UNSW Mathematics Society Presents

MATH1041 Workshop

Presented by Andrew William, Gerald Huang

Overview 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 Variables I - Discrete RVs
   Probability distribution function
   Mean and Variance of Discrete RVs
   The Binomial Distribution
4. Random Variable II – Continuous RVs
   Probability density function

Overview II

Mean and Variance of Continuous RVs
The Normal Distribution

5. Statistical Inference

   Estimation – Sample mean
   Confidence interval (known standard deviation)
   Confidence interval (unknown standard deviation)
   Hypothesis Testing and the Central Limit Theorem

6. Population Proportion

   Sample proportion
   Confidence interval for proportions
   Hypothesis Testing for Proportions

7. Two Parameters

8. Linear Regression

1. Descriptive Statistics

Types of data

Quantitative data

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

Examples : height, WAM.

Categorical data

Categorical data is data that exists naturally in *one or more categories.*

Examples: Colour, gender.

Why is this important?

How we visualize and summarize data is highly dependent on what types of data we deal with!

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. To compute this, take sum of all observed values, and divide by total number of observations.

Mode - Measure of Location

The **median** of an observed sample is the ”middle” of observed values. To compute this, you arrange all values in order, and take the middle value.

- Robust to outliers.

Numerical summary of a Quantitative Variable

Standard Deviation – Measure of Spread

The **standard deviation** is a measure of how much the data varies around the mean. Here is the formula to compute it:

$$\sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}}$$

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

$$IQR = Q_3 - Q_1.$$

- Robust to outliers.

Numerical summary of a Quantitative Variable

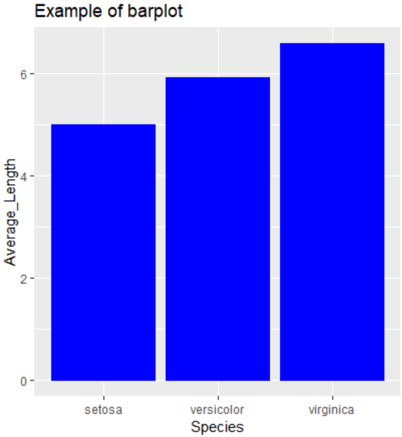
5 number summary

- List of 5 numbers that do well in describing data succinctly.
- Partitions data into 5 quartiles : min, Q_1, median, Q_3, max.
- Typically used in first few steps of exploratory data analysis to get feel of data at a glance.

Graphical summary of a Categorical Variable

Bar chart

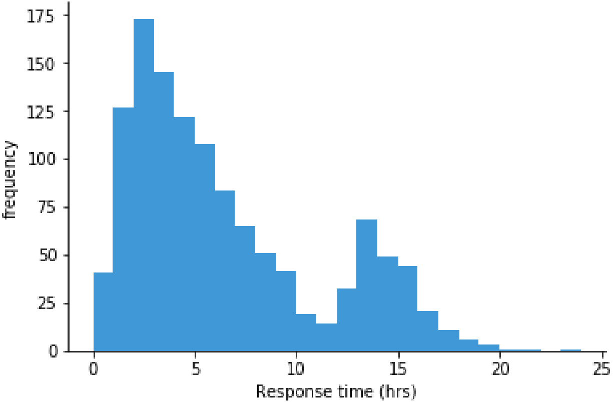
We can use a **bar chart** to graph the data across different categories.



Graphical summary of a Quantitative Variable

Histogram

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

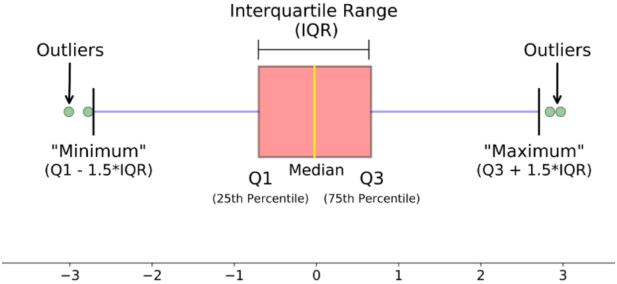


Graphical summary of a Quantitative Variable

Boxplot

A boxplot is a visual way of looking at the spread of data.

- Shows us Q_1, Q_2, Q_3.
- 'Whiskers' show us the "minimum" and "maximum" values of data that we consider to be non-outliers.
- Anything outside of that range can be considered an outlier.



2. 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 2018 S2 Q1i)

Assume 20% of people have a flu immunization at the start of winter. Suppose that 5% of people who have a flu immunization will contract the flu, and that 30% of those who do not have a flu immunization will contract the flu.

- What proportion of people will contract the flu?

Tree Diagram Approach

Tree Diagram Approach

- 20% people immunized.
- 5% of immunized people contract flu.
- 30% of non-immunized people born flu.

What proportion of people will contract flu?

There's a lot of information to unpack here, so one helpful thing to do is to write down everything that we know.

- 20% people immunized.
- 5% of immunized people contract flu.
- 30% of non-immunized people born flu.

What proportion of people will contract flu?

There's a lot of information to unpack here, so one helpful thing to do is to write down everything that we know.

Let I be the event that someone is immunized, and F be the event that someone gets flu.

Summarizing our information in probability terms, we have that:

$$\begin{aligned} P(I) &= 0.2 \\ P(F \mid I) &= 0.05 \\ P(F \mid I^c) &= 0.3 \end{aligned}$$

Someone having flu, by getting flu and being immunized, or getting flu and not being immunized.
We need to glue this together using probability rules.

$$P(F) = P((F \text{ and } I^c) \text{ or } (F \text{ and } I)).$$

Someone having flu, by getting flu and being immunized, or getting flu and not being immunized.

We need to glue this together using probability rules.

$$P(F) = P((F \text{ and } I^c) \text{ or } (F \text{ and } I)).$$

To make our lives easier, we'll say $P(A) = P(F \text{ and } I^c)$, and $P(B) = P(F \text{ and } I)$. Hence, we are trying to solve for $P(F) = P(A \text{ or } B)$.

Someone having flu, by getting flu and being immunized, or getting flu and not being immunized.

We need to glue this together using probability rules.

$$P(F) = P((F \text{ and } I^c) \text{ or } (F \text{ and } I)).$$

To make our lives easier, we'll say $P(A) = P(F \text{ and } I^c)$, and $P(B) = P(F \text{ and } I)$. Hence, we are trying to solve for $P(F) = P(A \text{ or } B)$.

By the additive law of probability, we have

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

Note that $P(A \text{ and } B)$ doesn't exist because you can't be both immunized and not at the same time! This is an example of a **mutually exclusive** event, which we'll talk more about later.

Our equation to solve becomes

$$P(F) = P(A) + P(B).$$

Let's work on $P(A) = P(F \text{ and } I^c)$ first.

Our equation to solve becomes

$$P(F) = P(A) + P(B).$$

Let's work on $P(A) = P(F \text{ and } I^c)$ first.
By the equation of conditional probability we know

$$\frac{P(F \text{ and } I^c)}{P(I^c)} = P(F \mid I^c).$$

So this means

$$\begin{aligned} P(F \text{ and } I^c) &= P(F \mid I^c) \times P(I^c) \\ &= 0.3 \times P(I^c) \end{aligned}$$

Here, we can apply our definition of complement rule to find out what $P(I^c)$ is. So

$$P(I^c) = 1 - P(I) = 1 - 0.2 = 0.8.$$

Plugging this back into the equation above, we get that

$$P(F \text{ and } I^c) = 0.3 \times 0.8 = 0.24.$$

Here, we can apply our definition of complement rule to find out what $P(I^c)$ is. So

$$P(I^c) = 1 - P(I) = 1 - 0.2 = 0.8.$$

Plugging this back into the equation above, we get that

$$P(F \text{ and } I^c) = 0.3 \times 0.8 = 0.24.$$

We can then work on the next part $P(F and I)$ in a similar fashion. So

$$\begin{aligned} P(A) &= P(F \text{ and } I) \\ &= P(F \mid I) \times P(I) \\ &= 0.05 \times 0.2 \\ &= 0.01. \end{aligned}$$

Finally, we can plug everything back into our original equation

$$\begin{aligned} P(F) &= P(A) + P(B) \\ &= 0.01 + 0.24 \\ &= 0.25 \end{aligned}$$

Hence, the probability of getting a flu is 25%!

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 $P(A) = P(H) = \frac{1}{2}$.

Observe that $P(A)$ is just the probability of flipping a head on the first go, so it is $P(A) = P(H) = \frac{1}{2}$.

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 $\frac{1}{2}$.

Observe that $P(A)$ is just the probability of flipping a head on the first go, so it is $P(A) = P(H) = \frac{1}{2}$.

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 $P(A \text{ and } B) = \frac{1}{4}$.

Observe that $P(A)$ is just the probability of flipping a head on the first go, so it is $P(A) = P(H) = \frac{1}{2}$.
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

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

3. Random Variables 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.

- Number of heads when flipping 3 coins. The sample space is all possible combinations, and
- 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 $P(X = x_1) = \frac{1}{6}$.
- If x_2 is the event of a head being thrown from a fair coin, then $P(X = x_2) = \frac{1}{2}$.

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:

$$0 \times 0.09 + 1 \times 0.37 + 2 \times 0.37 + 3 \times 0.17 = 1.62.$$

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 μ_X. Then:

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

MATH1041 2020 Final Exam Q1a) - Slightly modified

Let random variable X represent number of cups of coffee that Jane drinks on a normal day.
Below is the probability distribution of number of cups of coffee she drank over the past year:

Number of cups of coffee (per day)	0	2	5	8
Probability	0.4	0.4	0.1	0.1

1. Calculate the average number of cups of coffee she drinks per day, μ_X.
2. Let Y be the random variable $Y = 5 + 9X$. Compute the mean, μ_Y.

1. The mean would be

$$\mu_X = 0 \times 0.4 + 2 \times 0.4 + 5 \times 0.1 + 8 \times 0.1 = 2.1.$$

2. The mean of Y would be

$$\mu_Y = 5 + 9 \cdot \mu_X = 5 + 9 \times 2.1 = 10.5.$$

Variance of Discrete RVs

Definition 3.5: Variance of a Discrete RV

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

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

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

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

MATH1041 2020 Final Exam Q1a) - Slightly modified

Let random variable X represent number of cups of coffee that Jane drinks on a normal day.
Below is the probability distribution of number of cups of coffee she drank over the past year:

Number of cups of coffee (per day)	0	2	5	8
Probability	0.4	0.4	0.1	0.1

Calculate the variance σ_X^2.

Recall we found that $\mu_X = 2.1$, so

$$\begin{aligned}\sigma_X^2 &= (0-2.1)^2 0.4 + (2-2.1)^2 0.4 + (5-2.1)^2 0.1 + (8-2.1)^2 0.1 \\ &= 2.9571.\end{aligned}$$

Rules for Variances

Proposition

Suppose that $Y = aX + b$. Then

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

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

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

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 2018 S2 Q1ic)

At the end of the flu season four people, all of whom had received flu immunizations, are randomly chosen.

- What is the probability exactly two of the four people contracted flu?

* Recall we found that probability of any person contracting flu is 0.25.

First, let X be the random event r out of n people contracting flu. Then we have to recognize that X has a binomial distribution, with $n = 4$, $p = 0.25$. Therefore we need to compute $P(X = 2)$

First, let X be the random event r out of n people contracting flu. Then we have to recognize that X has a binomial distribution, with $n = 4$, $p = 0.25$. Therefore we need to compute $P(X = 2)$

$$P(X = 2) = \binom{4}{2} 0.25^2 \times 0.75^2 = \frac{27}{128}.$$

Computing Binomial Distributions

MATH1041 2018 S2 Q1ic) modified)

At the end of the flu season four people, all of whom had received flu immunizations, are randomly chosen.

- What is the probability that 3 or less people contracted flu?

Similar to the previous example, now we are trying to compute for $P(X \leq 3)$. Recall that

$$\underbrace{P(X=0)+P(X=1)+P(X=2)+P(X=3)}_{P(X\leq 3)}+P(X=4)=1.$$

So $P(X \leq 3) = 1 - P(X = 4)$.

Computing Binomial Distributions

MATH1041 2018 S2 Q1ic) modified)

At the end of the flu season four people, all of whom had received flu immunizations, are randomly chosen.

- What is the probability that 3 or less people contracted flu?

Similar to the previous example, now we are trying to compute for $P(X \leq 3)$. Recall that

$$\underbrace{P(X=0)+P(X=1)+P(X=2)+P(X=3)}_{P(X\leq 3)}+P(X=4)=1.$$

So $P(X \leq 3) = 1 - P(X=4)$. Then

$$1 - P(X=4) = 1 - \binom{4}{4} \times 0.25^4 \times 0.75^0 = \frac{255}{256}.$$

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

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

4. 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 μ_X and that $Y = aX + b$. 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

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

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

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

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, 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!

MATH1041 Final Exam T2 2021 Q8 Modified

Andrew decides to create a new random variable Y from a normal random variable X. X follows a normal distribution with mean 5 and standard deviation 2. Andrew then defines $Y = 4X^2 - 29X + 70$.

- Compute $E[Y]$.

Hint : Use the fact that $Var(X) = E[X^2] - [E[X]]^2$.

To get $E[Y]$, first note that we can break it down using the rules for expected values. So

$$\begin{aligned} E[Y] &= E[4X^2 - 29X + 70] \\ &= E[4X^2] + E[-29X] + E[70] \\ &= 4E[X^2] - 29E[X] + 70. \end{aligned}$$

To get $E[Y]$, first note that we can break it down using the rules for expected values. So

$$\begin{aligned} E[Y] &= E[4X^2 - 29X + 70] \\ &= E[4X^2] + E[-29X] + E[70] \\ &= 4E[X^2] - 29E[X] + 70. \end{aligned}$$

$$E[Y] = E[4X^2 - 29X + 70] = 4E[X^2] - 29E[X] + 70.$$

For the second term, $E[X]$ can be inferred from the question which states that

$$X \sim \mathcal{N}(5, 2).$$

Hence $E[X] = 5$ (also $Var(X) = 4$).

For the first term, $E[X^2]$, we can compute this using the hint the question gave us!

$$Var(X) = E[X^2] - [E[X]]^2.$$

So,

$$[E[X^2] = Var(X) + [E[X]]^2 = 2^2 + 25 = 29.$$

For the first term, $E[X^2]$, we can compute this using the hint the question gave us!

$$Var(X) = E[X^2] - [E[X]]^2.$$

So,

$$[E[X^2] = Var(X) + [E[X]]^2 = 2^2 + 25 = 29.$$

Finally, plugging everything back in, we get that

$$E[Y] = 4 \times 29 - 29 \times 5 + 70 = 41.$$

Using R output to find Normal Distributions

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

Example of calculating probability for standard normal distribution

Let Z be a standard normal random variable, so $Z \sim N(0, 1)$.

- Find $P(Z > 0.666)$.

Using R output to find Normal Distributions

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

Example of calculating probability for standard normal distribution

Let Z be a standard normal random variable, so $Z \sim N(0, 1)$.

- Find $P(Z > 0.666)$.

`pnorm(0.666)` calculates $P(Z < 0.666)$. So we have

$$P(Z > 0.666) = 1 - P(Z < 0.666) = 1 - \texttt{pnorm(0.666)} = 0.2527056.$$

Using R output to find Normal Distributions

Example of calculating probability for standard normal distribution

Let Z be a standard normal random variable, so $Z \sim N(0, 1)$.

- Find the value of c such that $P(Z < c) = 0.05$.

Using R output to find Normal Distributions

Example of calculating probability for standard normal distribution

Let Z be a standard normal random variable, so $Z \sim N(0, 1)$.

- Find the value of c such that $P(Z < c) = 0.05$.

To find the standard normal quantiles, we use `qnorm()`.
We simply use the `qnorm()` function in R on to find

$$P(Z < c) = 0.05 \implies c = \texttt{qnorm(0.05)} = -1.644854.$$

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

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

Illustration of Normal Distribution Symmetry

Example: Finding probability of standard normal RV using algebra

Compute the probability

$$P(-2 < Z < 2),$$

given that `pnorm(2)` $= 0.9772499.$

Algebraically, we can write

$$\begin{aligned} P(-2 < Z < 2) &= P(Z < 2) - P(Z < -2) \\ &= P(Z < 2) - P(Z > 2) \\ &= 2P(Z < 2) - 1 \\ &= 2 \times 0.9772499 - 1 \\ &= 0.9544998. \end{aligned}$$

Standardising a Normal Distribution

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

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

You can find this using the rules for expectations of random variables, as well as the rules for variances of random variables.

$$E[Z] = E[\frac{X - \mu}{\sigma}] = \frac{1}{\sigma} E[X] - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0.$$

$$Var[Z] = Var[\frac{X - \mu}{\sigma}] = \frac{1}{\sigma^2} Var[X] = \frac{1}{\sigma^2} \sigma^2 = 1.$$

MATH1041 2020 Midterm, Q22 modified

Suppose that $X \sim N(-3, 4)$ and that Z follows the standard normal distribution.
Suppose that $P(X < x) = 0.4$.

- Compute the value of x such that $P(X < x) = 0.4$.

First, we must standardise X. Notice that since x is an observed value from X, when we standardize X, we must thus apply the same to x. Observe that

$$\begin{aligned} P(X < x) &= P(X - (-3) < x - (-3)) \\ &= P\left(\frac{X+3}{4} < \frac{x+3}{4}\right) \\ &= P\left(Z < \frac{x+3}{4}\right) = 0.4. \end{aligned}$$

MATH1041 2020 Midterm, Q22 modified

Suppose that $X \sim N(-3, 4)$ and that Z follows the standard normal distribution.
Suppose that $P(X < x) = 0.4$.

- Compute the value of x such that $P(X < x) = 0.4$.

Then, this means that

$$\frac{x+3}{4} = \texttt{qnorm(0.4)}.$$

Finally, we have that

$$x = \texttt{qnorm(0.4)} \times 4 - 3$$

5. Statistical Inference

Sample mean

Let X and Y be random variables. Recall the following results:

- $\mu_{aX+bY} = a\mu_X + b\mu_Y$ (**no independence required**).
- $\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y$ (**independence required**).

Sample mean

Let X and Y be random variables. Recall the following results:

- $\mu_{aX+bY} = a\mu_X + b\mu_Y$ (**no independence required**).
- $\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y$ (**independence required**).

For a collection of random variables to be a random sample, we will assume that they are (1) **independent** and (2) **identically distributed**.

Sample mean, bias, variability

(Properties) Random sample

Let $X_1, X_2, \ldots, X_n$ be a random sample.

- Since each random variable is **identically distributed**, then the means and variances (and consequently, the standard deviation) are all the same.
- Each random variable is **independent** of one another.

Sample mean

Let $X_1, X_2, \ldots, X_n$ be a random sample; that is, they are identically distributed and independent. Consider the sample mean:

$$Y = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Sample mean

Let $X_1, X_2, \ldots, X_n$ be a random sample; that is, they are identically distributed and independent. Consider the sample mean:

$$Y = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Then

- **Mean**: $\mu_Y = \mu_X$.
- **Variance**: $\sigma_Y^2 = \dfrac{\sigma_X^2}{n}$.
- **Standard deviation**: $\sigma_Y = \dfrac{\sigma_X}{\sqrt{n}}$.

Confidence interval

Suppose that we didn't know σ_X and we wanted to estimate it instead.

Confidence interval

Suppose that we didn't know σ_X and we wanted to estimate it instead. We can instead construct intervals that tell us with a lot of confidence whether the true value is inside the interval.

Confidence interval

Suppose that we didn't know σ_X and we wanted to estimate it instead. We can instead construct intervals that tell us with a lot of confidence whether the true value is inside the interval.

- **Method 1**: $68 - 95 - 99.7$ rule.
- **Method 2**: Use a level of confidence.

68 − 95 − 99.7 rule

Roughly, 68% of data is within 1 standard deviation from the mean, 95% of data is within 2 standard deviations from the mean, and 99.7% of data is within 3 standard deviations from the mean.

$68 - 95 - 99.7$ rule

Roughly, 68% of data is within 1 standard deviation from the mean, 95% of data is within 2 standard deviations from the mean, and 99.7% of data is within 3 standard deviations from the mean. So we can construct an interval using this sort of information.

(2021 Term 2, Midterm) Q25

Student IQ scores collected from a school are approximately normally distributed with a mean of 96 and a standard deviation of 6.

1. What proportion of the student IQ scores are between 84 and 108?
2. What range contains the central 68% of IQ scores?
3. Ella has an IQ of 114. How many standard deviations above the mean is Ella's IQ?
4. Hence, what proportion of students in the school have an IQ higher than Ella's IQ?

Population mean CI (known σ)

If we don't have one of the three probabilities, we can construct our interval ourselves instead. Even if we do, we might want an exact confidence interval.

Population mean CI (known σ)

If we don't have one of the three probabilities, we can construct our interval ourselves instead. Even if we do, we might want an exact confidence interval.

1. Instead of using the *number* of standard deviations, we will use exact quantiles. That is, we will find z^* such that $P(-z^* < Z < z^*) = C$, where $Z \sim N(0, 1)$.
2. Then we compute the "number of standard deviations" from the mean:

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

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

Population mean CI (unknown σ)

When σ is not known, we use the t-distribution instead.

1. Replace z^* with t^* where t^* is the value computed such that, if $T \sim t(n-1)$, then $P(-t^* < T < t^*) = C$.
2. We no longer have the true standard deviation so we can replace it with the *sample standard deviation*:

$$\boxed{CI_C(\mu) = \left[\bar{X} - t^* \frac{S}{\sqrt{n}}, \bar{X} + t^* \frac{S}{\sqrt{n}}\right]}$$

where t^* is the value computed such that, if $T \sim t(n-1)$, then $P(-t^* < T < t^*) = C$ and S is the sample standard deviation

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

Hypothesis Testing

We have a particular claim that we want to prove about an *unknown* parameter. To prove this, we will use a hypothesis test.

General set up:

1. State the **null** and **alternative** hypothesis.
2. Calculate the value of the **test statistic** and its **null distribution**.
3. Calculate the *p*-**value** and conclude.

1. Null and alternative hypothesis

The first step is to state what the null and alternative hypotheses are. The alternative hypothesis is the claim we are trying to find evidence for (i.e. the one we *want to prove*) and the null hypothesis is its counter claim.
Notation: H_0 : null hypothesis, H_a : alternative hypothesis.

2. 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

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

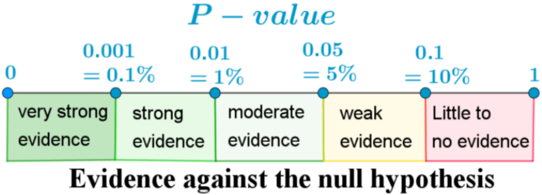
- Similar to the confidence interval, we use σ if it is known so that $T \sim N(0, 1)$ distribution.
- Otherwise, we can approximate it using a sample standard deviation by replacing σ with s where s is the *sample standard deviation.*

3. 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.

In MATH1041, we generally use the value of the p-value to determine how strong our evidence is.



(2016, Semester 2) 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.

Step 1. State the null and alternative hypothesis.

Step 1. State the null and alternative hypothesis.

$$H_0 : \mu_0 = 6.5, \quad H_a : \mu_0 < 6.5.$$

Step 1. State the null and alternative hypothesis.

$$H_0 : \mu_0 = 6.5, \quad H_a : \mu_0 < 6.5.$$

Step 2. Compute the value of the test statistic.

Step 1. State the null and alternative hypothesis.

$$H_0 : \mu_0 = 6.5, \quad H_a : \mu_0 < 6.5.$$

Step 2. Compute the value of the test statistic. Note that we're using a sample standard deviation so the test statistic is a t-distribution. We, therefore, have

$$T_{\text{obs}} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{6.3 - 6.5}{0.3/\sqrt{16}} = \frac{-0.8}{0.3} = -2.67.$$

Step 1. State the null and alternative hypothesis.

$$H_0 : \mu_0 = 6.5, \quad H_a : \mu_0 < 6.5.$$

Step 2. Compute the value of the test statistic. Note that we're using a sample standard deviation so the test statistic is a t-distribution. We, therefore, have

$$T_{\text{obs}} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{6.3 - 6.5}{0.3/\sqrt{16}} = \frac{-0.8}{0.3} = -2.67.$$

Now we need to compute the probability that the test statistic takes a value as (or more) extreme than the observed value under H_0. The null distribution is a $t(16-1)$ distribution; hence, we have that $T \sim t(15)$.

Step 3. Compute the p value.

Step 3. Compute the p value. This is equivalent to computing

$$P(T < T_{\text{obs}}) = P(T < -2.67) \approx 0.008797577....$$

You can find this value by computing `pt(-8/3, df = 15)` on R.

Step 3. Compute the p value. This is equivalent to computing

$$P(T < T_{\text{obs}}) = P(T < -2.67) \approx 0.008797577....$$

You can find this value by computing `pt(-8/3, df = 15)` on R.
Step 4. Conclude.

Step 3. Compute the p value. This is equivalent to computing

$$P(T < T_{\text{obs}}) = P(T < -2.67) \approx 0.008797577....$$

You can find this value by computing `pt(-8/3, df = 15)` on R.
Step 4. Conclude. There is strong evidence to reject the null hypothesis. In other words, the gardener should add more lime to the soil.

Central Limit Theorem

Assumptions

- $X_1, X_2, \ldots, X_n$ are **independent** and **identically distributed**.
- The sample size n is *sufficiently large*.

Under these assumptions, we have the Central Limit Theorem.

Definition (Central Limit Theorem)

The sample mean, $\bar{X}$, is approximately normal; that is,

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

Example CLT

(2018, Semester 2) Q2 i)

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 $\sigma = 8kg$.

1. Compute the probability that the average weight of the 16 women is more than 74 kg.
2. Hence compute the probability that the total weight of the 16 women is more than 1184 kg.

1. By the Central Limit Theorem, $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$. Thus, we have that
$$\bar{X} \sim N(72, 8/\sqrt{16}) = N(72, 2).$$

1. By the Central Limit Theorem, $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$. Thus, we have that

   $$\bar{X} \sim N(72, 8/\sqrt{16}) = N(72, 2).$$

   We want to compute $P(\bar{X} > 74)$. This is 1 standard deviation above the mean! By the $68 - 95 - 99.7$ rule, we have that $P(\bar{X} > 74) = 0.16$ (draw it out!).

2. We want to find $P(X_1 + X_2 + \cdots + X_{16} > 1184)$. We have that $\bar{X} = \frac{1}{16}(X_1 + X_2 + \cdots + X_{16})$, so that

   $$X_1 + X_2 + \cdots + X_{16} = 16\bar{X}.$$

   Hence,

   $$P(X_1 + X_2 + \cdots + X_{16} > 1184) = P(16\bar{X} > 1184) = P(\bar{X} > 74).$$

   This is the same as part a), so the probability is 16%.

6. Population Proportion

Sample proportion

Definition (Sample proportion)

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

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

Assumption

If $np \geq 10$ and $np(1-p) \geq 10$, then n is large enough to use the normal approximation to the binomial.

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.

Confidence interval for proportions

Confidence intervals for proportions can be determined by noting that

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

The level C confidence interval is

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

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

(2016, Semester 1) Q4 iv) – modified

466 of the 6272 Swedish men in the sample were diagnosed with prostate cancer.

1. Derive the 95% confidence interval for the true proportion with prostate cancer p using the following result

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

2. Hence construct the 95% confidence interval.

1. Since $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, then standardising the expression gives us

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

1. Since $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, then standardising the expression gives us
$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1).$$
Now, if we want to construct a 95% confidence interval, then we want this expression to lie between the 2.5 and 97.5 percentiles of the standard normal distribution. (Why?)

1. Since $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, then standardising the expression gives us

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

Now, if we want to construct a 95% confidence interval, then we want this expression to lie between the 2.5 and 97.5 percentiles of the standard normal distribution. (Why?)
We, therefore, require

$$-z^* \leq Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z^*$$

where z^* is the value computed such that
$P(-z^* < Z < z^*) = 0.95$. In other words, we have

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

2. Substituting $\hat{p} = \frac{466}{6272}$ and $n = 6272$ gives us the confidence interval

$$CI_C(p) = [0.0678, 0.808].$$

(2018 Semester 2) Q1 iib) – 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`.

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

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

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.

7. 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 χ^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

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

Under H_0, X_{obs}^2 follows a χ^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.

8. 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 β_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:

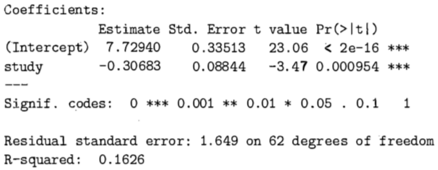
(a) Is the relationship between sleep and study statistically significant?

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

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

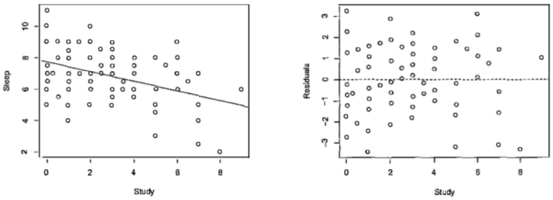
(d) Estimate the change in mean sleep for a one hour increase in study.

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) $\text{sleep} = 7.72940 - 0.30683 \times \text{study} + \text{residual}$
(c) Subbing $\text{study} = 6$ we obtain $\text{sleep} = 5.88842$
(d) This is simply the estimated slope, which is -0.30683.