- A parameter is a numerical fact about a population
- A census with 100% response rate and no response-bias is the ideal method of exactly determining population parameter
- In most cases it is impossible to determine the parameter exactly, can be estimated using a sample (part of the population)
- Factors for a good estimate (for generalising)
  1 Sampling frame must contain population of interest
  2 Probability sampling (probability of being chosen is non-zero and known)
  3 Must be large enough
  4 High response rate
- Factors 1 and 2 result in selection bias if not enforced
- estimate = parameter + bias + random error
- Bias is influenced by 1, 2 and 4: aim for minimal selection bias and non response
- Random error is influenced by 3. Larger sample -> smaller error
- Selection bias if 1 or 2 not enforced. All non probability sampling (e.g. convenience/ volunteer sampling) results in selection bias

PROBABILITY SAMPLING

Simple random sampling (use of chance)

- Draw units from the population at random without replacement
- Chance of selection at every stage changes, but chance of ending up in final sample is the same
- + Tends to be good representation of population
- - Subject to non response

Systematic sample

- Selecting units from a list by applying a selection interval (that is randomly derived)
- + simpler selection process, can be treated like a simple random sample if numbers are assigned randomly
- May not be representative if sampling list is non-random

Stratified sampling

- Population divided into subgroups (strata), and a random sample is taken from each strata
- Good to use for estimations within subgroups in addition to estimating within population parameter
- Estimate of parameter is done by taking weighted average of subgroup estimates
- Able to get representation from every strata
- Need information about sample frame and stratum

Cluster sampling

- Population is broken down into clusters, then randomly sample a fixed number of clusters
- All observations from selected clusters are included
- + less tedious, less time consuming
- - high variability due to dissimilar clusters or small number of clusters

NON-PROBABILITY SAMPLING

- Selection is done by human discretion rather than probability

Convenience sampling

- Researcher uses subjects that are most easily available to participate in the study
- - Prone to selection bias (some parts of demographic left out)
- - Vulnerable to non-response bias

Volunteer sampling

- Researcher seeks volunteers to participate in the study
- - Non-response bias, where people who do not volunteer are left out
- - Selection bias when some members are not included

VARIABLES

Independent vs dependent variable

- Independent: subject to manipulation in a study
- Dependent: variable hypothesised to change depending on how independent variable is manipulated

Categorical variables

- Take category of label values, each observation can only be placed in one label and labels are mutually exclusive
- **Ordinal**: variables are categories come with natural order and numbers are used to represent it (e.g. happiness)
- **Nominal**: have no intrinsic ordering (e.g. eye colour)

Numerical variables

- Takes numerical values for which arithmetic operation makes sense
- **Discrete**: possible values of the variable form a set of numbers with gaps (e.g. MCs for modules)
- **Continuous**: can take on all possible numerical values (e.g. time)

SUMMARY STATISTICS

- Measures of central tendencies: mean, median, mode
- Measures of dispersion: Standard deviation, interquartile range
- First quartile is the 25th percentile of data values

Standard deviation = square root of variance ≠ spread pattern

- Always non-negative with the same units
- **Adding a constant to all data points does not change standard deviation**
- **Multiplying a value c to all data points results in standard deviation being multiplied by |c|**

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

IQR ≠ spread pattern

- Interquartile range is the difference between the third quartile and first quartile
- Always non-negative
- **Adding a constant to all data points does not change the IQR**
- **Multiplying a value c to all data points results in the IQR being multiplied by |c|**

Mean

- Mean is defined as the sum of all data points divided by the number of data points
- Does not tell distribution over total, frequency
- **Adding a constant value to all data points changes that mean by that constant value**
- **Multiplying a value c to all data points results in the mean being multiplied by that constant c**

Median

- Median of a numerical value in a data set is the middle value of the variable after arranging all the values in ascending / descending order
- Does not indicate total value, frequency or distribution
- **Adding a constant value to all data points changed the median by that constant value**
- **Multiplying a value c to all data points results in the median being multiplied by constant c**

Mode

- Value that appears most frequently in the data set
- Mode is the ”peak” of the distribution graph of a discrete variable

TYPES OF STUDIES

Experimental studies

- Goal is to prove a cause and effect relationship
- Independent variable should be the only variable that results in a change in dependent variable
- Researchers assign participants to control and experimental groups (random assignment/ draw without replacement is best)
- Groups can have different sizes as long as sizes are quite large
- Single-blinded experiment: participants/ evaluators do not know whether they are in (or evaluating) the treatment or control groups
- Double-blinded experiment: both
- Researchers know

Observational studies

- Self assign to respective groups
- Only establishes association
- **A confounder is a third variable associated with both the independent and dependent variables**
- Confounder must be a different variable than the dependent and independent variables
- More confounders that the study can control and still show association → stronger evidence for a genuine relationship

RATES

- **Marginal** rate is the probability of an event occurring; not conditioned on another event occurring
- **Joint** rate is the probability of 2 events occurring together (i.e. intersection of the probabilities of 2 events, P(A and B); denominator is total number
- **Conditional** rate is the probability of event A occurring given that event B occurs, P(A|B); denominator is B (given condition)

ASSOCIATION

- A and B are associated with each other if rate (A|B)≠ rate (A | not B)
- A and B are positively associated if rate (A|B) > rate (A | not B)
- A and B are negatively associated if rate (A|B) < rate (A | not B)
- A and B are interchangeable due to symmetry of rates, i.e. rate (A|B) > rate (A| not B) iff rate (B|A) > rate (B| not A)
- Symmetry rule:
- rate(A|B) > rate(A|NB) ↔ rate(B|A) > rate(B|NA)
- rate(A|B) < rate(A|NB) ↔ rate(B|A) < rate(B|NA)
- rate(A|B) = rate(A|NB) ↔ rate(B|A) = rate(B|NA)

BASIC RULE OF RATES

- rate(A) is always between rate (A|B) and rate (A| NB)
- Given rate (A|B) = x and rate (A|C) = y, with B and C disjoint, min {x,y} ≤ rate(A |B ∪ C) ≤ max {x,y}
- The closer rate (B) gets to 100%, the closer rate (A) gets to rate (A|B)
- Rate (A) is exactly in between rate (A|B) and rate (A| not B) if rate (B) = 50%

SIMPSON'S PARADOX

- Relationship between rates in subgroups is reversed / disappears when subgroups are combined
- **Sure sign of confounder**
- To determine is a variable is a confounder, data must be collected on it
- Allocate factors proportionately to remove association between variable and treatment type
- Randomized assignment gives equal proportion most of the time, but is not always possible as people cannot be forced
- To control confounder, slicing is used
- Slicing: subgroup analysis is used (conclude based on subgroup numbers instead of misleading overall data)

To **calculate the marginal rate**, rate(Curly), we take the column totals of all Curly-haired persons (both Male and Female) divided by the grand total of everyone in the data set, $\frac{(77+72)}{373} \approx 39.95\%$

To **calculate the joint rate**, rate(non-Black and Female), we take the count of “Females with non-black hair” divided by once again the grand total of everyone in the data set, i.e.

$$\frac{(9+20+55+5+16+27)}{373} \approx 35.39\%$$

NUMERICAL DATA

- A distribution is an orientation of data points, broken down by frequency of occurrence

Histograms

1. Shape of a distribution: peaks/ skewness
- One distinct peak in a unimodal distribution
- Symmetrical distribution: peak is in the middle
- Left skewed: peak is shifted right
- Right skewed: peak is shifted left

2. Centre of a distribution: mean, median, mode
- Symmetrical: mean, median, mode close
- Left skewed: mean < median < mode
- Right skewed: mean > median > mode

3. Spread of a distribution: range, standard deviation
- Higher the variability, wider the range in which the data being spread across
- Most common measure is standard deviation

4. Outliers
- Mean can be pulled far in direction of skew, not a good measure of the central tendency
- Removal cause s.d. decrease; Q1Q3 same/ decrease; IQR increase/ decrease/ same

Boxplots

- Q1, median, Q3 and maximum to construct
- Outlier if value > Q1+1.5*IQR or
  value < Q1-1.5*IQR

1. Shape
- Deduced by comparing the variability in the upper half of the data to the lower half of the data
- Skewed right if lower half has less variability than the upper half

2. Centre
- Deduce the median value at a glance

3. Spread
- IQR gives an idea of the spread for the middle 50%

BIVARIATE EXPLORATORY DATA ANALYSIS

1. Scatter plot
- To have an idea of the pattern formed between 2 variables
- Direction (+ve, -ve, neither)
  - One increases, the other increases/decreases
- Form (general shape)
  - Linear form (data scattered about a line)
- Strength
  - How closely the data follow the form
- Outliers
  - Points deviating

2. Correlation coefficients (linear association)
- Range between -1 and 1
- (0-0.3: weak, 0.3-0.7: moderate, 0.7-1: strong)
- r > 0: positive association
- r < 0: negative association
- r = 0: no linear association
- Calculation
  - Convert data point into standard unit:
  - SUx = $\frac{x - average\ x}{sx}$, where sx is standard deviation
  - r value is the average of product of X and Y
- **r is not affected by interchanging two variables, adding a constant, multiplying a positive constant**
- r value might be present for no-linear association
- Outliers can increase/ decrease strength

3. Regression analysis
- Fit a line or curve to a data set and do predictions
- y = mx + c
- **gradient m = $\frac{sy}{sx}$r** (hence sign is the same)
- y obtained from subbing into equation cannot be used to predict x
- Prediction of y beyond observed range of x is wrong

PROBABILITY

- Sample space: A collection of all outcomes of a probability experiment
- Event: sub collection of the sample space
- Sample statistic: refers to the use of a sample to draw a conclusion about the population
- Rules of probability:
  - $0 \le P(E) \le 1$ for each event E
  - P(S) = 1 if S is the entire sample space
  - Mutually exclusive event: $P(E \cup F) = P(E) + P(F)$

Uniform probabilities and rates

- Equal probability to every outcome: $\frac{1}{size\ of\ sample\ space}$

Conditional probabilities

- P(E | F) = rate(E | F); if P(F) = 0, P(E | F) = 0
- P(E | F) = $\frac{P(E \cap F)}{P(F)}$
- P(A) is always between P(A | B) and P(A | not B)

Independence (means no association)

- $P(A) \times P(B) = P(A \cap B)$

RANDOM VARIABLES

Discrete random variables

- Points in the plot separated by gaps
- Mode is the x values of the highest point
- $P(X \ge 5) = P(5) + P(6)$

Continuous random variables

- Visualised with a density curve
- $P(0.3 \le Y \le 0.5)$ = shaded area under the curve in interval 0.3 to 0.5

Normal distributions
- N(x, y): normal distribution, **mean x and variance y**
- Properties:
  - Bell-shaped curve
  - Peak of curve occurs at the mean
  - Curve is symmetrical about the mean
  - Mean = median = mode
  - Area under curve = 1

STATISTICAL INFERENCE

- confidence of the sample statistic being used to estimate the population parameter
- provides a range of values that we are reasonably certain that the population parameter lies in
- C.I. for population proportion

$$p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$$

  - p* is the sample proportion
  - z* is the value from the standard normal distribution
  - n is the sample size

- Properties
  - larger the sample size, smaller margin of error and C.I.
  - tighter bound when C.I. % is lower (bigger % C.I. more samples, bigger range)
  - sample statistic will always be within C.I. (sample mean within C.I.)
- 5% will not contain the population mean
- 99% confidence interval [1.58, 1.80], we can infer that the sample mean is 1.69 (avg of the two), and margin of error is 0.11 (difference/2)
- Does not mean that there is a 95% chance that our population parameter will fall in C.I.
- Means that if we collect many random samples and construct a C.I. for each of them, about 95% of them would contain the population parameter
  - 95 of 100 C.I. contain population parameter

HYPOTHESIS TESTING

1. Null and alternative hypothesis
- Null hypothesis takes the stance of no effect
- Both hypotheses mutually exclusive

2. Calculate p-value
- p-value: probability of obtaining a test result at least as extreme as the result observed, assuming null is true
- OR probability of observing data at least as favourable to the alt as data set, if null is true

3. Conclusion about null hypothesis
- p-value < sig level, reject
- p-value ≥ sig level, do not reject

Chi-Squared Test

- Used to check whether 2 categorical variables are significantly associated
- Data must be counts for the categories of a categorical variable

One sample t-test

- Requires population distribution to be approximately normal (n ≥ 30)
- Used for testing for significant difference between sample mean and known mean

To **calculate the conditional rate**,
rate(Female | Gamer) = $\frac{48}{144}$ = 0 .33,
rate(Female | Non-Gamer) = $\frac{72}{136}$ = 0 .53.
Since rate(Female |Gamer) < rate(Female |Non-Gamer), there is negative association between being female and being a gamer.

Hypothesis test to find out whether X prevents cancer which affects 10% of dogs. Random sample of 100 dogs received X and 5 eventually had cancer
p-value = probability that 5 or less puppies out of 100 have cancer, given that the probability of cancer is 0.1 (result at least as extreme as observed data → 5 or more or less) & (find out if X prevents cancer → less than 5)

Determine **outlier from boxplot**:
IQR is 18 - 15 = 3
24 (max val) > 18 + 1.5 x 3 = 22.5
(falls 1.5 away from the max/ min value)
At least one high outlier
15 - 1.5 x 3 = 10.5 and there are no values smaller than 12 (min val)
No low outliers

Correlation coefficient of subgroups
1 subgroup has r = 0.8, another subgroup has r = 0.8
combined r may not be 0.8 (can be anything)

Linear regression

Variables in regression line

- Weak association between 2 variables corresponds to a non-deterministic relationship
- Regression equation gives us the predicted average Y values for a given X value and not the exact Y value
- Cannot draw conclusions of the predicted average values of Y for any X beyond the range of X values in data (no extrapolation)

Properties of r

- Gradient of regression line for Y vs X ≠ gradient of regression line for X vs Y (unless Sx = Sy)
- Adding any value to X (+6) and subtracting any value from Y (-3) does not change r
- Multiplying both X and Y by -1 does not change r

Location of points

- Each older sibling is heavier than younger sibling
- For each point (x, y), y > x → all the lines will lie above the line y = x in scatter plot

Summary statistics
40 students in a class; each student scored 5m less for math than science
- If A > B for math, then A > B for science
- S.d. for math and science are equal
- No difference in IQR
- Perfect positive correlation between math and science marks: y = x + 5