UNSW Mathematics Society presents

# MATH1041 Revision Seminar

## Life Sciences

Statistics For Life Sciences

T3, 2020

*Presented by Gerald Huang and Raymond Li*

# Table of Contents I

1. **Descriptive statistics**
   - Types of data
   - Numerical summaries
   - Graphical summaries
2. **Probability**
   - Basic definitions
   - Probability rules
   - Conditional probability
   - Independence and mutual exclusion
3. **Random Variable I – Discrete RVs**
   - Probability distribution function
   - Mean and Variance of Discrete RVs
   - The Binomial Distribution
4. **Random Variable II – Continuous RVs**
   - Probability density function

# Table of Contents II

- Mean and Variance of Continuous RVs
- The Normal Distribution

5 **Statistical Inference**

6 **Population Proportion**

7 **Two Parameters**

8 **Linear Regression**

# *Descriptive statistics*

# Types of data

## Quantitative data

**Quantitative data** takes *numerical* values where arithmetic makes sense!

**Examples**.

- temperature;
- WAM distribution.

## Categorical data

**Categorical data** places objects into separate *categories*. They are used to *describe* a particular data.

**Examples**.

- colour;
- gender.

# Numerical summary of a Categorical Variable

## Table of frequency

- List all possible categories.
- List all of the counts, percent, or proportion in each category.

| Categories | Frequency | Percentage (%) |
|---|---:|---:|
| Category A | 78 | 19.11 |
| Category B | 330 | 80.88 |

# Numerical summary of a Quantitative Variable

## Mean – Measure of Location

The **mean** of an observed sample is given by the *average* of the observed values.

- Add up all of the observed values and then divide by the number of values.

## IQR – Measure of Spread

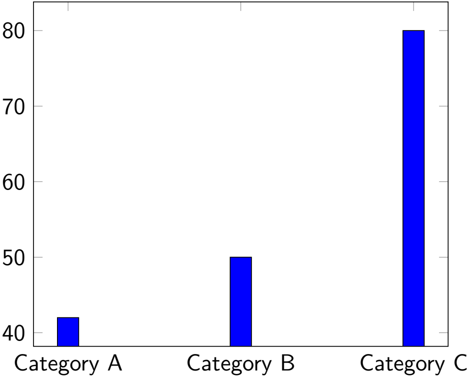
- Calculate $Q_3$ which is the third quartile.
- Calculate $Q_1$ which is the first quartile.
- The interquartile range is simply

$$IQR = Q_3 - Q_1.$$

# Graphical summary of a Categorical Variable

## Bar chart

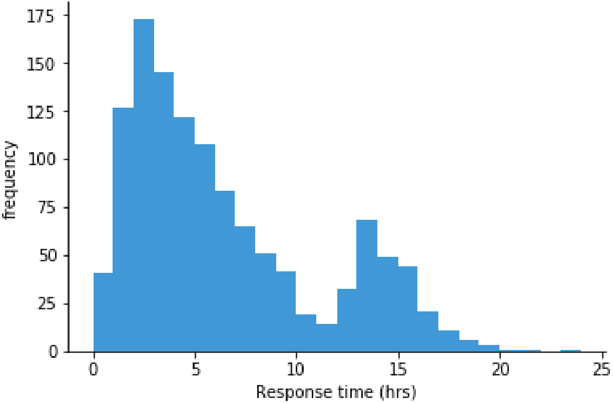
We can use a **bar chart** to graph the data across different categories.



# Graphical summary of a Quantitative Variable

## Histogram

A **histogram** is a graphical display of data using bars of different heights.



# *Probability*

# Events and Sample spaces

## Definition 2.1: Event

An **event** is an individual outcome.

## Definition 2.2: Sample space

A **sample space** is the set of *all possible events*. We normally denote a sample space as $S$.

## Definition 2.3: Probability of an event

If an event is denoted as $A$, then the **probability** of event $A$ is denoted by $P(A)$.

# Probability Rules

1. **Boundedness** – For any event $A$, the probability is bounded between 0 and 1:
$$0 \leq P(A) \leq 1.$$

2. **Covering** – The probability of the **sample space** is 1:
$$P(S) = 1.$$

3. **Additive rule*** – For any two events $A$ and $B$,
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

4. **Complement rule**: For any event $A$, the complement is denoted as $A^c$. The complement rule states that
$$P(A^c) = 1 - P(A).$$

# Probability Rules – Additive rule

**Be careful!**

You might be used to seeing

$$P(A \text{ or } B) = P(A) + P(B).$$

This is not generally true! We'll see this rule at a later time where we place some restrictions on $A$ and $B$.

For now, use the fact that

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

# Conditional Probability

We now look at probabilities where extra information is assured!

## Definition 2.4: Conditional Probability

For two events $A$ and $B$ (with $P(A) \neq 0$), the **conditional probability** of $A$ (or the probability of $A$ given that event $B$ has already occurred) is given by

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}.$$

We say $P(A \mid B)$ as *the probability of $A$* ***given*** *$B$*. In other words, we know that event $B$ has already occurred.

## MATH1041 2016 S2 Q2)

In the survey completed by MATH1041 students at the start of this year, one question asked was **gender** (male/female), and another question asked was **hemisphere** of birth (northern/southern).
60% of respondents were female.
28% of respondents were born in the northern hemisphere.
45% of respondents were female and born in the southern hemisphere.

- Show that the percentage of respondents who were male and born in the northern hemisphere is 13%.
- Suppose the chosen student is known to be male. What is the conditional probability that the student was born in the northern hemisphere, given that the student is male? That is, compute $P(N \mid M)$.

- 60% of respondents were female.
- 28% of respondents were born in the northern hemisphere.
- 45% of respondents were female and born in the southern hemisphere.

Show that the percentage of respondents who were male and born in the northern hemisphere is 13%.

Note that we want to find $P(N \text{ and } M)$. By our additive rule, we have

$$P(N \text{ or } M) = P(N) + P(M) - P(N \text{ and } M).$$

So,

$$P(N \text{ and } M) = P(N) + P(M) - P(N \text{ or } M).$$

Observe that the complement of "female and born in the southern hemisphere" is the same as "male or born in the northern hemisphere.

So

$$P(N \text{ or } M) = 1 - P(F \text{ and } S) = 1 - 0.45 = 0.55.$$

Hence, we have

$$\begin{aligned} P(N \text{ and } M) &= P(N) + P(M) - P(N \text{ or } M) \\ &= 0.28 + (1 - 0.6) - 0.55 \\ &= 0.28 + 0.4 - 0.55 \\ &= 0.13. \end{aligned}$$

So, the probability of respondents who were male and born in the northern hemisphere is 13%.

- 60% of respondents were female.
- 28% of respondents were born in the northern hemisphere.
- 45% of respondents were female and born in the southern hemisphere.

Suppose the chosen student is known to be male. What is the conditional probability that the student was born in the northern hemisphere, given that the student is male? That is, compute $P(N \mid M)$.

We need to calculate the conditional probability $P(N \mid M)$. By our conditional probability formula, this is equivalent to

$$P(N \mid M) = \frac{P(N \text{ and } M)}{P(M)}.$$

From our previous part, we proved that $P(N \text{ and } M) = 0.13$. On the other hand, we see that the complement of a "student being female" is "student being male". So

$$P(M) = P(F^c) = 1 - P(F) = 1 - 0.6 = 0.4.$$

So we have

$$P(N \mid M) = \frac{P(N \text{ and } M)}{P(M)} = \frac{0.13}{0.4} = \frac{13}{40}.$$

# Independence and mutually exclusive events

We now look closely at some restrictions on the events $A$ and $B$.

**Definition 2.5: Independence**

For events $A$ and $B$, we say that they are **independent** if either

$$P(A) = P(A \mid B) \quad \text{or} \quad P(B) = P(B \mid A).$$

By rewriting our conditional probability, we also have the equivalent form of independence.

$$P(A) = \frac{P(A \text{ and } B)}{P(B)} \iff P(A)P(B) = P(A \text{ and } B).$$

## MATH1041 2015 S2 Q2i)

A fair coin is flipped twice, so on each toss $P(H) = \frac{1}{2}$ and $P(T) = \frac{1}{2}$.
Let $A$ be the event: '*the first flip is H*'.
Let $B$ be the event: '*both flips have the same outcome*', i.e. $HH$ or $TT$.
Using the definition of independence, prove that the events $A$ and $B$ are **independent**.

We want to show that either

$$P(A \mid B) = P(A) \quad \text{or} \quad P(B \mid A) = P(B).$$

Observe that $P(A)$ is just the probability of flipping a head on the first go, so it is $P(A) = P(H) = 1/2$.
Also, we see that $P(B)$ is the probability that the both flips have the same outcome. We have 4 outcomes (HH, HT, TH, TT) with two of the flips the same (HH, TT). So the probability of $B$ is simply $\frac{1}{2}$. The probability of $A$ and $B$ happening together occurs when we have the outcome $HH$. Hence, we have $P(A \text{ and } B) = \frac{1}{4}$.
Hence, we have

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{1/4}{1/2} = \frac{2}{4} = \frac{1}{2} = P(B),$$

which is enough to show independence!

# Independence and mutually exclusive events

**Definition 2.6: Mutually exclusive events**

We say that events $A$ and $B$ are **mutually exclusive** if they are *disjoint* – that is, $P(A \text{ and } B) = 0$.

Be careful! Mutually exclusive events are NOT the same as independent events! Independent events occur when events $A$ and $B$ don't depend on each other. Mutually exclusive events occur when events $A$ and $B$ cannot occur at the same time!

# Back to our additive rule...

Recall that

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

But if $A$ and $B$ are *mutually exclusive*, then $P(A \text{ and } B) = 0$. So,

$$P(A \text{ or } B) = P(A) + P(B) - 0 = P(A) + P(B).$$

# Independence and mutually exclusive events

## In summary...

- If $A$ and $B$ are **independent** events, then

$$P(A)P(B) = P(A \text{ and } B) \quad \text{or} \quad P(A) = P(A \mid B).$$

- If $A$ and $B$ are **mutually exclusive** events, then

$$P(A \text{ and } B) = P(A) + P(B).$$

# *Random Variable I – Discrete RVs*

# A definition...

### Definition 3.1: Random Variable

A **random variable** is a function that takes an outcome from a sample space $S$ and maps it to a real number.

- We usually denote random variables by a capital letter (e.g. $X$ and $Y$).

### Definition 3.2: Discrete Random Variables

A **discrete random variable** is a random variable that takes up a *finite* set of values. In other words, we can count each outcome in the sample space.

**Examples**.

- Rolling a die;
- Number of people drinking coffee on a particular morning;
- Number generated by a random number generator.

# Probability distribution – pdf of discrete RVs

## Definition 3.3: Probability distribution

The **probability distribution function** of a discrete random variable is a function that lists out the *likelihood* of a particular event from happening.

**Examples**.

- If $x_1$ is the event of a six being thrown from a fair die, then $P(X = x_1) = \dfrac{1}{6}$.
- If $x_2$ is the event of a head being thrown from a fair coin, then $P(X = x_2) = \dfrac{1}{2}$.

# Mean and variance of Discrete RVs

## Definition 3.4: Mean of a Discrete RV

The **mean** of a discrete random variable is the weighted sum of the event and its probability

$$\mathbb{E}(X) = \mu_X = \sum_{x_i \in X} x_i P(X = x_i).$$

- You may sometimes hear it being referred to as the *expected value* of $X$.

### Example: Mean of a Discrete RV

Let $x_i$ be the number of cars owned per household. A table of values is given as below.

| Number of cars | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Probability | 0.09 | 0.37 | 0.37 | 0.17 |

The mean number of cars owned can be computed as follows:

$$0 \times 0.9 + 1 \times 0.37 + 2 \times 0.37 + 3 \times 0.17 = 1.62.$$

# Independence

We won't cover a formal definition in this course for independence, but loosely speaking...

Two random variables $X$ and $Y$ are *independent* if they don't influence each other in any way.

As an example, if $X$ is the event of rolling a number of a die and $Y$ is the event of tossing a coin, the two have no influence on one another. So we say that $X$ and $Y$ are independent!

But this will be important for when we talk about variance!

# Rules for expected values

We talked about computing expected values from a random variable! Now, suppose we have a random variable that is a combination another random variable! Let's look at $Y = aX + b$.

**Proposition: Rule for means / expected values**

Suppose that the mean of $X$ is given by $\mu_X$. Then:

$$\mu_Y = \mathbb{E}(Y) = \mathbb{E}(aX + b) = a \cdot \mathbb{E}(X) + b = a \cdot \mu_X + b.$$

## MATH1041 Midterm, Q19

A random variable $X$ can take values $0, 2, 5, 9$ with the following probability distribution:

| $x$ value | 0 | 2 | 5 | 9 |
|---|---|---|---|---|
| probability | 0.4 | 0.1 | 0.2 | 0.3 |

1. Compute the mean, $\mu_X$.
2. Let $Y$ be the random variable $Y = 5 + 9X$. Compute the mean, $\mu_Y$.

1. The mean is simply

$$\mu_X = 0 \times 0.4 + 2 \times 0.1 + 5 \times 0.2 + 9 \times 0.3 = 3.9.$$

2. The new mean is simply

$$\mu_Y = 5 + 9 \cdot \mu_X = 5 + 9 \times 3.9 = 40.1.$$

# Variance of Discrete RVs

### Definition 3.5: Variance of a Discrete RV

The **variance** of a discrete random variable is

$$\operatorname{Var}(X) = \sigma_X^2 = \sum_{x_i \in X} (x_i - \mu_X)^2 P(X = x_i).$$

The **standard deviation** is given by the square root of the variance

$$\sigma_X = \sqrt{\sigma_X^2}.$$

## MATH1041 Midterm, Q19 cont.

A random variable $X$ can take values $0, 2, 5, 9$ with the following probability distribution:

| $x$ value | 0 | 2 | 5 | 9 |
|---|---|---|---|---|
| probability | 0.4 | 0.1 | 0.2 | 0.3 |

Compute the variance, $\sigma_X^2$.

We found the mean before in the last example: $\mu_X = 3.9$, so

$$\begin{aligned}\sigma_X^2 &= (0-3.9)^2 0.4 + (2-3.9)^2 0.1 + (5-3.9)^2 0.2 + (9-3.9)^2 0.3 \\ &= 14.49.\end{aligned}$$

# Rules for Variances

## Proposition

Suppose that $Y = aX + b$. Then

$$\mathrm{Var}(Y) = \mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X).$$

Remember how we said that independence was important for variances? Well, it turns out that if two events $X$ and $Y$ are **independent**, then we have the following property.

If $X$ and $Y$ are **independent**, then

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

# The Binomial Distribution

We now look closely at the most famous type of discrete distribution: the **binomial distribution**.

Suppose that an experiment is repeated $n$ times with each trial being (1) independent. Each of these trials only have (2) two outcomes: a success and a failure. Finally, each of these trials have the (3) same probability $p$ for success.

If all three of these conditions are met, then we say that it is a binomial distribution with $X$ being the random variable of success.

# The Binomial Distribution

### Definition 3.6: Binomial Distribution

Let $X$ be the number of successes. Then the probability distribution of $X = x$ is given by

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x},$$

with $p$ being the probability of success.

We say that $X$ is a Binomial Distribution, which we can stylise as

$$X \sim B(n, p).$$

# Computing Binomial Distributions

## MATH1041 Midterm, Q20

A particular coin is biased. Each time it is flipped, the probability of a head is $P(H) = 0.9$ and the probability of a tail is $P(T) = 0.1$. Each flip is independent of the other flips. The coin is flipped twice. Let $X$ be the total number of times the coin shows a head out of **two** flips.

- Compute $P(X = 2)$.

- This question is a binomial distribution with $n = 2$ and $p = 0.9$. So we have

$$P(X = 2) = \binom{2}{2} 0.9^2 \times 0.1^0 = 0.9^2 = 0.81.$$

# Computing Binomial Distributions

## MATH1041 Midterm, Q20

A particular coin is biased. Each time it is flipped, the probability of a head is $P(H) = 0.9$ and the probability of a tail is $P(T) = 0.1$. Each flip is independent of the other flips. The coin is flipped twice. Let $X$ be the total number of times the coin shows a head out of **two** flips.

- What is the probability that $X < 2$?

Recall that $\underbrace{P(X = 0) + P(X = 1)}_{P(X<2)} + P(X = 2) = 1$. So

$$P(X < 2) = 1 - P(X = 2) = 1 - 0.81 = 0.19.$$

# Mean and Variance of a Binomial Distribution

### Expected value and variance of a binomial distribution

If $X$ is a binomial distribution, then

- the expected value is

$$\mu_X = \mathbb{E}(X) = np.$$

- the variance is

$$\sigma_X^2 = \text{Var}(X) = np(1-p).$$

# *Random Variable II – Continuous RVs*

# A definition...

**Definition 4.1: Continuous Random Variables**

A **continuous random variable** is a random variable that takes an interval set of values (instead of finite many).

**Examples**.

- time it takes for a bus to arrive.

# Probability density – pdf of continuous RVs

Similar to a probability distribution, we can determine the probability of event by a function for continuous random variables.

## Definition 4.2: Probability density function

The **probability density function** $f_X(x)$ is a curve that describes the random variable under a sample space. The probability of an event is the area under the curve that makes up the event.

# Mean and variance of continuous RVs [NON-ASSESSABLE]

## Definition 4.3: Mean of a Continuous RV

The **mean** of a continuous random variable is given by

$$\mathbb{E}(X) = \mu_X = \int_X x f_X(x)\, dx$$

## Definition 4.4: Variance of a Continuous RV

The **variance** of a continuous random variable is given by

$$\mathrm{Var}(X) = \sigma_X^2 = \int_X (x - \mu_X)^2 f_X(x)\, dx.$$

...Don't even WORRY about this! You're not going to need to know how to compute these.

# Rules for expected values

Follows the same as the discrete case!

**Proposition: Rule for means / expected values**

Suppose that the mean of $X$ is given by $\mu_X$. Then:

$$\mu_Y = \mathbb{E}(Y) = a \cdot \mu_X + b.$$

The same works for variances!

# Rules for Variances

### Proposition

Suppose that $Y = aX + b$. Then

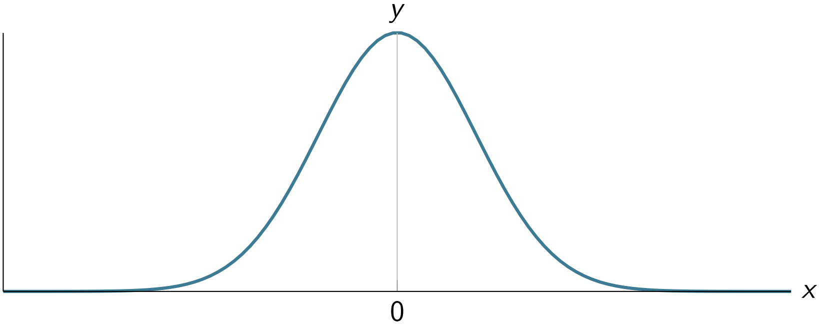
$$\mathrm{Var}(Y) = \mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X).$$

If $X$ and $Y$ are **independent**, then

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

# Normal Distributions

We now look at one of the most important family of curves in statistics: the **Normal Distribution**!



# A definition...

If a distribution looks like a bell curve like earlier, we say that it is *normally distributed*. We often write the distribution as

$$X \sim \mathcal{N}(\mu, \sigma).$$

If $\mu = 0$ and $\sigma = 1$, we say that $Z$ follows a *standard normal distribution*. We normally use $Z$ for a standard normal distribution!

# Using R output to find Normal Distributions

To find the standard normal probability, we use `pnorm()`.

**MATH1041 Midterm, Q23**

Let $Z$ be a standard normal random variable. That is, $Z \sim N(0, 1)$. Compute the value of $P(Z > 1.6)$.

`pnorm(1.6)` calculates $P(Z < 1.6)$. So we have

$$P(Z > 1.6) = 1 - P(Z < 1.6) = 1 - \texttt{pnorm(1.6)} = 0.05479929.$$

# Using R output to find Normal Distributions

To find the standard normal quantiles, we use `qnorm()`.

**MATH1041 Midterm, Q23**

Let $Z$ be a standard normal random variable. That is, $Z \sim N(0,1)$. Find the value of $c$ such that $P(Z < c) = 0.0301$.

We simply use the `qnorm()` function in R on to find

$$P(Z < c) = 0.0301 \implies c = \texttt{qnorm(0.0301)} = -1.879326.$$

# Algebra and Normal Distributions

We can use algebra to also find probability of intervals.

If $a < b$, then

$$P(a < Z < b) = P(Z < b) - P(Z < a).$$

Compute the value of $P(-1.26 < Z < 1.6)$ if $Z \sim N(0, 1)$.

$$\begin{aligned} P(-1.26 < Z < 1.6) &= P(Z < 1.6) - P(Z < -1.26) \\ &= \texttt{pnorm(1.6)} - \texttt{pnorm(-1.26)} = 0.841366. \end{aligned}$$

# Algebra and Normal Distributions

If $Z \sim N(0,1)$, then we also have that the probability density of $Z$ being symmetric about 0!
So...

If $Z \sim N(0,1)$, then

$$P(Z < -c) = P(Z > c).$$

So... we have

$$\begin{aligned} P(-c < Z < c) &= P(Z < c) - P(Z < -c) \\ &= P(Z < c) - P(Z > c) \\ &= P(Z < c) - (1 - P(Z < c)) \\ &= 2P(Z < c) - 1. \end{aligned}$$

## MATH1041 Midterm, Q28 [modified]

Compute the probability

$$P(-3 < Z < 3),$$

given that `pnorm(3)` $= 0.9986501$.

$$\begin{aligned} P(-3 < Z < 3) &= P(Z < 3) - P(Z < -3) \\ &= P(Z < 3) - P(Z > 3) \\ &= 2P(Z < 3) - 1 \\ &= 2 \times 0.9986501 - 1 = 0.9973002. \end{aligned}$$

# Standardising a Normal Distribution

If $X \sim N(\mu, \sigma)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

## MATH1041 Midterm, Q22

Suppose that $X \sim N(-15.9, 24.5)$ and that $Z$ follows the standard normal distribution.
Suppose that $P(X < x) = 0.55$.

- Compute the value of $z$ such that $P(Z < z) = 0.55$.
- Hence, compute the value of $x$ such that $P(X < x) = 0.55$.

- We can compute $z$ using `qnorm(0.55)` $= 0.1256613$.
- We can standardise $X$. Observe that

$$\begin{aligned} P(X < x) &= P(X - (-15.9) < x - (-15.9)) \\ &= P\left(\frac{X + 15.9}{24.5} < \frac{x + 15.9}{24.5}\right) \\ &= P\left(Z < \frac{x + 15.9}{24.5}\right) = 0.55. \end{aligned}$$

So $x =$ `qnorm(0.55)` $\times 24.5 - 15.9$.

# *Statistical Inference*

# Simple random sample

## Random sample

A simple random sample of size *n* consists of *n* individuals from the population, chosen in such a way that every possible combination of *n* individuals has equal chance to be the sample actually selected.

This can usually be done by generating random numbers with each element in the population assigned a number. Note that 'simple' in the definition means that the sampling is done with no replacement. The observations $X_i$ in a random sample (with replacement) are independent and identically distributed.

# Sample mean and standard deviation

## Sample mean

The sample mean is defined by

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$

The mean and standard deviation of the mean are given by

$$\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

# Sample standard deviation

## Sample standard deviation

The sample standard deviation is defined by

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

## Standard error

The standard error is given by

$$\frac{S}{\sqrt{n}}$$

# Population mean CI (known $\sigma$)

## Confidence Interval

Since for large sample sizes $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, the confidence interval for the mean is

$$CI_C(\mu) = [\bar{X} - z^* \frac{\sigma}{n}, \bar{X} + z^* \frac{\sigma}{n}]$$

where $z^*$ is the number such that $P(-z^* < Z < z^*) = C$, $Z \sim N(0, 1)$. For a 95% CI, take $z^* = 1.96$. This confidence interval should be used when the standard deviation of the population is known.

# Population mean CI (unknown $\sigma$)

## Confidence Interval

If the standard deviation is estimated by the standard error, then there is additional uncertainty. Instead we use the statistic

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

which follows a $t$-distribution with $n-1$ degrees of freedom. The CI is

$$CI_C(\mu) = [\bar{X} - t^* \frac{s}{n}, \bar{X} + t^* \frac{s}{n}]$$

where $t^*$ is the number such that $P(-t^* < Z < t^*) = C$, $Z \sim N(0,1)$.

For large values of $n$, the CI will be the same as the case where $\sigma$ is known. In this case the $Z$ statistic can be used instead.

# Hypothesis Tests

To formally test a claim, we must first set up the null hypothesis $H_0$ and alternative hypothesis $H_a$.

## Null and alternative hypotheses

$H_0$: the statement we are trying to reject
$H_a$: the statement we are trying to find evidence for

Suppose $H_0 : \mu = \mu_0$
The hypothesis test can be one-sided e.g. $H_a : \mu > \mu_0$. In this case, we want to check if the sample mean is large.
Instead, $H_a : \mu \neq \mu_0$ is a two-sided test, meaning that we are checking if $\mu$ is too small or too large.

# Test statistic

A test statistic will then need to be used to mathematically determine whether the sample mean is too extreme.

## Test statistic

The test statistic is usually of the form

$$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

The test statistic is assumed to follow the null distribution, i.e. the null hypothesis is assumed to be true. Assuming the $\bar{X} \sim N(\mu, \sigma^2)$, it follows that $T \sim N(0, 1)$. If $\sigma$ is not know, it can be estimated by $s$, and the statistic follows a $t$ distribution. The idea is to calculate the value of the test statistic under the null hypothesis, and if this statistic is too large or too small, there is evidence to reject $H_0$.

# $P$-value and significance value

## Definition of $P$-value

The $P$-value is the probability the test statistic takes a value as extreme or more extreme than the observed value under $H_0$.

What classifies as small or large will depend on the situation. More precisely, we need to define the significance level.

## Definition of $\alpha$

The significance level $\alpha$ is the probability of rejecting the null hypothesis when it is true.

If the $P$-value is smaller than $\alpha$, there is sufficient evidence to reject the null hypothesis in favour of the alternative.

# Example hypothesis test

**MATH1041 2016 S2 Q3**

A gardener needs to determine the pH of the soil in her garden, and she will add lime to the soil if the pH is less than 6.5. She takes 16 samples from random locations and finds that the average pH of the samples is 6.3 with a standard deviation of 0.3. Carry out a hypothesis test to determine whether the gardener must add lime to the soil.

# 2016 S2 Q3

First we state the null and alternative hypotheses

$$H_0 : \mu \geq 6.5$$

$$H_a : \mu < 6.5$$

Since we are testing for the mean, the test statistic is given by

$$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{6.3 - 6.5}{0.3/\sqrt{16}} = -2.67$$

Under the null hypothesis, $T$ follows a standard $t$ distribution with 15 degrees of freedom.

The $P$-value can be determined from the data tables, noting that this is a one-sided test.

$$P(T < -2.67) = 0.0087$$

Thus there is strong evidence to reject the null hypothesis. The gardener should add more lime to the soil.

# Central Limit Theorem

## Theorem

If $X_1, ...X_n$ are independent and identically distributed, and the sample size $n$ is large,

$$\bar{X} \overset{\text{approx}}{\sim} N(\mu, \sigma/\sqrt{n})$$

In other words, if the sample size is large enough, the mean can always be assumed to be normally distributed, regardless of its distribution. Hypothesis testing on the mean can be conducted regardless of whether the data is normally distributed. The sample size required varies based on distribution, with more skewed distributions needing larger sample sizes.

# Example CLT

### MATH1041 2018 S2 Q2i

Use the Central Limit Theorem and the 68-95-99.7 rule to answer this question.

A random group of 16 women attempt to squeeze into a lift.

Assume that the weights of women have a mean of $\mu = 72$ kg with a standard deviation of $\sigma = 8kg$.

(a) Compute the probability that the average weight of the 16 women is more than 74 kg.

(b) Hence compute the probability that the total weight of the 16 women is more than 1184 kg.

# 2018 S2 Q2i

(a) Using the CLT,

$$\bar{X} \sim N\left(72, \frac{8}{\sqrt{16}}\right) = N(72, 2)$$

Since 74 is one standard deviation above the mean, the probability is 16%.
(b) Since $1184 = 16 \times 74$, this probability is the same as above, which is 16%.

# *Population Proportion*

# Sample proportion

## Sample proportion definition

If $X \sim B(n, p)$ then

$$\hat{p} = \frac{X}{n}$$

is the sample proportion random variable.
Further,

$$\mu_{\hat{p}} = p \text{ and } \sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$$

Since $p$ is often not known, the standard error is approximated by

$$\mathrm{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

# Normal approximation

For large values of $n$, the binomial distribution can be approximated by the normal distribution with the same mean and variance.

## Normal approximation

$$X \overset{\text{approx}}{\sim} N\left(np, \sqrt{np(1-p)}\right)$$

and

$$\hat{p} \overset{\text{approx}}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

As a rule of thumb, the approximation is good if $np \geq 10$ and $np(1-p) \geq 10$.

# Continuity correction

When using a continuous distribution like the normal to approximate a discrete distribution like the binomial, continuity correction needs to be done.

While a binomial will have a probability mass at an integer $x$, we can imagine this to be equal to the area under the normal density from $x - 0.5$ to $x + 0.5$. For example $P(X < 5)$ can be approximated by the normal as $P(X < 4.5)$ and $P(X \leq 5)$ can be approximated as $P(X < 5.5)$

The easiest way to do this correctly is to think about it logically rather than remembering a formula.

# CI for proportions

Confidence intervals for proportions can be determined by noting that

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

## CI

The level $C$ confidence interval is thus

$$\mathrm{CI}_C(p) = \left[\hat{p} - z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

where $z^*$ is the number such that $P(-z^* < Z < z^*) = C$, $Z \sim N(0, 1)$.

# Example CI

### MATH1041 2016 S1 Q4iv modified

466 of the 6272 Swedish men in the sample were diagnosed with prostate cancer.
(a) Derive the 95% confidence interval for the true proportion with prostate cancer $p$ using the following result

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

(b) Hence construct the 95% confidence interval.

# 2016 S1 Q4iv

(a) It follows from the given result that

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

To construct a 95% confidence interval, we want this expression to lie between the 2.5 and 97.5 percentiles of the standard normal distribution.

$$-1.96 \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96$$

Rearranging the expression gives

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

# 2016 S1 Q4iv

(b) The confidence interval can be found by substituting

$$\hat{p} = \frac{466}{6272} = 0.074 \text{ and } n = 6272$$

into the confidence interval derived in part (a).

$$[0.0678, 0.808]$$

# Hypothesis tests for proportions

## Hypothesis test for $p$

$$H_0 : p = p_0$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

which follows a standard normal distribution under the null hypothesis.

Hypothesis tests can be conducted in the same way as for the population mean - including one-sided and two-sided tests.

# Example population inference

## MATH1041 2018 S2 Q1iib modified

Newspapers reported Newspoll results where voters were polled on their preference between the two major Australian political parties: Labor and Liberal. The proportion preferring Labor was 0.51. Suppose that the Newspoll sample size was n = 100. Making the necessary assumptions about the sampling process, construct a 95% confidence interval for the proportion of Australian voters preferencing Labor at the time of the poll.
The following RStudio output is given: qnorm(0.975) = 1.9599, qnorm(0.95) = 1.6448, qnorm(0.925) = 1.2815

# 2018 S2 Q1iib

The expected proportion

$$\hat{p} = 0.51$$

The standard error is given by

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.05$$

The 95% confidence interval is given by

$$\begin{aligned} \mathrm{CI}_{0.05}(p) &= [0.51 - 1.9599 \times 0.05, 0.51 + 1.9599 \times 0.05] \\ &= [0.412, 0.608] \end{aligned}$$

# *Two Parameters*

# Two-sample $t$-test

Suppose we have two samples and we want to investigate whether the population means are the same.

## Comparing population means

$$H_0 : \mu_1 = \mu_2$$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

where the pooled standard deviation

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Note that this assumes the populations have the same standard deviation and are normally distributed.

# CI for two-sample test $t$-test

Similar to the other tests, a confidence interval can be determined from the distribution of $T$.

## CI

$$\mathrm{CI}_C(\mu_1 - \mu_2) = \left[(\bar{X}_1 - \bar{X}_2) \pm t^* S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

# Paired $t$-test

The paired $t$-test is used when we want to test whether there is a difference in mean in two samples where values are paired with each other. Since the two samples are not independent, the two-sample $t$-test is not appropriate.

## Paired $t$-test

The null hypothesis is that the mean is the same in the two samples.

$$H_0 : \mu_D = X_1 - X_2 = 0$$

$$T = \frac{\bar{D}}{S_D/\sqrt{n}} \sim t(n-1)$$

where $D_i$ is the difference in the $i$-th pair of observation and $S_D$ is the standard deviation of this difference.

This test assumes that the difference is normally distributed, but this works practically if the sample size is large.

# Chi-squared test

The $\chi^2$ test of independence can be used to determine whether two random variables are independent.

$$H_0: \text{ the two variables are independent}$$

$$H_a: \text{ the two variables are not independent}$$

In order to test this, we first construct a two-way table, which has rows and columns for each of the two variables. Under the null hypothesis, the expected count under $H_0$ is given by

$$\text{Expected counts} = \frac{\text{row total} \times \text{column total}}{n}$$

# Chi-squared test

## Test statistic

The test statistic is given by

$$X^2_{obs} = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Under $H_0$, $X^2_{obs}$ follows a $\chi^2$ distribution with $(r-1)(c-1)$ degrees of freedom.

It is important that the sample size is always large enough that all the expected counts are greater than 10. Similar to other test statistics, a *p*-value can be determined to assess whether the null hypothesis should be rejected. Note that this is always a one-sided test.

# Example of Chi-Squared test

## MATH1041 2016 S2 Q4i modified

A simple random sample of 400 in all fields of employment was cross-classified by broad Employment Area and Gender.

| Employment | Female | Male |
|---|---|---|
| Commerce | 33 | 55 |
| Industry | 70 | 47 |
| Service | 137 | 58 |
| TOTAL | 240 | 160 |

Carry out a test of whether Employment Area and Gender are independent or not. You may assume that the observed $X^2 = 27.11$.

# 2016 S2 Q4i

(A) Clearly state the hypotheses $H_0$ and $H_1$.

$$H_0 : \text{Employment Area and Gender are independent}$$

$$H_a : \text{Employment Area and Gender are not independent}$$

(B) Under the null hypothesis, verify that the expected count of respondents that are employed in Service and are Female is 117.
Total in service $= 137 + 58 = 195$
Total female $= 240$

$$\text{Expected count} = \frac{195 \times 240}{400} = 117$$

# 2016 S2 Q4i

(C) Show that the contribution to $X^2$ from the table entry corresponding to employment in Service and Female is 3.42.

$$\begin{aligned}\text{Contribution} &= \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \\ &= \frac{(137-117)^2}{117} \\ &= 3.42\end{aligned}$$

(D) State the distribution that can be used to assess the significance of $X^2$ assuming that $H_0$ is true.

$$\chi^2(r-1)(c-1) = \chi^2(2)$$

# 2016 S2 Q4i

(E) Give an expression for and calculate the $P$-value.

$$P\text{-value} = P(X^2 \geq X^2_{obs}) = P(X^2 \geq 27.11)$$

By comparing with the data table, it can be seen that this is less than 0.005.

(F) State your conclusion in terms of $H_0, H_a$ and in simple language.

Since the $P$-value is very small, there is strong evidence against $H_0$ and in favour of $H_a$.

# *Linear Regression*

# Least square regression line

Least squares regression is finding the line that minimises the sum of squares of the residuals.

## Regression line

For a sample, the equation for the fitted line is given by

$$\hat{y} = b_0 + b_1 x$$

with

$$b_1 = r\frac{s_y}{s_x}, b_0 = \bar{y} - y_1 x$$

where $r$ is the Pearson coefficient and $s_x, s_y$ the sample standard deviations.

$$y = \hat{y} + \text{residual}$$

# Population regression line

The regression line given on the previous slide depends on the sample. The population regression line can be thought of as the true regression line for the population.

## Population regression line

The population regression line is denoted by

$$\mu_y = \beta_0 + \beta_1 x$$

and so

$$y = \mu_y + \text{error}$$

The assumption is that the error term is normally distributed with 0 mean and the same variance at each $x$ value.

# Inference of slope $\beta_1$

## Test statistic

$$\frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} \sim t(n-2)$$

The confidence interval can be given by

$$\mathrm{CI}_C(\beta_1) = [\hat{\beta}_1 - t^*\mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + t^*\mathrm{SE}(\hat{\beta}_1)]$$

We can test for whether there is a relationship between the variables using the null hypothesis $H_0 : \beta_1 = 0$.

# Analysing residuals

It is important to recognise the assumptions that are made when fitting a regression line - errors are normally distributed with zero mean and constant variance.

A residual vs fitted plot can be used to visually determine this - there should not be a pattern in the graph. A pattern suggests that there is some unaccounted for relationship i.e. a poor fit. Further, the residual quantiles can be plotted against the quantiles of a normal distribution - if residuals are normally distributed, a 45 degree line should be obtained.

# Example regression

## MATH1041 2018 S2 Q4 modified

For a random sample of 64 university students, the relationship between the reported hours of sleep during the previous 24 hours and the reported hours of study during the same period was tested. A linear regression model was fitted and some RStudio output is given on the next page. Use this RStudio output and plots provided to answer the following questions:

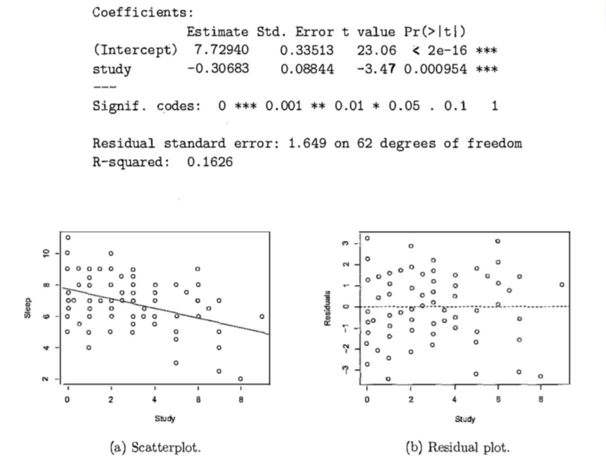
(a) Is the relationship between sleep and study statistically significant?

(b) Write down the linear regression equation for predicting sleep from study.

(c) Estimate the mean sleep if study is equal to 6 hours.

(d) Estimate the change in mean sleep for a one hour increase in study.

# Data



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.72940    0.33513   23.06  < 2e-16 ***
study       -0.30683    0.08844   -3.47 0.000954 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.649 on 62 degrees of freedom
R-squared:  0.1626
```

(a) Scatterplot.

(b) Residual plot.

Figure: RStudio output

# 2018 S2 Q4 modified

(a) We need to check if the slope is significantly different from zero. From the data, the $P$-value for the slope parameter is 0.000954 so there is very strong evidence for the slope being different from zero.
(b) sleep $= 7.72940 - 0.30683 \times$ study $+$ residual
(c) Subbing study $= 6$ we obtain sleep $= 5.88842$
(d) This is simply the estimated slope, which is -0.30683.