# Chapter 1: Design of Studies

QR Framework

* Frame, specify, collect, analyse, communicate

**Controlled Experiment**

Assigning Subjects to Treatment or Control Groups

Treatment group: Children with parental consent

* Control group: Children without parental consent
  + In view of the ethical issue that it is morally unjust to deny treatment to a few (although for the benefit of many)
* However, confounder: socio-economic status
  + Children in poor households contract polio early on in life, and many will recover completely; prior exposure gives immunity from future infections

Randomised Assignment

* Aim: Control and treatment groups have similar risk
* The larger the number of subjects, the more likely the similarity between the groups, minimising confounding
* Randomised, controlled, double-blind experiments
  + Placebo, blinding subjects and blinding doctors
* Non-randomised experiments’
  + Historical vs present data (different environment and situations)
  + Weak to control group, stronger to treatment group (overstatement of results)

**Observational Study**

* Some things can only be investigated through observational studies eg smoking on health

*Type of Observational Study*

* Cohort study
  + People selected by exposure (eg smokers vs non-smokers)
  + At the beginning, no one has HD (future response)
  + Can use both RR and OR
* Case-control study
  + People selected by response side (eg HD vs no HD)
  + How many of them smoked in the past (past data)
  + Can use only OR

*Issues in observational studies: Confounders*

* Associated with both exposure and the disease
* Obscures actual relationship between exposure and disease
  + In a controlled experiment, confounding minimised by randomised assignment
  + In an observational study, must try to control for confounders
* Controlling for confounders through slicing
  + Separating by confounder eg gender, age

*Issues in observational studies: Adherence*

* Although the initial experiment may be a randomised, controlled, double-blind experiment, whether or not subjects choose to adhere will be an observational study
  + Eg A trial drug is prescribed, but subjects can choose to take the drug every day
  + As it is an observational study, are there now confounders present?
    - Those who are more health conscious would adhere, also have a reduced risk of sickness
* Also shows success of blinding
  + If drug has no side effects, comparing with others, might break the blind

*Issues in observational studies: Simpson’s Paradox*

* Relationships between percentages in subgroups can be reversed when subgroups are combined
* Eg A higher percentage of men was accepted into a uni overall, but within each individual course there was equal percentage of men and women accepted
  + Confounder: choice of subgroup (affects probability of getting in (admission rates), affects the likelihood of you being a male)
  + Controlled by slicing by major
  + Can also be controlled by an adjusted admissions rate (if no. of men and women who applied by major was equal)

**RR and OR**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Heart Disease | No HD | Total |
| Smokers | A | B | A+B |
| Non-smokers | C | D | C+D |
| Total | A+C | B+D |  |

*Calculating Rate*

* Rate (HD|smokers) = A/(A+B)
* Rate (HD|non-smokers) = C/(C+D)
* Rate (smoking|HD) = A/(A+C)
* Rate (smoking|no HD) = B/(B+D)

Symmetry rule

* Positive association: If r (HD|smokers) > r (HD|non-smokers), then r(smokers|HD) > r (smokers|no HD) as long as 0 < rate A, B < 1
* Negative association: If r{A|B) < r(A|not B), r{B|A) < r(B|not A)
* Not associated: If rate (A|B) = rate (A|not B), then rate{B|A) = rate (B|not A)

*Risk Ratio*

* Risk: a more specific rate
* Risk (HD|non-smokers) = C/(C+D)
* RR between smokers and non-smokers =
  + If RR = 1, No Association
  + If RR > 1, higher risk in the first group
* Population RR can only be estimated accurately from cohort studies

*Odd Ratio*

* Odds (HD|smokers) = A/B
* Odds (HD|non-smokers) = C/D
  + Odds =
* OR between smokers and non-smokers =
* If OR = 1, RR = 1
* OR > 1, RR > 1
* Population OR can be estimated accurately from both cohort and case-control studies

# Chapter 2: Association

Bivariate (X, Y) data

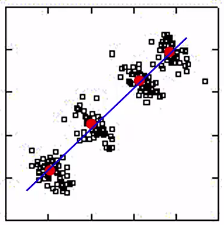
* Standard deviation (sd) shows the average distance from each data point to their average
* Typically represented by a scatter diagram
* Strength of linear association measured by correlation coefficient, r

**Correlation coefficient**

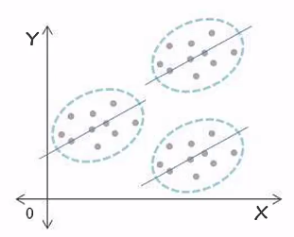
* 0 < r < 1: positive association (r > 0.7 = strong association)
* -1 < r < 0: negative association
* r = 0: no linear association
* Computation: Take the average of the product of the standard units (SU) of each data pair (SU = )
* r is not affected by
  + Swapping axes
  + Adding the same number to all values of a variable
  + Multiplying a positive number to all values of a variable
* Limitations
  + Correlation and causation: correlation doesn’t imply causation. R tells only association, not correlation or causation.
  + Outliers affect correlation
  + R only gives linear association

**Ecological Correlation**

* Correlation of data on a group basis
  + Variation of data is due to ecological/environmental differences (eg country)
  + Data of individuals 🡪 grouped 🡪 correlation between group aggregated data
* Association of individuals tends to be overstated if based on aggregated data, because variability among individuals will be reduced

*Ecological Fallacy*

* Deducing inferences on correlation about individuals based on aggregated data
* Eg if the graph is food consumption against health for different countries. Looking within one group, easy to say that less food = better health. However, looking overall: trend shows that for individuals, more food = better health.

*Atomistic Fallacy*

* Generalising correlation for a group based on individuals
  + Within each group (among the individuals), a positive linear correlation exists. However, no clear relationship between the groups of data.

**Attenuation Effect**

* Range restriction
  + When the range of one variable is limited, r tends to be smaller, association between variables is understated
* Removal of data
  + Changes the context of the data
* Regression fallacy
  + “Regression toward mediocrity”. The best group will, on average, fall back while the bottom group will, on average, improve in test-retest situations

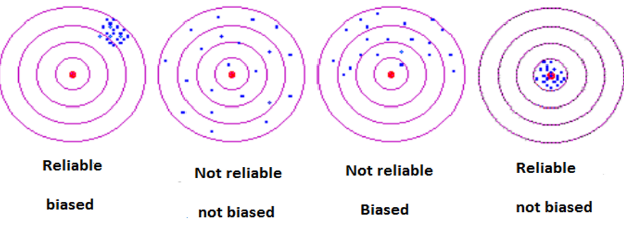
**Linear Regression**

* Predicting with a regression line
  + X: independent (predictor), Y: dependent. Y = a + bX
  + Equation can be obtained by excel. Obtained by minimising the distance from each point to the line.
* Note: Prediction of Y is a prediction of the *average* Y when X = x.
* Note: Prediction of Y beyond range of X based on the regression line will be inaccurate.

# Chapter 3: Measurement

A Common Language

* Variables
  + Categorical
    - Ordinal: Good, okay, bad or happy, okay, sad etc
    - Nominal: Names
  + Numerical
* Accuracy of measurement: Validity, reliability, bias
  + Validity: Extent to which a measuring tool measures what it is supposed to measure
  + Reliability: The same approximate result is obtained for the same subject
  + Bias: A measurement that is systematically off the mark in the same direction
* True score theory: useful for making good measurements
  + Every measurement, X = True value (T) + random error (er) + systematic error (es)
  + Minimise random error by averaging

Challenges when asking questions

* Deliberate bias
* Unintentional bias
* Desire to please
* Asking the uninformed
* Unnecessary complexity
* Ordering of questions
* Confidentiality and anonymity

# Chapter 4: Sampling

Data collection

* Unit (individual) vs Sample vs Population
* Generally a census of the whole population is preferred, but impossible
  + But taking a sample over a census has advantages: less time consuming, costly, also during blood tests a sample and not all your blood is taken

Getting a Good Sample

* Having a sampling frame that is representative of the population. Aim: extend the results of the study to the population
* Probability sampling
  + Every unit has a known probability of being selected
  + Simple random sampling (easiest)
    - Randomly selecting, every sample has the same chance of being selected
  + Systematic sampling
    - Can be used for populations with unknown exact size
    - Select the nth person from a list numbered randomly
  + Stratified sampling
    - Taking a probability sample from strata
    - More efficient, can compare within strata to eliminate confounders
  + Multistage sampling plan
    - Several stages of selection by probability sampling
* Difficulties in sampling
  + Using an imperfect sampling frame
    - Unwanted units included, desired units not included
  + Non-response
    - Uncontactable/unwilling selected units distorts results
  + Voluntary or self-selected samples
    - Volunteered opinions tend to be more biased
  + Convenience or haphazard samples
  + A judgement sample
    - Those who are “typical”/”representative” are selected
  + Quota samples
    - Having a quota for different categories; but this categorisation might not affect the results, unnecessary
* If different proportions of strata respond to a survey over different years, overall observed population changes can be different
  + eg starting salary – many non-professional degree holders respond in the first year, many professional degree holders in the second year

# Chapter 5: Uncertainty

*Relative Frequency and Personal Probability*

* Random circumstances: outcome not determined till observed
  + Relative frequency – Can be quantified exactly based on repeated observations (eg red or blue)
  + Personal probability – “will I do this today?”
* Relative frequency interpretation
  + Probability = relative frequency of event (3/20, 1/6)
  + Risk = probability
  + Probability of each outcome for a random circumstance with N equally likely outcomes = 1/N
* Personal probability interpretation
  + Circumstances are not repeatable, only apply to particular individuals, and will only happen once
  + Probability based on individual’s belief, past experiences, subjective

Probability Rules

* Mutually exclusive events
  + Events are mutually exclusive if they cannot occur at the same time
* Independent events
  + Events are independent of each other if they do not affect each other’s chance of occurrence
* Complement rule
  + P(event) = 1 – P(not event)
* Addition rule
  + When events are mutually exclusive, probability of either of these events occurring (or) is the sum of their individual probabilities
* Multiplication rule
  + The probability of two independent events occurring at the same time is the product of their individual probabilities
* Odds (event) =

*Average Values*

* If you have to answer 1 MCQ question (4 options), get $5000 if you get it right, and $0 if it’s wrong, then the average amount you get each time you answer is
  + Also called expected value or mean value
  + 1 Identify all the possible, mutually exclusive outcomes of the activity. 2 Determine the value of each outcome, as well as the probability of each outcome. 3 Average value = sum of V(A) x P(A) of every outcome.

P-value

* P-value: the probability of a ‘strange’ outcome happening, assuming null hypothesis holds
* Significance level (eg 0.05): If p-value > 0.05, then the null hypothesis at 5% sig. level is not rejected. If p-value < 0.05, then reject null hypothesis at 5% significance level

Conditional Probabilities

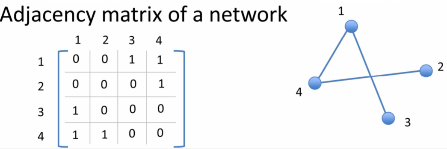
* P(A and B) = P(B) x P(A|B)
* If A and B are independent, then
  + P(A|B) = P(A)
  + P(A and B) = P(A) x P(B)
  + P(A|B) ≠ P(B|A) in general

Testing Rare Events

* Base Rate = P(disease)
* Sensitivity = P(positive| disease)
* Specificity = P(negative| no disease)
* Why should screening tests be done?
  + No alternative test
  + Good chance of successful treatment
  + Test is inexpensive

# Chapter 6: Networks

Terminologies

* Vertices: two points connected when they have something in common
* Edges: lines connecting vertices
* Degree: the number of edges a vertex is incident with
* Adjacent: when two networks are joined directly by an edge
  + Most of the time, degree of vertex A = no. of vertices adjacent to A, except when A and B are connected by more than 1 edge
* Order: Number of vertices in the network
* Size: Number of edges in the network
* Distance d(X,Y): smallest number of steps needed to be taken from X to Y
* Matrix: number each vertex, then draw a matrix where 1 represents the vertices are adjacent

Average distance of a vertex

* The average distance of that vertex from all other vertices

Centrality Measures

* Closeness: Ccen(u) = (sum of all distances from vertex u to all other vertices in the network)/(n-1) where n is the number of vertices in the network
* Degree: Dcen(u) = deg(u)/(n-1)
* Betweenness: Bcen(u)
  + List down all possible pairs of vertices other than u and find the shortest path between them
  + Calculate the proportion of shortest paths that pass through u for each pair
  + Average all proportions found