Psychol. Statistics using R

Unreliability & Assumptions

Changwei W. Wu, Ph.D.

Graduate Institute of Mind, Brain and Consciousness
Research Center of Brain and Consciousness
Taipei Medical University

Statistics

1. **Measuring Unreliability of Inference**
   - Central Limit Theorem
   - Variance & Standard Error of the Means (SEM)
2. **Small sample size**
   - t distribution
   - Degree of freedom
3. **[R] GGplot & Assumption check**
   - Normality & Homogeneity of Variance

Functions for Probability Distributions

	Random variables	*Probability density function (PDF)*	*Cumulative distribution function (CDF)*	*Quantile*
Normal distribution	***rnorm***	***dnorm***	***pnorm***	***qnorm***
Binomial distribution	***rbinom***	***dbinom***	***pbinom***	***qbinom***
χ^2 distribution	***rchisq***	***dchisq***	***pchisq***	***qchisq***
t distribution	***rt***	***dt***	***pt***	***qt***
F distribution	***rf***	***df***	***pf***	***qf***
		Given specific value (x), calculate corresponding probability (y)	Given specific value (x), calculate cumulative probability from $-\infty$	Given cumulative probability, calculate the corresponding value (x)

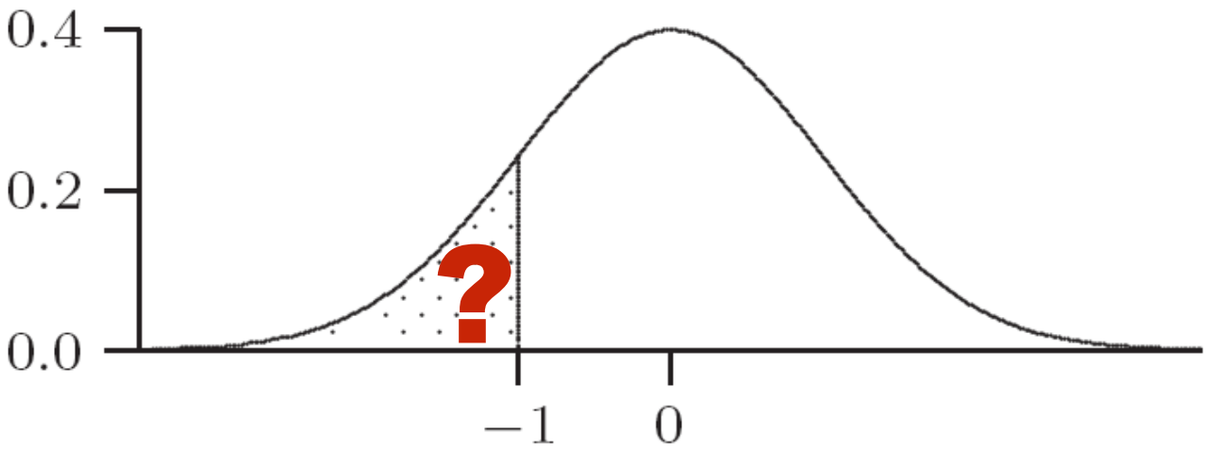
TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

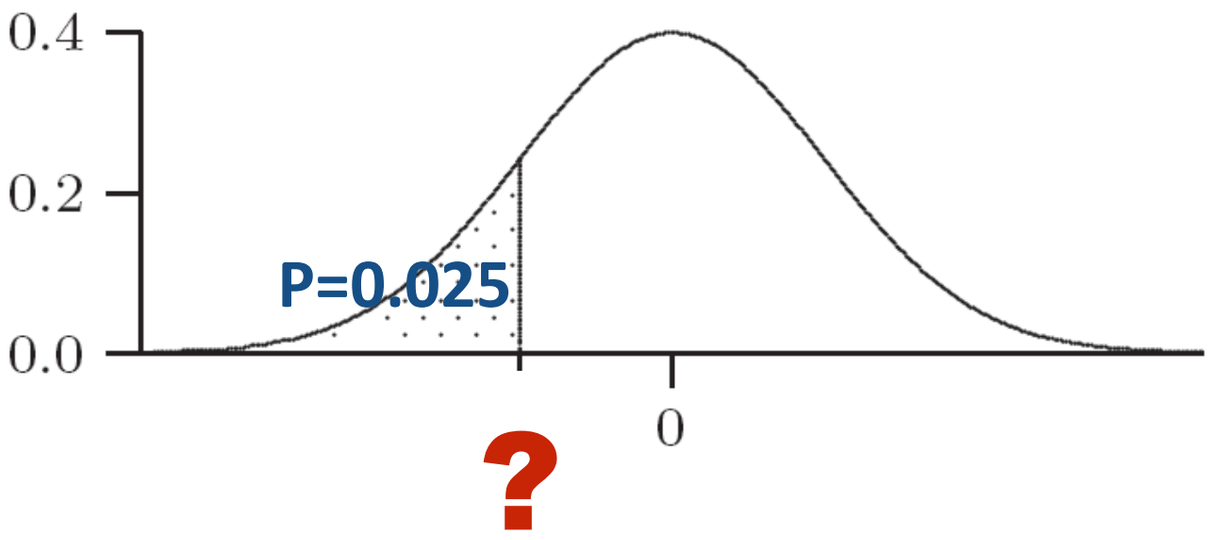
Standard Normal Distribution (Z Distribution)

DEFINITION 3.10. Let X be a normal random variable with mean μ, and standard deviation σ. The transformation

$$Z = \frac{X - \mu}{\sigma}$$

expresses X as the **standard normal random variable** Z with $\mu = 0$ and $\sigma = 1$.





➡ **Use R function to calculate shaded area**

➡ **Pr(Z ≤ a)=0.025, what is the 'a' value ?**

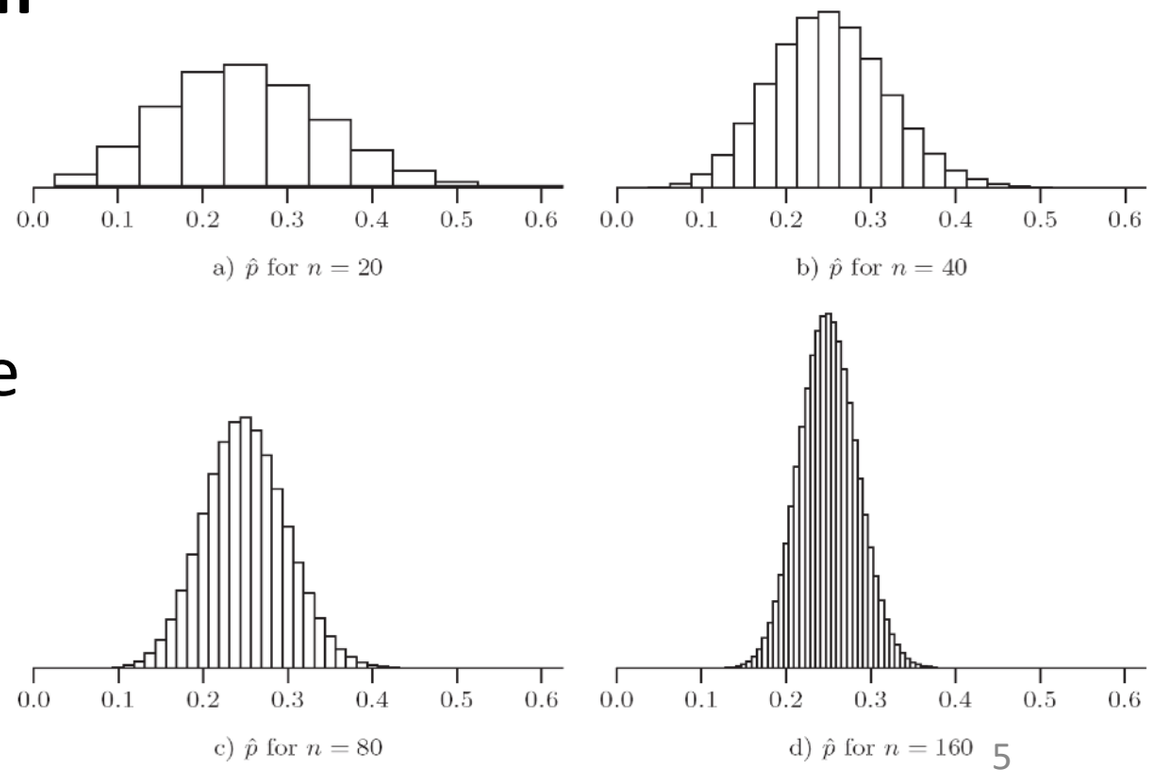
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Central Limit Theorem

Central-Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2. Then for large n, $\overline{X} \dot{\sim} N\left(\mu, \sigma^2/n\right)$ even if the underlying distribution of individual observations in the population is not normal. (The symbol $\dot{\sim}$ is used to represent "approximately distributed.")

- In probability theory, **central limit theorem** (**CLT**) states conditions under which the mean of a sufficiently large number of independent random variables (**n≧30**), each with finite mean and variance, will be approximately normally distributed.



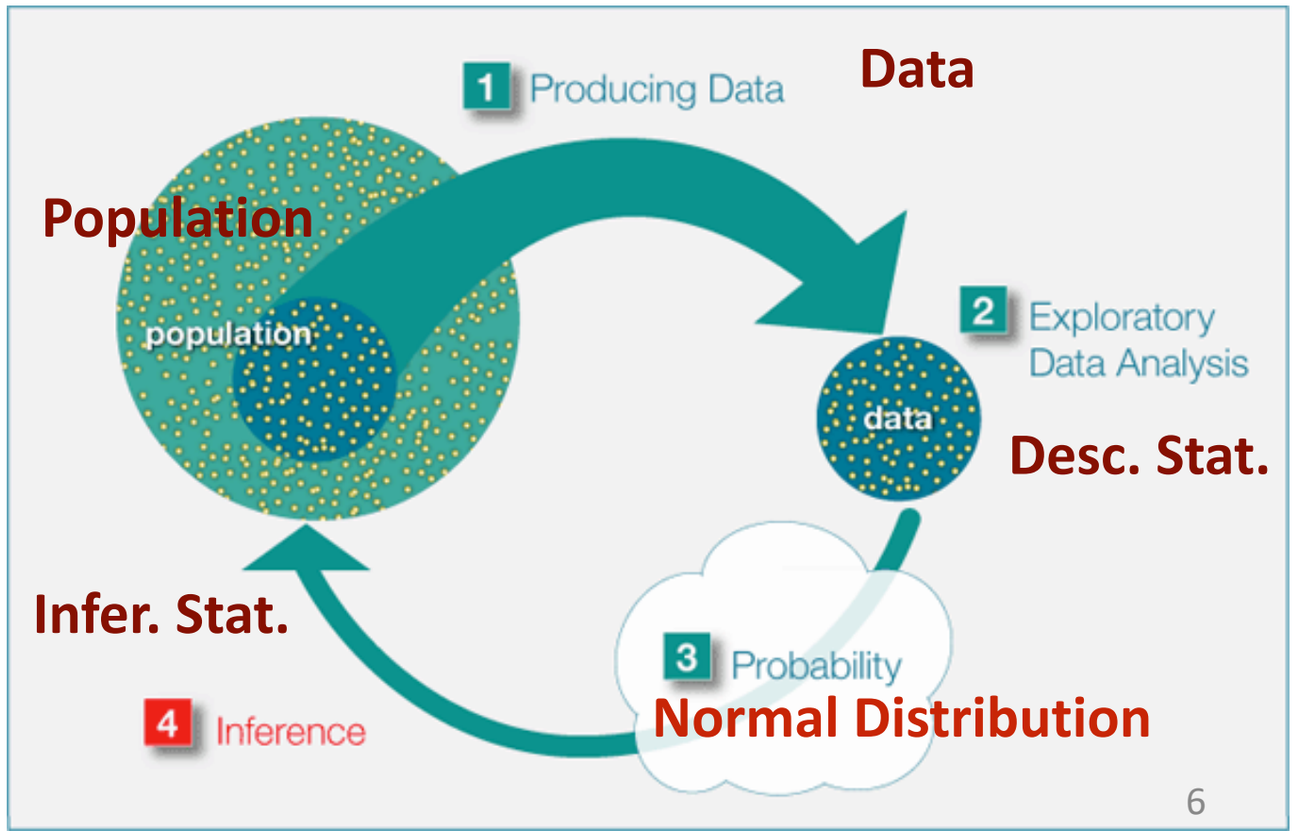
臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

Inference on Central Limit Theorem

Central-Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2. Then for large n, $\overline{X} \dot{\sim} N\left(\mu, \sigma^2/n\right)$ even if the underlying distribution of individual observations in the population is not normal. (The symbol $\dot{\sim}$ is used to represent "approximately distributed.")

- If it is normal distribution, how do we know our inference to population is correct?
- How to measure the **Unreliability** of inference from the random samples (the dataset)?



https://bolt.mph.ufl.edu/6050-6052/unit-4/

TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Measure of Unreliability

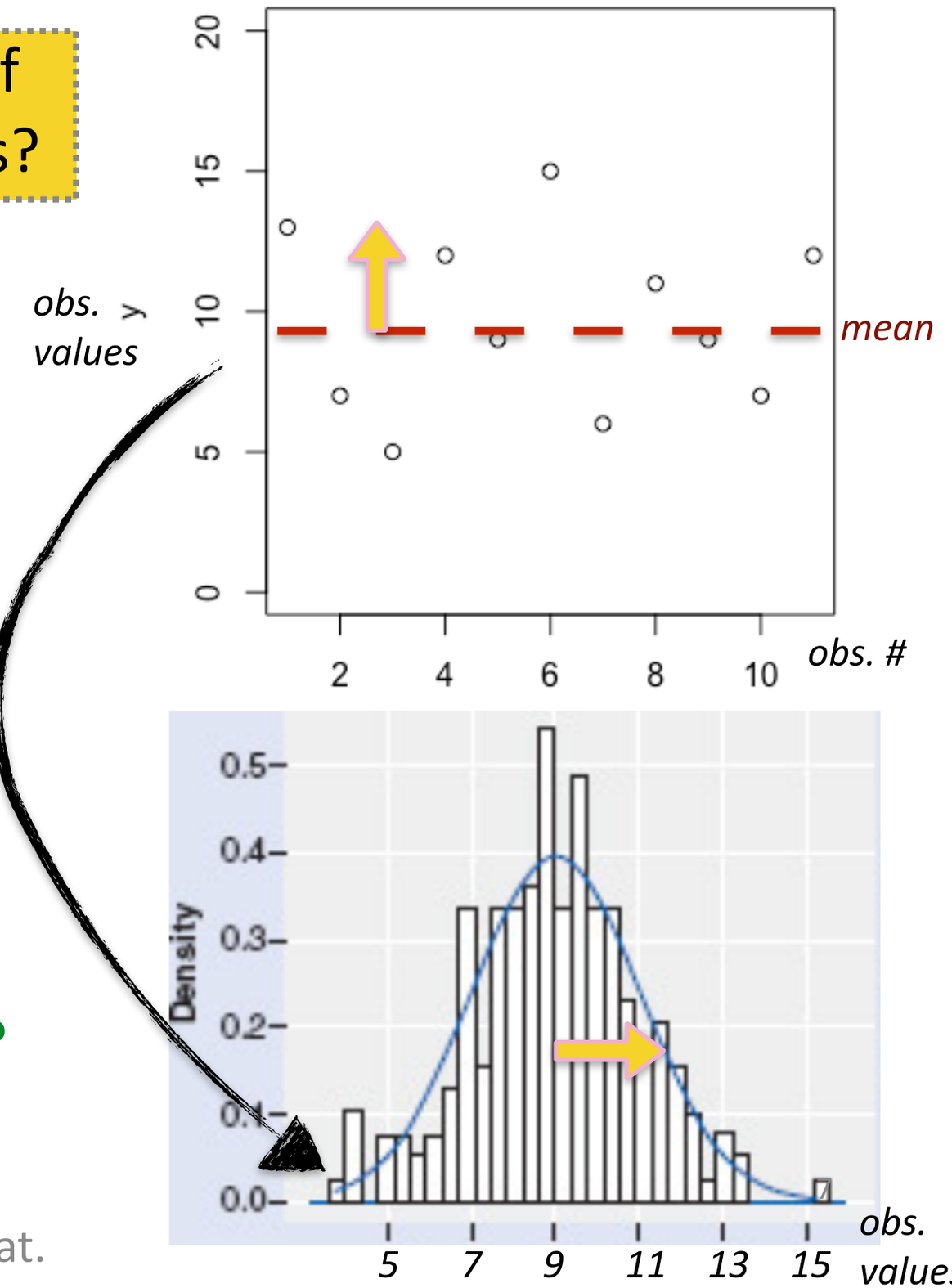
- How to measure the **Unreliability** of inference from the random samples?

- Unreliability depends on:
  - ***variance*** (s^2)

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

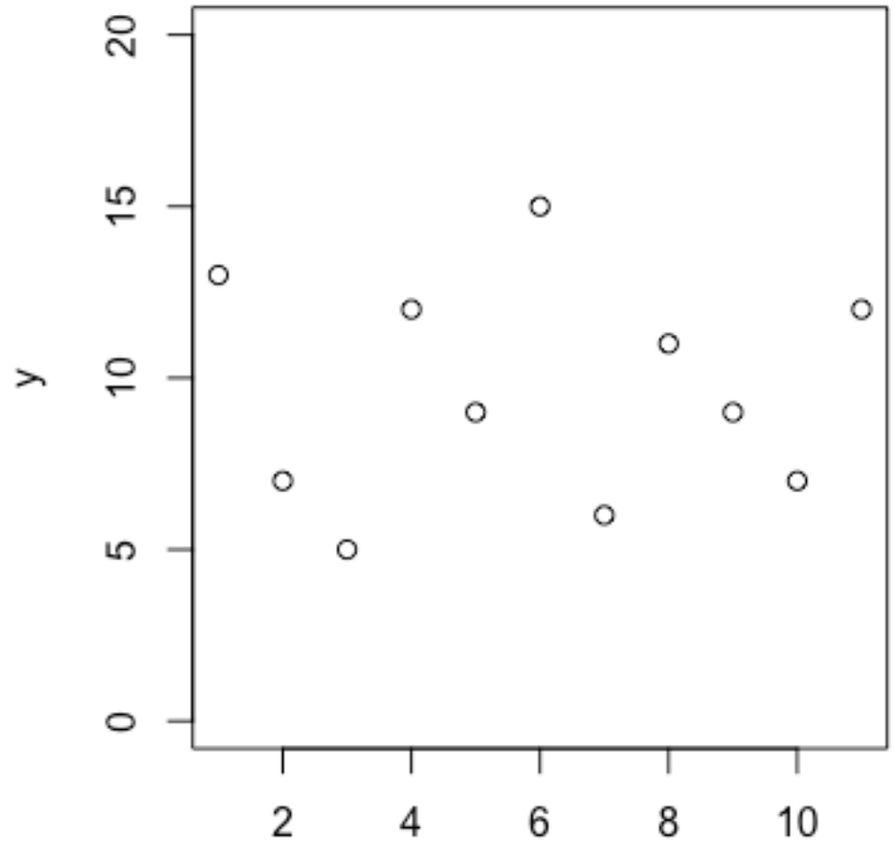
  - ***sample size*** (n)

Variance? Standard Deviation (SD)?



Measure of Unreliability

- How to measure the **Unreliability** of inference from the random samples?



- Unreliability depends on:
  - ***variance*** (s^2)

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

  - ***sample size*** (n)

More subject number, getting closer to the population

… but we already consider "sample size" in variance…

Standard Error of the Means (SEM)

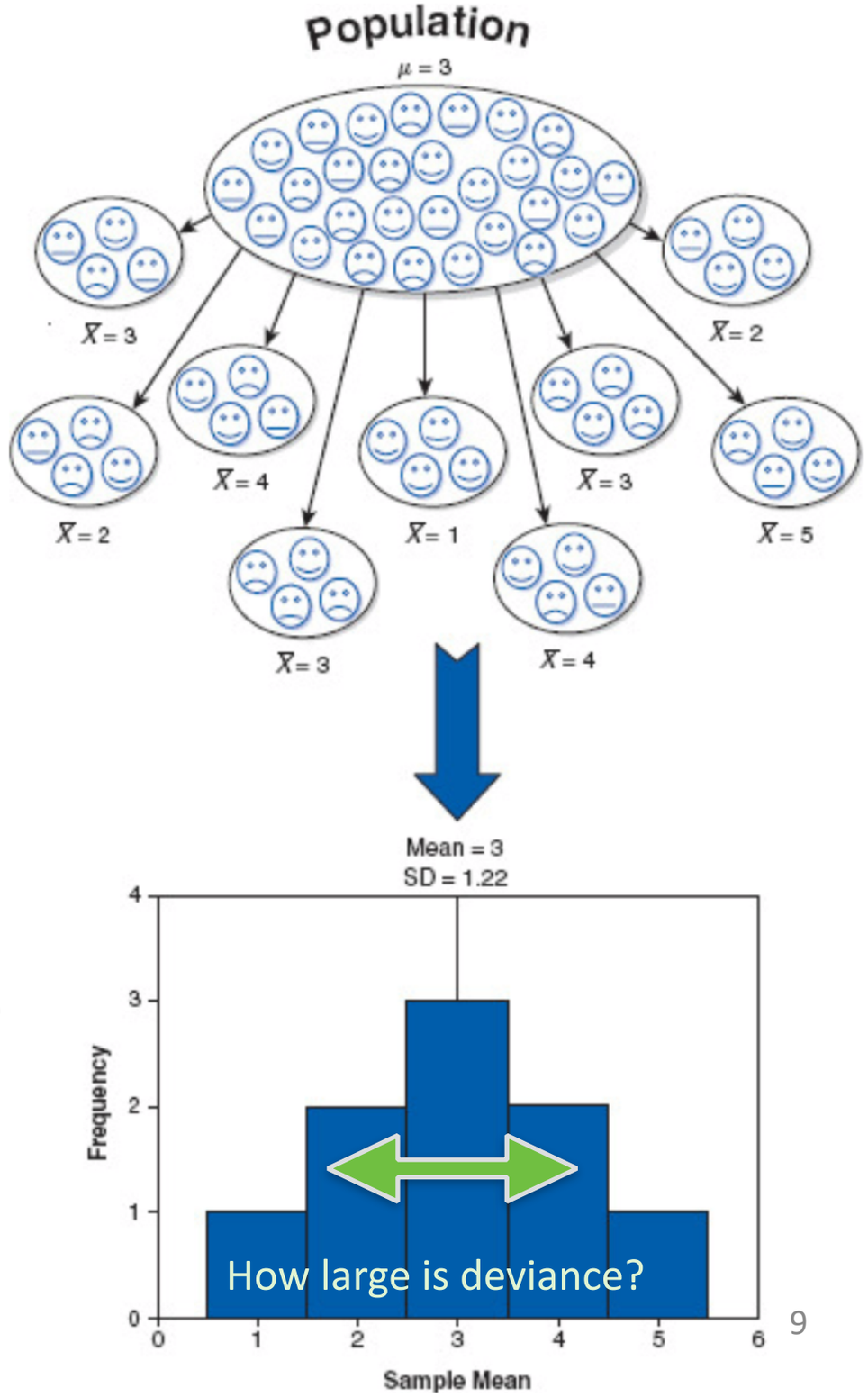
- How to measure the **Unreliability** of inference from the random samples?

- Unreliability depends on:
  - ***variance*** (s^2)
  - ***sample size*** (n)

- **Standard error** (SE) indicates:

  ➡ The variability *across* sample means from the same population

  ➡ ***Large SE***: the given sample is not one accurate reflection of the population. (high unreliability)



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Standard Error of the Means (SEM)

- How to measure the **Unreliability** of inference from the random samples?

- Unreliability depends on:
  - ***variance*** (s^2)
  - ***sample size*** (n)

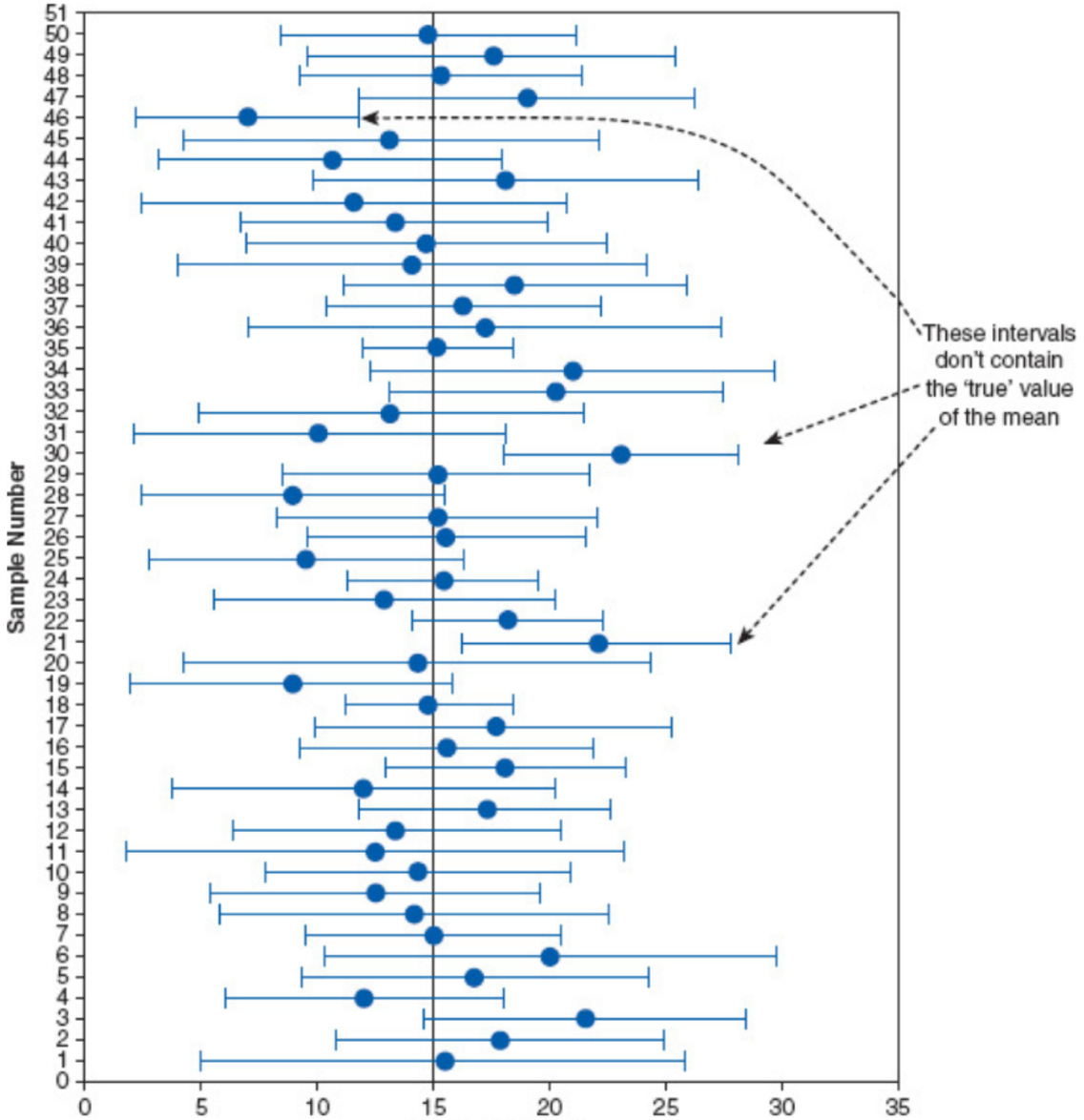
$$\text{unreliability} \propto s^2$$

Level of representative based on sample size

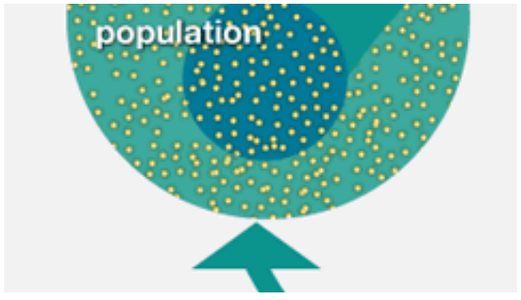
$$\text{unreliability} \propto \frac{s^2}{n}$$

To keep the same unit of the parameter…

$$SE_{\bar{y}} = \sqrt{\frac{s^2}{n}}$$



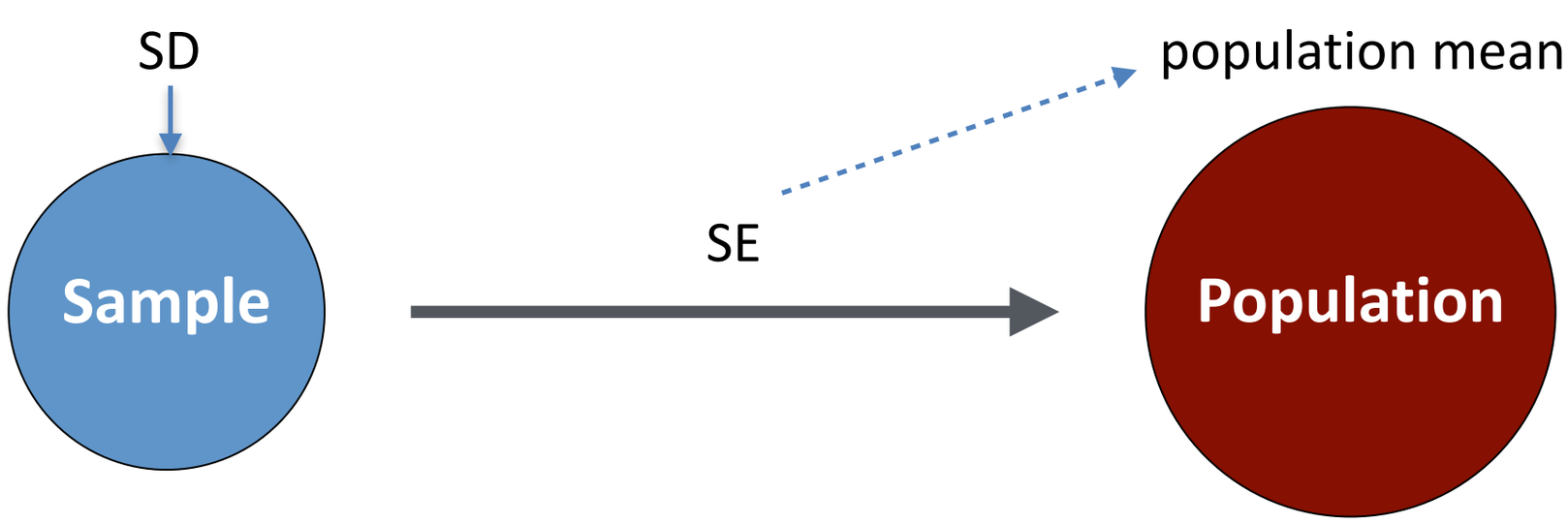
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY



Question

What is standard deviation?

What is standard error?

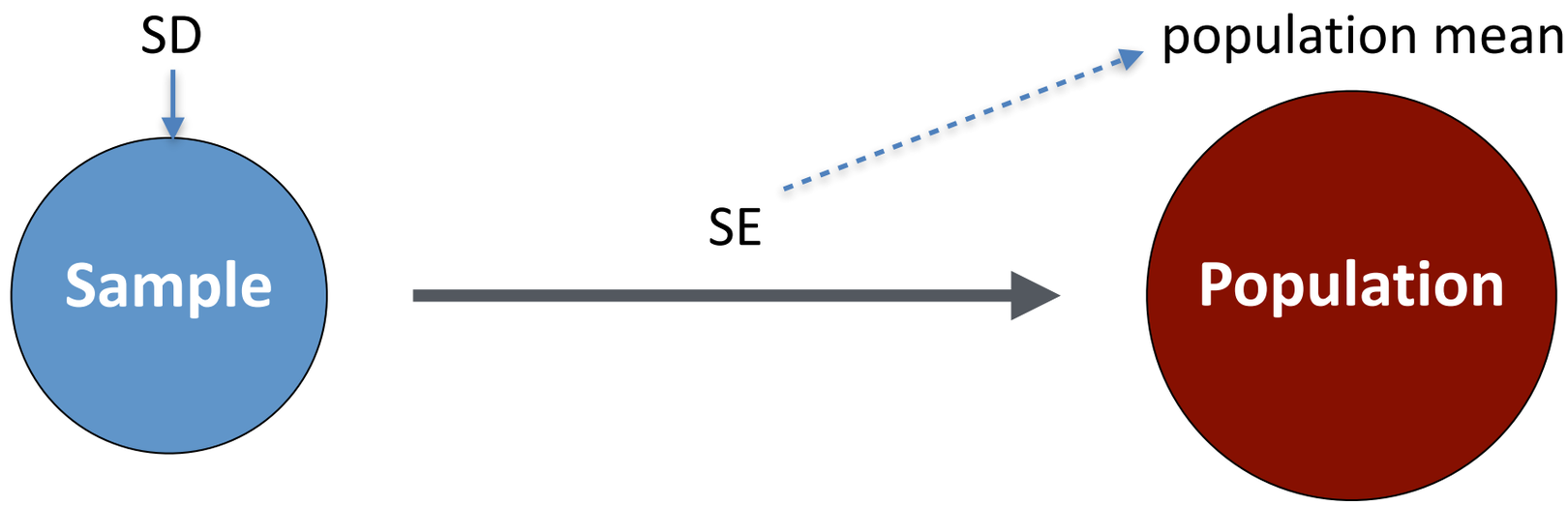


Is it possible that SE equals 0?

Standard Error of the Means

- Average of all sample means → the population mean.
- **Standard error** is the standard deviation of sample means.
- It is a measure of **how representative a sample is likely to be of the population**.
- Large standard error means lots of variability between the means of different samples and so the sample we have in hand might not be representative of the population.
- Small standard error indicates that most sample means are similar to the population mean and so our sample is likely to be an accurate reflection of the population.

$$SE_{\bar{y}} = \sqrt{\frac{s^2}{n}}$$



Confidence Interval

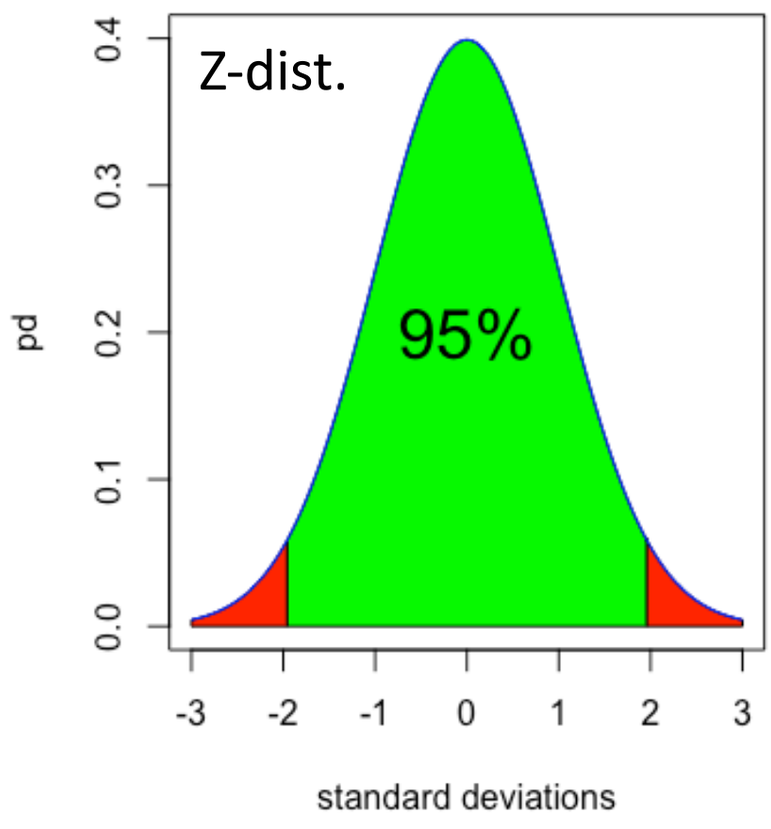
- *Confidence interval*: the likely range in which the mean would fall if the sampling exercise were to be repeated.

$$\text{unreliability} \propto \sqrt{\frac{s^2}{n}}$$

  - Typical range: confidence level of ***CDF*** = 95% (or 99%)

➡ **If the distribution is known, we can set a "confidence factor" for decision.**

➡ **Use R to calculate the value for the shaded area.**



Confidence Interval

- *Confidence interval*: the likely range in which the mean would fall if the sampling exercise were to be repeated.

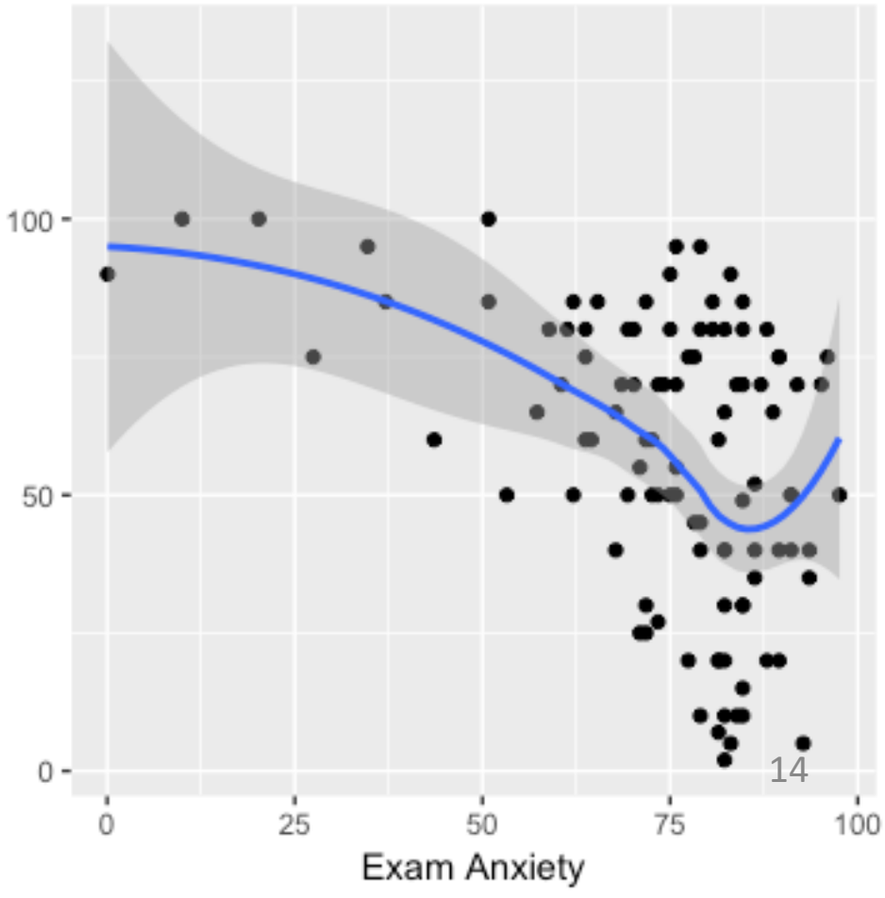
$$\text{confidence interval} \propto \sqrt{\frac{s^2}{n}}$$

- Typical range: confidence level of ***CDF*** = 95% (or 99%)

$$C\left(\overline{X} - 1.960\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.960\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Using standard normal distribution (Z) if the sample size is large enough.

Confidence factor (Normal distribution)



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Confidence Interval

Notice that to determine a confidence interval requires four different numbers:

1. a point estimate, here the sample mean $\overline{X}$;
2. a measure of variability, here the standard error of the mean, $\frac{\sigma}{\sqrt{n}}$;
3. a desired level of confidence $1-\alpha$, in this case $1-\alpha = 0.95$, so $\alpha = 0.05$;
4. and the sampling distribution of the point estimate, here the standard normal distribution, which provided the confidence factor with which to adjust the variability for the desired level of confidence, in this case 1.960.

Using this language, the endpoints of the confidence interval have the form

$$\text{point estimate} \pm (\text{confidence factor})(\text{standard error}). \tag{4.3}$$

$$\text{lower boundary of confidence interval} = \bar{X} - (1.96 \times SE)$$
$$\text{upper boundary of confidence interval} = \bar{X} + (1.96 \times SE)$$

An Introduction to Biostatistics 2008, 2nd Ed. By Thomas Glover and Kevin Mitchell. Ting Lung Book Co.

In Small Sample Size

- What if the sample size is not that “large”?
  - **Student’s *t* - distribution** (1908)
    - An approximation of normal distribution.

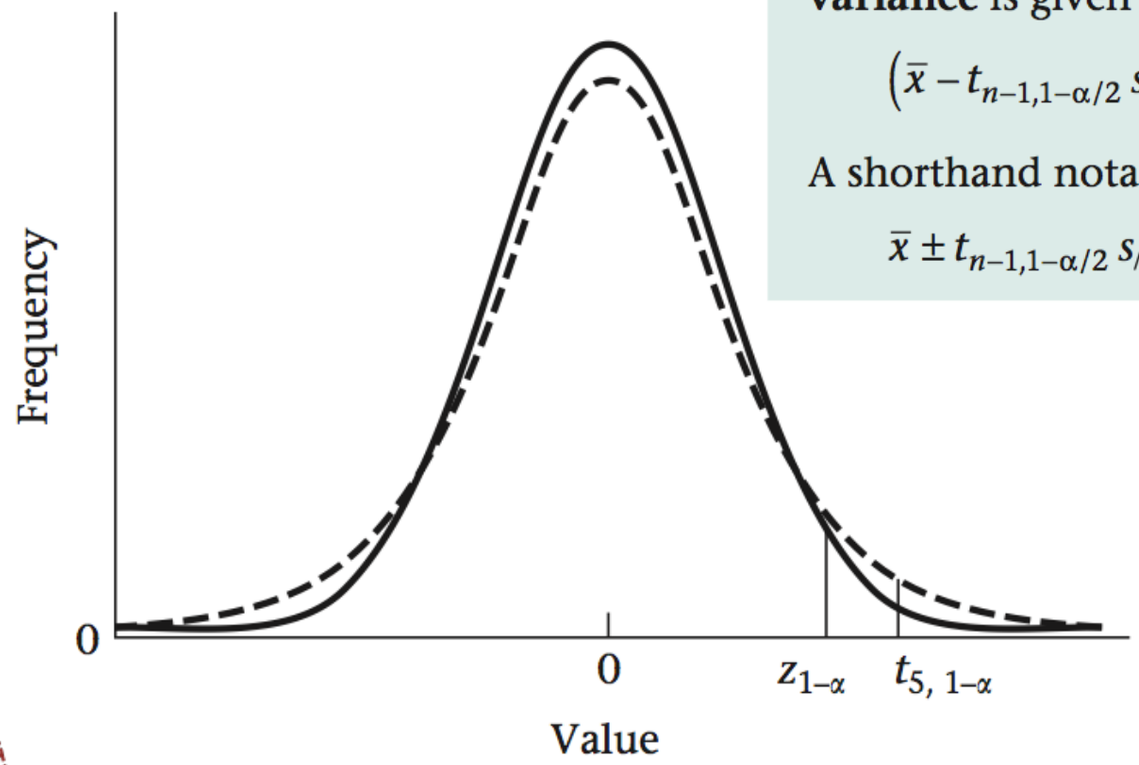
Confidence Interval for the Mean of a Normal Distribution

A $100\% \times (1 - \alpha)$ CI for the mean μ of a normal distribution with unknown **variance** is given by

$$\left(\bar{x} - t_{n-1,1-\alpha/2}\, s/\sqrt{n}\,, \bar{x} + t_{n-1,1-\alpha/2}\, s/\sqrt{n}\right)$$

A shorthand notation for the CI is

$$\bar{x} \pm t_{n-1,1-\alpha/2}\, s/\sqrt{n}$$



Associated with the sample size (n)

→ Degree of Freedom (n-1)

Degree of Freedom

$$\text{variance} = \frac{\text{sum of squares}}{\text{degrees of freedom}} = \frac{\sum (y - \bar{y})^2}{n - 1}$$

- **DOF: Number of observations that are free to vary:**
  - *Example*: A sample of five numbers and their average was **4**.

2	7			

2	7	4	0	→ what this number will be?

  - The key is the usage of **sample mean**. We cannot calculate the variance until we know the value of the sample mean.
  - The mean value is *a parameter estimated from the data*, so we lose one degree of freedom as a result.
  - So when calculating variance, we divide by the number (***n-1***).

In Small Sample Size

- What if the sample size is not that “large”?
  - **Student’s *t* - distribution** (1908)
  - An approximation of normal distribution.

$$P(-t_0 \le t \le t_0) = 1 - \alpha.$$

That is, what two values $-t_0$ and t_0 cut the t distribution such that a middle area of size $1 - \alpha$ will lie between $-t_0$ and t_0? See Figure 4.4.

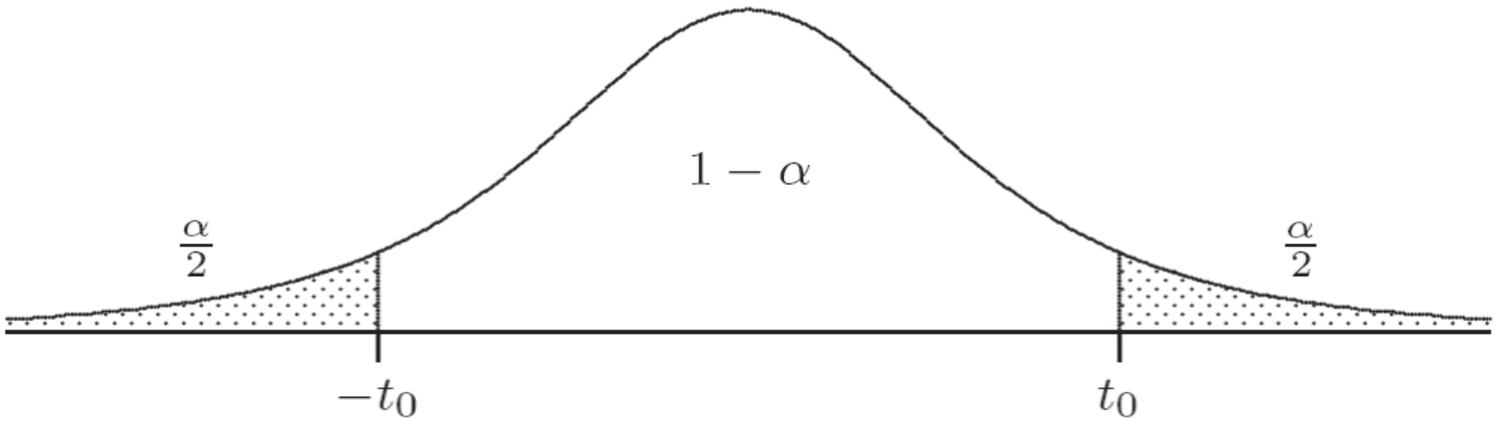
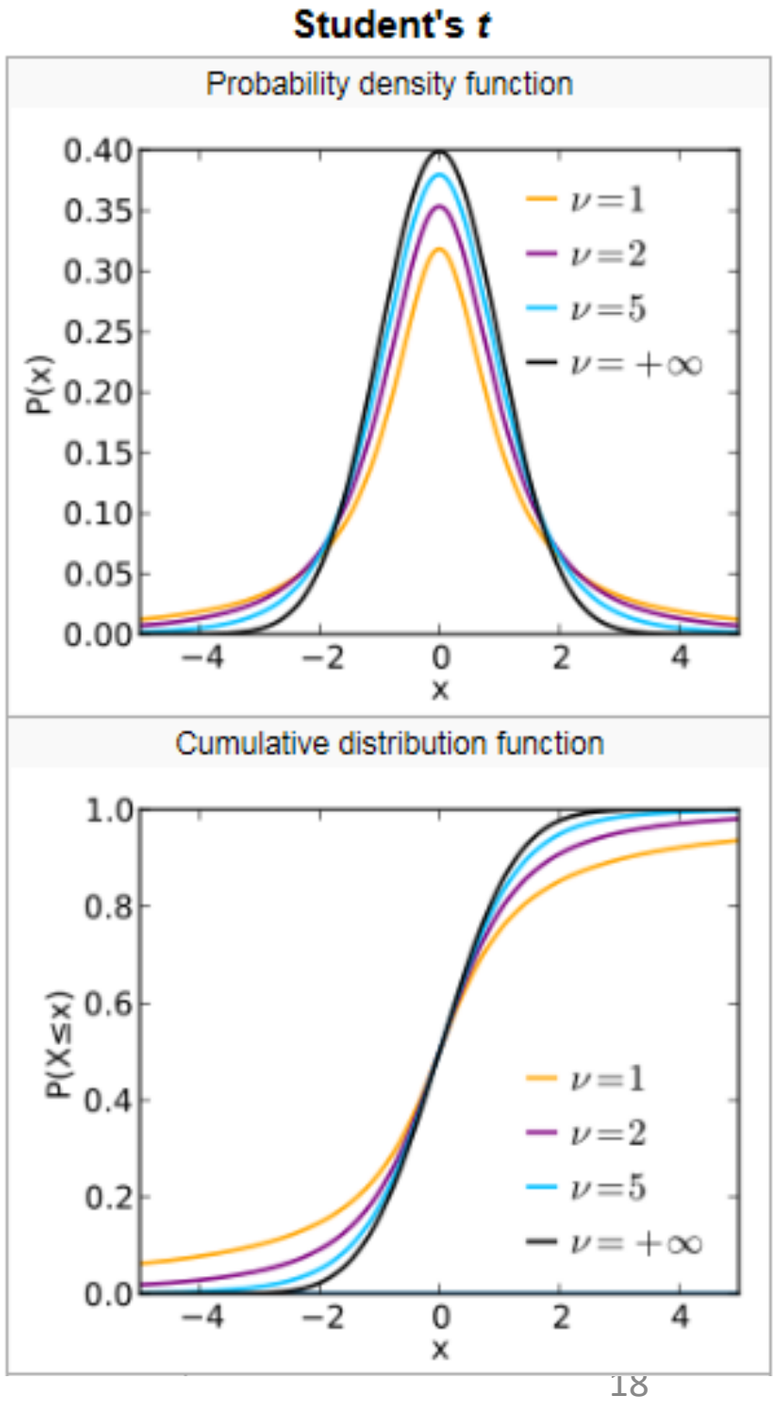
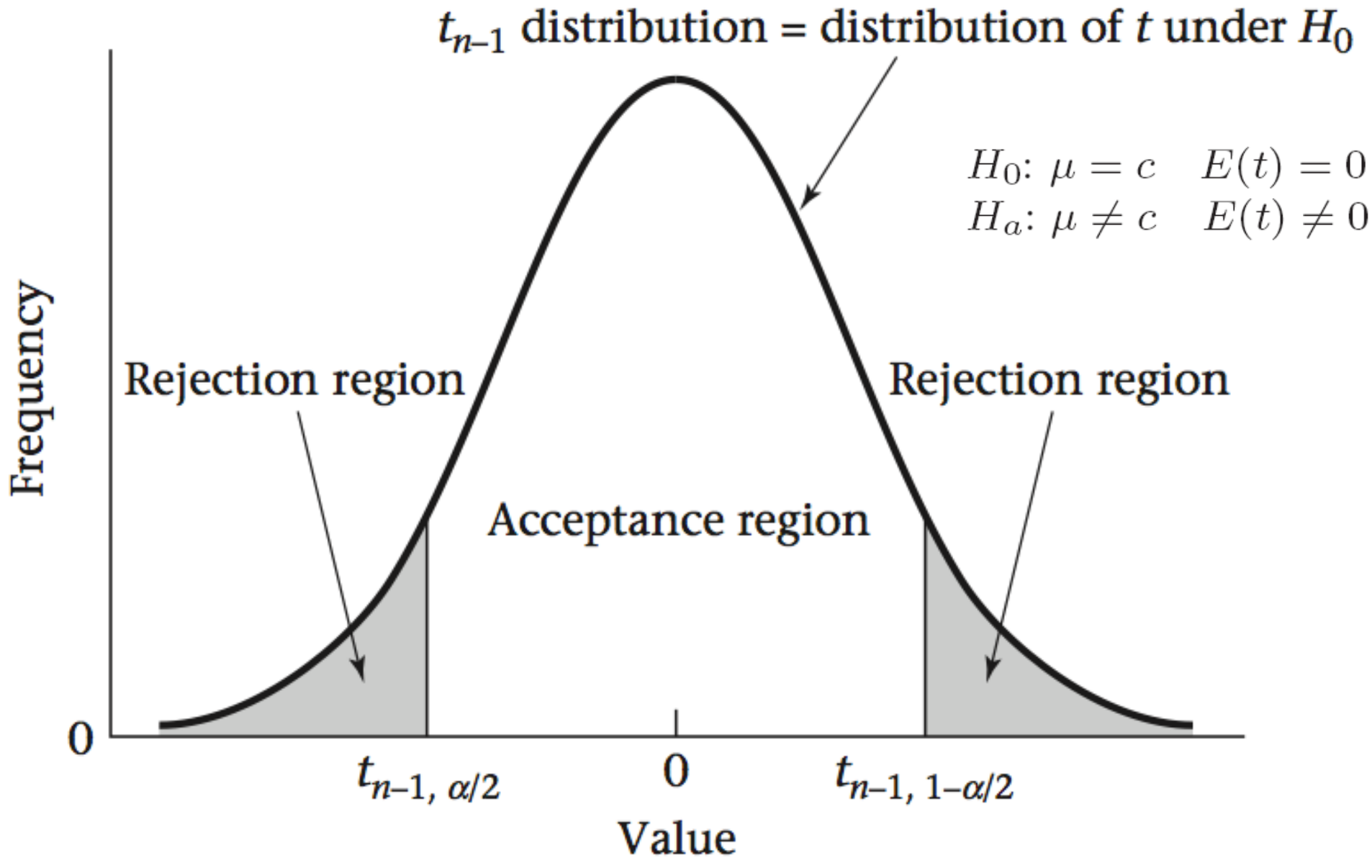


FIGURE 4.4. Locating the middle area of size $1 - \alpha$ in a t distribution.



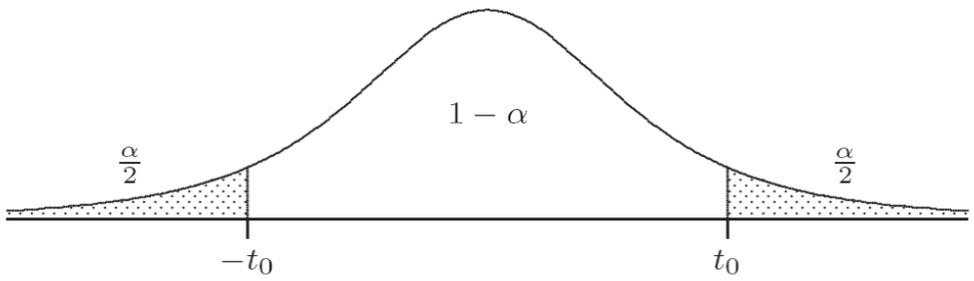
p-value: *5% CDF*

- **p-value** is the **cumulative probability** *(shaded area under the PDF curve)* beyond the decision threshold.



In Small Sample Size

$$\text{confidence interval} = t\text{-value} \times \text{standard error}$$



$$\text{CI}_{95\%} = t_{(\alpha=0.025,\text{d.f.}=9)}\sqrt{\frac{s^2}{n}}$$

qt(.025,9)

FORMULA 4.1. A $(1-\alpha)100\%$ confidence interval for the population mean μ is given by

$$C\left(\overline{X} - t_0\frac{s}{\sqrt{n}} \le \mu \le \overline{X} + t_0\frac{s}{\sqrt{n}}\right) = 1-\alpha,$$

where t_0 has $\nu = n-1$ degrees of freedom. Thus, the interval endpoints are

$$L_1 = \overline{X} - t_{(1-\frac{\alpha}{2},\nu)}\frac{s}{\sqrt{n}} \qquad \text{and} \qquad L_2 = \overline{X} + t_{(1-\frac{\alpha}{2},\nu)}\frac{s}{\sqrt{n}}.$$

Compared with Z distribution…

$$\text{lower boundary of confidence interval} = \bar{X} - (1.96 \times SE)$$
$$\text{upper boundary of confidence interval} = \bar{X} + (1.96 \times SE)$$

Key Functions

1. **GGPLOT2 library**
   - figure plots (***ggplot*** / ***geom_****)
   - figure separation (***facet_****)
   - QQ plot (***qplot***)

2. **Assumption Check**

pastecs
   - QQ plot (***qqnorm***)
   - Normality check (***stat.desc*** / ***Shapiro.test***)

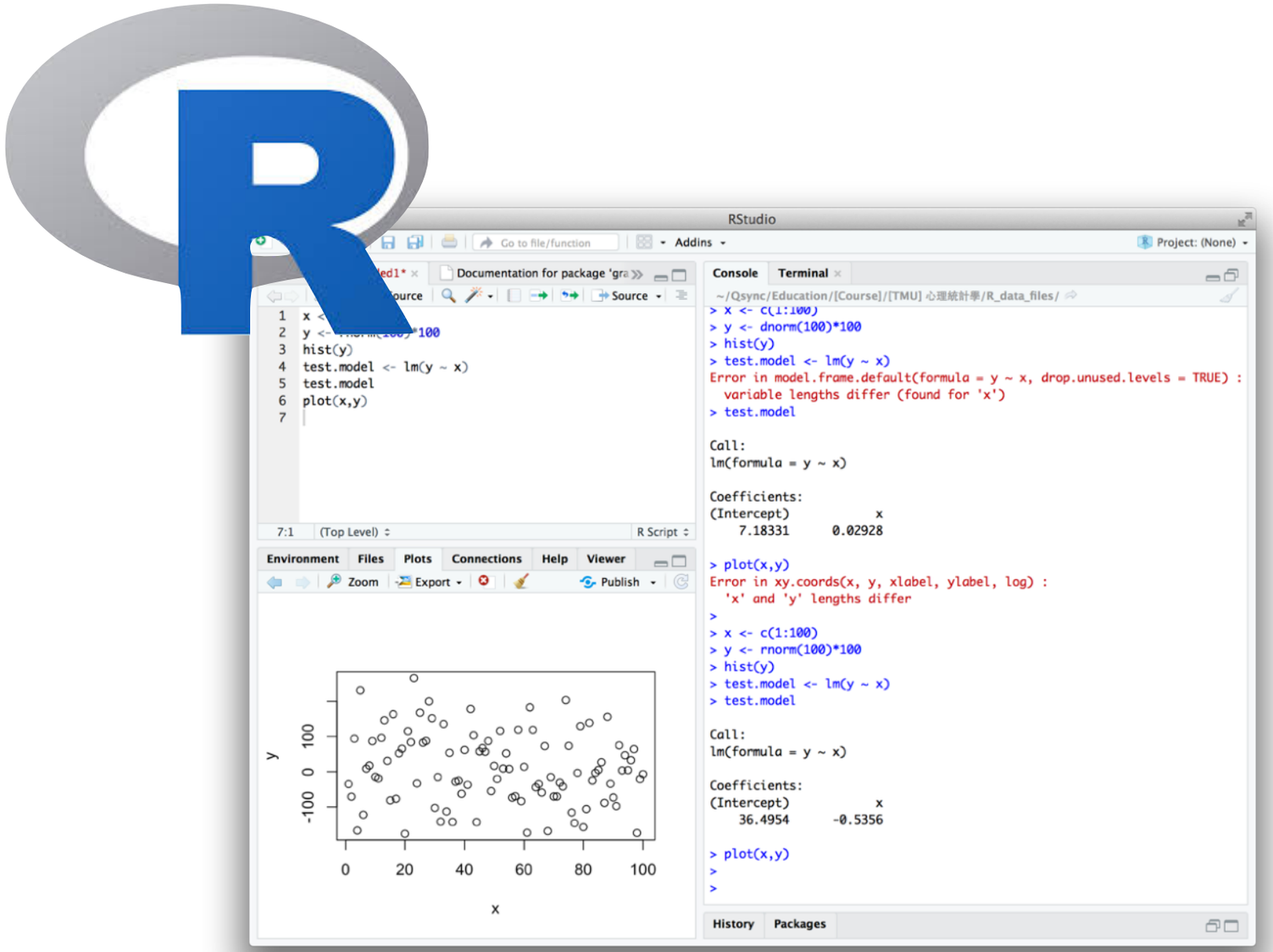
car
   - Homogeneity of Variance (***leveneTest***)

臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Check if you have these functions in R.

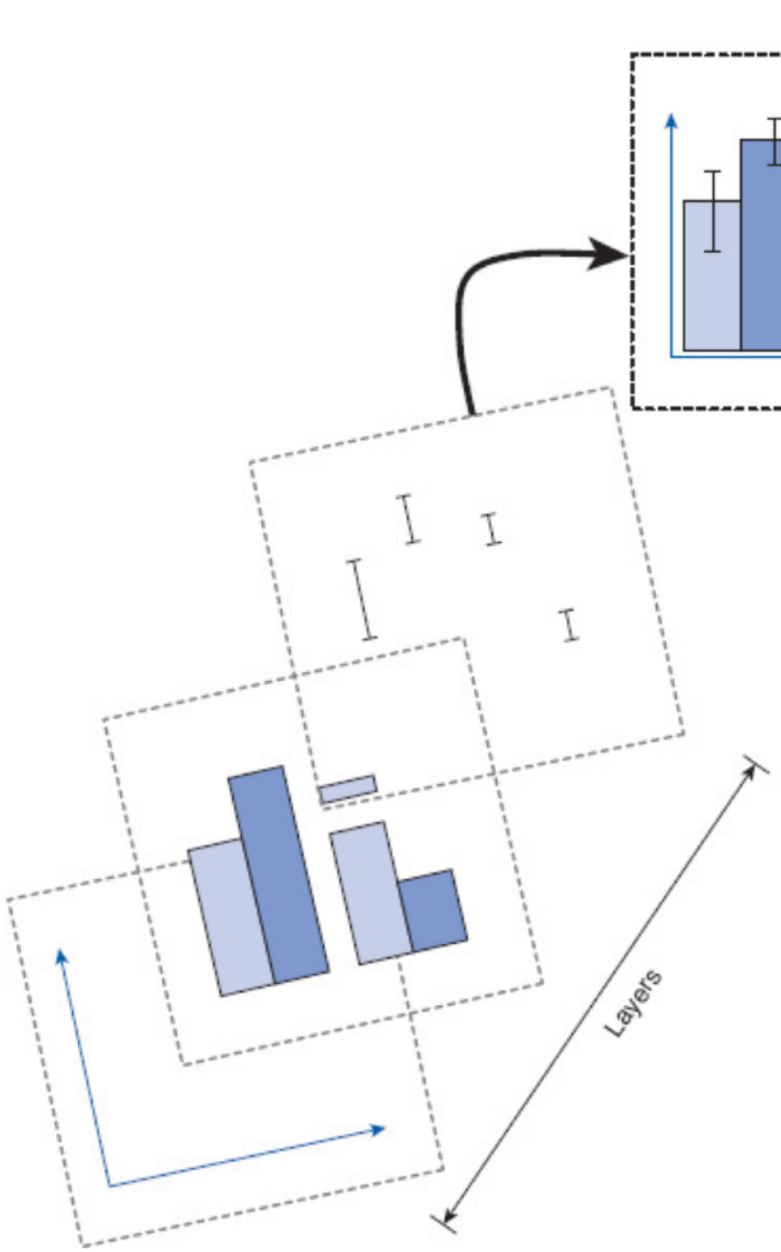
ggplot2

① GRAPHS USING GGPLOT2



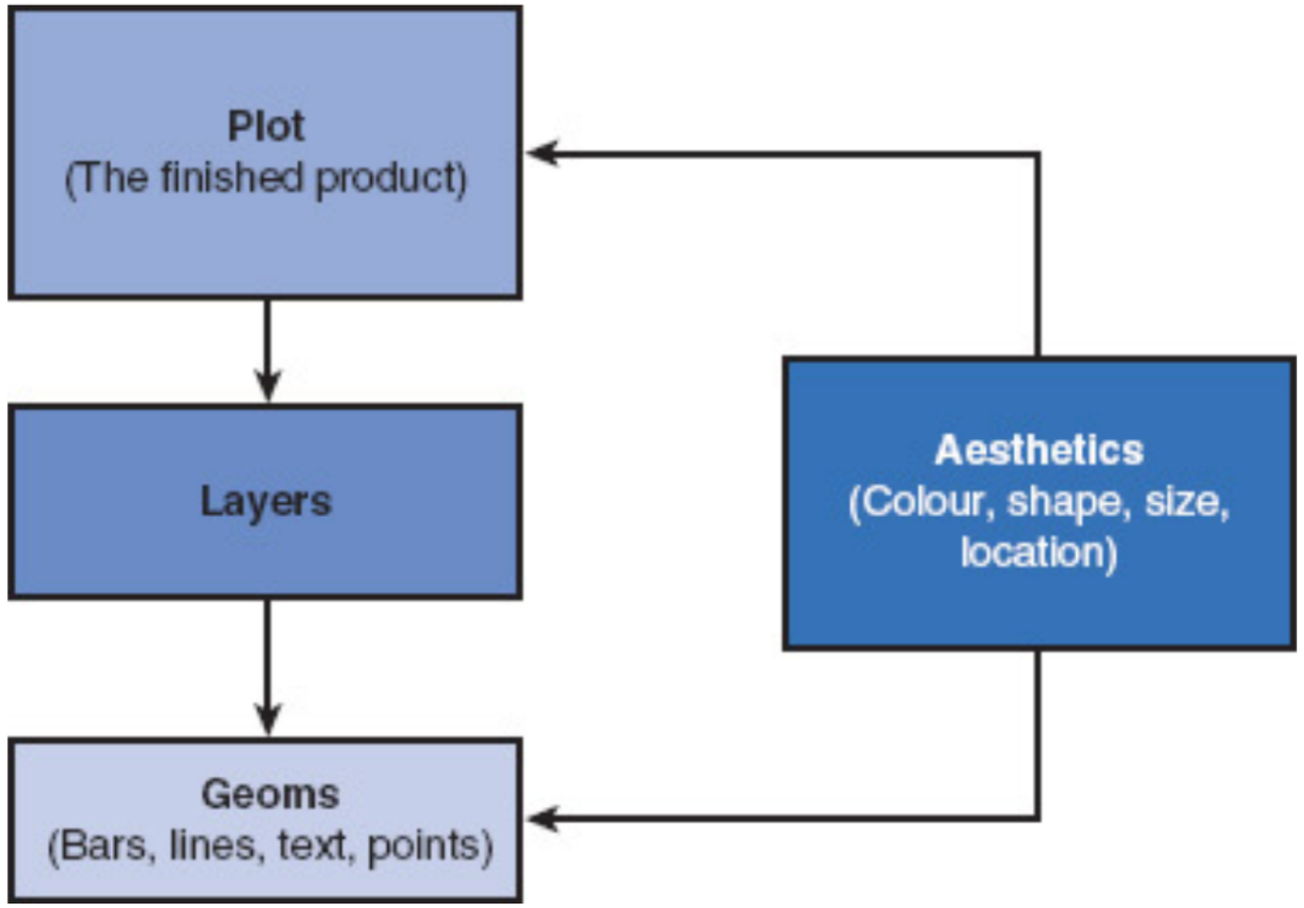
TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

R Graphs



Concept of Layers

(like photoshop)



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

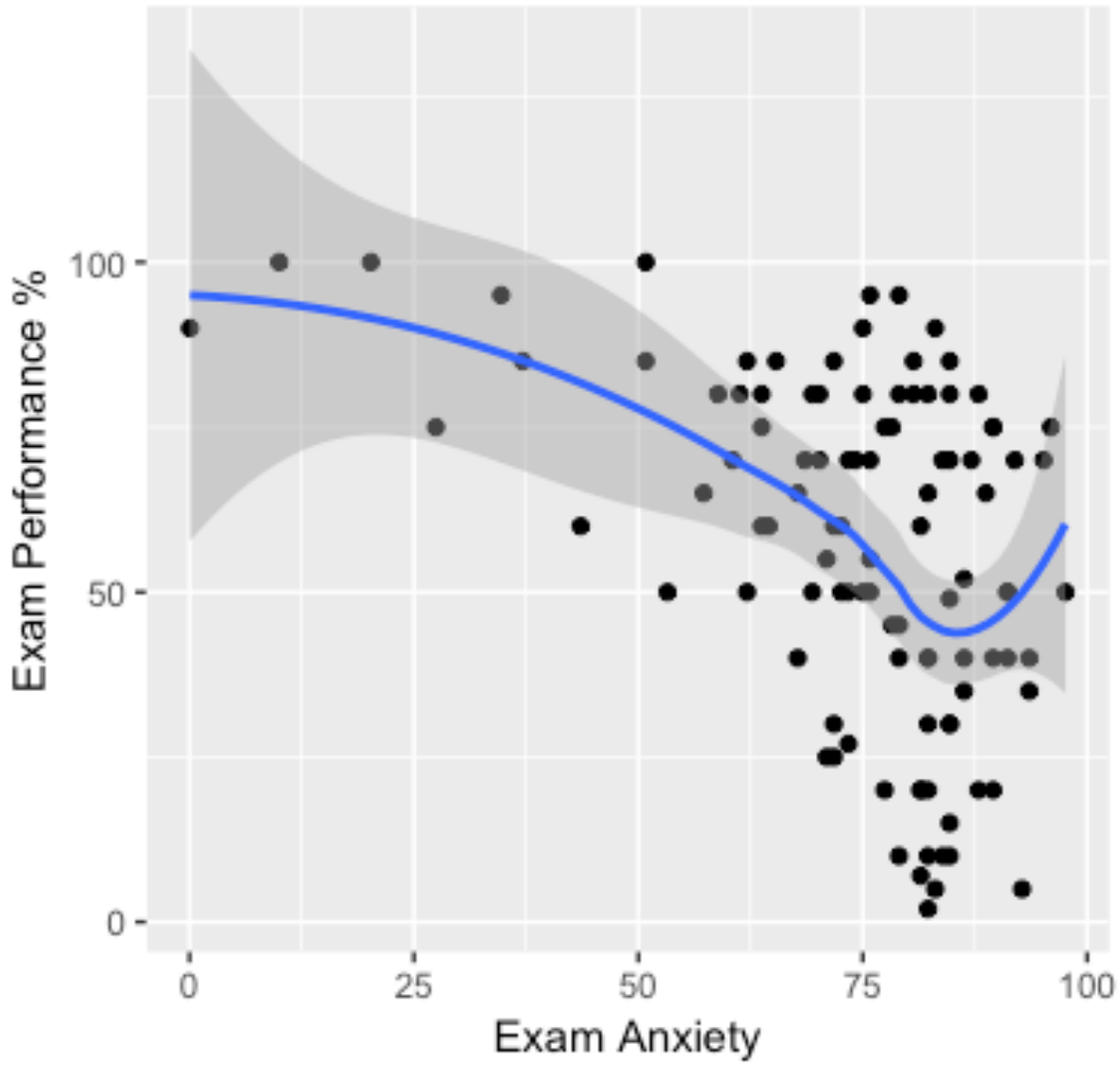
DEMO

R Graphs

- library(**ggplot2**)

- Scatter plot
- Histograms
- Boxplots
- Bar charts
- Line charts

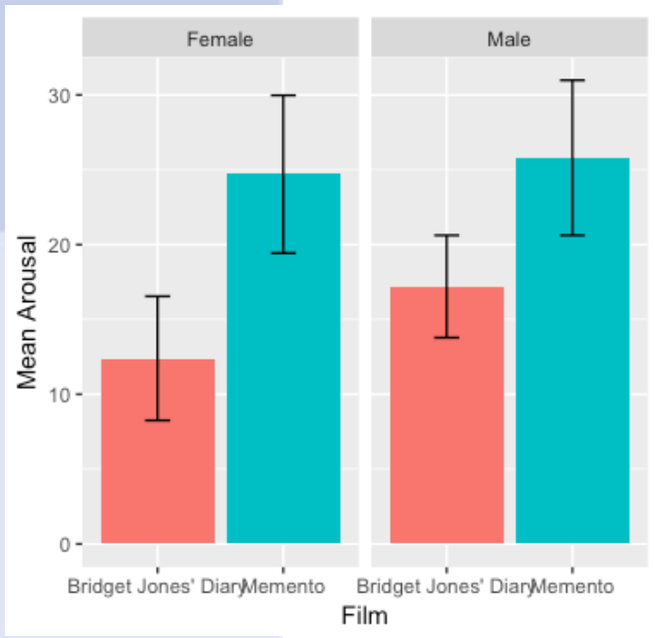
- For your practice
  - check “Rcode_03.R” file



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

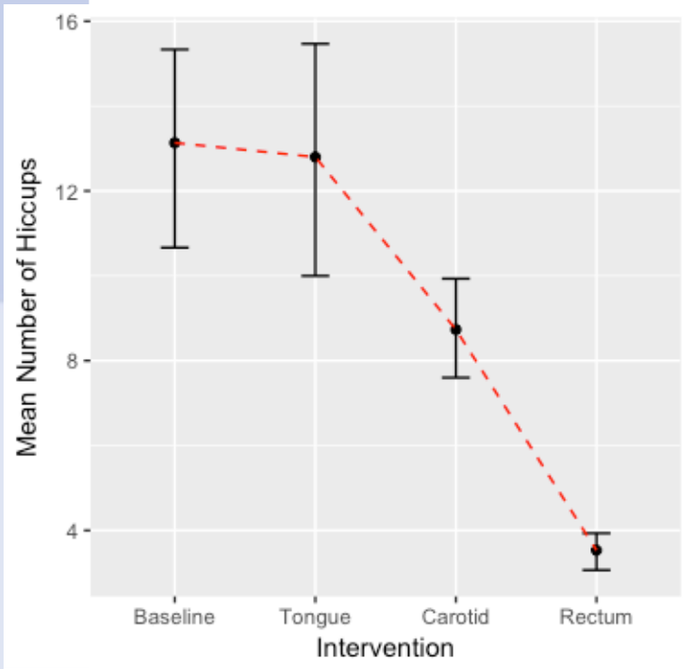
R Graphs

	Required	*Optional*
geom_bar()	x: the variable to plot on the *x*-axis	colour size fill linetype weight alpha
geom_point()	x: the variable to plot on the *x*-axis y: the variable to plot on the *y*-axis	shape colour size fill alpha
geom_line()	x: the variable to plot on the *x*-axis y: the variable to plot on the *y*-axis	colour size linetype alpha
geom_smooth()	x: the variable to plot on the *x*-axis y: the variable to plot on the *y*-axis	colour size fill linetype weight alpha



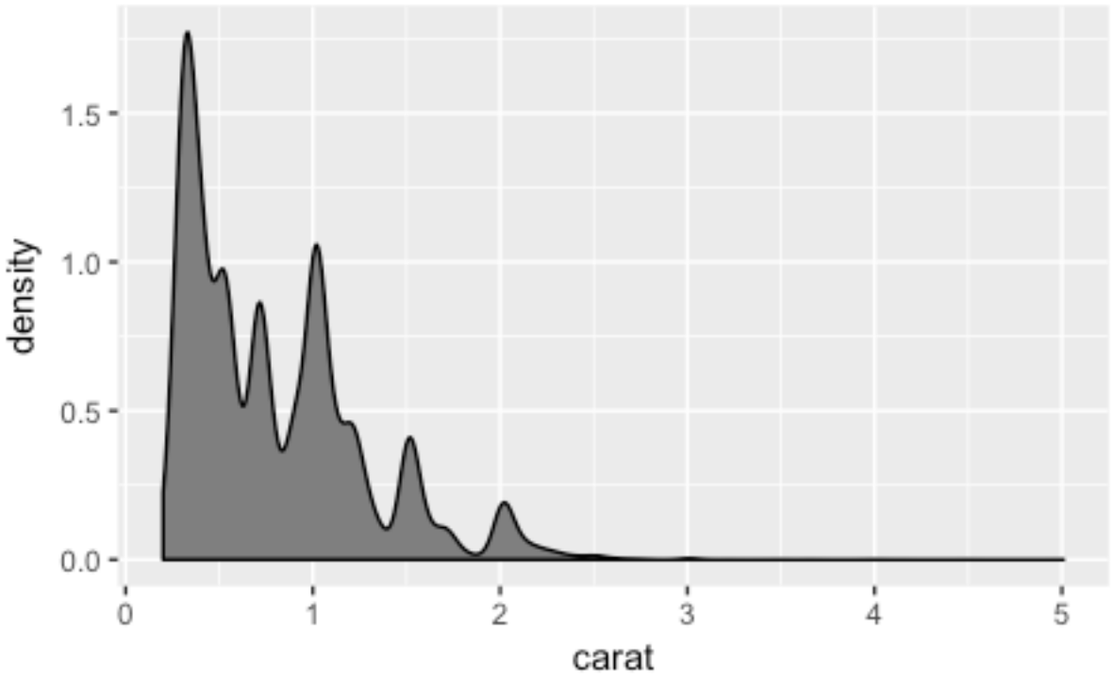
R Graphs

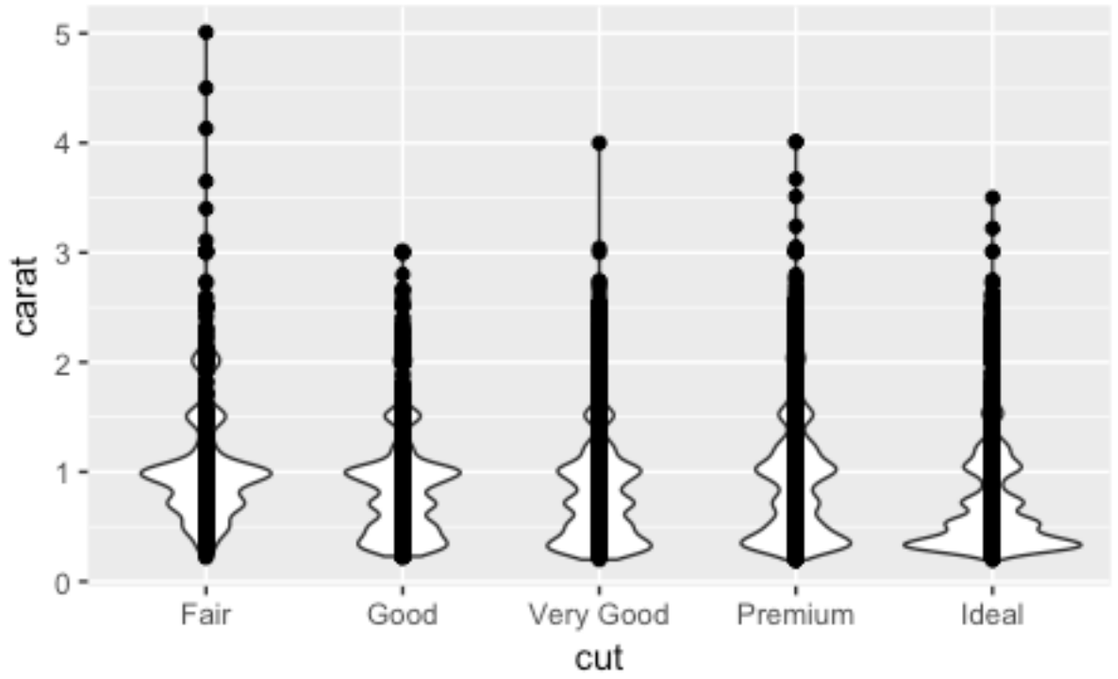
geom_histogram()	x: the variable to plot on the *x*-axis	colour size fill linetype weight alpha
geom_boxplot()	x: the variable to plot ymin: lower limit of 'whisker' ymax: upper limit of 'whisker' lower: lower limit of the 'box' upper: upper limit of the 'box' middle: the median	colour size fill weight alpha
geom_text()	x: the horizontal coordinate of where the text should be placed y: the vertical coordinate of where the text should be placed label: the text to be printed all of these can be single values or variables containing coordinates and labels for multiple items	colour size angle hjust (horizontal adjustment) vjust (vertical adjustment) alpha
geom_density()	x: the variable to plot on the *x*-axis y: the variable to plot on the *y*-axis	colour size fill linetype weight alpha

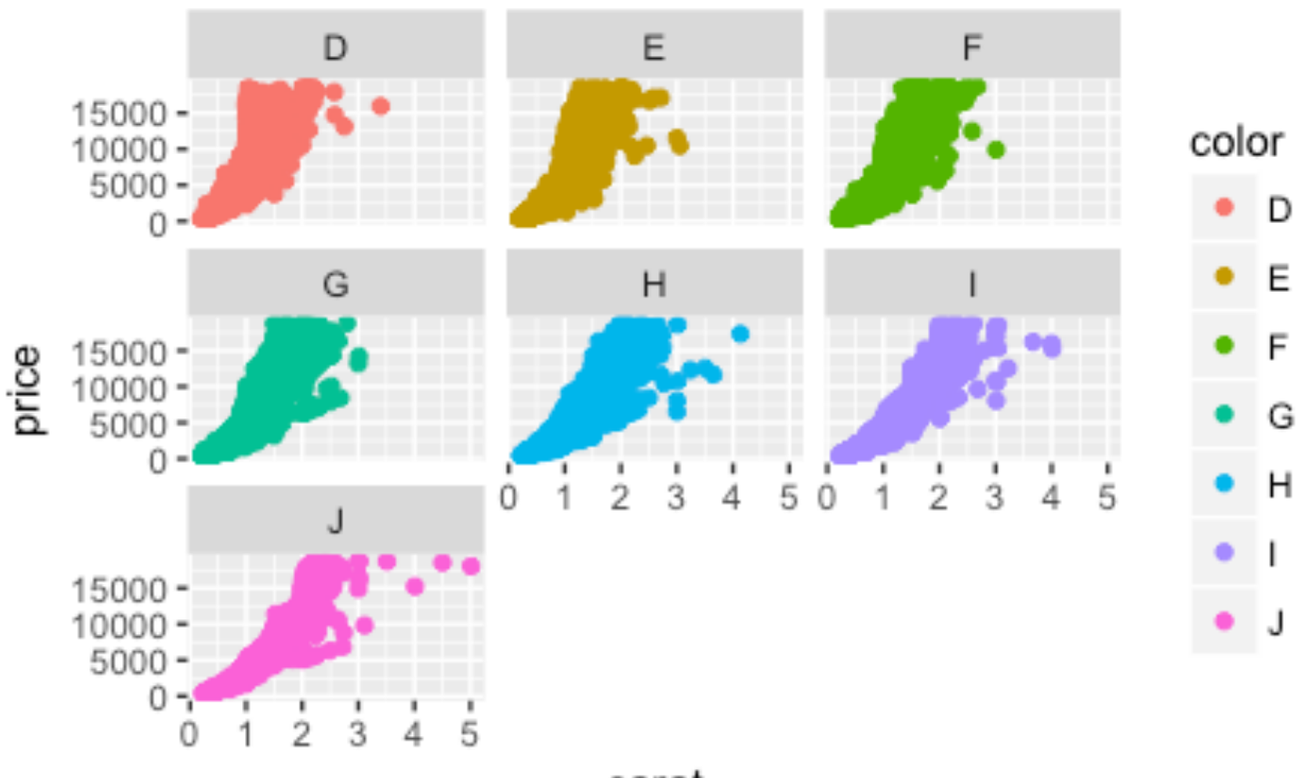


DEMO

GGPLOT2





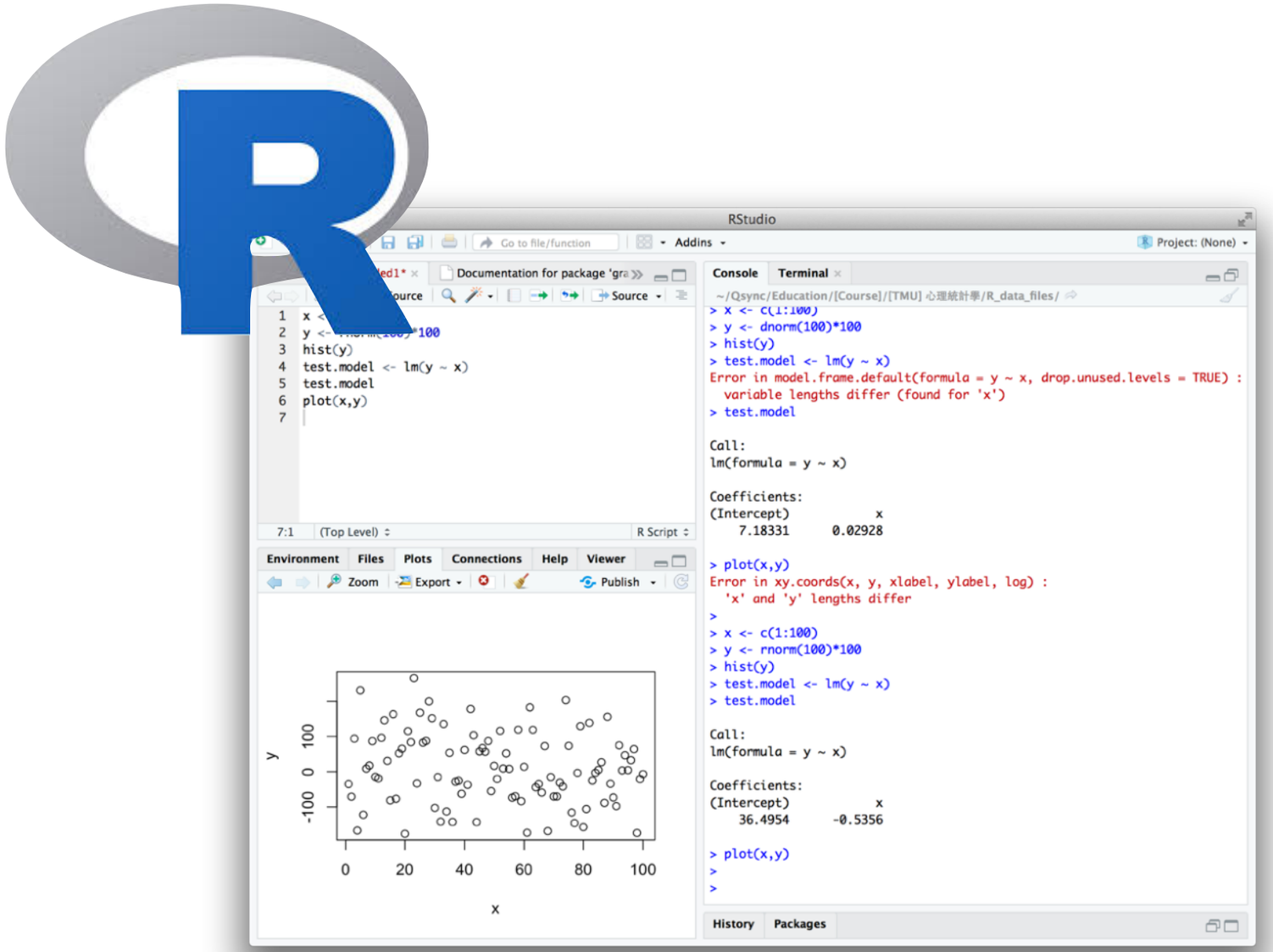


TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY
1960

shapiro.test

leveneTest

② ASSUMPTION CHECK



TAIPEI MEDICAL UNIVERSITY

臺北醫學大學

TAIPEI MEDICAL UNIVERSITY

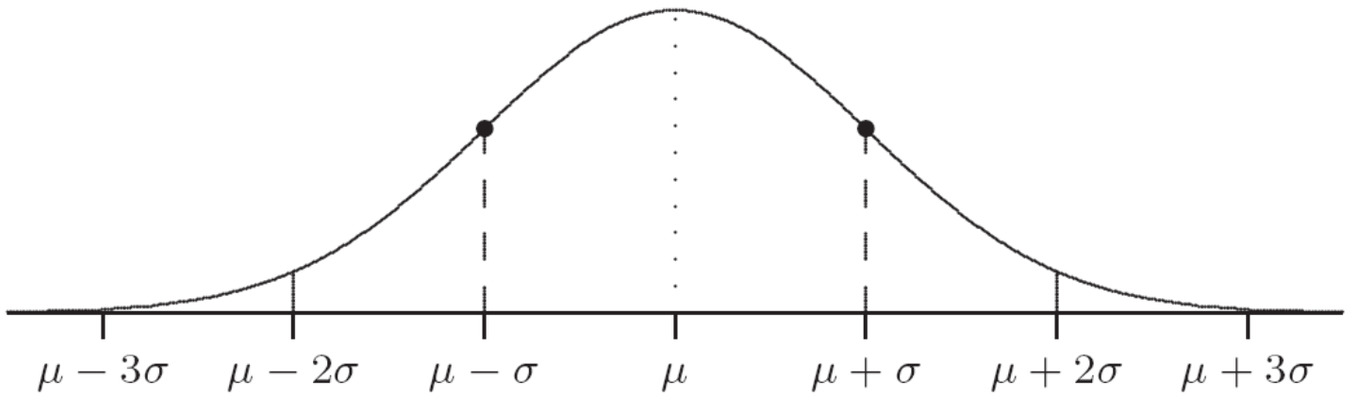
Check for Normal Distribution (Normality Assumption)

- Normal distribution is a general assumption for parametric testing (e.g., t-test, regression).
- If the **normality assumption** is **not met**, the testing statistics is unreliable.

DEFINITION 3.9. The **probability density function for a normal random variable** has the form

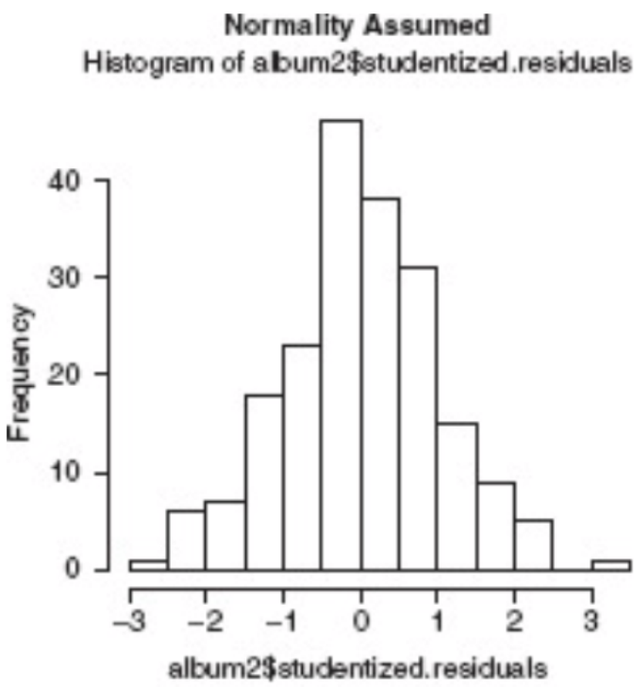
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

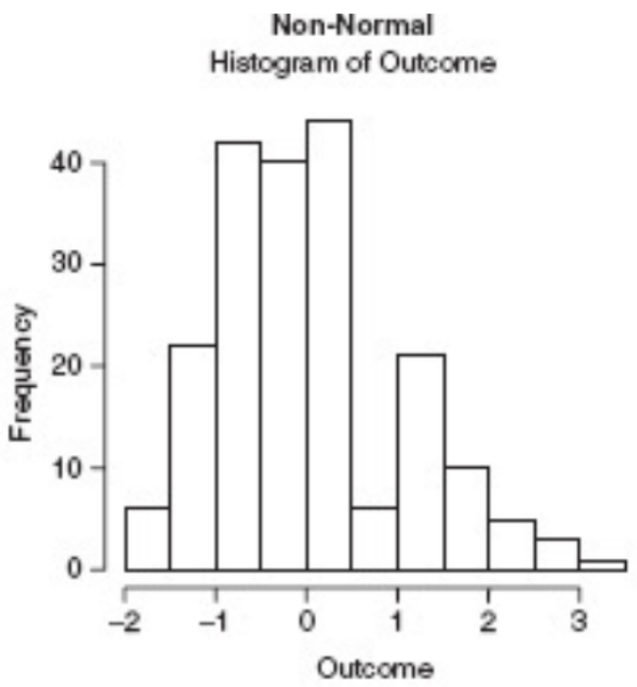
where σ is the standard deviation of the random variable and μ is its mean.

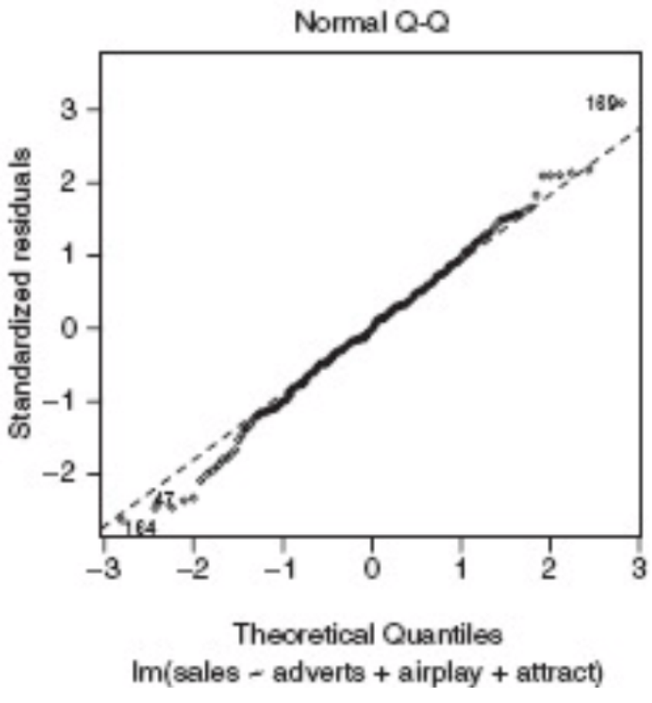


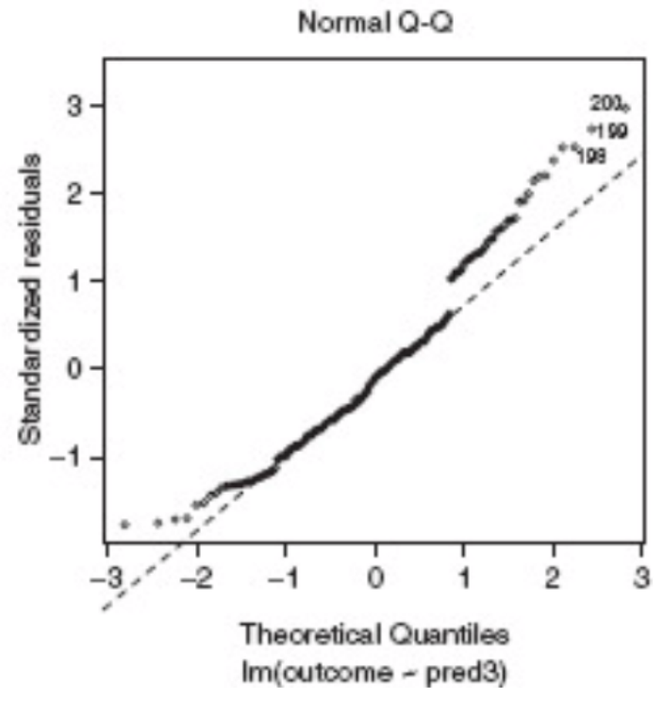
Visual Inspection: Q-Q Plot

- **Quantile-quantile plot**
  - A graph plotting the quantiles of a variable against the quantiles of a particular distribution (often *normal dist.*).
  - If values fall on the diagonal of the plot then the variable shares the same distribution as the one specified.
  - Deviations from diagonal show deviations from the dist. of interest.







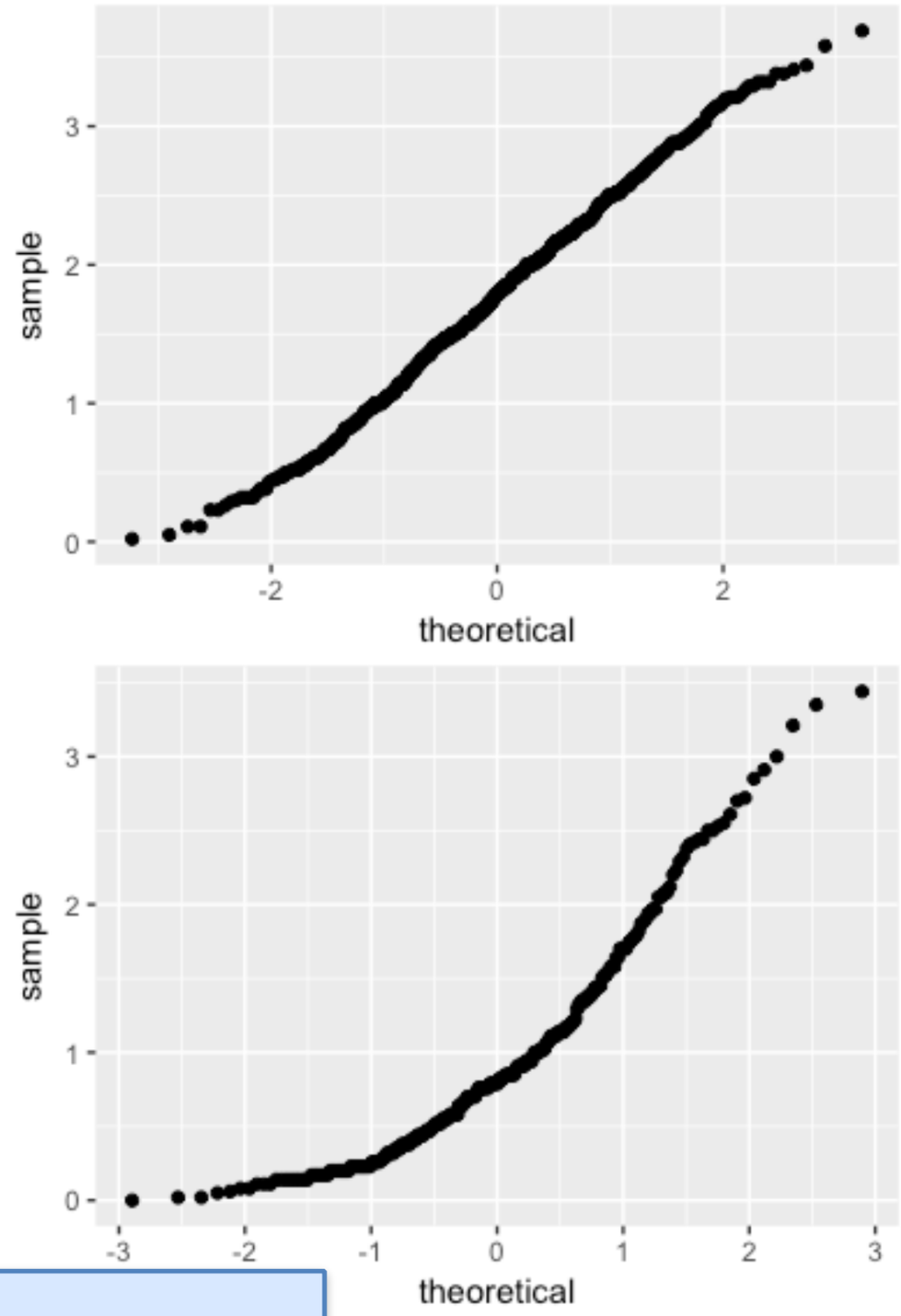


臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

DEMO

Normal Distribution

- Normal distribution is a general assumption for parametric testing (e.g., t-test, regression).



- **Q-Q plot** (quantile-quantile plot)
  - *rank sorted data and compared to expected values in normal distri.*
  - *qqnorm(dlf$day1)*
  - *qplot(sample=dlf$day1)*
- normal distri. = straight diagonal line

Load: **DownloadFestival.dat**

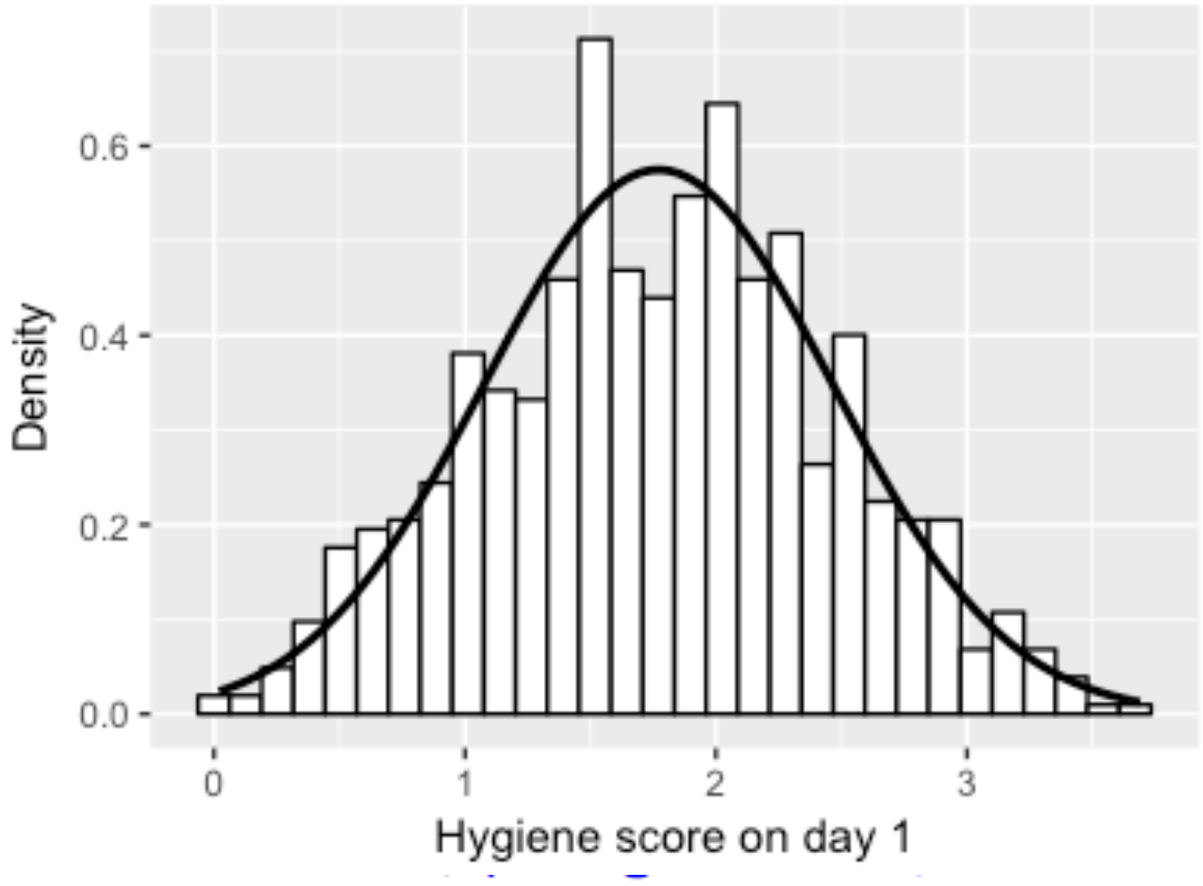
DEMO

Normal Distribution

- Normal distribution is a general assumption for parametric testing (e.g., t-test, regression).
- Check descriptive statistics from the data first

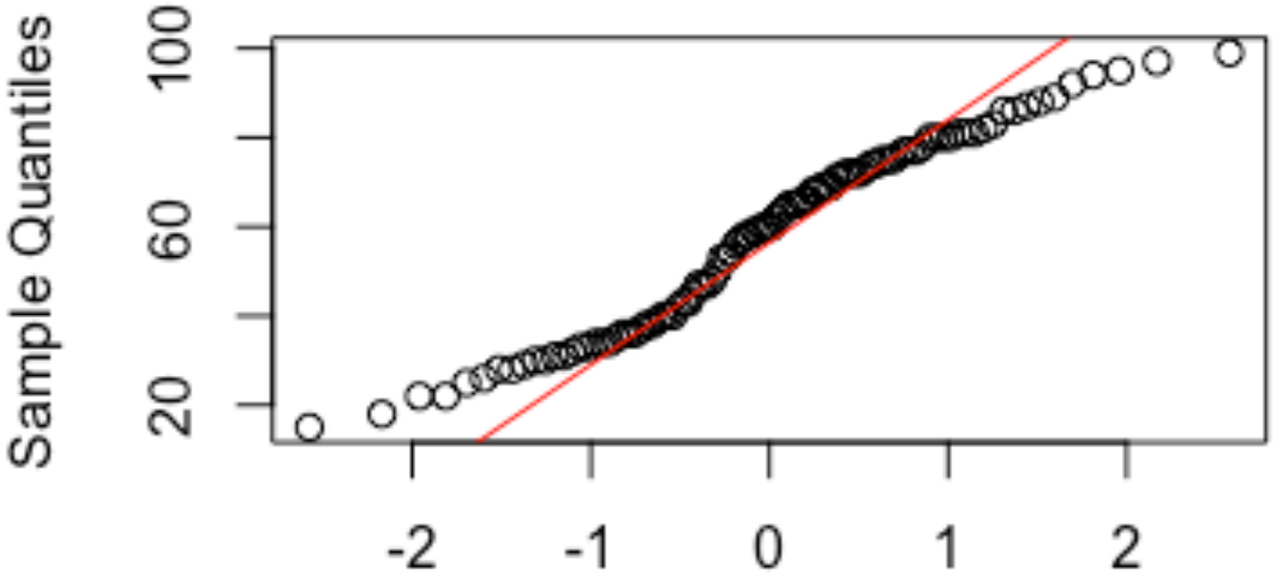
Load**: DownloadFestival.dat**

- **Plot and calculate**
  - *hist(dlf$day1)*
- reference packages
  - *library(pastecs)*
  - *stat.desc(dlf$day1, basic = FALSE, norm = TRUE)*



	day1	day2	day3
median	1.790	0.790	0.760
mean	1.771	0.961	0.977
SE.mean	0.024	0.044	0.064

Shapiro-Wilk Normality Test



Load:
rexam.dat

- H_0: sample is normally distributed.
- H_a: sample is not normally distributed.

- **QQ plot**
  - *qqnorm(rexam$exam);*
  - *qqline(rexam$exam,col='red')*

- **Shapiro-Wilk Normality Test**
  - *shapiro.test(rexam$exam)*

```
        Shapiro-Wilk normality test

data:  rexam$exam
W = 0.96131, p-value = 0.004991
```

reference to p<0.05 in general

- **p-value < 5% of probability**
- **Reject H_0, the distribution deviates from normality.**

TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Homogeneity of Variance

- Equal variance across groups is another general assumption for tests.
- **Levene's test** (Levene, 1960)
  - Test the **H_0** that variances in different groups are equal (diff=0).
  - Compare the variance between universities, what is conclusion?

Load: **rexam.dat**

- **Usages in R**
  - *library(car)*
  - *leveneTest(variable, group factor, center=median/mean))*

```
Levene's Test for Homogeneity
      Df F value Pr(>F)
group  1  2.0886 0.1516
      98
```

reference to p<0.05 in general

- **p-value > 0.05 of probability**
- **Accept H_0, the variances were similar between universities.**

Discussion

1. **Mean & Unreliability**
   - t distribution & Degree of freedom
2. **Variance**
   - χ^2 distribution & Confidence interval
3. **[R] GGplot & Assumption check**
   - Normality & Homogeneity of Variance

https://en.wikipedia.org/wiki/Levene%27s_test

Homogeneity of Variance

- **Test for equality of variance (*2 samples*):**
  - F-ratio $\mathrm{H}_0: \sigma_1^2 = \sigma_2^2 \; vs \; \mathrm{H}_1: \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2}$$

  - Usages in R

stats

- *var.test(variable1, variable2, alternative=c(two.sided, less, greater))*

- **Test for equality of variance (*multiple samples*):**
  - Levene's test (Levene, 1960): 1-way ANOVA testing whether the 'average' absolute deviation from the median is the same across groups.

$$\mathrm{H}_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2}$$

  - Usages in R

car

- *leveneTest(variable or model, data frame, center=**median**/mean))*

- k is the number of different groups to which the sampled cases belong,
- N_i is the number of cases in the ith group,
- N is the total number of cases in all groups,
- Y_{ij} is the value of the measured variable for the jth case from the ith group,
- $Z_{ij} = \begin{cases} |Y_{ij} - \bar{Y}_{i\cdot}|, & \bar{Y}_{i\cdot} \text{ is a mean of the } i\text{-th group,} \\ |Y_{ij} - \tilde{Y}_{i\cdot}|, & \tilde{Y}_{i\cdot} \text{ is a median of the } i\text{-th group.} \end{cases}$

Testing After Stepping into Probability

	Hypothesis Testing	Icons
Step 1	• Restate the question as a research hypothesis (H_1) and a **null hypothesis** (H_0) to answer the research question.	① Hypothesis
Step 2	• Determine the characteristic of the distribution and check **model assumption**	② Assumption
Step 3	• Compare the sample position and the critical value (threshold) on the general distribution (**P-value**)	③ Testing
Step 4	• Report the **effect size** of the samples	④ Effect Size
Step 5	• Make the **final decision** • Reject H_0 or accept H_1	⑤ Decision

TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY