- 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
- Includes convenience sampling and volunteer sampling

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
- Standard deviation and IQR does not relate to spread pattern
- First quartile is the 25th percentile of data values
- Standard deviation = square root of variance ≠ spread pattern

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

Mean
- Mean is defined as the sum of all data points divided by the number of data points
- **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
- The mode is that value that appears most frequently in the data set
- When describing distribution of points of a discrete variable, mode is the ”peak” of the distribution graph

IQR ≠ spread pattern
- Interquartile range is the difference between the third quartile and first quartile
- IQR is 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|**

Experimental studies

- Primary goal is to prove a cause and effect relationship between 2 variables
- To establish cause and effect, the independent variable should be the only variable that results in a change in the dependent variable
- Researchers can assign participants to control and experimental groups (random assignment is best)
- Random assignment: random draw without replacement
- Groups can have different sizes as long as sizes are quite large

Blinding

- Blinding is done to guard against human bias.
- In a single-blinded experiment, either participants or evaluators do not know whether they are in (or evaluating) the treatment or control groups
- In a double-blinded experiment, both participants and evaluators do not know whether they are in (or evaluating) the treatment or control groups
- Researchers still know which group is subjected to which treatments

Observational studies

- In an observational study, participants self assign to each of the respective groups (may be unethical to assign)
- Observational studies can only establish association between variables
- Association does not imply causation
- **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
- The more confounders that the study can control and still show association results in stronger evidence for a genuine relationship

Comparison

- Comparison method is used to see the effect of treatments on outcomes
- Comparison of outcomes between a treatment and control group
- Control and treatment groups should be similar
- Subjects should be put in control and treatment groups randomly
- A large number of participants should be used. Law of large numbers ensures all other variables are almost equally present in both groups

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)

	Hair type				
	Straight		Curly		
Colour	Male	Female	Male	Female	Total
Red	7	9	8	5	29
Brown	35	20	12	16	83
Blonde	51	55	38	27	171
Black	22	25	19	24	90
Total	115	109	77	72	373

The marginal rate, rate(Curly), is _(a)_; while the joint rate, rate(non-Black and Female) is _(b)_ %.

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

	Female	Male	Total
Gamer	48	96	144
Non-Gamer	72	64	136
Total	120	160	280

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.