Statistical Method
Hypothesis Testing

I-Chen Lee, STAT, NCKU

Sept 26, 2023

Frame Title

- Sections 7 of Maurits; Van Den Heuvel Kaptein (Edwin). (2022). Statistics for Data Scientist: An Introduction to Probability, Statistics, and Data Analysis.
- Chapter 10 of Akinkunmi, M. (2019). *Introduction to statistics using R. Synthesis Lectures on Mathematics and Statistics*, 11(4), 1-235.

Overview

Univariate Data

- Univariate data is used to describe one characteristic or variation from observation to observation.
- To describe patterns, some ways including graphical methods, measures of central tendency, and measures of variability could be conducts.
  - Graphical methods: bar chart, histogram, pie chart,...
  - Location: mean/median/mode
  - Dispersion: variance/standard deviation/range,...
- The whole picture of univariate dataset could be addressed by a distribution.

Textbook: Akinkunmi, M. (2019). *Introduction to statistics using R. Synthesis Lectures on Mathematics and Statistics*, 11(4), 1-235.

Thinking univariate data together with probability distributions helps good guess of your data

	Univariate data	Probability
Distribution	Bar chart/histogram	pmf/pdf
Distribution	empirical cdf	cdf
Measure of location	sample mean/median	expectation/median
Measure of dispersion	sample variance	variance

Review of Homework 2

- Give the expressions of given distributions.
- Match data to the given distributions (Multiple selection).

Examples are as the following pages:

Question 1: Normal distribution

If a random variable X follows the normal distribution (μ, σ^2), then

- The parameters are μ and σ^2.
- The pdf is

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

- The cdf is

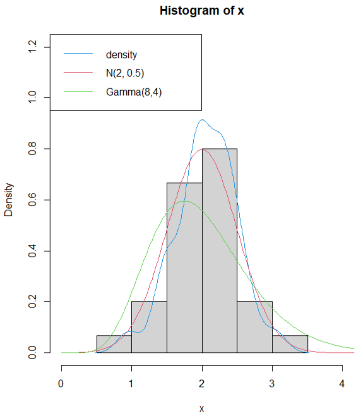
$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(s-\mu)^2/(2\sigma^2)} ds, \quad -\infty < x < \infty.$$

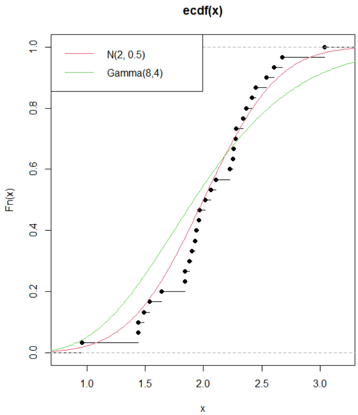
- The expectation is $E(X) = \mu$.
- The variance is $Var(X) = \sigma^2$.

Question 2: Given a dataset and distributions

- Empirical Distributions with given distributions

1. Normal($\mu = 2,\ \sigma = 0.5$)
2. Gamma(shape = 8, rate = 4)





Statistical hypothesis testing

Purpose:
Use a statistical way to examine a conceptual guess or verification.
Keywords:

- Null hypothesis (H_0)
- Alternative hypothesis (H_A or H_1)
- Type-I error(Pro$\{$reject H_0 is True$|$under H_0 is True$\}$) / Type-II error
- Rejection region
- p-value

	H_0 is rejected	Fail to reject H_0
H_0 is True	Type-I error (probability is α) (level of significance)	Correct decision
H_0 is not True	Correct decision (power = 1-β)	Type-II error (probability is β)

Relationship between samples and population

Fig. 7.4 Types of errors in hypothesis testing

Sample (decision)	Population: H_0 is true	Population: H_0 is not true
Do not reject H_0	No Error	Type 2 Error β
Reject H_0	Type 1 Error α	No Error

Null and alternative hypotheses

A hypothesis is a statement provided the result of a research study, which can be used to describe the population parameter or comparison.

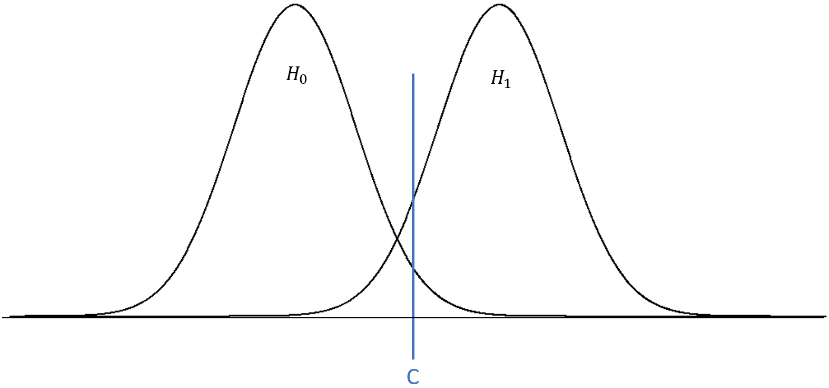
- The hypothesis that in favor of the assumption is called the "null hypothesis (H_0)". (Very important!!!)
- Usually, the hypothesis (H_0) under investigation we are trying to disprove (reject).
- The hypothesis that the null hypothesis fails is called "alternative hypothesis (H_1)". (Sometimes is important.)

For example:
After you flipped a coin 100 times, claim that the coin is fair.

$$H_0 :$$

$$H_1 :$$

Null and alternative hypotheses



Type-I error and Rejection region

- A type-I error occurs when rejecting the null hypothesis when null hypothesis is true. This error is also known as a false positive.

$$p = \text{Pro}\{\text{reject } H_0 \text{ is True} | \text{under } H_0 \text{ is True}\}.$$

A significance level (α) of 0.05 indicates a 5% risk of concluding that H_0 is not true when H_0 is true. Usually, we set $\alpha = p = 0.05$.

- Under H_0, we can set (or evaluate) a critical value c to ensure the type-I error is at the significant level (α).

$$R_T = \{\mathbf{X} : T(\mathbf{X}) > c\},$$

where $T(\mathbf{X})$ is the test statistic and $\mathbf{X}$ denotes the information from the random samples.

Illustration of rejection regions

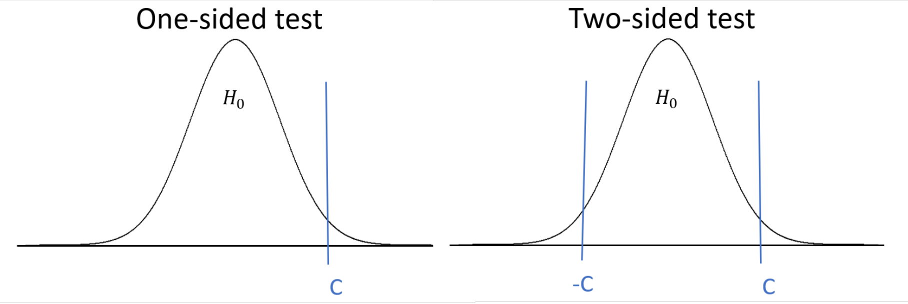


Illustration of rejection regions

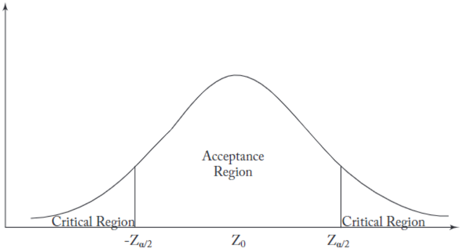


Figure 10.1: Two-tail normal test.

```
> ### z-score ###
> qnorm(0.025, mean = 0, sd = 1, lower.tail = TRUE)
[1] -1.959964
> qnorm(0.025, mean = 0, sd = 1, lower.tail = FALSE)
[1] 1.959964
> ### p-value ###
> pnorm(-1.066, mean = 0, sd = 1, lower.tail = TRUE)
[1] 0.1432118
> pnorm(1.066, mean = 0, sd = 1, lower.tail = FALSE)
[1] 0.1432118
```

Rejection region and p-value

- Under H_0, we can set (or evaluate) a critical value c to ensure the type-I error is at the significant level (α).

$$R_T = \{X : T(\mathbf{X}) > c\},$$

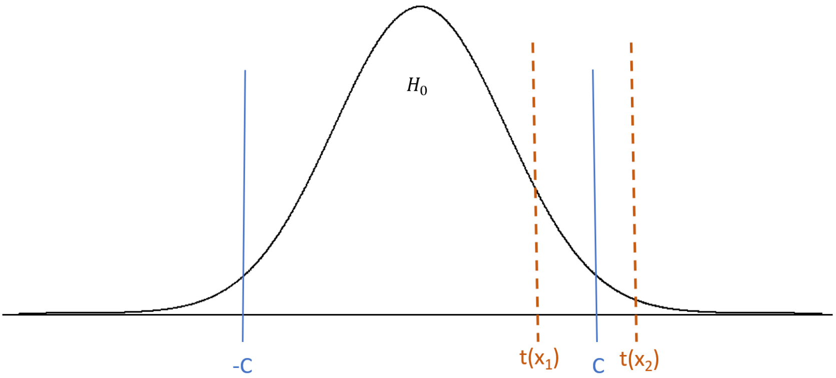
where $T(\mathbf{X})$ is the test statistic and $\mathbf{X}$ denotes the information from the random variable.

- $\alpha = Pro\{\mathbf{X} \in R_T\} = Pro\{T(\mathbf{X}) > c\}$.
- We can collect a set of samples $\mathbf{x}$, and substitute the information to the test statistic, denoted by $t(\mathbf{x}) = T(\mathbf{x})$. Then, the p-value is

$$p - \text{value} = Pro\{T(\mathbf{X}) > t(\mathbf{x}) | \text{ under } \ H_0)\}.$$

- The null hypothesis H_0 is rejected if the p-value is less than or equal to a predefined α.

Rejection region and *p*-value



Procedure for hypothesis testing

An unknown research hypothesis needs to be examined and explored.

1. State the null and alternative hypotheses, and specify the level of significance.
2. Choose an appropriate testing. (Usually we call the testing is the statistical method.)
3. Evaluate or develop the **test statistic**.
4. Under H_0, compare the test statistic to a critical value and decision rule from the corresponding distribution.
5. Show the rejection region or evaluate the p-value.
6. Draw the appropriate conclusions.

Common examples of hypothesis testing

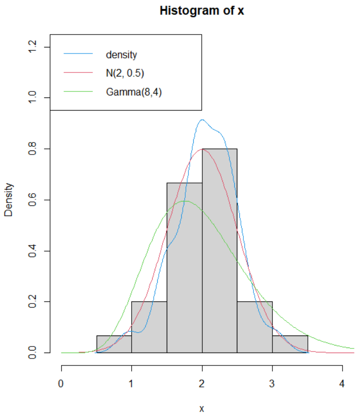
1. Let $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ be random samples from a distribution.
   - Test if $\mathbf{X}$ follows the given distribution F.
   - Test if the population mean is μ_0.
2. Let $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ be random samples from two distributions.
   - Test if the distributions of $\mathbf{X}$ and $\mathbf{Y}$ are the same.
   - Test if the population means of $\mathbf{X}$ and $\mathbf{Y}$ are the same.

Testing for distribution

- Empirical Distributions with given distributions

1 Normal($\mu = 2,\ \sigma = 0.5$)

2 Gamma(shape $= 8$, rate $= 4$)



Testing for distributions (Goodness of fit)

To check model assumption, there are following situations:

- two theoretical CDFs,
- two empirical CDFs and
- an empirical CDF to a theoretical CDF.

$$H_0 : \text{the samples are from the distribution } F_0$$

$$H_1 : \text{the samples are not from the distribution } F_0$$

Methods:

- Kolmogorov–Smirnov test (most commonly-uesd)
- Anderson–Darling test
- Cramer-Von Mises Test

Testing for distributions (Goodness of fit)

Let the empirical CDF (ecdf) define as

$$\hat{F}_n(x) = \frac{\text{number of samples } \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[-\infty,x]}(X_i),$$

where $\mathbf{1}$ is an indicator function.

- Kolmogorov–Smirnov test: the test statistic is

$$D_n = \sup_x |F_n(x) - F_0(x)|.$$

- Anderson–Darling test: the test statistic is

$$A^2 = \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x).$$

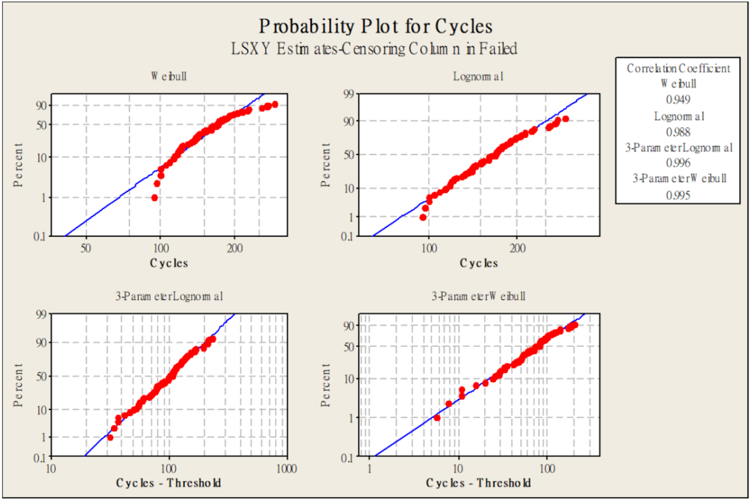
- Cramer-Von Mises Test

$$\omega^2 = \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dF(x).$$

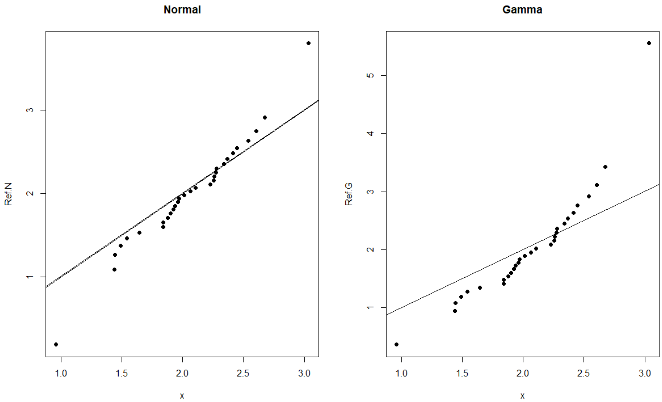
Testing for distributions (Graphical method)

- To check model assumption, the graphical goodness of fit is related to P-P plot or quantile-to-quantile (Q-Q) plot.
- Idea: plotting quantiles of collected data with the quantiles of a theoretical distribution.
- If the distribution fits better, the dots form a straight line (identy line $y = x$).
- Perform together with the corresponding p-value of a test.

Testing for distributions (Graphical method)



Question 2 (conti)




```
> round(dist, 4)
          normal  gamma
KS.test   0.3659 0.0423
AD.test   0.5899 0.0686
CVM.test  0.4817 0.0659
```

An example of hypothesis testing

In a pharmaceutical company, the operations manager claimed that the mean of the drugs produced by the company is 100 mg. If a random sample of 60 drugs is with mean of 98 mg. Assuming that the standard deviation of 14 mg is known, test the hypothesis to justify the operations manager's claim with the level of significant (α) 0.05.

- Descriptive statistics:
- Population:
- Sample:
- Null hypothesis (H_0):
- Alternative hypothesis (H_1):
- Method and test statistic: One sample z-test
    1. Test statistic.
    2. Rejection region.
    3. Conclusion.

An example of hypothesis testing (conti.)

Method and test statistic: One sample z-test

1. Test statistic:
$$z_0 = \frac{(98 - 100)}{14/\sqrt{60}} = -1.1066.$$
2. Under H_0, the test statistic (Z) follows a standard normal distribution $N \sim (0, 1)$.
3. Acceptance region (by theorem): $-1.96 < z_0 < 1.96$
4. Rejection region (by theorem): $|z_0| > 1.96$.
5. p-value (by your data): $p = \text{Pro}\{|Z| > z_0\} = 0.27$.
6. Conclusion:

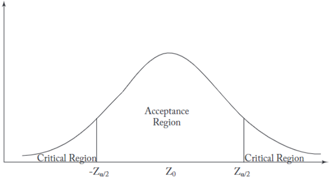


Figure 10.1: Two-tail normal test.

Values from a distribution in R

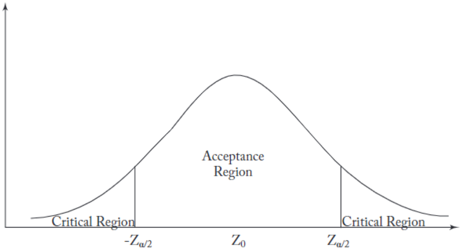


Figure 10.1: Two-tail normal test.

```
> ### z-score ###
> qnorm(0.025, mean = 0, sd = 1, lower.tail = TRUE)
[1] -1.959964
> qnorm(0.025, mean = 0, sd = 1, lower.tail = FALSE)
[1] 1.959964
> ### p-value ###
> pnorm(-1.066, mean = 0, sd = 1, lower.tail = TRUE)
[1] 0.1432118
> pnorm(1.066, mean = 0, sd = 1, lower.tail = FALSE)
[1] 0.1432118
```

Exercise: specify the null hypothesis

Research questions:

1. Claim that the mean of the drugs produced by the company is 100 mg.
2. A stockbroker claimed that weekly average return on a stock is normal with an average of return of 0.5%.
3. The National Bureau of Statistics (NBS) claimed that equal to 10% of the graduate youths are unemployed.

*Level of significance is 0.05.

Exercise: identify the possible distributions

Collected data:

1. If a random sample of 60 drugs is chosen with mean of 98 mg. and standard deviation of 14 mg.
2. He took the 20 previous weeks return and found that the weekly average returns was 0.48% with standard deviation 0.08%.
3. To ascertain the validity of the claim, a random sample of 10,000 graduate youths are selected in which 1,250 graduate youths are unemployed.

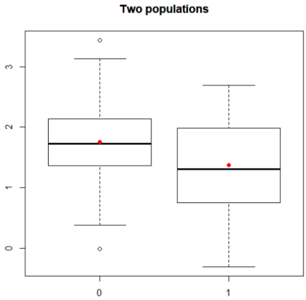
Possible names of testing for mean of one population

Sample size is matter! If the sample size is larger, we can use approximated version of test.
Notice that the assumptions of a test.
Common used tests are:

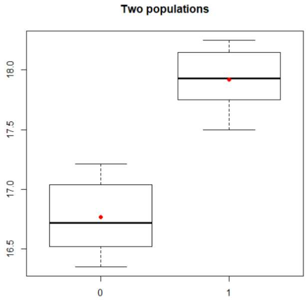
- z-test: testing for population mean. But, the observations are follow a normal distribution with the known variance.
- t-test: testing for population mean. The observations are random samples from a continuous distribution with unknown mean and variance, and the random samples are continuous or ordinal variables.
- Binomial test: testing for the proportion of event occurring. The observations are random samples from a Bernoulli distribution $Ber(p)$. If the sample size is large, the test statistic is approximated to a normal distribution with mean p and variance $\frac{p(1-p)}{n}$.

Why using testing hypothesis?

Difference between two populations?



mean of Group 0 > mean of Group 1?

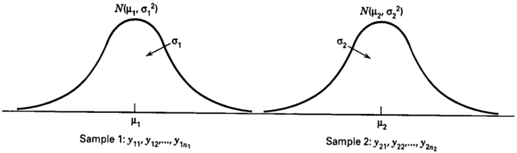


mean of Group 0 < mean of Group 1?

Hypothesis testing framework

Statistical hypotheses:

$$y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, 2, j = 1, \dots, n.$$



(Null hypothesis)	$H_0: \mu_1 = \mu_2$
(Alternative hypothesis)	$H_1: \mu_1 \neq \mu_2$

Assumptions for the independent two-sample test

Two populations are assumed that $Y_{11}, \dots, Y_{1n_1}$ are random samples from $N(\mu_1, \sigma_1^2)$, and $Y_{21}, \dots, Y_{2n_2}$ are random samples from $N(\mu_2, \sigma_2^2)$.

1. Random samples (also between two populations)
2. Homogeneity: $\sigma_1^2 = \sigma_2^2 = \sigma^2$
3. Normality distributed (why?)

Statistical hypotheses:

(Null hypothesis) $H_0: \mu_1 = \mu_2$

(Alternative hypothesis) $H_1: \mu_1 \neq \mu_2$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is known

(Null hypothesis) $H_0: \mu_1 = \mu_2$

(Alternative hypothesis) $H_1: \mu_1 \neq \mu_2$

Under $H_0: \mu_1 = \mu_2$,

$$Z_0 = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{y}_1 - \bar{y}_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1),$$

$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ estimates the population mean μ_i, $i = 1,2$.

If $|Z_0| > \Phi^{-1}(0.975)$, where $\Phi(z)$ is the cdf of a standard normal distribution, then reject $H_0: \mu_1 = \mu_2$.

Values from a distribution in R

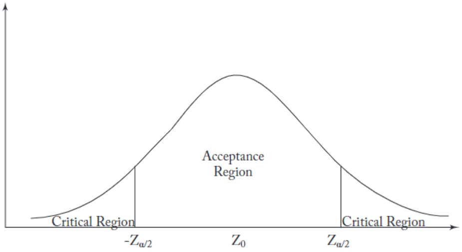


Figure 10.1: Two-tail normal test.

```
> ### z-score ###
> qnorm(0.025, mean = 0, sd = 1, lower.tail = TRUE)
[1] -1.959964
> qnorm(0.025, mean = 0, sd = 1, lower.tail = FALSE)
[1] 1.959964
> ### p-value ###
> pnorm(-1.066, mean = 0, sd = 1, lower.tail = TRUE)
[1] 0.1432118
> pnorm(1.066, mean = 0, sd = 1, lower.tail = FALSE)
[1] 0.1432118
```

Group members (Week 07: Oct 17)

Randomly assigned:

GROUP	Topics								
1	Levene's test & F-test	水利所 1 碩	游博凱	數據所 1 碩	劉仁忠	統計系 4	葉詠馨	統計所 1 碩	林承寬
2	Chi-squared test	水利所 1 碩	林奎瑱	數據所 1 碩	徐仁瓏	統計系 4	潘翠婷	統計所 1 碩	吳思蒨
3	Independent t-test	災碩士學程 1	鄭柏武	數據所 1 碩	張立勳	統計系 4	黃筱云	統計所 1 碩	陳沛群
4	Welch's test	工設所 2 碩	陳柏瑋	數據所 1 碩	李易庭	統計系 4	林詠晴	統計所 1 碩	黃群翔
5	Mann-Whitney U-test	工設所 2 碩	江沛晴	數據所 1 碩	侯登耀	統計系 4	郭旻霏	統計所 1 碩	李品嫻
6	Kruskal-Wallis test	環工所 1 碩	林炫君	數據所 1 碩	吳明軒	統計系 4	陳亭瑄	統計所 1 碩	王媛鈺
7	Paired t-test	會計所 2 碩	黃敬涵	數據所 1 碩	曾文海	統計所 1 碩	洪瑞廷	統計所 1 碩	李宗祐
8	Wilcoxon signed-rank test	企管所 1 碩	陳冠瑜	數據所 1 碩	黃亮臻	統計所 1 碩	張伊萱	統計所 1 碩	施文千
9	McNemar's test	環醫所 4 博	劉光威	數據所 1 碩	李家銘	統計所 1 碩	李承祐	統計所 1 碩	尹子維
10	Friedman test	統計所 2 博	蔡昇翰	統計所 1 碩	黃皓謙	統計所 1 碩	黃懷玲		

Contact:

https://docs.google.com/spreadsheets/d/1Mr_pXbooHfSb9Tz7xeA-PqBuR_LmAlnXjrD5_2-SZ8U/edit?usp=sharing