University of Otago
ŌTĀKOU WHAKAIHU WAKA

STAT115
Tutoring Materials

Disability Information and Support (DI&S)

July 2025

Tutor
Eden Li (he/him)
Ph.: +64 27 361 4776
Email: eden.li@otago.ac.nz

- **Population**: the entire group we want to learn about.
- **Sample**: the subset of that population we actually observe.
- **Parameter** (population quantity) vs. **Statistic** (sample-based estimate).
- μ - population mean
  σ - population standard deviation
  π - population proportion.
- $\bar{x}$ - sample mean
  s - sample standard deviation
  $\hat{p}$ - sample proportion.
- **Proportion**: fraction of the (sample or population) total in a given category ($0 \leq \hat{p} \leq 1$).
- **Ratio**: numerator and denominator have the *same* units (e.g. waist/hip).
- **Rate**: numerator and denominator have *different* units (e.g. km per hour; cases per 1,000 person-years).
- **Random variable** X: an unknown quantity described by a probability distribution.
- **Observed (realised) value** x: the concrete outcome recorded in the data.
- **Variable types**
  - **Quantitative**
    - *Continuous*: can take any value on an interval (e.g. height, blood pressure).
    - *Discrete*: isolated values, usually counts (e.g. number of GP visits).
  - **Categorical**
    - *Binary / dichotomous*: two categories (e.g. pass vs. fail).
    - *Nominal*: ≥ 2 unordered categories (e.g. blood type A/B/O/AB).
    - *Ordinal*: ordered categories (e.g. pain score 0–10, Likert scale).
- **Censored data**
  - **Right-censored**: true value is *greater* than a known limit (e.g. patient still alive at study end; age > 90).
  - **Left-censored**: true value is *smaller* than a detection limit (e.g. viral load < 10 copies/mL).
  - **Interval-censored**: true value lies between two known bounds (e.g. infection occurs between two clinic visits two years apart).

- **Getting help & packages**
  - Install once: `install.packages("tidyverse")` *(data wrangling / plots)*
  - Load every session: `library(tidyverse)`
  - Function help: `?lm`, worked example: `example(t.test)`
- **Data import & quick checks**
  - CSV: `df <- read.csv("myfile.csv", stringsAsFactors = FALSE)`
  - Peek: `head(df)`, `str(df)`, `summary(df)`
  - Subset rows: `dplyr::filter(df, Group == "A")`
- **Descriptive statistics**
  - Centre: `mean(x)`, `median(x)`
  - Spread: `sd(x)`, `IQR(x)`, `var(x)`
  - Always add `na.rm = TRUE` if missing values exist
  - Correlation: `cor(x, y)` (number) — `cor.test(x, y)` (CI + p)
- **Base R graphics**
  - Histogram: `hist(x, breaks = 20, main = "Histogram")`
  - Scatterplot: `plot(df`X, `df`Y, `main = "Scatterplot")`
- **Key distribution helpers**

  Normal $Z \sim N(0,1)$
  - Density: `dnorm(z)`
  - Tail area: `pnorm(q)` $(= P(Z \leq q))$
  - Quantile: `qnorm(p)`
  - Random draw: `rnorm(n)`

  t**-dist** T_ν
  - `dt(x, df)`, `pt(t, df)`, `qt(p, df)`, `rt(n, df)`

  Binomial $X \sim \text{Bin}(n, \pi)$
  - Point prob: `dbinom(x, n, pi)`
  - Cumulative: `pbinom(q, n, pi)`
  - Quantile: `qbinom(p, n, pi)`
  - Random draw: `rbinom(N, n, pi)`

  χ^2 **&** **F**
  - χ^2 tail: `pchisq(q, df, lower.tail = FALSE)`
  - Critical χ^2: `qchisq(0.95, df)`
  - F tail: `pf(F, df1, df2, lower.tail = FALSE)`

  - Critical F: `qf(0.95, df1, df2)`

- **Confidence intervals & t-tests**
  - One-sample mean: `t.test(x, mu = mu0)`
  - Two independent groups: `t.test(y  g, data = df)` (`var.equal = TRUE` for pooled)
  - Paired: `t.test(before, after, paired = TRUE)`
  - Exact one-prop CI / test: `binom.test(x, n)`

- **Two-way tables & χ^2 / Fisher**
  - Build: `tab <- table(df`A, df`B); totals: `addmargins(tab)`
  - χ^2 test: `chisq.test(tab)`
  - Small expected counts? use `fisher.test(tab)`

- **Proportion tests**
  - One / two props (large n): `prop.test(x = c(18,12), n = c(30,30))`

- **Simple & multiple linear regression**
  - Fit: `fit <- lm(Y  X1 + X2, data = df)`
  - Inspect: `summary(fit)`; 95% CI: `confint(fit)`
  - Predict: `predict(fit, newdata = data.frame(X1 = 10, X2 = 5), interval = "confidence")`

- **Logistic regression (STAT115 Weeks 10-11)**
  - Binary outcome: `logit <- glm(case  age + sex, family = binomial, data = df)`
  - Odds ratios: `exp(coef(logit))`; CI: `exp(confint(logit))`

- **One-way ANOVA & multiple comparisons**
  - Overall model: `a1 <- aov(y  group, data = df)`
  - Summary table: `summary(a1)`
  - Pairwise Tukey: `TukeyHSD(a1)` *(controls family-wise error)*

- **Simulation snippets**
  - Reproducibility: `set.seed(123)`
  - 1 000 $N(0,1)$ draws: `x <- rnorm(1000)`
  - Central-limit-theorem demo: `ybar <- replicate(1e4, mean(rnorm(50)))` *(hist to visualise)*

- **Workspace utilities**
  - Clear memory: `rm(list = ls())`
  - Save history: `savehistory("my_hist.Rhistory")`

- **Subjective probability** – a personal degree of belief (e.g. "I'm 80 % sure it will rain tomorrow").
- **Objective / long–run probability** – the proportion of times an event occurs in a very large number of identical trials (e.g. coin toss heads ≈ 0.5).
- **Sample space** S – all possible outcomes of an experiment (fair die: $S = \{1, 2, 3, 4, 5, 6\}$).
- **Event** A – a subset of S (e.g. "even number" $= \{2, 4, 6\}$).
- **Complement**: $P(A) + P(\overline{A}) = 1$.
- **Addition rule** (two events): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- **Multiplication rule / conditional prob.**: $P(A \cap B) = P(A)\, P(B \mid A)$.
- **Independent events** – knowing one tells us nothing about the other. Equivalent checks:

$$P(A \cap B) = P(A)\, P(B) \quad \Longleftrightarrow \quad P(B) = P(B \mid A) \quad \Longleftrightarrow \quad P(A) = P(A \mid B).$$

  - A = person *has* the disease, $\overline{A}$ = person *does not*.
  - B = test is *positive*, $\overline{B}$ = test is *negative*.

Sensitivity	$P(B \mid A)$ – probability the test detects the disease.
Specificity	$P(\overline{B} \mid \overline{A})$ – probability a healthy person tests negative.
False-positive rate	$1 -$ specificity $= P(B \mid \overline{A})$.
Positive Predictive Value (PPV)	$P(A \mid B)$ – "If the test is positive, how likely is disease?"
Negative Predictive Value (NPV)	$P(\overline{A} \mid \overline{B})$.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})}.$$

Tip: Low disease prevalence ($\downarrow$) $\Rightarrow$ PPV tends to be low even when sensitivity and specificity are high.

	Disease A	No disease $\overline{A}$	Total
Test + B	a	b	$a + b$
Test – $\overline{B}$	c	d	$c + d$
Total	$a + c$	$b + d$	n

  - Sensitivity $= a/(a + c)$, Specificity $= d/(b + d)$.
  - PPV $= a/(a + b)$, NPV $= d/(c + d)$.

- **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 -¿ 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:**

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

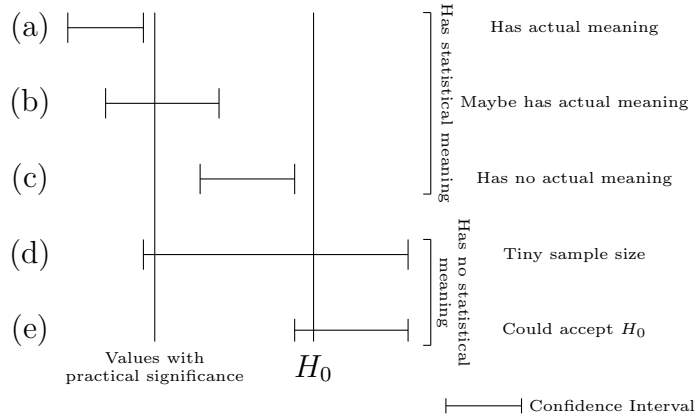
- **Key formula for Confidence interval for relative risk:**

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

- **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.

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

- **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**:

$$\text{Sample mean} = \underbrace{\text{Population (true) mean}}_{} + \underbrace{\text{Error}}_{\text{Systematic error}\quad\text{Random error}}$$

  - **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 the 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 that 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, and the lead researcher.

- **Pros and cons for RCT**

  - Advantages: Experiment - the best way to test a 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**:

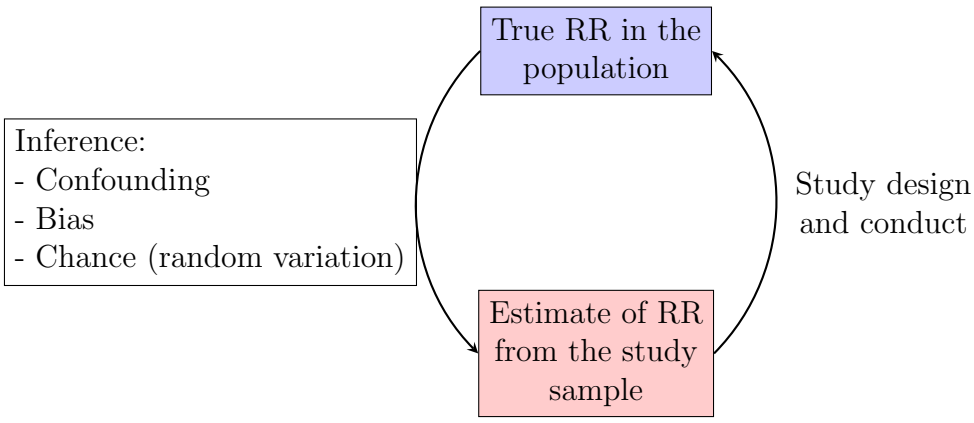
  - Pros: Clear chronological order from reason to outcome. Can evaluate the relationship between multiple results and factor(s).
  - Cons: Large time consumption. Bias affects. Small sample size.

- **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 outcomes 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**



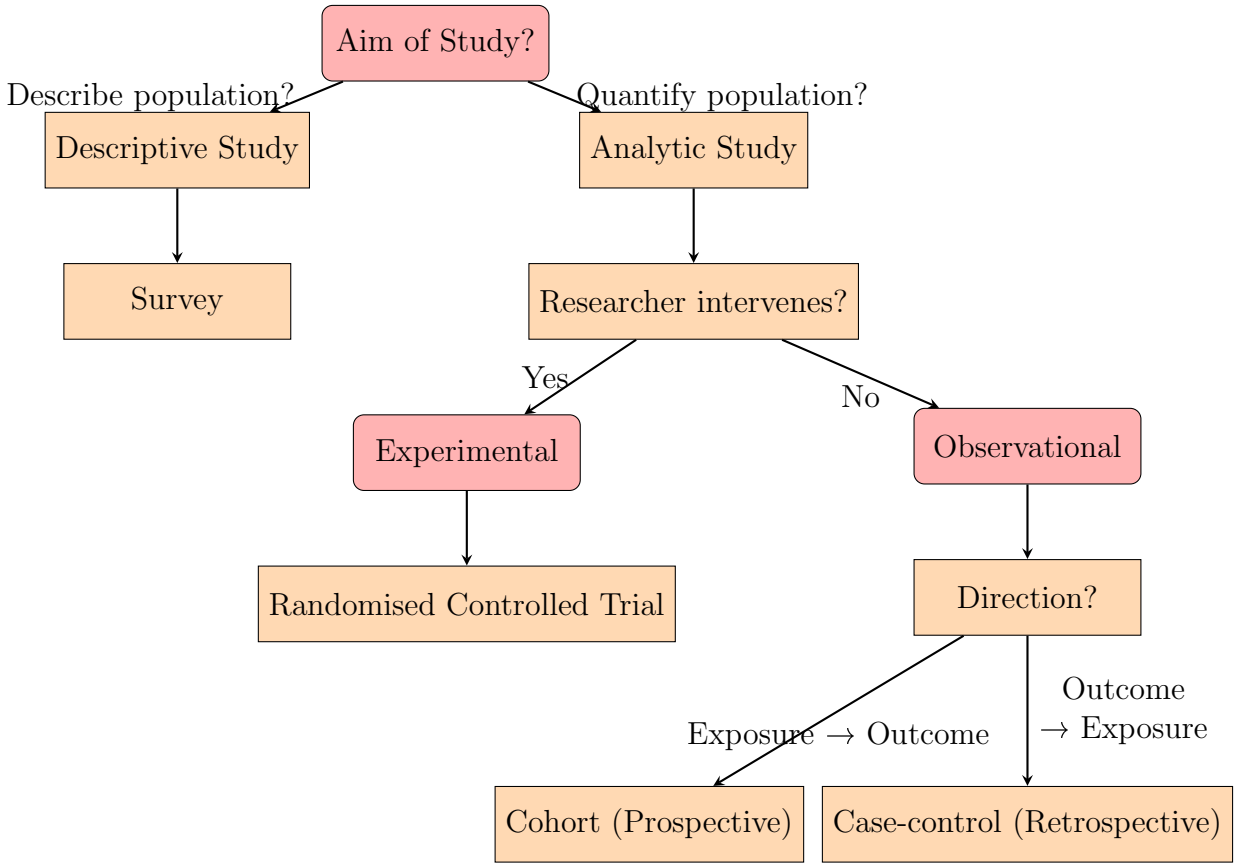
- **What is Confounding?**: Confounding is a distortion of the association between exposure and outcome caused by the presence of a third factor. A confounder is a variable which causes this distortion.

- **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).

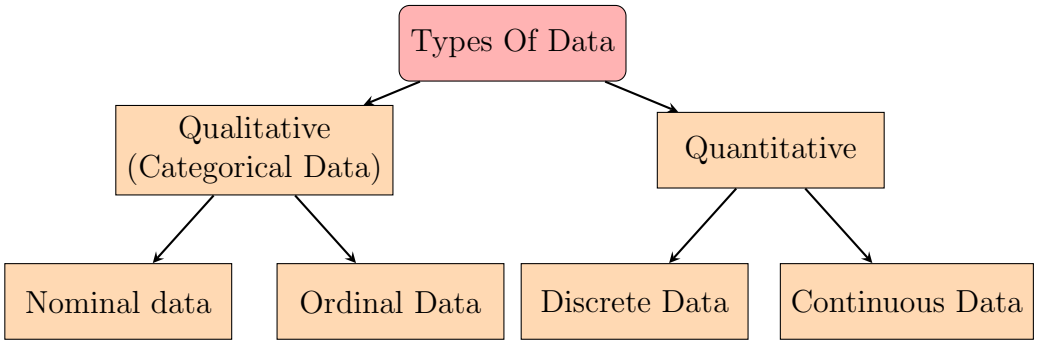
- **RCT, Cohort study, Case-control study**:
  - **Randomised controlled trial**: Analytic, experimental, prospective.
  - **Cohort study**: Analytic, observational, usually prospective.
  - **Case-control study**: Analytic, observational, retrospective.

- **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.

- **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 ± multiplier ± standard error for the mean
- **The standard normal critical value for a 99% interval**: 2.58
- **Multiplier formula**:

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

- **What is the α 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 σ_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**:

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

- ν **(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, π is not close to 0 or 1 (this increases symmetry)

- **Formula for estimating** π:

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

,

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

, 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, π 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 $\mu_P = \pi$, **sd** $= \sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$
- **95% confidence interval for** π (use the sample proportion (p) to estimate the unknown true population proportion (π)):

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

- **Margin of error**:

$$multipliers * sd$$

- **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**:

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

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

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 π 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):

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

- μ **(normal distribution) moves the curve but does not change its shape**.

- σ **spreads the curve more widely about X =** μ **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**: μ is estimated by the sample mean. σ 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 x(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

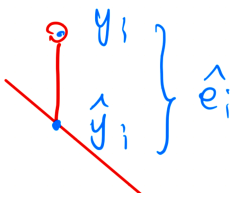
$$\mu_Y = \beta_0 + \beta_1 x$$

  - **Y** is the numerical outcome variable (continuous or approximately so).
  - **x** is the explanatory variable.
  - β_0 is the intercept or constant (where the line crosses the y-axis).
  - β_1 is the slope of the line.
  - **e** (often denoted ϵ) is the random error or residual term.
- **SLR equation for estimating:**

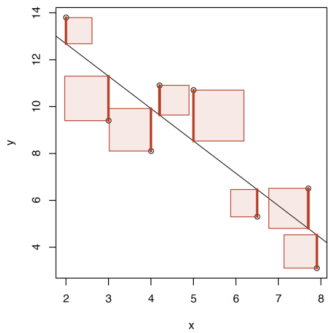
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- **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 β_1 and β_0**:

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

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

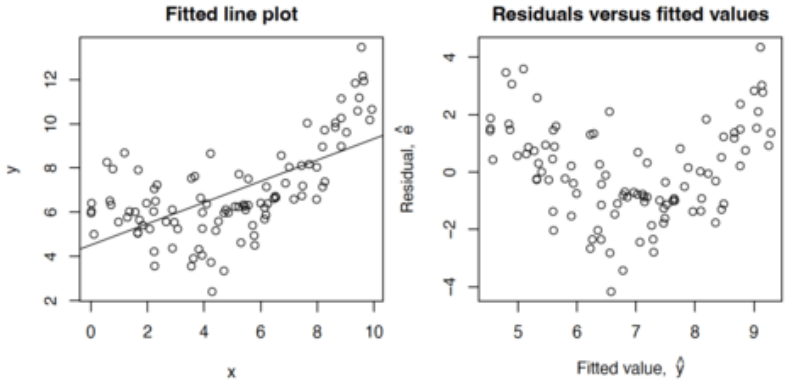
- **Example for how to calculate regression (Stress and Blood Pressure)**:
  - Get n, n = 6
  - Find the explanatory and outcome
  - Calculate β_1 and β_0
  - Get the regression equation
  - Using R for SLR

- **Assumptions for Simple Linear Regression (LINE)**:
  - Linearity: The relationship between the mean response μ_Y and x is described by a straight line.
  - Independence: The responses $Y_1, Y_2, ..., Y_n$ are statistically independent.
  - Normality: The error terms $e_1, e_2, ..., e_n$ come from a normal distribution.
  - Equal variance: The error terms all have the same variance, σ^2 ('homoscedastic').

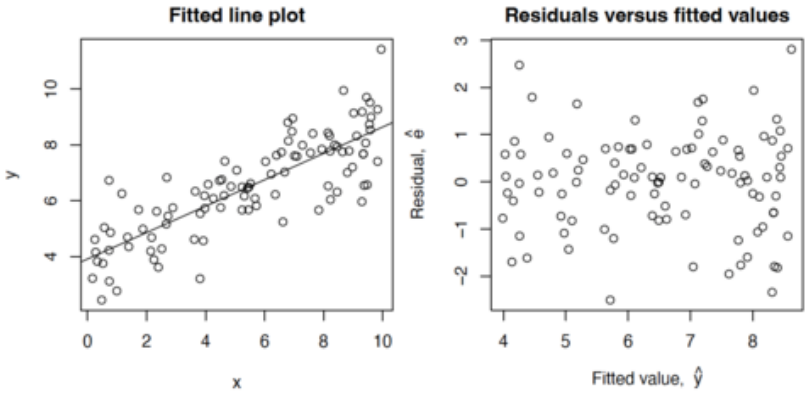
- **What diagram is used for checking linearity**: Residual plot.

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

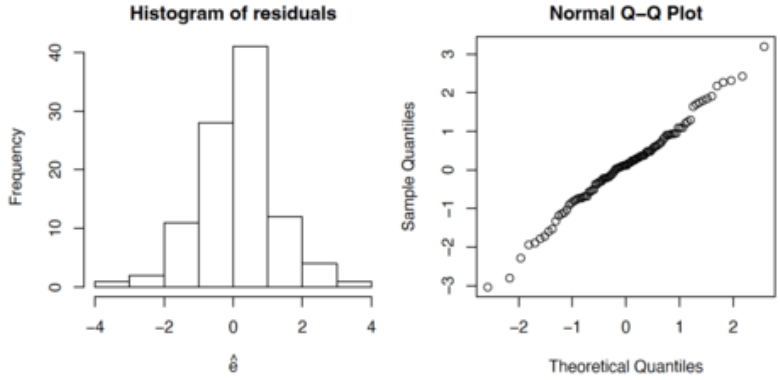
- **Failure of linearity assumption**:



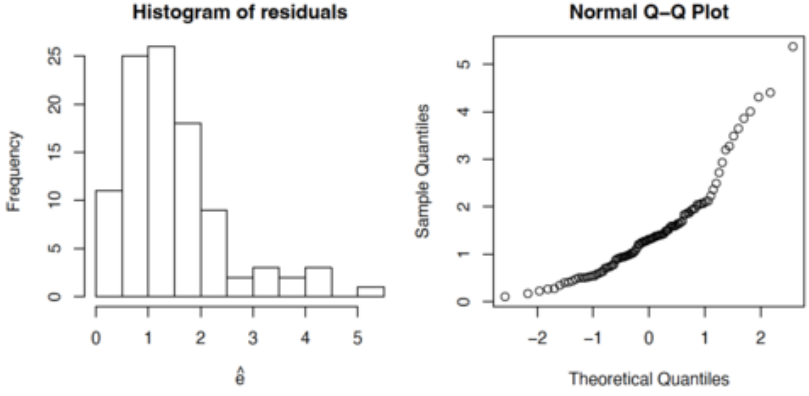
- **Linearity assumption holds**:



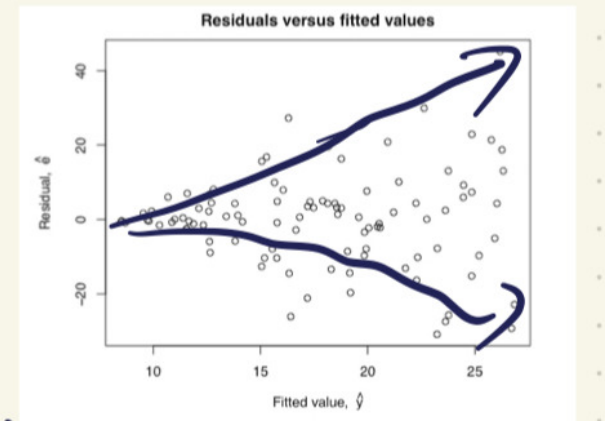
- **Checking independence assumption**: May get insight by thinking about the study design (Ask yourself questions).
- **Plot for Checking the normality assumption**: Q - Q plot.
- **Pass of normality assumption**:



- **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**:

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

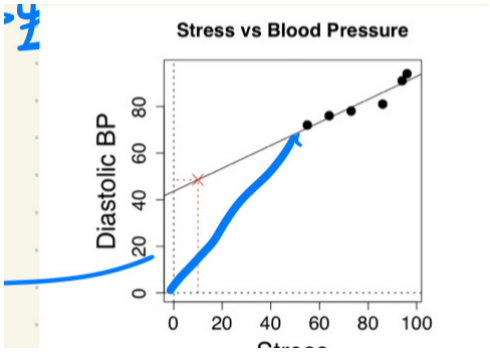
- **What is the SE for SLR's CI**:

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

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

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

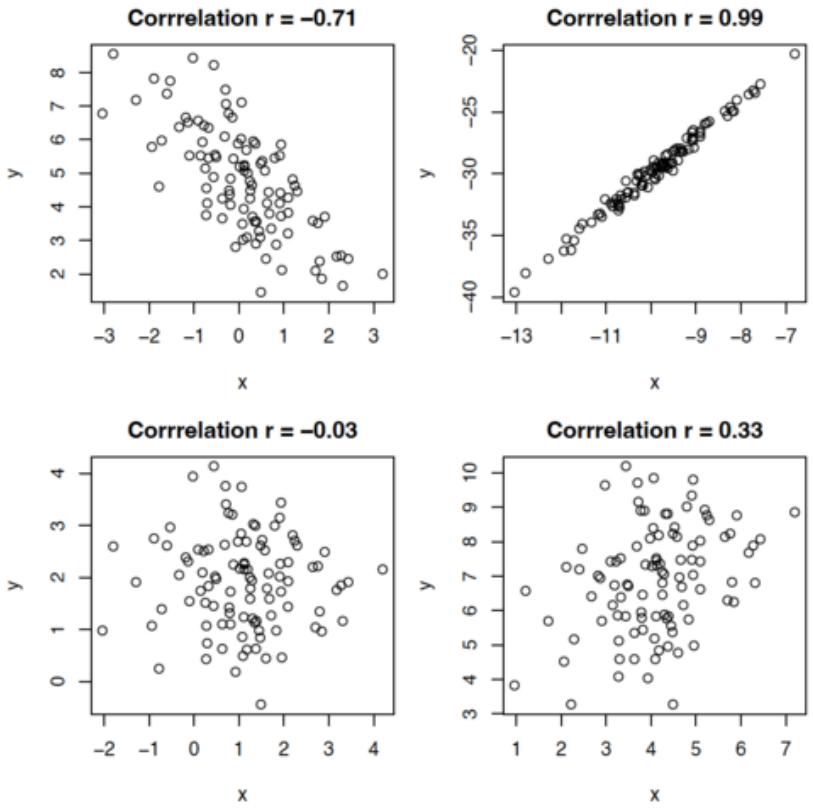
- $\beta_1 = 0$ **indicates what**: That the response is not (linearly) related to the predictor. So the estimated slope will (almost) always be non-zero: $\hat{\beta}_1 \neq 0$.
- **Steps to test to assess the strength of evidence in the data for** $\beta_1 \neq 0$:
  - Setting up the hypotheses: $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$.
  - 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.
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  - $\mathbf{r} \in [-1, 1]$. 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**:



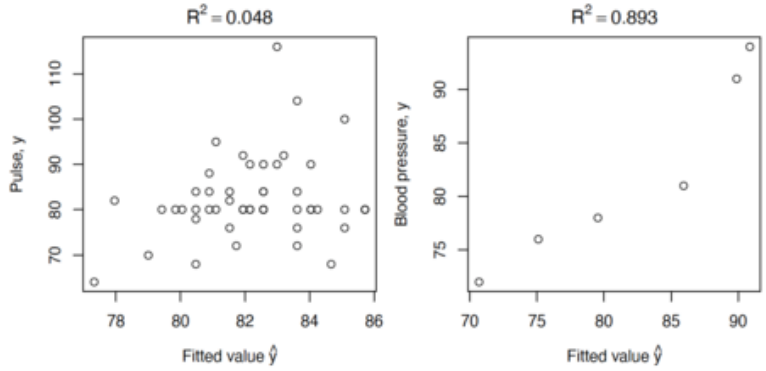
- **Re-write for r**: s_x and s_y are sample standard deviations for x and y variables. s_{xy} is the sample covariance between x and y.

$$= \frac{S_{xy}}{S_x S_y}$$

- **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.

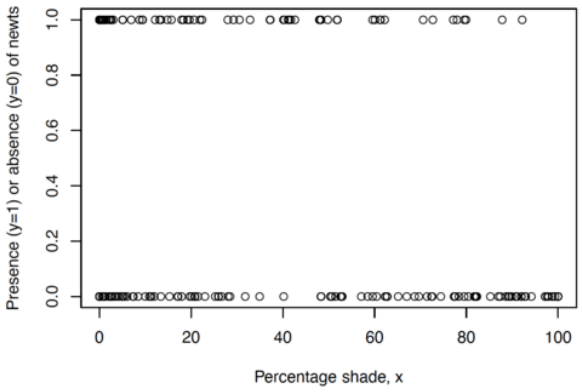
$$ESS = TSS - RSS$$

- **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**:

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

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 β_0, β_1 are the regression coefficients. β_0 is intercept and β_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^{β_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 σ_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 α we reject H_0. If the p-value is greater than or equal to α we do not reject H_0.
- **Test statistic (t - statistic)**: A test statistic is the standardised value of the sample value.
$$T = \frac{observed\ sample\ value - null\ value}{estimated\ standard\ error}$$
When studying hypothesis testing, we usually use π 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 (σ). Basis: Uses population standard deviation (σ).
  - t-statistic: When to use: Small samples (n < 30) or unknown population standard deviation (σ). 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 χ^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:**

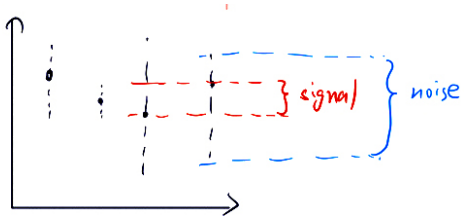
$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- **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 **Analysis of Variance**.
- **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 H0; 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**:

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

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

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

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

- **Sample mean for the ith group:**

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

- **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:

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

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

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

- **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).