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:**

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

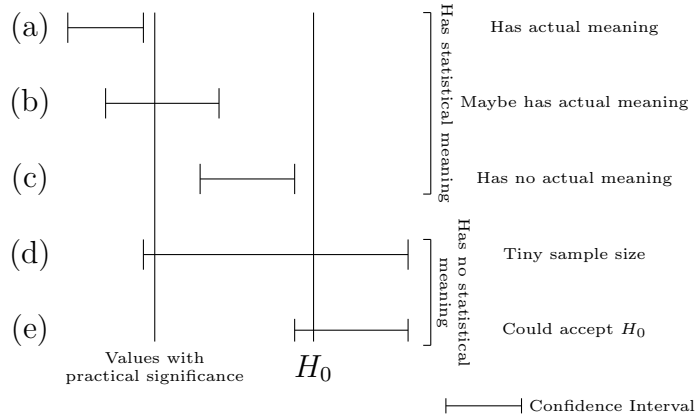
- **Key formula for Confidence interval for relative risk:**

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

- **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.

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

- **Bernoulli** ($X \sim \text{Bern}(p)$) : one trial, outcome 0/1

$$E(X) = p, \qquad \text{Var}(X) = p(1-p)$$

- **Binomial** ($X \sim \text{Bin}(n, p)$) : n independent Bernoulli trials

$$E(X) = np, \qquad \text{Var}(X) = np(1-p)$$

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

  Conditions: binary outcome, fixed n, independent trials, p constant.

- **Normal family** : $X \sim N(\mu, \sigma^2)$
  - Changing μ shifts the curve; changing σ stretches / shrinks it.
  - Standard normal: $Z \sim N(0, 1)$.
  - Convert any normal value to a Z–score: $Z = (X - \mu)/\sigma$.

- t–**distribution** : T_ν has thicker tails than $N(0, 1)$; use when population σ is unknown and sample size is moderate / small. As $\nu \to \infty$, $T_\nu \to N(0, 1)$.

- χ^2 **&** F : arise from squared Z's and variance ratios; used later for goodness-of-fit, contingency tables, and ANOVA.

- **Central Limit Theorem (CLT)** For a simple random sample of size n, the *sampling distribution* of $\bar{X}$ is *approximately* normal for "large enough" n (rule-of-thumb $n \geq 30$ if the parent distribution is not too skew).

- Mean and variance of $\bar{X}$:

$$E(\bar{X}) = \mu, \qquad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \qquad \text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- **Key take-away** : bigger $n \Rightarrow$ smaller SE $\Rightarrow$ more precise estimate of μ.

- A *relative-frequency histogram* shows sample data; a *probability density function* describes the population. Estimate parameters by $\hat{\mu}$ = sample mean and $\hat{\sigma}$ = sample sd.

- **Standardising (Z-score)**: $z = \dfrac{X - \mu}{\sigma}$ — converts any $N(\mu, \sigma^2)$ value to the standard normal scale.

- **Goal**: give a plausible range for a population parameter (mean μ, proportion π, RR, OR, ...) based on a random sample.
- 95 % CI when population standard deviation is *known*:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

  General form: estimate ± multiplier × standard error.

- Replace σ by sample s and use the *t-distribution*:

$$\bar{x} \pm t_{1-\alpha/2,\,\nu} \frac{s}{\sqrt{n}}, \qquad \nu = n - 1$$

  - Works if data are (approximately) normal, *or* $n \geq 30$ (CLT).

- **99 % vs. 95 %**: larger confidence level ↑ ⇒ larger critical value (1.96 → 2.58) ⇒ wider interval.
- **Difference of means** $(\mu_1 - \mu_2)$:

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

  Use Welch's ν (software handles this).

- Sample proportion: $\hat{p} = x/n$.

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

  Conditions: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ (ensures normal approximation to binomial).

- **Margin of error (ME)** = multiplier × SE. Desired ME ⇒ solve for n:

$$n = \left(\frac{z_{1-\alpha/2}\,\sigma}{\text{ME}}\right)^2, \qquad \text{round up.}$$

- 95 % CI means: "If we *repeated* this study many times, 95 % of the calculated intervals would contain the true parameter." It does <u>not</u> say the parameter itself is random.
- Wider interval ⇔ more uncertainty (small n, large s) – always report n alongside the CI.

- **Null hypothesis** (H_0): no effect / no difference / no association.
- **Alternative hypothesis** (H_A): there *is* an effect / difference / association.
- **Generic test statistic**
$$\frac{\text{estimate} - \text{null value}}{\text{standard error}}$$
  - Use Z when σ known or n large (≥ 30).
  - Use t ($\text{df} = n - 1$) when σ unknown and sample moderate / small.
- **p-value**: probability of obtaining the test statistic (or more extreme) if H_0 is true. Reject H_0 when $p < \alpha$ (convention $\alpha = 0.05$).
- **Five-step workflow**
  1. State H_0 and H_A (specify one- or two-sided).
  2. Compute test statistic (Z or t).
  3. Find p-value.
  4. (Optional) Build $(1 - \alpha)$ CI — cross-checks decision.
  5. Conclude in plain language.
- **Errors & power**
  - Type I (α): reject a true H_0 — false positive.
  - Type II (β): fail to reject a false H_0 — false negative.
  - Power $= 1 - \beta$ — boosted by larger n or bigger effect size.
- **Common two-sample tests**

Scenario	Test	R command
Means, indep.	t-test (Welch)	`t.test(y ~ g)`
Means, paired	Paired t-test	`t.test(b, a, paired=TRUE)`
Proportions	Z-test	`prop.test(x, n)`

- **χ^2 test for independence**
  1. H_0: variables independent; H_A: associated.
  2. Compute expected counts: $E_{ij} = r_i c_j / n$.
  3. $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$.
  4. $\text{df} = (r - 1)(c - 1)$; get p with `pchisq(..., lower.tail=FALSE)`.
  5. If any $E_{ij} < 5$, use `fisher.test` instead.
- **CI vs. test linkage:** At the same α, a two-sided test and its CI agree: CI excludes null value $\Leftrightarrow$ reject H_0.

- **Main families of regression**
  - *Linear* – outcome Y *quantitative*.
  - *Logistic* – outcome binary (0 / 1).
  - *Cox* – time-to-event in survival analysis (covered later in STAT115).
- **Terminology**
  - Explanatory variable X: covariate / predictor / independent var.
  - Outcome variable Y: response / dependent var.
- **SLR model**: $Y = \beta_0 + \beta_1 x + \varepsilon$
  - $\mu_{Y|x} = \beta_0 + \beta_1 x$ – mean response at x.
  - β_0: intercept; β_1: slope; ε: random error.
- **Estimated line**: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where residuals $\hat{e}_i = y_i - \hat{y}_i$.
- **Least-squares estimates**

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- **Key assumptions ("LINE")**
  - **L**inear relationship between $\mu_{Y|x}$ and x.
  - **I**ndependent observations.
  - **N**ormal errors ε.
  - **E**qual error variance (homoscedasticity).
- **Diagnostics**
  - Residual plot – check linearity & equal variance.
  - Q-Q plot – check normality of residuals.
  - Leverage / Cook's distance – spot influential points.
- **Error variance estimate**

$$S_e^2 = \frac{\text{RSS}}{n-2}, \qquad \text{RSS} = \sum \hat{e}_i^2.$$

- **Slope inference**

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}, \qquad \text{df} = n - 2, \qquad \text{SE}(\hat{\beta}_1) = \frac{S_e}{\sqrt{\sum(x_i - \bar{x})^2}}.$$

- **Prediction at x_0**

$$\hat{y}_0 \pm t_{1-\alpha/2,\, n-2}\, S_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}.$$

- **Correlation coefficient**

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \ \sum(y_i - \bar{y})^2}}, \qquad -1 \le r \le 1.$$

- **Coefficient of determination**: $R^2 = 1 - \text{RSS/TSS}$ – fraction of variance in Y explained by the model.

- **Logistic model**

$$\log\left(\tfrac{p}{1-p}\right) = \beta_0 + \beta_1 x, \qquad p = P(Y = 1).$$

- **Interpretation**: one-unit increase in x multiplies the *odds* by $\exp(\beta_1)$.

- **Inference on** β_1

$$z = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}, \qquad p\text{-value from } N(0,1).$$

- **Multiple regression model**

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon.$$

  Uses least squares; df for error $= n - k - 1$.

- Typical goals: adjust for confounders, prediction, rank important predictors.

- **A Nalysis O f VA riance** (ANOVA) compares *means* of a quantitative response across $K \geq 3$ groups with *one* global test instead of many pairwise t-tests.
- Model: $Y_{ij} = \mu_i + \varepsilon_{ij}$, $\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ (same variance in every group).
- Sample means: $\bar{y}_{i\cdot} = \dfrac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$, overall mean: $\bar{y}_{\cdot\cdot}$.

$$\underbrace{\text{TSS}}_{\text{total}} = \underbrace{\text{GSS}}_{\text{between}} + \underbrace{\text{RSS}}_{\text{within}}$$

$$\text{GSS} = \sum_{i=1}^{K} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2, \qquad \text{RSS} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2.$$

$$F = \frac{\text{GSS}/(K-1)}{\text{RSS}/(n-K)} \quad \sim \; F_{K-1,\, n-K} \;\text{ under } H_0.$$

Reject $H_0 : \mu_1 = \cdots = \mu_K$ when F is large $\Rightarrow$ right–tail p-value from F-distribution.

Source	SS	df	MS = SS/df
Groups	GSS	$K-1$	$\text{GSS}/(K-1)$
Residuals	RSS	$n-K$	$\text{RSS}/(n-K)$
Total	TSS	$n-1$	

- Normality of residuals (Q–Q plot) and equal variance (residuals vs. fitted). Mild departures are OK if n_i are similar and n is moderate.
- **Multiple comparisons**: identify which groups differ. Classic choice – Tukey's HSD controls family–wise error rate.
  - `a1 <- aov(y ~ group, data = df)` → global F-test.
  - `summary(a1)` → ANOVA table + p-value.
  - `TukeyHSD(a1)` → pairwise CIs and adjusted p-values.
- Repeating m independent tests inflates Type I error: FWER $= 1 - (1-\alpha)^m$. ANOVA keeps the overall α at the planned level.
- Adding a *blocking variable* (e.g. batch, sex) can remove extraneous variation and reduce RSS, increasing power.

- **Analytic studies** – test hypotheses such as "Does a Mediterranean diet increase life-expectancy?" Key principles: *replication* (separate real effect from chance) and *control* (context for the effect of interest).
- **Descriptive studies** – simply characterise *person–place–time* (e.g. lifestyle patterns in NZ).
- A **well-defined population** is precise in space *and* time (e.g. "all colorectal-cancer cases in NZ, 2015").
- **Sampling frame**: list of all eligible units. If incomplete, bias may arise.
- **Probability sampling** – selection chances are *known*. Types:
  - Simple random sample (SRS).
  - Stratified sample – improves precision; allows dis-proportionate strata sizes.
  - Cluster / multi-stage – cheaper but less precise.
- **Error in a sample estimate**

$$\text{Sample mean} = \underbrace{\text{True mean}}_{\text{target}} + \underbrace{\text{Systematic error}}_{\text{bias}} + \underbrace{\text{Random error}}_{\text{chance}}.$$

  - Random error $\downarrow$ when $n \uparrow$.
  - Systematic error (bias) *cannot* be cured by larger n.
- **Experimental study** – investigator intervenes (e.g. randomised controlled trial, RCT). *Randomisation* balances unmeasured factors.
- **Observational study** – investigator only observes (cohort, case-control). Cannot fully rule out confounding.
  - **RCT** (gold standard): analytic, experimental, prospective. Pros – best for causality; Cons – feasibility/ethics.
  - **Cohort**: analytic, observational, usually prospective. Pros – clear time order; Cons – long & costly.
  - **Case-control**: analytic, observational, retrospective. Pros – efficient for rare disease; Cons – higher bias potential.
- **Confounder** – variable associated with *both* exposure and outcome, distorting the association.
- **Two main bias classes**
  - Selection bias – non-comparable groups.
  - Information bias – systematic measurement error.
- **Qualitative (categorical)**
  - Nominal – no order (eye colour).

  - Ordinal – natural order (Likert scale).

- **Quantitative**
  - Discrete – counts (children in family).
  - Continuous – any real value in range (height).

- **Classifying a study**

$$\text{Aim? } \Rightarrow \begin{cases} \text{Describe} \rightarrow \text{Descriptive (survey)} \\ \text{Test} \rightarrow \begin{cases} \text{Intervene? yes} \rightarrow \text{Experimental (RCT)} \\ \text{Intervene? no} \begin{cases} \text{Forward follow-up} \rightarrow \text{Cohort} \\ \text{Backward look} \rightarrow \text{Case-control} \end{cases} \end{cases} \end{cases}$$