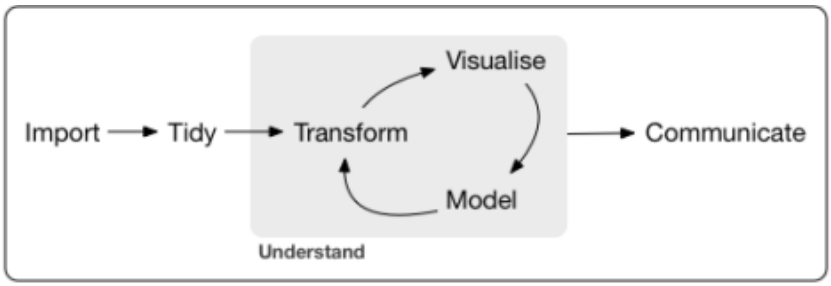
TOPIC 1 – Summarising and Visualizing Data

What is Statistics?

- The science of collecting, organising, interpreting, visualising, analysing and reporting data.



- Most data we collect has variability/uncertainty (often referred to as noise) which can make it difficult to work out what is really going on.

The Scientific Method

1. Formulate the research question
2. Collect relevant data
3. Do statistical analysis
4. Draw Conclusion

Samples and Populations

- A population is the set of all possible measurements of interest.
- A sample is a subset of measurements from the population.
- We use a sample to make a statement about the population.

Types of Data/Variables

- Categorical/Qualitative
  - Nominal (unordered) – information given are categories or names (e.g. gender)
  - Ordinal (ordered) – the categories can be ordered (e.g. good, average, bad)
- Numeric/Quantitative
  - Discrete – $\in \mathbb{N}$ from counting, and only integer values are possible
  - Continuous - $\in \mathbb{R}$ or part of R such as R^+ so that any value in the range is possible (e.g. length measurement)

Data Matrices

- Way to record and store data

	spam	num_char	line_breaks	format	number
1	0	21.705	551	1	small
2	0	7.011	183	1	big
3	1	0.631	28	0	none
50	0	15.829	242	1	small

Data Entry

- Data should be entered into a spreadsheet, database or statistical software package
- Ideally create template before collecting data
- We want to keep data tidy and clean

Tidy Data

- 'Tidy' is a standard way of storing data that is used throughout R
- It includes:
  - Each variable is in a column
  - Each observation is in a row
  - Each value is a cell
- Each row is known as a record (one for each subject) containing information collected for the subject.`

Introduction to R

- Creating an object:

```
object_name <- value
```

- All assigned variables are stored until overwritten or explicitly deleted using the command rm()
- To see what variables are stored use command ls()
- The command c() creates a vector

```
x <- c(4,1,5,7,6,2,1,9,5,2,1,8,4)
```

- sort(x) for sorting the vector into numerical order

Importing a dataset

- In order to import the file "email50.csv"

```
library(tidyverse)
email50 <- read_csv('email50.csv')
```

- In order to access each individual variable

```
head(email50$num_char)
```

Small and Large Datasets

- For a sample size *n,* observations are denoted by x_1, x_2, x_3,… x_n generally
- If n < 30 then the dataset is small; if n > 30 the dataset is large

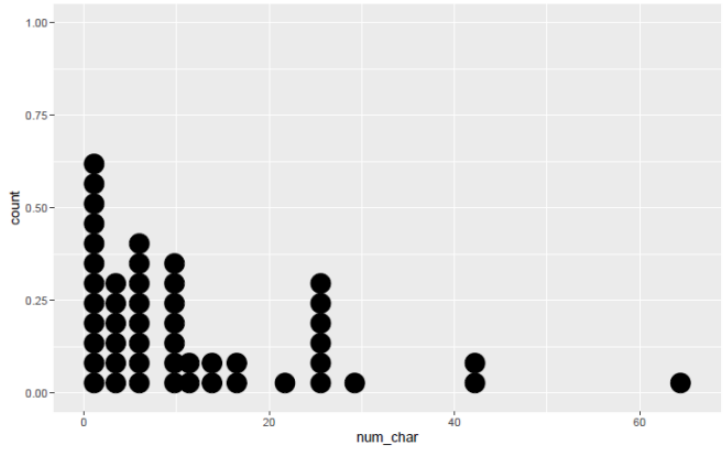
Data Visualisation with ggplot2

- ggplot() creates a coordinate system that you can add layers to
- The first argument in ggplot() is the data set being used in the graph
- geom_dotplot() creates a dotplot
- The general form/template for creating a graph is:

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

- For example:

```
library(tidyverse)
email50 <- read_csv("email50.csv")
ggplot(email50) +
  geom_dotplot(mapping = aes(x = num_char))
```



Stem-and-leaf displays

- For small datasets we can use a stem-and-leaf plot

```
stem(email50$num_char)
```

Frequencies and Proportions

- The frequencies with which certain outcomes occur within a certain dataset is given by:

```
table(email50$number) # base R
email50 %>% count(number) #tidyverse;
```

- The proportion or relative frequency is $\hat{p} = \frac{f_i}{n}$
- In R,

```
table(email50$number)/sum(table(email50$number))
email50 %>% count(number, sort = TRUE) %>% mutate(prop = n/sum(n))
```

- To find out more about any function in R, use ?. E.g. ?count
- Use mutate() to create new variables with existing ones

Bar Charts

- If you have a categorical/discrete variable, you can draw a bar chart to visualise it:

```
ggplot(data = email50) +
  geom_bar(aes(x = number))
```

- When you have two categorical variables (bivariate data) you can create a cluster bar chart:
  - First create a bivariate table:

```
table(email50$spam, email50$number)
```

  - Then to create the bluster bar chart:

```
ggplot(data = email50) +
  geom_bar(aes(x = number, fill=spam), position = "dodge")
```

- Switching x and fill can make the graph easier to interpret (up to personal preference)

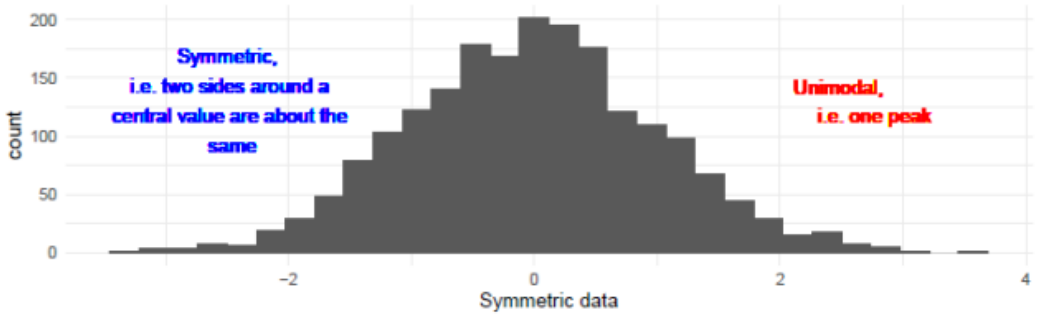
Histograms

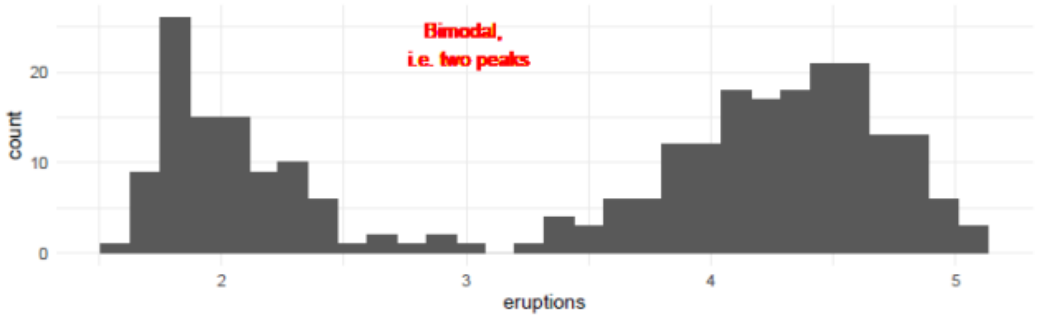
- Histograms give a summary of how data is distributed or spread out.
- Achieved by grouping data into intervals and the frequency of individuals in each interval is plotted
- Intervals all called **bins**
- For each interval or bin, a rectangle is constructed such that the frequencies are represented by the area and not the height.
- To construct a histogram:

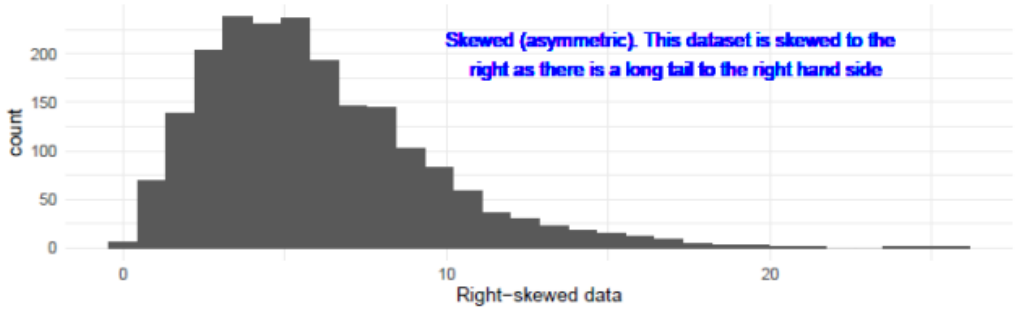
```
ggplot(data = email50) +
  geom_histogram(aes(x = num_char), bins = 6)
```

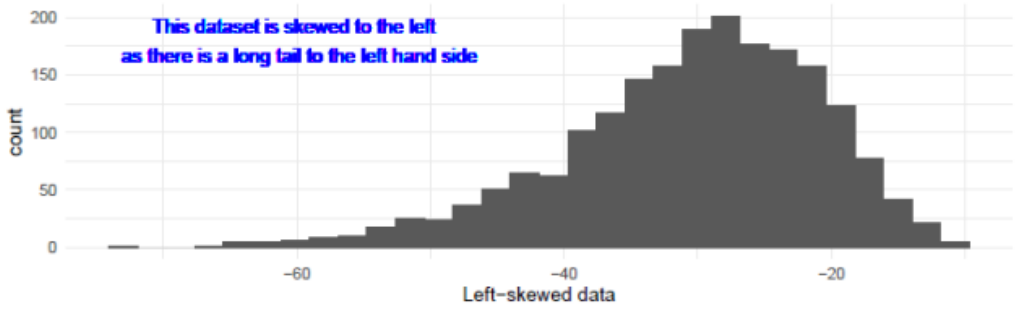
- It is important to determine the number of ins appropriately:
  - Fewer bins – sacrificing information
  - Too many bins – swamped by details and lose the overall picture

Shape of Data









Five Number Summary

- The minimum = $x_{(1)}$ and the maximum = $x_{(n)}$.
- The range = $x_{(n)} - x_{(1)}$
- The median is the middle value when the set is numerically ordered
- The lower quartile, Q_1, is the 25^{th} percentile

- The upper quartile, Q_3, is the 75th percentile.
- We denote the min, median and max as Q_0, Q_2 and Q_4.
- The **interquartile range** (IQR) = $Q_3 - Q_1$. It covers approx. 50% of the obs.
- In R the 5 number summary is obtained via:

```
summary(email50$num_char)
```

- Percentiles are obtained using:

```
quantile(email50$num_char, c(0.00, 0.25, 0.50, 0.75, 1))
```

- The values in the vector give the percentiles to be displayed.

Five Number Summary by hand

- The median can be calculated for:

n odd: $\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$

n even: $\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$

- If n/4 is an integer, then set k = n/4 and:

$$Q_1 = \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right), \quad Q_3 = \frac{1}{2}\left(x_{(n-k)} + x_{(n-k+1)}\right).$$

- If n/4 is not an integer then set $k = \lceil \frac{n}{4} \rceil$ and:

$$Q_1 = x_{(k)}, \quad Q_3 = x_{(n-k+1)}.$$

EXAMPLE:

- Suppose the data set of size n = 8 is:

  `# 1 2 3 4 5 6 7 8`
  - As 8/4 is an integer, it is easy to cut the dataset into 4 equal proportions.

    `# 1 2 | 3 4 | 5 6 | 7 8`
  - Then, for simplicity, we use the midpoint of 4 and 5 as the median of the dataset. i.e. $\tilde{x} = 4.5$.
- Suppose the dataset of size n = 9 is:

  `# 1 2 3 4 5 6 7 8 9`
- Obviously 9/4 is no longer an integer
  - The index for the median is $\frac{n+1}{2} = 5$.
  - The index for the lower quartile is $\lceil \frac{n}{4} \rceil = \lceil 2.25 \rceil = 3$.
  - To find the upper quartile simply count $\lceil \frac{n}{4} \rceil$ from the max.
    - Therefore, $Q_1 = 3$, $Q_2 = 5$, $Q_3 = 7$.

Boxplot

- Within tidyverse, you have to define the five number summary manually:

```
email50 %>%
  summarise(Q0 = min(num_char),
            Q1 = quantile(num_char, 0.25),
            Q2 = median(num_char),
            Q3 = quantile(num_char, 0.75),
            Q4 = max(num_char))
```

- How to draw a boxplot:
  - Draw an axis with a sensible scale
  - A box that stretches from Q_1 to Q_3.
  - A line in the middle that displays the median
  - Visual points that display observations that fall more than 1.5 times the IQR from either edge of the box.
  - A line (or whisker) that extends from each end of the box and goes to the farthest non-outlier point in the dataset.
- Formally, outliers are points that are more than 1.5 X IQR beyond [Q_1, Q_3]
- Hence:
  - LT (Lower threshold) = $Q_1 - 1.5$ X IQR
  - UT (Upper threshold) = $Q_3 + 1.5$ X IQR
- In R, we use:

```
ggplot(data = email50) +
  geom_boxplot(aes(y = num_char)) + coord_flip()
```

- The mapping is to the y-axis by default; coord_flip() flips the coordinate axes.
- Box plots give an impression of the shape of the dataset (symmetry, skewed and unusual obs)
- Boxplots are useful to compare a continuous variable (e.g. length) with a nominal variable (e.g. treatment).
- We can do this by adding in the x-axis as the nominal variables:

```
ggplot(data = tooth) +
  geom_boxplot(aes(y = length, x = dose))
```

- If you want to reorder the variables on the x-axis, you will have to convert them into factor variables with:

```
head(factor(tooth$dose))
```

- And then to reorder do:

```
head(fct_relevel(factor(tooth$dose), "High", after = 2))
```

```
ggplot(data = tooth) +
  geom_boxplot(aes(y = length,
                   x = fct_relevel(factor(dose), "High", after = 2))) +
  xlab("dose")
```

Pipe %>%

- Used to express a sequence of multiple operations
- Writes out code in a way that is easier to understand (rather than embedding code in brackets)

Review of Sigma Notation

- For the values $x_1 = 3$, $x_2 = 4$, $x_3 = 5$, $x_4 = 3$:

$$\sum_{i=2}^{n-1}(2x_i + 3) = 2\sum_{i=2}^{3} x_i + 3\underbrace{[(n-1)-(2-1)]}_{2} = 18 + 6 = 24;$$

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i, \quad \text{for any constant } c;$$

$$\sum_{i=1}^{n} c \cdot 1 = c\sum_{i=1}^{n} 1 = cn.$$

Sample Mean

- The sample mean is the average of observations
- For obs x_1, x_2,…, x_n:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- Given constants $c, d \in \mathbb{R}$ and obs $x_1, x_2, \ldots, x_n$. Then the mean of the transformed observations

$$e_i = c \cdot x_i + d, \quad i = 1, 2, \ldots, n,$$

is

$$\overline{e} = c \cdot \overline{x} + d.$$

Proof.

Write down the left and side of the equations and begin with the definition:

$$\begin{aligned}
\overline{e} &\overset{def.}{=} \frac{1}{n}\sum_{i=1}^{n} e_i \overset{expand}{=} \frac{1}{n}(e_1 + e_2 + \ldots + e_n) \\
&\overset{replace}{=} \frac{1}{n}((cx_1 + d) + (cx_2 + d) + \ldots + (cx_n + d)) \\
&\overset{group}{=} \frac{1}{n}(c(x_1 + x_2 + \ldots + x_n) + nd) \\
&\overset{simplify}{=} \frac{c}{n}\sum_{i=1}^{n} x_i + \frac{n}{n}d \\
&= c \cdot \overline{x} + d.
\end{aligned}$$

□

Change of working origin and unit

- The following helps to transform data to a new working origin, a, and a new working unit, h:

$$d_i = \frac{x_i - a}{h} = \frac{1}{h}x_i - \frac{a}{h}, \quad i = 1, 2, \ldots, n.$$

Thus,

$$\overline{d} = \frac{1}{h}\overline{x} - \frac{a}{h}.$$

Solving for $\overline{x}$ yields

$$\overline{x} = h\overline{d} + a.$$

Mean vs Median

- If the data is symmetric, the mean will approximately equal the median.
- Median is robust against outliers and incorrect readings, whereas mean is not.

Trimmed Mean

- Mean is not robust against outliers, so one way to get around that is to exclude (or trim) a certain proportion of observations from each end.
- e.g. 5% trim means the top and bottom 5% is excluded and the rest is averaged
- In R, you can specify a fraction (0 to 0.5) to the trim argument mean()

```
c(mean(x), mean(x, trim = 0), mean(x, trim = 0.25))
```

The Mode

- The mode is the most common value in the dataset or the value with the highest frequency
- Often used for categorial data

Variance and standard deviation

- We need a measure on how spread our or compressed the data is.
- If our mean is the 'target', we prefer our data to be compressed around the mean.

Squared Deviations

- The Squared deviation from a constant a is:

$$S(a) = \sum_{i=1}^{n} (x_i - a)^2$$

- To see what value would minimise this function we differentiate it:

$$\begin{aligned} S(a) &= \sum (x_i^2 - 2ax_i + a^2) = \sum x_i^2 - 2a\left(\sum x_i\right) + n \cdot a^2 \\ &= \sum x_i^2 - 2an\overline{x} + na^2 \\ \Rightarrow \frac{\partial S(a)}{\partial a} &= S'(a) = -2n\overline{x} + 2na. \end{aligned}$$

- $S'(a)$ equals 0 if $a = \overline{x}$.

- The mean minimises the function

(Sample) Variance and Standard Deviation

- For data $x_1, x_2, \ldots, x_n$, the sample variance s^2 is:

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

- The sample standard deviation is:

$$s = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- An alternative for s^2:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2\overline{x}\cdot x_i + \overline{x}^2)\\ &= \sum_{i=1}^{n}x_i^2 - 2\overline{x}\sum_{i=1}^{n}x_i + n\overline{x}^2\\ &= \sum_{i=1}^{n}x_i^2 - 2\overline{x}\cdot n\overline{x} + n\overline{x}^2, \quad \text{with } \sum x_i = n\overline{x};\\ &= \sum_{i=1}^{n}x_i^2 - n\overline{x}^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\end{aligned}$$

- Hence,

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\right).$$

Changing of working origin and unit

- If we transform data to a new working origin a and a new working unit h i.e.

$$d_i = \frac{x_i - a}{h} \quad \text{or equicalently} \quad x_i = h\cdot d_i + a$$

- Then the variance of $x_1, x_2, \ldots, x_n$, s_x^2 is equal to h^2 times the variance od $d_1, d_2, \ldots, d_n$, s_d^2 i.e.

$$s_x^2 = h^2 \times s_d^2.$$

- Therefore, $s_x = hs_d$.
- Notice that a plays no part in the result

NOTE: In an exam, use 2 more significant figures than what is in the data.

TOPIC 2 – Probability

Classical (unconditional) probability

- Classical probability is used for a random experiment in which the possible outcomes are equally likely.
- We can define the probability for certain events.

Sample Spaces

- The set of all possible outcomes of an experiment is called a sample space
- We use the notation Ω
  - Coin: Ω = {H, T};
  - Dice: Ω = {1,2,3,4,5,6}
  - Weight: $\Omega = R^+$
- Weight is different as it has an infinite number of outcomes
- An empty set is denoted as Ø, i.e. Ø = { }.

Counting

- We want to count objects for any event.
  - Any subset A of the sample space Ω, denoted by $A \subset \Omega$ is called an event
  - Any event that can be broken into other events is called a compound event
- A non-compound event is called a simple or elementary event.
- The counting operator N(A) is a set function that counts how many elements belong to the set A. E.g.

  Coin: $\Omega = \{H, T\} \Rightarrow N(\Omega) = 2;$

  Dice: $\Omega = \{1, 2, 3, 4, 5, 6\} \Rightarrow N(\{1, 2, 5\}) = 3;$

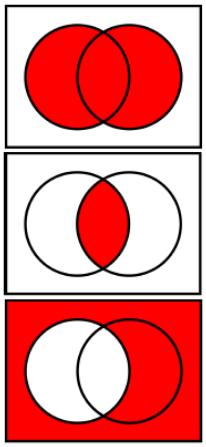
  Weight: $\Omega = \mathbb{R}^+ \Rightarrow N(\mathbb{R}^+) = \infty$
- **Component Experiment** -> To count the sample space for a component experiment we can use the multiplication principle: $N(\Omega_2) = N(\Omega_1) \times N(\Omega_1)$

Notation in set theory and probability

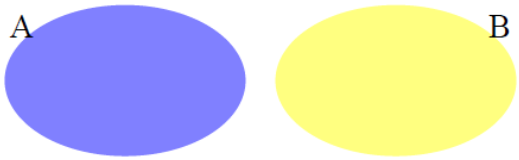
Symbol	set theory	probability
$A \cup B$	union A and B	event A or B
$A \cap B$	intersection of A and B	event A and B
$A^C = \Omega \backslash A$	complement of A	not event A
Ω, $\emptyset$	largest, empty set	certain, impossible event

- A Venn diagram is commonly used to illustrate some concepts in set theory

- Denote a Union as $A \cup B$ which means $\omega \in A \cup B \Rightarrow \omega \in A$ or $\omega \in B$.
- Denote an Intersection as $A \cap B$ which means $\omega \in A \cap B \Rightarrow \omega \in A$ and $\omega \in B$.
- Denote a Complement of A as A^C, so that $\omega \in A^C$ means that $\omega \in \Omega$ but $\omega \notin A$.



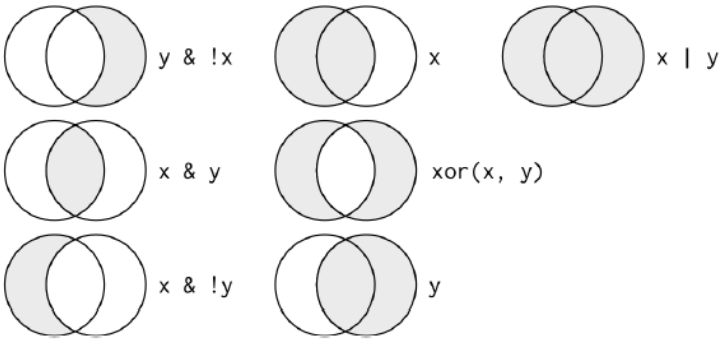
- If $A \cap B = \emptyset$ (the empty set or an impossible event), then A and B are mutually exclusive (me).



- It is convenient to use the notation: $A \setminus B = A \cap B^c$ (read as A and not B)

Comparison and Logical Operators in R

- Set operations such as union and intersection are also referred to as logical/Boolean operators.
- In R, they are used to combine different test or comparison conditions and allow us to express complex logic.
- **Comparisons in R**: R provides the standard suite: >, >=, <, <=, != (not equal) and == (equal).
- You can then combine various comparison operators with Boolean operators: & is “and”, | is “or”, and ! is “not”.



Filter()

- We can use the filter() function in tidyverse in order to subset observations based on their values.
- The first argument is the name of the data frame

- The second and subsequent arguments are the expressions that filter the data frame.
- Using email50 as an example:

```
filter(email50, spam == "no", number == "small", line_breaks < 50)
```

- This will only display the conditions specified in the function.
- Alternatively, you can use pipes %>%:

```
email50 %>% filter(spam == "no", number == "small", line_breaks < 50)
```

What is Probability?

1. **Subjective Probability** expresses the strength of one's belief.
2. **Classical Probability** has equally likely outcomes
    a. Suppose there are '*n*' equally likely possibilities of which one much occur
    b. '*s*' of these outcomes are regarded as favourable (=success)
    c. Then the probability P of a success is given by s/n.
3. The frequency interpretation of probability:
    a. The probability of an event (or outcome) is the proportion of times the event occurs in a long run of repeated experiments.

Mathematical Formulation of Probability

- Given a sample space Ω, an event A ⊂ Ω, we define P(A), the probability of A to be a value of a non-negative additive set function that satisfies the following three axioms:

  A1 For any event A, $P(A) \geq 0$,

  A2 $P(\Omega) = 1$,

  A3 If A and B are mutually exclusive events, then
  $$P(A \cup B) = P(A) + P(B)$$

  A3' If A_1, A_2, A_3, ..., is a finite or infinite sequence of mutually exclusive events in Ω, then
  $$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

- **Addition Theorem:** If A and B are any events in Ω, then:
  $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Further Properties of Probability

- Using the 3 Axioms, we can obtain the following results:

  $P(A^C) = 1 - P(A)$ since

  $$A \cap A^C = \emptyset \quad \Rightarrow \quad 1 = P(\Omega) = P(A \cup A^C) = P(A) + P(A^C)$$

  $P(\emptyset) = 0$ because

  $$\emptyset = \Omega^C \quad \Rightarrow \quad P(\emptyset) = 1 - P(\Omega)$$

  If $A \subset B$, then $P(A) \leq P(B)$.

  - Since $B = (A \cap B) \cup (A^c \cap B)$, and $(A \cap B)$, $(A^c \cap B)$ are mutually exclusive.
  - So if we apply axiom 3, we have

    $$P(B) = P(A \cap B) + P(A^c \cap B).$$

  - As $A \subset B$, we have $A \cap B = A$ and

    $$P(B) = P(A) + \underbrace{P(A^c \cap B)}_{\geq 0} \geq P(A).$$

  For any event A, $0 \leq P(A) \leq 1$ as $A \subset \Omega$ and $P(\Omega) = 1$.

  If $A_1, A_2, \ldots, A_k$ are mutually exclusive such that $\cup_{i=1}^k A_i = \Omega$ (we call this exhaustive), then

  $$P(\cup_{i=1}^k A_i) = \sum_{i=1}^{k} P(A_i) = P(\Omega) = 1.$$

Three Set Union and Beyond

- For any event A, B, C in Ω:

  $$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

More on Counting

- The elementary process of counting can become quite sophisticated.
- Methods of counting are used in order to construct probability assignments on finite sample spaces, although they can be used to answer other questions as well.
  - For example: To win the Lotto, you have to correctly pick 6 numbers out of the numbers 1 to 45. To be able to calculate the probability of winning, we first must count how many groups of 6 numbers can be chosen.

Factorial Notation

- By definition:

  $$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots 2 \cdot 1.$$

- Recall that 0! = 1

- The factorial notation is used to answer the total number of ways to select without replacement of *n* distinct objects.

Permutations

- nPr is used to compute the number of ordered arrangements, called permutations, or r objects selected from *n* distinct objects **without replacement**.
- It is given by:

$$^nP_r = n(n-1)\dots(n-r+1) = \frac{n!}{(n-r)!}.$$

Combinations

- The number of distinct subsets or combinations of size *r* that can be selected from *n* objects **without replacement** is given by: (unordered)

$$^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

- Note that:

$$^nC_r = {}^nC_{n-r}$$

Sampling in R

- We first create the sample space:

```
n <- 158
x <- 1:n
set.seed(6) # set random seed to 6 to reproduce results
```

- To obtain a random permutation of numbers 1, 2, …, 158:

```
sample(x)
```

- To choose 10 numbers without replacement:

```
sample(x, size = 10)
```

- To choose 10 numbers with replacement:

```
sample(x, 10, replace = TRUE)
```

```
email50 %>% sample_n(2)
```

- If you want to sample *n* rows of data from a data matrix, you can use the sample_n() function.
- Suppose a lotto type barrel contains 10 balls numbered 1, 2, …,10. Three balls are drawn.
    - How many unordered distinct samples can be drawn (without replacement)?

$$^{10}C_3 = \binom{10}{3} = \frac{10 \times 9 \times 8}{1 \times 2 \times 3} = 120.$$

    - What is the probability of event A: all numbers less than or equal to 7?

$$^7C_3 = \binom{7}{3} = \frac{7 \times 6 \times 5}{1 \times 2 \times 3} = 35.$$

$$\text{Therefore } P(A) = \tfrac{35}{120} = \tfrac{7}{24}.$$

- What is the probability of Event B: all numbers are even?

$$P(B) = \frac{\binom{5}{3}}{120} = \frac{10}{120} = 1/12$$

- Union of A and B

  $P(A \cup B)$?

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{7}{24} + \frac{1}{12} - \frac{1}{120} \\ &= \frac{44}{120} = \frac{11}{30} = 0.3667. \end{aligned}$$

- Intersection of A and B

  $P(A \cap B)$?

$$A \cap B = \{2, 4, 6\} \Rightarrow P(A \cap B) = 1/120.$$

Partitioning

- The number of ways in partitioning *n* distinct objects onto *k* groups containing n_1, n_2, … n_k objects, respectively, given by:

$$\frac{n!}{n_1! n_2! \ldots n_k!}$$

  Where:

$$n = n_1 + n_2 + \ldots n_k.$$

- Notice that the number of ways of partitioning things into two groups is just the same as the number of combinations between them.
- This is also called "multinomial coefficients"

Conditional Probability

Motivating/demotivating question

- Probability depends on the underlying sample space Ω.
- If it is unclear what sample space the event A refers to, we make it clear by writing:

  $P(A|\Omega)$ instead of $P(A)$
- This is read as "the conditional probability of A given Ω"
- If A and B are events in Ω, then the conditional probability of A given B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- From the definition, we can see conditional probabilities have the following properties:
  - $0 \leq P(A|B) \leq 1$;
  - $P(A|B) = 0 \Rightarrow P(A \cap B) = 0 \Rightarrow A, B$ are m.e.
  - $P(A|B) = 1 \Rightarrow P(A \cap B) = P(B) \Rightarrow B \subseteq A$.

General Multiplication Rule of Probability

- If A and B are any two events in Ω, then
$$P(A \cap B) = P(B) \times P(A|B), \quad \text{if } P(B) \neq 0.$$
- If we switch A and B yields:
$$P(A \cap B) = P(A) \times P(B|A), \quad \text{if } P(A) \neq 0.$$
- These hold because
$$P(A|B) := \frac{P(A \cap B)}{P(B)} \text{ etc.}$$

Statistical Independence

- If A and B are any two events in a sample space, we say that **A is independent of B** if and only if P(A|B) = P(A).
- From the general multiplication rule it follows that if P(A|B) = P(A), then P(B|A) = P(B) and we can simply say that A and B are independent.
- From the special multiplication rule that A and B are **independent if and only if**:
$P(A \cap B) = P(A) \cdot P(B)$.
- Notice that if A and B are independent then A and B^c are also independent.
- But if A and B are mutually exclusive, then A and B are dependent.
- Knowing A has occurred means B can't occur, and that is information between A and B. Thus they are not independent.

Law of Total Probability

- Suppose we have two events A and B, then:
$$\begin{aligned} P(A) &= P((A \cap B) \cup (A \cap B^C)) \\ &= P(A \cap B) + P(A \cap B^C) \\ &= P(A|B)P(B) + P(A|B^C)P(B^C). \end{aligned}$$
- This is equivalent to:
$$P(B) = P(B|A)P(A) + P(B|A^C)P(A^C)$$
- This can be extended into more sophisticated partitions:
  - B_j, $j = 1, 2, \ldots, n$ are mutually exclusive with $P(B_j) > 0$ and $\cup_{j=1}^{n} B_j = \Omega$;
  - A_i, $i = 1, 2, \ldots, m$ are mutually exclusive with $P(A_i) > 0$ and $\cup_{i=1}^{m} A_i = \Omega$.
- Then:
$$P(A_i) = \sum_{j=1}^{n} P(A_i|B_j)P(B_j), \quad \text{for } i = 1, 2, \ldots, m.$$

Tree Diagrams

- Tree diagrams are a tool to organise outcomes and probabilities around the structure of the data.
- They are most useful when two or more processes occur in a sequence and each process is conditioned on its predecessors.

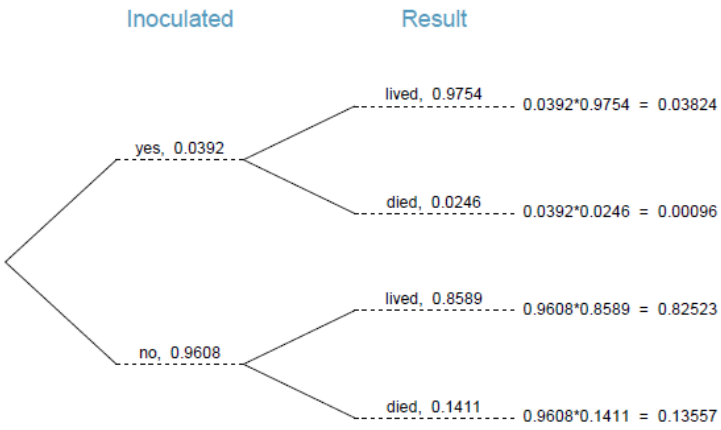
Cross-tab in R

- To produce a cross tab, we can use the *tabyl()* function in the janitor package (rather than the *table()* function which isn't in tidyverse)
- For example:

```
library(janitor)
smallpox %>%
  tabyl(inoculated, result) %>%
  adorn_totals(where=c("row","col")) %>%
  adorn_percentages(denominator = "all") %>%
  adorn_rounding(digits = 4) %>%
  adorn_ns() %>%
  knitr::kable() #optional
```

inoculated	died	lived	Total
no	0.1356 (844)	0.8252 (5136)	0.9608 (5980)
yes	0.0010 (6)	0.0382 (238)	0.0392 (244)
Total	0.1366 (850)	0.8634 (5374)	1.0000 (6224)

- This data structure is generally reflected in a tree diagram.
- The above table is visualised as:



- Tree diagrams are often annotated with marginal, conditional and joint probabilities.
- The joint probabilities can be computed using the general multiplication rule:

For instance,

$$\begin{aligned} &P(\text{inoculated} = \text{yes and result} = \text{lived}) \\ &\quad = P(\text{inoculated} = \text{yes}) \times P(\text{result} = \text{lived} | \text{inoculated} = \text{yes}) \\ &\quad = 0.0392 \times 0.9754 = 0.0382 \end{aligned}$$

Bayes' Rule

- Suppose we have two events A and B, then:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

- Together with the Law of total probability it becomes:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^C)P(B^C)}.$$

- It extends to:

$$P(B_j|A_i) = \frac{P(A_i|B_j)P(B_j)}{\sum_{j=1}^{n} P(A_i|B_j)P(B_j)}$$

  - if we allow more sophisticated partitions as before i.e. B_1, B_2, ..., B_n are mutually exclusive events such that $B_1 \cup B_2 \cup \ldots \cup B_n = \Omega$

TOPIC 3 – Discrete Random Variables

Discrete (integer-valued) random variables

- Many observed numbers are the **random** results of many possible numbers
- A random variable X is a real-valued function of the elements of a sample space Ω (i.e. X : Ω → R)
- We can map elements on the sample space on the real line
- For instance, when we flip a coin twice we can map:

$$X = \begin{cases} 0, & \text{if the results is two Heads} \\ 1, & \text{if the results is one Head \& one Tail} \\ 2, & \text{if the results is two Tails} \end{cases}$$

- Note that such functions are denoted with upper-case letters, (e.g. *X*) and their outcomes are denoted with lower-case letters (e.g. *x*)
- The domain of a random variable is the set of all possible values it can take.

Distribution of Random Variables

- The probability distribution of a discrete random variable X is the lost of possible values of X together with their probabilities:

$$p_i = \mathrm{P}(X = i) \geq 0 \quad \text{and} \quad \sum_i p_i = 1.$$

- The set of p_i is often denoted as the probability mass function (pmf) of the discrete random variable X.
- The pmf (and subsequently the probability density function (pdf)) can also be denoted as *f(x)*.

EXAMPLE:

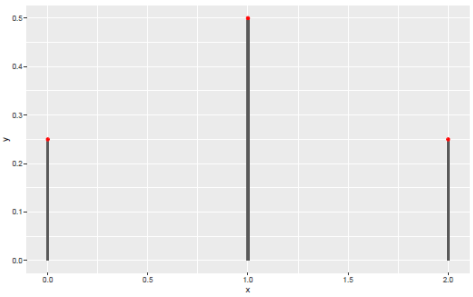
- This **experiment** consists of noting the genders of the children in a two-child family.
- The **response** random variable X is the number of female children.

$$\Omega = \{\mathrm{MM},\ \mathrm{MF},\ \mathrm{FM},\ \mathrm{FF}\} \quad \& \quad X = \{0, 1, 2\}.$$

  - i.e. X can take values 0, 1, 2.
- Let's say:

$$p_0 = P(X = 0) = 0.25, \quad p_1 = P(X = 1) = 0.50,$$
$$p_2 = P(X = 2) = 0.25, \quad P(X = x) = 0, \quad \text{for all other } x.$$

  - It is important to formally state what happens to all other x in the real field.
- For the random variable, you can summarise it by:
  - a graph



ise

  - (A bar chart of the probability, with ZERO width)

Cumulative Distribution Function

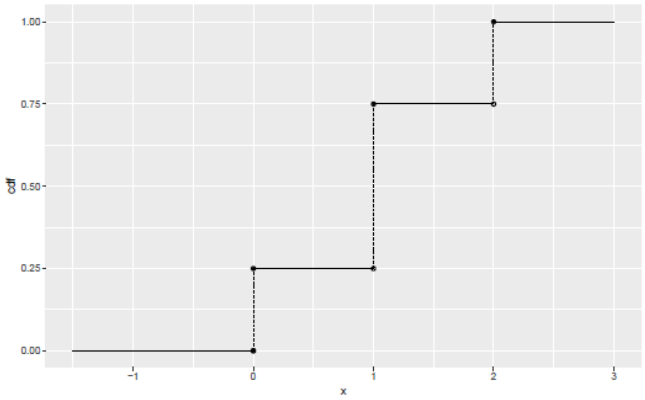
- The probability that a value of a random variable X is less than or equal to x, that is:

  $$F(x) = \mathrm{P}(X \leq x),$$

  is called the cumulative distribution function (cdf) or just the distribution function).
- In the discrete case, we have:

  $$F(x) = \mathrm{P}(X \leq x) = \sum_{i \leq x} P(X = i)$$
- In the previous examples, the corresponding cdf would be:

  $$F(x) = \begin{cases} 0, & x < 0; \\ 0.25, & 0 \leq x < 1; \\ 0.70, & 1 \leq x < 2; \\ 1, & x \geq 2. \end{cases}$$
- And graphically:



- We use filled and unfilled circles to represent the value is included or excluded respectively.

Bernoulli Trials and Binomial Distribution

Bernoulli Trials

- Bernoulli trials satisfy the following assumptions:
  - There are only two possible outcomes for each trials
  - The probability of success is the same for each trial
  - The outcomes from different trials are independent
  - There are a fixed number *n* of Bernoulli trials conducted

Binomial Distribution

- We can generalise the result of a Bernoulli trial for any $n \geq 1$ and any probability p between 0 and 1.
- The probability distribution of the number of successes X = i in n independent Bernoulli trials is called the **binomial distribution**.

$$p_i = \mathrm{P}(X = i) = \binom{n}{i} p^i (1-p)^{n-i}.$$

- To denote that the random variable X has the binomial distribution with parameters n and p we write $X \sim B(n,p)$
- In this notation, we have a success probability p and number of trials n.
- To see that $\sum p_i = 1$, recall the binomial expansion:

$$(a+b)^n = \sum_{i=1}^{n} \binom{n}{i} a^i b^{n-i} = \binom{n}{0} a^0 b^n + \binom{n}{1} a^1 b^{n-1} + \ldots + \binom{n}{n} a^n b^{n-n}.$$

- If we reverse engineer it:

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = \Big(p + (1-p)\Big)^n = 1^n = 1.$$

EXAMPLE:

- Roll a fair dice 9 times. Let X be the probability of sixes obtained.
- Then X ~ B(9,1/6), that is:

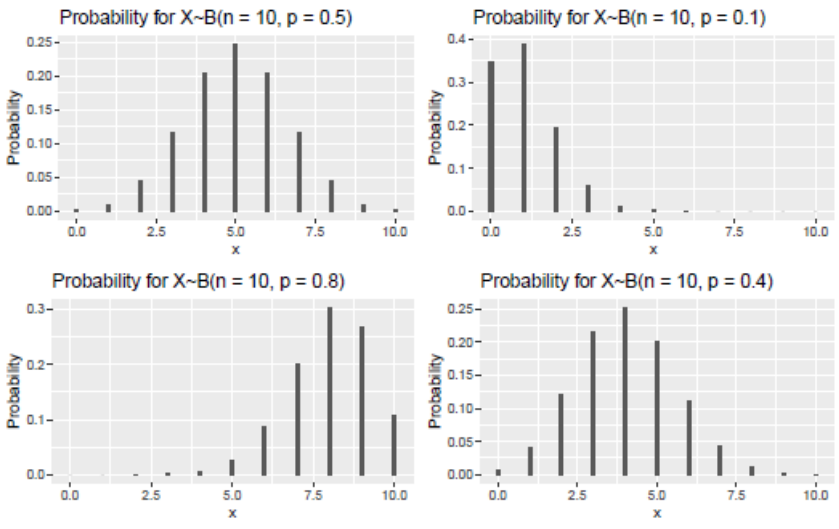
$$p_i = \mathrm{P}(X = i) = \binom{9}{i} \left(\frac{1}{6}\right)^i \left(1 - \frac{1}{6}\right)^{9-i} = \binom{9}{i} \frac{5^{9-i}}{6^9}$$

- In order to evaluate this in R:

```
n <-  9
p <- 1/6
round(dbinom(0:n, n, p), 4)
```

Shape of the Binomial Distribution

- The binomial distribution is "centered" at n × p
- The closer p is to ½, the more symmetric the distribution/histogram is.
- The larger n is, the closer to the shape is to bell (normal)



- To produce this in R:
- So we get a binomial distribution if:
    - We are counting something over a fixed number of trials or repetitions
    - The trials are independent
    - The probability of the outcome of interest is constant across trials

Mean and Variance of Distribution

Mean of a Distribution

- For a random variable X taking values 0, 1, 2, … with P(X=i) = p_i , i = 0, 1, 2, …, the mean or **expected value** of X is defined to be:

$$\mathrm{E}(X) = \sum_i i \cdot p_i.$$

```
ggplot(data=data.frame(x = 0:10, bar_height = dbinom(0:10, 10, 0.5))) +
  geom_bar(aes(x = x, y= bar_height), stat = "identity", width=0.1) +
  ggtitle("Probability for X~B(n = 10, p = 0.5)") + ylab("Probability")
```

- This type of sum is referred to as a **weighted sum**: we weight the face-values *i* based on their corresponding probability p_i with $\sum_i p_i = 1$
- We also refer to E() as the expected value operator
- We often set μ = E(X).
- Interpreting E(X):
    - Long run average of observations of X
    - Center of balance of the probability mass function/density/histogram
- It can be generalised for a random variable X taking values x_1, x_2, … that:

$$\mathrm{E}(X) = \sum_i x_i \cdot \mathrm{P}(X = x_i).$$

- For any function g(x), we can define the expected value E(g(X)) with:

$$\mathrm{E}(g(X)) = \sum_i g(x_i)\mathrm{P}(X = x_i)$$

EXAMPLE:

- Two books are assigned for a statistics class: a textbook and its corresponding study guide.
- The publisher determined:
    - 20% of enrolled students do not buy either book
    - 55% buy the textbook that costs $137
    - 25% buy both books that costs $170 in total
- What is the expected revenue per student for this unit?

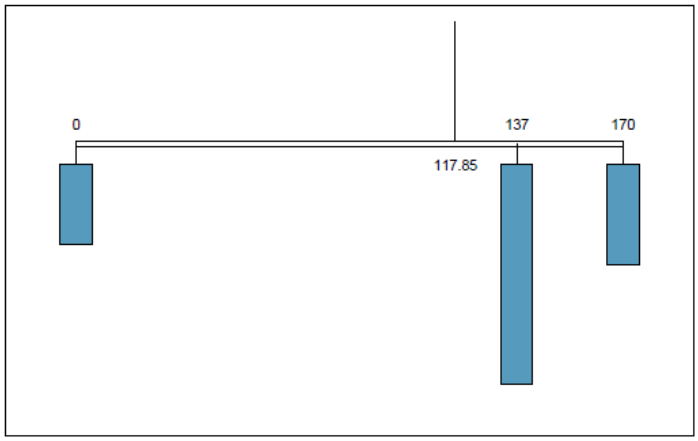
- The probability distribution table:

i	1	2	3	Total
x_i	0	137	170	–
$\mathrm{P}(X = x_i)$	0.20	0.55	0.25	1.00

- The expected value E(X) is:

$$\begin{aligned}\mathrm{E}(X) &= 0 \times \mathrm{P}(X = 0) + 137 \times \mathrm{P}(X = 137) + 170 - \times \mathrm{P}(X = 170) \\ &= 0 \times 0.2 + 137 \times 0.55 + 170 \times 0.25 = 117.85 \quad (\$).\end{aligned}$$

- $117.85 represents the average amount the publisher expects to make from a single student.
- Graphically, if we use the weight system to represent the probability distribution for X, the mean is the point of balance
    - The "string" hold the distribution at the mean to keep the system balanced.
- The mean is not necessarily an observable value:



E() is a linear operator

- For constants a and b:

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b.$$

- **Proof:**

$$\begin{aligned}\mathrm{E}(aX + b) &= \sum_{\text{all } i} g(x_i)\mathrm{P}(X = x_i), \quad \text{where } g(x) = ax + b; \\ &= \sum_{\text{all } i} \Big((ax_i + b) \times \mathrm{P}(X = x_i) \Big) \\ &= a \sum_{\text{all } i} x_i \times \mathrm{P}(X = x_i) + b \sum_{\text{all } i} \mathrm{P}(X = x_i) \\ &= a \cdot E(X) + b.\end{aligned}$$

Expected Value of X ~ B(n,p)

- If X ~ B(n,p), then E(X) = n × p.
- **Proof**:

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{i=0}^{n} i \cdot p_i = \sum_{i=0}^{n} i \times \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^{n} i \times \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \\
&= \sum_{i=1}^{n} i \times \frac{n \cdot (n-1)!}{i \cdot (i-1)!(n-i)!} p^i (1-p)^{n-i} \\
&= n \cdot p \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!((n-1)-(i-1))!} p^{i-1} (1-p)^{(n-1)-(i-1)} \\
&= n \cdot p \sum_{j=0}^{n-1} \frac{(n-1)!}{j!((n-1)-j)!} p^j (1-p)^{(n-1)-j}, \quad \text{letting } j = i-1; \\
&= n \cdot p \underbrace{\sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j}}_{\text{sums to 1 as prob. from } Y \sim B(n-1,p)} = n \cdot p.
\end{aligned}$$

EXAMPLE:

- Suppose in a quiz, there are 20 questions and each question has 5 possible answers.
- A students decides to answer the questions by selecting an answer at random.
- What is the expected number of correct answers?
  - Let X ~ B(20, 0.2)
  - Then the expected number of correct answers is E(X) = np = 4
- What is the probability that the student has more than 10 correct answers?

$$\begin{aligned}
\mathrm{P}(X > 10) &= 1 - \mathrm{P}(X \leq 10) \\
&= 1 - \Big(\mathrm{P}(X=0) + \ldots + \mathrm{P}(X=10)\Big) \\
&= 1 - 0.9994 \quad \text{using } 1-\text{pbinom}(10,\ 20,\ 0.2) \\
&= 0.0006
\end{aligned}$$

- If the marking guide stated that the student scores 4 for a correct answer but -1 for a wrong answer, what is the expected score?

$$\mathrm{E}\left[4 \times X + (-1) \times (20 - X)\right] = \mathrm{E}(5X - 20) = 5\mathrm{E}(X) - 20 = 0.$$

Variance of Distribution

- $\mathrm{E}(X^2) = \sum g(x_i) p_i = \sum x_i^2 p_i$

- The variance of the random variable X is defined by:

$$\begin{aligned}
\mathrm{Var}(X) = \sigma^2 &= \mathrm{E}(X-\mu)^2 \\
&= \mathrm{E}(g(X)), \quad \text{where } g(x) = (x-\mu)^2; \\
&= \sum_i (x_i - \mu)^2 \mathrm{P}(X = x_i) \\
&= \mathrm{E}(X^2) - \mu^2.
\end{aligned}$$

where $\mu = E(X)$ and σ^2 is a measure of spread.

- The **standard deviation** of X is σ.

- Proof:

$$\begin{aligned}\mathrm{E}(X-\mu)^2 &= \mathrm{E}(X^2 - 2\mu X + \mu^2) \\ &= \underbrace{\mathrm{E}(X^2) - 2\mu\mathrm{E}(X) + \mu^2}_{\text{expected value of a sum = sum of the expectation}} \\ &= \mathrm{E}(X^2) - 2\mu^2 + \mu^2 \\ &= \mathrm{E}(X^2) - \mu^2, \quad \text{as } E(\mu^2) = \mu^2.\end{aligned}$$

- For any constants *a* and *b*,

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

- If X ~ B(n, p), then

$$\mathrm{Var}(X) = n \cdot p \cdot (1-p)$$

  - If p = 0 or 1, then variance is 0.
  - The largest variance is when p = 0.5.

The Poisson Approximation to the Binomial

- Poisson Distribution: A theoretical model for counts which do not have a natural upper bound.
- Examples:
  - The number of accidents, crashes, etc
  - Radioactivity measured by Geiger counter
  - Number of soldiers killed by horse-kicks each year
  - Rare events (meteorites, earthquakes)
- The Poisson distribution can be seen as the limiting distribution of B(n,p).

▶ Let $n \to \infty$, $p \to 0$ (as the event is very rare), but $np = \lambda \in (0, \infty)$
i.e. the average number of rare events stays constant.

▶ For $X \sim B(n,p)$, we know that

$$\mathrm{P}(X=k) = \underbrace{\binom{n}{k}p^k}_{=(\star)}\underbrace{(1-p)^{n-k}}_{=(\star\star)}.$$

▶ Then

$$\begin{aligned}(\star) &= \binom{n}{k}p^k = \binom{n}{k}\frac{\lambda^k}{n^k} = \frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{\lambda^k}{k!} \\ &= \frac{n}{n}\cdot\frac{n-1}{n}\cdots\frac{n-k+1}{n}\frac{\lambda^k}{k!} \to \frac{\lambda^k}{k!}\end{aligned}$$

▶ To see $\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$, we first recall the Taylor's series expansion

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$$

is valid for $|x| \le 1$ and $x \neq -1$.

▶ Consider taking log := $\log_e$

$$\begin{aligned}n\log\left(1-\frac{\lambda}{n}\right) &= n\left(-\frac{\lambda}{n} - \frac{\lambda^2}{2n^2} - \frac{\lambda^3}{3n^3} - \dots\right) \\ &= -\lambda - \frac{\lambda^2}{2n} - \frac{\lambda^3}{3n^2} - \dots \\ &\to -\lambda, \quad \text{as } n \to \infty.\end{aligned}$$

- **Approximation is good if np^2 is small**
- Therefore:

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad \text{for } k = 0, 1, 2, \ldots.$$

- We denote this as X ~ P(λ)
- This implies that these events:
  - Occur randomly over time but at a constant rate per unit of time
  - Each occurrence is independent of each other
- Property of the Poisson distribution: If X ~ P(λ):

$$E(X) = \lambda, \quad \text{and} \quad \text{Var}(X) = \lambda$$

EXAMPLE:

- A certain factory averages 3 industrial accidents a week.
- Let X be the number of accidents in a week, then: X ~ P(3).
- How likely is it that no accidents will occur in a week?

$$P(X = 0) = \frac{e^{-3}3^0}{} = e^{-3} = 0.0498$$

$$P(X \geq 2) = 1 - P(X \leq 1)$$

$$= 1 - [P(X = 0) + P(X = 1)] = 0.8009$$

- How likely is it that 2 or more accidents will occur in a given week?

A different time unit

- Often we have to adjust the rate based on the length of the "exposure"
- For instance, how likely is it that only one accident will occur in a four-week period?
- If the rate of accidents is on average 3 per week, that is the same as having 12 accidents over a four-week period.
- This means that if *Y* is the number of accidents in a four-week period, then Y ~ P(12) and

$$P(X = 1) = \frac{e^{-12}12^1}{1!} = 7.373 \times 10^{-5}.$$

Geometric Distribution

- The binomial distribution is just on possible integer-valued random variable.
- Suppose we have an infinite sequence of independent trials, each of which gives a success with probability *p* and failure with probability *q = 1 – p*.
- The geometric distribution with parameter *p* has probabilities for the number of failures X before the first success:

$$p_i = P(X = i) = q^i p, \quad i = 0, 1, 2 \ldots.$$

- This is denoted as **X ~ geom(p)**
- Note that the probabilities add to 1:

$$p_0 + p_1 + \ldots = p + qp + q^2 p + \ldots = p(1 + q + q^2 + \ldots)$$

$$= p\left(\frac{1}{1-q}\right) = 1$$

as p + q = 1, and we know that by induction:

- Notice that the geometric distribution can also be defined as the number of trials X required to get to (and including) the first success.

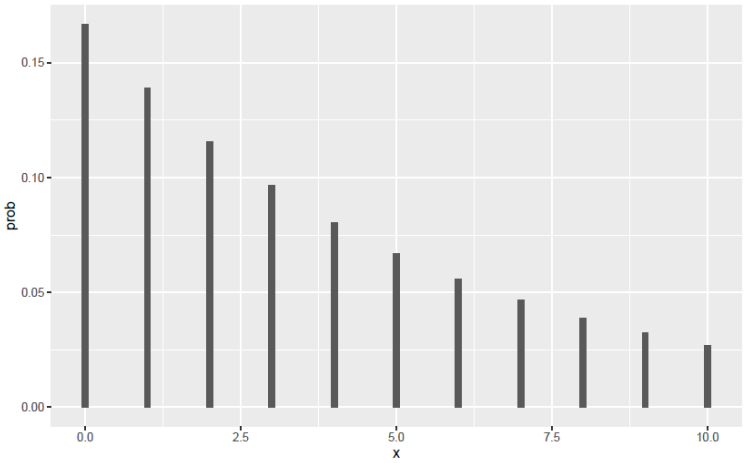
$$1 + q + \ldots + q^n = \frac{1 - q^{n+1}}{1 - q} \to \frac{1}{1 - q} \quad \text{as } n \to \infty \text{ with } |q| < 1.$$

EXAMPLE:

- A fair die is thrown repeatedly until it shows a six.
- What is the probability that more than 7 failures are required?

$$P(X > 7) = 1 - P(X \leq 6) = 1 - \sum_{i=0}^{7} \left(\frac{5}{6}\right)^i \frac{1}{6} = 0.2326 (4dp)$$

with `1-pgeom(7,1/6)` or with `1-sum(dgeom(0:7, 1/6))`.



- Is it more likely that an odd number of failures is required, or an even number?

Because $0 \leq P(X = i) \leq 1$ and $F(\infty) = 1$, we can find

$$\begin{aligned} P(\text{even}) - P(\text{odd}) &= \sum_{j=1}^{\infty} P(X = 2(j-1)) - \sum_{k=1}^{\infty} P(X = 2k - 1) \\ &= \sum_{j=1}^{\infty} (P(X = 2(j-1)) - P(X = 2j - 1)) \\ &= \sum_{j=1}^{\infty} \left(q^{2(j-1)} p - q^{2j-1} p\right) \\ &= \sum_{j=1}^{\infty} q^{2j-2} p \underbrace{(1-q)}_{\geq 0} \end{aligned}$$

  - This implies that an even number of failures (odd number od throws and one success) is more likely.

Expected Value of the Geometric Distribution

- Let X be the number of failures before the first success.

$$\begin{aligned}\mathrm{E}(X) &= \sum_{x=0}^{\infty} x \cdot pq^x = \sum_{x=1}^{\infty} x \cdot pq^x \\ &= pq\sum_{x=1}^{\infty} x \cdot q^{x-1} = pq\sum_{x=0}^{\infty} \frac{d}{dq} q^x \\ &= pq\frac{d}{dq}\sum_{x=0}^{\infty} q^x,\end{aligned}$$

interchanging the order of the sum and $\frac{d}{dq}$; ok as the sum converges

$$= pq\frac{d}{dq}\left(\frac{1}{1-q}\right) = pq \cdot \frac{1}{(1-q)^2} = \frac{q}{p}.$$

- And the variance:
  $\mathrm{Var}(X) = \frac{q}{p^2}$

Extending the Geometric Distribution

- The distribution for the number of failures before a target number of successes.
- This is known as the negative binomial distribution, denoted as X ~ NB(k,p)
- The corresponding probability function is:
  $$\mathrm{P}(X = x) = \binom{k+x-1}{k-1} p^k (1-p)^x, \quad x = 0, 1, 2, \ldots.$$
  - This is the definition R uses.
- It can be shown that the mean and variance are:
  $$\mathrm{E}(X) = \frac{k(1-p)}{p}, \quad \text{and} \quad \mathrm{Var}(X) = \frac{k(1-p)}{p^2}.$$
- The negative binomial is the sum of *k* independent geometric distributions, and that is why the mean and variance of NB is *k* times that of a geometric distribution.

EXAMPLE:

- Suppose we want the probability that 3 or more failures are needed to get 3 '6's on a standard die.
  Here $X \sim NB(3, 1/6)$ and $\mathrm{P}(X = x) = \binom{x+2}{2}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^x$,
  $x = 0, 1, 2, \ldots$.

  This means that
  $$\begin{aligned}&\mathrm{P}(X \geq 3) = 1 - \mathrm{P}(X \leq 2) \\ =&1 - (\mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \mathrm{P}(X = 2)) \\ =&1 - \left(\binom{2}{2}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^0 + \binom{3}{2}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^1 + \binom{4}{2}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2\right) \\ =&0.9645.\end{aligned}$$

  with `1-pnbinom(2, 3, 1/6)` or with `1-sum(dnbinom(0:2, 3, 1/6))`

- Just like the geometric distribution can be alternatively defined based on the total number of trials to get the first *k* success.
- Negative binomials can also be defined alternatively as the total number of trials to get *k* successes.
- The corresponding probability function is then:

$$\mathrm{P}(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, k+2, \ldots.$$

Hypergeometric distribution

- Suppose that an urn contains *N* balls, of which *r* of them are black *N* – *r* are white.
- Let X denote the number of black balls drawn when taking *n* balls without replacement.
- By using simple counting rules, we have:

$$\mathrm{P}(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}.$$

- Notice that we can draw:
  - n balls in total so $0 \leq x \leq n$;
  - at most r black balls so $0 \leq x \leq r$;
  - at most $N - r$ white balls so $n - x \leq N - r \Rightarrow x \geq n - (N - r)$
  - by combining all these inequalities we have $x = [\max(0, n - (N - r)), \ldots, \min(r, n)]$.

Properties of hypergeometric distribution

- The mean and variance of a hypergeometric distribution are:

$$\mathrm{E}(X) = \frac{nr}{N} = n \times \frac{r}{N}$$

$$\mathrm{Var}(X) = \frac{r(N-r)n(N-n)}{N^2(N-1)} = \frac{N-n}{N-1} \times \frac{r}{N}\frac{N-r}{N}$$

- If we let $p = \frac{r}{N}$, the proportion of black balls, then:

$$\mathrm{E}(X) = \frac{nr}{N} = np$$

$$\mathrm{Var}(X) = \frac{N-n}{N-1} \times np(1-p),$$

  - Note: they look similar to the mean and variance of a binomial distribution.
- The term $\frac{N-n}{N-1}$ in the variance is often referred to as the **sample without replacement factor** as the variance is reduced due to sample without replacement.

EXAMPLE:

- There are 7 vacant tutor positions to fill and there are 20 applicants.
- Of the applicants, 12 are male and 8 are female.
- Let X be the number of females being appointed.
- What is the probability that there are 3 OR 4 women being appointed?

$$\begin{aligned} \mathrm{P}(X = 3 \text{ or } 4) &= \mathrm{P}(X = 3) + \mathrm{P}(X = 4) \\ &= \frac{\binom{8}{3}\binom{12}{4} + \binom{8}{4}\binom{12}{3}}{\binom{20}{7}} = \frac{(56 \times 495) + (70 \times 220)}{77520} = 0.5562. \end{aligned}$$

Or you can calculate directly using `dhyper(3, 8, 12, 7) + dhyper(4, 8, 12, 7)`.

Multivariate hypergeometric distribution

- The hypergeometric can easily be extended to more than just white and black balls in the population.
- We will look at a simple example.
- We have 20 balls in the box: 6 are red, 4 are green, 7 are blue and 3 are orange.
- If we draw 8 balls out without replacement, the probability that we get exactly two each of the 4 colours would be:

$$\mathrm{P}(R = 2, G = 2, B = 2, O = 2) = \frac{\binom{6}{2}\binom{4}{2}\binom{7}{2}\binom{3}{2}}{\binom{20}{8}} = 0.04501.$$

TOPIC 4 – Continuous Random Variables

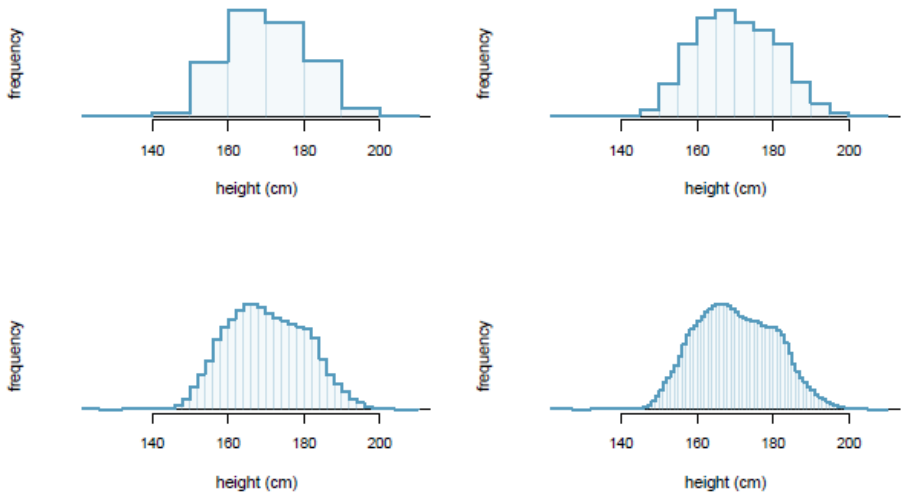
Continuous Random Variables

- These are random variables such that **any** value is possible.
- Random variables give values in the Real field (e.g. speed of a car, temperature)
- Key Property: A random variable X attains **zero probability** for any single value.
- That is:

$$P(X = x) = 0$$

Visualising the Transition

- Figures below show histograms of the height of 3 million US adults with a different number of bins.



- The more bins, the smoother the curve.

Interpreting Mathematically

- Use P(X = x) = 0 for all x ∈ R
- Focus on event X ∈ (a, b] which is an interval of length $b - a > 0$
- If X is a continuous random variable then,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

$$P(a < X \leq b) = \int_a^b \underbrace{f(x)}_{\text{=probability density function}} dx.$$

Probability Density Function

- A probability density function (pdf) (or probability density) is any non-negative function *f(x)* such that:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

- We also define a **cumulative distribution function** (cdf):

$$F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(t)\,dt.$$

- As a consequence, a cdf *F(x)* is any function that satisfies:

  (i) $0 \leq F(x) \leq 1$ (as F represents some probability)

  (ii) $F(x)$ is a non-decreasing function of x.

  (iii) If $a < b$, then
  $\mathrm{P}(a < X \leq b) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) = F(b) - F(a)$.

  (iv) $F(-\infty) = 0$, $F(+\infty) = 1$.

Scaling of Non-negative Functions

- Find *c* such that the following non-negative function is a probability density of a continuous random variable:

$$f(x) = \begin{cases} 0, & \text{for } x \leq 0; \\ ce^{-4x}, & \text{for } x > 0. \end{cases}$$

- From the definition of *f* we have:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \Rightarrow \int_{0}^{\infty} ce^{-4x}\,dx = 1.$$

Let $y = 4x \Rightarrow dy = 4dx \Rightarrow dx = \frac{1}{4}dy$, and we have

$$\int_{0}^{\infty} ce^{-4x}\,dx = \int_{0}^{\infty} \frac{c}{4}e^{-y}\,dy = \frac{c}{4}\underbrace{\int_{0}^{\infty} e^{-y}\,dy}_{=1}$$

Hence $\frac{c}{4} = 1 \Rightarrow c = 4$.

Mean of continuous random variables

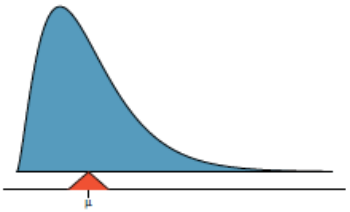
- Let *g* be any continuous function
- The expected value of g(X) of a continuous variable having probability density *f* is defined by:

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

- The expected value or the mean of X is given by

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)\,dx.$$

- The mean is the **point of balance**.



Variance of continuous variables

- The variance of X is given by:

$$\begin{aligned}\sigma^2 &= \mathrm{E}[(X-\mu)^2] \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \\ &= E(X^2) - (E(X))^2 \\ &= \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2.\end{aligned}$$

Quantiles (or Percentiles)

- The *q*th percentile is the value of x such that:

$$\int_{-\infty}^{x} f(t)\,dt = q\%.$$

Uniform Distribution

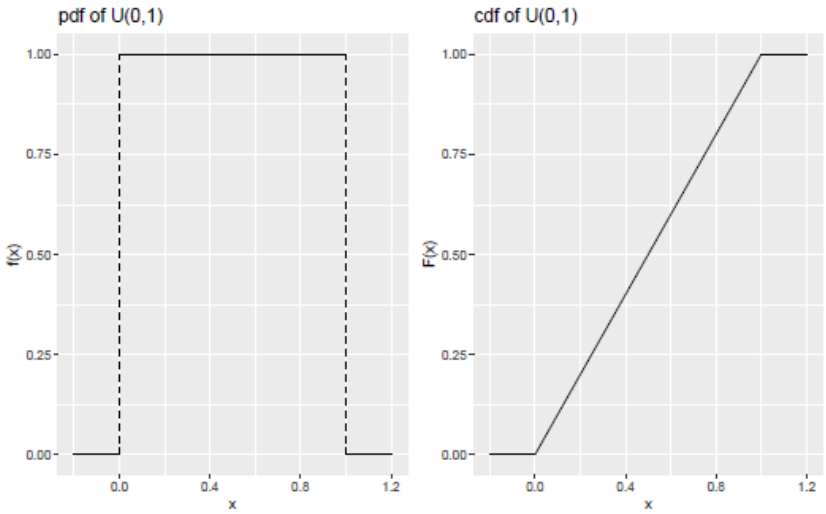
- The uniform distribution, with parameters *a* and *b* with *a* < *b* has pdf:

$$f(x) = \frac{1}{b-a} \cdot 1_{(a,b)}(x) = \begin{cases} \frac{1}{b-a}, & \text{for } a < x < b; \\ 0, & \text{elsewhere.} \end{cases}$$

- The corresponding cdf is:

$$F(x) = \begin{cases} 0, & \text{for } x \le a; \\ \frac{x-a}{b-a}, & \text{for } a < x < b; \\ 1, & \text{for } x \ge b. \end{cases}$$

- The shorthand notation is ***X ~ U(a, b)***
- The standard uniform is ***U(0, 1)***



- The function $1_A(x)$ is called the indicator function of the set A. It has value of 1 if x ∈ A and a value of 0 if x ∉ A
- The uniform distribution is potentially useful to model or to be applied in conjunction with rounding errors, generating random variables or simulation studies.

Expectation and Variance of X ~ U(a,b)

If $X \sim U(a,b)$, then $\mathrm{E}(X) = \mu = \frac{a+b}{2}$.

- Proof:

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x)\,dx \\
&= \int_a^b \frac{x}{b-a}\,dx \\
&= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_{x=a}^{b} \\
&= \frac{b^2-a^2}{2(b-a)} \\
&= \frac{(b+a)(b-a)}{2(b-a)} \\
&= \frac{a+b}{2}.
\end{aligned}$$

- One can show that $\mathrm{E}(X^2) = \frac{a^2+ab+b^2}{3}$.
- This means that

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 \\
&= \frac{a^2+ab+b^2}{3} - \left(\frac{(a+b)}{2}\right)^2 \\
&= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{4a^2+4ab+4b^2-3a^2-6ab-3b^2}{12} \\
&= \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}.
\end{aligned}$$

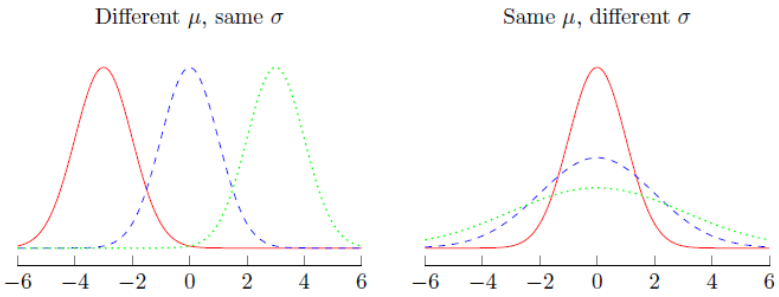
Normal Probability Density

- The normal probability density is given by:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

- To say that the random variable X has a normal distribution with parameters μ and σ^2 we write:

$$X \sim N(\mu, \sigma^2).$$



- Normal pdf is always symmetric about the μ.
- Larger σ is, the more spread out the pdf is.

Standard Normal Random Variable

- The normal with mean 0 and variance 1 is called the standard normal random variable and is generally denoted by Z.

- Thus

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

with

$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty.$$

  - The density function of the normal distribution has the shape of a symmetric bell curve.
  - Its maximum is at x = μ and it has inflection points at μ ± σ

- Let X ~ N(μ, σ^2), the centered and standardised random variable:

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

- As a result:

$$\mathrm{P}(Z \le z) := \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dz.$$

- This also means that:

$$\mathrm{P}(X \le x) = \mathrm{P}\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \mathrm{P}\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

- Using this result, we can calculate the probabilities of the standard normal distribution (available in the Standard Normal Probabilities Table)

- We often denote the *z-score* as:

$$z = \frac{x - \mu}{\sigma}$$

Standardisation

- If *X* is any random variable with mean μ and variance σ^2, then the standardised version of X is:

$$Y = \left(\frac{X - \mu}{\sigma}\right)$$

▶ If $Y = \left(\frac{X-\mu}{\sigma}\right)$ with $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then

$$\mathrm{E}(Y) = 0, \quad \text{and} \quad \mathrm{Var}(Y) = 1.$$

- **Proof:**

$$\mathrm{E}(Y) = \mathrm{E}\left(\frac{-\mu}{\sigma} + \frac{1}{\sigma}X\right) = -\frac{\mu}{\sigma} + \frac{1}{\sigma}\mathrm{E}(X) = -\frac{\mu}{\sigma} + \frac{\mu}{\sigma} = 0.$$

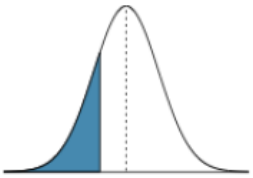
$$\mathrm{Var}(Y) = \mathrm{Var}\left(\frac{-\mu}{\sigma} + \frac{1}{\sigma}X\right) = \mathrm{Var}\left(\frac{X}{\sigma}\right)$$

$$= \frac{1}{\sigma^2}\mathrm{Var}(X), \quad \text{as } \mathrm{Var}(cX) = c^2\mathrm{Var}(X);$$

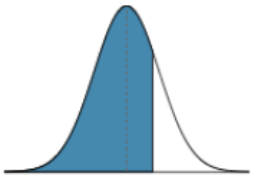
$$= 1.$$

Standard Normal Probability Table

- We use the standard normal table to calculate the probability that is linked to any particular z-score.
- The table gives the area under the standard normal curve to the left of z.



Negative z



Positive z

EXAMPLE:

- What is P(Z ≤ 0.43)?
  - Looking at the table and finding the intersection between 0.4 and 0.03, we get: $\mathrm{P}(Z \leq 0.43) = 0.6664.$
- If we are finding P(Z ≤ 0.435), we end up in the middle of two cells.
- In this case, we should take the average of the two adjacent cells.

$$\mathrm{P}(Z \leq 0.435) \approx \frac{1}{2}\left(\mathrm{P}(Z \leq 0.43) + \mathrm{P}(Z \leq 0.44)\right)$$

$$= \frac{1}{2}(0.6664 + 0.6700) = 0.6682$$

Non-standard Normal

- Suppose X ~ N(3, 2^2). Find P(X ≤ 4).
- We have:

$$Z = \frac{X-3}{2} \sim N(0,1).$$

- This means the z-score for P(X ≤ 4) is

$$z = \frac{4-3}{2} = 0.50.$$

- Therefore:

$$P(X \leq 4) = P(Z \leq 0.50) = 0.6915.$$

- If we want P(X ≥ x), we use the complement rule.
- i.e. **P(X ≥ x) = 1 - P(X ≤ x)**
- And further: **P(a ≤ X ≤ b) = P(X ≤ b) – P(X ≤ a)**

Useful Identities for the Normal

- $\phi(-z) = \phi(z)$, because the symmetry of normal, ϕ
- $\Phi(-z) = 1 - \Phi(z)$, because the symmetry of normal, ϕ, and $\int \phi(t)\, dt = 1$.
- For any $z > 0$, $P(|Z| \leq z) = 2\Phi(z) - 1$ because
  $P(|Z| \leq z) = P(-z \leq Z \leq z) = \Phi(z) - \Phi(-z)$.
- This also means that $P(|Z| \geq z) = 2(1 - \Phi(z))$.

Finding Quantiles

- Let X ~ N(5, 3^2). Find a constant *c* such that:

$$P(X > c) = 0.1.$$

- We know that:

$$0.1 = P(X > c) = 1 - P(X \leq c) \quad \Rightarrow \quad P(X \leq c) = 0.9.$$

- Then we standardise:

$$P\left(Z \leq \frac{c-5}{3}\right) = 0.9.$$

- And then we do the reverse of before to solve for z (using the table):

$$P(Z \leq z) = 0.9 \quad \Rightarrow \quad z = 1.28.$$

- Then:

$$\frac{c-5}{3} = 1.28 \quad \Rightarrow \quad c = 5 + 3 \times 1.28 = 8.84.$$

68-95-99.7 rule

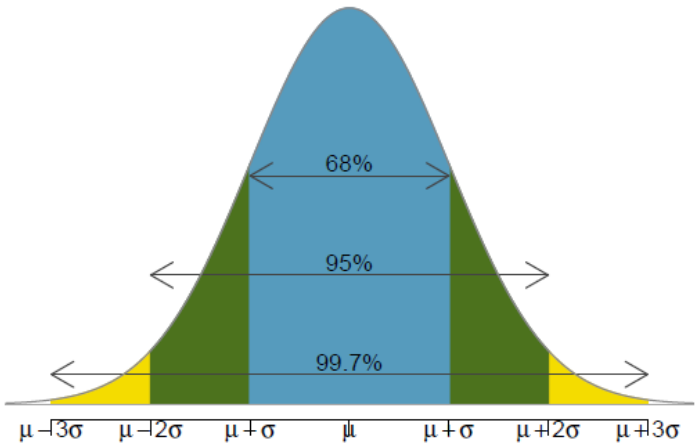
- This is a useful rule for finding the probability of falling within 1, 2 and 3 standard deviation from the mean in the normal distribution.
- This will be useful when trying to make a quick estimate without a calculator or the normal probability table.
- What percentage of data will lie within one standard deviation of the mean?

$$\begin{aligned} \mathrm{P}(\mu-\sigma < X < \mu+\sigma) &= \mathrm{P}\left(\frac{\mu-\sigma-\mu}{\sigma} \le Z \le \frac{\mu+\sigma-\mu}{\sigma}\right) \\ &= \mathrm{P}(-1 \le Z \le 1) \\ &= \mathrm{P}(Z \le 1) - \mathrm{P}(Z \le -1) \\ &\quad [\text{ or } 2\mathrm{P}(0 \le Z \le 1) = 2 \times (\mathrm{P}(Z \le 1) - 0.5)] \\ &= 0.8413 - 0.1587 \\ &= 0.6826. \end{aligned}$$

- Similarly:

$$\mathrm{P}(\mu-2\sigma < X < \mu+2\sigma) = 0.9544$$
$$\mathrm{P}(\mu-3\sigma < X < \mu+3\sigma) = 0.9974$$

- This means 68.3%, 95.4% and 99.7% of the observations from any normal distribution will lie within one, two and three standard deviations of the mean respectively.



Chebyshev's Inequality

- This inequality links the three notions of probability, mean and variance together. If a random variable X has mean μ and variance σ^2, then for any positive constant *c,*

$$\mathrm{P}(|X-\mu| \ge c\sigma) \le \frac{1}{c^2}.$$

- The proof uses the fact that P(A) = E(1_A(X)), where 1() is the indicator function. This is because:

$$E1_A(X) = \int_{-\infty}^{\infty} 1_A(x) \cdot f(x)\,dx = \int_A f(x)\,dx = \mathrm{P}(A)$$

- **Proof:**

$$
\begin{aligned}
&\mathrm{P}\left(\frac{|X-\mu|}{\sigma} > c\right) \\
=&\mathrm{E}\left(1_{\frac{|X-\mu|}{\sigma}\geq c}(X)\right) \\
\leq&\mathrm{E}\left(\left(\frac{X-\mu}{c\sigma}\right)^2 1_{\frac{|X-\mu|}{\sigma}\geq c}(X)\right), \quad \text{as } \left|\frac{X-\mu}{c\sigma}\right| \geq 1 \\
\leq&\mathrm{E}\left(\left(\frac{X-\mu}{c\sigma}\right)^2\right), \quad \text{replace } 1_A(\cdot) \text{ with } 1; \\
=&\frac{1}{c^2}\mathrm{E}\left(\left(\frac{X-\mu}{\sigma}\right)^2\right) \\
=&\frac{1}{c^2}\mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{c^2}.
\end{aligned}
$$

- We can interpret this inequality as: the probability of being more than *c* standard deviation away from the mean is at more $1/c^2$.
- Note: If you want to find the bounds for discrete distributions, you have to be careful about the end-points of the interval.

Calculations in R

- To generate samples of independent continuous random variables:

```
set.seed(2020)
n <- 200
```

- We use *rnorm()* and *runif()* to generate pseudo random numbers for standard normal and standard uniform distributions, respectively.

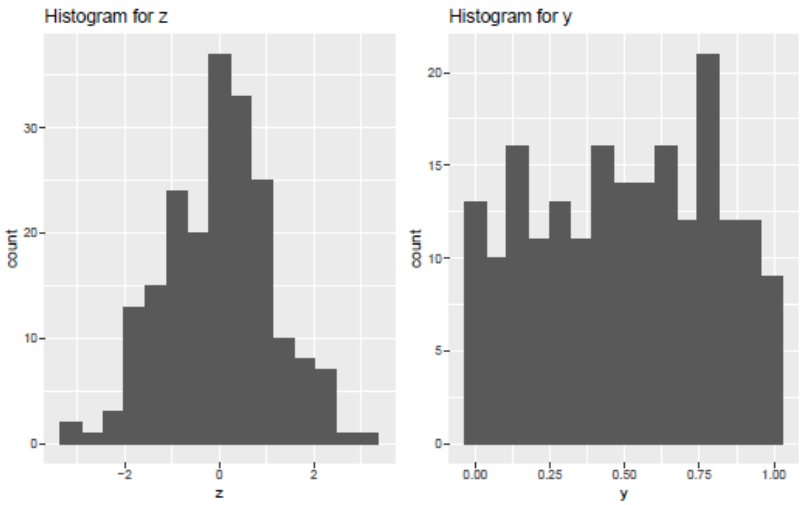
```
z <- rnorm(n)
head(z)
```

```
# [1]  0.3769721  0.3015484 -1.0980232 -1.1304059 -2.7965343  0.7205735
```

```
y <- runif(n)
head(y)
```

```
# [1] 0.2328378 0.6098221 0.8027201 0.5844282 0.3458712 0.5559263
```

```
ggplot(data = data.frame(z=z)) +
  geom_histogram(aes(x= z), bins = 15) +
  labs(subtitle = "Histogram for z")
ggplot(data = data.frame(y=y)) +
  geom_histogram(aes(x= y), bins = 15) +
  labs(subtitle(subtitle = "Histogram for y"))
```



- The *r* in front of *norm* and *unif* signifies drawing random numbers.
- For the normal distribution, we also have *rnorm(), dnorm(), pnorm()* and *qnorm()* where:
  - d signifies density
  - p signifies probability: $P(X \leq x)$
  - q returns the quantile.

```
pnorm(1.96)
```

```
# [1] 0.9750021
```

```
qnorm(0.975)
```

```
# [1] 1.959964
```

- **For non-standard normal:**

```
pnorm(95, mean = 100, sd = 10)
```

```
# [1] 0.3085375
```

```
qnorm(0.3085375, mean = 100, sd = 10)
```

```
# [1] 95
```

- To find the upper tail probability:

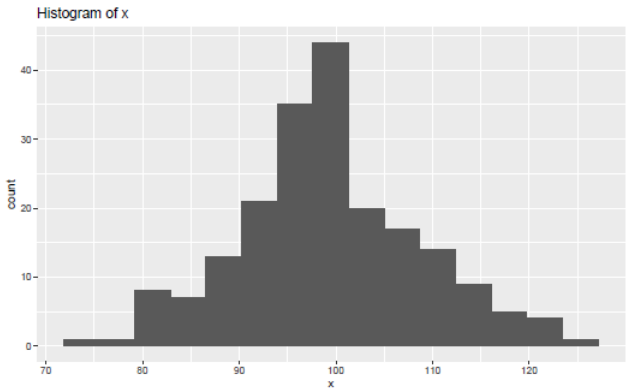
```
1- pnorm(95, mean= 100, sd=10)
```

```
# [1] 0.6914625
```

```
pnorm(95, mean = 100, sd = 10, lower.tail = FALSE)
```

```
# [1] 0.6914625
```

```r
x <- rnorm(n, mean = 100, sd = 10)
ggplot(data=data.frame(x=x)) +
  geom_histogram(aes(x=x), bins=15) + labs(title= "Histogram of x")
```



TOPIC 5 – Sampling Distribution

Sample: a sequence of random variables

- This topic is dealing with multiple random variables rather than just one.
- We have to define rules and discuss certain properties when dealing with multiple random variables at the same time.
- Let X be a real-valued random variable and x ∈ R, then:

  $A = \{X \leq x\}$

  represents an event.
- Let Y be another real valued group of random variables:

  $B = \{Y \leq y\}, \quad y \in \mathbb{R}.$

Independence of random variables

- Definition of independent of events: A and B and independent if and only if
  $P(A \cap B) = P(A)P(B)$
- Two random variables X and Y are independent if and only if for any numbers x and y, the events {X ≤ x} and events {Y ≤ y} are independent events.
- Examples:
  - $(X = \texttt{height},\ Y = \texttt{weight})$ from a random person are not independent.
  - $(X_1 = \texttt{lottery number next draw}$ and $X_2 = \texttt{lottery numbers in three weeks time})$ are independent.
  - $(X_1 = \texttt{today's rainfall}$ and $X_2 = \texttt{tomorrows rainfall})$ are not independent.

Joint Distribution

- From the independence of random variables definition, we can easily get the joint cumulative distribution function and joint probability density/mass function of independent random variables.
- Let $F_X(x) = P(X \leq x)$ and $F_Y(y) = P(Y \leq y)$ be the cumulative distribution function of the independent random variables X and Y, then the joint cumulative distribution function is
  $F_{X,Y}(x, y) := P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y).$
- If $F_{X,Y}(x,y)$ is the joint cumulative distribution function of two random variables X and Y, then $F_x(x)$ and $F_Y(y)$ are called the *marginal cumulative distribution* function of X and Y respectively.
- We can recover the marginal distribution from the joint distribution, even when they are not independent, via
  $F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y)$
- This is consistent with:
  $\lim_{y \to \infty} F_Y(y) = 1.$

Joint Densities

- Let $f_x(x)$ and $f_y(y)$ be the probability density function of the independent random variables X and Y, the joint probability density function is given by:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

Properties of E() and Var()

- Let X and Y be independent variables for any real-valued constants a and b.
  1. $\mathrm{E}(aX) = a\mathrm{E}(X)$ (independence not needed).
  2. $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$ (independence not needed).
  3. $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$ (independence not needed).
  4. $\mathrm{E}(aX+bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$ (independence not needed).
  5. $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ (independence needed).
  6. $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ (independence needed).
  7. $\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$ (independence needed).
- **Proof of property 3:**

$$\begin{aligned}
\mathrm{E}(X+Y) &= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty}(x+y)\cdot\mathrm{P}(X=x,Y=y) \\
&= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty}x\mathrm{P}(X=x,Y=y) + \sum_{x=0}^{\infty}\sum_{y=0}^{\infty}y\mathrm{P}(X=x,Y=y) \\
&= \sum_{x=0}^{\infty}x\sum_{y=0}^{\infty}\mathrm{P}(X=x,Y=y) + \sum_{y=0}^{\infty}y\sum_{x=0}^{\infty}\mathrm{P}(X=x,Y=y) \\
&= \sum_{x=0}^{\infty}x\mathrm{P}(X=x,Y\in\mathbb{N}) + \sum_{y=0}^{\infty}y\mathrm{P}(X\in\mathbb{N},Y=y) \\
&= \sum_{x=0}^{\infty}x\mathrm{P}(X=x) + \sum_{y=0}^{\infty}y\mathrm{P}(Y=y) = \mathrm{E}(X)+\mathrm{E}(Y).
\end{aligned}$$

Sampling Distribution

- Many observed phenomena can be modelled as a function of several random variables:
  - Total weight of passengers in a lift
  - Average waiting time in a queue
- These simple statistics are calculated based on a sample.
- If we take a different sample (from the same population), these statistics (and sample mean and variance) would be different too.
- This means that these statistics would be random and have their own distributions.

Mean and Variance of the Sample mean

- Let X_1, X_2, …, X_n ne n random variables that are independent, and each have the same distribution (they are identically distributed)
- They are denoted as i.i.d (independent and identically distributed)
- This means the share the same mean μ and variance σ^2.
- Then for the sample mean, that is $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ we have:

  i) mean: $\mu_{\overline{X}} = \mathrm{E}(\overline{X}) = \mu.$

  ii) variance: $\sigma^2_{\overline{X}} = \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$

  iii) standard deviation: $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$

- The standard deviation of the mean is often referred to as the standard *error* of the mean.
- **Proof:**

$$\begin{aligned}
\mathrm{E}(X_1 + \ldots + X_n) &= \mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_n) \\
&= \underbrace{\mu + \mu + \ldots + \mu}_{n \text{ times}} = n\mu \\
\mathrm{E}(\overline{X}) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \times n\mu = \mu. \\
\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n) \\
&= \underbrace{\sigma^2 + \ldots + \sigma^2}_{n \text{ terms}} = n\sigma^2. \\
\mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_1\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}.
\end{aligned}$$

Sum of Independent Random Variables

- X_1, X_2, X_3 are i.i.d random variables with distribution:

i	0	1	3
p_i	1/3	1/3	1/3

$T_2 = X_1 + X_2 \Rightarrow$

i	0	1	2	3	4	6
p_i	1/9	2/9	1/9	2/9	2/9	1/9

$T_3 = X_1 + X_2 + X_3 \Rightarrow$

i	0	1	2	3	4	5	6	7	9
p_i	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{4}{27}$	$\frac{6}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{1}{27}$

- Note: the distribution of T_3 clusters around the mean $\mathrm{E}(T_3) = 4$.

Sum of Independent Normal Variables

- It is easy to find the mean and variance of two or more independent random variables, without first finding the distribution of their sum
- However, in the case of independent normal variables, we have the following important results concerning the distribution of the sum.
- If X_1 and X_2 are independent, $X_1 \sim N(\mu_1,\sigma_1^2)$ and $X_2 \sim N(\mu_2,\sigma_2^2)$
- More generally:

$$\sum_{i=1}^{n} a_i X_i \sim N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2).$$

- Using the additive property, we obtain the following result:
  - If X_1 and X_2 are independent, then:

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2).$$

  - If all X_i are i.i.d. $N(\mu,\sigma^2)$, then

$$T = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \quad \text{and} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

EXAMPLE:

- Steel rods, made with diameter $R \sim N(4.9, 0.03^2)$, are fit into sockets, made with diameter $S \sim N(5, 0.04^2)$.
- For a satisfactory fit, the socket diameter should exceed the rod diameter, but by no more than 0.2cm.
- *If a rod and socket are taken at random, what is the probability that the fit is unsatisfactory.*
  - For a satisfactory fit, we need $S > R \rightarrow S - R > 0$ and $S - R < 0.2$.
  - The distribution for $S - R$ is:

$$S - R \sim N(5.00 - 4.90, (0.03^2 + 0.04^2)) = N(0.10, 0.05^2).$$

  - The probability for a satisfactory fit is therefore:

$$\begin{aligned} \mathrm{P}(\text{satisfactory}) &= \mathrm{P}(0 < S - R < 0.20) \\ &= \mathrm{P}\left(\frac{0 - 0.1}{0.05} < Z < \frac{0.20 - 0.10}{0.05}\right) \\ &= \mathrm{P}(-2 < Z < 2) = \mathrm{P}(Z < 2) - \mathrm{P}(Z < -2) \\ &= 0.9772 - 0.0228 = 0.9544. \end{aligned}$$

$$\mathrm{P}(\text{unsatisfactory}) = 1 - 0.9544 = 0.0456.$$

EXAMPLE:

- The tibia length of a certain species of beetle can be modelled by L ~ N(7.8, 0.3^2)
- What is the probability that the average length of 25 independent tibia lengths will be less than 7.6mm?

L_i are i.i.d. $N(7.8, 0.3^2)$, then $\overline{L} = \frac{1}{n}\sum_{i=1}^{n} L_i \sim N(7.8, \frac{0.3^2}{25})$, so $\text{Var}(\overline{L}) = 0.06^2$.

$$\begin{aligned} \text{P}(\overline{L} < 7.6) &= \text{P}\left(Z < \frac{7.6 - 7.8}{0.06}\right) \\ &= \text{P}(Z < -3.33) \\ &\leq \text{P}(Z < -2.99) \\ &= 0.0019. \end{aligned}$$

From R, $\text{P}(Z < -3.33) = 0.0004$.

- What is the probability that the average will differ from 7.8mm by more than 0.1mm?

$$\begin{aligned} \text{P}(|\overline{L} - 7.8| > 0.1) &= \text{P}(|Z| > 0.1/0.06) \\ &= P(|Z| > 1.67) \\ &= 2P(Z < -1.67) \\ &= 2 \times 0.0475 \\ &= 0.095. \end{aligned}$$

Central Limit Theorem (CLT)

- If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with (common) mean μ and variance $0 < \sigma^2 < \infty$, then

$$\text{P}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \leq x\right) \to \Phi(x) = \text{P}(Z \leq x) \quad \text{as } n \to \infty;$$

$$\text{P}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right) \to \Phi(x) = \text{P}(Z \leq z) \quad \text{as } n \to \infty.$$

- Thus, for large *n* (here n ≥ 25) the following are approximately true:

$$T = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

$$\overline{X} = \frac{1}{n} T \sim N\left(\mu, \sigma^2/n\right).$$

- The closer the distribution of X_i is to the normal, the better the approximation for small *n* values.
- CLT is an asymptotic result which means that it works better as *n* gets larger.
  - But the approximation is so good that for n ≥ 25, the common distribution of X_i's has virtually no effect on the distribution of the mean.
  - The approximation could also be good for smaller *n* if the common distribution of the X_i's does not depart grossly from normality.
- CLT works for i.i.d rvs as long as $0 < \sigma^2 < \infty$.

EXAMPLE:

- Systolic blood pressure readings for pre-menopausal, non-pregnant women aged 35-40 have a mean of 122.6 mm Hg and a standard deviation of 11 mm Hg.
- An independent sample of 25 women is drawn from this target population and their bp recorded.
- What is the probability that the average bp is greater than 125 mm Hg?
  - We know that X_i, $i = 1, 2, \ldots, 25$ are iid with $E(X_i) = 122.6$ and $Var(X_i) = 11^2$.
  - Then $\overline{X} \sim N\left(122.6, \frac{11^2}{25}\right)$ approximately by CLT.
  - This means

$$P(\overline{X} \geq 125) \simeq P\left(Z > \frac{125 - 122.6}{11/\sqrt{25}}\right)$$
$$= P(Z > 1.09) = 1 - \Phi(1.09) = 1 - 0.8621 = 0.1379.$$

- If the sample size increases to 40, what would the answer to the first part be now?

$$P(\overline{X} \geq 125) \simeq P\left(Z > \frac{125 - 122.6}{11/\sqrt{40}}\right)$$
$$= P(Z > 1.38) = 1 - \Phi(1.38) = 1 - 0.9162 = 0.0838.$$

Simulating the Effect of CLT

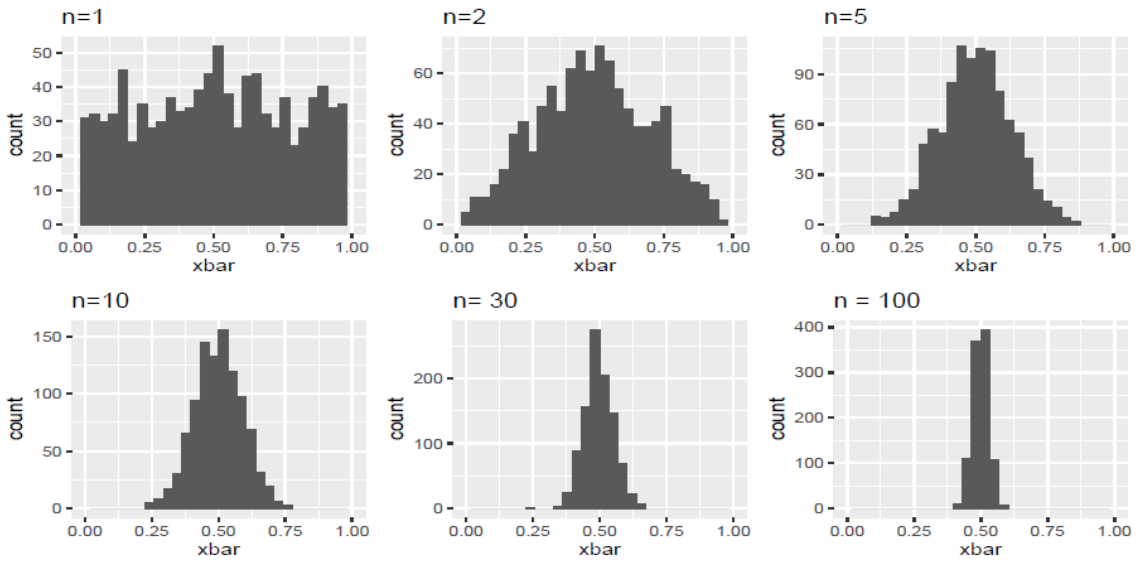
- We are going to conduct a simulation study to see how CLT works on various distributions.
- For each distribution, we are going to:
  - Obtain a sample of size *n* from it
  - Calculate the sample mean for that sample
  - Then repeat 1000 times to collect a sample of the sample mean
  - Draw a histogram based on the sample of the sample mean so that we can observe the sampling distribution of the sample mean
- Then we change *n* from 1 to larger numbers and observe how the changing sample size affects the shape of the histogram.

CLT in R

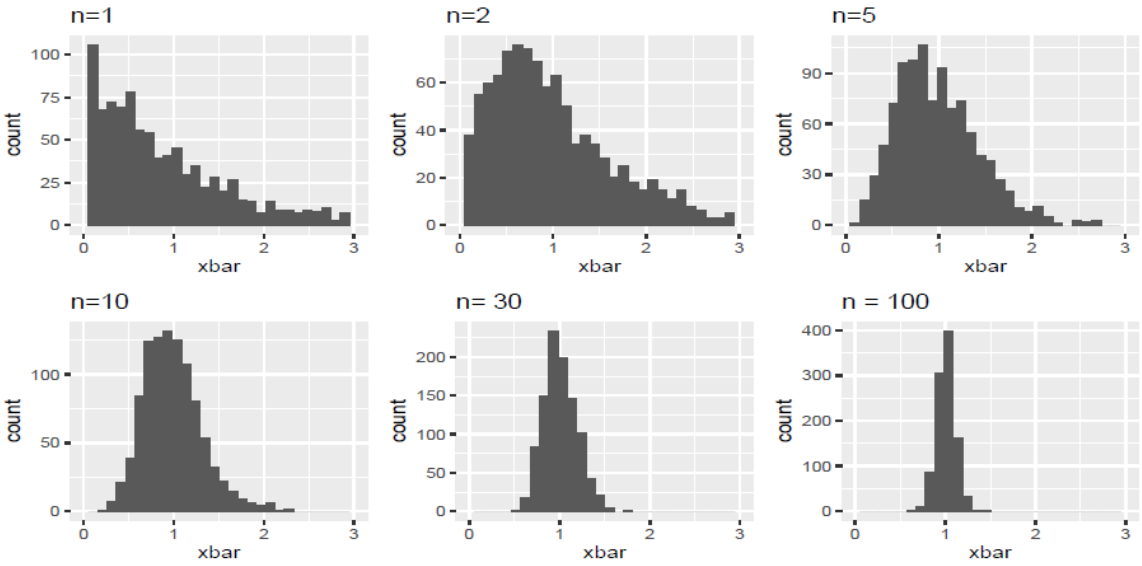
```
library(tidyverse)
library(purrr)
# Uniform, n=1
n <- 1
map(1:1000, ~ runif(n)) %>% map_df(~(data.frame(xbar = mean(.x)))) %>%
  ggplot() + geom_histogram(aes(x= xbar)) + xlim(0,1) + ggtitle("n=1")

# Exponential, n=1
1:1000 %>% map(~ rexp(n)) %>% map_df(~(data.frame(xbar = mean(.x)))) %>%
  ggplot() + geom_histogram(aes(x= xbar)) + xlim(0,3) + ggtitle("n=1")
```

- And then change *n* to {1, 2, 5, 10, 30, 100}

CLT to X ~ U(0, 1)



CLT for X ~ E(1)



Normal Approximation to the Binomial

- The CLT is applies to both continuous random variables and discrete random variables, as long as $E(X^2) < \infty$.
- Looking at the binomial distribution:
- Let X_i be independent random variables (outcomes of Bernoulli trials), defined as:

$$X_i = \begin{cases} 1, & \text{if the } i\text{th trail is a success } S, \\ 0, & \text{if the } i\text{th trail is a failure } F. \end{cases}$$

and let p = P(S) on the ith trial.

- Then $E(X_i) = p$ and $Var(X_i) = p(1 - p)$
- This means when n is sufficiently large,

$$X \overset{approx}{\sim} N(np, np(1-p)).$$

- This approximation is quite **good if $np \geq 5$ and $n(1-p) \geq 5$.**
- The closer p is to 0.5, the better the approximation is for small n.

EXAMPLE:

- Approximating P(X = 3) where X ~ B(12, 0.5)
- Suppose X ~ B(12, 0.5), then:

$$P(X = 3) = \binom{12}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^9 = 0.0537$$

- Comparing to the area approximately under the curve:

$$X \simeq Y \sim N(12 \times 0.5, 12 \times 0.5 \times (1 - 0.5)) = N(6, 3)$$
$$P(X = 3) \simeq P(3 - \lambda < Y < 3 + (1 - \lambda)); \quad \lambda \in [0, 1].$$

- Generally, we choose λ = ½ which is closer to the true value of P(X = 3) in the above example.
- Note that:

```
pnorm(3.5176, mu, sd) - pnorm(2.5176, mu, sd)
```

```
# [1] 0.05371257
```

comes very close to the *dbinom(3, 12, 0.5)* but is only optimal for this particular example.

Continuity Correction

- To approximate binomial probabilities using the normal, consider areas of corresponding rectangles.
- Adjust the normal probability statement by adding or subtracting 0.5 to the constant to increase the area under the normal curve.

$$P(X = x) \simeq P(x - 0.5 < Y < x + 0.5))$$
$$= P\left(\underbrace{\frac{x - 0.5 - \mu}{\sigma}}_{z_l} < Z < \underbrace{\frac{x + 0.5 - \mu}{\sigma}}_{z_u}\right)$$
$$= \Phi(z_u) - \Phi(z_l).$$

- For P(X ≥ x) repeat the above step noting that:

$$P(X \geq x) = \sum_{i \geq x} P(X = i) \quad \Rightarrow \quad P(X \geq x) \simeq P\left(Y \geq x - \frac{1}{2}\right)$$

- I.e. you should always aim to have an = sign in your inequality.

EXAMPLE:

- If X ~ B(12, 0.5), find P(2 ≤ X < 5)
- Approximating normal is Y ~ N(6,3), with np = n(1 – p) = 6 so the approximation is reasonable:

$$\begin{aligned} \mathrm{P}(2 \le X < 5) &= \mathrm{P}(2 \le X \le 4) = P(X \le 4) - \mathrm{P}(X \le 1) \\ &\simeq \mathrm{P}(Y \le 4.5) - \mathrm{P}(Y \le 1.5) \\ &= \mathrm{P}\left(Z \le \frac{4.5-6}{\sqrt{3}}\right) - \mathrm{P}\left(Z \le \frac{1.5-6}{\sqrt{3}}\right) \\ &= \Phi(-0.87) - \Phi(-2.60) \\ &= 0.1885. \end{aligned}$$

- The exact answer is :

```
sum(dbinom(2:4,12,0.5))
```

```
# [1] 0.1906738
```

which indicates that it is a good approximation even for small samples spaces.

EXAMPLE:

- Suppose that 80% of patients with a certain disease can be cured with a certain drug.
- What is that probability that amongst 150 patients with the disease, at most 37 of them cannot be cured with the drug?
  - Let X denote the number of patients that can be cured
  - Assume the patients respond independently:

    $X \sim B(150, 0.8).$
  - The approximating normal is:

    $Y \sim N(150 \times 0.8, 150 \times 0.8 \times 0.2) \quad \Rightarrow \quad Y \sim N(120, 24)$

    As $150 \times 0.8 = 120 > 5$ and $150 \times (1 - 0.8) = 30 > 5$, the approximation is appropriate and

$$\begin{aligned} \mathrm{P}(X \ge 113) &= 1 - \mathrm{P}(X \le 112) \simeq 1 - \mathrm{P}(Y \le 112.5) \\ &\quad (\text{alternatively } \simeq \mathrm{P}(Y \ge 112.5) = 1 - \mathrm{P}(Y \le 112.5)) \\ &= 1 - \mathrm{P}\left(Z \le \frac{112.5-120}{\sqrt{24}}\right) = 1 - \Phi(-1.53) \\ &= 0.9370. \end{aligned}$$

Proportions of Success of a Binomial

- Any binomial problem can also be worked out in terms of the proportions rather than counts
- Recall that if X ~ B(n,p), then X is the sum/total of *n* independent Bernoulli trials.
- Then if P is the proportion of successes in the sample then:

$$P = \frac{X}{n} = \frac{T}{n} \quad (\text{basically the average of those trails})$$

- This means the if the normal approximation/CLT is appropriate on *X*, it would be appropriate for P as well.

- Then:

$$\mathrm{E}(P) = \mathrm{E}\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

$$\mathrm{Var}(P) = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

- This means that:

$$P \overset{approx}{\sim} N\left(p, \frac{p(1-p)}{n}\right).$$

- We can apply continuity correction here as well
- The correction for P would be $\pm\frac{1}{2n}$

EXAMPLE:

- The proportion of children having a particular type of birth defect born to women from a particular ethnic group is 0.05.
- Calculate the probability that in 785 independent births, no more than 4% of children have the birth defect.
    - The approximating normal for the curve is therefore:
      $P \simeq Y \sim N\left(0.05, \frac{0.05\times(1-0.05)}{785}\right) = N(0.05, 0.00778^2).$
    - This approximation is appropriate as both np and n(1 – p) are greater than 5.
    - Thus:

$$\begin{aligned}\mathrm{P}(P \leq 0.04) &\simeq \mathrm{P}(Y \leq 0.04 + \frac{1}{2\times 785}) \\ &= \mathrm{P}\left(Z \leq \frac{0.03 + \frac{1}{2\times 785} - 0.05}{0.00778}\right) \\ &= \Phi(-1.20) = 0.1151.\end{aligned}$$

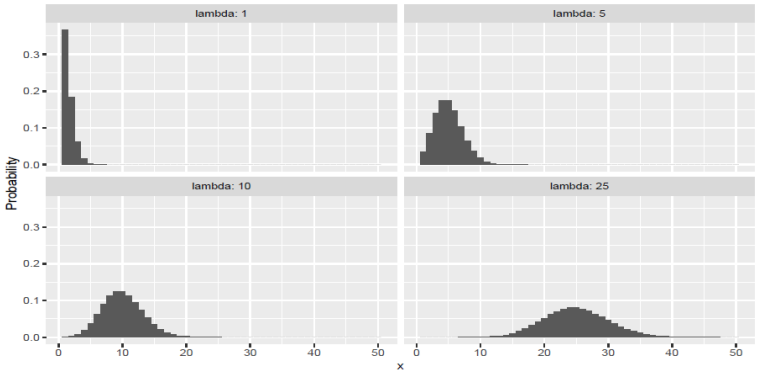
Normal Approximation to the Poisson

- The approximation is appropriate when λ is large (we set λ > 9)
- Recall that if X ~ Pois(λ) then E(X) = Var(X) = λ
- This means that when λ is large,

$$X \simeq Y \sim N(\lambda, \lambda).$$

- Poisson is another discrete distribution so continuity correction should be used.

Visualising the Approximation



- The distribution is converging to a normal distribution as λ increases

EXAMPLE:

- The number of bread rolls ordered at a bakery is random but follows a Poisson distribution with an average rate of 120 per day.
- What is that probability that between 130 and 144 rolls are ordered on a particular day?
    - Let X be a rv to represent the number of bread rolls ordered, then X ~ Pois(120)
    - The exact probability is:

$$\mathrm{P}(130 \leq X \leq 144) = \sum_{x=130}^{144} \frac{e^{-120}120^{x}}{x!}$$

    - If we use the normal approximation, then

$$X \simeq Y \sim N(120, 120).$$

    - The exact probability can be approximated by:

$$\begin{aligned}\mathrm{P}(130 \leq X \leq 144) &\simeq \mathrm{P}\left(130 - \frac{1}{2} \leq Y \leq 144 + \frac{1}{2}\right) \\ &= \mathrm{P}\left(\frac{130 - \frac{1}{2} - 120}{\sqrt{120}} \leq Z \leq \frac{144 + \frac{1}{2} - 120}{\sqrt{120}}\right) \\ &= \mathrm{P}\left(0.87 \leq Z \leq 2.24\right) \\ &= \Phi(2.24) - \Phi(0.87) \\ &= 0.9875 - 0.8078 \\ &= 0.1797.\end{aligned}$$

    - Hence, there is an 18% chance of getting between 130 and 144 orders on any particular day.
- The exact probability can be found using R:

```
sum(dpois(130:144, 120))
```

```
# [1] 0.1772363
```

Normal Approximation to the Negative Binomial

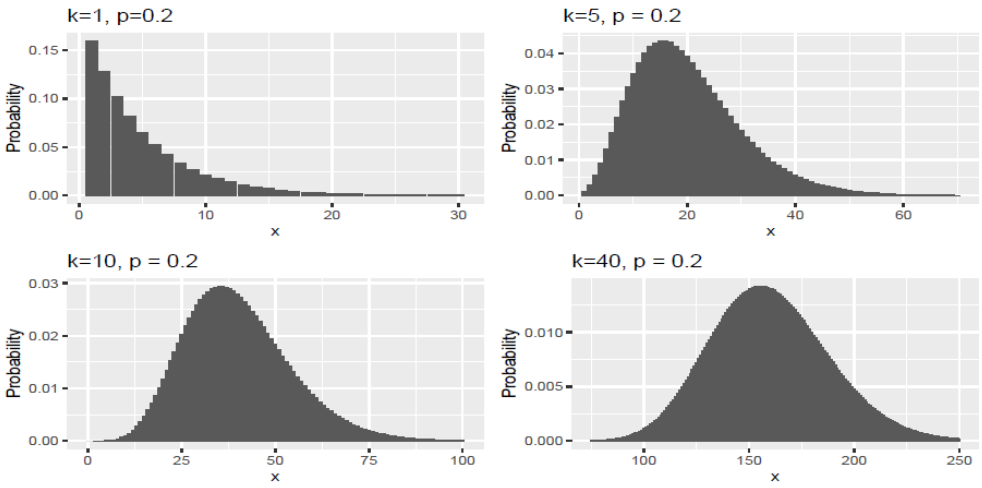
- The negative binomial can be thought of as the sum of *k* i.i.d geometric rvs (identical so they all have the same success prob. p)
- The rule of thumb for large *k* is when $k > \frac{9}{1-p}$
- If X ~ NB(k,p), then:

$$\mathrm{E}(X) = \frac{k(1-p)}{p} \quad \text{and} \quad \mathrm{Var}(X) = \frac{k(1-p)}{p^2}$$

- This means when *k* is large, the corresponding normal approximation is:

$$X \simeq Y \sim N\left(\frac{k(1-p)}{p}, \frac{k(1-p)}{p^2}\right)$$

Visualising the Approximation when p = 0.2



EXAMPLE:

- Suppose that during a game of golf, there is an 8% chance the golfer loses a ball into the water hazard.
- When the golfer loses a ball, the game is over, and the golfer has to start again.
- Assuming all attempts are independent, what is the probability that the golfer can complete 200 games prior to losing all 12 of his golf balls?
    - Let X be the number of completed games until all 12 balls are lost
    - Then X ~ NB(12,0.08)
    - The exact probability is:

$$\mathrm{P}(X \geq 200) = \sum_{200}^{\infty} \binom{x+12-1}{12-1} (0.08)^{12}(1-0.08)^{x}.$$

    - If we use the normal approximation (which is appropriate as *k* is large), then:

$$X \simeq Y \sim N\left(\frac{k(1-p)}{p}, \frac{k(1-p)}{p^2}\right) = N(138, 1725)$$

    - The exact probability can be approximated by:

$$\begin{aligned}\mathrm{P}(X \geq 200) &= 1 - \mathrm{P}(X \leq 199) \\ &\simeq 1 - \mathrm{P}(Y \leq 199.5) \\ &= 1 - \mathrm{P}\left(Z \leq \frac{199.5 - 138}{\sqrt{1725}}\right) \\ &= 1 - \Phi(1.48) \\ &= 0.0694.\end{aligned}$$

Sampling Distributions

- How do statistics vary across samples?
- Height for randomly selected *n* = 4 adult.
  What is the distribution of $\overline{X}$ and S^2?
- **Model**: Assume 4 independent readings of H ~ N(178, 8^2)

- **Observations**:

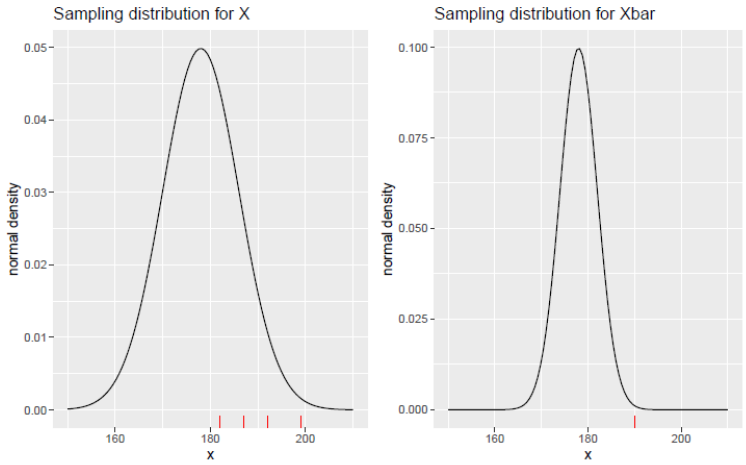
$$X_1, X_2, X_3, X_4 \quad \Rightarrow \quad \overline{X} = \frac{1}{4}\sum_{i=1}^{4} X_i$$

- **The Mean:** because $E(X_i) = 178$ and $Var(X_i) = 8^2$, it follows that:

$$\overline{X} = N(178, 4^2).$$

Interference

- Knowing the sampling distribution helps identify unusual statistics values.
- For instance, if X_{bar} (for four basketball players) and we want to see where they are in the distribution of X_i and X_{bar}:



- Using a sample, we can make inferences about unknown population parameters.
- We can conclude that there must be something special about those 4 observations (e.g. sampling may be from a different population where the average height is not 178.

TOPIC 6 – Statistical Inference

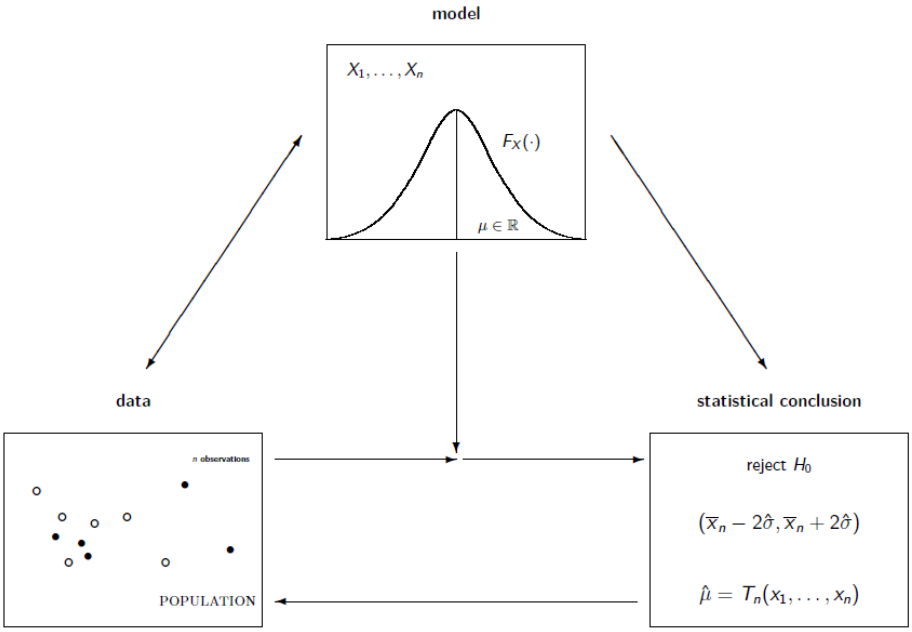
Statistical Inference

- Linking of observed data with possible statistical models
- Based on some statistical models:
    - Make decisions
    - Produce estimates
    - Make predictions

Random Sample

- Statistical inference is inference about a population from a random sample drawn from it
- A set of observations (random variables) X_1, X_2, …, X_n constitutes a random sample of size *n* from the infinite population with cumulative distribution function $F(x) = P(X \leq x)$ if
    - Each X_i is a rv with identical cdf given by F(x)
    - These *n* rvs are independent

Statistical Inference Visualised



Three Basic Questions

- If we can only choose one parameter value based on some sample data, which parameter value would be our best guess for an unknown model parameter?
    - Point estimation
- Which possible parameter values of the statistical model are compatible with the sample data?
    - Interval estimation or confidence intervals
- Is there enough evidence based on the sample data to reject a pre-specified model parameter value?
    - Hypothesis testing

***yrbss* Dataset**

- Consider the dataset called *yrbss* which represents all 13,583 high school students in the Youth Risk Behaviour Surveillance System from 2013.
- The table below is a portion of the table representing the data (missing values are represented with NA):

	age	gender	grade	height	weight	helmet_12m	physically_active_7d	strength_training_7d
1	14	female	9	NA	NA	never	4	0
2	14	female	9	NA	NA	never	2	0
3	15	female	9	1.73	84.37	never	7	0
4	15	female	9	1.60	55.79	never	0	0
13582	17	female	12	1.60	77.11	sometimes	5	NA
13583	17	female	12	1.57	52.16	did not ride	5	NA

Name	Description
age	Age of the student.
gender	Sex of the student.
grade	Grade in high school
height	Height, in meters.
weight	Weight, in kilograms
helmet	Frequency that the student wore a helmet while biking in the last 12 months.
active	Number of days physically active for 60+ minutes in the last 7 days.
lifting	Number of days of strength training (e.g. lifting weights) in the last 7 days.

- Then, taking a **random sample of this population** (referred to as the *yrbss_samp* data set), we get the following:

	age	gender	grade	height	weight	helmet_12m	physically_active_7d	strength_training_7d
5653	16	female	11	1.50	52.62	never	0	0
9437	17	male	11	1.78	74.84	rarely	7	5
2021	17	male	11	1.75	106.60	never	7	0
2325	14	male	9	1.70	55.79	never	1	0

- We will use this sample to draw conclusions about the population

Point Estimate

- We would like to estimate some features of the high schoolers in *yrbss* using the sample including:
  - What is the average height of the *yrbss* high schoolers?
  - What is the average weight of the *yrbss* high schoolers?
- In order to estimate the population mean, we simply take the sample mean:

$$\overline{x}_{height} = \frac{1.5 + 1.78 + \ldots + 1.7}{100} = 1.6969.$$

- This sample mean is called a **point estimate** of the population mean.
- If we recompute with a new sample we will get slightly different answers

Standard Error of an Estimate

- Point estimates only approximate the population parameter, and they vary from one sample to another.
- The **sample variation** can be described in detail by the sampling distribution of the point estimate, or roughly by the standard error.
- The standard deviation associated with an estimate is called the **standard error**.
- It describes the typical error or uncertainty associated with the estimate.
- When considering the point estimate, there is no obvious way to estimate its standard error/deviation from a single sample as one sample only yields one point estimate.

- However, given *n* independent observations from a population with standard deviation σ, the standard error of the sample mean is equal to:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Interval Estimation

- We know that point estimates are not exact and there is always some variation attached to each estimate.
- The logical next step would be to provide a plausible range of values for the parameter.
- A plausible range of values for the population parameter is called the **confidence interval**.
- Using a confidence interval is more effective than using a point estimate and gives a greater chance of having the correct approximation.

Confidence Interval

- Given a sample $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu,\sigma^2)$
- Assuming we know that σ is then,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = Z \sim N(0,1)$$

and thus, $P(-1.96 \leq Z \leq 1.96) = 0.95$

- We will then substitute *Z* to solve for μ:

$$\begin{aligned} 0.95 &= \mathrm{P}\left(-1.96 \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \\ &= \mathrm{P}\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) \\ &= \mathrm{P}\left(1.96\frac{\sigma}{\sqrt{n}} \leq \mu - \overline{X} \leq -1.96\frac{\sigma}{\sqrt{n}}\right) \\ &= \mathrm{P}\left(\overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} - 1.96\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

- Therefore, the random interval:

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

covers μ with a probability of 0.95

- In other words, we are 95% confident that the calculated confidence interval includes μ
  - This statements should be stated in terms of the **confidence interval** itself being the random variable
  - We cannot say that there is a 95% chance that μ lies within the confidence interval as this implies μ is random (which it isn't)
- The sample mean will always be at the centre of the interval as it is still our best single point estimate for μ.
- The CI (confidence interval) that relies on the fact that:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = Z \sim N(0,1)$$

requires that

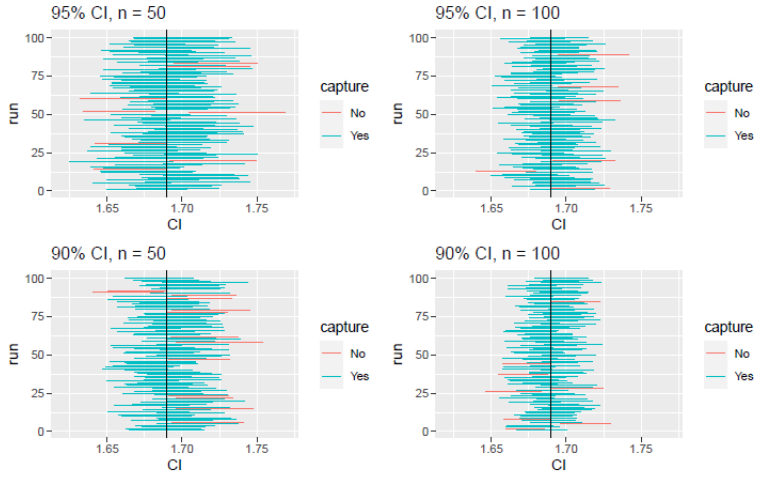
  - X_i be i.i.d normally distributed with known variance, or
  - *n* be large enough (for CLT to be appropriate) with a known variance.

- We can also look at what to do when some of conditions are not satisfied

CI for μ if σ is known

- The $100(1-\alpha)\%$ CI for μ is given by
  $$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
  and is constructed by finding $z_{\alpha/2}$ such that (and solving for μ):
  $$1-\alpha = \mathrm{P}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$$
- $z_{\alpha/2}$ represents the top 100(α/2)-percentile of the standard normal distribution
- α = 0.05 (or a 95% confidence) is by far the most commonly used
  - If $\alpha = 0.05$, then $z_{\alpha/2} = z_{0.025} = 1.96$ is the top 2.5 percentile of the standard normal distribution.
- The CI covers the true μ with relative frequency approximately (1 – α).
- This means that for:
  - a 90% CI, we have $z_{\alpha/2} = z_{0.1/2} = z_{0.05} = 1.645$.
  - a 99% CI, we have $z_{\alpha/2} = z_{0.01/2} = z_{0.005} \approx 2.58$.
- The CI gets wider as you increase the confidence level, i.e. make (1 – α) larger.

Simulated CIs for *yrbss*



- 100 random samples of size *n* = 50 or 100 were taken from *yrbss*.
- For each sample, a confidence level was created to capture the average height of students.

- To get the first plot, we used:

```
library(tidyverse)
yrbss <- read_csv("yrbss.csv")
mu <- round(mean(yrbss$height, na.rm=TRUE), 2)
n <-  50
CI_sim <- 1:100 %>%
  purrr::map_df(~ yrbss %>% filter(!is.na(height)) %>% sample_n(n) %>%
                  summarise(lb = mean(height)-1.96*sd(height)/sqrt(n),
                            ub = mean(height)+1.96*sd(height)/sqrt(n))) %>%
  mutate(capture = ifelse(lb<mu & ub>mu, "Yes", "No"), run = 1:100)
ggplot(CI_sim) +
  geom_segment(aes(y = run, yend=run, x = lb, xend=ub, color= capture)) +
  geom_vline(xintercept = mu, color="black") +ggtitle("95% CI, n = 50") +
  xlab("CI") + xlim(1.62, 1.77)
```

- Notes for the R code above:
  - `purrr::map_df` is used to repeat the code 100 times and condense the results into a data frame.
  - `summarise` is used to compute one CI per sample.
  - `mutate` and `ifelse` is combined to see whether the CI contains the true mean.
  - `geom_segment` is used to plot the CIs; `geom_vline` plots a vertical line.

CI for *height*

- Assuming the variable *height* is normally distributed, and from the full dataset we know that σ = 0.1047.
- The from *yrbss.samp*, we have:

```
yrbss.samp %>% summarise(mu = mean(height), n = n())
```

```
#       mu   n
# 1 1.6969 100
```

- Using the above information, we can construct a 90% and 99% CI for the average height from the sample.
- The 90% CI for μ: Find z such that $0.90 = P(-z \leq Z \leq z)$, that is $P(Z > z)$. From the z-table: z = 1.645.
  CI calculates to
  $1.6969 \pm 1.645 \times \frac{0.1047}{\sqrt{100}} = 1.6969 \pm 0.0172 = (1.6797, 1.7141)$
- The 99% CI for μ: $0.99 = P(-z \leq Z \leq z)$ implies that z = 2.5758, and:
  CI calculates to
  $1.6969 \pm 2.5758 \times \frac{0.1047}{\sqrt{100}} = 1.6969 \pm 0.027 = (1.6699, 1.7141)$

Margin of Error and Minimum Sample Size

- In the construction of the CI, the $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ part is often referred to as the **margin of error**.
- The margin of error can be interpreted as how accurate our estimates are for a given confidence interval.
- A larger *n* will give a smaller margin of error.
- Sometimes (to save resources), we are satisfied if the margin of error is within a certain limit (i.e. a maximum margin or error) which implies our estimates would be accurate enough.

- We can then plan ahead in our data collection process to find the necessary sample size.
    - This is known as sample size calculations.

EXAMPLE:

- Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest.
- The average systolic blood pressure for people in the US is about 130 mmHg with a standard deviation of about 25 mmHg.
- How large of a sample is necessary to estimate the average systolic blood pressure with a margin of error of 4 mmHg using a 95% confidence level?
    - The margin of error for a particular sample size n is:

      $$1.96 \times \frac{25}{n}$$

      and that has to be less than or equal to 4
    - This means that:

      $$1.96 \times \frac{25}{n} \leq 4 \quad \Rightarrow \quad 1.96 \times \frac{25}{4} \leq \sqrt{n}$$
      $$\Rightarrow \quad n \geq \left(1.96 \times \frac{25}{4}\right)^2$$
      $$\Rightarrow \quad n \geq 150.0625.$$
    - This means we should choose a sample size of at least 151 employees.
- We round up to the next whole number.

Hypothesis Testing

- A hypothesis, H, is a statement about the unknown parameter (e.g. μ) of the population
- A null hypothesis, H_0, is a hypothesis set up to nullified or refuted in order to support and alternative hypothesis, H_1.
- A hypothesis testing framework is built for a sceptic to consider a new claim
- The sceptic will not reject the null hypotheses unless the evidence in favour of the alternative hypothesis is so strong that H_0 is rejected in favour of H_1.
- The hypothesis testing framework is a very general framework
- The hallmarks of hypothesis testing are found in the Australian Legal System.
    - “Innocent until proven guilty”

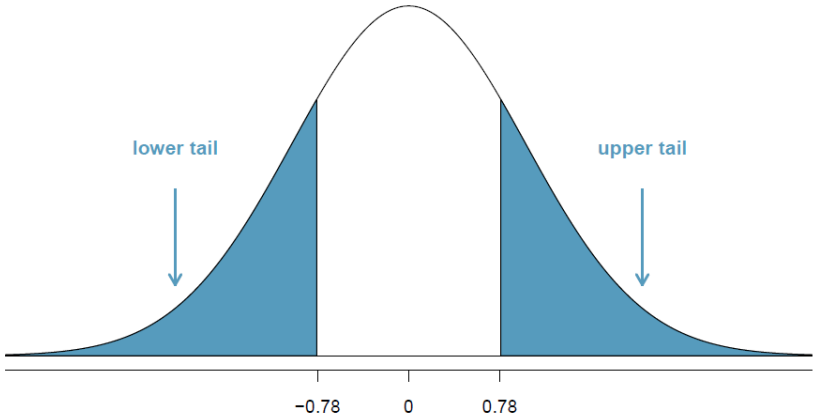
After the Hypothesis

- General Strategy: Find some statistics (τ) to access if data is consistent with H_0 and calculate a corresponding P-value.
- P-value is the probability that the observed value is an extreme or unusual observation based on the assumption that H_0 is true.
    - i.e. the lower the P-value, the more likely that H_0 is wrong.
- Uncertainty in the results: Because observations vary from sample to sample, we can never say for sure whether H_0 is true or not.

EXAMPLE:

- Suppose the same Youth Risk Behaviour Surveillance System (YRBSS), survey was also completed in 2008.
- The mean height for students at the time is 1.69m.
- We want to determine if the sample data set provides strong evidence that *yrbss* students in 2013 are taller or shorter than in 2008, versus the other possibility that there has been no change.
- We can simplify these options into two competing hypotheses:
    - H_0: The mean height of YRBSS students was the same for 2008 and 2013
    - H_1: The mean height of YRBSS students was different for 2008 and 2013
- This is generally referred to as a two-sided test as we are interested in both an increase and decrease of mean height between the two datasets.
- If we let μ_{height} be the mean height of YRBSS students in 2013, then these hypothesis can also be described in mathematical notation:
    - $H_0: \ \mu_{height} = 1.69.$
    - $H_1: \ \mu_{height} \neq 1.69.$
- 1.69 is also called the null value since it represents the value of the parameter if the null hypothesis is true.
- To access the mean height for 2013, we would use the sample mean from the *yrbss_samp* dataset.
- Assuming the readings within the height variable are independent, identical and normally distributed as $N(\mu,\sigma^2)$.
- This means our test statistics τ is:
$$\overline{X} \sim N\left(\mu, \frac{(0.0881)^2}{100}\right)$$
- We can then define the test statistics τ based on the z-score corresponding to the sample mean:
$$\tau: \quad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$
- The test statistics evaluated for the particular sample (called the observed value of the test statistics) is:
$$\tau_{obs} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{1.6969 - 1.69}{0.0881/\sqrt{100}} = 0.78$$
- Here, μ takes the null value as the test is conducted under the assumption that H_0 is true.
- Then the question becomes: Is a sample mean of 1.6969 considered as unusual?
- Larger or smaller values of τ, means more support for H_1
    - That is equivalent to values of $\overline{X}$ larger than or smaller than 1.69.
    - i.e. only the distance away from μ is important but not the direction. (evidence in either direction is favourable to H_1)

- We consider the event of $|Z| \geq 0.78 = (Z < -0.78 \; or \; Z > 0.78)$
    - Starting from τ_{obs} and move to values that are even more unusual)



- The probability of observing the above event is the P-value:
$$P\text{-value} = \mathrm{P}(|Z| \geq 0.78) = \mathrm{P}(Z \leq -0.78) + \mathrm{P}(Z \geq 0.78)$$
$$= 2\mathrm{P}(Z \leq -0.78) = 0.4354.$$
- The P-value is relatively large, so we should not reject H_0.
- This is, if H_0 is true, it would not be very unusual to see a sample mean this far from 1.69 due to sampling variation.
- Thus, we do not have sufficient evidence to conclude that the height is different to 1.69m.

P-value, decision & conclusion

- If the P-value is small enough, then we have evidence against H_0 in favour of H_1.
- In the height example, we would conclude that there is no significant changes to the previous mean-height. Why?
- How small does the P-value have to be in order for us to decide in favour H_1?
- There is not set value, but:
$$P\text{-value} \leq \alpha = 0.05 = 1/20.$$
is often used in practise
- This cut-off is called the **significance level**.
    - Note: this is the same alpha α in a 100(1 – α)% CI. That is why 95% CI is the most commonly used CI in practise.
- Other choices are 0.1, 0.01 or 0.001 according to 'innocent until proven guilty' principle.
- Conclusions:
    - If the P-value is smaller than α, report that there is evidence against H_0.
    - If the P-value is larger than α, simply state that the data is consistent with H_0.
- We should always state the conclusion in plain language so the general public can understand the results.
- Note: there is no final proof that H_0 is true or false.
- We know that the smaller the P-value, the stronger the evidence against H_0 in favour of H_1.
- Avoid temptations to change pre-specified values:
    - α must be pre-specified from the beginning of the study
    - Hypotheses must be setup before observing data.

The Eight Point Check

1. Null Hypothesis, H_0. The claim against evidence is searched for.
2. Alternative hypothesis, H_1. The alternative you will consider if H_0 is false.
3. Test statistics, τ.
4. Sampling distribution of τ when H_0 is true
5. Which values of τ would argue against H_0?
6. Observed value of τ from the sample. i.e. τ_{obs}
7. P-value
8. Write a statistical and contextual conclusion.

Assumptions for a test

- Continuing from the previous example, so far we know that:
  $P(|\overline{X} - 1.69| > |1.6969 - 1.69|) = 0.4354$
- This is based off various assumptions. In particular, $\overline{X}$ is normally distributed:
  - X_i is a random sample of normally distributed rvs.
  - *n* is large enough for CLT
- σ is known and is equal to 0.0881
- If any of these assumptions are not valid, then P-value is just a number and it will not have the same meaning or interpretation anymore.
- It is important to check your model assumptions whenever you can.

Rejected Regions

- We would reject the null hypothesis when the P-value is too small (i.e. ≤ α), but that is equivalent to τ_{obs} being too unusual/rare in the sampling distribution of the test statistics.
- The area of τ_{obs} that is considered as unusual, is denoted as the rejected region.
  - For instance, in the YRBSS height example with α = 0.05, we can reject the null hypothesis if τ_{obs} is in top and bottom 2.5% of the distribution
  - We would not reject H_0 as τ_{obs} = 0.78 does not lie in the rejection region.

Confidence Interval

- Recall that a 100(1 – α)% CI provides a plausible range of values of the population parameter.
- The 95% CI for μ is the range of values of μ_0 that would be retained if we carried out a test with a 5% significance level:
    - Retain H_0 if CI contains μ_0;
    - Reject H_1 if CI does not contain μ_0;
- Generally speaking, the 100(1 – α)% CI is the range of values of μ_0 that would be retained if we carried out a test on H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$ with a significance level of α.

EXAMPLE:

- From the YRBSS height example, a two-sided 95% CI for μ would be:
$$1.6969 \pm 1.96 \times \frac{0.0881}{\sqrt{100}} = (1.6796, 1.7142).$$
- In the test H_0: μ = 1.69 vs H_1: $\mu \neq \mu_0$, we would retain H_0 as 1.69m which lies within the 95% CI.
- This means that there is insufficient evidence to be able to conclude that the true mean is different from 1.69m.
- Obviously, we would reject H_0 if the null value does not lie in the 95% interval.

Decision Errors

- Hypothesis tests are not flawless, since we can make a wrong decision in tests based on data.
- Recall that if the P-value is small then either:
    - H_0 is true but we have observed an unlikely event OR
    - H_0 is false
- But we don't know which one has happened.
- However, in a statistical hypothesis test, we have the tools necessary to quantify how often we make such errors.
- There are two competing hypotheses: the null and the alternative. We make a statement about which one might be true, but we might choose incorrectly.
- Therefore, there are 4 possible scenarios:

		Decision	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type-I error
	H_A true	Type-II error	okay

- A **Type-I Error** is rejecting the null hypothesis when H_0 is actually true.
- A **Type-II Error** is failing to reject the null hypothesis when the alternative is actually true.

Probability of Type-I and Type-II errors

- Type-I errors means that if H_0 is true, but we reject H_0

$$\text{P(Type I error)} = \alpha, \quad \text{(i.e. the significance level)}$$

since the test reject α of the z-scores as being too extreme.

- Type-II error means that if H_0 is false, we do not reject it, because we obtained an unusual example with a mean similar to the hypothesised null value, and:

$$\text{P(Type-II error)} = \text{P(Not rejecting } H_0 \text{ when } H_0 \text{ is false)} = \beta$$

which we shall show later that β depends on α, the true parameter value and of course, the sample size n.

Choosing α

- Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05.
- However, we can sometimes select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.
- If making Type-I errors is dangerous (or costly), we should choose a small significance level (e.g. 0.01)
- Under this scenario, we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favouring H_1 (i.e. a very small P-value)
- If a Type-II error is relatively more dangerous, then we should choose a higher significance level.
- Overall, the significance level should reflect the consequences associated with Type-I and Type-II errors.

EXAMPLE:

- A car manufacturer is considering a higher quality more expensive supplier for window parts in its vehicles.
- They sample a number of parts from their current supplier and also parts from the new supplier. They decide that if the high-quality parts will last more than 12% longer, it makes financial sense to switch to this more expensive supplier.
- Is there a good reason to modify the significance level in such a hypothesis test?
    - This decision is just one of the many regular factors that have a marginal impact on the car and company. The standard significance level of 0.05 seems reasonable since neither a Type-I or Type-II error should be dangerous or much more expensive.
- The same car manufacturer is considering a slightly more expensive supplier for parts related to safety, not windows.
- If the durability of these safety components is shown to be better than the current supplier, they will switch manufacturers.
- Is there a good reason to modify the significance level in such an evaluation?
    - Because safety is involved, the car company should be eager to switch to a slightly more expensive manufacturer (reject H_0) even if the evidence of increased safety is only moderately strong. A slightly larger significance level, such as α = 0.10, might be appropriate.

Power

- A Type-I error is simple but β could be messy.
- Given that we can't minimise both α and β at the same time, we can instead minimise β while keeping α fixed.
- We often denote 1 – β as the power of the test, i.e.
  $\text{Power of the test} = \text{P(Rejecting } H_0 \text{ when } H_0 \text{ is false)} = 1 - \beta.$
- Obviously, we want to power of the test to be as high as possible.
- In order for us to evaluate the power of a test, we want to make the test even simpler: by making it one-sided.

One-sided Tests

- A video encoder algorithm has a rendering time for an industry standard video that is normally distributed with a mean of 19.25 seconds and a variance of 2.25 (σ = 1.5)
- A new video encoder is developed, and we need to find out: does this new encoder have a faster encoding time compared to the current encoder?
- Take a sample of 10 different computers and obtain the encoding times in minutes for each using the new encoder.
- Data: *y*

18.3	17.9	19.1	16.8	18.9	17.4	19.6	18.3	19.6	16.3

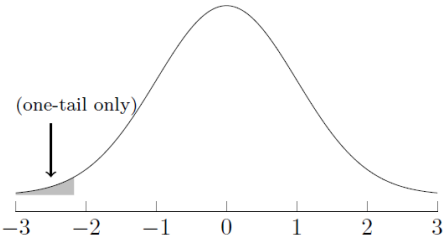
$\overline{y} = 18.22$ and $s = 1.132$

- Is there evidence that the new encoder has a population mean of 19.25, OR has a population mean that is smaller than 19.25.
- We are not interested, and therefore not gathering evidence, for the scenario μ > 19.25.
- We assume that the standard deviation is unchanged
- The test procedure is the same, with the exception that only smaller values of τ provide more support for H_1.

Notation	Terminology
$H_0 : \mu = 19.25$	Null hypothesis
$H_1 : \mu < 19.25$	(One-sided) Alternative hypothesis
$\alpha = 0.05$	Significance level

- Our test statistics is still $\tau = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}$.

$$\tau_{obs} = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}} = \frac{18.22 - 19.25}{1.5/\sqrt{10}} = \frac{-1.03}{0.4743416} = -2.17$$



- $P\text{-value} = P(Z < -2.17) = 0.015.$
- Therefore, we reject the H_0 at the 5% significance level as the P-Value is less than 0.05, and there is sufficient evidence at the 5% level of significance to conclude that the true mean encoding time for the new encoder is NOT 19.25 minutes but significantly less than 19.25 minutes (it was 18.22 in the sample)

- The rejected region in this example would be to reject H_0 if $\tau_{obs} < -z_\alpha = -1.645$, i.e. the bottom 5% of values.
- One-sided tests are often a matter of perspective.
    - If you are a competitor and want to find evidence that the new encoder is not better, then you will have an alternative hypothesis of the form: $\boldsymbol{H_1 : \mu > 19.25}$
- In this case, $\overline{y} < 19.25$, then there is no evidence that $\mu > 19.25$
- We will retain H_0 with a strong belief and there is no need to formally do the test and compute the P-value, etc.
- But in the case that $\overline{y} > 19.25$, we can carry out the test similarly to before with a few changes:
    - Larger values of τ, more support for H_1
    - Adjust the P-value expression accordingly

Two-sided or One-sided

- Depends on the research question
- When you are interested in checking for an increase or a decrease, but not both, use a one-sided test
- When you are interested in any difference, use a two-sided test
- Switching a two-sided test to a one-sided test after observing the data is dangerous because it can inflate the Type-I error rate.

One-sided Confidence Interval

- So far, we have seen confidence intervals which are two-tailed that are consistent with two-sided tests.
- We also need a one-sided CI to be consistent with one-sided tests.
- Suppose we are testing H_0: $\mu = \mu_0$ vs H_1: $\mu < \mu_0$
- We would reject H_0 if τ_{obs} is too unusual (too small)
- This also means to be consistent with H_0, CI would consist of a range of values τ_{obs} that are not too small.
- From the standard normal table, we know that:
$$P(Z > -1.645) = 0.95, \quad \text{i.e. } -z_\alpha = -1.645.$$
- We also know that:
$$\tau = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = Z \sim N(0,1).$$
- Substituting that in the P-value expression to get:
$$0.95 = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > -1.645\right) = P\left(\overline{X} - \mu > -1.645 \times \frac{\sigma}{\sqrt{n}}\right)$$
$$= P\left(\overline{X} + 1.645 \times \frac{\sigma}{\sqrt{n}} > \mu\right)$$
$$\left[= P\left(\mu < \overline{X} + 1.645 \times \frac{\sigma}{\sqrt{n}}\right)\right]$$
- This gives an upper bound on the plausible values of μ but is unbounded below.

- The 95% CI for μ is therefore:

$$\left(-\infty,\ \overline{X} + 1.645 \times \frac{\sigma}{\sqrt{n}}\right)$$

- In order to retain H_0, $\overline{X}$ **– μ cannot be too small** which implies that **μ cannot be that much larger than** $\overline{X}$ which implies that **CI has an upper bound**.
- If μ_0 lies in the CI, we retain H_0.
- If μ_0 does not lie in the CI, we reject H_0 in favour of H_1.
- Similarly, for a test on $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$, the corresponding one-sided CI for μ is

$$\left(\overline{X} - z_\alpha \times \frac{\sigma}{\sqrt{n}}, \infty\right)$$

Changing the Sample Size

- Considering the video encoder example:
- If we have a sample size of n = 5:

$$\begin{aligned} \tau_{obs} &= \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{18.22 - 19.25}{1.5/\sqrt{5}} \\ &= \frac{-1.03}{0.67} = -1.535 \end{aligned}$$

$$P\text{-value} = P(Z < -1.535) = 0.062.$$

  - Not enough evidence to reject at the 5% level
- Suppose we have a sample size n = 1000:

$$\begin{aligned} \tau_{obs} &= \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{18.22 - 19.25}{1.5/\sqrt{1000}} \\ &= \frac{-1.03}{0.0474} = -21.73 \end{aligned}$$

$$\text{P-Value} = P(Z < -21.73) = 3.279278 \times 10^{-98} \approx 0$$

  - Reject H_0 at any level because we are very confident that the new encoder has a decreased encoding time.
- The larger the sample size, the more confident we can be about the sample mean as an estimate of μ.
- We can work out in advance what sort of difference between the sample mean and μ is needed for the test to be significant, given a particular sample size.
- Given:

$$\tau_{obs} = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{x} - 19.25}{1.5/\sqrt{1000}}$$

this will be significant if $\tau_{obs} < -1.645$

- The minimum value of $\overline{x}$ that achieves this is:

$$\overline{x} < 19.25 - 1.645 \times \frac{1.5}{\sqrt{1000}} = 19.17$$

- Any observed sample mean less than 19.17 (0.08 units lower than 19.25) will be deemed significantly less than 19.25 at the 5% level of significance.

Practical vs Statistical Significance

- Does a difference of 0.08 minutes (4.8 seconds) mean anything?
- If a result is **statistically significant**, we are saying that we are confident that the true mean is different from the hypothesized value.
- However, the magnitude of that difference may not be of interest to us.
- Therefore, before conducting an experiment, the meaningful difference can also be decided:
  - This is the **practical significance** amount and it depends on the context.
- Denote **δ** as the minimum practical significance

- Suppose marketing decide a new video encoder is only competitive if the average decrease in time in 0.5 minutes:
  - They require δ = 0.5
- If we observe a difference lower than 0.5, it is negligible.
- Now we aim to find a sample size such that we can detect a practically significant result as the 5% level of significance.

Computing the required sample size

- We want minimal difference δ = 0.5 to satisfy:

$$\tau_{obs} = \frac{\overline{y} - \mu}{1.5/\sqrt{n}} < -1.645$$

- Rearranging to get:

$$\frac{-1.645 \times 1.5}{\sqrt{n}} > -0.5\,(= 18.75 - 19.25)$$

$$\Rightarrow \quad \sqrt{n} > \frac{-1.645 \times 1.5}{-0.5} = 4.935 \quad \Rightarrow \quad n > (4.935)^2 = 24.35423.$$

- This means we need at least 25 observation to detect an encoding time of 30 seconds quicker and be able to conclude the 5% significance level.
- If we have a lower sample size the result won't be statistically significant.
- General formula for sample size calculations:

$$n > \left(\frac{z_{critical} \times \sigma}{\delta}\right)^2$$

Revisiting Power

- Consider the video encoder example, but we want to test a new hypothesis:

  $H_0 : \mu = 17.25$

  $H_1 : \mu > 17.25$
  - With α = 0.05 and σ = 1.5
  - Also using the same summary statistics: $\overline{y}$ = 18.22 and n = 10
- As power is related to the rejection of H_0, we have the figure out the rejection region based on $\overline{y}$.

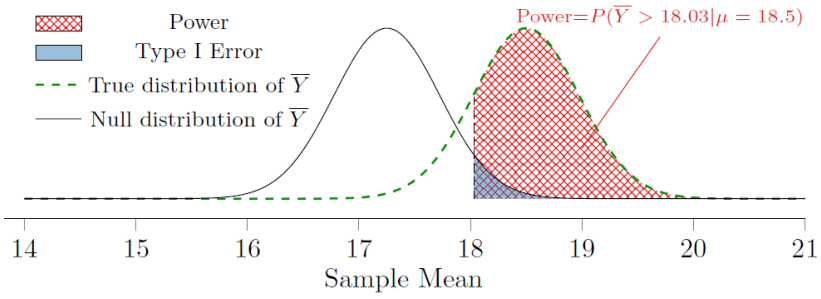
- Critical value (minimum $\overline{y}$ that will give a statistically significant result at the 5% level) given by:

$$\frac{\overline{y} - 17.25}{1.5/\sqrt{10}} > 1.645 = z_\alpha = z_{0.05}$$

- Solving gives $\overline{y} > 18.03$
- Thus, for our statistical procedure with a sample size of 10, we will:
  - Reject H_0 if $\overline{y} > 18.03$
  - Not reject H_0 if $\overline{y} < 18.03$
- We can then work out the probability of getting $\overline{y} > 18.03$ for particular values of μ.
  - This is called the **power of the test** for various values of μ
- Computing the power when the true mean is μ = 18.5 (Power of the test when μ = 18.5):

$$= \mathrm{P}(\overline{Y} > 18.03 | \mu = 18.5)$$
$$= \mathrm{P}\left(\frac{\overline{Y} - 18.5}{1.5/\sqrt{10}} > \frac{18.03 - 18.5}{1.5/\sqrt{10}}\right)$$
$$= \mathrm{P}(Z > -0.99)$$
$$= 0.84$$

- If the true mean is 18.5, there is an 84% chance that H_0 : μ = 17.25 is rejected in favour of H_1 : μ > 17.25



Computing the Power when μ → μ_0

Power at $\mu = 18.0$

$$= \mathrm{P}(\overline{Y} > 18.03 | \mu = 18.0)$$
$$= \mathrm{P}\left(Z > \frac{18.03 - 18.0}{1.5/\sqrt{10}}\right)$$
$$= \mathrm{P}(Z > 0.063)$$
$$= 0.476$$

Power at $\mu = 17.25$

$$= \mathrm{P}(\overline{Y} > 18.03 | \mu = 17.25)$$
$$= \mathrm{P}\left(Z > \frac{18.03 - 17.25}{1.5/\sqrt{10}}\right)$$
$$= \mathrm{P}(Z > 1.645)$$
$$= 0.05, \text{ as expected}$$

- If the true mean is 18, there is a 48% chance that H_0 is rejected
- If the true mean is 17.25 (H_0 is true), there is a 5% chance that we reject H_0
- **Note**: the greater the distance between the true mean and the mean claimed in H_0, the higher the power is.
- **Note**: Power are increases as the sample size *n* increases.

Power for two-sided tests

- For a two-tailed test, the sample principles apply but now we have two rejected regions.

- Reject H_0 at the 5% level if $|\tau_{obs}| > 1.96$.
  - Either $\tau_{obs} < -1.96$ or $\tau_{obs} > 1.96$.
  - Either $\overline{y} < 17.25 - 1.96 \times 1.5/\sqrt{10} = 16.23$.
  - Or $\overline{y} > 17.25 + 1.96 \times 1.5/\sqrt{10} = 18.18$.
- Power $= P(\overline{Y} > 18.18) + P(\overline{Y} < 16.32)$ for the chosen true μ.

TOPIC 7 – Inference about a Single Population Mean

Tests for the mean μ

- Statistical tests can be developed to test claims about the population mean, but it has a few assumptions:
- Assumption 1: Normality
  - The population we are interested in has a normal distribution.
  - The test to use for making an inference about μ is the sample mean.
  - This leads to two scenarios, either σ is known or σ is unknown.
- Assumption 2: σ is known
  - If the population variance is known, σ^2, is known, then the sampling distribution of the average is:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  - This is called the Z-test

Z-test

- Test $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, where μ_0 is a given null value.
- If H_0 is true with σ known then,

$$\overline{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

- If the observed sample average is $\overline{x}$, then the P-value is:

$$P(\overline{X} \geq \overline{x}) = P\left(Z \geq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right), \quad \text{where } Z \sim N(0,1).$$

- The test statistics of the Z-test is the z-score based on the observed sample mean:

$$\tau = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

- Whether the underlying data is normal or not, if we have a large sample size (n), the CLT will enable us to calculate an approximate P-value to test the hypothesis.

EXAMPLE:

- In a random sample of 128 arterioles taken from SIDS victims, the mean muscle thickness as a percentage of total arteriole diameter was 9.10.
- Assume that the percentage of muscle thickness can be modelled by:

$$X_i \sim N(\mu, 2.15^2).$$

- For normal children of the same age, we have μ = 6.04.
- *Is there evidence that the muscle thickness is greater in SIDS victims?*
  - Test $H_0 : \mu = 6.04$ against $H_1 : \mu > 6.04$ with test statistic:

$$\tau = \frac{\overline{X} - 6.04}{2.15/\sqrt{128}} \sim N(0,1), \quad \text{if } H_0 \text{ is true.}$$

- The observed value of the test statistic is:

$$\tau_{obs} = \frac{\overline{x} - 6.04}{2.15/\sqrt{128}} = \frac{9.10 - 6.04}{2.15/\sqrt{128}} = 16.10.$$

- Then the P-value is:

$$\begin{aligned} P - value &= P(Z \geq 16.10) \\ &< P(Z \geq 2.99) \\ &= 1 - P(Z < 2.99) = 0.0014. \end{aligned}$$

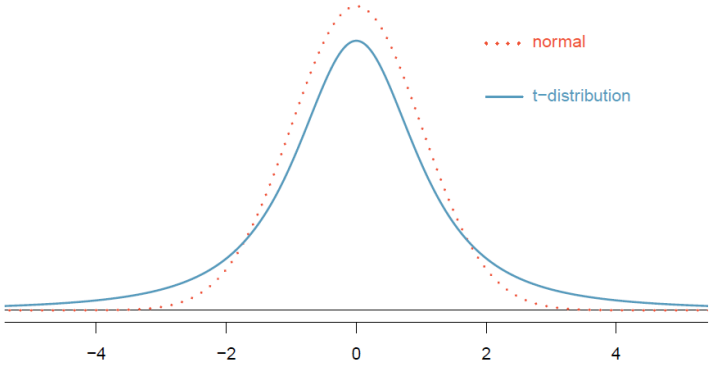
- The P-value is very small, so there is strong evidence against H_0
- i.e. There is strong evidence that the muscle thickness is greater in SIDS victims.

One sample t-test

- If σ is unknown, replace the fixed parameter σ in the test statistics with the sample standard deviation *S*.
- Formally, this procedure is called a t-test.
- The test statistic for a t-test is given by:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

  - i.e. it is a random variable having the t-distribution with *n – 1* degrees of freedom
- A t-distribution has a bell shape and is centred at zero. However, its tails are thicker than a normal distribution.
- This means observations are more likely to fall beyond two standard deviations from the mean than in a normal distribution.



- When the degrees of freedom is greater than 30, the t-distribution is indistinguishable from a normal distribution.
- As n → ∞, t_{n-1} approaches a normal distribution.

EXAMPLE:

- The birthweights of a random sample of n = 14 boys born to mothers who smoked heavily during pregnancy were recorded.
- The data is as below:

```
x <-  c(79, 92, 88, 98, 109, 109, 112, 88, 105, 89, 121, 71, 110, 96)
```

- It is believed that on average, boys born to mothers who smoke have a lower birthweight than the national average of 109 ounces.
- *Is there any evidence that birthweight is lower for boys born to mothers who smoke?*
    - We assume that the birthweight for mothers who smoke is modelled by:
      $W \sim N(\mu, \sigma^2)$
    - From the data we have:
      $\overline{w} = 97.6429$ and $s = 14.0582$.
    - Test $H_0 : \mu = 109$ against $H_1 : \mu < 109$ using a t-test.
    - Test statistic is:

      $$\tau = \frac{\overline{w} - 109}{s/\sqrt{14}} \sim t_{n-1} = t_{13}, \quad \text{if } H_0 \text{ is true.}$$

    - The observed value of the test statistics is:

      $$\frac{97.6429 - 109}{14.0582/\sqrt{14}} = -3.0227.$$

    - And from the t-table we can find the range of the P-value:
      $P(t_{13} < -3.012) = P(t_{13} > 3.012) = 0.005$ and

      $P(t_{13} < -3.852) = P(t_{13} > 3.852) = 0.001$ which means

      $$0.001 < P\text{-value} < 0.005.$$

- Hence, we have strong evidence that the birthweight for boys born to mothers who smoke, have birthweights lower than the national average.

Confidence Interval for a t-test

- The two-sided $100(1 - \alpha)\%$ CI is given by:
  $$\overline{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$
- When testing $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$, the one-sided CI has the form:
  $$\left(-\infty, \quad \overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}\right)$$
- When testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, the one-sided CI has the form:
  $$\left(\overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}, \quad \infty\right)$$

Assessing Normality of Data

- Whenever we assume a normal distribution, we should determine whether our assumption is acceptable.
- Up to now, we can obtain a histogram and look for:
    - A symmetric distribution
    - An approximate bell-shape
- There are proper significance tests for testing for a normal distribution, but they require a reasonably large amount of data before we can make a conclusion with any confidence.

- For small data sets, significance tests for normality may be inconclusive.
- However, we can get an idea of normality by obtaining and then observing the normal **quantile-quantile plot** of the data.
- This well give us a subjective measure of how closely our data matches a normal distribution.

Normal Quantile-Quantile Plot

- Also known as a QQ-plot. These are the steps required to create one:

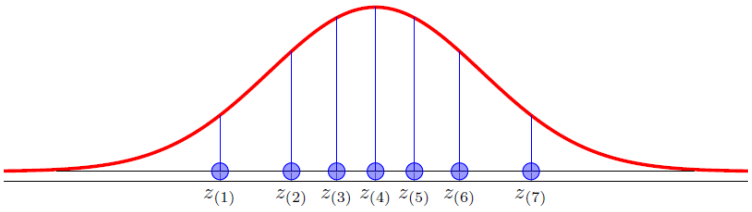
1. Calculate the normal scores:
    - These are the values along the axis that split the bell curve into equal sized areas
2. Match up the normal scores with the sample data:
    - Pair up ordered sample data with the ordered normal scores (i.e. smallest sample observed paired with smallest normal score.
3. Plot the pairs on a scatterplot:
    - If the data on the scatterplot is close to a straight line, this indicates that our data is very close to a normal distribution.
    - However, a perfect straight line is suspicious of fake data.

- We don't expect the line to be straight as this would indicate that there is no randomness.
- We expect random variance about a straight line with no consistent patterns.

EXAMPLE:

X	10.37	8.7	8.52	9.7	9.51	11.16	10.76

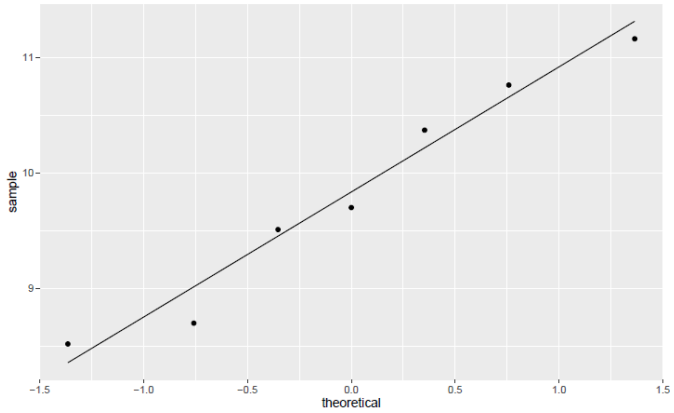
- Is it reasonable to assume this data comes from a normal distribution?
    - Calculation of normal scores: There are 7 observations, so that area under the standard normal curve is divided into 8 equal areas.
    - The normal scores are the values that divide the area in equal parts (i.e. $z_{(1)}$, $z_{(2)}$,…, $z_{(7)}$)
    - The $z_{(i)}$ values are not equally spaced, it is the areas that are equally spaced.



    - Matching the normal scores against the data:

Before Sorting		After sorting	
data	**normal score**	**data**	**normal score**
10.37	0.37	8.52	-1.47
8.7	-0.79	8.7	-0.79
8.52	-1.47	9.51	-0.37
9.7	0.00	9.7	0.00
9.51	-0.37	10.37	0.37
11.16	1.47	10.76	0.79
10.76	0.79	11.16	1.47

- Plotting normal scores against the data:



- This data roughly follows a straight line, so normal assumption is appropriate.

QQ-Plot using R

- *geom_qq()* creates the QQ-plot. *geom_qq_line()* draws the reference line on the diagonal.
- From the previous example, we have:

```
library(ggplot2)
dat = data.frame(x = c(10.37, 8.70, 8.52, 9.70, 9.51, 11.16, 10.76))
ggplot(data = dat, aes(sample = x)) + geom_qq() + geom_qq_line()
```

TOPIC 8 – Inference about Two Population Means

Independent samples from two populations

- Until now, we have looked at single samples and tested the population mean μ being a particular value μ_0.
- What if we want to see whether there is a difference between two populations?
- Define μ_1 and μ_2 as the population mean of each population being tested.
- In this case, we are testing:
  - $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, which is equivalent to
    $$H_0 : \mu_1 - \mu_2 = 0 \quad \text{against} \quad H_1 : \mu_1 - \mu_2 \neq 0.$$
- There are two possibilities if random samples are drawn from populations with the same mean:
  - The sample means will be close together (more likely)
  - The sample means will be far apart (less likely)
- i.e. The more different sample means are, the less likely they are from the same population.
- In order to test this, we need to find the sampling distribution of $\overline{X}_1 - \overline{X}_2$

Sampling Distribution of $\overline{X}_1 - \overline{X}_2$

- Suppose we have two independent samples of size n_1 and n_2 from two separate populations
  - Denote them as $X_{11}, \ldots, X_{1n1}$ and $X_{21}, \ldots, X_{2n2}$ respectively
- We further assume that the two populations have their own normal distribution. i.e.
  $$X_{1j} \sim N(\mu_1, \sigma_1^2), \quad j = 1, 2, \ldots, n_1$$
  $$X_{2j} \sim N(\mu_2, \sigma_2^2), \quad j = 1, 2, \ldots, n_2.$$
  - where σ_i is the true standard deviation of population *i*.
- This implies that:
  $$\overline{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_2}\right), \quad \overline{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$
- Thus, **the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is**:
  $$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$
- Standardising gives:
  $$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$
- Given that we assume H_0 is true (i.e. $\mu_1 = \mu_2$), the sampling distribution reduces to:
  $$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$
- This can be further simplified using the equal variance assumption (assuming $\sigma_1 = \sigma_2 = \sigma$):
  $$\frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1).$$

When σ_1 and σ_2 are known

- If σ_1 and σ_2 are known, we can carry out a two-sample z-test to test the hypotheses:

$$H_0 : \mu_1 = \mu_2 \quad \text{against} \quad H_1 : \mu_1 \neq \mu_2$$

with the following test statistic:

$$\tau = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1), \quad \text{if } H_0 \text{ is true.}$$

- In practice, population variances are rarely known.
- In this care, we will estimate the population variances with the samples variances and then carry out a t-test.

When σ_1 and σ_2 are unknown

- Th exact t-distribution only allows a single estimate of variance, so as a result we generally assume the populations have equal variances.
- The equal variance assumption in usually reasonable as different treatments often affect the mean but not the variance.
- To estimate the common variance, we use the **pooled variance** estimate:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

  - where s_1 and s_2 are the sample standard deviations of each population.
- This has $n_1 + n_2 - 2$ degrees of freedom.

Two-sample t-test

- To test:

$$H_0 : \mu_1 = \mu_2 \quad \text{against} \quad H_1 : \mu_1 \neq \mu_2.$$

- The test statistic is:

$$\tau = \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}, \quad \text{if } H_0 \text{ is true}$$

- Given that both a large and small value of τ_{obs} would argue against H_0, the P-value of the two-tailed test is:

$$\text{P-value} = \text{P}(|t_{n_1+n_2-2}| \geq |\tau_{obs}|).$$

- To test a one-sided alternative:

$$H_1 : \mu_1 < \mu_2, \ \text{P-value} = \text{P}(t_{n_1+n_2-2} \leq \tau_{obs});$$

$$H_1 : \mu_1 > \mu_2, \ \text{P-value} = \text{P}(t_{n_1+n_2-2} \geq \tau_{obs}).$$

EXAMPLE:

- Twenty identical lab rats are divided into two groups, and each group is fed a different diet.
- After a certain period of time, the weight gain of each rat was taken and is shown below:

diet	n	mu	sd
<chr>	<int>	<dbl>	<dbl>
Diet1	9	100.2222	19.34411
Diet2	11	126.5455	18.50602

- We want to test if there is any difference in average weight between the two diets.
- The pooled sample variance and sd is:

$$\begin{aligned} s_p^2 &= \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \\ &= \frac{(9-1)\times 19.3441^2+(11-1)\times 18.506^2}{9+11-2} \\ &= 356.5713 \end{aligned}$$

$$s_p = \sqrt{s_p^2} \approx 18.8831$$

- The test statistic is:

$$\tau = \frac{\overline{X}_1-\overline{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2} = t_{18} \quad \text{if } H_0 \text{ is true.}$$

- The observed value of the test statistic is:

$$\begin{aligned} \tau_{obs} &= \frac{\overline{x}_1-\overline{x}_2}{s_p\sqrt{1/n_1+1/n_2}} \\ &= \frac{100.2222-126.5455}{18.8831\times\sqrt{\frac{1}{9}+\frac{1}{11}}} \\ &= \frac{-26.3233}{8.4873} \\ &= -3.1015 \end{aligned}$$

- The P-value is (two-sided):

$$\begin{aligned} \text{P-Value} &= P(|t_{18}| > |\tau_{obs}|) \\ &= 2P(t_{18} > |\tau_{obs}|) \\ &= 2P(t_{18} > |-3.1015|) \\ &= 0.0061591 \text{ (exact)} \\ &< 2\times 0.005 = 0.01 \text{ (using table)} \\ &< 0.05 \end{aligned}$$

- We can conclude at the 5% significance level that there is evidence against H_0. i.e. The two diets do not have an equal effect on weight gain of the rats.

Confidence Interval for $\mu_1 - \mu_2$

- Confidence intervals of a parameter are in the form:
  $\text{estimate} \pm \text{critical value} \times \text{estimated s.e.(estimate)}$
- In the diet example:
  - $\text{Parameter} = \mu_1 - \mu_2$
  - $\text{Estimate} = \overline{x}_1 - \overline{x}_2 = 100.2222 - 126.5455 = -26.3233$
  - $\text{Critical value} = t_{n_1+n_2-2,\alpha/2} \overset{\alpha=0.05}{=} t_{18,0.025} = 2.101$
  - $\text{S.E. Estimate} = s_p\sqrt{1/n_1+1/n_2} = 8.4873$

- So the 95% CI is:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{n_1+n_2-2,\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$
$$= -26.3233 \pm 2.101 \times 8.4873 = -26.3233 \pm 17.832$$
$$= (-44.155, -8.491).$$

- Outcomes of a test as level α can be determined by checking if *zero* lies inside the 100(1 – α)% CI.
  - Reject $H_0 : \mu_1 - \mu_2 = 0$ at level α if zero lies outside $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$
  - Don't Reject $H_0 : \mu_1 - \mu_2 = 0$ at level α if zero lies inside $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$

Assumptions for the two sample t-test

- The two samples are independent.
- Observations in each sample are independent of each other.
- The data comes from a normal or approximate normal distribution (can be checked with a QQ-plot)
- The standard deviations for both populations are the same
  - Check this using a boxplot
  - Check how close s_1 and s_2 are.

Checking Normality

- When checking normality of the data for two sample t-tests, we need to obtain a separate normal QQ-plot for each sample.
- We may assume same variance, but they could have different means, hence different normal distributions.

Checking Equal Variance

- Rule of thumb: The ratio of standard deviations should be less than 2.
- Further, the length of a boxplot also provides a rough idea on the spread of data so equal variance assumption is appropriate if the length of the longer boxplot is less than twice the length of the shorter one.

Two sample t-test with Unequal Variances

- If the ratio of the standard deviations is too large, we can't pool the variances together (equal variance assumption is not appropriate)
- The test statistic in this case is:

$$\tau_{obs} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- There is no exact sampling distribution for this statistic, but it can be shown that it is approximately a t-distribution with degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

  - Known as the Welch-Satterthwaite modification

Experimental Design

Randomised Experiments

- Randomised experiments are the gold standard for data collection, but they do not always ensure an unbiased perspective into the cause and effect relationship.
- In the diet example, we assumed at the design stage of the experiment that the 20 rats came from a homogeneous background, i.e.:
  - They were all roughly the same age
  - They had similar diets
  - Their living conditions were the same
- The rats were also allocated totally at random to the two treatment groups to the only systematic difference was their diet.
- We then concluded that the diet had different effects, but this is only true if there are **no other systematic differences** between the two groups.

Confounding

- Consider the following extreme case: Diet 1 rats were all female and Diet 2 rats were all male.
- Then the difference between the two sample average weight gains could be due to sex differences OR diet differences and it would be impossible to separate the two effects.
- So we cannot make a valid conclusion about the diet in this case.
- This is an example of **confounding**.
- As a result of this, in setting up the experiment the two samples need to be identical at the start so the only systematic differences between them is the treatment applied.
- In this example, sex is called the **nuisance factor** (a factor that may affect the result but is not of interest)

Blocking

- One strategy to control a nuisance factor is to first group individuals based on this variable into groups, and then randomise cases within each group to the treatment groups.
  - Known as Blocking
- For instance, we are investigating the effect of a drug on heart attacks, but we suspect a patient being low or high risk would have some impact on the study.
- Instead of allocating treatment at random from the beginning, we might first split the patients in the study into low-risk and high-risk groups/blocks.
- We can then randomly assign half the patients from each block to the control group and the other half to the treatment group.

- This strategy ensures that each treatments group has roughly an equal number of low-risk and high-risk patients.

Paired data and Paired t-test

- Suppose the rats in our earlier experiment do not come from a homogeneous background.
  - For instance, the rats could have very different initial weights.
- To account for this, we can pair the rats up as follows:
  - Pair 1: Two fattest rats
  - Pair 2: Next two fattest rats
  - ⋮
  - Pair n: Thinnest two rats.
- This way we get n pairs of rats with similar initial weight
- We assume the two rats in each pair are the same and randomly allocate one to Diet 1 and the other to Diet 2
- In this situation, only the difference in weight gain within each pair is analysed.

- If the treatments have the same effect, the differences in weight gain should have a distribution with zero mean. i.e. $\mu_D = 0$.
- We now have a sample of n differences d_i
- We can then conduct a one-sample t-test on the differences, if the assumptions of a one-sample t-test are satisfied.
- We are testing:
  $H_0 : \mu_D = 0 \quad \text{against} \quad H_1 : \mu_D \neq 0.$
  with test statistic:
  $$\tau = \frac{\overline{D}}{S_D/\sqrt{n}} \sim t_{n-1}, \quad \text{if } H_0 \text{ is true.}$$
- In the test static:
  - $\overline{D}$ is the sample average of the differences;
  - S_D is the sample standard deviation of the differences;
  - there are n pairs of data.
- The test is carried out the same as a one-sample t-test.
- Confidence intervals are of the form:
  $\text{estimate} \pm \text{critical value} \times \text{estimated s.e.}(\text{estimate})$

Assumptions for the Paired t-test

- We can assume the differences are normally distributed (can check with QQ-plot)
- We are making no assumptions about the original observations.
- We are assuming that the differences are independent.

Paired vs Two-sample t-tests

- To do a paired t-test, two groups of data collected must be paired
- For two independent groups/samples, a two sample t-test is appropriate.
- Examples of paired data:
    - Before and after
    - The left and right side of a person
    - Two individuals at the same age/weight
- One way to determine if a design is paired or not is to ask the question:
    - Would any information be lost if the data were shuffled?
    - If it is lost, do not use paired test.

TOPIC 9 – Inference Regarding Proportions

Motivating Example

- In past years, each year 15% of people who insured their car made a claim.
- This year, out of a random sample of 400 car insurance policies, 76 of those made a claim.
- *Is there any evidence that the population proportion of people that made a claim has increased?*
    - Let X be the number of people who claimed insurance in the sample:
      $X \sim B(n, p),$
- We can base our test on X as large values would argue against the population proportion has increased.

One-sided Tests for Proportions

- Consider tests of $H_0 : p = p_0$ against alternatives of $H_1 : p > p_0$ or $H_1 : p < p_0$ for the distribution family B(n,p)
- We can either carry out an exact test or an approximated one.

Binomial Exact Test

- In the above example, we are testing:
  $H_0 : p = 0.15 \quad \text{against} \quad H_1 : p > 0.15.$
- The test statistic is X (the number of people who made a claim in the sample).
- Then:
  $\tau = X \sim B(400, 0.15), \quad \text{under } H_0.$
- The observed value of the test statistic is 76.
- P-value is given by:
  $$P\text{-value} = \mathrm{P}(X \geq 76) = \sum_{i=76}^{400} \mathrm{P}(X = i) = 0.0171.$$
- As the P-value is small, there is sufficient evidence against H_0. i.e. there is evidence that the population proportion of people who made a claim has increased.

Binomial Approximated Test

- For X ~ B(n,p) with sufficiently large *n*, the CLT can be applied:
  $X \overset{approx}{\sim} N(np, np(1-p).)$
- This means that:
  $$\frac{X - np}{\sqrt{np(1-p)}} \overset{approx}{\sim} N(0, 1).$$
- We can then re-do the car insurance example with an approximated P-value.
- Using CLT we have:
  $X \simeq Y \sim N(np, np(1-p)) = N(60, 51), \quad \text{under } H_0$

- The approximated P-value is:

$$\begin{aligned}
\mathrm{P}(X \geq 76) &= 1 - \mathrm{P}(X \leq 75) \\
&\simeq 1 - \mathrm{P}(Y \leq 75.5), \quad \text{where } Y \sim N(60, 51); \\
&= 1 - \mathrm{P}\left(Z \leq \frac{75.5 - 60}{\sqrt{51}}\right), \quad \text{where } Z \sim N(0, 1); \\
&= 1 - \mathrm{P}\left(Z \leq 2.1704\right) \\
&= 1 - 0.985 \\
&= 0.015.
\end{aligned}$$

- (Not required now) It should be noted that there is a slight variation of this approximated test due to a different continuity correction which has test statistic:

$$\tau = \frac{\left(|X - np| - \frac{1}{2}\right)^2}{np} + \frac{\left(|n - X - n(1-p)| - \frac{1}{2}\right)^2}{n(1-p)},$$

Two-sided Test for Proportion

- General alternative to $H_0 : p = p_0$ is $H_1 : p \neq p_0$
- Continue to use $X \sim B(n, p_0)$ as the test statistic under H_0.
- Large values of $|X - np_0|$ argue against H_0 (i.e. values that are far away from np_0)

EXAMPLE:

- A company claims that 93% of all items produced are non-defective. A random sample of 100 items is taken.
- If the observed number of defectives in the sample was 11, is there any reason to doubt the 93% claim?
- Let p be the probability of a defective item. Then we are testing $H_0 : p = 0.07$ against $H_1 : p \neq 0.07$.
- As we have a large sample size (n = 100), we will use the binomial approximated test.
- Under H_0:

$$X \sim B(100, 0.07) \simeq Y \sim N(np, np(1-p)) = N(7, 6.51).$$

- Large value of $|X - 7|$ will argue against H_0.
- P-value:

$$\begin{aligned}
P\text{-value} &= \mathrm{P}(|X - 7| \geq |11 - 7|) \\
&= \mathrm{P}\left(|X - 7| \geq 4\right) \\
&\simeq \mathrm{P}\left(|Y - 7| \geq 3.5\right), \quad \text{using continuity correction;} \\
&= \mathrm{P}\left(\frac{|Y - 7|}{2.5515} \geq \frac{3.5}{2.5515}\right) \\
&= \mathrm{P}\left(|Z| \geq 1.37\right), \quad \text{where } Z \sim N(0, 1); \\
&= 2\left(1 - \Phi(1.37)\right) \\
&= 0.1706.
\end{aligned}$$

CI for Proportions

- The confidence intervals for proportions are based on the standardised scores:

$$Z' = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}},$$

  - Where $\hat{p} = \frac{X}{n}$ is the sample proportion.
- If *n* is sufficiently large, then:
  $Z' \simeq N(0,1)$
- The standard error of $\hat{p}$ depends on the unknown parameter *p* so we use $\hat{p}$ as our estimate for *p* in:

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{p(1-p)}{n}.$$

- The 100(1 – α)% CI for *p* is:

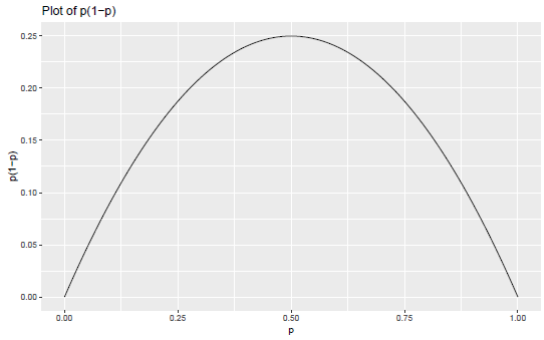
$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

- Any plausible *p* must be in the interval (0, 1) as it is a proportion. Any bound of a CI that is outside that range must be manually reset to 0 or 1.
- We cannot simply use this CI to test for H_0, which means that if you are using a CI to test for $H_0 : p = p_0$, you should use:

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}}.$$

Conservative CI for Proportion

- It can be time consuming to calculate a new CI every time we test a different p_0.
- However, based on the structure of the s.e.($\hat{p}$), it is possible to make the CI as wide as possible, which makes the CI **conservative**.
- The following is a plot of *p(1 – p)* against *p*:



- The function is maximised at p = 0.5. As a result:

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \leq \sqrt{\frac{1}{n} \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{\frac{1}{4n}}.$$

- So the conservative CI is given by:

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{4n}}$$

One-sided CI for a proportion

- For a one-sided CI, the boundaries will be 0 or 1 instead of $\pm\infty$
- The approximate one-sided 100(1 – α)% CO for p:

  For $H_1 : p < p_0$: $\left(0, \hat{p} + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}\right)$.

  For $H_1 : p > p_0$: $\left(\hat{p} - z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}, 1\right)$.

- If you are just interested in the range of plausible population proportion and not testing against a particular p_0, replace p_0 with $\hat{p}$.

Calculating Sample Sizes

- There are circumstances where the sample size estimate is needed in advance when planning a survey or experiment.
- The CI allows us to link the maximum margin of errors to the minimum required sample size.
- The normal approximation CI has the form:

  $$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

- The issue is that we have no knowledge of p or $\hat{p}$. Thus, we can alternatively use the conservative CI (worst case scenario) to solve for sample size n:

  $$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{4n}}$$

EXAMPLE:

- A radio station wishes to carry out a survey to determine what proportion of their listeners would stop listening if they no longer broadcast news on the half hour.
- *What sample is required if they want their estimate to be within 0.04 of the actual proportion with probability equal to 0.90?*
- We use the conservative estimate to get:

  $$z_{0.05}\sqrt{\frac{1}{4n}} = 1.645\sqrt{\frac{1}{4n}} \leq 0.04 \quad \Rightarrow \quad \frac{1.645}{2 \times 0.04} \leq \sqrt{n}$$
  $$\Rightarrow \quad n \geq (20.5625)^2 = 422.8164.$$

- So in this case, the minimum sample size needed is 423.

Two-sample test of Proportions

- Given two independent samples/populations, suppose we want to test:

  $H_0 : p_1 - p_2 = d_0$ against $H_1 : p_1 - p_2 \neq d_0$,

  i.e. we are assuming the two proportions are different from the beginning
- We can test for an one-sided alternative as well i.e. $H_1 : p_1 - p_2 < d_0$ or $H_1 : p_1 - p_2 > d_0$.
- We will then base our test on our best estimate of $p_1 - p_2$ which is $\hat{p}_1 - \hat{p}_2$.

- If n_1 and n_2 are sufficiently large (for CLT to apply), then:

$$\hat{p}_1 = \frac{X_1}{n_1} \overset{approx}{\sim} N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \text{and}$$
$$\hat{p}_2 = \frac{X_2}{n_2} \overset{approx}{\sim} N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

- Given the two samples are independent, we have:

$$\hat{p}_1 - \hat{p}_2 \overset{approx}{\sim} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$
$$= N\left(d_0, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right), \quad \text{under } H_0.$$

- Thus we have a test statistic of (using $\hat{p}_1$ and $\hat{p}_2$ as estimates):

$$\tau = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \overset{approx}{\sim} N(0,1), \quad \text{under } H_0.$$

- The observed value of the test statistic, P-value and conclusion all work the same as the previous hypothesis test.
- We can set $d_0 = 0$ if we want to test $H_0 : p_1 = p_2$ (the equality of two proportions)

Testing for the Equality of Two Proportions

- We are testing:

$$H_0 : p_1 = p_2 (= p) \quad \text{against} \quad H_1 : p_1 \neq p_2.$$

- Then:

$$\hat{p}_1 - \hat{p}_2 \overset{approx}{\sim} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$
$$= N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right), \quad \text{under } H_0.$$

- By assuming $p_1 = p_2$ under H_0, the variance would be equal across the two populations.
- But we still need to estimate the unknown common proportion p.

Pooled Sample Proportion

- We use the pooled sample proportion to estimate p:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{n_1}{n_1 + n_2}\hat{p}_1 + \frac{n_2}{n_1 + n_2}\hat{p}_2 = \frac{X_1 + X_2}{n_1 + n_2}$$

  - It is a weighted average between $\hat{p}_1$ and $\hat{p}_2$
  - The more we sample from a population over the other, more weight will be allocated

- Using this pooled estimate, we obtain the following test statistic:

$$\tau = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1), \quad \text{under } H_0,$$

  where $\hat{p}_i = \frac{X_i}{n_i}$ and $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$.

- This can be done one-sided as well.
- We don't apply continuity correction when we are testing for equality of proportions.

EXAMPLE:

- The number of students that pass at two different schools is recorded:
  - 40 out of 70 pass in School 1
  - 45 out of 100 pass in School 2
- Is there any difference between the two schools in their overall pass rates?
- Let p_i be the passing proportion for School *i*. Then we are testing:

$$H_0 : p_1 = p_2 (= p) \quad \text{against} \quad H_1 : p_1 \neq p_2.$$

- Based on the info provided, we have:

$$\hat{p}_1 = 0.5714 \text{ and } \hat{p}_2 = 0.45.$$

- Under H_0, the pooled proportion is:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{40 + 45}{70 + 100} = 0.5.$$

- The test statistic is:

$$\tau = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1), \quad \text{under } H_0,$$

- The observed value of the test statistic is:

$$\tau_{obs} = \frac{0.5714 - 0.45}{\sqrt{0.5(0.5)\left[\frac{1}{70} + \frac{1}{100}\right]}} = 1.56.$$

- Both a large and small value of τ would argue against H_0 in favour of H_1.
- The P-value is:

$$\begin{aligned} P\text{-value} &= \mathrm{P}(|Z| \geq |\tau_{obs}|) \\ &= 2(1 - \mathrm{P}(Z \leq 1.56)) \\ &= 2 \times 0.0594 \\ &= 0.1188. \end{aligned}$$

- As the P-value is large, there is insufficient evidence to reject H_0. i.e. there is insufficient evidence to conclude that there is a difference in the pass rate between the two schools.

Confidence Interval for $p_1 - p_2$

- The CI is given by:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

TOPIC 10 – Simple Linear Regression

Bivariate Data

- So far we have been focusing on univariate data only (observation of a single variable)
- Now we will be looking for the relationship between two or mare variables (multivariate data)
- Bivariate data specifically looks at only 2 variables and will be the focus of this topic.
- Scatterplots are a type of graph used to study the relationship between two numerical variables. They simply plot the points (x_1, y_1), … (x_n, y_n).

Creating a Scatterplot with *geom_point()*

- The function geom_point() of ggplot2 adds a layer of points, which create a scatterplot.
- The code is as follows:

```
ggplot(data = cheese) +
  geom_point(aes(x= Lactic, y = taste))
```

Correlation Coefficient

- The correlation coefficient, *r*, is a numerical index that measures the degree of linear association between *x* and *y*.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2) \cdot (\sum_{i=1}^{n}(y_i - \overline{y})^2)}}$$

- With:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = (n-1)s_x^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = (n-1)s_y^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

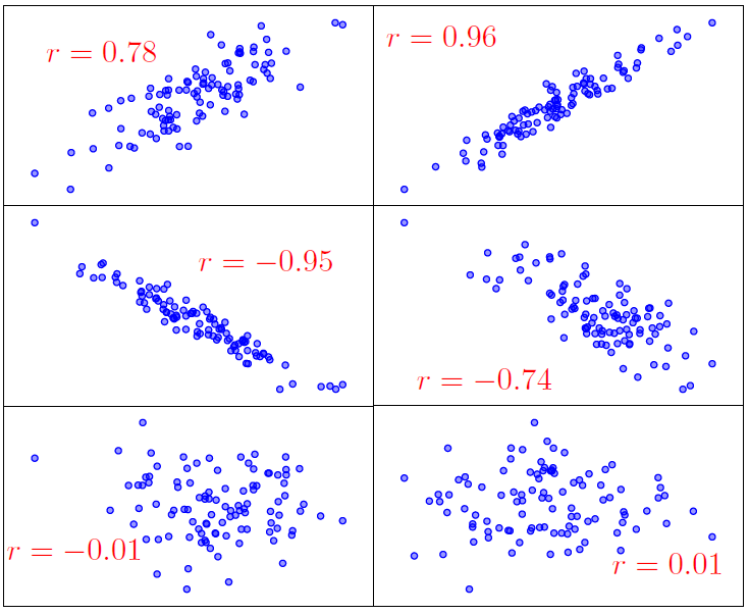
- So it follows that:
  - $S_{xx} = sum(x^2) - \frac{[sum(x)]^2}{n}$
  - $S_{yy} = sum(y^2) - \frac{[sum(y)]^2}{n}$
  - $S_{xy} = sum(xy) - \frac{[sum(x)]*[sum(y)]}{n}$
- Linear rescaling *x* or *y* does not change the correlation coefficient *r*.
- By choosing $d_i = \frac{x_i - a}{h}$, $e_i = \frac{y_i - b}{g}$ we can show that:

$r_{de} = r_{xy}$.

Properties of the Correlation Coefficient

- The correlation coefficient is always between -1 and 1: $r \in [-1, 1]$.
- If $0 < r \leq 1$, the association is positive
  - If x increases, y increases (and vice versa)
- If $-1 \leq r < 0$, the association is negative
  - If x increases, y decreases (and vice versa)
- If $r = 0$, there is no linear association, but that does not imply that there is no relationship between x and y!
- If $r = 1$, there is a perfect positive linear relationship between x and y
  - i.e. all obs (x_i, y_i) lie on a straight line with a positive slope.
- If $r = -1$, there is a perfect negative linear relationship between x and y.
  - i.e. all obs (x_i, y_i) lie on a straight line with a negative slope.

Visualising Correlation



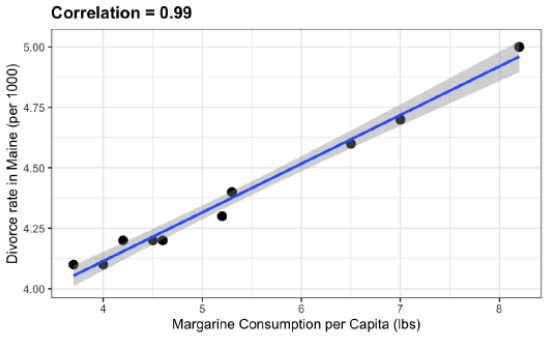
- Correlation should only be used to measure the strength of a **linear relationship**. If the relationship is non-linear, it cannot be measured by *r*.

Anscombe's Quartet

- Anscombe's quartet comprises four data sets that:
  - Have nearly identical descriptive statistics
  - Yet have very different distributions and appear different when graphed
- This is a famous example showing the limitations of the correlation coefficient.
- As such, it is always important to visualise data as part of the analysis process.

Zero Correlation & Independence

- If variables are independent, they can't be associated and thus have zero correlation.
- However, the reverse of this is not true.
- If two variables' correlation coefficient is close to 0, it doesn't necessarily mean they are independent.
- Further, two variables can be related without either one causing the other to change.
    - For example, the following graph shows a very strong correlation between divorce rates and margarine consumption:



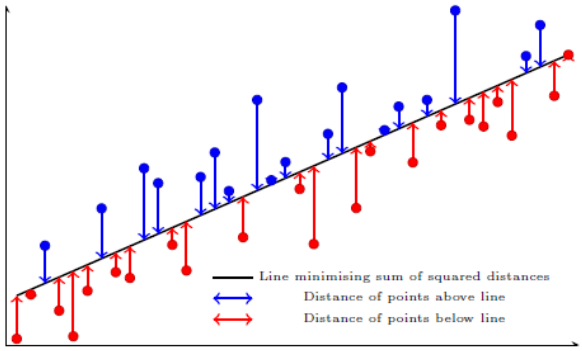
    - However, it is obvious that neither of these cause one another.

Simple Linear Regression

- Regression considers two continuous variables measured on each observation:
    - *Y* (the dependent/response variable) is considered as varying about
    - *X* (the independent/predictor variable)
- Simple Linear Regression:
    - **Simple**: the response variable depends on only one predictor variable
    - **Linear**: the relationship is a straight line
- When modelling a linear relationship, we can use the following:

$$Y = a + bX$$

- Now we must determine the values of *a* and *b*, i.e. we need to quantify a method to find the line of best fit.
- Firstly, find the vertical gap between each point and the line. These vertical gaps are called residuals, and the line of best fit should have the smallest residuals.



- Now, we define our line of best fit as the line that minimises:

$$\sum_{i=1}^{n} \text{residual}^2$$

- The line obtained by this method is called the **least squares regression line**.

Terminology and Formula for the Line

- A more accurate description of reality for the relationship of *X* and *Y* may be the following probabilistic model:
  $$Y = \underbrace{\alpha + \beta X}_{\widehat{Y}} + \varepsilon = \hat{Y} + \varepsilon.$$
- Where:
  - α: True intercept.
  - β: True slope.
  - ε: Random variation/error away from the line, according to a specified probability distribution with mean zero.
- Note that:
  $$E(Y|X = x) = \alpha + \beta x,$$
  denotes the mean of *Y* given *X* = *x*

Least Squares Regression Line

- Suppose we have observed bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$.
- By the method of least squares, choose $\hat{\alpha}$ and $\hat{\beta}$ such that:
  $$M(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}^2.$$
  is minimised
- The estimated regression line is:
  $$\widehat{Y} = \hat{\alpha} + \hat{\beta} X$$
  Where:
  $$\text{Slope/regression coefficient:} \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}}$$
  $$\text{Intercept:} \quad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}.$$

Regression Line: Slope

- In general, $\hat{\beta}$ is the expected increase in *Y* for a 1-unit increase in *X*.
- From the correlation coefficient, we get:
  $$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{xx}}} \frac{\sqrt{S_{yy}}}{\sqrt{S_{yy}}} = r\sqrt{\frac{S_{yy}}{S_{xx}}}$$
- This shows that $\hat{\beta}$ and *r* have the same sign

Regression Line: Intercept

- $\hat{\alpha}$ is the intercept, the value of $\hat{Y}$ when *X* = 0.
- If you need to add a regression line to a scatter plot, you need an additional point to form the line
- The regression line passes through the component-wise mean $(\overline{x}, \overline{y})$ as well

Testing Regression

- Using regression to test a prediction is useless if there is no actual linear relationship between *X* and *Y*, but rather the true slope of the regression line is 0.
- When the fitted line is flat, there is no reason to use *X* in order to predict *Y*.
- To do this formally, use the test hypotheses:
  $H_0 : \beta = 0 \quad \text{against} \quad H_1 : \beta \neq 0.$
- So we need to know the sampling distribution of $\hat{\beta}$

Distribution Assumption on ε

- The least squares method doesn't depend on the distribution of ε, but we will still have to specify one to find the sampling distribution of $\hat{\beta}$.
- In this case we set the model to:
  $$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$
  where ε_i are i.i.d. rvs with
  $$\varepsilon_i \sim N(0, \sigma^2).$$
- The parameter σ measures the inherent variability of the response variable *Y* at a given *X*, i.e. how far the *Y* values are typically away from the regression line.
- Σ can be estimated by $s_{Y|X}$ (the sample standard deviation of the observed residuals), where
  $$s_{Y|X} = \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}} = \sqrt{\frac{S_{yy} - \hat{\beta} S_{xy}}{n-2}}$$

Testing β

- One can show that:
  $$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$
- We do not have any information on σ, but we can estimate it with $s_{Y|X}$
- Thus, we will be using a t-test.
- We are testing:
  $H_0 : \beta = 0 \quad \text{against} \quad H_1 : \beta \neq 0$
- We assume the linear regression model is appropriate
- Test statistic is:
  $$\tau = \frac{\hat{\beta}}{s_{Y|X}/\sqrt{S_{xx}}} \sim t_{n-2} \quad \text{under } H_0.$$
- The rest of the procedure is similar to a regular two-sided t-test.
- This also means that:
  $$\text{s.e.}(\hat{\beta}) = \frac{s_{Y|X}}{\sqrt{S_{xx}}}.$$

Confidence Interval for β

- A two-sided CI for β is:
  $$\hat{\beta} \pm t_{n-2,\alpha/2} \times \text{s.e.}(\hat{\beta}) = \hat{\beta} \pm t_{n-2,\alpha/2} \times \frac{s_{Y|X}}{\sqrt{S_{xx}}}$$
- Note: α here is the significance level (not the intercept)

Test for the Correlation Coefficient

- Suppose we want to test:
  $H_0 : \rho = 0 \quad \text{and} \quad H_1 : \rho \neq 0$
  where *p* is the population correlation coefficient between *X* and *Y*.
- It can be shown that the test statistic for testing H_0 : β = 0 can be written as:
  $$\frac{\hat{\beta}}{s_{Y|X}\sqrt{S_{xx}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$
- This means that the test on *p* is the same test on β (i.e. testing H_0 : β = 0 is the same as testing H_0 : *p* = 0)
- Thus, if there is a significant result, there is a significant regression.

Test for α

- There is also a test (and confidence interval) for the intercept.
  - H_0 : α = 0 vs H_1 : α ≠ 0 (testing if line goes through origin)
- Testing α has a similar structure to testing β:
- Observed value of the test statistic is:
  $$\tau_{obs} = \frac{\hat{\alpha}}{\text{s.e.}(\hat{\alpha})}, \quad \text{where} \quad \text{s.e.}(\hat{\alpha}) = s_{Y|X}\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}.$$
- P-value:
  $P(|t_{n-2}| \geq |\tau_{obs}|)$.

Test Conclusion and Interval Interpretation

- Based on the outcome of β, we can decide whether making predictions is reasonable.
- If we do not reject H_0 : β = 0, we cannot go any further (no regression)
- If we do reject H_0 (the regression has relevance), then we can use the regression line to estimate *Y* for particular values of *X*.
- Note that:
  - Regression theory treats *Y* as the response variable (varying about *X*).
  - Therefore: **cannot use *Y* to estimate *X*.**
  - Can only use the equation for the range of values of *X* that we have already observed (do NOT extrapolate)

Fitted or Predicted Values

- We have rejected H_0, so we can use the regression line for prediction.
- Given that:
  $$Y = \underbrace{\alpha + \beta X}_{\hat{Y}} + \varepsilon; \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$
- Estimate $\hat{Y}$ with $\hat{y} = \hat{\alpha} + \hat{\beta}x$
- Since ε has zero mean, this implies:
  $\mu_{Y|X=x} = \text{E}(Y|X = x) = \hat{Y}$.

Confidence Interval for Fitted Values

- Fitted value is an estimate of the mean of *Y* when *X* = *x*.
- We want a CI of this estimation, so we must first find the standard error of $\hat{y}$:
  - Firstly:

$$\text{s.e.}(\overline{Y}) = \sqrt{\frac{s^2_{Y|X}}{n}}, \quad \text{and} \quad \text{s.e.}(\hat{\beta}) = \sqrt{\frac{s^2_{Y|X}}{S_{xx}}}$$

  - So then:

$$\begin{aligned}\text{s.e.}(\widehat{Y}|X = x) &= \text{s.e.}(\hat{\alpha} + \hat{\beta}x) \\ &= \text{s.e.}(\overline{Y} - \hat{\beta}\overline{x} + \hat{\beta}x) \\ &= \text{s.e.}(\overline{Y} + \hat{\beta}(x - \overline{x})) \\ &= \sqrt{s^2_{Y|X}\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)},\end{aligned}$$

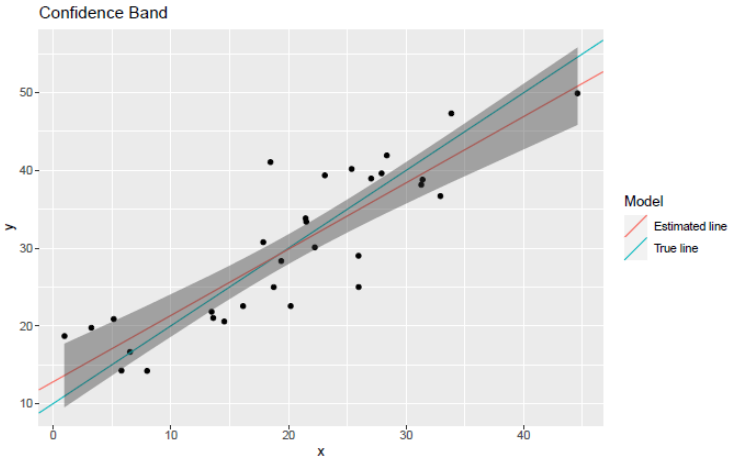
- Note: the s.e. of $\hat{Y}$ depends on *x*
  - Minimised at $\overline{x}$ and increases as *x* moves further away from $\overline{x}$

Confidence Bands for Estimated Line

- Confidence interval for $E(\hat{Y}|X = x)$ is a variable depending on *x*
- The term **confidence band** is used to represent the uncertainty in an estimate of the entire curve or function.
- The 100(1 – α)% CI for $\hat{Y}|X = x$ is:

$$\hat{\alpha} + \hat{\beta}x \pm t_{n-2,\alpha/2} \times s_{Y|X}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$

- The confidence interval for each true mean will have a different width depending on the distance along the line.



Prediction Bands for an Individual

- We can also use the regression line:

$$Y = \alpha + \beta X + \varepsilon; \quad \varepsilon \sim N(0, \sigma^2)$$

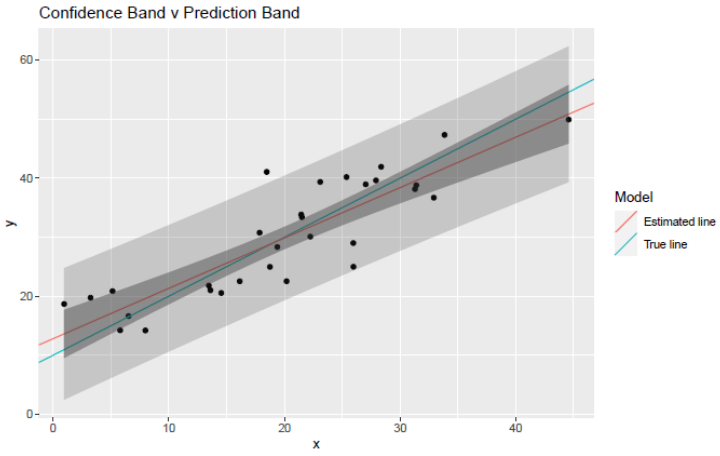
  to estimate the value of *Y* for a new individual
- The predicted value for a new individual at *X* = *x* is the same value as the mean at *X* = *x*.
  - i.e. the point on the line

- However, an individual will have a larger standard error. For an individual:

$$\text{s.e.}(Y|X=x) = \text{s.e.}([\hat{Y}|X=x] + \varepsilon) = s_{Y|X}\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$

- The 100(1 – α)% CI for $Y|X = x$:

$$\hat{\alpha} + \hat{\beta}x \pm t_{n-2,\alpha/2} s_{Y|X}\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$



Regression Diagnostics

- There are no assumptions about the distribution of predictors (*X* is measured without error)
- Assume:
  - The relationship between *X* and *Y is linear*
  - The *Y* values are normally distributed about the regression line.
  - Residuals have constant variance
- Residuals are normally distributed:
  - Assume errors to be normally distributed with, ***ε ~ N(0, σ²)***
- If there is only a linear relationship with *X* and *Y*, the remaining variation in the residuals is basically noise with constant variance as $\varepsilon \sim N(0, \sigma^2)$.
  - If there is no obvious pattern in the residual plot, the assumptions of linearity and constant variance are valid

Coefficient of Determination

- *Total Sum of Squares = Residual S.S. + Regression S.S.*
  - Thus:

$$1 = \underbrace{\frac{\text{Residual S.S.}}{\text{Total S.S.}}}_{\substack{\text{proportion of variation that} \\ \text{didn't explained by regression}}} + \underbrace{\frac{\text{Regression S.S.}}{\text{Total S.S.}}}_{\substack{\text{proportion of variation that} \\ \text{explained by regression}}}$$

- Both terms can be interpreted as a **ratio of variation**
- It can be shown that:

$$\frac{\text{Regression S.S.}}{\text{Total S.S.}} = \hat{\beta}\, S_{xy}\frac{1}{S_{yy}} = \frac{S_{XY}^2}{S_{XX}} \times \frac{1}{S_{YY}} = \frac{S_{XY}^2}{S_{XX}S_{YY}} = r^2$$

- This means that r^2 represents the percentage of variation in *Y* explained by the linear regression of *Y* on *X*.
- r^2 is called the **coefficient of determination**. It is a measure of how much variation in *Y* can be explained by knowing the *X* value
- The closer r^2 is to 1, the more worthwhile the regression is. (i.e. predicted values will be more accurate)
- Note: you can get a significant regression with a relatively low r^2
- As a rough rule, we aim for r^2 of **0.7 or more**.

TOPIC 11 – Categorical Data Analysis

Testing Categorical Data

- In many applications, data is classified into distinct categories.
- These categories need not have a natural numerical ordering.
- For example, in an experiment involving dihybrid cross of flies, 148 progeny were classified by phenotype as follows:

AB	Ab	aB	ab	Total
87	31	25	5	148

- We want to evaluate whether the data resembles a particular distribution
- For instance:
  - We have a ratio of 9:3:3:1 for AB:Ab:aB:ab.
  - In terms of probability the ratio is $\frac{9}{16}:\frac{3}{16}:\frac{3}{16}:\frac{1}{16}$
- If we use groups 1 to 4 to represent the phenotypes, then the probability model can be represented as:

$$p_1 = \frac{9}{16}, \quad p_2 = \frac{3}{16}, \quad p_3 = \frac{3}{16}, \quad p_4 = \frac{1}{16}.$$

where p_i is the probability of an observation to fall into group i, such that:

$$\sum_{i=1}^{g} p_i = 1.$$

Motivational Setting

- Suppose we have n independent trials with X successes and $n - X$ failures. i.e. $X \sim B(n, p)$
- We wish to test $H_0 : p = p_0$ against $H_1 : p \neq p_0$. If H_0 is true then:

	Success	Failure
Observe	$O_1 = X$	$O_2 = n - X$
Expect	$E_1 = np_0$	$E_2 = n(1 - p_0)$

- Large values of $|X - np_0|$ support H_1
- Assuming that normal approximation is appropriate, then large values of

$$\tau = \frac{(X - np_0)^2}{np_0(1 - p_0)}, \quad \text{support } H_1.$$

- Notice that:

$$\begin{aligned}(O_2 - E_2)^2 &= [(n - X - (n - np_0))]^2 \\ &= (X - np_0)^2 = (O_1 - E_1)^2.\end{aligned}$$

- Also:

$$\frac{1}{np_0} + \frac{1}{n(1 - p_0)} = \frac{1}{np_0(1 - p_0)}.$$

- Thus,

$$\tau = \frac{(X - np_0)^2}{np_0(1-p_0)} = (X - np_0)^2 \left[\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right]$$
$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

- This is a special case of Pearson's x^2 statistic.

Pearson's x^2 Goodness of Fit Test

- Assume we have *g* categories (not just success/failure) and H_0 specifies a model giving expected frequencies for each category.
- For instance, we could be testing:

$$H_0 : p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}, \quad \text{against} \quad H_1 : \text{not } H_0,$$

- It is not necessary that all the proportions are difference for H_1 to be true, only that at least one is not as specified.
- We can test the claim by comparing the observed frequency with the expected frequency under H_0.
- The Pearson's x^2 test-statistic (without continuity correction) is:

$$\tau = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i},$$

where:
  - O_i is the observed frequency in the ith category;
  - $E_i = np_i$ is the expected frequency if H_0 is true;
  - $n = \sum O_i = \sum E_i$ is the total number of observations.

- If the data supports the proposed model, then we expect (O_i, E_i) to be close together and $(O_i - E_i)^2$ wouldn't be too large.
- Essentially, we reject the model if the test statistic is too large.
- The sampling distribution of the statistic has a chi-squared distribution with $g - 1$ degrees of freedom:

$$P\text{-value} = \mathrm{P}(\chi^2_{g-1} \geq \tau_{obs}).$$

- The x^2 test should only be used when expected frequencies are all greater than 5.

Continuity Correction

- Categorical data is discrete, but the test involves the test statistic to a continuous distribution.
- Therefore we should use the continuity correction, and then test statistic of the x^2 goodness of fit test becomes:

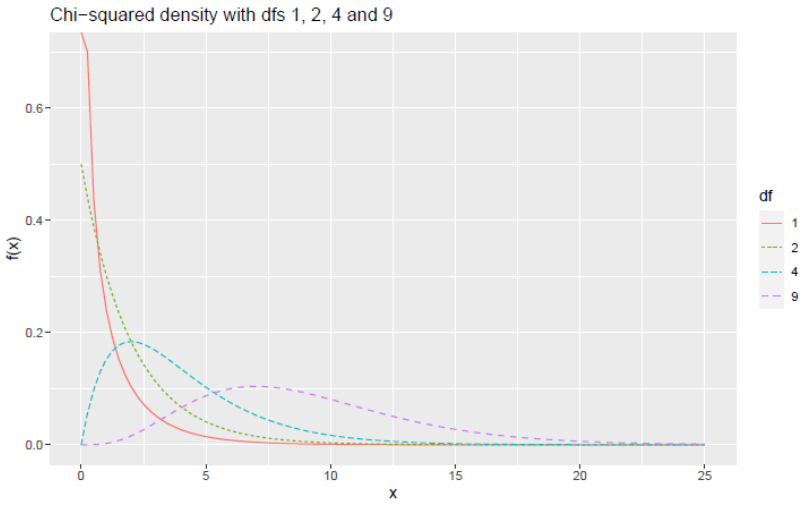
$$\tau = \sum_{i=1}^{g} \frac{\left(|O_i - E_i| - \frac{1}{2}\right)^2}{E_i}.$$

- This is known as **Yates' continuity correction**

- Note: the continuity correction is always going to make the test statistic smaller, and the test becomes more conservative.
- In some rare cases, the correction will be reduced so that it is not bigger than the differences themselves.

The x^2 distribution

- A x^2 random variable can only take non-negative values.
- The distribution is not symmetric, but rather is right-skewed.



- Note:
  - $x_1^2 = Z^2$, where Z ~ N(0,1)
  - If $X \sim x_v^2$ then E(X) = *v* and Var(X) = 2*v*

The x^2 Table

- The principal interest in the x^2 distribution is the calculation of P-values of the Goodness of Fit test.
- The x^2 tables typically give:
  $$P(\chi_\nu^2 \geq x) = p$$
  where *v* is the degrees of freedom (row), *p* is the upper tailed probabilities (column) and *x* is given in the body of the table.
- In R, the following function are useful:
  - pdf: `dchisq(x, df = nu);`
  - cdf: `pchisq(q, df = nu);`
  - quantile or critical values: `qchisq(p, df = nu);`
  - random numbers: `rchisq(n, df = nu).`

EXAMPLE:

- Consider the Phenotype example:

group	1	2	3	4	Total
phenotype	AB	Ab	aB	ab	
O_i	87	31	25	5	148

- We are testing:

$$H_0 : p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16} \quad \text{against} H_1 : \text{not } H_0.$$

- Under H_0, the model specifies the following expected frequencies:

group	1	2	3	4	Total
E_i	$\frac{9}{16}\times 148 = 83.25$	$\frac{3}{16}\times 148 = 27.75$	$\frac{3}{16}\times 148 = 27.75$	$\frac{1}{16}\times 148 = 9.25$	148

- The test statistic of the test is:

$$\tau = \sum_{i=1}^{4} \frac{\left(|O_i - E_i| - \frac{1}{2}\right)^2}{E_i} \sim \chi^2_{g-1} = \chi^2_4, \quad \text{under } H_0.$$

- The observed value of the test statistic is:

$$\begin{aligned}\tau_{obs} &= \frac{\left(|87 - 83.25| - \frac{1}{2}\right)^2}{83.25} + \frac{\left(|31 - 27.75| - \frac{1}{2}\right)^2}{27.75} \\ &\quad + \frac{\left(|25 - 27.75| - \frac{1}{2}\right)^2}{27.75} + \frac{\left(|5 - 9.25| - \frac{1}{2}\right)^2}{9.25} \\ &= 0.1269 + 0.2725 + 0.1824 + 1.5203 \\ &= 2.1021.\end{aligned}$$

- The P-value for testing the fit of the model is:

$$P\text{-value} = \mathrm{P}(\chi^2_3 \geq 2.1021) > 0.1.$$

- Since the P-value is large, we can conclude that the data is consistent with H_0, i.e. the observed ratio is not significantly different from 9:3:3:1

EXAMPLE:

- The number of fatal accidents on NSW roads in months with 31 days in 1993 were:

Jan	Mar	May	July	Aug	Oct	Dec
44	56	37	42	59	59	63

- Test the claim that the accident rate is the same for all months
- Let p_i denote the probability that a fatal accident is allocated to month *i*.
- We are testing:

$$H_0 : p_i = \frac{1}{7}, \quad i = 1, 2, \ldots, 7, \quad \text{against} \quad H_1 : \text{not } H_0.$$

- The total number of accidents is 360. Thus E_i = 360/7 = 51.43
- The test statistic is:

$$\tau = \sum_{i=1}^{4} \frac{\left(|O_i - E_i| - \frac{1}{2}\right)^2}{E_i} \sim \chi^2_{g-1} = \chi^2_6, \quad \text{under } H_0.$$

- Then we get the following information:

Month	Jan	Mar	May	July	Aug	Oct	Dec	Total
O_i	44	56	37	42	59	59	63	360
E_i	51.43	51.43	51.43	51.43	51.43	51.43	51.43	360
$\frac{(\lvert O_i - E_i \rvert - \frac{1}{2})^2}{E_i}$	0.93	0.32	3.77	1.55	0.97	0.97	2.38	10.89

- This means $\tau_{obs} = 10.89$. Thus:
  $P\text{-value} = P(\chi^2_6 \geq 10.89) > 0.1.$
- Large P-value means the data is consistent with H_0.

Further Application of Goodness of Fit Test

- If we want to check the fit of a model that involves unknown parameters, we first have to estimate the parameters with the data.
- Since we use the same data to estimate parameters and the test fit, we find the sampling distribution of x^2 has to be adjusted.
- The degrees of freedom are reduced to $g - k - 1$, where:
  - g is the number of categories
  - k is the smallest number of parameters that need to be estimated using the data.

EXAMPLE:

- 200 groups of 5 insects each were inspected.
- For each group, the number of infected insects (x) was counted.
- The data was condensed into the table below, writing x_i for the number of infected and f_i for the corresponding frequency:

x_i	0	1	2	3	4	5	Total
f_i	20	62	55	38	20	5	200

- *Does the binomial model fit the data?*
- We are testing:
  $H_0 : X \sim B(5, p) \quad \text{against} \quad H_1 : \text{not } H_0.$
- We need to estimate p. There were 1000 insects in total and 391 of them were infected. SO an estimate for p would be:
  $\hat{p} = \frac{391}{1000} = 0.391.$
- The test can be summarised into the following table:

i	0	1	2	3	4	5	Total
p_i	0.0838	0.2689	0.3453	0.2217	0.0712	0.0091	1
O_i	20	62	55	38	20	5	200
$E_i = np_i$	16.76	53.78	69.06	44.34	14.24	1.82	200

- However, one of the cells has an expected value < 5, so the x^2 test is NOT valid!
- If any E_i falls below 5, we can
  - Get a large n
  - Pool classes in a sensible way

- Here we will combine the last two cells together to get a single category for > 4, and the table becomes:

i	0	1	2	3	≥ 4	Total
p_i	0.0838	0.2689	0.3453	0.2217	0.0803	1
O_i	20	62	55	38	25	200
$E_i = np_i$	16.76	53.78	69.06	44.34	16.06	200
$\frac{(\lvert O_i - E_i \rvert - \frac{1}{2})^2}{E_i}$	0.4479	1.1082	2.6625	0.7692	4.4355	9.4233

- The test statistic of the test is:

$$\tau = \sum_{i=1}^{4} \frac{(|O_i - E_i| - \frac{1}{2})^2}{E_i} \sim \chi^2_{g-1-1} = \chi^2_3, \quad \text{under } H_0,$$

as g = 5 now

- Hence, τ_{obs} = 9.4233 and
  $P\text{-value} = P(\chi^2_3 \geq 9.4233) < 0.025.$
- The P-value is small, so we have evidence against H_0, i.e. there is a significant difference between the proposed binomial model and the data.

- A similar procedure can be used to test the fit of other discrete distributions such as Poisson and negative binomial.

Testing the Fit of a Normal Model

- Given a dataset $x_1, x_2, \ldots, x_n$ we want to test if the data come from a $N(\mu, \sigma^2)$ population.
  - We first calculate the sample mean, $\overline{x}$, and the sample variance, s^2.
  - Form a grouped frequency table and summarise the data with (ideally) 5 to 10 categories.
  - To check against normal population, work out the expected frequencies for each category by fitting $N(\overline{x}, s^2)$.
  - Calculate the x^2 test statistic as usual.
  - The calculate the P-value, use $g - 2 - 1$ df.

EXAMPLE:

- We have 30 observations corresponding to Sydney's annual rainfall from 1941-1970:

26.74	48.29	50.74	31.04	46.47	36.05
41.45	38.83	66.26	86.63	53.15	59.19
40.86	41.29	72.46	67.33	27.13	59.19
59.67	51.01	57.08	44.90	80.11	43.30
36.01	48.40	52.78	24.56	56.94	43.42

- Test if the rainfall follows a normal distribution.
- We are testing:
  $H_0 : X \sim N(\mu, \sigma^2), \quad \text{against} \quad H_1 : \text{not } H_0.$
- We estimate μ and σ^2 with:
  $\overline{x} = 49.71 \text{ and } s^2 = 229.15$
- The grouping the data into a frequency table:

Interval	$x \leq 40$	$40 < x \leq 50$	$50 < x \leq 60$	$x \geq 60$	Total
Frequency	7	9	9	5	30

- We now calculate the expected frequencies using:
  $X \sim \mathcal{N}(49.71, 229.15)$.
- Then:

$$p_1 = P(X \leq 40) = \quad \mathrm{P}\left(Z \leq \frac{40 - 49.71}{\sqrt{229.15}}\right) = P(Z \leq -0.64) = 0.2611.$$

- Thus $E_1 = 30 \times 0.2611 = 7.833$
- And similarly for the rest:

$$p_2 = P(40 < Y \leq 50) = P(-0.64 < Z \leq 0.019) = 0.2469$$
$$E_2 = 30 \times 0.2469 = 7.407.$$

  Similarly,
$$E_3 = 30 \times 0.2437 = 7.311 \text{ and}$$
$$E_4 = 30 - 7.833 - 7.407 - 7.311 = 7.449.$$

- The test statistic is:

$$\tau = \sum_{i=1}^{4} \frac{\left(|O_i - E_i| - \frac{1}{2}\right)^2}{E_i} \sim \chi^2_{g-2-1} = \chi^2_2, \quad \text{under } H_0.$$

- The observed value of the test statistic is:

$$\begin{aligned}\tau_{obs} &= \frac{\left(|7 - 7.833| - \frac{1}{2}\right)^2}{7.833} + \frac{\left(|9 - 7.407| - \frac{1}{2}\right)^2}{7.407} \\ &\quad + \frac{\left(|9 - 7.311| - \frac{1}{2}\right)^2}{7.311} + \frac{\left(|5 - 7.449| - \frac{1}{2}\right)^2}{7.449} \\ &= 0.0142 + 0.1613 + 0.1934 + 0.5099 \\ &= 0.8788.\end{aligned}$$

- Here $g = 4$ and $k = 2$, so we have 1 d.f.
- The P-value is:
  $P(\chi^2_1 \geq 0.8788) = 0.204 > 0.1$
- As the P-value is large, data is consistent with H_0, i.e. data is consistent with the normal model

- This procedure can be modified to test the goodness of fit of other continuous distributions as well
- The procedure is not unique as the number of categories is not fixed, and there are also many ways to define the boundaries of these categories

Tests for Independence

- If we have data classified according to two attributes, then we can construct a **contingency table** or a **two-way classification table** which is a convenient way of presenting the group frequencies.

- For example, we have data on 422 drivers and motorcyclists killed in NSW in 1988. We classify people by blood alcohol level and sex.

Alc (g/100ml)	0	(0, 0.08)	[0.08, 0.15)	≥ 0.15	Total
Male	206	37	35	76	354
Female	53	5	4	6	68
Total	259	42	39	82	422

- Test the claim that gender affects blood alcohol level (i.e. testing whether the two categorising variables are dependent)
- We would be testing:
  H_0 : the two variables are independent against H_1 : not H_0.
- Recall that independence means that the joint probabilities equal the product of the marginal probabilities, that is:
  $\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x) \cdot \mathrm{P}(Y = y)$
- Let p_{ij} denote the probability of a victim being sex *I* and alcohol level group *j*, the independence model says:
  $p_{ij} = p_{i\cdot} \times p_{\cdot j}$,
- We will use the following notation:
  - O_{ij}, observed number of being of sex i and alcohol level group j;
  - $O_{i\cdot} = \sum_{j=1}^{c} O_{ij}$ observed number in row i, i.e. row marginal total
  - $O_{\cdot j} = \sum_{i=1}^{r} O_{ij}$ observed number in column j, i.e. column marginal total
- We estimate $p_{i\cdot}$ and $p_{\cdot j}$ by the marginal proportions, i.e.

$$\hat{p}_{i\cdot} = \frac{O_{i\cdot}}{n}, \quad \hat{p}_{\cdot j} = \frac{O_{\cdot j}}{n}$$

- If H_0 is true (if row and column variables are independent), then the expected number E_{ij} in cell (*i, j*), can be estimated by:

$$E_{ij} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j} = n\left(\frac{O_{i\cdot}}{n} \times \frac{O_{\cdot j}}{n}\right) = \frac{O_{i\cdot} \times O_{\cdot j}}{n}$$
$$= \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

- The expected frequencies for the accident data are:

Sex/Alcohol Level	0	(0, 0.08)	[0.08, 0.15)	>=0.15
Male	217.265	35.232	32.716	68.787
Female	41.735	6.768	6.284	13.213

- The test statistic for this test is:

$$\tau = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(|O_{ij} - E_{ij}| - \frac{1}{2}\right)^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}, \quad \text{under } H_0.$$

- We lose 1 d.f. for each factor because we have used the marginal totals in calculating the expected values.
- There are generally two ways to organise all these calculations:

- Separate table for O_{ij}, E_{ij} and $\frac{(|O_{ij}-E_{ij}|-\frac{1}{2})^2}{E_{ij}}$
- Put all info in a single table but each cell has

O_{ij}
(E_{ij})
$\frac{(\lvert O_{ij}-E_{ij}\rvert-\frac{1}{2})^2}{E_{ij}}$

- Using the first method, we get:

Sex/Alcohol Level	0	(0, 0.08)	[0.08, 0.15)	>=0.15
Male	0.5334	0.0456	0.0973	0.6551
Female	2.7767	0.2376	0.5065	3.4106

- As a result:

$$\tau_{obs} = \sum_i \sum_j \frac{(|O_{ij} - E_{ij}| - \frac{1}{2})^2}{E_{ij}} = 8.2628.$$

- In this data set we have $r = 2$ and $c = 4$, so the df for the test is $(2 – 1)(4 – 1) = 3$.
- Thus, P-value is:

$$P\text{-value} = \mathrm{P}(\chi^2_3 \geq 8.2628) \in (0.01, 0.025)$$

- As P-value is small, we have evidence against H_0, i.e. there is strong evidence to suggest blood alcohol level and sex are related in accident victims.