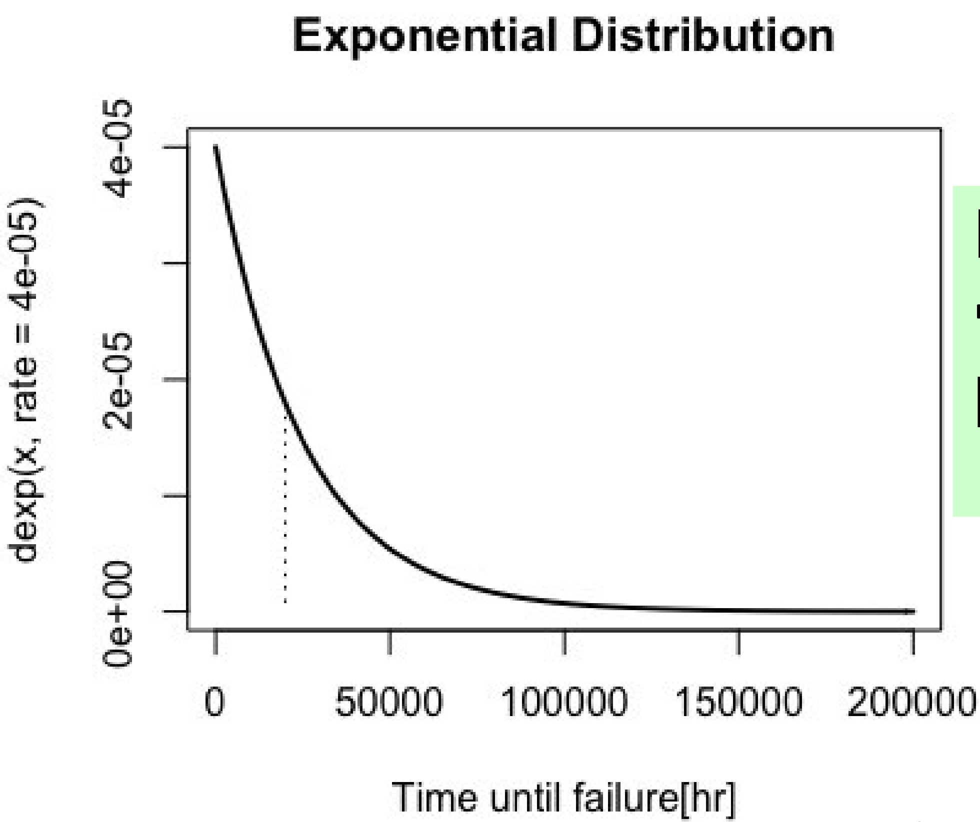
- A variable x is said to have an **exponential distribution** with parameter $\lambda > 0$ if the density function for x is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Continuous Distributions

連續函數Exponential Distribution來描述使用燃料引擎的風扇，其可正常運作的壽命。

$$f(x) = \lambda e^{-\lambda x}; \lambda = 0.00004$$



P(x ≤ 20,000)
➔
pexp(20,000, rate = 0.00004)
= 0.5506

Definition:

- A variable x has a **Weibull distribution** with parameters α and β if the density function of x is

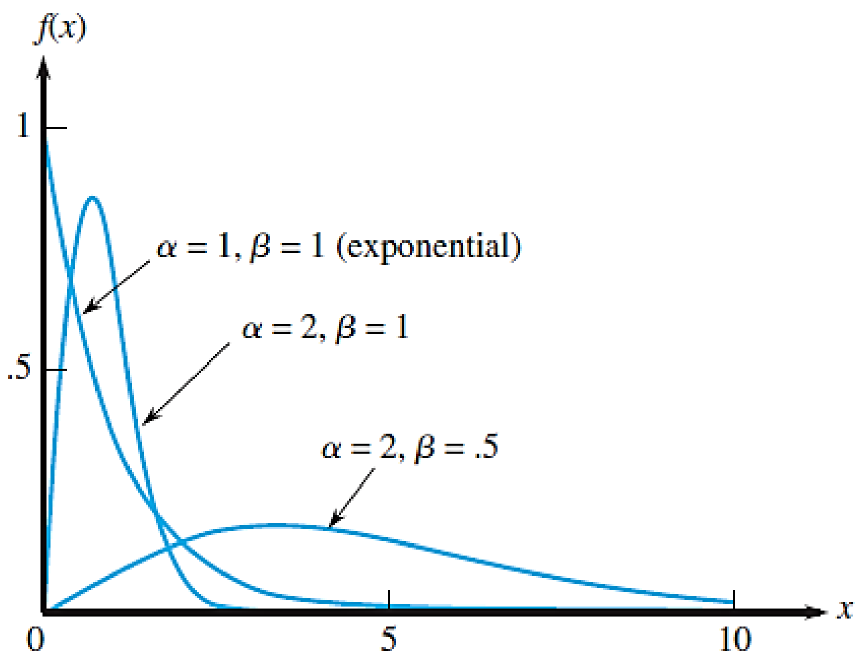
$$f(x) = \begin{cases} \dfrac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

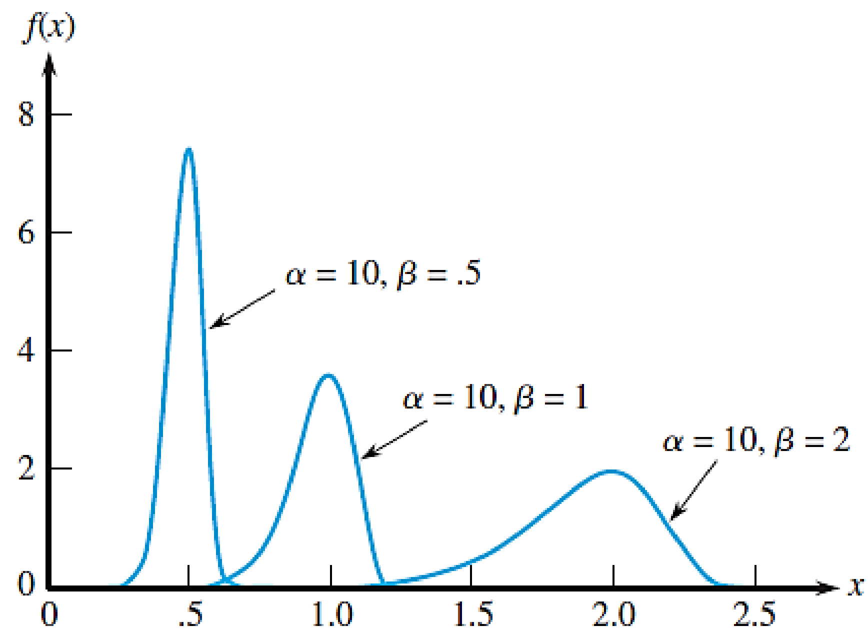
Area under density curve to the left of t = $\int_0^t f(x)dx$
= 1 - $e^{-(t/\beta)^{\alpha}}$

Continuous Distributions

1. Positively or Negative skewed
2. α= β = 1 for exponential distribution

dweibull(x, shape =, scale =)
Here:
shape = alpha
scale = beta

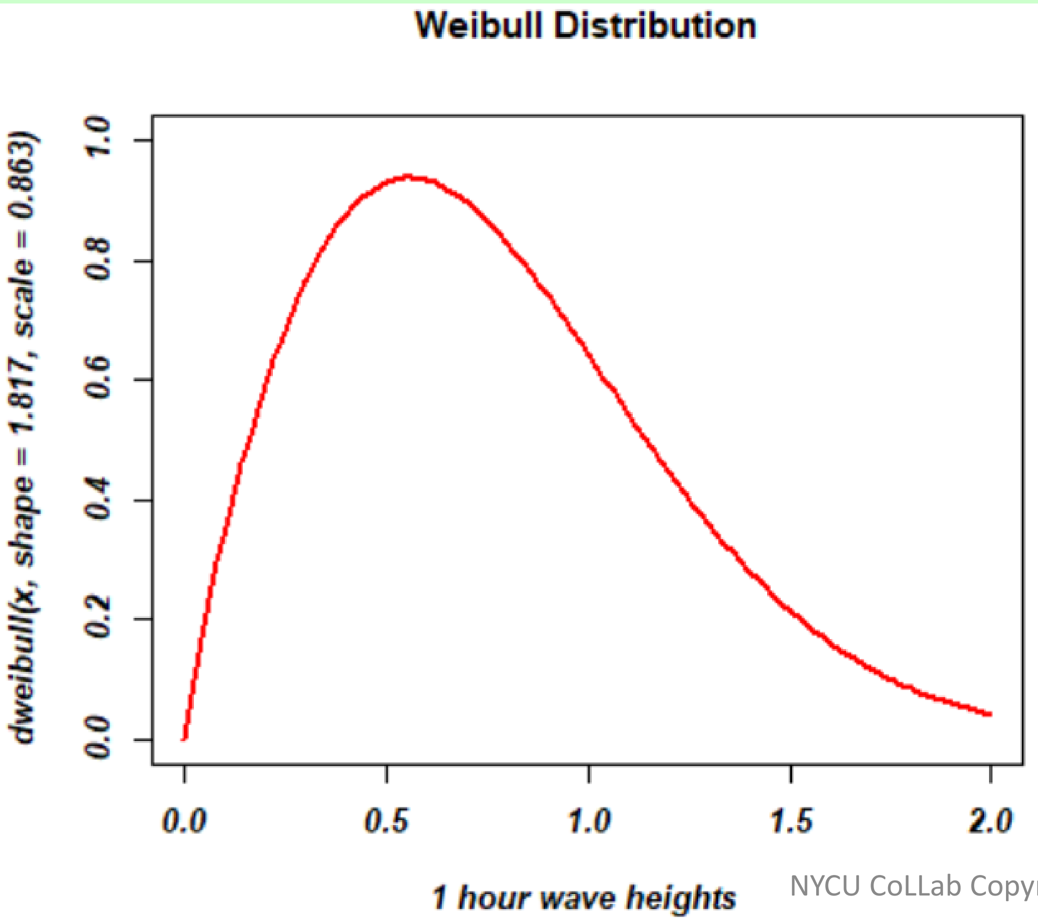




Continuous Distributions

前人研究發現離岸風機基座承受小時特定波浪高度的機率可以使用Weibull(alpha = 1.817, beta = 0.863)來描述。試著回答下列問題。

```
dweibull(x, shape = 1.817, scale = 0.863)
```



$$f(x)=\begin{cases}\dfrac{\alpha}{\beta^{\alpha}}x^{\alpha-1}e^{-(x/\beta)^{\alpha}} & x>0\\[2ex] 0 & x\le 0\end{cases}$$

Discrete Distributions

(特定時間內，某情況發生的次數)

- Usually used as a model for the number of times an "event" occurs during a specified time period or particular region of space
- The **Poisson mass function** is

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0, 1, 2, 3, \ldots$$

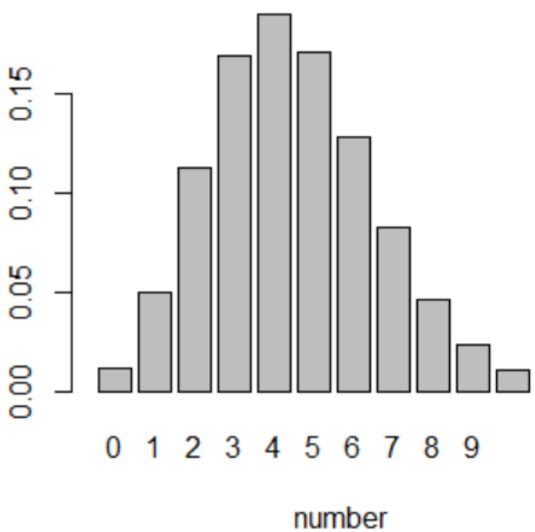
where the parameter λ must satisfy $\lambda > 0$.

Discrete Distributions

在一定的時間內，生物被捕捉的數量數據可以使用Poisson

Example 1.22 Let x denote the number of creatures of a particular type captured in a trap during a given time perio[redacted], so, on average, traps will contain 4.5 creatures. [The article "Dispersal Dynamics of the Bivalve *Gemma Gemma* in a Patchy Environment (*Ecological Monographs*, 1995: 1–20) suggests this model; the bivalve *Gemma gemma* is a small clam]. The proportion of traps with five creatures is

$$(\text{proportion with } x = 5) = \frac{e^{-4.5}(4.5)^5}{5!} = .1708$$




```
dpois(x = 0:10, lambda =4.5)
```

x:	0	1	2	3	4	5	6
$p(x)$:	.0111	.0500	.1125	.1687	.1898	.1708	.1281
x:	7	8	9	10	11	12	
$p(x)$:	.0824	.0463	.0232	.0104	.0043	.0016	

TRY
it
in
R

R: Distribution
Poisson distribution

R_distribution_d.R

Discrete Distributions

✓Let x be a Poisson variable with parameter λ.

✓**The mean value of x is λ itself.**

$$\mu_x = \sum_{x=0}^{\infty} x \cdot p(x) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}$$

✓ If we now let **y = x – 1**, the range of summation is from y = 0 to infinite

$$\lambda \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = \lambda$$

Discrete Distributions

$$\sigma^2 = \sum_{x=0}^{\infty} (x-\lambda)^2 \frac{e^{-\lambda}\lambda^x}{x!} = \lambda$$

Discrete Distributions

1組4個電池，其中品質良好的數量為x的機率為p(x)

成功機率 p = 0.9 則失敗機率q = 1-p = 0.1

Example 1.14 Consider a package of four batteries of a particular type, and let x denote the number of satisfactory (i.e., nondefective) batteries in the package. Possible values of x are 0, 1, 2, 3, and 4. One reasonable distribution for x is specified by the following mass function:

$$p(x) = \frac{24}{x!(4-x)!}(.9)^x(.1)^{4-x} \qquad x = 0, 1, 2, 3, 4$$

where "!" is the factorial symbol (e.g., $4! = (4)(3)(2)(1) = 24$, $1! = 1$, and $0! = 1$). This looks a bit intimidating, but there is an intuitive argument leading to $p(x)$ that we will mention shortly. Substituting $x = 3$, we get

$$p(3) = \frac{24}{(6)(1)}(.9)^3(.1)^1 = .2916$$

Discrete Distributions

超過0.9機率至少有兩個電池是良好的!!

That is, roughly 29% of all packages will have three good batteries. Substituting the other x values gives us the following tabulation:

x:	0	1	2	3	4
$p(x)$:	.0001	.0036	.0486	.2916	.6561

The proportion of packages with at least two good batteries is

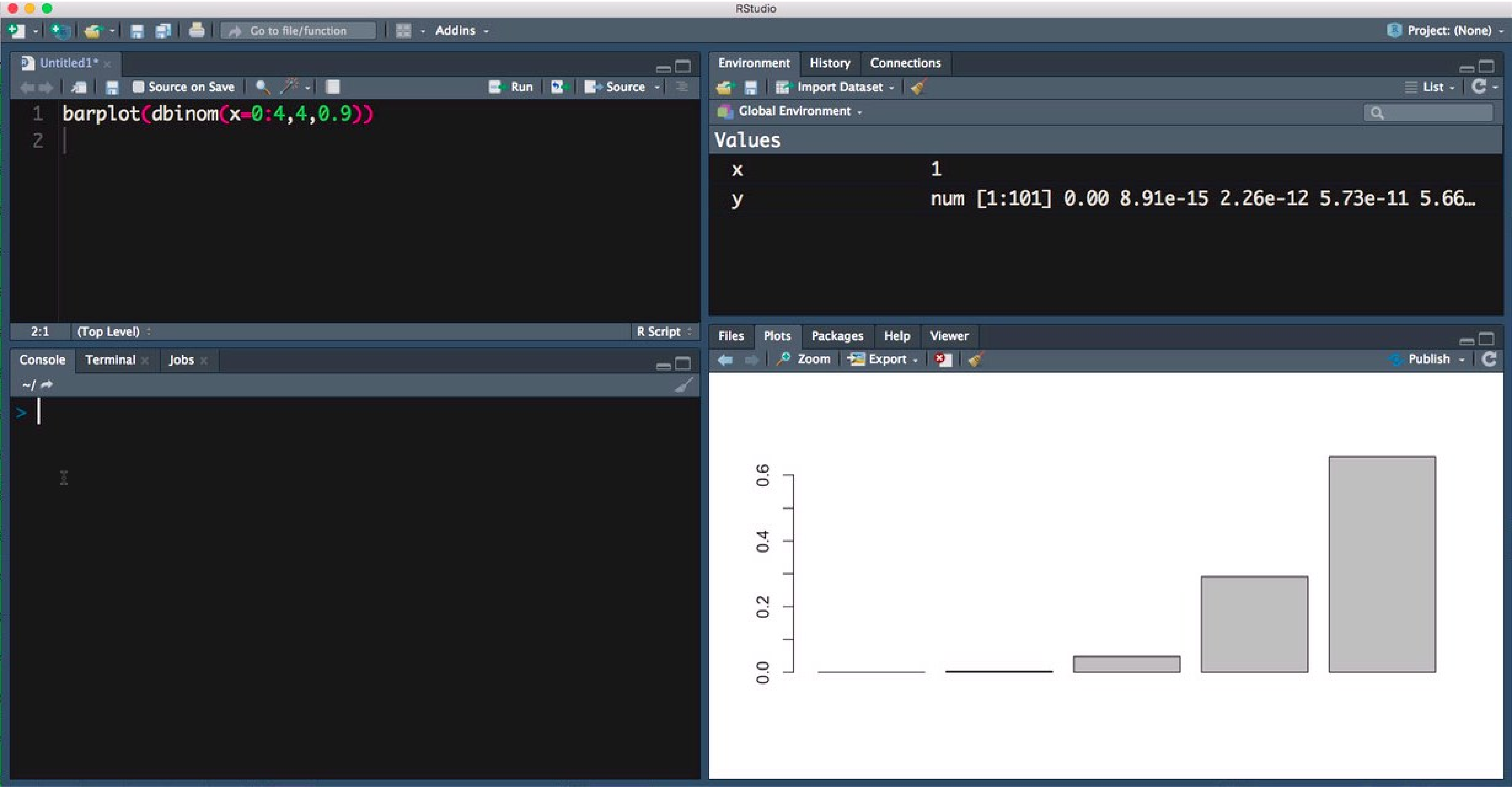
$$\begin{array}{l}\text{proportion of packages with } x \\ \text{values between 2 and 4 (inclusive)}\end{array} = p(2) + p(3) + p(4) = .9963$$

More than 99% of all packages have at least two good batteries.

Discrete Distributions

$$p(x) = C_x^n p^x q^{n-x}, x = 0,1,2,\ldots,n$$

dbinom(x,n,p)



Discrete Distributions

- If x is a binomial variable with parameters n = group size and π = success proportion,
- then $\mu_x = n\pi$.

$$\mu_x = \sum x \cdot p(x) = \sum x \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x}$$

$$= n\pi$$

Discrete Distributions

$$\sigma^2 = \square\,(x - n\pi)^2 \square p(x) = \square\,(x - n\pi)^2 \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$$

$$= n\pi(1-\pi)$$

✓ <u>When π=0.5, is maximized for σ value</u>

Discrete Distributions

(數量大且發生機率小)

- Many applications of Poisson distribution are in fact based on an underlying binomial situation without explicitly-stated values of and n and π.

The Poisson Approximation to the Binomial Distribution

Often a binomial scenario involves a group size n that is quite large in combination with a success proportion π close to zero. Under such circumstances, the binomial mass function can be well approximated by the Poisson mass function with $\lambda = n\pi$. In particular, if $n \geq 100$, $\pi \leq .01$, and $\lambda = n\pi \leq 20$, then

$$\frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x} \approx \frac{e^{-\lambda}\lambda^x}{x!}$$

A more formal statement of this result is that the Poisson mass function on the right-hand side is the limit of the binomial function on the left as $n \to \infty$, $\pi \to 0$ in such a way that $n\pi \to \lambda$.

R: Distribution

Appropriate distribution

"MASS"

R: Distribution

Appropriate distribution

fitdistr(x, densfun)

R: Distribution Appropriate distribution

-**densfun** is distribution of x

- "beta"
- "normal"
- "chi-squared"
- "exponential"
- "f"
- "gamma"
- "lognormal"
- "Poisson"
- "t"
- "weibull"

R: Distribution
Appropriate distribution

-Output value

- estimate: the parameter estimates
- **sd**: the estimated standard errors
- **vcov**: the estimated variance-covariance matrix
- **loglik**: the log-likelihood (數值越大代表函數擬合的程度越好)

distribution fitting process

Distribution

Fit your data to likely distribution

Check the fit of the data

Selecting an appropriate distribution

研究員提出設計方法使得樁擁有更好的工作效率，以下為17筆現地測試之樁的長度與直徑比值資料

Example 2.17 There has been recent increased use of augered cast-in-place (ACIP) and drilled displacement (DD) piles in the foundations of buildings and transportation structures. In the article "Design Methodology for Axially Loaded Auger Cast-in-Place and Drilled

Displacement Piles" (*J. Geotech. Geoenviron. Engr.*, 2012: 1431–1441) researchers propose a design methodology to enhance the efficiency of these piles. The authors reported the following length-diameter ratio measurements based on 17 static-pile load tests on ACIP and DD piles from various construction sites. The values of p for which z percentiles are needed are $(1 - .5)/17 = .029$, $(2 - .5)/17 = .088, \ldots,$ and $.971$.

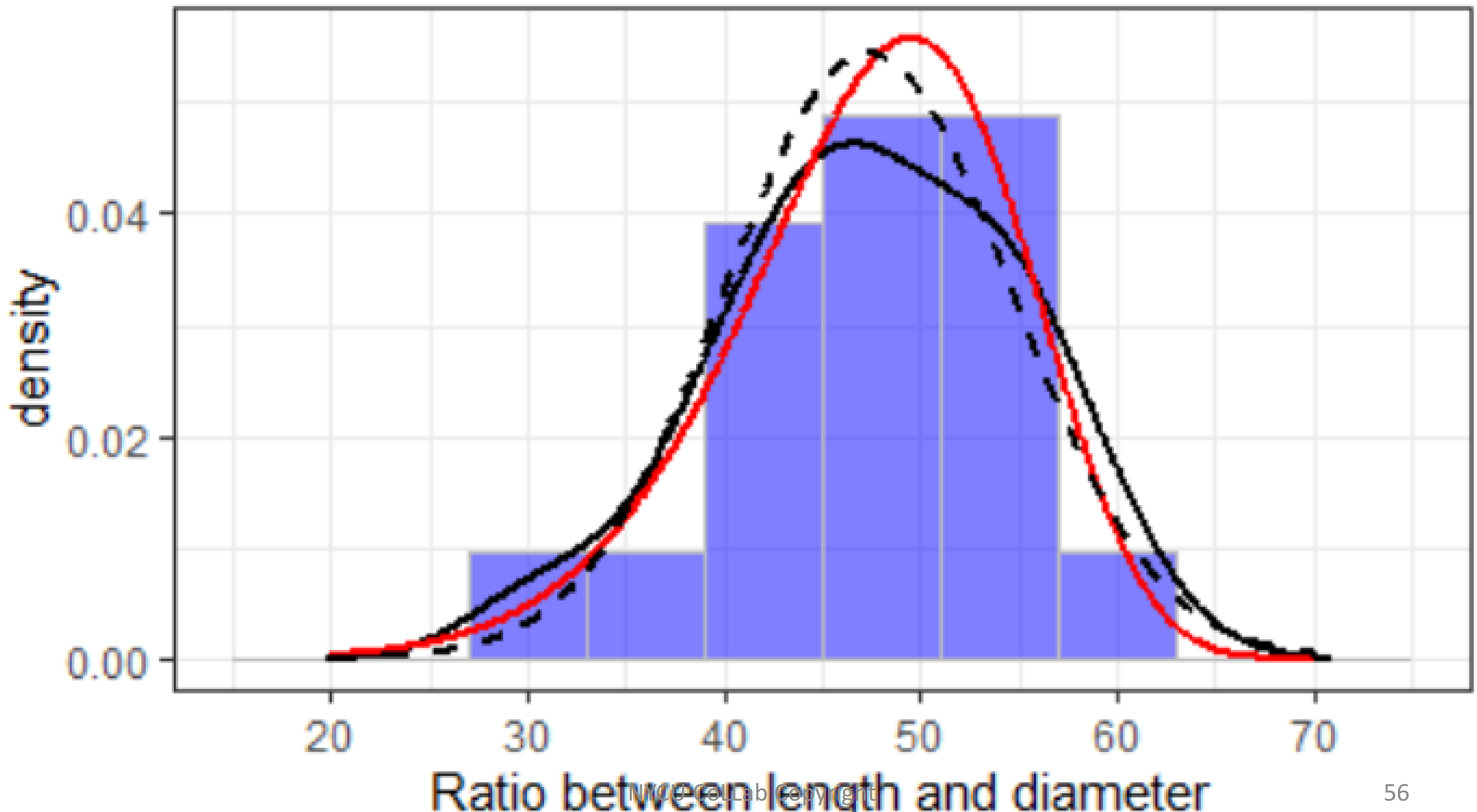
$x_{(i)}$:	30.86	37.68	39.04	42.78	42.89	42.89	45.05	47.08	47.08
z percentile:	−1.89	−1.35	−1.05	−0.82	−0.63	−0.46	−0.30	−0.15	0.00
$x_{(i)}$:	48.79	48.79	52.56	52.56	54.8	55.17	56.31	59.94	
z percentile:	0.15	0.30	0.46	0.63	0.82	1.05	1.35	1.89	

TRY
it
in
R

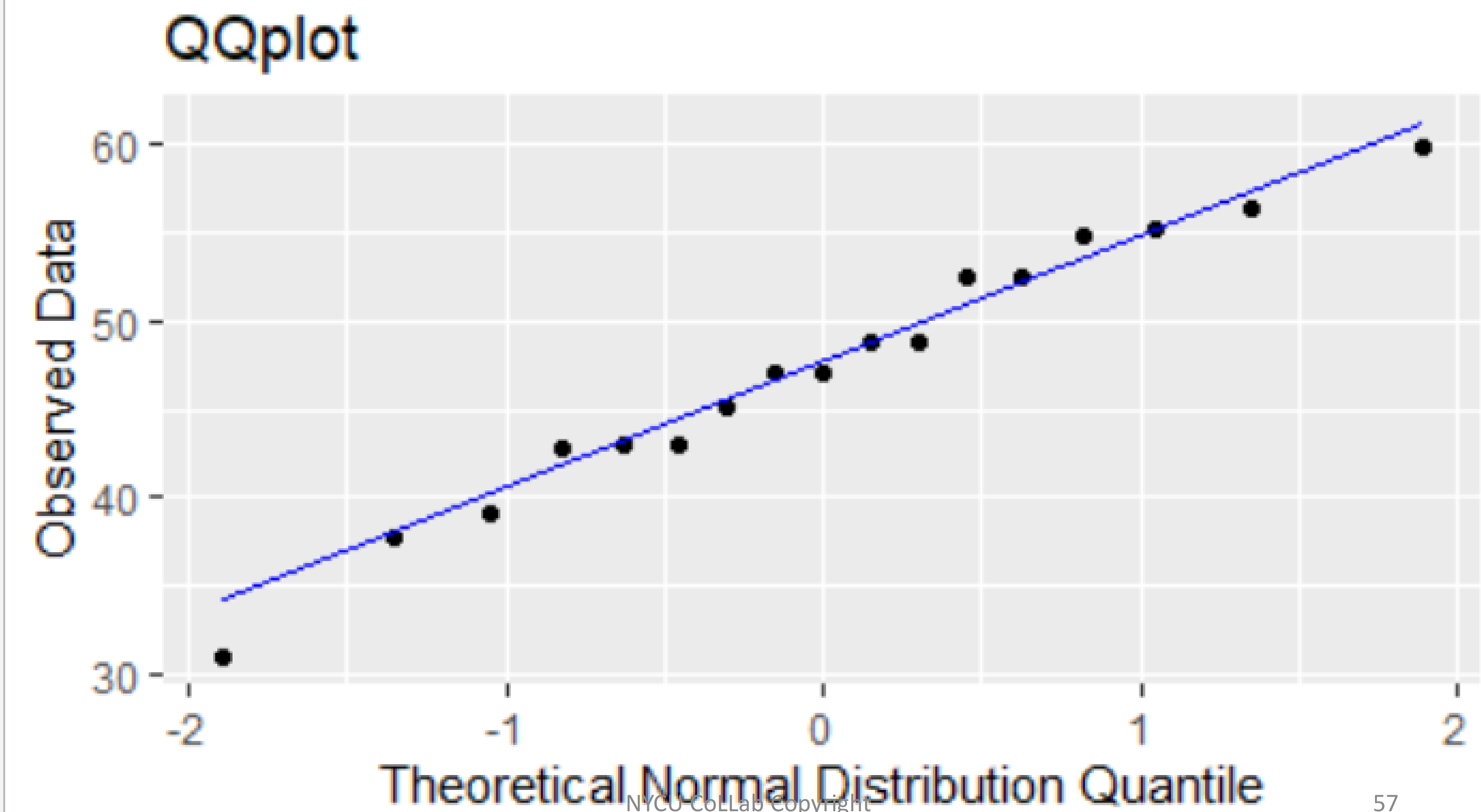
R: Distribution fitting distribution

R_distribution_e.R

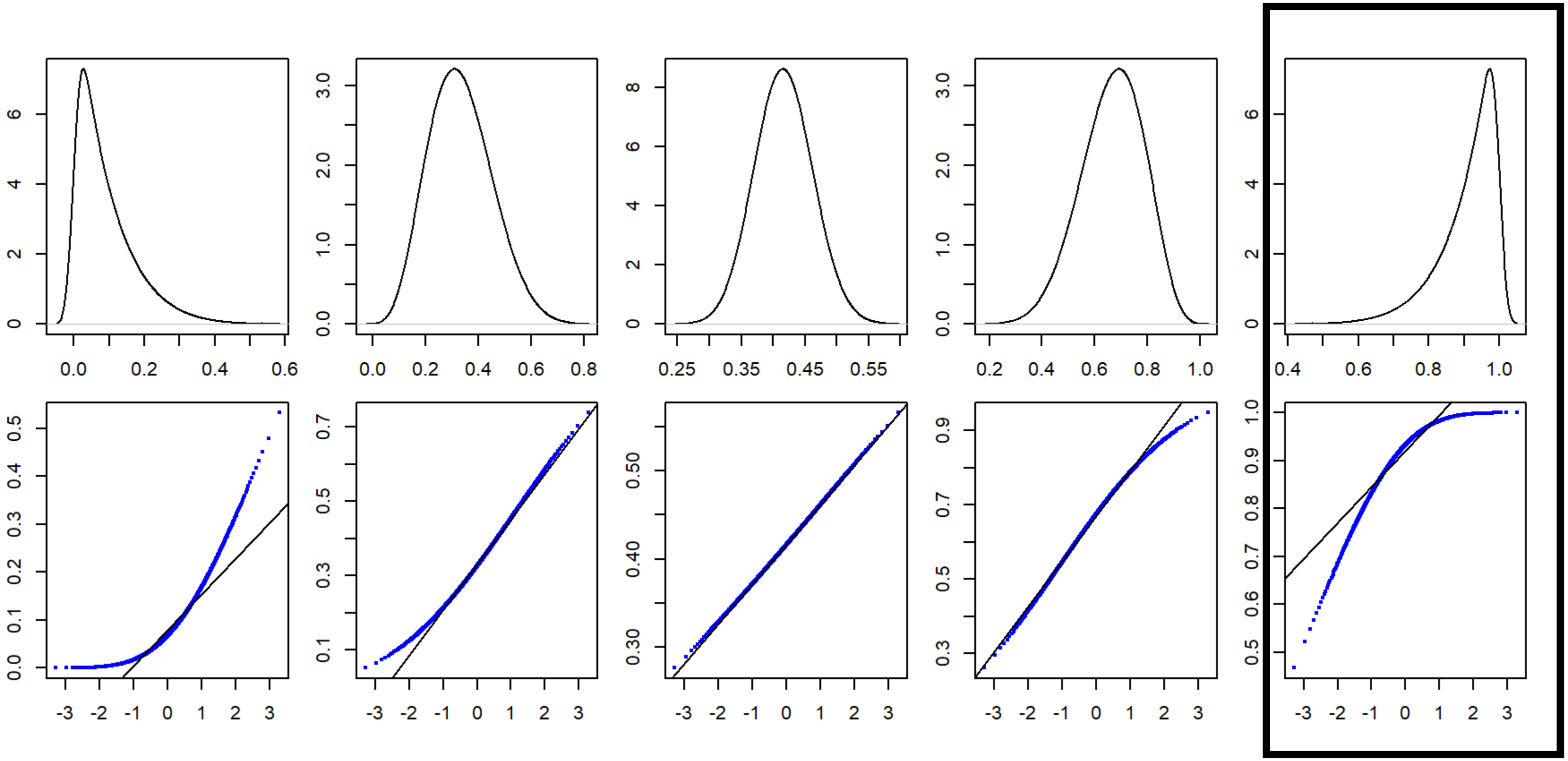
R: Distribution fitting distribution



R: Distribution fitting distribution



R: Distribution fitting distribution



課堂練習: 學號-姓名-ch6-Distribution.R

25個圓柱狀試體的彈性模數大小數據:

37.0 37.5 38.1 40.0 40.2 40.8 41.0 42.0 43.1 43.9 44.1 44.6 45.0 46.1 47.0 50.2 55.0 56.0 57.0 58.0 62.0 64.3 68.8 70.1 74.5

單位: MPa

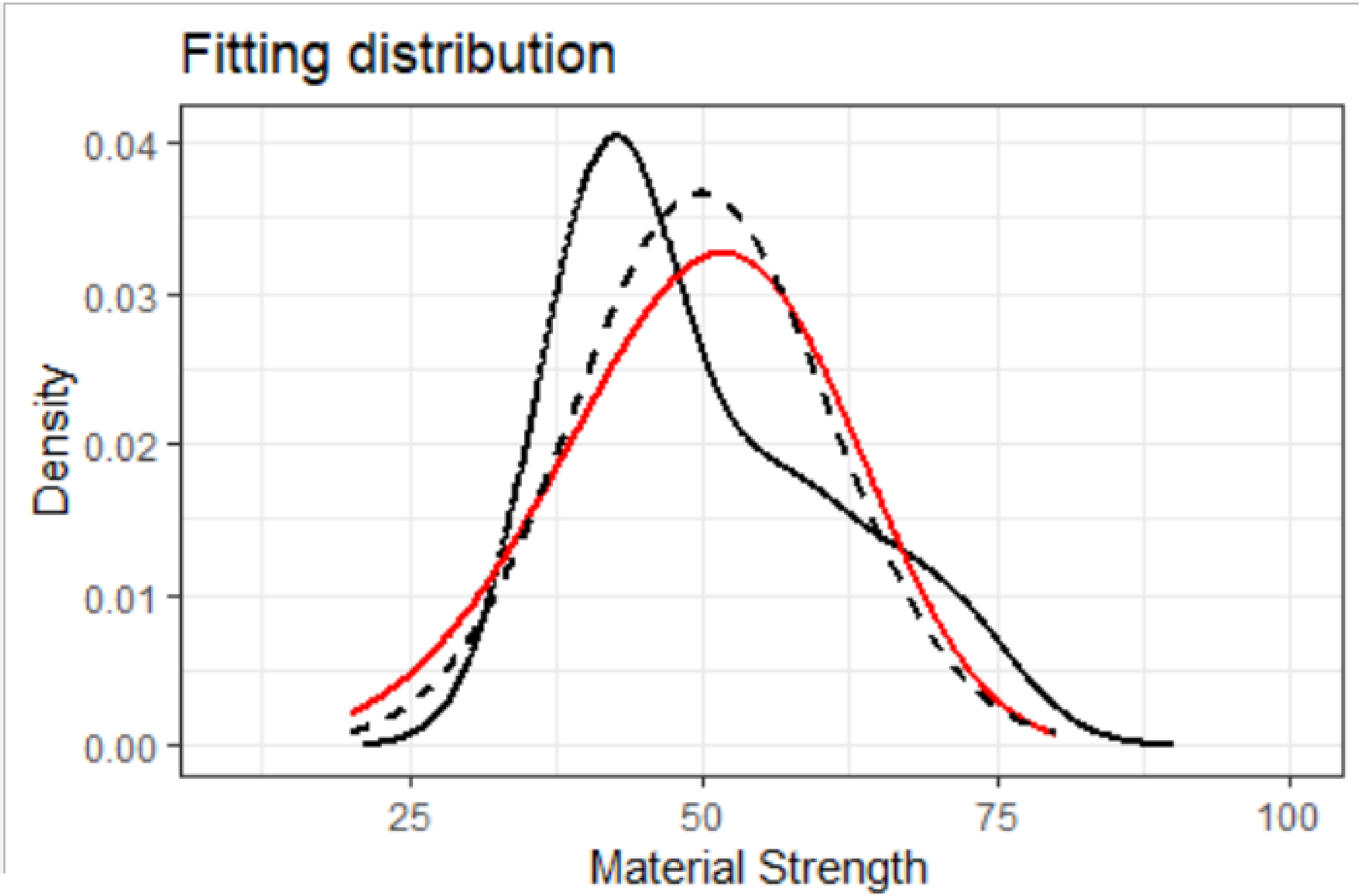
試著回答以下問題:

(1) Fitting the data with "Weibull" and "Normal" distributions and then comparing them with density curve. (red line for Weibull, dashed line for Normal, solid line for density curve)

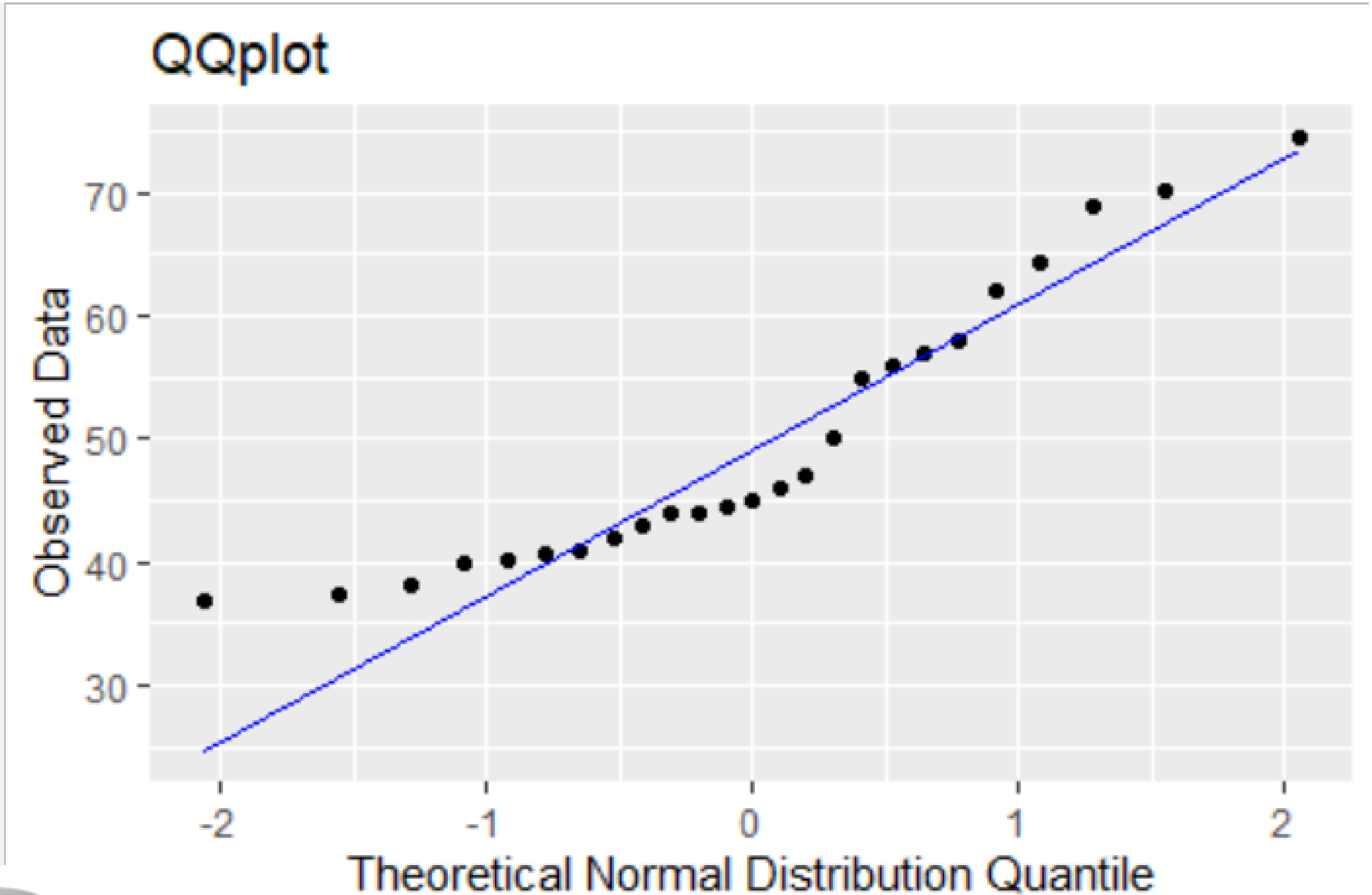
(2) Construct a normal quantile plot(檢驗分布是否為常態分布)

(3) Make comment on the plausibility of a normal population distribution(描述自己觀察到的結果)

課堂練習: 學號-姓名-ch6-Distribution.R



課堂練習: 學號-姓名-ch6-Distribution.R



R: Distribution fitting distribution