

# STAT110 / STAT115 Tutoring Materials

INSERT\_NAME\_OF\_STUDENT

Disability Information and Support (DI&S)

July 2024

**Tutor**

Mr Eden Li

Ph.: +64 27 361 4776

Email: [eden.li@otago.ac.nz](mailto:eden.li@otago.ac.nz)

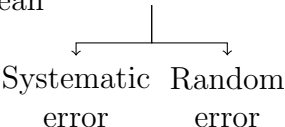




- $\mu$ : Population mean ( $\mu$ ).
- $\sigma$ : Population standard deviation ( $\sigma$ ).
- $\pi$ : Population proportion.
- $\hat{\mu}$ : An estimate of the parameter  $\mu$ .
- $\bar{x}$ : A statistic and an estimate (for STAT110/115 paper only).
- Proportion: Fraction of one quantity when compared to the whole.
- Ratio: Fraction given by one quantity over another.  
**N.B.**, both quantities have the same units.
- Rates (the difference between rates and ratio): Rates are like ratios for quantities with different units.
- Random variables are described by: Probability distributions.
- Observed values of random variables are: Data.
- Difference between random variables and observed/realised value:  
Random variables: unknown quantity varies unpredictably;  
Observed/realised value: got the actual quantity of the unknown quantity.
- Types of variables:
  - Continuous - can be expressed on a continuous scale in which every value is possible.
  - Discrete - can be in one-to-one correspondence with the counting numbers.
  - Categorical - restricted to one of a set of categories. For example 'Heads' or 'Tails'.
    - \* type 1: 0 - 1 binary, A/B/O/AB more than two
    - \* type 2: A/B/O/AB nominal, pass/fail ordinal
- Types of censored data: Right censored, Left censored, Interval-censored  
**N.B.**, censored data are categorised by two variables: the censor type and the censoring point or interval.
  - Right censored: The true value is known to be larger than a recorded value.  
e.g., we know that someone lived until at least 31 Dec 2017. 50+
  - Left censored: The true value is known to be smaller than a recorded value.  
e.g., we know that a measurement is less than a known limit of detection. 10-
  - Interval-censored: The true value is known to lie between two values.  
e.g., we know the date of infection with HPV is after a negative test and before a positive test 2 years later.

- **Purpose of Analytic Studies:** To test hypotheses (quantify population). For example, do government subsidy programs impact the profitability of fisheries? Does IT investment impact productivity in industry? Does a Mediterranean diet impact life expectancy?
- **Function of Replication:** To allow us to separate out true effects from chance effects.
- **Function of Control:** Provides context for evaluating the effect of interest.
- **Descriptive Studies:** The characteristics of people with a disease (person; place; time); lifestyle patterns in a population; attitudes to health care.
- **Well-defined and Not Well-defined Population:**
  - Well-defined: The collection of words in poems by W. B. Yeats. All patients diagnosed with colorectal cancer in New Zealand in 2015.
  - Not well-defined: The population of New Zealand. Right now? Past? Future?(Time). Target population for a particular cancer treatment. Which type/stage of cancer? Existing or future patients(time)? Over a certain age? On other medications?
- **Traits of Random Sampling:** Known-chance, Equal-chance. If doesn't have those traits, define the sample has representative or not.
- **Sampling Frame Definition:** List of all eligible sampling units from which the sample will be drawn. For example, to draw 200 out of 10,000 employees to form a sample, the roster of 10,000 employees, is the sampling frame. May be sourced from census, company data base or other secondary data. Completeness may be an issue when sourcing Sampling Frame. Sometimes need to use/combine multiple sample frames. Non-probability sampling techniques don't require a sampling frame.
- **Sources of Error for Sample Mean:**

$$\text{Sample mean} = \text{Population (true) mean} + \text{Error}$$



```

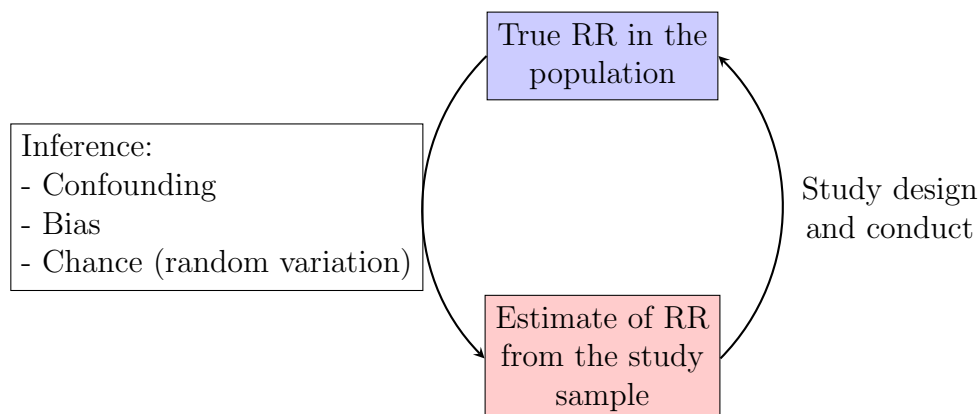
      |
      v
  Systematic  Random
  error       error

```

- **Random Error:** Due to natural variability. Increasing the sample size will reduce the random fluctuations in the sample mean. Statistical methods allow us to quantify the influence of random error on our estimate.
- **Systematic Error in a Descriptive Study (Bias):** Due to aspects of the design or conduct of the study which systematically distort the results. Occurs if a sample is not representative of the population (Selection bias). Occurs if the information collected from the sample members is incorrect (Information bias). Cannot be reduced by increasing the sample size.

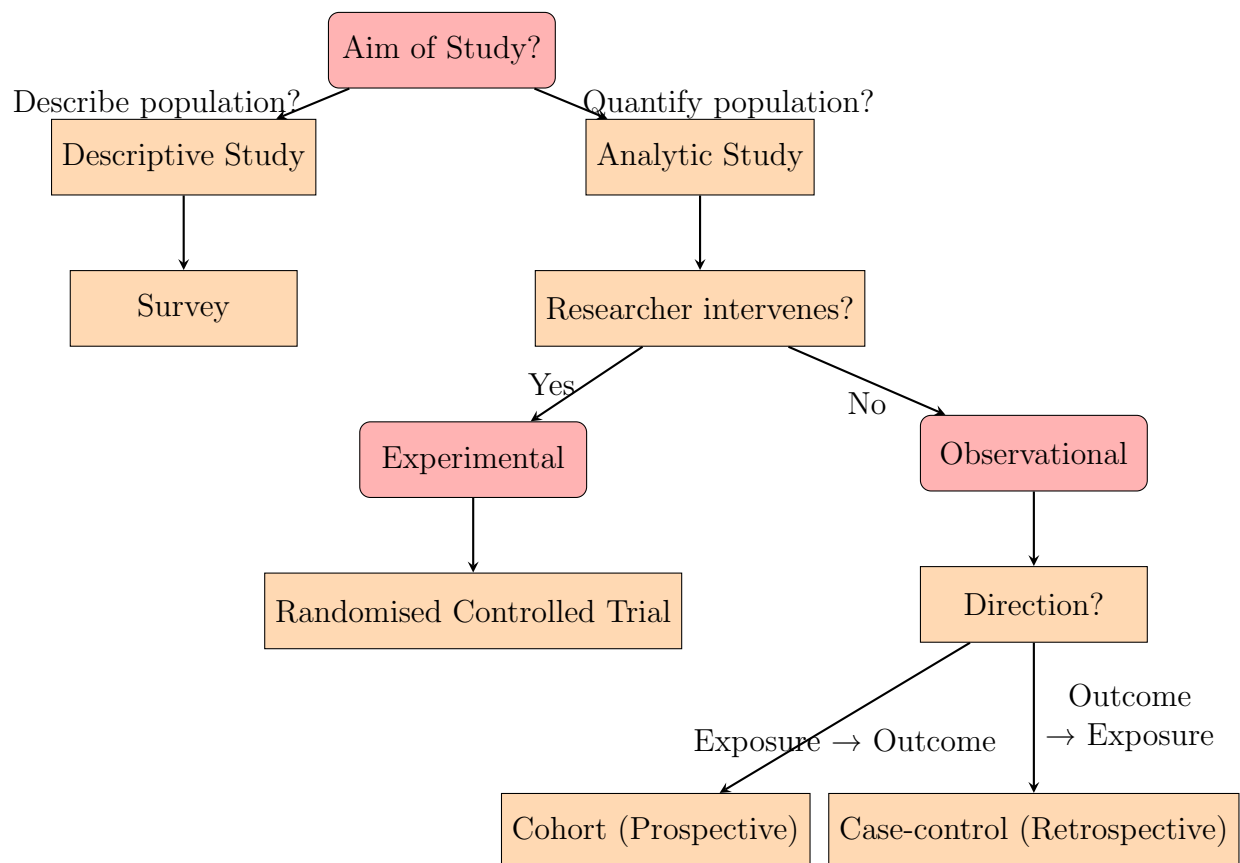
- **Probability Sampling:** We want our sampling frame to match the population of interest and provide a way to draw a sample. Probability sampling is important because it helps to justify the statistical models. For a finite population of size  $N$  draw a sample of size  $n$  such that each possible sample has the same probability of being selected.
- **Key Characteristic of Probability Sampling:** The key characteristic is that we know the probability of being selected for everyone in the sample frame.
- **Simplest Form of Probability Sampling:** Simple random sampling.
- **Types of Probability Sampling:** Simple random sampling, stratified random sampling, cluster sampling.
- **Traits of Simple Random Sampling:** Same chance of selection (e.g., Lotto).
- **Advantages of Stratified Sampling:** More precise estimate than for the same sample size from a simple random sample. Can take different sized samples from different strata (a device for reducing overall variability). Useful if you are interested in the results for each stratum and some of the strata are small. Example: colon cancer treatment, samples of colon cancer patients, stratified by ethnicity.
- **Types of Stratified Sampling:** Proportionate stratified sample, disproportionate stratified sample (equal number from each stratum).
- **Cluster Sampling:** The population may be composed of similar and naturally occurring groups. Dividing the population into a group/cluster (then selecting a sample from each cluster).
- **Types of Cluster Sampling:** One-stage, two-stage. Pros and cons for two stage cluster sampling: reduce cost & time, less precise.
- **Experimental Studies:** The researcher manipulates the conditions (intervenes in a natural process) and records the results. The aim is to control all other factors to isolate the effects of the intervention. Best way to study causation. Why randomisation? Randomisation can be used to ensure that effects of unmeasured factors are equalised across the intervention and control groups. Why NOT experimental studies? Ethical problems.
- **Observational Studies:** The investigator does not intervene, simply observes a naturally occurring process, and collects information. The idea is to get as close as possible to the information that would have been obtained if the experimental study could have been done. Cons: We can't know the confounding factors.
- **Case Control Study:** Outcome trace back to reason.
- **Traits of Randomised Controlled Trial (RCT):**
  - Is considered the "gold standard" analytic study.
  - **Randomisation** - or random allocation, is used to create two comparable groups, one who will have the placebo treatment and the other the experimental treatment. At the end of follow-up any difference between the groups can be attributed to the difference in treatment.

- **Control group** - is used to isolate the effects of the intervention.
- **What is blinding?:** Blinding refers to not knowing whether the participant is in the intervention or the control group. Several people may be blinded to the allocation including the participants, the people caring for patients, the people measuring outcomes, the lead researcher.
- **Pros and cons for RCT:**
  - Advantages: Experiment - the best way to test an hypothesis. If the trial is well conducted, differences in outcome can be attributed to the intervention.
  - Disadvantages: May not be ethical or feasible.
- **Example cohort study:** British doctors and smoking. Aim: to investigate the relationship between smoking and lung cancer.
- **Pros and cons for cohort study:**
  - Pros: Clear chronological order from reason to outcome. Can evaluate the relationship between multiple results and factor(s).
  - Cons: Large time consumption. Bias affect. Small sample size.
- **Characteristics of a case-control study:** Generally carried out to test hypotheses. Participants are chosen on the basis of their outcome status: a group with the outcome (cases) and a group without (controls). Information is collected from people with and without outcome about exposures that occurred in the past (retrospective). i.e. in general before disease was diagnosed.
- **Pros and cons for Case-control study:**
  - Advantages: Relatively quick. Smaller than cohort studies, particularly for rare outcomes. Can examine the effects of multiple exposures.
  - Disadvantages: Events have already occurred so the potential for bias is higher. It is very hard (if not impossible) to remove all the effects of confounding.
- **Sources of error in analytic studies**



- **What is Confounding?:** Confounding is a distortion of the association between exposure and outcome caused by the presence of a third factor. A confounder is a variable which causes this distortion.

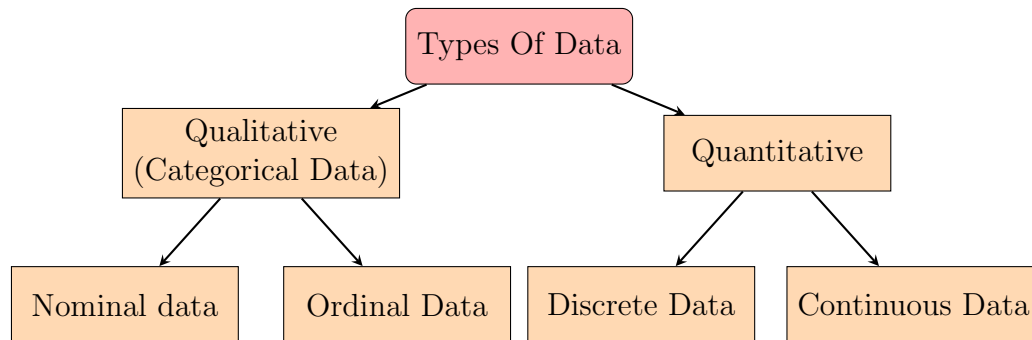
- **A variable must be both ( ) to become a confounder:**
  - associated with the exposure (independent of outcome);
  - and associated with the outcome (independent of exposure).
  
- **Bias in an analytic study:**
  - **Selection bias:** arising from the way participants are selected for inclusion in the study. In an analytic study, selection bias occurs if the selection processes cause a systematic difference between the groups of participants selected for the study. Prospective analytic studies rarely obtain participants through random sampling from a population. The issue of representativeness must be considered, but for analytic studies we consider it a generalisability issue rather than bias.
  - **Information bias:** arising from the way study information is obtained, interpreted and recorded. In an analytic study, information bias is a particular problem if there are systematic differences in the information obtained from groups under comparison in the study. Information bias may be introduced by the observer, the study individual (respondent), instruments used to collect the data (e.g., badly-designed questionnaire), or missing measurements (e.g., from loss to follow-up in a prospective study).
  
- **RCT, Cohort study, Case-control study:**
  - **Randomised controlled trial:** Analytic, experimental, prospective.
  - **Cohort study:** Analytic, observational, usually prospective.
  - **Case-control study:** Analytic, observational, retrospective.
  
- **Summary for the classification:**



- **Discrete:** A type of variable that can only take on specific values. These values are typically whole numbers or counts and cannot be subdivided further. For example, the number of children in a family is a discrete variable because it can only be a whole number (e.g., 1, 2, 3, etc.).
- **Categorical:** Represent data that falls into specific categories or groups. The categories in nominal variables do not have any inherent order or ranking. Examples of nominal variables include gender (e.g., male, female), eye colour (e.g., blue, brown, green), or types of fruit (e.g., apple, banana, orange).
- **Continuous:** Measurements that can take on any value within a specific range. They can be subdivided infinitely, and there are no gaps or interruptions in the possible values. Examples of continuous variables include height, weight, temperature, and time. These variables are often represented by real numbers and can include decimal values.
- **Ordinal:** Similar to categorical variables, but they have an inherent order or ranking associated with their categories. The order represents the relative magnitude or importance of the categories, but the actual differences between the categories may not be uniform or measurable. Examples of ordinal variables include educational attainment (e.g., high school, bachelor's, master's, Ph.D.), socioeconomic status (e.g., low, medium, high), or survey ratings (e.g., strongly agree, agree, neutral, disagree, strongly disagree).
- **If a data set is Categorical, must it also be Nominal?:** No. All nominal data is categorical data, but not all categorical data is nominal data. Nominal data refers specifically to categorical data without any order or hierarchy.

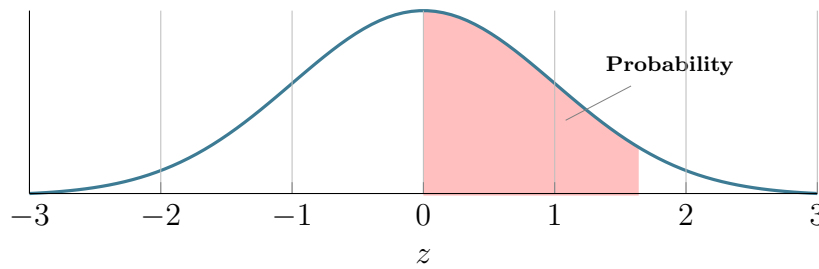


- **Types of data**



- **How to identify whether a study uses probability sampling?:** To find sampling frame.
- **Note for Stratified sampling:** Stratified sampling involves dividing the population into distinct subgroups (strata) based on certain characteristics.
- **Why Non-response can cause bias in surveys?:** because non-respondents tend to(maybe) behave differently compared to people who respond.

- **Usage of `pnorm()`:** `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)`

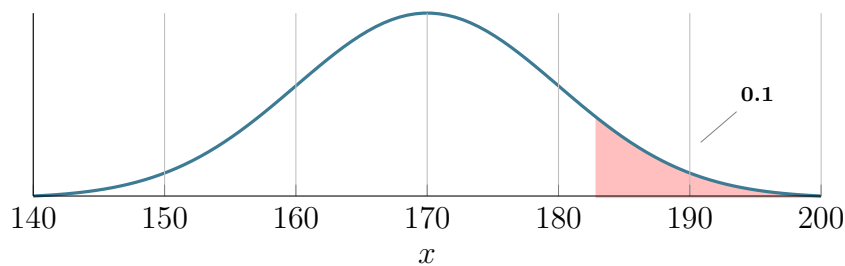


Find  $\Pr(0 < Z < 1.64)$   
`pnorm(1.64) - pnorm(0)`

This function is used to find the probability in a specific area under a normal distribution curve. The parameter 'q' is the "value" from which you want the area above/below.

Example question: Find the proportion of students with a height between 180-190cm.

- **Usage of `qnorm()`:** `qnorm(p, mean = 0, sd = 1, lower.tail = TRUE)`

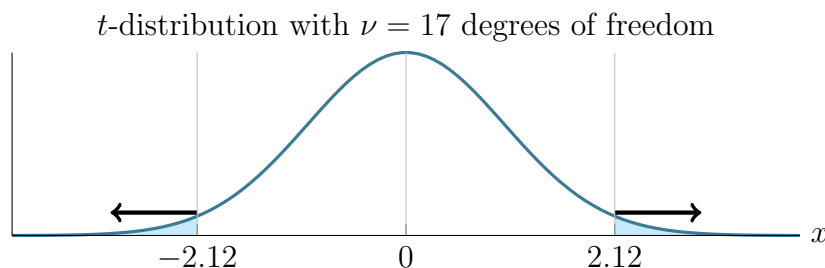


Find the height which is exceeded by 10% of students.  
`qnorm (0.1, mean = 170, sd=10, lower.tail = FALSE)`  
 or  
`qnorm (0.9 ,mean = 170, sd = 10)`

This function is used to find the value "q" in `pnorm()`.

Example question: Find the height which is exceeded by 10% of students.

- **Usage of `dbinom()`:** `dbinom(x=15, size=20, prob=0.75)`  
 This function will provide the individual binomial probabilities associated with a given outcome, provided the number of trials (size) and the probability of 'success' (prob).
- **Usage of `pbinom()`:** `pbinom(q=10,size=20,prob=0.75)`  
 This function will provide the sum of all individual binomial probabilities less than or equal to a given outcome q, provided the number of trials (size) and the probability of 'success' (prob).  
 Example question: Find the probability that 10 or fewer live with both parents.
- **Usage of `pt()`:** `2*pt(q=2.12, df=17, lower.tail=FALSE)`



This function is used to get the p-value. Because there are two tails, so we need to \*2.

- **Usage of `qt()`:** `qt(p, df)`  
This function is used to find the multiplier.
- **Usage of `pchisq()`:** `1-pchisq(q = 9.70, df = 1)`  
This function is used to find the p-value of a chi-square.
- **Usage of `qchisq()`:** `qchisq(p = 0.95, df = 1)`  
This function is used to find the critical value of chi-square.
- **Usage of `pf()`:** `pf(1.0242, df1=3, df2=28, lower.tail=F)`  
This function is used to find the p-value of the F-distribution.
- **Usage of `qf()`:** `qf(0.05, 3, 15, lower.tail=FALSE)`  
This function is used to find the critical value of F statistic.
- **Output for regression model fit:**

```
Call:
lm(formula = y ~ x1 + x2, data = mydata)

Residuals:
    Min       1Q   Median       3Q      Max
-1.1411 -0.3145  0.0187  0.3178  1.1867

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2154     0.1321   16.77  <2e-16 ***
x1            -0.7307     0.0917   -7.97  <2e-16 ***
x2             0.3263     0.0862    3.78  <2e-16 ***

Signif. codes:
*** p<0.001 ** p<0.01 * p<0.05 ' ' p>0.05

Residual standard error: 0.5209 on 37 degrees of freedom
Multiple R-squared:  0.7995, Adjusted R-squared:  0.7833
F-statistic: 49.79 on 2 and 37 DF, p-value: 5.757e-11
```

- The "Residual standard error" represents the standard deviation of the residuals, which is an estimate of the average distance between the observed and predicted values.

- The "Multiple R-squared" and "Adjusted R-squared" values indicate the goodness of fit of the model. They represent the proportion of variance in the response variable explained by the predictors. Adjusted R-squared takes into account the number of predictors and the sample size.
- The "F-statistic" is a measure of the overall significance of the model. It assesses whether the regression model as a whole is statistically significant.
- The associated p-value indicates the probability of obtaining such an F-statistic by chance.

• **Output for a t test:**

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	1.3456	1.3456	12.463	0.00123 **
x2	2	0.7621	0.3811	3.527	0.03812 *
Residuals	36	3.5584	0.0988		

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

- The "data" line indicates the name of the variable or group being tested.
- The "t" value represents the calculated t-statistic for the test. It measures the difference between the sample mean and the hypothesised mean relative to the variability in the data.
- The "df" value stands for degrees of freedom, which measures the amount of information available for the test. It represents the sample size minus one.
- The "p-value" is the probability of obtaining the observed test statistic (t-value) or a more extreme value under the null hypothesis. It indicates the level of statistical significance.
- The "alternative hypothesis" states the alternative to the null hypothesis.
- The "95 per cent confidence interval" provides a range of values within which we can be 95% confident that the true population mean lies. It is calculated based on the sample data and reflects the estimate's precision.
- The "sample estimates" section presents the estimated mean of the variable or group being tested.

• **Output for an ANOVA table:**

data: x

t = 3.5897, df = 98, p-value = 0.0005379

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

0.05743578 0.19679155

```
sample estimates:
mean of x
0.1271137
```

- The "Response" line shows the name of the dependent variable.
- The "Df" column represents the degrees of freedom associated with each factor or source of variation.
- The "Sum Sq" column shows the sum of squares associated with each factor or source of variation. It represents the total variability explained by each factor.
- The "Mean Sq" column represents the mean square, which is calculated by dividing the sum of squares by the degrees of freedom. It represents the average variability explained by each factor.
- The "F value" column displays the F-statistic, which is calculated by dividing the mean square of each factor by the mean square of the residuals. It measures the ratio of explained variation to unexplained variation and is used to test the significance of each factor.
- The "Pr(>F)" column shows the p-value associated with each factor. It indicates the probability of obtaining the observed F-statistic or a more extreme value under the null hypothesis.

- **Subjective Probability:** The individual's personal estimate of the chance of loss.
- **Probability as the Degree of Belief:** In a statement.
- **Objective Probability:** The long-run relative frequency of an event based on the assumptions of an infinite number of observations and of no change in the underlying conditions.
- **Sample Space Definition:** The sample space is the set of all possible outcomes of an experiment, e.g., in the case of a fair die,  $\{1, 2, 3, 4, 5, 6\}$ .
- **Complementary Events:**

$$Pr(A) + Pr(\bar{A}) = 1$$

- **Mutually Exclusive Events:** There is no intersection between the two events

$$Pr(A \cap B) = 0$$

e.g., Māori and Chinese.

- **Conditional Probability Multiplication Rule:**

$$Pr(A \cap B) = Pr(A)Pr(B|A)$$

- **Conditional Probability Addition Rule:**

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

- **Independent Events:** The outcome of one event does not affect the outcome of the second event.
- **Independence Check** (true if):

$$Pr(B) \neq Pr(B|A)$$

or

$$Pr(A) \neq Pr(A|B)$$

- **Definitions of Diagnostic Tests:** Sensitivity, Specificity, False positive fraction, Positive Predictive Value, Negative Predictive Value
- **Sensitivity (Diagnostic Tests):**

$$Pr(B|A)$$

The probability that a person with the disease has a positive test. Think of this as "the probability of a positive test result, given the person actually has the disease.

- **Specificity (Diagnostic Tests):**

$$Pr(\bar{B}|\bar{A})$$

The probability that a person without the disease has a negative test. Think of this as "the probability of a negative test result, given the person does NOT have the disease.

- **Negative Predictive Value:**

$$Pr(\bar{A}|\bar{B})$$

The proportion of patients with negative test results who don't have the disease.  
The proportion of patients with negative test results who are correctly diagnosed.

- **False positive fraction (Diagnostic Tests):**

$$1 - \textit{specificity} = Pr(\bar{B}|\bar{A})$$

- **Positive Predictive Value:**

$$Pr(A|B)$$

The proportion of patients with positive test results who have the disease. The proportion of patients with positive test results who are correctly diagnosed.

- **Relative Risk (RR):** Ratio of two probabilities. RR gives the risk of an outcome relative to "exposure". It is calculated as the ratio of the risk of an outcome for an exposed and an unexposed group.

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

Meaning of the RR value:  $RR = 1$  there is no association between outcome and exposure (e.g. rugby position and injury).  $RR < 1$  first row happens less likely than the second row.  $RR > 1$  first row happens more likely than the second row.

- **Risk Difference (RD):** Difference between two probabilities. The RD is given by the difference in the risk for the two groups.

$$RD = \frac{a}{a+b} - \frac{c}{c+d}$$

- **Odds Ratio (OR):** Ratio of two odds. The OR compares the odds of an outcome for two groups. Ratio of the odds of the outcome for the exposed group to that for the unexposed group.

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

. There is no mathematical distinction between exposure and outcome variables - it makes it particularly useful for quantifying associations between binary variables where there is no "direction" e.g. alcohol consumption (Yes/No) and smoking (Yes/No).

- **Confidence Interval for Difference Between Two Proportions:**

$$p1 = \frac{a}{r1}$$

,

$$p2 = \frac{c}{r2}$$

,

$$(p_1 - p_2) \pm Z_{(1-\frac{\alpha}{2})} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- **Steps to Calculate the Confidence Interval for Relative Risk:**

- Get the RR value.
- Get the  $\ln(RR)$ .
- Calculate the SE of  $\ln(RR)$  (with formula).
- Calculate the CI for  $\ln(RR)$  (with formula).
- Calculate the CI for RR ( $\exp()$  function).

- **Standard error for Confidence interval for relative risk:**

$$S_{\ln(RR)} = \sqrt{\frac{1}{a} - \frac{1}{r_1} + \frac{1}{c} - \frac{1}{r_2}}$$



- **Key formula for Confidence interval for relative risk:**

$$\ln(RR) \pm Z_{(1-\frac{\alpha}{2})} \cdot S_{\ln(RR)}$$

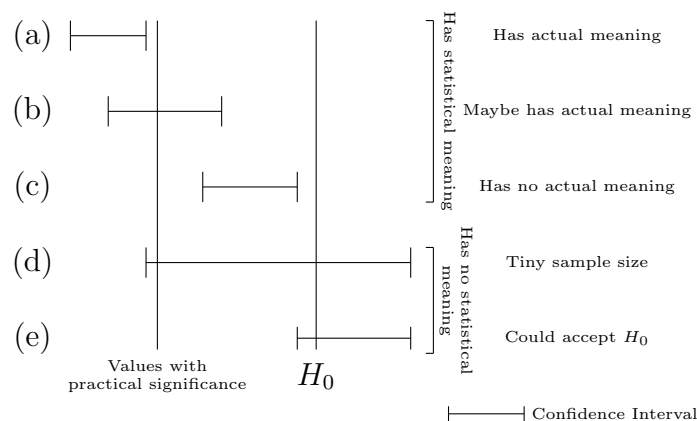
- **Steps to Calculate the Confidence Interval for Odds Ratio:**

- Get the OR value.
- Get the  $\ln(OR)$ .
- Calculate the SE of  $\ln(OR)$  (with formula).
- Calculate the CI for  $\ln(OR)$  (with formula).
- Calculate the CI for OR ( $\exp()$  function).

- **Standard error for Confidence interval for Odds Ratio:**

$$S_{\ln(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

- **The meaning for range of CI:**



- **Risk Difference in Terms of the Number of Cases Per x People:** To get the risk difference in terms of the number of cases per x people, we need to multiply this answer by x. For example, express your answer in terms of the extra number of cases of cancer among 1000 people who eat red or processed meat four or more times per week.

$$\frac{2341}{191678} - \frac{277}{68601} = 0.008175$$

To get the risk difference in terms of the number of cases per 1000 people, we need to multiply this answer by 1000.

$$RD = \left( \frac{2341}{191678} - \frac{277}{68601} \right) * 1000 = 8.175$$

- **The standard normal critical value for a 95% interval:** 1.96
- **Confidence interval formula:**

$$\bar{x} \pm Z_{(1-\frac{\alpha}{2})} \times \frac{\sigma_X}{\sqrt{n}}$$

estimate for the mean  $\pm$  multiplier  $\pm$  standard error for the mean

- **The standard normal critical value for a 99% interval:** 2.58
- **Multiplier formula:**

$$z = \frac{x - \text{mean}}{\text{sd}}$$

- **What is the  $\alpha$  in the multiplier:** tail probability
- **Multiplier pattern:** when CI is bigger (e.g., 95% to 99%), the multiplier will be bigger
- $s_X$ : sample standard deviation
- **In practice the true standard deviation  $\sigma_X$  is not known:** We estimate it with the sample standard deviation.
- **This means our critical values must now come from the 't' distribution, not the standard normal.**
- **t distribution CI:**

$$\bar{x} \pm t_{(1-\frac{\alpha}{2}, \nu)} \times \frac{s_X}{\sqrt{n}}$$

- **$\nu$  (degree of freedom) for t-distribution:**  $\nu = n - 1$
- **The t-distribution will be the correct sampling distribution if:** either the underlying distribution of X is normal, and/or the sample size is sufficiently large (Central Limit Theorem holds).
- **What is the degree of freedom in t-distribution:** to replace the mean and sd in a normal distribution (because t-distribution is always standardised)
- **When to use t-distribution:** when the sample size is small
- **Calculate the estimate sample size when knows the CI:** assuming knows the sd and mean (normally given in the question), solve the equation, rounding UP
- **Comparing means with CI:**

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(1-\frac{\alpha}{2}, \nu)} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- **Using CLT to test appropriate normal distribution:**

$$n\pi \pm 3\sqrt{n\pi(1-\pi)}$$

gives two values between 0 and n, if not, then fails the test. This approximation is good only when: n is large,  $\pi$  is not close to 0 or 1 (this increases symmetry)

- **Formula for estimating  $\pi$ :**

$$P = \frac{X}{n}$$

,

$$p = \frac{x}{n}$$

,  $x$  is the observed value of  $X$ . (and more) Using the Central Limit Theorem, the resulting distribution of these proportions is approximately normal if,  $n$  is large enough,  $\pi$  far enough from 0 or 1. As before, we judge this using:

$$n\pi \pm \sqrt{n\pi(1-\pi)}$$

gives values between 0 and  $n$ .

- **Derivation of the mean of the sampling distribution:** If  $P = \frac{X}{n}$ , then  $\mu_P = \pi$ ,  
 $\text{sd} = \sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$
- **95% confidence interval for  $\pi$**  (use the sample proportion ( $p$ ) to estimate the unknown true population proportion ( $\pi$ )):

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

- **Margin of error:**

$$\text{multipliers} * \text{sd}$$

- **Note for CI:** This confidence interval (and margin of error) is correct only if the normal approximation to the binomial is appropriate. In practice, bias due to non-response should also be considered in our interpretation of an estimate.

- **Mean of the binary (Bernoulli) distribution:**

$$\mu = p$$

- **Variance of the binary (Bernoulli) distribution:**

$$\sigma^2 = p(1 - p)$$

- **Difference between binary distribution and binomial distribution:**

$$n = 1 \Rightarrow \text{binary distribution}$$

$$n > 1 \Rightarrow \text{binomial distribution}$$

**Mean of the binomial distribution:**

$$\mu = np$$

**Variance of the binomial distribution:**

$$\sigma^2 = np(1 - p)$$

- **Conditions for binomial distribution:** Outcome is binary. We have n independent trials. The number of trials is fixed. The probability of success  $\pi$  must stay constant.
- **Probability of x successes in n trials:**

$$\Pr(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

- Binomial coefficient ( $\binom{n}{k}$ ):

$$\frac{n!}{(k!)(n - k)!}$$

**Standard normal distribution(Z):**

$$Z \sim N(\mu = 0, \sigma^2 = 1)$$

- $\mu$  (normal distribution) moves the curve but does not change its shape.
- $\sigma$  spreads the curve more widely about  $X = \mu$  but does not alter the centre.
- **Compare a relative frequency histogram with a probability distribution:** Relative frequency histogram represents a sample (smaller number of individuals). The probability density function represents a population (a large number of individuals).
- **How to estimate the value of the parameters if estimating a probability distribution curve from a relative frequency histogram:**  $\mu$  is estimated by the sample mean.  $\sigma$  is estimated by the sample standard deviation, s.

- **What do the areas under the normal distribution curve represent? Probabilities.**
- **What is Z-score (Z-value)?** Number of standard deviations away from the mean.
- Any normal distribution value,  $X \sim N(\mu_X, \sigma_X^2)$ , can be put on the standard normal scale,  $Z \sim N(0, 1)$ . The Z-score follows a standard normal distribution.
- **Formula for Z-Value:**

$$Z = \frac{(X - \mu_X)}{\sigma_X}$$

- **When will the sampling distribution of the mean will follow a normal distribution?** If  $n$ (the samples, not  $X$ ) is large enough.
- **Central Limit Theorem (CLT):** The sampling distribution derived from a simple random sample will be approximately normally distributed.
- **What is the mean of the sampling distribution? Population mean,  $\mu_{\bar{X}} = \mu_X$ .**
- **Variance of the sampling distribution:**

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

The variability of sample means.

- **Notes on the sampling distribution:** If sample size  $n$  is greater, then the standard error of the mean is smaller (more compact distribution, greater precision). If  $X$  is normal, then  $X_{bar}$  is normal (for any  $n$ ). If  $X$  is not normal, then  $X_{bar}$  is approximately normal for large  $n$  (central limit theorem).

- **Types of regression:** Linear (continuous data), Logistic (categorical data), Cox (categorical data in a survival analysis).
- **Explanatory variable (X):** Also known as a covariate, predictor, or independent variable.
- **Outcome variable (Y):** Also known as response or dependent variable.
- **Simple Linear Regression (SLR):** Looks at a relationship between two continuous variables where the relationship between the two variables is approximately a straight line.
- **SLR equation:**

$$Y = \beta_0 + \beta_1 x + e$$

- This implies that the mean response is related to  $x$  by

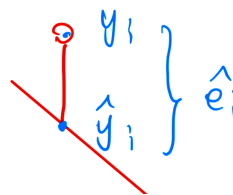
$$\mu_Y = \beta_0 + \beta_1 x$$

- $\mathbf{Y}$  is the numerical outcome variable (continuous or approximately so).
- $\mathbf{x}$  is the explanatory variable.
- $\beta_0$  is the intercept or constant (where the line crosses the y-axis).
- $\beta_1$  is the slope of the line.
- $\mathbf{e}$  (often denoted  $\epsilon$ ) is the random error or residual term.
- **SLR equation for estimating:**

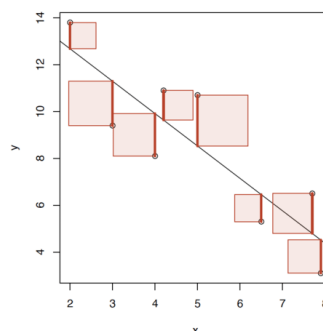
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- **Residual ('estimated error') term:**

$$\hat{e}_i = y_i - \hat{y}_i$$



- **How to find regression line:** The line of best fit minimises the sum of the squares of the residuals.



- **Equation for how to find regression line:**

$$\sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- **How to calculate  $\beta_1$  and  $\beta_0$ :**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- **Example for how to calculate regression (Stress and Blood Pressure):**

- Get n, n = 6
- Find the explanatory and outcome
- Calculate  $\beta_1$  and  $\beta_0$
- Get the regression equation
- Using R for SLR

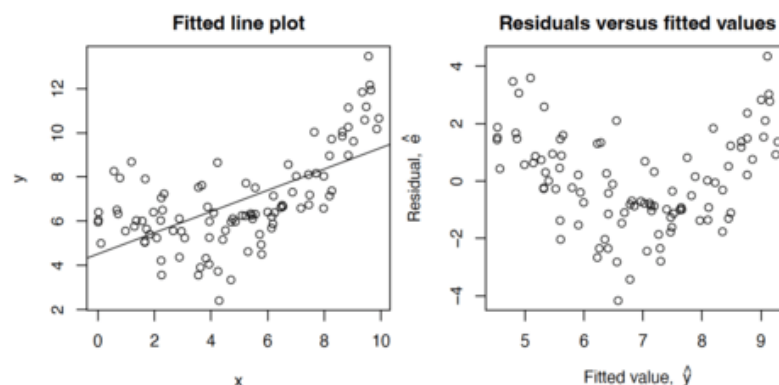
- **Assumptions for Simple Linear Regression (LINE):**

- Linearity: The relationship between the mean response  $\mu_Y$  and x is described by a straight line.
- Independence: The responses  $Y_1, Y_2, \dots, Y_n$  are statistically independent.
- Normality: The error terms  $e_1, e_2, \dots, e_n$  come from a normal distribution.
- Equal variance: The error terms all have the same variance,  $\sigma^2$  ('homoscedastic').

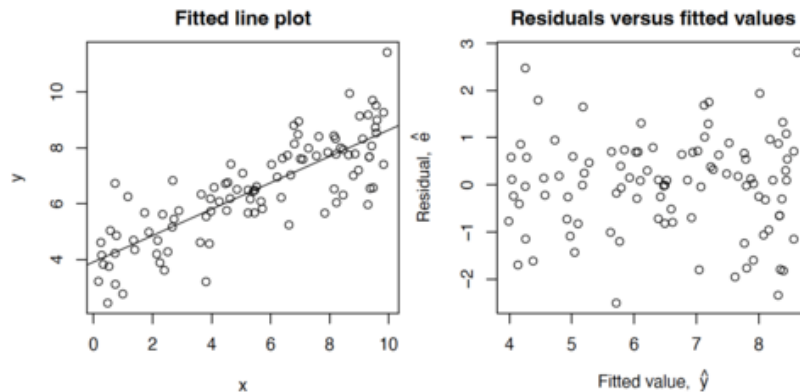
- **What diagram is used for checking linearity: Residual plot.**

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

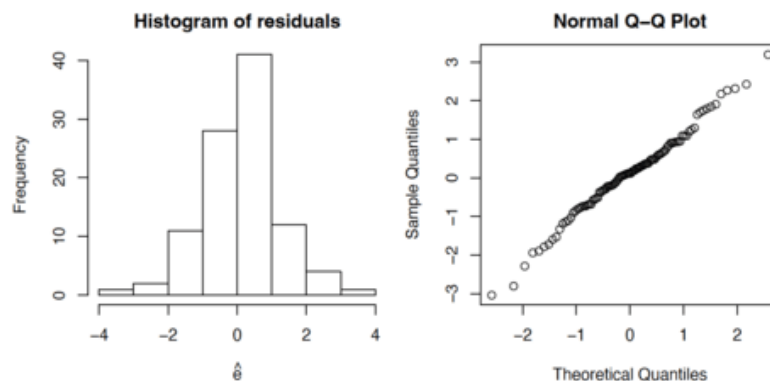
- **Failure of linearity assumption:**



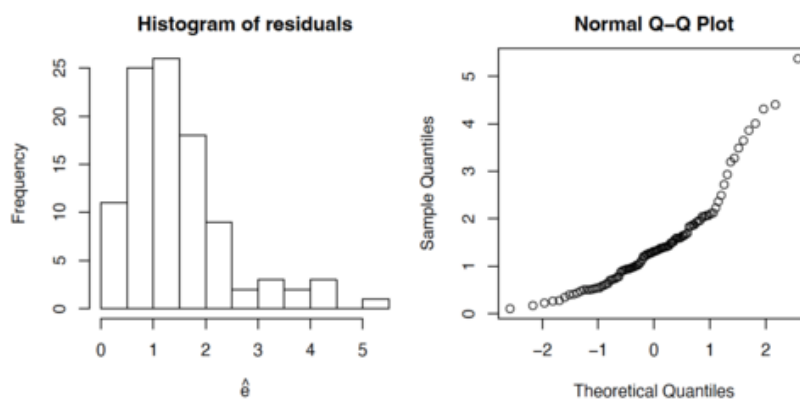
- **Linearity assumption holds:**



- **Checking independence assumption:** May get insight by thinking about the study design (Ask yourself questions).
- **Plot for Checking the normality assumption:** Q - Q plot.
- **Pass of normality assumption:**

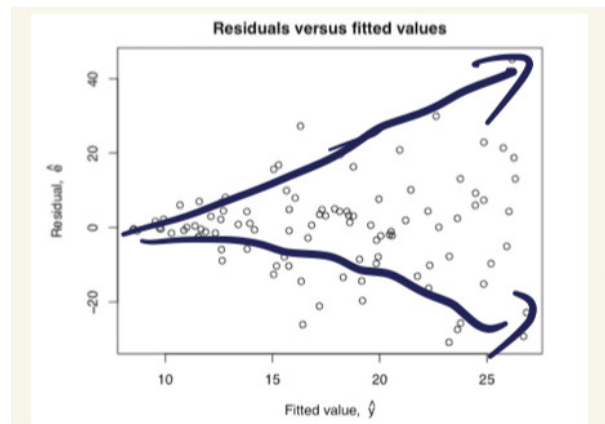


- **Fail of normality assumption:**





- **Checking equal variance assumption (homoscedasticity):** Pass if the residual plot is not like this.



- **What is the impact if Fail of the linearity assumption:** Critical. If that assumption fails, all conclusions drawn from the model will be invalid.
- **What is the impact if Fail of independence or equal variance assumptions:** Remain valid. However, estimates can be inefficient. Follows that the fitted regression line is useable. Any test results or confidence intervals based on the regression model will be invalid.
- **What is the impact if Fail of normality assumption:** Typically least important. Effects validity of confidence intervals and test results when the sample size  $n$  is small.
- **What to do with outliers:** The first thing to do is check that the data are correctly recorded. If data cannot be corrected, try refitting regression with outliers removed, but still investigate the cause of outliers - may be very important.
- **Estimate of error variance:**

$$S_e^2 = \frac{1}{n-2} \sum_{i=1}^n (\hat{e}_i^2) = \frac{RSS}{n-2}$$

where  $RSS = \sum_{i=1}^n \hat{e}_i^2$  is the residual sum of squares.

- **Degree of freedom for SLR's CI:**  $v = n - 2$  because there're two parameters.
- **What is the multiplier for SLR's CI:**

$$t = \frac{\text{estimate} - \text{null}}{\text{std.error}}$$

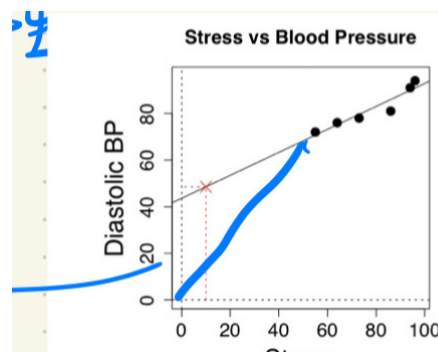
- **What is the SE for SLR's CI:**

$$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- **Using R to find SLR's CI:**

```
> confint(model1)
                2.5 %      97.5 %
(Intercept) 24.8300345 62.4009555
X            0.2557407 0.7284774
```

- $\beta_1 = 0$  **indicates what:** That the response is not (linearly) related to the predictor. So the estimated slope will (almost) always be non-zero:  $\hat{\beta}_1 \neq 0$ .
- **Steps to test to assess the strength of evidence in the data for  $\beta_1 \neq 0$ :**
  - Setting up the hypotheses:  $H_0 : \beta_1 = 0$ ,  $H_A : \beta_1 \neq 0$ .
  - Calculating The test statistic
$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$
  - Computing the p-value
  - Draw conclusion with rejecting or not  $H_0$
- **When predicting the data:** Ignore  $e_0$ .
- **Why not recommend extrapolating when predicting data:** The plot may not be linear.



- **Prediction error:** The prediction error is analogous to a standard error, but takes account of both sources of uncertainty. For prediction at  $x_0$ , the prediction error is:

$$PE(\hat{y}_0) = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- **Prediction interval formula:**

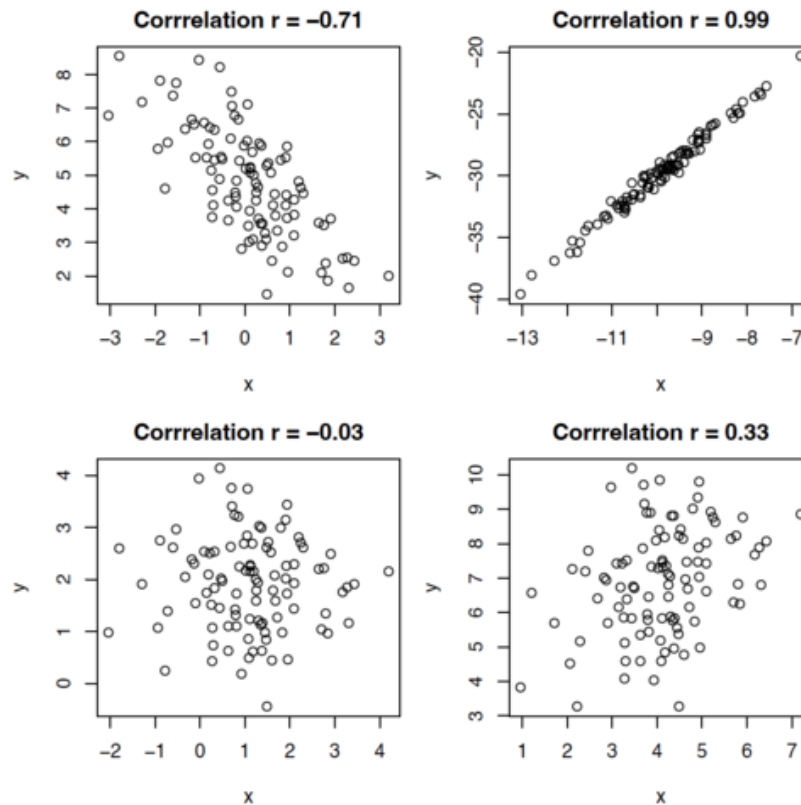
$$\hat{y}_0 \pm t_{(1-\frac{\alpha}{2}, n-2)} \times PE(\hat{y}_0)$$

- **Correlation coefficient (r):** Summarises the strength of a linear relationship between variables. It is a measure of linear association between variables. It describes both the strength and direction of the relationship.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- $r \in [-1, 1]$ . A positive value of  $r$  means that  $Y$  and  $X$  increase together. A negative value of  $r$  means that as  $X$  increases,  $Y$  decreases (and vice-versa).
- The strength of the linear relationship increases as  $r$  tends towards 1 or -1.  $r = 0$  corresponds to no linear relationship between the variables.

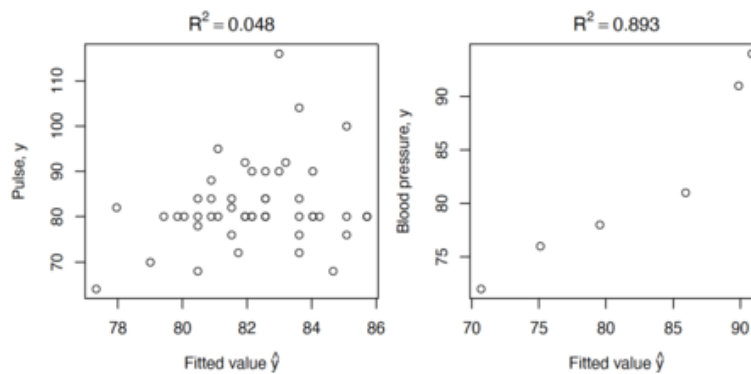
- **Scatterplots for  $r$ :**



- **Re-write for  $r$ :**  $s_x$  and  $s_y$  are sample standard deviations for  $x$  and  $y$  variables.  $s_{xy}$  is the sample covariance between  $x$  and  $y$ .

$$r = \frac{s_{xy}}{s_x s_y}$$

- **Correlation coefficient versus regression models:** The correlation coefficient is a summary of the data. Unlike linear regression, the correlation coefficient does not specify a model for the data, and cannot (for example) be used for prediction. The correlation coefficient is symmetric in the variables. That is, the correlation between  $x$  and  $y$  is the same as the correlation between  $y$  and  $x$ . In regression, the variables are not handled symmetrically. Regression models look at variation in  $Y$  for fixed values of  $x$ .
- **Coefficient of determination ( $R^2$ ):**  $R^2$  is a measure of how well a regression model describes the data.  $R^2$  is the squared correlation between the observed and predicted responses.  $R^2 \in [0, 1]$ .
- **Meaning for the value of  $R^2$ :** A high value of  $R^2$  (close to 1) indicates a regression model that describes the data very well. Conversely, a low value of  $R^2$  (close to 0) indicates a regression that describes the data poorly.



- What describes the overall variation in the response variable?: Total sum of squares.

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

- What describes the total variation of the data points about the regression line?: Residual sum of squares (RSS can be thought of as variation not explained by the regression model).

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

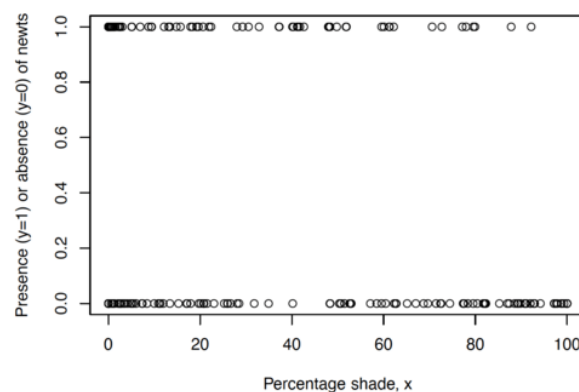
- What describes as the amount of variation in the response that is explained by the regression model?: Explained sum of squares.

$$ESS = TSS - RSS$$

- Equation of  $R^2$ :

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- **Correlation does not equal causation:** e.g., just because there's more ice cream in the summer and more drowning in the summer doesn't mean there's a link between ice cream and drowning.
- **Logistic regression:** Outcome variable is binary.



- **Equation for logistic regression:**

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

$Y$  is the binary outcome variable,  $Y = 1$  or  $Y = 0$  for each observation.  $p$  is the probability that specified category will occur; i.e.  $p = \Pr(Y = 1)$ .  $x$  is the explanatory variable. Parameters  $\beta_0, \beta_1$  are the regression coefficients.  $\beta_0$  is intercept and  $\beta_1$  slope 'on the logit scale'. In the formula, log is the natural logarithm (log to base  $e$ ).

- **Which technique do we use when estimating the regression coefficients?:** Maximum likelihood estimation.
- **What will increase  $x$  by one unit result in?:** A multiplicative change of  $e^{\beta_1}$  to the odds.
- **Formula for logistic curve for the probability  $p$ :**

$$p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- **Testing in logistic regression:**

- Define the hypotheses:  $H_0 : \beta_1 = 0$  and  $H_A : \beta_1 \neq 0$ .
- The test statistic is:

$$z = x \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

- Get the corresponding p-value.
- Reject/not reject  $H_0$ .
- Conclusion.

- **Multiple regression model:**

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e$$

- **Mean value of the Multiple regression model:**

$$\mu_Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- **Applications of multiple regression:**

- Adjusting for the effect of confounding variables.
- Establishing which variables are important in explaining the values of the response variable.
- Predicting values of the response variable.
- Describing the strength of the association between the response variable and the explanatory variables.

- **Least squares estimates:**

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- **RSS for  $\hat{e}_i$ :**

$$RSS = \sum_{i=1}^n \hat{e}_i^2$$

To estimate the error variance  $\sigma_e^2$ .

- **Usual estimate:**

$$s_e^2 = \frac{RSS}{n - k - 1}$$

- $H_0$ : Null hypothesis. The hypothesis is that there is no association, no effect or no difference.
- $H_A$ : Alternative hypothesis. The hypothesis is that there is an association, effect or difference.
- **P-value**: We measure the "consistency" of the observed data with the claim using a p-value. The p-value is the probability of observing the value of the test statistic, or a value more extreme, calculated under the assumption that  $H_0$  is true. A small p-value indicates we would be unlikely to see the data we did if the null hypothesis were true. i.e., the smaller the p-value is, the easier to reject  $H_0$ . If the p-value is less than  $\alpha$  we reject  $H_0$ . If the p-value is greater than or equal to  $\alpha$  we do not reject  $H_0$ .

- **Test statistic (t - statistic)**: A test statistic is the standardised value of the sample value.

$$T = \frac{\text{observed sample value} - \text{null value}}{\text{estimated standard error}}$$

When studying hypothesis testing, we usually use  $\pi$  to denote the value of the overall parameter, instead of using p.

- **Z Statistic vs t-statistic**:

- Z-statistic: When to use: Large samples ( $n > 30$ ) or known population standard deviation ( $\sigma$ ). Basis: Uses population standard deviation ( $\sigma$ ).
- t-statistic: When to use: Small samples ( $n < 30$ ) or unknown population standard deviation ( $\sigma$ ). Basis: Uses sample standard deviation (s).

- **How to conduct a hypothesis test**:

1. Set  $H_0$  and  $H_A$ .
2. Calculate t-statistic(if it is sample) or Z-statistic(if it is population).
3. Calculate p-value.
4. Calculate 95% CI with (3).
5. Draw a conclusion based on the p-value (reject  $H_0$  or not).

- **The difference between p and p\***:

- p (Sample Proportion): Frequency or proportion of events in a sample. Example: If 50% support a policy in a sample, (  $p = 0.5$  ).
- p\* (Estimate of Overall Proportion): Estimate of the overall proportion based on sample data. Example: If 50% support a policy in a sample, (  $p^* = 0.5$  ).

- **When to use chi square**: When trying to control groups in experiments - looking for differences between men and women in each group, etc. Looking for differences between categorical variables - maybe you want to know if there is a difference between men and women for their favourite type of ice cream.

- **How to conduct chi-square test**:

1. Define the Null-Hypothesis and Alternative Hypothesis.  $H_0$  : The treatment and response are independent (i.e. no association).  $H_A$  : The treatment and response are dependent in some way (i.e. there is some association).
2. Calculating expected cell counts.
3. Calculating the  $\chi^2$  test statistic.
4. Get the degree of freedom.
5. Calculate the p-value with R.
6. Reject / not reject  $H_0$ .
7. Draw conclusion.

- **Expected cell counts:**

$$E_{(row\ i, col\ j)} = \frac{r_i \times c_j}{n}$$

- $r_i$  is the row total, for row i
- $c_j$  is the column total, for column j
- $n$  is the total number (of trials, patients, etc.)

We worked out what we would have expected to see under the null hypothesis in each cell given the observed row and column totals.

- **Formula for chi-square:**

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

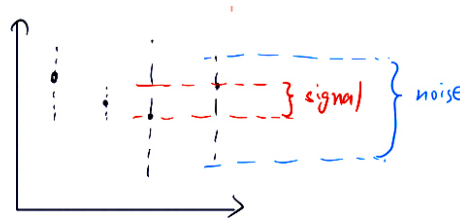
- **Degree of freedom for chi square:**  $\nu = (number\ of\ rows - 1) * (number\ of\ columns - 1)$
- **Range for p-value:** (0, 1)
- **Belief (interpretation/decision):**

	Fail to reject $H_0$	Reject $H_0$
Null is true	Correct interpretation (No error)	Error (Type I)
Null is false	Error (Type II)	Correct interpretation (No error)

- **Type I error (a false positive result):** Concluding that there is an association between exposure and outcome, where there is not. Type I error is controlled when we set the significance level (usually 0.05).
- **Type II error (a false negative result):** Concluding that there is not an association between exposure and outcome, where there is. Type II error is primarily controlled through the sample size. Ideally, power should be between 80 and 90%.



- **ANOVA:** Abbreviation of **A**nalysis of **V**ariance.
- **Methods for comparing means** of continuous responses between multiple groups.
- **F-ratio:** Signal/noise



- **Reasons why using "2+2+2=3" is undesirable:**
  - It's more work than we need to do. Three tests may not seem too bad, but to compare 10 groups we would have to do 45 different pairwise t-tests.
  - It can lead to lots of false positive results. Every test has the potential to incorrectly reject  $H_0$ ; i.e. falsely identify a difference between a pair of groups. If we do lots of tests then we risk generating lots of false positives.
- **ANOVA model:**

$$Y_{ij} = \mu_i + e_{ij}$$

$\mu_i$  is the true mean response for the  $i$ th group at the population level.  $e_{ij}$  is the error term for the  $j$ th response in the  $i$ th group. The error terms are assumed to be independent and to follow a  $N(0, \sigma^2)$  with constant variance. The number of different groups is denoted  $K$ , and the number of responses in the  $i$ th group is denoted  $n_i$ .

- **Est. mean for ANOVA:** The "." = est. Value.

$$\hat{\mu}_i = \bar{y}_{i.}$$

- **Sample mean for the  $i$ th group:**

$$\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

- **Formula for residual sum of squares in ANOVA** (no need to memorise):

$$RSS = \sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

- **Total sum of squares in ANOVA** (no need to memorise):

$$TSS = \sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

- $y_{..}$  is the sample mean overall the data.

- **Formula for GSS in ANOVA** (no need to memorise):

$$GSS = TSS - RSS$$

$$GSS = \sum_{i=1}^k n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

GSS can be interpreted as a measure of the variation that is explained by differences between groups.

- **Setting up the hypotheses to test ANOVA:**

As usual, the null hypothesis will be the 'no difference' hypothesis:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_K$$

The alternative is simply an expression that the null is incorrect:

$$H_A : \mu_1, \mu_2, \dots, \mu_K \text{ not all equal.}$$

- **Equation for F statistic:**

$$F = \frac{GSS/(K-1)}{RSS/(n-K)}$$

- **GMS:**  $GSS/(K-1)$  is the group mean square.
- **RMS:**  $RSS/(n-K)$  is the residual mean square.
- **What situations would let  $H_0$  fail:** Large differences between group means. Relatively large value of GSS. A large value of F.
- **ANOVA table:**

Source	SS	DF	MS
Groups	GSS	$K-1$	$\frac{GSS}{K-1}$
Residuals	RSS	$n-K$	$\frac{RSS}{n-K}$
Total	TSS	$n-1$	

- **P-value is right censored.**
- **Blocking variable:** A second treatment variable that when included in ANOVA analysis will have the effect of reducing the SSE term (noise).

