Chapter 0: Quantitative Reasoning Framework

Quantitative Reasoning (QR) Framework. There are a total of 5 steps that are involved in the QR Framework.

1. Frame
2. Specify
3. Collect
4. Analyse
5. Communicate

Framing the Question (Step 1). QR questions need to be framed *suitably* to leverage on quantitative data or methods to produce *precise* answers. A good QR question is quantitative and precise.

Specifying the Relevant Quantity (Step 2). Are there any data points that we should exclude? How do we measure our quantities of interest?

Collecting the Data (Step 3). How do you collect the data? Do you take a *census* or a *sample*? How do we decide who and how many to collect?

Analyse the Data (Step 4). How do we understand the data collected? What tools can I use to gain some understanding from the data? (E.g., correlation coefficient)

Communicate the Findings. Answer the question that was framed. It is important to communicate the findings in *non-technical* language. Recognise the limitation of our study.

Chapter 1: Design of Studies

Cause and Effect. Some causal relationships are hard to establish.

Control Groups and Treatment Groups. In any *controlled experiment*, we need to ask if control and treatment groups are *different*, as their differences may result in biasness in the results.
If the control group and the treatment groups are similar, then the difference in response can be attributed to the treatment. Otherwise, the treatment effect may be confounded with some other factor.

Rate. Used to control for unequal group sizes. (i.e., number of polio cases per 100,000 children)

Randomised Assignment. An many subjects, it is likely that the two groups are similar in all aspects.

Randomised Control Double-blind Experiment. Experiment that has a *placebo*, *blinding subjects*, and *blinding doctors*.

Placebo. A substance with no effect that cannot be differentiated easily from the treatment.

Blinding Subjects. Subjects that do not know if they are in the treatment or the control group.

Blinding Doctors. Doctors who make diagnoses do not know if the subject is from the control group or treatment group.

Observational Study. Investigator merely records what happens to the subjects, who choose whether they are in the control or treatment group.

Observational Study vs Controlled Experiments. The key different between an observational study and control experiment is the assignment of subjects.

	Controlled Experiment	Observational Studies
Assignment by…	Investigators	Subjects

Association is not Causation. Association does not always arise from causation.

Two-by-two Table. A two-by-two table records the data and allows us to calculate the rate.

	Disease	No Disease	Row Total
Smoker	**A**	**B**	**A + B**
Non-smoker	**C**	**D**	**C + D**
Column Total	**A + C**	**B + D**	**A + B + C + D**

Rate of disease among:

1. Smoker = $\frac{A}{A+B}$ = rate(disease | smoker)
2. Non-smoker = $\frac{C}{C+D}$ = rate(disease | non-smoker)

Rate of smoking among:

1. Disease = $\frac{A}{A+C}$ = rate(smoker | disease)
2. No disease = $\frac{C}{B+D}$ = rate(smoker | no disease)

Proving Association. In general, if A, B are population characteristics with

$$0 < \text{rate}(A) < 1 \text{ and } 0 < \text{rate}(B) < 1$$

If rate(A | B) > rate(A | not B) or rate(B | A) > rate(B | not A), then we can say that A and B are positively associated.

If rate(A | B) < rate(A | not B) or rate(B | A) < rate(B | not A), then we can say that A and B are negatively associated.

If rate(A | B) = rate(A | not B) or rate(B | A) = rate(B | not A), then we can say that A and B are not associated. (This rarely occurs)

Confounder. A confounder is associated with *both* exposure and disease. The actual relationship between exposure and disease can be obscure by confounders. In an observational study, it is important to control for confounders.

Controlling for Confounders. Slicing is the technique to control for a confounder. Breaks the data into smaller more homogeneous groups to reduce the effect of the confounders. Another statistical method is linear regression.

Chapter 2: Association

Deterministic Relationship. The value of a variable can be determined if we know the value of the other variable.

Statistical Relationship. Natural variability exists in measurement/outcomes of two variables. The *average* pattern of one variable can be described given the value of the other variable.

Bivariate Data. Data that has *two* variables.

Standard Deviation(sd). Used to indicate the spread around the average.

Scatter Diagram. Used to visually observe the relationship between two variables. This includes the *nature* of the relationship (e.g., linear), *direction* of the relationship and *strength* of the relationship.

Correlation Coefficient (r). A measure of *linear* association between two variable**s.** Ranges between -1 and 1. It summarises the direction and strength of *linear* relationship.

Sign of Correlation Coefficient. The direction of the linear association can be deduced from the sign of the correlation coefficient.

(a) If r > 0, then there is a positive *linear* association.
(b) If r < 0, then there is a negative *linear* association.
(c) If r = 0, then there is no *linear* association.

Magnitude of Correlation Coefficient. The strength of the *linear* association can be deduced from the magnitude of the correlation coefficient.

(a) If r is 1 or -1, there is a perfect *linear* association.
(b) If 0.7< | r | < 1, there is a strong *linear* association.
(c) If 0.3< | r | < 0.7, there is a moderate *linear* association.
(d) If | r | < 0.3, there is a weak *linear* association.

Importance of Drawing a Scatter Diagram. The scatter diagram is necessary to help us visualise some details that may be missed out from a simple measure of correlation coefficient.

Computing Correlation Coefficient. The steps for computing the correlation coefficient are as follows.

1. Convert each variable to *standard unit (SU).* $\text{SU} = \frac{\text{X}-\bar{\text{X}}}{\sigma_{\text{x}}}$
   a. The standard unit tells us the number of standard deviations a value is from the mean.
   b. A positive SU means that an observation is above average, vice versa.
2. Take the product of the standard units for each pair of observation.
3. r = average of all the products.

In general, $r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\text{X}-\bar{\text{X}}}{\sigma_{\text{X}}}\right)\left(\frac{\text{Y}-\bar{\text{Y}}}{\sigma_{\text{Y}}}\right)$.

Features of Correlation Coefficient.

1. r is a pure number without units.
2. r will not be affected by
   a. *Interchange* of the two variables.
   b. *Adding* a number to all values of a variable
   c. *Multiplying* a positive number to all values of a variable

Correlation Coefficient Limitation: Causality. Correlation *does not* imply causation. Strong correlation merely indicates a strong evidence of linear association between two variables.

Correlation Coefficient Limitation: Outliers in the Data Set. An outlier is a data point that is "*unusually*" far away from the bulk of the data. Outliers can inflate or deflate the *r* value. Correlation coefficient can be very sensitive to outliers.

Correlation Coefficient Limitation: Non-linear Association. Correlation coefficient only gives evidence to linear association and does not give evidence to non-linear association.

Excluding Outliers. It is dangerous to exclude outliers from an analysis for the sake of creating an association without understanding the causes of the occurrence. Outliers must be handled with care as they may contain important information.

Ecological Correlation. Correlation computed based on aggregate data, such as group averages or rates. Useful as aggregated data is usually easier to obtain than data on individuals. Association is "***overstated***" based on aggregate data as compared to individual data.

Ecological Fallacy. Deduce the inferences on *correlation* about individuals based on aggregate data.

Atomistic Fallacy. Generalise the *correlation* base on individuals towards the aggregate-level correlation.

Attenuation Effect. Caused by *range restrictions* where a bivariate data set are formed base on certain criteria on one variable resulting in the data for the other variable to be only available for a limited range. Attenuation effect is when range restriction causes ***decrease*** in magnitude of the correlation coefficient.

Removal of Some Data. Important information is lost during exclusion of data. This may cause the result of analysis to be skewed.

Regression Line. Depends on the independent(predictor) and dependent(predicted) variable. Can be expressed by the equations $Y = a + bX$ where *a* is the intercept and *b* is the slope.

Least Square Method. Line is determined so that the overall distance from each point to the line is *minimised*.

Predicting Using the Regression Line. Given an *X* value, we can use the regression line to predict the Y's average value. The X value given must be within the range of the data used to generate the regression line (no extrapolation allowed).

Chapter 3: Sampling

Unit. Elements from which measurements are to be taken from.

Population. Collection of all *units*.

Sample. A portion of the population that are selected to be in the study.

Census. Measurement would be taken from every unit in the population.

Advantages of a Sample Over a Census. In a lot of scenarios, there are advantages to taking a sample over taking a census.

1. When a census is not possible (E.g., blood test)
2. Speed
3. Cost
4. Accuracy

Purpose of Using a Sample. Results from the sample can be extended to the population from which the sample was drawn.

A Good Sample.

1. *Every* unit in the population has a possibility of being selected.
2. Selection process is not biased.

Basic Steps of Taking a Good Sample.

1. Sampling Frame
2. Probability Sampling
3. 100% response rate

Sampling Frame. Sampling frame is a list of sampling units intended to identify all units in the population. The simplest sampling frame consists of a list of units in the population.

Characteristics of a Good Frame. A good sampling frame covers exactly or bigger than the population so that every unit in the population has a chance of being selected. It also needs to be up-to-date and complete.

Probability Sampling Plan. Every unit in the population must have a known probability of being selected into the sample.

Simple Random Sampling (SRS). Every possible sample of the same size has the same chance of being selected.

How to Select a Simple Random Sample.

1. Assign a number from 1 to N to each sampling unit in the sampling frame, where N is the number of sampling units.
2. Pick a random number from 1 to N.
3. The sampling unit that has been assigned the number chosen will be selected into the sample.

Systematic Sampling. Selects units from a list through the application of a selection interval, *K*, so that every K^{th} unit on the list, following a random start, is included in the sample.
It can be treated as an SRS when the sampling units are arranged *randomly*. However, when variable of interest in the arrangement of the sampling units and the number K have the same cyclical effect, it will give us an undesirable sample.

Stratified Sampling Plan. We first divide the population of units in groups (strata) and then we take the probability sample from each group. Used when population units fall into natural groupings. This may have accuracy and cost benefits.

Multistage Sampling Plan. In a lot of studies, it might take several stages of selection before reaching the population units. At each stage, probability sampling plan is implemented. If everyone in the cluster is selected, it is called *cluster sampling*. If a probability sampling plan is used on the cluster, it is called *two-stage sampling*.

Difficulties in Sampling.

- Imperfect Sampling Frame
  - A sampling frame that includes unwanted units or excludes desired units.
- Non-response
  - Not all selected units are contactable.
  - Not all selected units are willing to take part in the study.
  - Non-response distorts the results of the study as non-respondents differ from respondents.
  - Incentives may be used to convince the selected sampling units to participate.

Disasters in Sampling.

- Using a volunteer or self-selected sample.
  - Typically biased since respondents are usually people with strong view on the issue.
- Using a convenience or haphazard sample.
  - The information we can gather from respondents who are easily available is normally different from the hard ones to get.
- Using a judgement sample.
- Selecting a quota sample
  - The process of selection in which elements are chosen using prearranged categories of sample elements to obtain a predetermined number of cases in each category.

Parameter. A numerical fact about the population. We can use a sample to give us an *estimate* about the parameter.

Estimation Equation. The estimation equation is a model to help us understand why it is common to see that our sample's estimate differs from the population's parameter.

$$Estimate = Parameter + Bias + Random\ Error$$

Random error can be quantified and detected easily if we know the sample size and standard deviation of the variable we are measuring. Bias is difficult to quantify and detect.

Selection Bias. The systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample. Caused by imperfect sampling frame or non-probability sampling methods.

Non-response Bias. Systematic tendency from subjects who do not respond to the survey or questionnaire. Caused by differences between non-respondents and respondents. In this module, we consider a response rate of less than 70% to be low.

Random Error. Error in the sample estimate due to sampling variability.

Sample Size and Random Error. When estimating a population's rate or average, a sample with larger size is likely to have smaller random error.

Confidence Interval. A range of values that we are reasonably certain our unknown parameter lies in. We can report the confidence interval by stating "We are 95% confident that the range from 0.18 and 0.22 contains the population parameter".

Interpretation of Confidence Interval. With repeated sampling, *about* 95% of the samples collected will have the population parameter lie within each respective sample's confidence interval.

Sample Size and Confidence Interval. When estimating a population's rate or average, a sample with a larger size is like to have a smaller confidence interval range.

Chapter 4: More on Observational Studies

Risk. The risk of an uncertain outcome is defined as its rate in a population. It is a number between 0 and 1.

Risk Ratio (RR). Risk ratio is the ratio of the risk of an outcome in an exposed group to the risk of an outcome in an exposed group. This is also known as *relative risk*. It is a useful measure of *association*.

Odds. $odds(A) = \frac{A}{A'} = \frac{risk}{1-risk}$. Odds is always larger than risk. If risk is very small, then $odds \approx risk$

RR Example.

	Diabetic (D)	Healthy	Row Total
Female (F)	72,000	140,000	216,000
Male (M)	52,000	156,000	208,000
Column Total	124,000	300,000	424,000

$$risk(diabetes) = \frac{124{,}000}{424{,}000} \approx 0.29 = 29\%$$

$$risk(D|F) = \frac{72{,}000}{216{,}000} \approx 0.33 > 0.25 = \frac{52{,}000}{208{,}000} = risk(D|M)$$

$$Risk\ ratio = \frac{risk(D|F)}{risk(D|M)} \approx \frac{0.33}{0.25} \approx 1.33$$

$$Odds(D|F) = \frac{72{,}000}{144{,}000} = 0.5$$

$$Odds(D|M) = \frac{52{,}000}{156{,}000} \approx 0.33$$

$$Odds\ ratio\ between\ females\ and\ males = \frac{Odds(D|F)}{Odds(D|M)} = \frac{0.5}{0.33} \approx 1.5$$

Interpreting Odds Ratio.

1. $OR = 1$. No difference in disease risk between the two groups ($RR = 1$)
2. $OR > 1$. Higher risk in the first group. $RR > 1$
3. $OR < 1$. Lower risk in the first group. $RR < 1$

Cohort Study, Case-Controlled Study Summary

	Cohort Study	Case-Controlled Study
Collection Methodologies	Takes a simple random sample separately from the exposure groups	Takes a simple random sample separately from the disease groups.
Uses		Good for rare diseases.
Population Risk?	Yes	No
Population Risk Ratio?	Yes	No
Population Odds	Yes	No
Population Odds Ratio	Yes	Yes

* This hold even if fraction sampled are different among the exposure groups.

Cross-Product-Ratio. The cross product will result in the population odds ratio for a cohort study, case controlled study and even the entire population.

	Diabetic	Healthy
Female	364	142
Male	256	158

$$Odds\ Ratio\ between\ Females\ and\ Males = \frac{364 \times 158}{256 \times 142}$$

Chapter 5: Uncertainty

Random Circumstance. A random circumstance is one in which the outcome is uncertain. The outcome is not determined until we observe it.

Probability. Probability is a measure of how likely something will happen. It is a value between 0 and 1 assigned to an outcome of a random circumstance.

Relative Frequency Interpretation of Probability. Based on repeated observation on repeated observation of outcomes. Probability can either be computed exactly by making an assumption about the physical world related to the circumstance or it can be approximated by observing the proportion of times an outcome occurs *over the long run*.

Personal Probability Interpretation of Probability. Based on our own personal belief. Circumstances are often not repeatable, only apply to particular individuals and only happen once and will never happen again. Personal probability is the degree to which one believes the outcome will happen.

Equally Likely Outcomes. In general, for a random circumstance with exactly N outcomes, where each outcome is equally likely, the probability of each outcome is $\frac{1}{N}$.

Mutually Exclusive Outcomes. Two events are mutually exclusive if they cannot occur at the same time.

Independent Outcomes. If two events do not affect each other's chance of occurrence, the events are said to be independent of each other.

Probability Rules

Addition Rule	When two (or more) events are mutually exclusive, the probability of either one of these events occurring is the sum of their individual probabilities.
Complement Rule	The probability of an event not occurring is 1 minus the probability of the event occurring.
Multiplication Rule	Given two independent events, the probability of these events both occurring at the same time is the product of their individual probabilities.

Average Values. Assume that all outcomes A_i are mutually exclusive where $i = 1, 2, \dots, k$. Then the average value associated to the activity is

$$Average\ value = V(A_1) \times P(A_1) + \cdots + V(A_k) \times P(A_k)$$

Average values are also known as expected values in statistics. It gives an indicated measurement over the long run. It does not represent any typical value for any one occurrence of the activity.

P-values. The probability of obtaining an outcome equivalent or more extreme than the observed.

Interpreting the P-value.

P-value ***smaller*** than the level of statistical significance.	Do not reject the null hypothesis and we *cannot* conclude that the *alternative hypothesis* is true
P-value ***larger*** than the level of statistical significance.	Reject the null hypothesis and we can conclude that the *alternative hypothesis* is true.

Conditional Probability. Conditional probability involves two events. The uncertain event A and the certain event B that we use as additional information.

$$P(A|B) = \frac{P(A\ and\ B)}{P(B)}$$

Or equivalently

$$P(A\ and\ B) = P(B) \times P(A \mid B)$$

In general,

$$P(A \mid B) \neq P(B \mid A)$$

Conditional Probability and Independence. If events A and B are independent, then

$$P(A \mid B) = P(A)$$

Base Rate. Probability that an individual has the disease.

Sensitivity. Condition probability that an individual is corrected tested positive given that he has the disease.

$$P(positive \mid disease)$$

Specificity. Condition probability that an individual is corrected tested negative given that he does not have the disease.

$$P(negative \mid no\ disease)$$

	Test positive	Test negative
Have the disease	True positive	False negative
Do not have the disease	False positive	True negative

To Test or Not to Test

To Test	Not to Test
No alternative test	Alternative more reliable test
Test is inexpensive and more expensive 2nd test	Test is expensive
Good chance of successful treatment	Unreliable treatment

Notes

1. The key difference between controlled experiment and observational study is who gets to allocate the subjects.
2. Volunteers may not be representative of the population.
3. Even with randomisation, blinding is important.
4. Group sizes need not be equal.
5. Be careful about the Simpson's paradox.