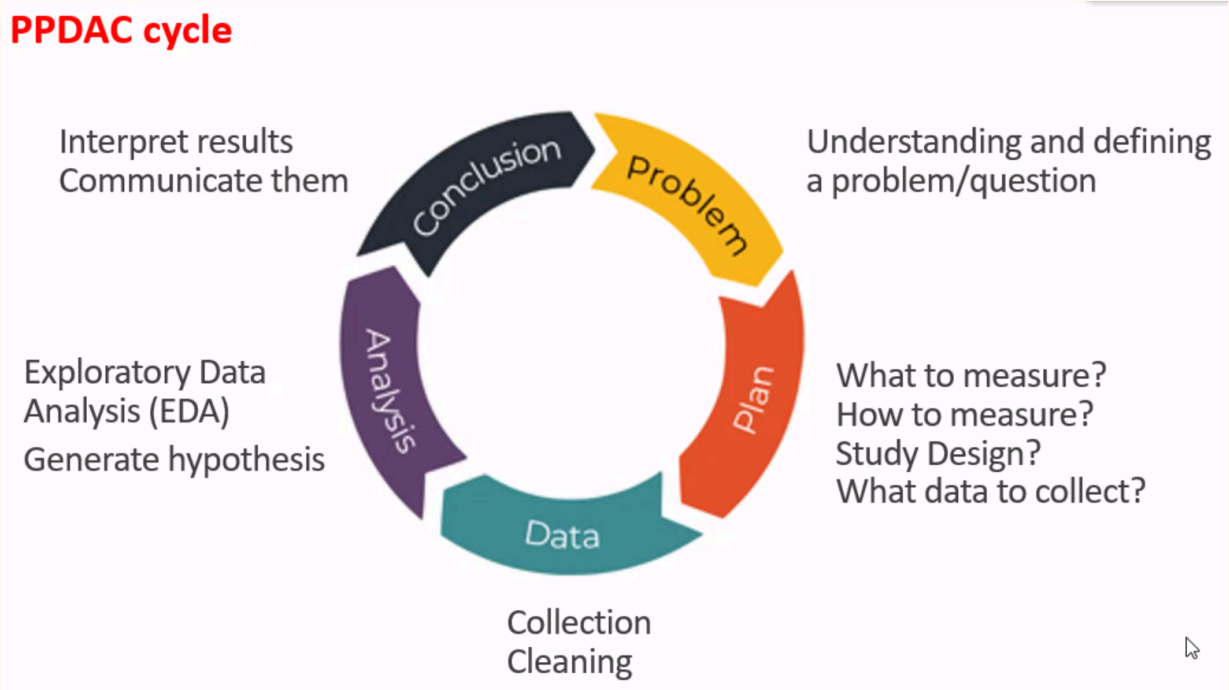
GEA1000

Chapters



Other Stuffs

Generally, a bar graph or a line graph can be used to show the general trend of the annual 'simple median' COE prices for Category C cars. Note that since a time variable is involved, a line graph is commonly preferred over a bar graph.

Chapter 1 - Sampling and Studies

Chapter 1.1 - Sampling

Exploratory Data Analysis (EDA)

1. Generate questions about data
2. Search for answers by visualising, transforming and modelling data
3. Use what is derived from the data to either refine our existing questions or generate new questions and delve deeper into the data
4. Can go in cycles.

Research Questions

Type of Research Questions	Examples
Make an estimate about the population	What is the average number of hours that students study each week?
	What proportion of all Singapore students is enrolled in a university?
Test a claim about the population	Is the average course load for a university student greater than 20 units?
	Does the majority of students qualify for student loans?
Compare two sub-populations	In university X, do female students have a higher GPA score than male students?
	Are student athletes more likely than non-athletes to do final year projects?
Investigate a relationship between two variables in the population	Is there a relationship between the average number of hours students spend each week on Facebook and their GPA?
	Does drinking coffee help students pass the math exam?

If the study is for a certain period of time/across certain categories, must state that also.

- How have the COE premium prices, across Categories A to E, changed for vehicle buyers from 2010 to 2022 in Singapore?

Sampling

Population vs Sample

- Population of Interest - A group in which researchers have interest in drawing conclusions of the study.
- Population Parameter - A numerical fact about a population, it's a constant.
- Sample - A proportion of the population selected in the study.
- Sample frame - is the list from which the sample was obtained.
- Estimate - An inference about the population's parameter, based on information obtained from a sample

Since going through the whole population for research isn't feasible (it is for certain cases, attempting to sample the whole population is called census), samples of the population are gathered and analysis is applied to it. With a good sample, the finding from the sample would be representative of the population, that is to say it gives a good enough estimate of the population and can be generalised to the whole population.

Sampling Frame

Sample frame is the list from which the sample was obtained, like how the population is reached.

- The sampling frame may not cover the population of interest.
- The sampling frame may include other data from outside the population of interest, as these can be filtered out before analysis. thus will not affect or bias study.
- A bad sampling frame is not one that covers more than necessarily, a bad frame is one that covers less than supposed to.
- So, for the sample and result to be a good representation of the population of interest, the sampling frame must be greater than or equal to the population of interest

Sample vs Census

- Census is when there is an attempt to gather data from the whole population. The data collected may not be of the whole population as for example people may not want to take part in the study but it is the attempt that makes it a census. While the sampling is just a proportion of the population.
- Taking samples usually done when census data is not available, it is cheaper as there is no need to collect so much data and make sense of them.
- Sampling is also faster, eg news agency may use sampling aapporch as they would want to get info out asap
- If a census has too few responses it will suffer non response bias also as those who responded may have certain characteristics.

Bias

Selection Bias	Non-response Bias
Associated with the researcher's biased selection of units	Associated with the participants' non-disclosure of information related to the study
• Imperfect Sampling Frame • Non-Probability Sampling	• Disinterested • Inconvenient • Unwilling to disclose sensitive information

Selection Bias

Selection bias associated with the researcher's biassed selection of units, individual could be unintentional or internally left out of the study because the sampling frame of the study excluded them, this means the result of the study would be biassed to those who are included in the sampling frame.

Imperfect sampling frame

- A good sampling frame would be bigger than or equal to the population of interest.
- For example, we are studying road satisfaction and using the number plate as a sampling frame for car owners. This will leave out drivers that drive cars belonging to other people, and thus the study would be biassed to those who are car owners.

Non-Probability Sampling

- This means that the sampling method does not involve the use of chance, i.e. not random, in the selection of individuals, so certain individuals of the population could be selected over others.

Non-Response Bias

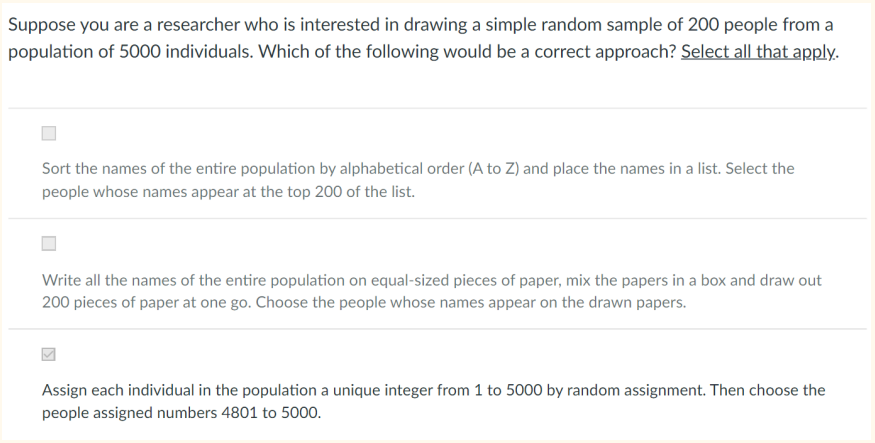
- Associated with the participants' non-disclosure of information related to the study, individuals who don't want to take part in the study.
- This non-disclosure may distort our understanding of the population parameter in the study.
- Non-response bias may occur regardless of whether the sampling method is probability or nonprobability in nature, if the response rate is too low.

Probability Sampling

- Sampling scheme such that the selection process is via a known randomised mechanism.
- It is important that every unit in the sampling frame has a known non-zero probability of being selected but the probability of being selected does not have to be the same for all the units.
- The randomised mechanism is important as it introduces an element of chance in the selection process so as to eliminate biases.

Simple Random Sampling (SRS)

Units are selected randomly without replacement from the sampling frame, each unit is assigned a number then n numbers 1 to size of sample frame is randomly generated (RNG).

- A SRS of size n consists of n units from the population chosen in such a way that every set of units has equal chance to be the sample actually selected.
- While it is expected that different samples sampled from the same sampling frame using SRS would be different, the variability between the samples is entirely due to chance.
- Advantage: Sample tends to be a good representation of population
- Disadvantage: Subject to non-response; accessibility of information (as RNG need a whole list of population to assign each a number and the whole list of population is hard to get)
- Examples:
    - number written on equally sized pieces of paper taken out of a well shaken box
    - **Example 1.2.11** Suppose we would like to sample 500 households in Singapore and find out how many household members there are in each household. Let us assume that every household has a unique home phone number. If we have a listing of all such phone numbers and list them from 1 to n, we can use a random number generator to select 500 phone numbers from the list to form our sample. Unique phone calls (i.e. sampling without replacement) can then be made to these households to survey the number of household members. This is another example of simple random sampling. Notice that this example also illustrates a common shortcoming of SRS, in that it can possibly be subjected to non-response from the units that are sampled.
        - Population of interest = households in singapore
        - Sample frame = households with home number (assumed to be the same as population in this case)
        - Sample = 500 households who picked up the phone
        - The study could be biassed to those who picked up the phone as people who didn't pick up the phone may be working and only have 1 member living in the house. Thus household member = 1 may be underrepresented
    - 

        - can take out a bunch all at once also

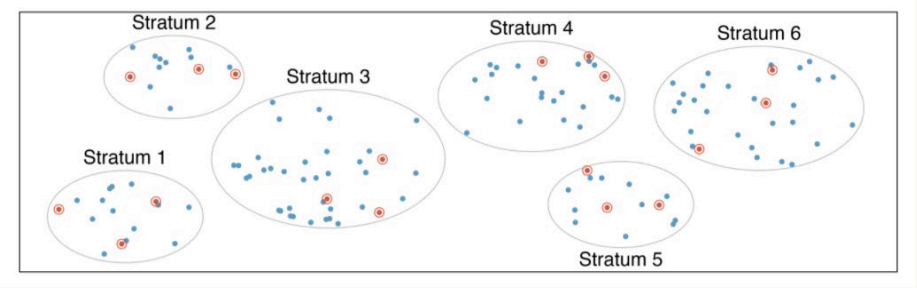
Systematic Sampling

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110

Method of selecting units from a list by applying a selection interval k and a random starting point from the first interval.

- Suppose we know how many sampling units there are in the population (denoted by n);
    - We decide how big we want our sample to be (denoted by k).
    - This means that we will select one unit from every n/k units;
    - from 1 to n/k , select a number at random, say r;
    - With this, the sample will consist of the following units from the list:
    - $$r,\quad r+\frac{n}{k},\quad r+\frac{2n}{k},\quad \cdots, r+\frac{(k-1)n}{k}$$
- However, it is often that we do not know the number of sampling units n in the population. In such a situation, systematic sampling can still be done by deciding on the selection interval k and randomly selecting a unit from the first k units and then subsequently every kth unit will be sampled. For example, if k = 10, we can sample the 5th, 15th, 25th units and so on.
- Advantage: Compared to simple random sampling, systematic sampling is a simpler sampling process as we do not need to know how many sampling units there are exactly.
- Disadvantage: if the listing is not random, but instead contains some inherent grouping or ordering of the units, then it is possible that a sample produced by systematic sampling may not be representative of the population.

Stratified Sampling

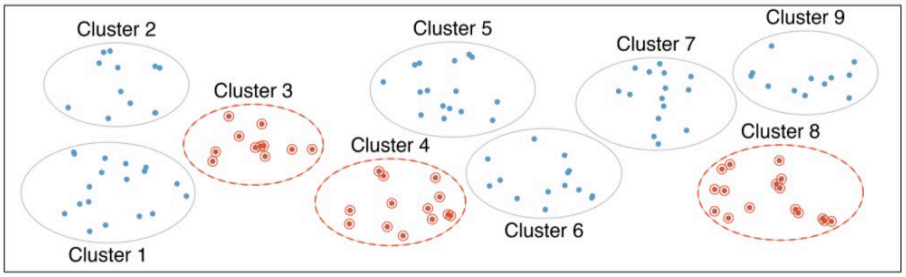


Method where the population is divided into groups called strata. Each stratum is similar in that they share similar characteristics but the size of each stratum does not necessarily have to be the same. We then apply simple random sampling to each stratum to generate the overall sample. Number of samples to be taken from each stratum using SRS could be such that the sample has the same proportion of strata as the population or could be the same across the strata but the final result needs to be taken from the weighted average of the result from each stratum. This is to ensure the sample taken is representative of the population.

- Advantage: able to get a representative sample from every stratum
- Disadvantage:
    - Need information about the sampling frame to form the strata, in some situations this is not possible.
    - Stratification may be difficult due to the non-binary nature of the characteristic so will have ambiguity during determine which stratum a particular unit belongs to
- Example:

Example 1.2.13 An example of stratified sampling can be seen during elections, for example, a Presidential Election. Voters visit their designated polling stations to cast their votes for the candidate that they wish to support. In countries where the number of voters is very large, it may take a long time before all the votes are counted. Stratified sampling can be employed if we wish to make a reasonably good prediction of the outcome. This is done by taking a simple random sample of the voters at each polling station (stratum) and then computing the *weighted average* of the overall vote count, based on the size of each stratum, for each candidate. This way, we would be able to have a reasonably good estimate of the total votes each candidate would receive.

Cluster Sampling



Method where the population is divided into clusters. A fixed number of clusters are then selected using simple random sampling. All the units from the selected clusters are then included in the overall sample. The clusters are usually naturally defined which makes it easy to determine which cluster a unit belongs to.

- Advantage: simpler, less costly and not as resource intensive than other probability sampling methods
- Disadvantage:
  - Depending on which clusters are selected, we may see high variability in the overall sample if there are largely dissimilar clusters with distinct characteristics. Clusters need to be reasonably heterogeneous with no cluster specific characteristic.
  - If the number of clusters sampled is small, there is also a risk that the clusters selected will not be representative of the population. Need a larger sample size to achieve a low margin of error.
- Example:

Example 1.2.14 Suppose a study wants to survey the mental wellness of Primary school students in Singapore. Cluster sampling can be done by treating each Primary school as a cluster and this way of clustering the population of interest is natural and unambiguous since all students in the population belongs to exactly one Primary school. A number of schools are then selected using simple random sampling for this survey and all the students in the selected schools will be part of the sample while those not in the selected schools will not be included. Another approach is of course to apply simple random sampling with the list of all students (from all Primary schools) as the sampling units. If this was done, then there is a possibility that all schools will have students forming part of the sample. Cluster sampling would not provide such a characteristic.

Stuff from Tutor

Here are the probabilities of selecting each unit for 2 probability sampling plans, just for your info.

For simple random sampling, each unit has equal probability of being selected. E.g., if we want to select 100 out of 1000 people to be in sample, then the probability is 100/1000 = 0.1 for each person.

For stratified sampling, suppose in a population of 10,000 people, there are 6000 Chinese, 2000 Malays, 1000 Indians and 1000 others. Let the different strata be the different ethnic groups. One way to choose your sample is to choose randomly 600 chinese, 200 Malays, 100 Indians and 100 others, so that the sample has same proportion of different ethnic groups as that of the population. Here the probability of selecting each unit is 0.1.

Another possible way to choose your sample is to choose 100 people randomly from each of the four groups. If so, then the probability of each Chinese being selected will be 100/6000, probability of a Malay being selected is 100/2000, and so on. Clearly the probabilities are not the same. But this is still probability sampling. Note however, for this case you will need to take a weighted average of the results for the 4 subgroups, since there are 60% Chinese, 20% Malays, 10% indians and 10% others in the population. That is, suppose the average results for the 4 subgroups are A, B, C and D respectively. Then the final result of the sample can be 0.6* A + 0.2* B + 0.1* C + 0.1* D, so that we can try to extend the result to the population. Hope this helps clarify why it is possible that probability of selecting each unit can be different.

Probability of to be chosen by SRS = Number of samples / Total amount
(All the different ways of being chosen, like ordering, is accounted for)

Summary

There is no single universally best probability sampling method as each has its advantages and disadvantages. All probability sampling methods can produce samples that are representative of the population (that is, the sample is unbiased). However, depending on the situation, some methods would further reduce the variability, resulting in a more precise sample.

Sampling Plan	Advantages	Disadvantages
Simple Random Sampling	Good representation of the population	Time-consuming; accessibility of information and sampling frame
Systematic Sampling	Simpler selection process as opposed to simple random sampling as number of sampling units can be unknown	Potentially under-representing the population if the list is ordered
Stratified Sampling	Good representation of the sample by stratum	Require sampling frame and criteria for classification of the population into stratum, this process may face ambiguity from non-binary nature of characteristic
Cluster Sampling	Less time-consuming and less costly	Require clusters to be reasonably heterogeneous and not have cluster-specific characteristics Need a larger sample size to achieve a low margin of error.

Non-Probability Sampling

The selection of individuals/units were not done by randomisation, but by human discretion.

Convenience Sampling

It is a non-probability sampling method in which the researcher uses the subjects that are most easily available to participate in the research study.

- Example: Mall surveys with population of interest = people living in that area

- Issue 1: Demographics of mall goers – teenagers, retired people, people who are more affluent (got disposable income). Other groups (non-teenagers and retirees, and the not so affluent) are left out. This is a good example of selection bias, the sampling frame is bad.
- Issue 2: Individuals asked to do the survey may not respond because they came to the mall to do shit not fill out stupid questionnaires. This could lead to non-response bias.

Volunteer Sampling

It is a non-probability sampling method in which the researcher actively seeks volunteers to participate in the study. Volunteer sampling happens when subjects volunteer themselves into a sample. Such a sample is also known as a self-selected sample and very often, the sample contains subjects who have a stronger opinion (either positive or negative) on the research question than the rest of the population. Such a sample is unlikely to be representative of the population of interest.

- Example 1: Online Polls
  - Those who did not respond are left out of the study. This presents to us the clear problem of non-response bias, in a volunteer sample.
- Example 2: Rating of show asked by host
  - a sample is unlikely to be representative of the population of interest. For example, the host of a "popular" radio talkshow may wish to find out how well received is his show. To do this, he asked his listeners to go online and submit a rating of this show, out of a score of 10. Each listener can voluntarily decide if they wish to be part of this rating exercise or not. By collecting a sample of opinions this way, it is likely that the sample will be skewed towards a high rating because listeners who did not like the talkshow would not even be aware of such a survey and therefore their opinions would have been left out. On the other hand, listeners who are strong supporters of this show would be more enthusiastic to go online to support their favourite radio show.

Summary for Sampling (asktutor)

let's say population of interest is a neighbourhood of 100 house, and studying satisfactory of neighbourhood environment:

- Simple random sampling = RNG house number to get 10 houses then knock their doors to ask for info
- Systematic sampling = simply choose starting point and determine interval then get the house number then knock their doors to ask for info
- Stratified sampling = sort houses into race/average age then for each strata do SRS then knock their doors
- Clusters sampling = sort houses into rows then select clusters using SRS then knock all the doors of the chosen rows
- Convenience sampling = wait in neighbourhood park and ask who ever walking in park
- Volunteer sampling = send out survey form and whoever reply will be included in the study

General Approach to Sampling

Sampling are used when census is not possible

1. Design a sampling frame.
   a. Recall that a sampling frame should ideally contain the population of interest so that every unit in the population has a chance to be sampled.
2. Decide on the most appropriate sampling method to generate a sample from the sampling frame.
   a. Probability sampling methods are generally preferred over non-probability sampling methods as non-probability sampling methods have a tendency to generate a biassed sample.
3. Remove unwanted units that aren't from the target population.

Generalisability Criteria

To ensure sample gathered is a good representation of the population of interest:

- Good sampling frame (>= population of interest) and probability based sampling is used to reduce selection bias
- Large sample size to reduce variability and amount of random error in the sample.
  - if sample size too small cannot generalise.
  - no "best" size, no certain "percentage" as for population that's not too variate small percentage is good enough.

- Minimise non-response rate
    - If non-response rate is high then generalizability to the population is difficult
    - At best only can generalise finding to people in the proportion of population of interest who would have responded to the study
    - Those who responded may have certain characteristic

Chapter 1.2 - Variable

A variable is an attribute that can be measured or labelled.

- A data set is a collection of individuals and variables pertaining to the individuals. Individuals can refer to either objects or people.
- In a research question where we are examining relationships between variables, there is usually a distinction between which are independent and which are dependent variables.

Research question	Dependent variable/Independent variable
Do NUS students who make notes using pen and paper score better in GEA1000 than those who use laptops?	Independent variable : Method of note taking for GEA1000
	Dependent variable : GEA1000 grade.
Does amount of caffeine consumed per day affect the quality of sleep amongst Singaporean adults?	Independent variable : Amount of caffeine consumed per day
	Dependent variable : Quality of sleep

Independent variables

- Variables that may be subjected to adjustments, either deliberately or spontaneously, in a study.

Dependent variables

- Variables that are hypothesised to change depending on how the independent variable is adjusted in the study.
- It is important to note that the dependent variable is hypothesised to change when the independent variable is adjusted. It does not mean that the dependent variable must change. It is perfectly possible that any changes to the independent variable does not result in any change in the dependent variable.

Categorical Variables

Variables that take on categories or label values. The categories or labels are mutually exclusive, meaning that an observation cannot be placed in two different categories or given two different labels at the same time.

- Among categorical variables, there are generally two subtypes.

Ordinal Variable

Categorical variable where there is some natural ordering and numbers can be used to represent the ordering.

- Example 1 : Year 1, 2, 3, 4 in University can be ordered but doing numerical operation on them means nothing
- Example 2 : Happiness index can be numbered 1-10 but the difference between 5 to 6 and 6 to 7 cannot be assumed to be the same so cannot calculate average or do arithmetic operation
    - The subjective nature of happiness means that one's 5 is different from another's 5.
    - but if need any comparison between different level of happiness then can calculate what percentage of respondents have responded in each of the different level (i dun get this i copy from lecture vid)

Nominal Variable

Categorical variable where there is no intrinsic ordering.

- Example 1: Blue eyes, brown eyes. Anything yes/no.
- Example 2: stuff that's in a cycle is nominal too, like quartiles of a year as Q1 Q2 Q3 Q4 can also be Q4 Q1 Q2 Q3 so no clear order.

Numerical Variables

Variables that take on numerical values and we are able to meaningfully perform arithmetic operations like adding and taking average.

- Among numerical variables, there are also generally two subtypes.

Discrete Numerical Variable

Variable where there are gaps in the set of possible numbers taken on by the variable.

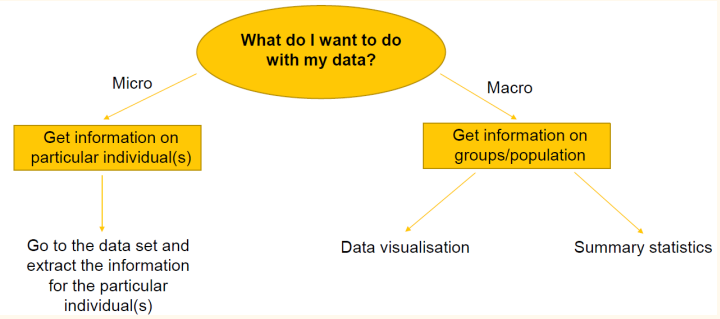
- Exampole: number of student passing a subject

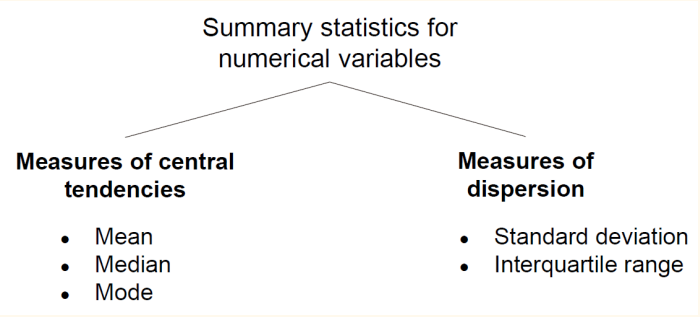
Continuous Numerical Variable

Variable that can take on all possible numerical values in a given range or interval.

- Examples: weight, height, length

Chapter 1.3 - Central Tendencies and Spread





- central tendency is the central or typical value for a probability distribution.

Mean

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Properties of mean

- $x_1 + x_2 + \cdots + x_n = n\overline{x}$
  - Possible calculate their sum of sample if we know their mean and the number of data points (n) that is used to compute the mean, without knowing the each samples
- $x_1 + c, \quad x_2 + c, \quad \ldots, \quad x_n + c$
  - Adding a constant value c (positive or negative) to all the data points changes the mean by that constant value.
- $cx_1, cx_2, \ldots, cx_n$
  - Multiplying a constant value of c (positive or negative) to all the data points will result in the mean being changed by the same factor of c.
- Mean does not show the distribution of sample, just that the mean cannot be the maximum or minimum value.
- Mean doesn't mean 50% more than mean and 50% less than mean.

Overall Mean of Groups

	Number of students	Average Mark
School A	349	32.21
School B	46	30.72
Overall	395	?

- Cannot just add the mean and divide by 2 because the number of samples from each school is different.
- Need to take the weighted average such that the weight of the school with more samples is higher.

- In general mean is bigger than the smallest and smaller than the biggest
- $\bar{x} = (n_a \bar{x_a} + n_b \bar{x_b}) / (n_a + n_b)$

Maen Disguised as Proportion

	New drug	Existing drug
Number of patients	500	1000
Total asthma attacks	200	300

Proportion of asthma attacks with new drug = 0.4
Proportion of asthma attacks with existing drug = 0.3

- New drugs seem to be doing worse but to conclusively say it's not better to know how the clinical trial was designed.
- Proportion can be thought of as a special case of mean if we let people who get asthma attacks be 1 and those who didn't get asthma attacks be 0.

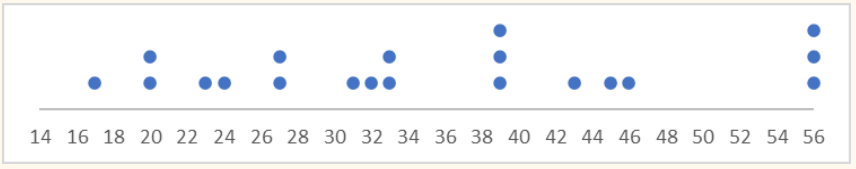
Standard Deviation

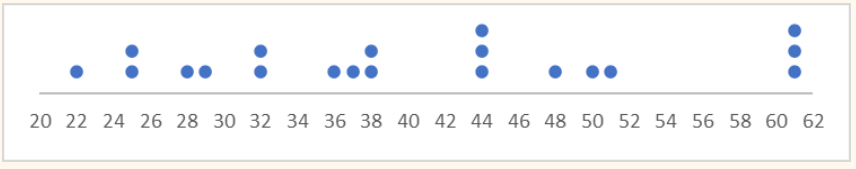
$$\text{Sample Variance, Var} = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n-1}; \quad \text{Standard Deviation, } s_x = \sqrt{\text{Var}}.$$

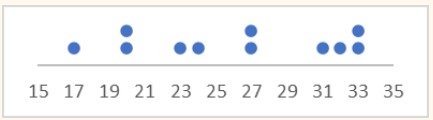
The standard deviation is one way of quantifying the "spread" of the data about the ***mean***. The formula is derived via the variance.

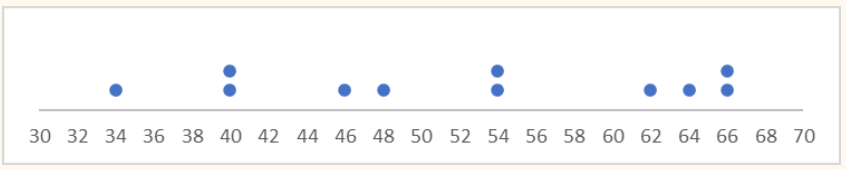
Remark 1.5.2 You may wonder why, in the computation of sample variance, we divide the sum of the squares $(x_i - \overline{x})^2$ by $n-1$ instead of n, since we have n data points and not $n-1$. The reason is because $x_1, x_2, \ldots, x_n$ are assumed to be a sample taken from a population. We are using the variance observed in such a sample to estimate the variance at the population level, which is usually unknown. You can think of dividing by $n-1$ instead of n as a 'correction' to make since our data is only a sample of the population. More detailed discussion on this is beyond the scope of this module.

Properties of Standard Deviation

- Always non-negative (>=0) and of the same unit as the samples and mean.
- Adding a constant value, c (positive or negative) to all the data points does not change the standard deviation. This just shift the data up/down the number line, the spread doesn't change
  - 

  - 

- Multiplying all the data-points by a constant value c (positive or negative) results in the standard deviation being multiplied by |c|
  - 

  - 

Coefficient of Variation

$$\text{coefficient of variation} = \frac{s_x}{\overline{x}}$$

- Coefficient of Variation is used to quantify the degree of spread relative to the mean.
- The coefficient of variation is a useful statistic for comparing the degree of variation across different variables within a data set, even if the means are drastically different from one another.

Example - Penguins

	Mean mass	Standard deviation of mass
Chinstrap	3733g	384.3g
Adelie	3710g	458.6g
Gentoo	5076g	504.1g
Overall	4201g	802.0g

- Observe that the overall mean mass 4201g is indeed between the group with the highest mean mass (Gentoo) at 5076g and the group with the lowest mean mass (Adelie) at 3710g. This is consistent with our earlier discussion.
- Even though the overall mean mass is 4201g with standard deviation 802g, it does not imply that the heaviest penguin weighs 4201 + 802 = 5003g.
- Suppose we wish to investigate whether the Adelie and Chinstrap species are similar in terms of their mass.
    - First, we observe that the mean mass of these two groups are rather similar with the Adelie species having a mean mass of 3710g while the Chinstrap species has a mean mass of 3733g.
    - However, the standard deviation of mass for these two species are rather different.
- To examine further on the difference in physical attributes between the Adelie and the Chinstrap species, we need to delve into other factors or variables that we have information on from the data set, for example, variables like age, gender, location and so on. This is Exploratory Data Analysis in action, where we start off with a few questions about the data set and with exploration into the data, we ask new questions and go back to the data set to look more closely at the data in an attempt to answer the new questions. In data analysis, this process is often repeated several times. In relation to this penguin data set, here are some further questions that can be asked:
    - Are male penguins heavier than female penguins?
    - Is there a relationship between bill length and bill depth across all species?
    - Do heavier penguins come from colder locations?
    - Can findings in this data be generalised to all of the three species?

Median

The median of a numerical variable in a data set is the middle value of the variable after arranging the values of the data set in ascending or descending order. If there are two middle values (when there are an even number of data points), we will take the average of the two middle values as the median.

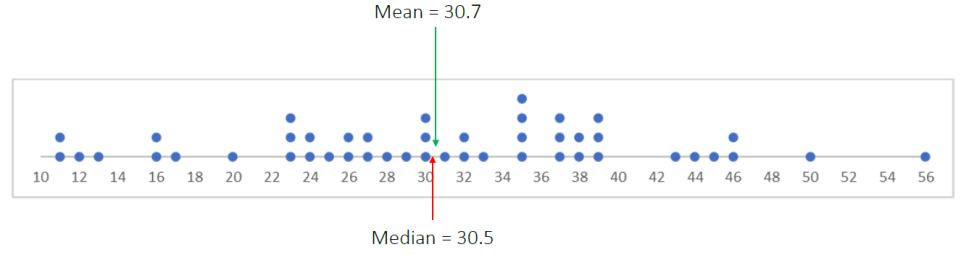
Properties of Median

- Adding a constant value c (positive or negative) to all the data points changes the median by that constant value.
- Multiplying a constant value of c (positive or negative) to all the data points will result in the median being changed by the same factor of c.
- Median does not show the distribution of the sample, just that the median cannot be the maximum or minimum value.
- Median means 50% of the sample is more than it and 50% less than it.

	Median score	Mean score
School A	32	32.21
School B	30.5	30.72
Combine schools A and B	32	32.04

-
    - Similar to what we observed for means, the overall median score (32) lies between the subgroup with the higher median (32) and subgroup with the lowest median (30.5)
    - However, if we know each of the subgroup medians, it is not possible to use this information to derive the overall median, unlike the case for mean.

Example



The mean score for school B was 30.72, which is very close to the median score. The main reason for this is because the spread of the scores are quite symmetrical about the mean and the median.

Interquartile range

- We have seen that the median represents a numerical value where 50% of the data is less than or equal to this value. This is also known as the 50th percentile of the data values.
- The first quartile, denoted by Q1, is the 25th percentile of the data values, while the third quartile, denoted by Q3 is the 75th percentile of the data values.
- This means that 25% of the data is less than or equal to Q1 while 75% of the data is less than or equal to Q3.
- The interquartile range, denoted by IQR is the difference between the third and first quartiles, so
    - IQR = Q3 − Q1.

Properties of IQR

- IQR and standard deviation share similar properties. For example, we know that IQR is always non-negative since Q3 is always at least as large as Q1 and so Q3 − Q1 ≥ 0.
- If we add a constant c (positive or negative) to all the data points, not only does the median value increase by c, Q1 and Q3 are increased by c as well. Thus, there will be no change in IQR.
- If we multiply all data points by a constant c, then IQR will be multiplied by |c|.

Getting Q1 and Q3

- If there are even number of data:
    - Median = average of the middle two value
    - Q1 = median of subset of data less than median
    - Q3 = median of subset of data more than median
    - IQR = Q3 - Q1
- If there are odd number of data:
    - Median = middle value
    - Q1 = median of subset of data less than median excluding the median
    - Q3 = median of subset of data more than median excluding the median
    - IQR = Q3 - Q1

Example 1: even number data

Example 1.6.8 Let us consider two simple data sets and compute the first quartile, median, third quartile and interquartile range. The first data set consists of an even number of data points as follows:

$$16,\ 30,\ 5,\ 1,\ 9,\ 22,\ 19,\ 8,\ 10,\ 28.$$

We arrange these 10 data points in increasing order:

$$1,\ 5,\ 8,\ 9,\ 10,\ 16,\ 19,\ 22,\ 28,\ 30.$$

1. Since there are 10 data points, the median is the average of the 5^{th} and 6^{th} ranked data points, so median is $\frac{1}{2}(10+16) = 13$.
2. To find the first and third quartiles, we divide the data set into the lower half (1^{st} to 5^{th} ranked data points) and upper half (6^{th} to 10^{th} ranked data points). The first quartile is the median of the lower half

   $$1,\ 5,\ 8,\ 9,\ 10,$$

   which is the 3^{rd} ranked data point in this lower half, so $Q_1 = 8$. The third quartile is the median of the upper half

   $$16,\ 19,\ 22,\ 28,\ 30,$$

   which is the 3^{rd} ranked data point in this upper half, so $Q_3 = 22$.

Example 2: odd number data

Let us consider the second data set which consists of an odd number of data points as follows:

$$5.6,\ 1.5,\ 3.3,\ 8.7,\ -3.1,\ 9.2,\ 15.5,\ 2.6,\ 11.5.$$

We arrange these 9 data points in increasing order:

$$-3.1,\ 1.5,\ 2.6,\ 3.3,\ 5.6,\ 8.7,\ 9.2,\ 11.5,\ 15.5.$$

1. Since there are 9 data points, the median is the 5^{th} ranked data point, so median is 5.6.
2. To find the first and third quartiles, we divide the data set into the lower half (1^{st} to 4^{th} ranked data points) and the upper half (6^{th} to 9^{th} ranked data points). Note that we have **not** included the median in **both** lower and upper halves. The first quartile is the median of the lower half

   $$-3.1,\ 1.5,\ 2.6,\ 3.3,$$

   which is the average of 1.5 and 2.6, so $Q_1 = 2.05$. The third quartile is the median of the upper half

   $$8.7,\ 9.2,\ 11.5,\ 15.5,$$

   which is the average of 9.2 and 11.5. So $Q_3 = 10.35$.
3. The interquartile range is $Q_3 - Q_1 = 10.35 - 2.05 = 8.3$.

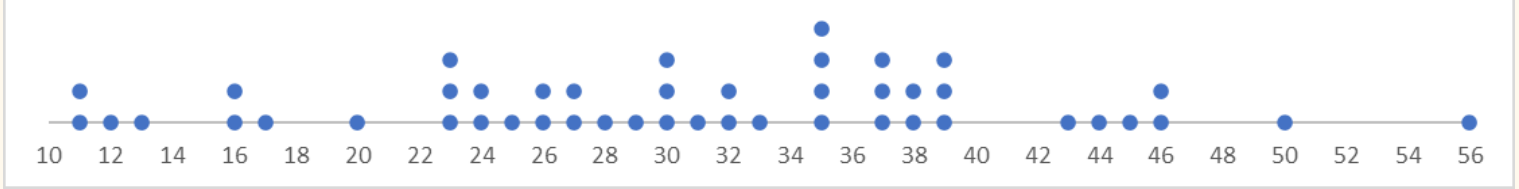
Remarks

1. In the example above, when the data set has an odd number of data points, we have not included the median in both the lower and upper halves. This is not a universal practice. You may encounter some texts that include the median in both halves.
2. In reality, when the number of data points is large, summary statistics like median and quartiles are not computed manually but instead, they are computed using softwares. However, even softwares do not adopt the same algorithm in computing these statistics. The good news is that we do not have to worry too much about finding the exact value of the quartile since for large data sets, all the different methods give pretty close answers and the small difference is not an issue. For small data sets, it is also not really meaningful to summarise the data since we have complete information of the entire data set anyway.

Mode

The mode of a numerical variable is the numerical value that appears most often in the data. For categorical data, a mode is the category that has the highest occurrence in the data.

- Mode is the only summary statistic that can be used for both numerical and categorical variables



- When we are describing the distribution of points of a discrete variable, the mode can be interpreted as a “peak” of the distribution.

- In the context of probability, a peak of the distribution, refers to the value that has the highest probability of occurring.

Mean & Standard Deviation VS Median & IQR

For a numerical variable, we can always use the mean and standard deviation as a pair of summary statistics to describe the central tendency as well as the dispersion and spread of the data. Similarly, the median and IQR can also be used. There is no clear cut answer but very often, the choice depends on the distribution of the data. Generally speaking, the median and IQR is preferred if the distribution of the data is not symmetrical or when there are outliers.

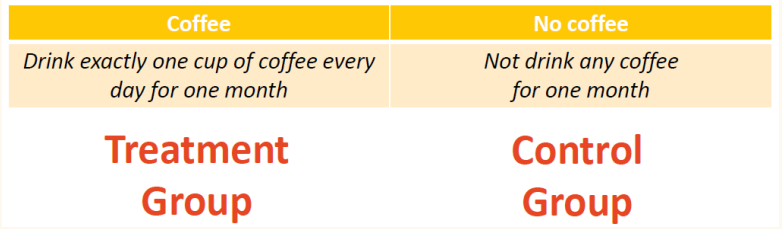
Chapter 1.4 - Designing Studies

Type of Research Questions	Examples
Make an estimate about the population	What is the average number of hours that students study each week?
	What proportion of all Singapore students is enrolled in a university?
Test a claim about the population	Is the average course load for a university student greater than 20 units?
	Does the majority of students qualify for student loans?
Compare two sub-populations	In university X, do female students have a higher GPA score than male students?
	Are student athletes more likely than non-athletes to do final year projects?
Investigate a relationship between two variables in the population	Is there a relationship between the average number of hours students spend each week on Facebook and their GPA?
	Does drinking coffee help students pass the math exam?

This chapter focuses on the last type of question, to compare two subpopulations / to investigate a relationship between two variables in the population.

Experimental Study

In an experimental study (sometimes also known as a controlled experiment or simply an experiment), we intentionally manipulate one variable (the independent variable) to observe whether it has an effect on another variable (the dependent variable). The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.

- Example: Does drinking coffee helps student pass mathematics exams
    - 

    - We can set up an experimental study by dividing the subjects, that is, the students taking the examination, into two groups.
    - The first group will be required to drink exactly one cup of coffee every day for a month.
        - This group is known as the treatment group since they are thought to be put through the “treatment” of drinking coffee.
    - The second group will not drink any coffee for one month.
        - This group who does not drink coffee is known as the control group.
        - Control group could receive some other standard treatment also.
    - The control group provides a baseline for comparison with the treatment group, a reference.
- Let say one group does better than the other it could be by luck that:
    1. One group is better at mathematics
    2. One group had more time to study
    3. Some other factor affect the experiment
- To remove all the other factors affecting the experiment, i.e. making sure the independent variable is the only factor that impacts the dependent variable, we use a statistical method known as random assignment.

Random Assignment

Random assignment is an impartial procedure that uses chance (or probability) to allocate subjects into treatment and control groups.

- One way to do this is put all paper with names in box and take out one by one
- By doing random assignment, the other factors can cancel out each other leading to similar characteristics for the groups.

Criteria for random assignment to work

- Both groups must be large but no necessarily the same size
    - If the number of subjects is large, by the law of probability, the subjects in the treatment and control groups will tend to be similar in all aspects.
- When we use the term “random” in random assignment, we do not mean that the assignment is haphazard. The term random in this case is used in relation to the use of an impartial chance mechanism to assign the subjects into two (or more) groups.
    - if for example the paper draw method is used and some papers are bigger than others, thus more likely to be drawn it'll be haphazard and not random.

Placebo

There is another important consideration when it comes to designing a controlled experiment. If we make it known to the control group that they are indeed the control group, and therefore not going to receive any form of treatment, this could possibly lead to bias.

- If the subjects in the control group are told that they will not be assigned any coffee for a month, when we are testing if coffee helps a student pass the mathematics examination, students in the control group may feel disadvantaged and therefore lack confidence and motivation to study. This may in turn result in these students not doing well in the examination and performing poorer than their friends in the treatment group who were given coffee. Any observed difference in passing rate between the two groups of students may not be the result of coffee at all. If this happens, the effect of coffee may be overstated.
- On the other hand, to the students in the control group, knowing that they will not be given coffee may actually cause them to take certain measures for their own benefit of passing the examination. For example, they may study harder and spend more time on their revision which may then result in the control group performing better than the treatment group in passing the examination. Again, any observed difference in passing rates between the two groups of students may not be the result of coffee at all. If this happens, the effect of coffee may be understated.
- One way to reduce this effect on the control group which could influence the study on the effects of coffee drinking is to give the subjects in the control group another beverage which tastes and smells the same as coffee but is without the active ingredients in coffee that is believed to improve one’s cognitive ability. This is known as a placebo.
    - A placebo is an inactive substance or other intervention that looks the same as, and is given the same way as, an active drug or treatment being tested. In the context of an experiment, a placebo is something given to the control group that in actual fact, has no effect on the subjects in the group.
    - However, it has been observed that in some instances, subjects in the control group upon receiving the placebo still showed some positive effects which is likely caused by the psychology of believing that they are actually being “treated”. This is known as the placebo effect.
    - Receiving something that does nothing may do something nonetheless

Single Blinding

One way to prevent the placebo effect from interfering with our experiment and observation on the benefits (if any) of the treatment is to blind the subjects involved in the experiment. By blinding the subjects, we mean that they do not know whether they belong to the treatment or control group. To do this, a placebo that is “similar” to the treatment is given to the control group so that the two treatments appear identical to the subjects. As a result, subjects do not know which group they belong to. If we can do this, we would have achieved single blinding.

Double Blinding

To take blinding one step further, other than blinding the subjects, it may be necessary to consider blinding the researchers conducting the study as well, especially if measuring the effects of the treatment may involve subjective assessments of the subjects. This is because the assessors may subconsciously believe that the treatment is

effective and this could introduce bias in the outcome. Thus, we should also blind the assessors so that they do not know whether they are assessing the treatment or the control group. We would have achieved double blinding if subjects and assessors are blinded about the assignment.

- For example, in the coffee experiment, if the assessors marking the students' answers are aware of which group each student belongs to, they may be inclined to award higher marks/ more lenient to students in the treatment group than those in the control group.

Note: sometimes it may not be possible to blind both the subjects and the assessors but when done right, double blinding can be very effective in reducing bias in the outcome of the experiment.

Observational Study

- An observational study observes individuals and measures the variables of interest, usually without any direct/deliberate manipulation of the variables by the researchers.
- Observational studies are alternatives to experiments that can be used when we are faced with ethical issues in experiments.
- As researchers usually do not attempt to directly manipulate or change one variable to cause an effect in another variable, observational studies do not provide convincing evidence of a cause-and-effect relationship between two variables.
- Example 1:
  - **Discussion 1.7.10** Besides an experimental study, another study design is an observational study. Consider the following research question: Does vaccination help reduce the effects of the coronavirus?

    If we were to design a controlled experiment, would the following be a possible and reasonable approach?

    - Enrol a group of participants into the study and inject all the participants with low dosages of the virus strain.
    - Perform random assignment to divide the group of subjects into the treatment group and control group.
    - Inject the treatment group with the vaccine and inject a harmless liquid (similar in colour, smell etc to the vaccine) into the control group, without revealing what they are being injected with.
    - Observe the number of participants in each group who develop symptoms similar to a coronavirus patient.

  - You probably realise by now that it is not so straightforward to design a controlled experiment like this. There are obvious *ethical issues* that need to be addressed. Some immediate questions that need to be answered are

    1. Should we inject such a virus into humans in the first place?
    2. How should we decide who is to be assigned to the treatment group and who is to be assigned to the control group?
    3. Is it fair not to let the subjects know if they are injected with the vaccine or with a placebo? Should we obtain consent from the subjects at the beginning of the study?

    Experiments can give us useful evidence for a cause-and-effect relationship. However, not all research questions are suitable to be investigated using an experiment, sometimes due to ethical issues like those listed above. Therefore, we need to consider the pros, cons and feasibility of an experimental study before deciding if we should proceed.
- Example 2:

Example 1.7.13 We would like to investigate whether exercising regularly (defined as exercising at least 3 times a week, at least 30 minutes of strenuous exercise each time) is associated[2] with having a healthy body mass index (BMI) (defined as between 18.5 to 22.9 kg/m^2) for Singaporean men between the ages of 30 to 40 years old.

Participants were recruited into the study and by their own declaration, they were classified into either the "treatment" group (those who exercise regularly) or the "control group" (those who do not). Participants were then told to proceed with their usual lifestyle habits and their body mass index were measured after 3 months. The following table summarises the findings at the end of the study.

	Treatment (Exercise regularly)	Control (Do not exercise regularly)
Healthy BMI range	320	127
Outside Healthy BMI range	101	191

This is an example of an observational study. Do you think there is sufficient evidence of association between exercising regularly and having a healthy BMI? We will discuss more questions like this in subsequent chapters.

- Example 3: smoking and heart disease, can't force people to smoke their entire lives without facing ethical issues.

Experimental VS Observational Study

- Not all research questions can be studied practically using an experiment. For example, if we would like to investigate if long term smoking is linked to heart disease, it is extremely difficult to design an experiment and put subjects into the treatment group where they will be required to smoke for the long term, even if this is against their will. This is challenging and unethical. An observational study may be more suitable for such an investigation.
- For observational studies, there is no actual treatment being assigned to the subjects but we normally still use the term treatment and control in the same way as though we are dealing with an experiment. For the investigation on smoking and heart disease, smokers who are observed to be smoking over a long period of time will be in the treatment group while non-smokers are in the control group. Sometimes, we may use the term exposure group instead of treatment group and non-exposure group instead of control group.
- For experimental studies, subjects are assigned into either the treatment or control group by the researcher. For observational studies, subjects assign themselves into either the treatment or control group.
- Observational studies cannot provide evidence of cause-and-effect relationships. On the other hand, experimental studies can provide such evidence if it has the features of randomised assignment and blinding (preferably double blinding).
- If an experiment is well-designed, can the conclusion of the experiment based on a sample be generalised to the population from which the sample was drawn?
    - Having a good design is not the only important piece of the puzzle.
    - In order to generalise the results from a sample to a bigger population, there are other factors that are equally important, for example, the sampling frame, sampling method, sample size and response rate.

	Experiments	Observational Studies
Assignment	By researchers	Participants assign themselves
Randomisation	Preferable	Not possible
Ethical issues	Possible *(if intervention may be harmful)*	Unlikely
Confounders	Unlikely *(if randomisation is done on enough participants)*	Usually exist many
Possible to show causation	Yes (in the ideal case)	Very difficult
Able to show association	Yes	

Quiz 1 Stuff

Professor Lim would like to find out if including a peer review component would affect students' final grades. He decided to get a sample of the students in his tutorial classes and place them into 2 groups. He assigned his Monday, Tuesday and Wednesday morning classes into the 'assessment with peer review' group and his Wednesday afternoon, Thursday and Friday classes in the 'assessment without peer review' group.

Which of the following best describes the type of sampling employed?

- None of the other options.
- Systematic sampling.
- Volunteer sampling.
- Cluster sampling.

Answer = none of the other options, because never say about sampling?
Probability sampling will require deliberate use of chance in the sampling process. In this case, the assignment of individuals has been pre-determined by the Professor. Within the types of non-probability sampling methods, this is not an example of volunteer sampling, as all students from both sub-groups were selected by the Professor to do the study and not self-selected.

May, an owner of a tuition center, wishes to find out if using iPads during tuition class improves her students' academic performance. She decided to conduct an experiment as follows:

1. She groups all the students in her center according to the day they come for tuition. For simplicity's sake, we can assume each student only goes for tuition once per week, there is at least one class of tuition every day in her center, and no student drops out halfway.
2. Every student who goes for tuition on weekends will be given an iPad to use during class. The students who go for tuition on weekdays will not be given an iPad.
3. She then keeps track of all her students' academic performance for the next 6 months.

Which of the following statements is/are true?

(I) She used a probability sampling method.

(II) This is a controlled experiment without random assignment.

- Only (I).
- Only (II).
- Neither (I) nor (II).
- Both (I) and (II).

Statement (I) is incorrect. Probability is not used in the selection of students into treatment/control. In fact, a census, not sampling, is conducted in this case. Statement (II) is correct. The students who go for tuition on weekends will be in the treatment group, and those who go on weekdays will be in the control group. There is no random assignment involved here.

Suppose you are a researcher who is interested in drawing a simple random sample of 200 people from a population of 5000 individuals. Which of the following would be a correct approach? Select all that apply.

- ☐ Sort the names of the entire population by alphabetical order (A to Z) and place the names in a list. Select the people whose names appear at the top 200 of the list.
- ☐ Write all the names of the entire population on equal-sized pieces of paper, mix the papers in a box and draw out 200 pieces of paper at one go. Choose the people whose names appear on the drawn papers.
- ☑ Assign each individual in the population a unique integer from 1 to 5000 by random assignment. Then choose the people assigned numbers 4801 to 5000.

Can take a bunch all at once out also

A multiple-choice mid-term examination was conducted for 2000 students in a General Education module GEB1000. There were 20 questions. Students were awarded 1 mark for each correct answer and received 0 mark for any wrong answer. There was no partial credit awarded for all questions. A teaching assistant helped with the collation of the scores of the paper, and provided the following summary statistics:

- Minimum = 2.0
- 1^{st} Quartile = 7.5
- Median = 11.5
- Mean = 9.0
- Mode = 12.0
- 3^{rd} Quartile = 13.2
- Maximum = 20.0

Which of the following statements is/are true? Select all that apply.

- ☐ None of the other statements is true.
- ☐ The 3^{rd} Quartile is incorrect.
- ☐ Based on the above information, we can conclude that the coefficient of variation is 2.
- ☑ Based on the above information, we can conclude that the range is 18.0.

Q3 is wrong because the quartiles of integers cannot have .2, at most is .5

A researcher wants to know the average weight of year 1 students in University A. The researcher does not have access to such information, hence he decided to do a survey.

All University A year 1 students have to take a compulsory module in the first semester of their studies, and hence have to be present for an in-person examination on 24th April at 1pm. The researcher stood outside the examination venue's only exit with a weighing scale and waited for the examination to end.

There were too many students for the researcher to weigh. Hence, to decide whom to weigh, while students were exiting, the researcher used a random integer generator to produce a random integer for each student. If the random integer was even, the researcher will measure the student's weight. If the random integer was odd, the researcher will not measure the student's weight. Assume that all students exited the venue orderly in a line and were compliant with the researcher. There were 800 students exiting the venue, and the random integer generator produced 200 even numbers and thus 200 students were weighed.

What is an/are issue(s) that the study is likely to face? Select all that apply.

- [ ] Bias present due to the random integer generator producing unlikely random numbers.
- [ ] Bias present due to the non-probability sampling method.
- [ ] Bias present due to a poor sampling frame chosen.
- [x] None of the other options.

Each student had a ½ chance of being chosen, it is random sampling.

Question 7 1 / 1 pts

To study the overall satisfaction levels of students staying in a college, a researcher obtained a simple random sample of 100 students from the list of all students staying in the college and sent out a survey form to these 100 students. Which of the following must be true about the study? Select all that apply.

- [x] There will not be any selection bias in this study.
- [ ] The sample is representative of all the students staying in the college.
- [ ] There will not be any non-response bias in this study.
- [ ] The results obtained from the sample cannot be generalised to all the students staying in the college.

Must be true about the study
to be a good representation, must have large enough sample size, but since population variability and size is unknown, it can't be said that sample must be representative of the population

cannot conclude that the result must cannot be generalised to all student, as if sample size big enough and non response rate is low, then it is generalisable, thus not a must.

Chapter 2 - Categorical Variable Analysis

- Chapter 2.1 - Rates
- Chapter 2.2 - Association
- Chapter 2.3 - Rules on Rates
- Chapter 2.4 - Simpson's Paradox
- Chapter 2.5 - Confounders
- Quiz 2

Example:
Suppose a new patient with kidney stones were to be treated with treatment X or Y, based on past results choose the better treatment.

Size of stone	Gender of patient	Treatment type (X or Y)	Outcome
Large	Male	X	Success
Large	Male	X	Success
Small	Male	Y	Success
Large	Male	Y	Failure
Small	Male	X	Success
Large	Male	Y	Success

Figure 2.0.1: snippet of data set consisting of 1050 data points
The variables are:

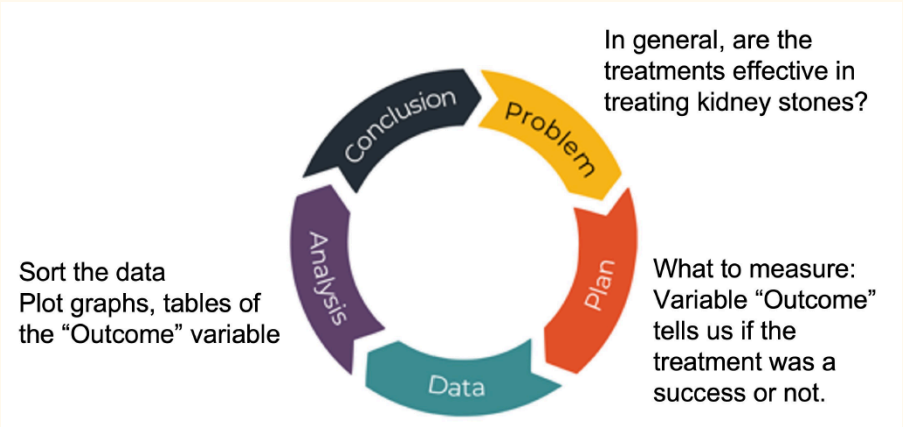
1. The size of the kidney stone. This is an ordinal categorical variable that has two categories. The kidney stones can be classified as either small or large.
2. The gender of the patient. This is a nominal categorical variable that has two categories, male or female.
3. The treatment that the patient underwent. Again, this is a nominal categorical variable and there are two categories, namely treatment X and treatment Y.
4. The outcome of the treatment is also a nominal categorical variable. The categories are success and failure, success means removed entirely or reduced significantly and failure means stone cannot be removed

Questions to be asked:

1. Are the treatments given to the patients successful? In other words, should this new patient receive treatment?
2. Which treatment is better for the new patient?
3. Answering previous questions could lead to more questions

Chapter 2.1 - Rates

Q1: Are the treatments given to the patients successful?



Using Table

Categories of the "Outcome" variable	Count	Rate	Percentage
Success	831	$\text{rate(Success)} = \frac{831}{1050} = 0.791$	$0.791 \times 100\% = 79.1\%$
Failure	219	$\text{rate(Failure)} = \frac{219}{1050} = 0.209$	$0.209 \times 100\% = 20.9\%$
Total	1050	$\frac{1050}{1050} = 1$	$1 \times 100\% = 100\%$

Figure 2.1.1: Table of overall success/failure for both treatments and the rates.

- A preliminary conclusion, based on the limited and best information currently available, is that we should generally recommend the new patient to go for treatment since there are more successful outcomes than failed outcomes.

Bar Graphs

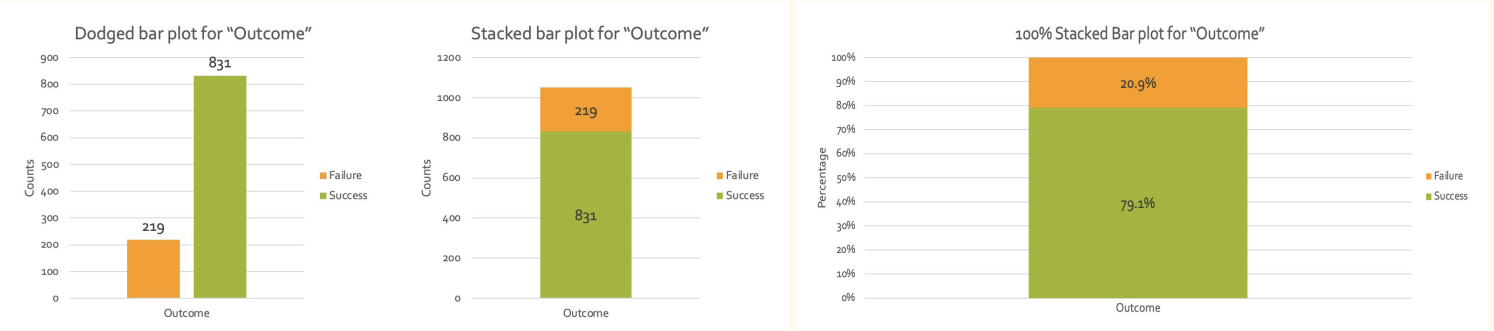


Figure 2.1.2: Dodged bar plot (left), stacked bar plot with absolute number (centre) and percentage (right)

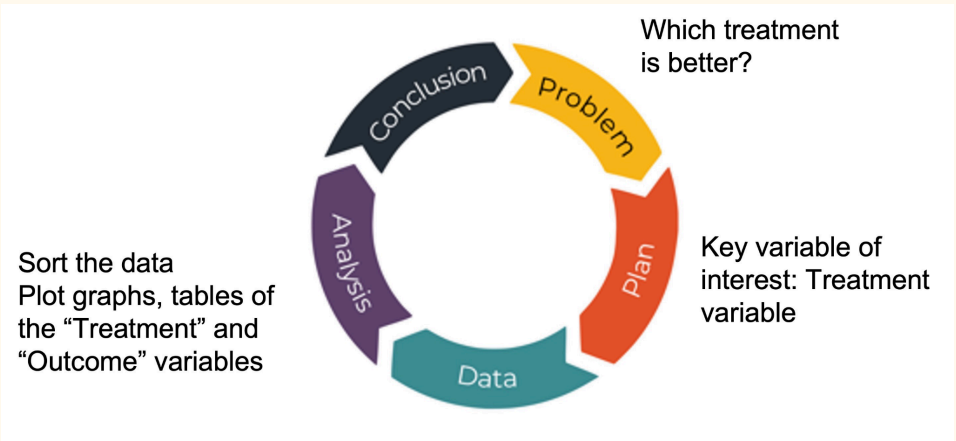
- Dodged bar plot.
  - The x-axis indicates the variable Outcome whereas the y-axis shows the number (that is, the count) of successes and failures in the variable Outcome.
  - Two bars, one for success counts and the other for failure counts, are placed next to each other.
  - Such an illustration is useful in comparing the relative numbers in the categories
- Stacked bar plot
  - The x and y-axes are similar to the dodged bar plot but instead of two bars, we now have only one bar where the counts of failure (219) are stacked on top of the counts of success (831).
  - Such an illustration is useful in comparing the occurrences of each category as a percentage or fraction of the total number of responses.
  - Instead of showing the absolute numbers in each category, it is also possible to show the percentages directly in the plot itself. However, it should be noted that the y-axis is now giving the percentages rather than the actual numbers. (Normalised y axis to be 0-100%)

Regardless of which bar plot is used, we can see that there are many more successes than failures and based on this, it is reasonable to recommend our patient to go for some form of treatment based on the information that we have at this stage.

Note: Kidney stone removal causes little morbidity and mortality in general. If the treatment is of something of higher risk than conclusion may be different.

Now that we are rather convinced that the new patient should receive treatment, the PPDAC cycle brings us back with new questions that arise from our investigation into the data set of 1050 previous patients. It is reasonable to ask the next question of which treatment type is better?

Q2: Which treatment is better for the new patient?



Variable of interest = Treatment type and Outcome

2x2 Contingency Table (before rates)

By convention, the dependent variable is placed on the columns on the table while the independent variable is placed on the rows.

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

Comparing absolute numbers doesn't give a meaningful conclusion, just because X had more success than Y doesn't necessarily mean X is better treatment than Y, it could mean that more patients went to treatment X. Thus the concept of normalisation by using rate is needed when the number of patients going for each treatment is different.

Marginal Rates

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

What proportion of the total number of patients underwent Treatment Y?

$$rate(Y) = 350/1050 = 1/3 = 0.333$$

What proportion of the total number of patients had a successful treatment?

$$rate(Success) = 831/1050 = 0.791$$

These calculations only used 2 numbers in the margin of the table that relates to 1 categorical variable each time thus its called marginal rates.

Conditional Rates

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

Those who underwent Treatment X, what proportion of them had a successful treatment?
OR
Given that the patient underwent treatment X what is the rate of success?

$$rate(Success\ given\ X) = rate(Success \mid X) = 542/700 = 0.774$$

Rate is based on the condition given thus conditional rates. Then based on the condition the rate is limited to the subpopulation.

Joint Rates

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

What is the proportion of patients who chose Treatment Y and had a failure?

$$rate(Y \cap Failure) = 61/1050 = 0.0581$$

This is very different from the conditional rate, joint rates use the whole population.

Example:
What proportion of patients were given treatment Y and had an unsuccessful outcome?

$$rate(Y \cap Failure) = 61/1050 = 0.0581$$

What proportion of patients given treatment Y had an unsuccessful outcome?

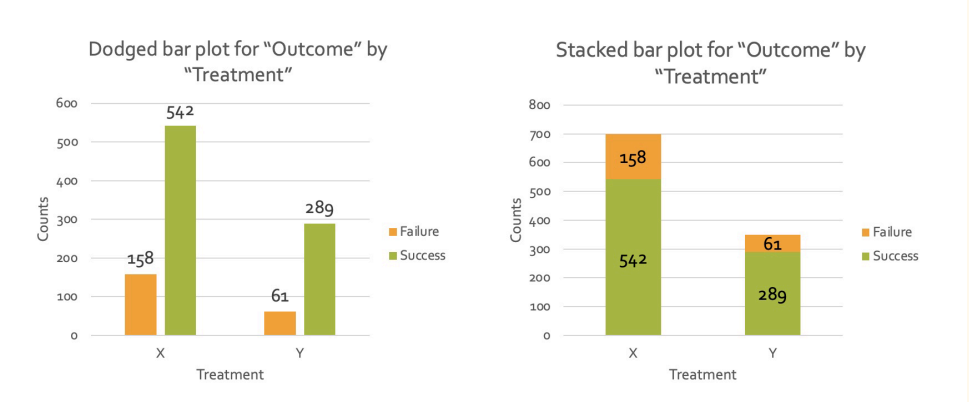
$$rate(failure \mid Y) = 61/350 = 0.174$$

2x2 Contingency Table (after rates)

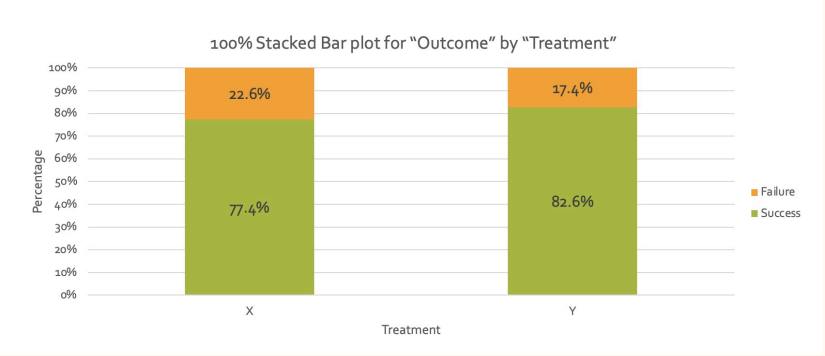
Treatment \ Outcome	Success	Failure	Row Total
X	542 (77.4%)	158 (22.6%)	700 (100%)
Y	289 (82.6%)	61 (17.4%)	350 (100%)
Column Total	831(79.1%)	219 (20.9%)	1050 (100%)

- Can see that the rate(Success|Y) is higher than the rate(Success|X) thus can conclude that treatment Y is positively associated with the success of the treatment.
- The percentages are normalised by row (calculated rates across the rows) thus called row percentages.

Bar Graphs



- Both bar plots tell us that there are a lot more successful treatments in treatment X than in treatment Y, which may lead to the conclusion that treatment X is more effective (since the green bars for treatment X are bigger than the green bars for treatment Y). However, it is also obvious from the stacked bar plot that these two treatments have very different numbers of patients (represented by the height of the two bars).
- Thus the need to plot rates instead of absolute numbers in order to make a fair comparison



- In this plot, notice that both the treatment X and treatment Y bars have been normalised to the same height (which is 100%). We are no longer comparing absolute numbers, but instead comparing the rates of success (the height of the green bars, as a proportion of the total height) between the two treatments. We can see immediately that treatment Y has a higher rate of success (taller green bar) compared to treatment X.
- This means that treatment Y is positively associated with the success of the treatment.

Chapter 2.2 - Association

- Positively/negatively association between two variables X and Y, just means that when Y increases X also increases/decreases.
- This does not imply a causal relationship as we don't know whether change in X is entirely due to Y because data is from an observational study.

Proving Associative Relationship

Establishing association	
Positive association between A and B: (any of the following)	Negative association between A and B: (any of the following)
$\text{rate}(\text{A} \mid \text{B}) > \text{rate}(\text{A} \mid \text{NB})$	$\text{rate}(\text{A} \mid \text{B}) < \text{rate}(\text{A} \mid \text{NB})$
$\text{rate}(\text{B} \mid \text{A}) > \text{rate}(\text{B} \mid \text{NA})$	$\text{rate}(\text{B} \mid \text{A}) < \text{rate}(\text{B} \mid \text{NA})$
$\text{rate}(\text{NA} \mid \text{NB}) > \text{rate}(\text{NA} \mid \text{B})$	$\text{rate}(\text{NA} \mid \text{NB}) < \text{rate}(\text{NA} \mid \text{B})$
$\text{rate}(\text{NB} \mid \text{NA}) > \text{rate}(\text{NB} \mid \text{A})$	$\text{rate}(\text{NB} \mid \text{NA}) < \text{rate}(\text{NB} \mid \text{A})$

Any row can be used, the inequality is mathematically equivalent.

- compare the same rates given different conditions

$rate(A \mid B) = rate(A \mid NB)$
No association.

$rate(A \mid B) > rate(A \mid NB)$
A is positively associated with B, presence of A when B is present is stronger compared to when B is absent.

$rate(A \mid B) < rate(A \mid NB)$
A is negatively associated with B, presence of A when B is present is weaker compared to when B is absent.

Using the kidneys tone example with A = Success, NA = Failure, B = Treatment X and NB = Treatment Y:

$$rate(A \mid B) = rate(Success \mid X) = 542/700 = 0.774$$
$$rate(A \mid NB) = rate(Success \mid Y) = 289/350 = 0.826$$
$$rate(A \mid B) < rate(A \mid NB)$$

Therefore, success of treatment is said to be negatively associated with X and positively associated with Y.

Chapter 2.3 - Rules on Rates

Symmetry Rule

Symmetry Rule Part 1:

$$\text{rate}(\text{A} \mid \text{B}) > \text{rate}(\text{A} \mid \text{NB}) \Leftrightarrow \text{rate}(\text{B} \mid \text{A}) > \text{rate}(\text{B} \mid \text{NA}).$$

Symmetry Rule Part 2:

$$\text{rate}(\text{A} \mid \text{B}) < \text{rate}(\text{A} \mid \text{NB}) \Leftrightarrow \text{rate}(\text{B} \mid \text{A}) < \text{rate}(\text{B} \mid \text{NA}).$$

Symmetry Rule Part 3:

$$\text{rate}(\text{A} \mid \text{B}) = \text{rate}(\text{A} \mid \text{NB}) \Leftrightarrow \text{rate}(\text{B} \mid \text{A}) = \text{rate}(\text{B} \mid \text{NA}).$$

- LHS is true if and only if the RHS is true AND RHS is true if and only if the LHS is true.
- Showing one side is true implies the other side is also true

Derivation for Part 1:

rate(A | B) > rate (A | NB) ⇔ rate(B | A) > rate(B | NA)

	B	Not B	Row Total
A	w	x	w + x
Not A	y	z	y + z
Column Total	w + y	x + z	w + x + y + z

$$\frac{w}{w+y} > \frac{x}{x+z} \qquad \frac{w}{w+x} > \frac{y}{y+z}$$
$$w(x+z) > x(w+y) \qquad w(y+z) > y(w+x)$$
$$wx + wz > xw + xy \qquad wy + wz > yw + yx$$
$$wz > xy$$

Logical Reasoning for Part 1:

1. Rate of A given B is more than the rate of A given NB.
2. Positive association between A and B.
3. More likely to see A when B is present as compared to when B is absent.
4. Also more likely to see B when A is present as compared to when A is absent.
5. Rate of B given A is more than the rate of B given NA.
6. Reverse is also true

Example:

Outcome / Treatment	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

Using the kidneys tone example with A = Success, NA = Failure, B = Treatment X and NB = Treatment Y:

$$rate(A \mid B) = rate(Success \mid X) = 542/700 = 0.774$$
$$rate(A \mid NB) = rate(Success \mid Y) = 289/350 = 0.826$$
$$rate(A \mid B) < rate(A \mid NB)$$

$$rate(B \mid A) = rate(X \mid Success) = 542/831 = 0.652$$

$rate(B \mid NA) = rate(X \mid Failure) = 158/219 = 0.721$
$rate(B \mid A) < rate(B \mid NA)$
as predicted by symmetry

Basic Rule On Rates

Basic rule on rates:

The overall rate(A) will always lie between rate(A | B) and rate(A | NB).

Consequence 1:

The closer rate(B) is to 100%, the closer rate(A) is to rate(A | B).

Consequence 2:

If rate(B) = 50%, then rate(A) = $\frac{1}{2}$[rate(A |B) + rate(A | NB)].

Consequence 3:

If rate(A | B) = rate(A | NB), then rate(A) = rate(A | B) = rate(A | NB).

$$rate(A) = \frac{n_A}{n_{Total}} = \frac{rate(A \mid B) * n_B + rate(A \mid NB) * n_{NB}}{n_{Total}} = weighted\ average\ of\ rate(A \mid B)\ and\ rate(A \mid NB)$$

The basic rule on rates states that the overall rate(A) is always between rate(A | B) and rate(A | NB)

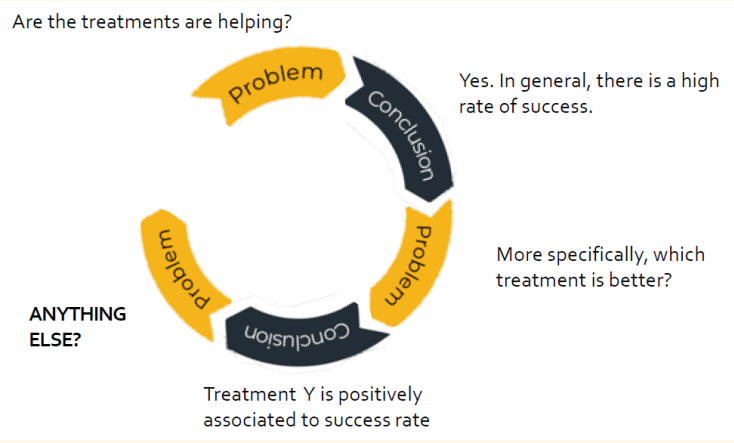
1. The first consequence gives us a little more indication of where the overall rate(A) is going to be. If rate(B) is closer to 100% (than rate(NB)), then rate(A) is going to be closer to rate(A | B) compared to rate(A | NB).
    - If B takes up higher proportion then it will have a greater weight in the weighted average thus moving rate(A) towards rate(A | B)
2. The second consequence specifically states that if rate(B) is exactly 50%, then rate(A) will be exactly the midpoint between rate(A | B) and rate(A | NB).
3. Finally, the third consequence states that if the two conditional rates, namely rate(A | B) and rate(A | NB) are the same, then the overall rate(A) will also take the same value of the two conditional rates.
    - This means A is not associated with B, independent.

Example:

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

Using the kidneys tone example with A = Success, NA = Failure, B = Treatment X and NB = Treatment Y:
$rate(A \mid B) = rate(Success \mid X) = 542/700 = 0.774$
$rate(A \mid NB) = rate(Success \mid Y) = 289/350 = 0.826$
$rate(A) = 831/1050 = 0.791$
$rate(B) = 700/1050 = 2/3 = 0.667$
Thus rate(A) is closer to rate(A | B) then rate(A | NB)

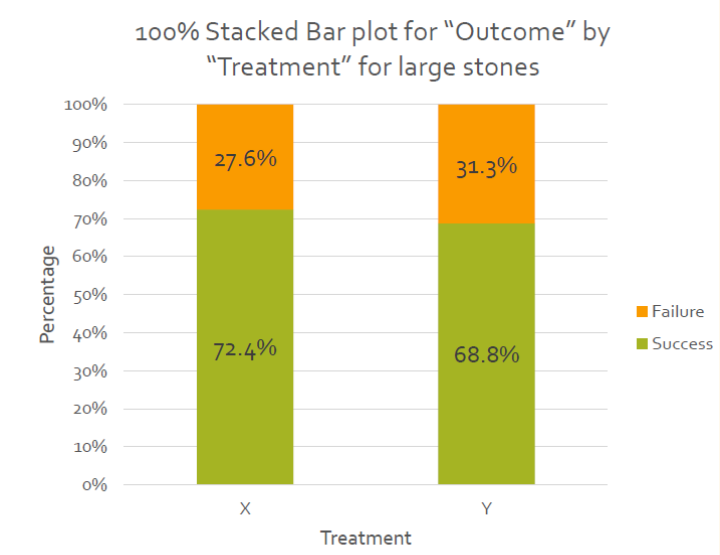
Chapter 2.4 - Simpson's Paradox



Size	Gender	Treatment	Outcome
Large	Male	X	Success
Large	Male	X	Success
Small	Male	Y	Success
Large	Male	Y	Failure
Small	Male	X	Success
Large	Male	Y	Success

- So far only considered the relationship between the variables of interest, treatment type and outcomes, while ignoring the other variables like gender and kidney stone size.
- Now let's plot some graphs with stone size large and small, using a stacked bar chart as comparison is done with binary categorical variables.

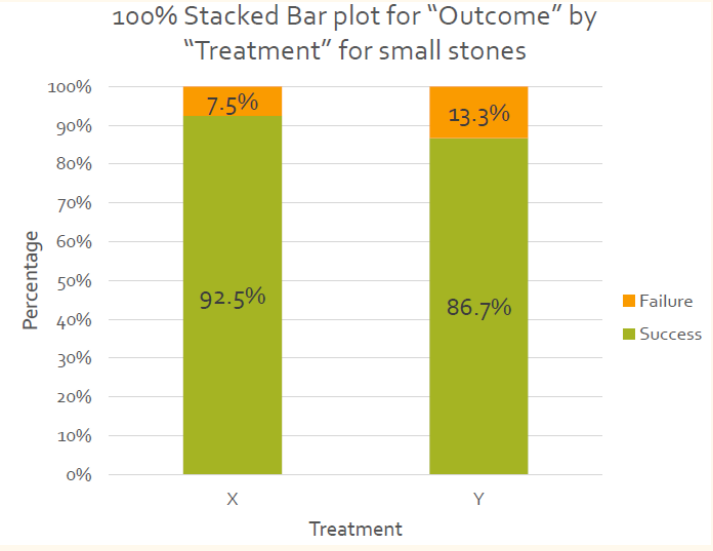
Large Stone



Large stones	Success	Failure	Total
X	381	145	526
Y	55	25	80
Total	436	170	606

- rate(Success | X) > rate(Success | Y)
- Across large stones, treatment X is better as it is positively associated with successful treatment.

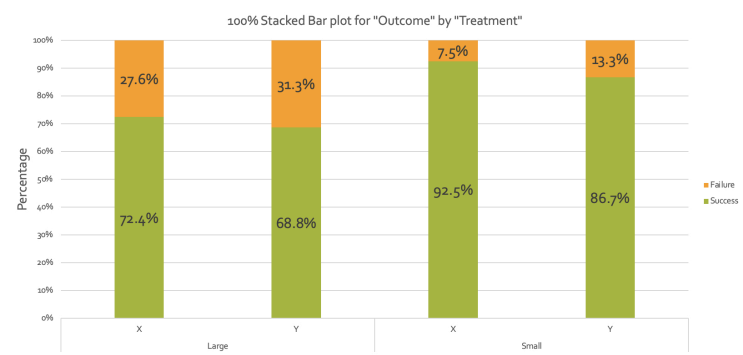
Small Stone



Small stones	Success	Failure	Total
X	161	13	174
Y	234	36	270
Total	395	49	444

- rate(Success | X) > rate(Success | Y)
- Across small stones, treatment X is also better as it is positively associated with successful treatment.

Large and Small Stone



- The bar chart above is called a *sliced stacked bar plot.* It is showing 3 variables at once: stone size, treatment outcome and treatment type.
- How can it be that when considering large stone and small stone separately, treatment X is better but when considering overall, treatment X is performing worse. This phenomenon is known as Simpson's Paradox.
- Now, how to recommend which treatment to use for our new patient. If the patient is under a large stone category, Treatment X is better. If the patient is under the small stone category, Treatment X is also better. Thus Treatment X is the better option either way, case by case.

Simpson's Paradox

> Simpson's Paradox is a phenomenon in which a trend appears in more than half of the groups of data but disappears or reverses when the groups are combined. Here, "disappears" means the two variables in question (say A and B) are no longer associated, that is, rate(A | B) = rate(A | NB).

If only two categories, majority means two. (+ - na cannot say got paradox, + + na means got paradox)

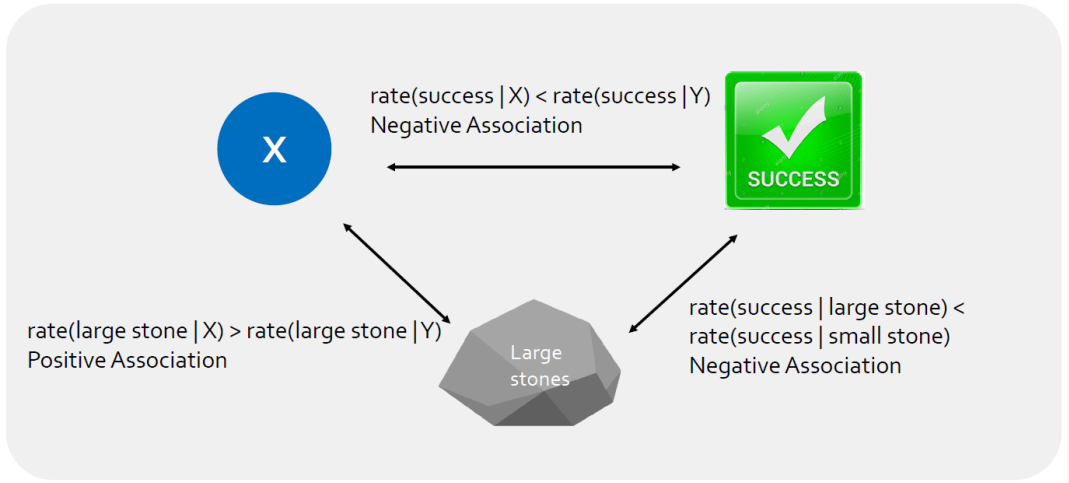
To see why this occurs, the data will need to be separated by category and form subgroups, this type of subgroup analysis is called slicing. Which we did when we analysed large and small stones separately.

	Large stones			Small stones			Total (Large+Small)		
	Succ.	Total trt.	R(succ.) in %	Succ.	Total trt.	R(succ.) in %	Succ.	Total trt.	R(succ.) in %
X	381	**526**	**72.4%**	161	**174**	**92.5%**	542	700	77.4%
Y	55	**80**	68.8%	234	**270**	86.7%	289	350	**82.6%**

- Looking at blue numbers:
  - Treatment X seems to be used more for removing large kidney stones compaore to small ones.
  - By basic rule on rates, the overall success rate of X must lie in between 72.4% - 92.5% while closer to 72.4%, which is true as overall treatment X success rate is 77.4%.
- Looking at orange numbers:
  - Treatment Y seems to be used more for removing small kidney stones compaore to big ones.
  - By basic rule on rates, the overall success rate of Y must lie in between 68.8% - 86.7% while closer to 86.7%, which is true as overall treatment Y success rate is 82.6%.
- Another observation would be that the success rate for treatment of large kidney stones is lower between 68.8% (treatment Y) and 72.4% (treatment X). And higher for small kidney stones between 86.7% (treatment Y) and 92.5% (treatment X). This tells us that treatments for large stones have a lower rate of success compared to small stones, which is not unreasonable to believe.
- Based on these 3 observations, it is no surprise that the overall success rate of X is lower than that of Y.
- In conclusion, we can explain Simpson's Paradox in the following way.
  - Treatment X is in fact a better treatment than Y.
  - However, because patients have been using Treatment X to treat more difficult cases (large kidney stones), this lowers the overall success rate of treatment X.
  - It does not change the fact that in the individual subgroups, regardless of stone size, treatment X achieves a higher success rate than treatment Y.
  - Slicing the data into the small and large stone subgroups will reveal that treatment X is indeed a better treatment.

    - Note that even though it is known that treatment X is better, overall treatment X is still negatively associated with successful treatment.
- In this case, this means that stone size is a third variable that was associated with the other two variables whose relationship we were initially investigating, thus affecting the conclusion of our initial study. Such a variable is called a confounder
- When Simpson's Paradox is observed, it implies that there is definitely a confounding variable present, that is a third variable that is associated with the two variables whose relationship we are investigating.
- However, the existence of a confounder does not necessarily lead to Simpson's Paradox.

Confounding Variable



- Treatment X is positively associated with large kidney stones.
- Large kidney stones are negatively associated with successful treatment.
- Thus this makes Treatment X to be negatively associated with successful treatment.

Chapter 2.5 - Confounders

A confounder is a third variable that is associated to both the independent and dependent variable whose relationship we are investigating.

Note: the direction of association with the variables of interest doesn't matter, as long as they have association.

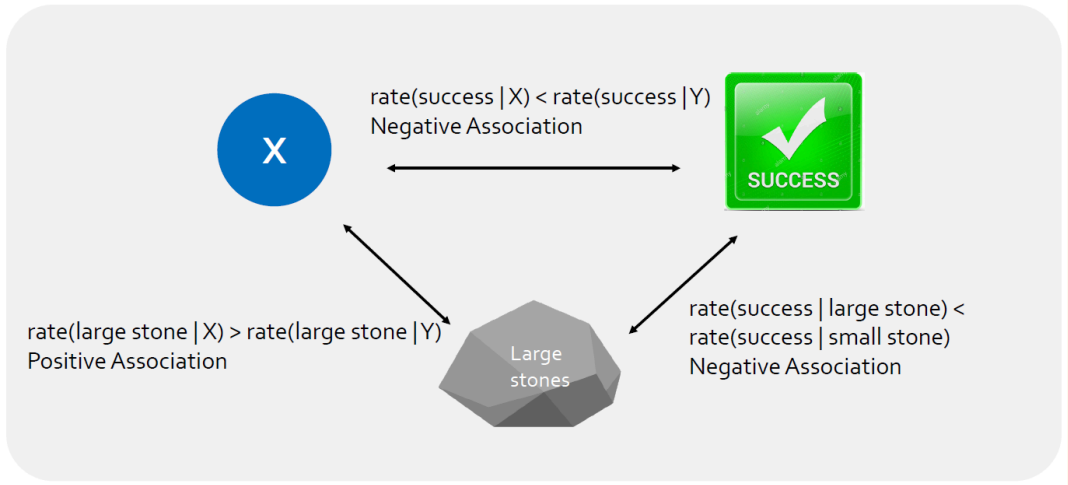
Proving Stone Size Is A Confounding variable

	Large	Small	Total
X	526	174	700
Y	80	270	350
Total	606	444	1050

$rate(Large \mid X) = 526/700 = 0.751$
$rate(Large \mid Y) = 80/350 = 0.229$
rate(Large | X) > rate(Large | Y)
X is positively associated with Large kidney stone

	Success	Failure	Total
Large	436	170	606
Small	395	49	444
Total	831	219	1050

$rate(Success \mid Large) = 436/606 = 0.719$
$rate(Success \mid Small) = 395/444 = 0.890$
rate(Success | Large) < rate(Success | Small)
Large kidney stone is negatively associated with success of treatment



Thus stone size is a confounding variable making Treatment X to be negatively associated with the success of treatment.

Dealing WIth Confounding Variables

- To remove confounding variables, all that needs to be done is to cut one of the associations it has to the variables of interest.
- This can either be done by slicing or random assignment

Slicing

Slicing is where data is segregated by the confounding variable and the association between the dependent and independent variables is studied separately for each case.

- Example: this is done by investigating the association between the dependent and independent variables for large stone cases separately from the small stone cases.
- This method requires researchers to collect more data than the two variables of interest, any variable that may be relevant in the study will need to be collected.
    - This is why surveys also collect background information, however participants seeing this may be discouraged to participate. Must strike a balance of enough information but not too much.
    - Collecting information on variables is costly in practice. Even if we do manage to collect all the information we need, the analysis can be complicated if the data needs to be sliced along many different variables.
    - For non-randomised designs like observational studies, we can never be totally sure that every single confounder has been identified and controlled for. Thus, observational studies offer only a limited conclusion in providing evidence of association and not causation.
- If the confounding variable is never collected, then there is no way of knowing the confounding variable.

Random Assignment

Fundamentally, confounding variables occur due to association which is a consequence of having an unequal proportion of variables in the two groups that we are trying to compare.

- In the kidney stone example, stone size was a confounder because patients with large stones were disproportionately allocated to treatment X instead of treatment Y.
- Now, if the allocation to treatment is to be done randomly, with a big enough group of patients, the two groups of patients will have the same proportion of large and small kidney stones. Thus the association between size and treatment will be removed, and the confounding variable is also removed.
- However, random assignment may not be feasible due to ethical reasons. Patients should have the right to choose which treatment they want.
- Then the next best solution to deal with confounding variables would be slicing.

Quiz 2

d. Upon completing parts (b) and (c) of this question, your friend Tammy told you that the answer to part (c) could have been deduced from (b). She explains that sex must be a confounder in the study because it is associated with whether a student is enrolled in a STEM first degree, which is then associated with the year of enrolment. Hence, sex is associated with both a student's year of enrolment and whether a student is enrolled in a STEM first degree. Based on Tammy's logic alone, is Tammy's claim necessarily correct?

For sex to be a confounder it must be proven that it is associated with both STEM and year directly, thus Tammy is a dummy. From her statement, it can only be said that STEM is the confounder when studying association between sex and year.

In general, association between categorical variables is not transitive: X is associated with Y and Y is associated with Z does not mean that X is associated with Z.

A study was conducted among 100 subjects to see if Covid-infection status was associated with increased self-reported loneliness (indicated as Yes/No) during lockdown restrictions. The results are summarised in the table below.

	Yes		No	
	Female	Male	Female	Male
Infected	25	25	3	10
Non-infected	10	12	8	7

Which of the following statements is/are correct? Select all that apply.

- [ ] This is an experimental study.
- [ ] Simpson's paradox is observed when examining the association between infection status and increased self-reported loneliness.
- [ ] There is a positive association between being female and being Covid-infected.
- [ ] Gender is a confounder in the association between infection status and increased self-reported loneliness.
- [x] Overall, there is a positive association between being Covid-infected and having increased self-reported loneliness.

Just because no simpson's paradox doesn't mean no confounder. gender is a confounder because female and infection is -ve associated and infected and loneliness is positively associated.

Chapter 3 - Numerical Variable Analysis

- Chapter 3.1 - Univariate EDA

- Chapter 3.2 - Bivariate EDA
- Chapter 3.3 - Correlation coefficient
- Chapter 3.4 - Linear regression

EDA - Exploratory Data Analysis is a process of summarising or understanding the data and extracting insights or main characteristics of the data.

Univariate data		Bivariate data	
Overall pattern	Deviation from the pattern	Overall pattern	Deviation from the pattern
1) Shape 2) Center 3) Spread	Outliers	1) Direction 2) Form 3) Strength	Outliers

Chapter 3.1 - Univariate EDA

- Distribution
- Histogram
- Boxplot
- Histogram vs Boxplot

Distribution

Distribution can be described by its overall pattern and deviation from the pattern

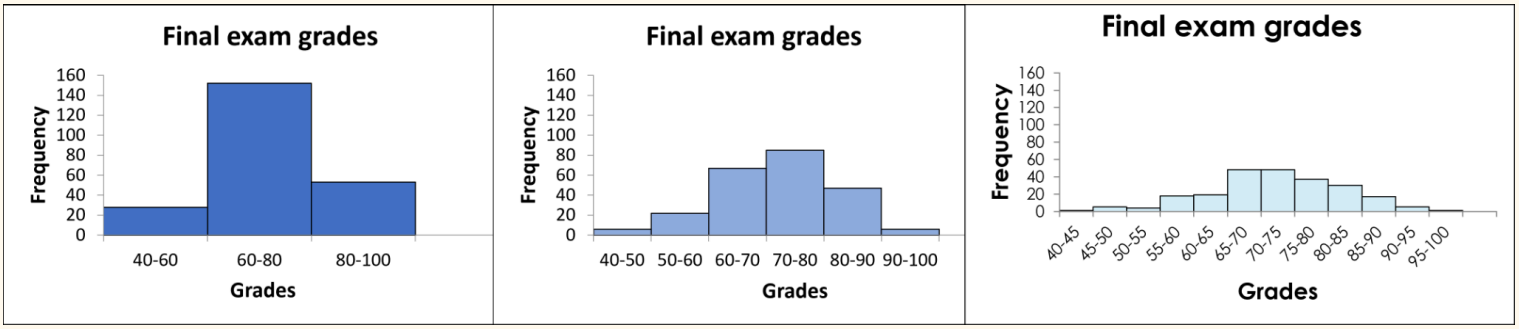
- Shape, its peaks and skewness, which can be gotten from histogram/box plot
- Centre and spread like mean, median, mode, standard deviation, IQR.
- Outliers which can be picked out easily using box plots.

Histogram

A histogram is a graphical representation that organises data points into ranges or bins.

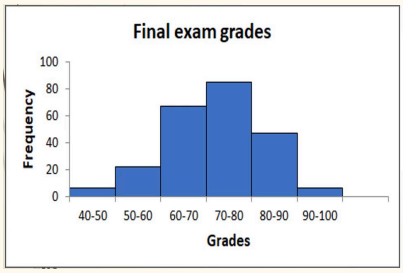
- Graphical display of distribution
- It is particularly useful when we have large data sets.
- To create a histogram, the variable values are “grouped” into equal size intervals called bins. With the height of each bar representing the frequency for that bin range, the highest bar would represent the most frequently occurring range of values.
- Many ways to fill the bins, convention in GEA1000 is to exclude the left end point and include the right endpoint.

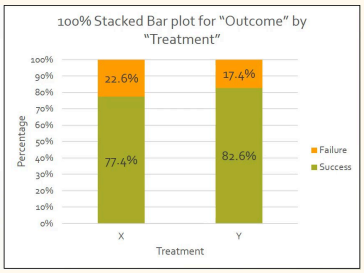
Size of Bins



- If bin widths are too large, it will result in only a few bins and information in the data will be lost when data points are grouped together into a small number of groups/bins.
- If bin widths are too small, there may be bins that have very few data points (or none) that do not give us a sense of the distribution.
- The initial choice of bin width may not be the most appropriate. Different histograms with various bin widths should be created before deciding which one is the most useful and informative.

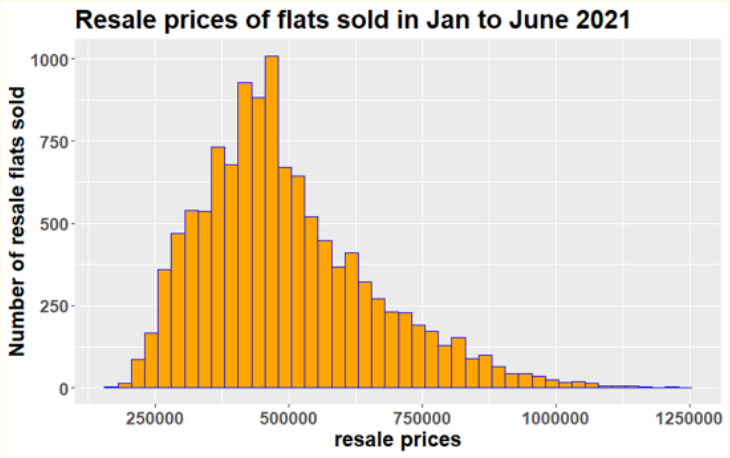
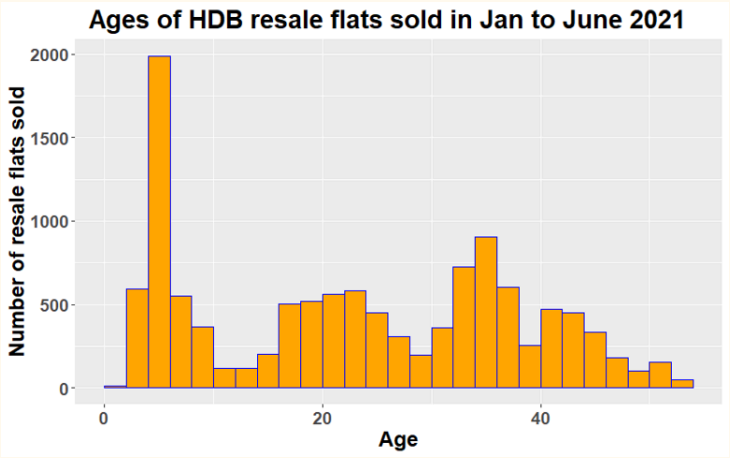
Histogram VS Bar Graph





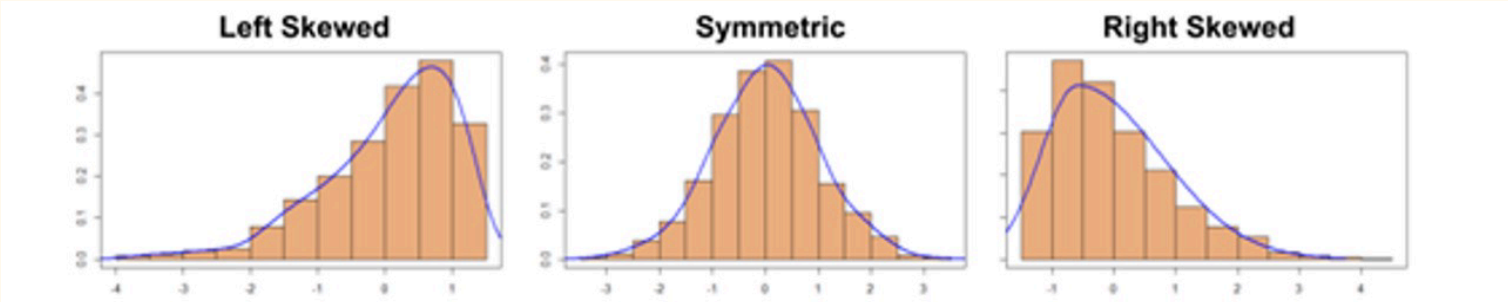
- Histogram show the distribution of numerical variable across a numberline
- Bar graph makes comparison across categories of a variable
- Order of bars in histogram cannot be changed due to the inherent ordered manner of numerical variable
- There are also usually no gaps between the bars in a histogram.

Peak

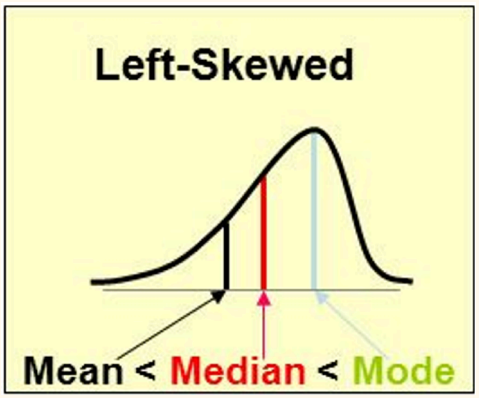
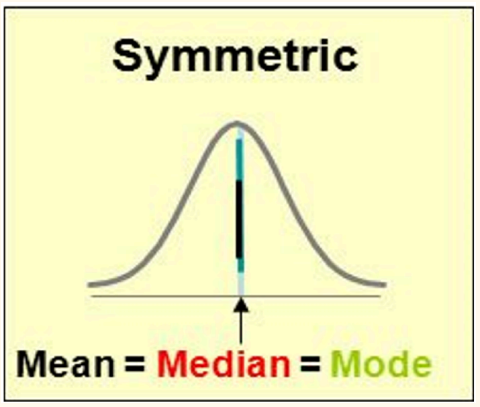
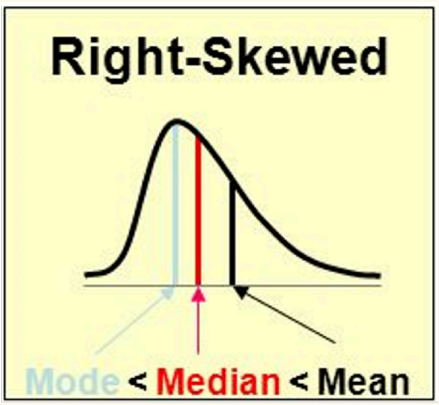
Unimodal	**Multimodal**
	
One distinct peak = unimodal	More than one distinct peak = multimodal Two distinct peaks = bimodal

Peaks are those that are most frequently in their immediate neighbourhoods of age ranges.

Skewness and Central Tendencies



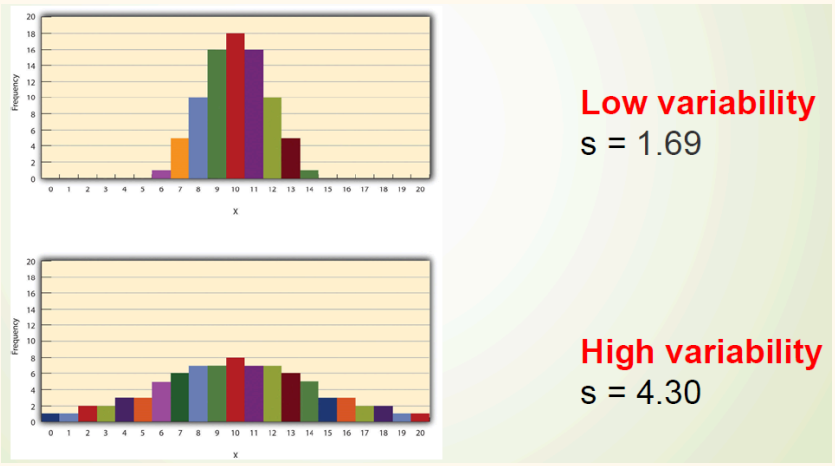
Left Skewed Distribution	**Symmetrical Distribution**	**Right Skewed Distribution**

		
Usually but not always mean < median < mode This is due to the small number of extremely small values which contributes to the long tail on the left, will push down the mean, as compared to the median which is less affected by these extremely small values.	The mean, median and mode will be very close to each other near the peak of the distribution Normal distribution is an example.	mode < median < mean This is due to the small number of extremely big values which contributes to the long tail on the right. These big values will push up the mean as opposed to the median which is less affected by these extremely large values.

For skewed distribution, the median is used for the central tendency to avoid the effects from the extreme values.

Spread

The spread of a distribution refers to how the data vary around the central tendency.



- Standard deviation, IQR and range are ways of measuring spread
- Range is susceptible to outliers.

Boxplot

A boxplot is a graphical representation that uses the five-number summary. Namely the minimum, quartile 1 (Q1), median (Q2), quartile 3 (Q3) and maximum.

Outlier

Outlier is an observation that falls well above or below the overall bulk of the data.

- if there isn't much difference to the result of study when the outlier is taken out, then it is good to take them out of the analysis.
- if there's a great effect on the result then cannot remove and must find out why there is such an outlier. Example, if its data entry error then can take out, if it leads to interesting insights into the behaviour of the data then cannot take out
- If they have minimal effect on the conclusions and if we cannot figure out why they are there, such outliers may possibly be removed. However, if they substantially affect the results, then we should not drop them without justification.

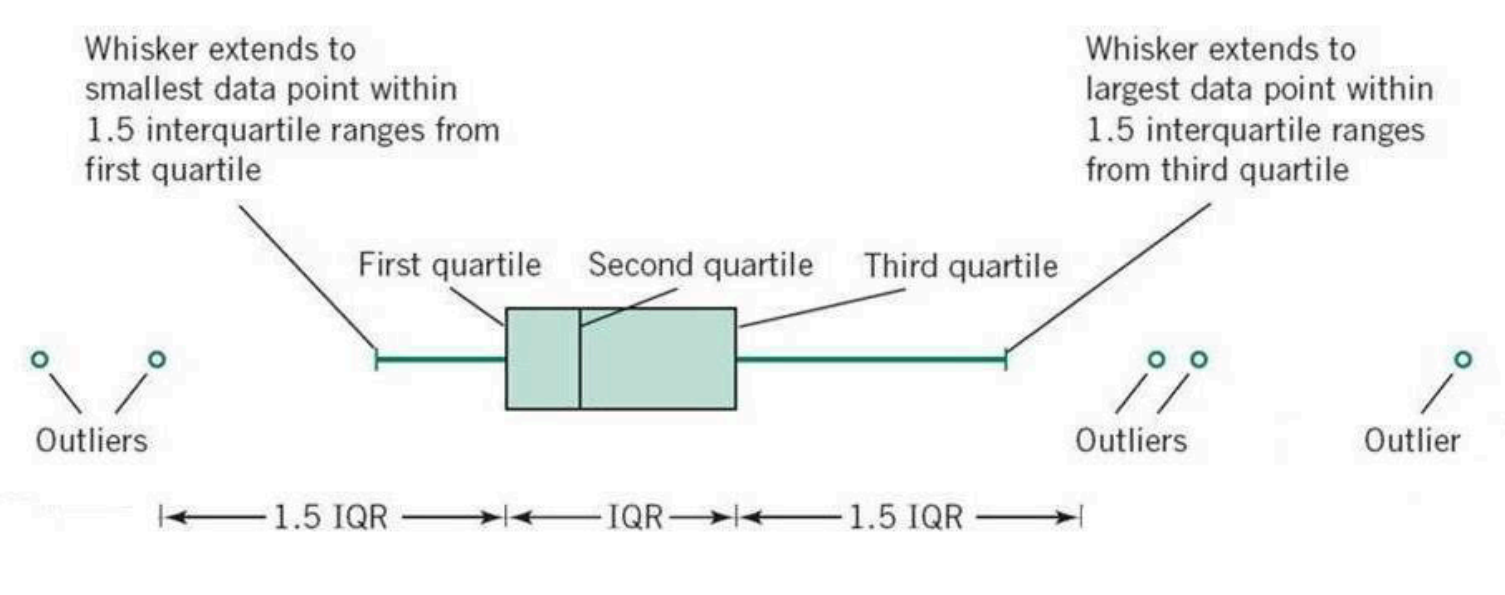
$$4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 300.$$

	Mean	Median	Mode	Standard deviation
Without removing 300	30	5.5	5	85.03
With 300 removed	5.45	5	5	1.04

- Without removing the outlier, the mean is pulled away in the direction of the skew (in this example, the distribution is skewed to the right).
- In such cases, mean may no longer be a good measure of the central tendency of the distribution. We call the median and the mode robust statistics.
    - IQR can increase, decrease and remain the same after the outlier is removed.

A data point is considered an outlier if it satisfies one of the following conditions:
- The value of the data point is greater than Q3 + 1.5 × IQR
- The value of the data point is less than Q1 − 1.5 × IQR

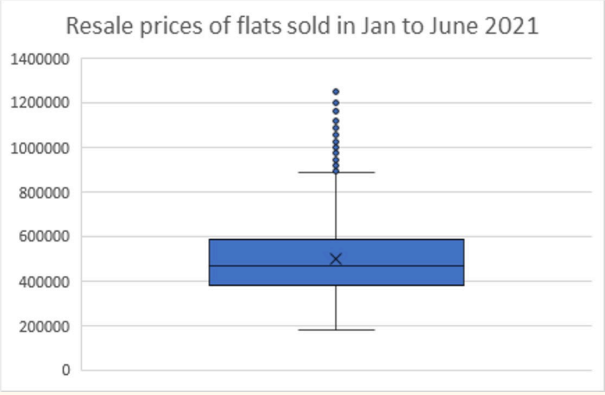
How to Draw Boxplot



1. Draw a box from Q1 to Q3.
2. Draw a vertical line in the box where the median (Q2) is located.
3. Identify all the outliers by using the consideration above.
4. Extend a line from Q1 to the smallest value that is not an outlier and another line from Q3 to the largest value that is not an outlier. These lines are called whiskers.
5. Mark each of the outliers with dots or asterisks.
6. In Excel, the mean is marked by an X.

If there are no whiskers, it means Q1/Q3 = min/max.

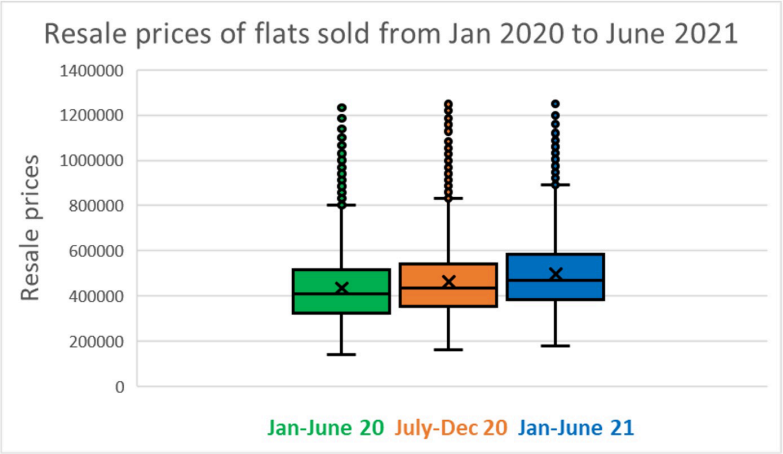
How to Interpret Boxplots



- Shape
    - From the boxplot, we see that the variability in the upper half of the data, given by (Max − Median) is significantly larger than the variability in the lower half of the data which is equal to (Median − Min).

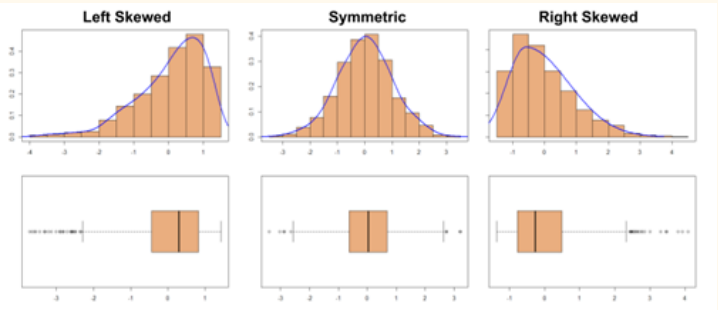
    - This means that the distribution is skewed to the right and there is a relatively long tail to the upper end of the distribution due to the existence of outliers.
- Center
    - The centre, described by the median, is easily observed in the box plot, unlike in a histogram. We can also compare the relative positions of the median and the mean from the boxplot.
- Spread
    - The IQR of 204000 gives us an idea of the spread for the middle 50% of the data set. On its own it may not be immediately informative but this would be a meaningful measure to compare across different distributions.

Comparing Box Plots

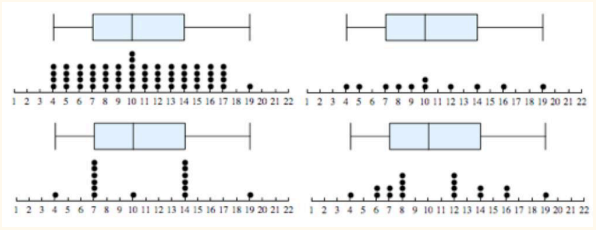


- All three distributions are right skewed as the upper halves of the data have greater variability than the lower halves, due to (large-valued) outliers.
- However, upon a closer look, it is also apparent that the upper half variability in period P1 is greater than the upper half variability in P2 which in turn is greater than the upper half variability in P3.
- The middle 50% (that is, the IQR) box of resale prices is lowest in P1, followed by P2 and then P3. Hence, the overall resale prices have increased over time. The spread (given by the height of the boxes) appears to be similar between P1 and P2 while slightly higher in P3.
- There appears to be more outliers in P1 and P2 compared to P3. The largest outlier is about the same while lowest outliers is lower in earlier periods shows that sales of expensive flats are not too affected by economic conditions.

Histogram vs Boxplot



- A histogram typically gives a better sense of the shape of the distribution of a variable, compared to a boxplot. When there are great differences among the frequencies of the data points, a histogram will be able to illustrate this difference better than a boxplot.
- If we wish to compare the distributions of different data sets, putting the different boxplots side by side is more illustrative than using histograms.
- To identify and indicate outliers, boxplots do a better job than histograms.
- The number of data points we have in a data set is better shown in a histogram than in a box plot.
    - In fact, two distributions with very different number of data points can have almost identical boxplots
    - On the other hand, this difference is apparent by comparing the histograms



The bottom line is that different graphics and summary statistics have their advantages and disadvantages and they are often used together to complement each other.

Chapter 3.2 - Bivariate EDA

Deterministic relationships

If two variables have a deterministic relationship, then the value of one variable can be determined exactly if the value of the other variable is known.

- ie: there's a fixed formula
- eg: Fahrenheit to Degree Celsius, C = (F − 32) × 5/9

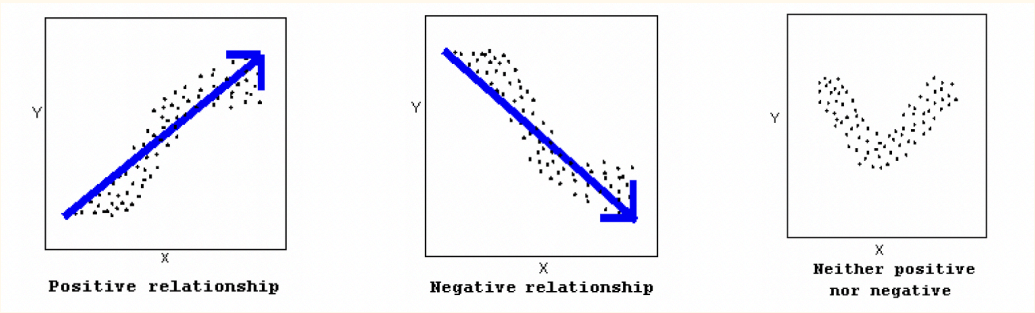
Non-deterministic relationships or Associations

If the relationship between two variables is not deterministic in nature, then the value of one variable cannot be determined exactly if the value of the other variable is known, best that can be done is predicting the average value of the other variable.

- Also called statistical or non-deterministic.
- Example of such relationships between variables is associations.

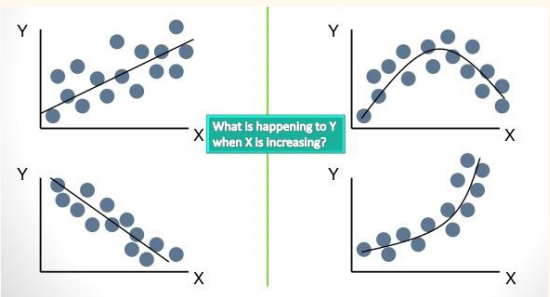
Scatter Plots

Direction



- The direction of the relationship can be either positive, negative or neither.
- A positive relationship between two variables when an increase in one of the variables is associated with an increase in the other variable.
- On the other hand, a negative relationship between two variables means that an increase in one variable is associated with a decrease in the other.
- Not all relationships can be classified as either positive or negative and there are those that do not behave in one way or the other.

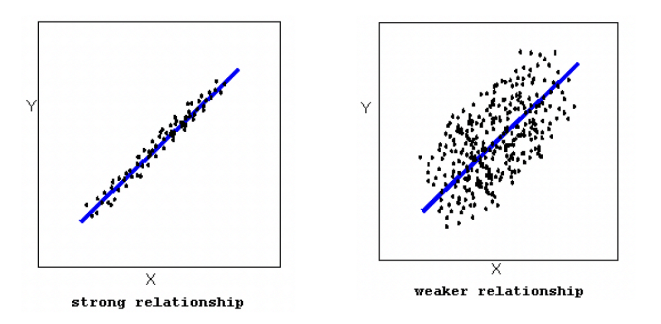
Form

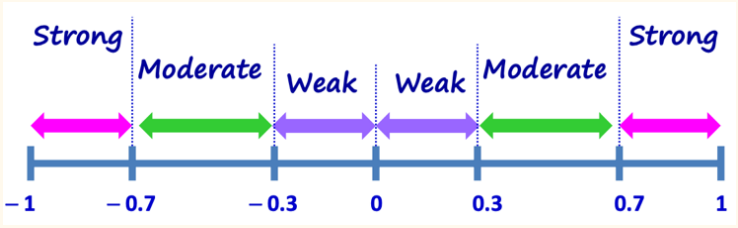


- The form of the relationship describes the general shape of the scatter plot.

- In general the form of the relationship is either linear or nonlinear.
- The form of the relationship is linear when the data points appear to scatter about a straight line.
- When the data points appear to scatter about a smooth curve, we say that the form of the relationship is non-linear.

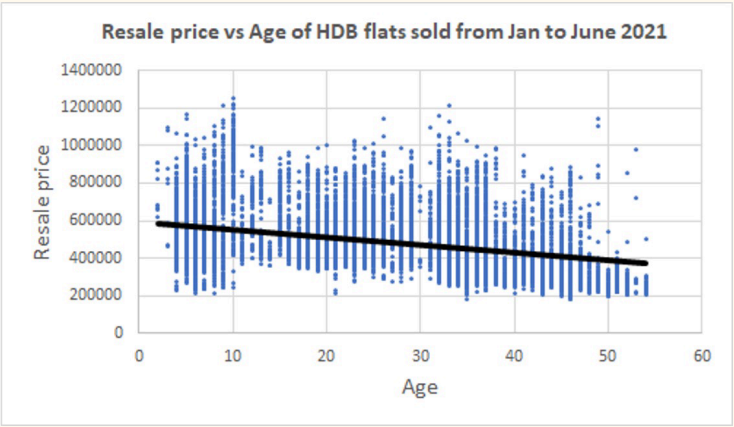
Strength



- The strength of the relationship indicates how closely the data follow the form of the relationship.
- Both scatter plots above suggest that there is a positive, linear relationship between the two variables.
- However, the scatter plot on the left shows the data points lying very close to the straight line.
- This indicates that the strength of the relationship is strong.
- The scatter plot on the right shows the data points scattered loosely around the straight line and thus the strength of the relationship is weaker than that in the scatter plot on the left.
- Correlation coefficient is used to quantify the strength.
- 

    - Weak, moderate and strong

Example



The trendline suggests that as the age of the HDB flat increases, the resale price decreases linearly on average, in the period of January to June 2021. Strength is subjective if seen from the scatter plot, so need a better definition of strength, i.e. the correlation coefficient.

Chapter 3.3 - Correlation Coefficient

- Calculating Correlation Coefficient
- Properties of Correlation Coefficient
- Interpreting Correlation Coefficient
- Limitations of Correlation Coefficient

Correlation coefficient

- The correlation coefficient between two numerical variables is a measure of the linear association between them.

- The correlation coefficient, denoted by r, always ranges between −1 and 1.
- Used to summarise the direction and strength of linear association between two variables.

Calculating Correlation Coefficient

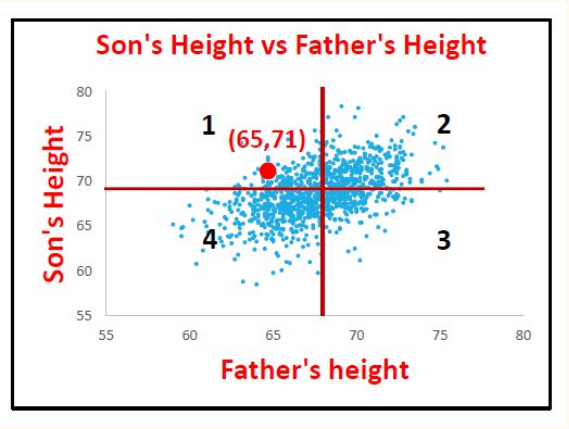
$$Standard\ unit:\ SU_x = \frac{x-\bar{x}}{s_x}\ \&\ SU_y = \frac{y-\bar{y}}{s_y}$$

$$Correlation\ coefficient = r = \frac{\sum(SU_x \times SU_y)}{n-1}$$

$$s_x = standard\ deviation$$

$$n = number\ of\ data\ points$$

Standard unit means how many sample standard deviations from the sample mean, if + then is SUx SD above sample mean.



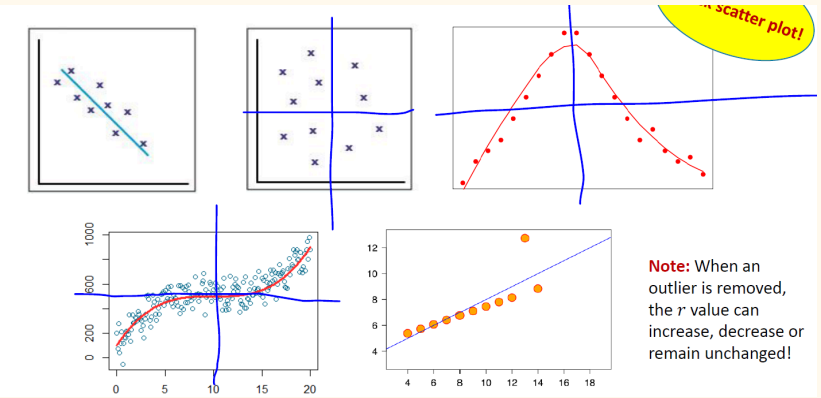
In the quadrants:

Q1: the product of standard unit is negative	SUx < 0	SUy > 0
Q2: the product of standard unit is positive	SUx > 0	SUy > 0
Q3: the product of standard unit is negative	SUx > 0	SUy < 0
Q4: the product of standard unit is positive	SUx < 0	SUy < 0

So if there is a point at the same y but the x is reflection along the x mean vertical line, the standard unit would be negative of another and will cancel out in the calculation of correlation coefficient.

Therefore if the points are scattered with no association they are likely to reflect along the lines so r is close to 0. If more points lie in Q2 and Q4, the r value will be very positive and thus a strong positive association. SImilarly, if more points lie in Q1 and Q3, the r value will be very negative and thus strong negative association

However this will not take into consideration where the points are on a line thus scatter plot must be used to check the form or the relationship. This also explains why for non-linear relationships there could also be high magnitudes of r.

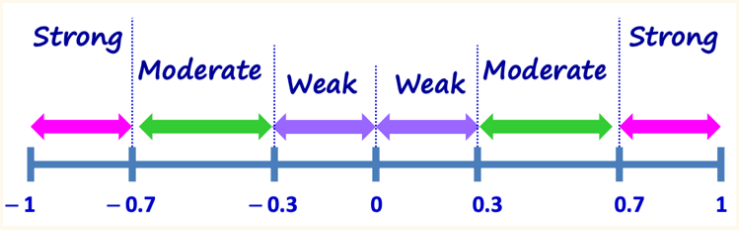


Properties of Correlation Coefficient

1. The correlation coefficient r is not affected by interchanging the x and y variables.
2. The correlation coefficient r is not affected by adding a number to all values of a variable.
3. The correlation coefficient r is not affected by multiplying a positive number to all values of a variable.
    a. multiplying negative values will make correlation coefficient r to have opposite signs.

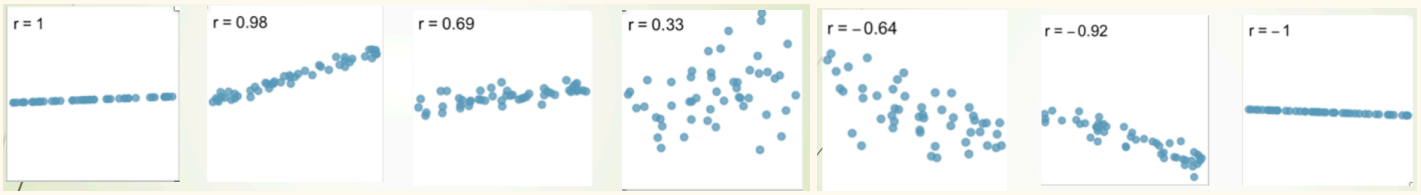
All of these can be shown by using the formula for standard unit and correlation coefficient and properties of mean and standard deviation.

Interpreting Correlation Coefficient

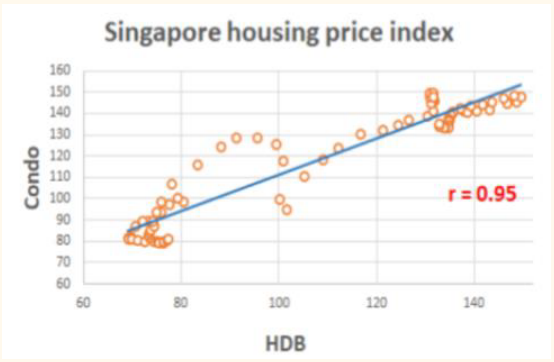
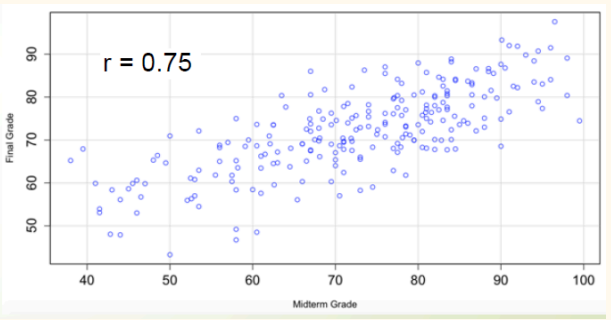
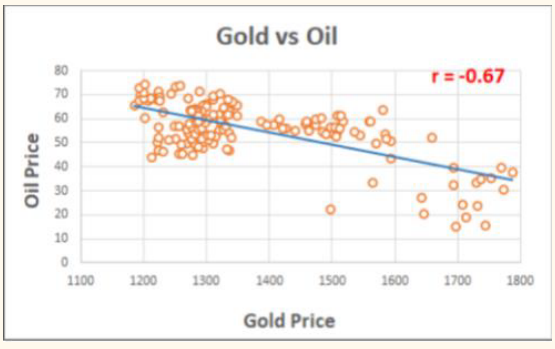
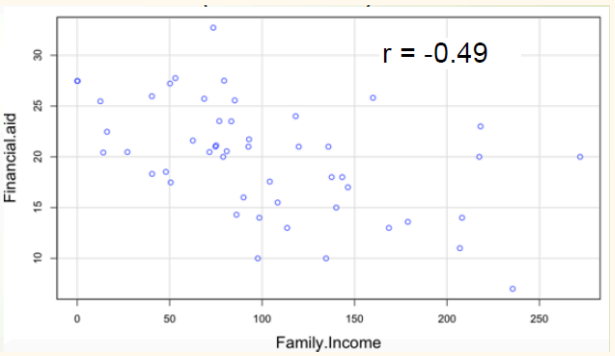


(rule of thumb only)

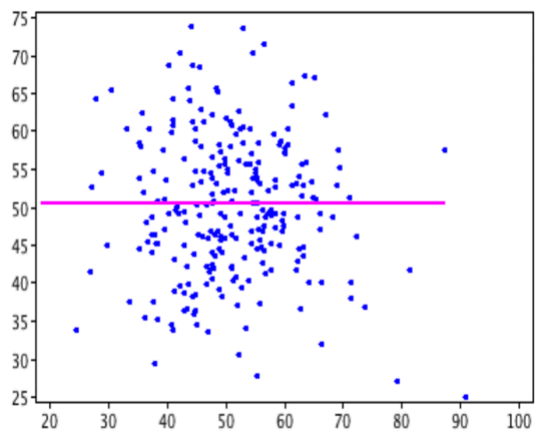
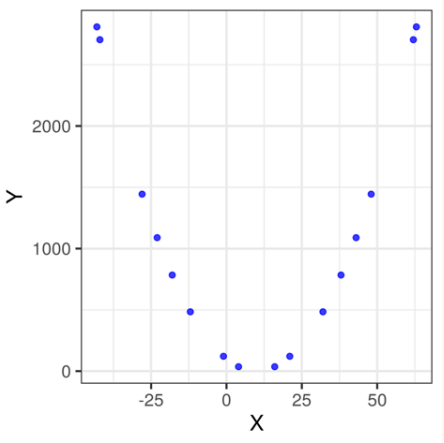
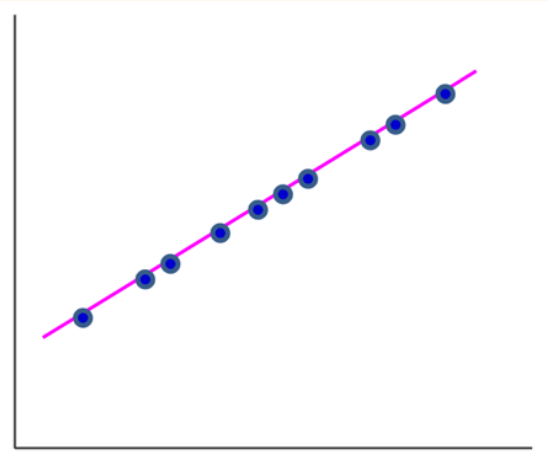
- The sign of r indicates the direction of the linear association.
- The magnitude of r indicates the strength of the linear association between two numerical variables.

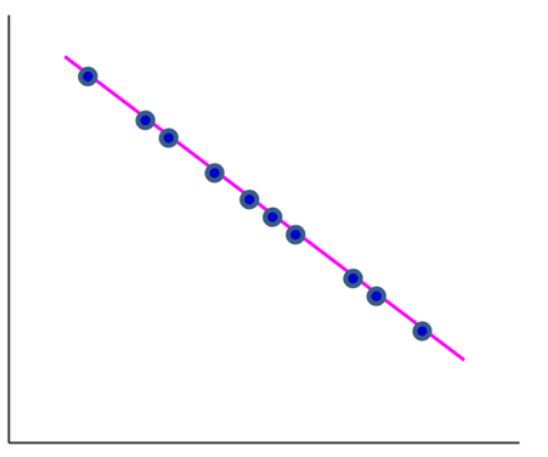


Examples

r	Example 1	Example 2
$r > 0$ positive association		
$r < 0$ negative association		

<table>
<tr>
<td>$r = 0$
no linear association

no linear association doesnt means no association, just that the form is non-linear</td>
<td>  </td>
<td>  </td>
</tr>
<tr>
<td>$r = 1$
perfect positive association</td>
<td>  </td>
<td rowspan="2">If all lies in a line that is not horizontal or vertical then r = 1,-1

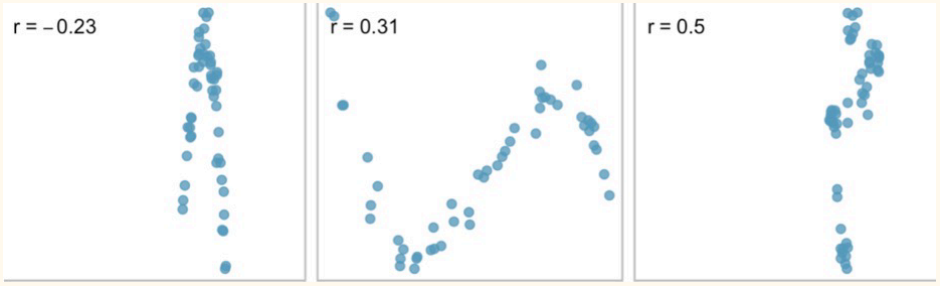
If all lies in a horizontal or vertical line, then r = 0. Because a change in one variable doesn't relate to a change in another variable, thus no association.</td>
</tr>
<tr>
<td>$r = -1$
perfect negative association</td>
<td>  </td>
</tr>
</table>

Limitations of Correlation Coefficient

Association is not causation

- Very often when there is a strong association between two variables, with a correlation coefficient of r that is close to 1 or −1, it is mistakenly concluded that any change in the explanatory variable, say x, will cause the response variable y to change.
- This is incorrect as what we can conclude is only a statistical relationship between x and y and not a causal relationship. There could be confounders.
- Correlation does not imply causation.

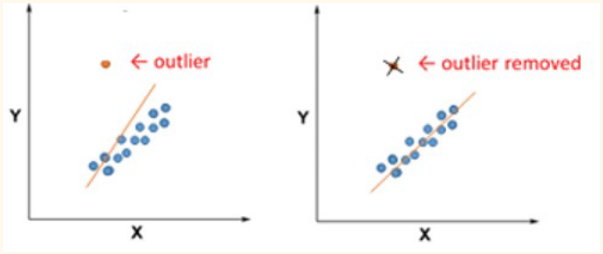
Correlation Coefficient does not tell us anything about non-linear association

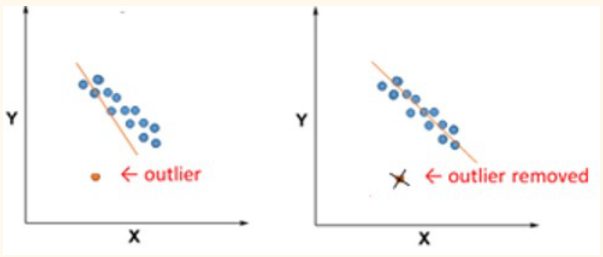


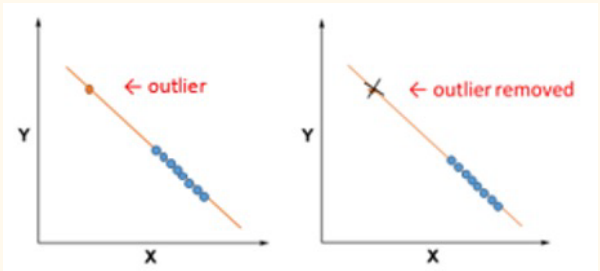
- r is only for linear relationship just because r is small doesn't mean it is a weak linear
- It could be something else so must look at scatter plot

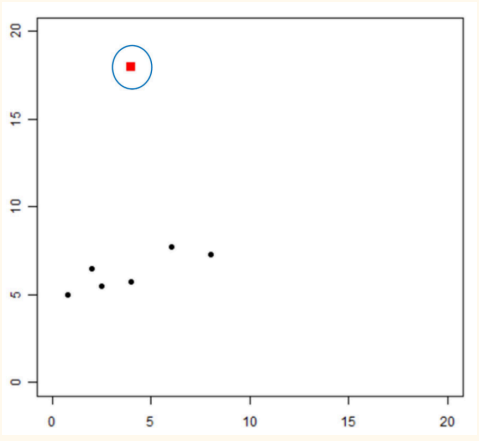
Outliers can affect the correlation coefficient significantly

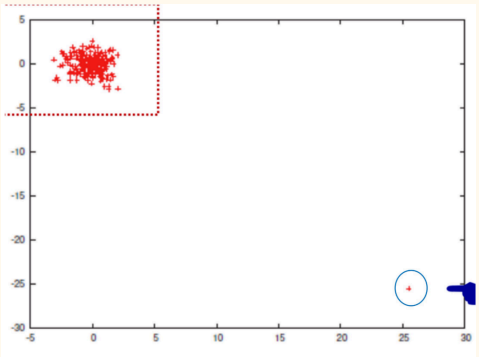
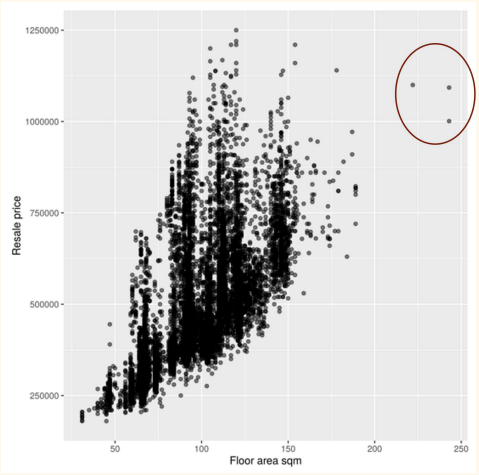
The removal of outliers from a data set can have different effects on the correlation coefficient, depending on how the outlier is positioned in relation to the rest of the data points.







Scatter Plot	r before removal of outlier	r after removal of outlier
	0.22	However, when we remove the outlier, we see that there is a strong positive linear association between the remaining data points. Thus, in this case, the presence of the outlier decreases the strength of the correlation, compared to when the outlier is removed.

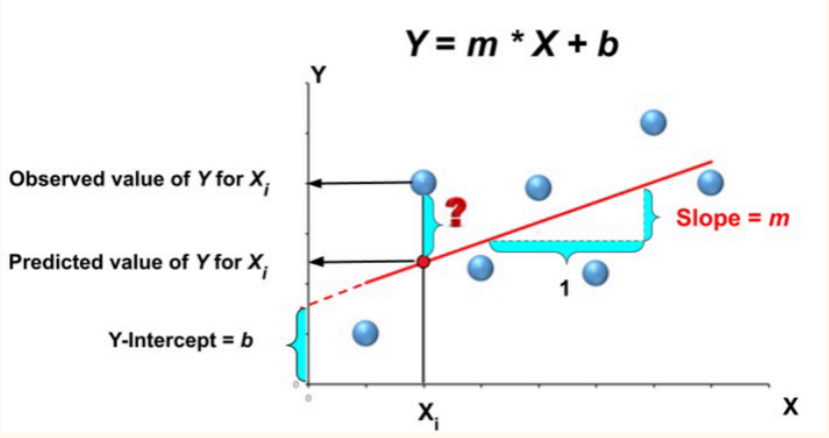
	−0.75	When the outlier is removed, the remaining data points give a correlation coefficient of 0.01. Thus, in this case, the presence of the outlier actually increases the strength of the correlation, compared to when the outlier is removed.
	0.626	After the outliers are removed, the correlation coefficient becomes 0.625, which is practically the same as before.

Chapter 3.4 - Linear regression

- Linear Regression
- How to Use Linear Regression
- Finding Linear Regression for Linear Relationship
- Finding Linear Regression for Non-Linear Relationship

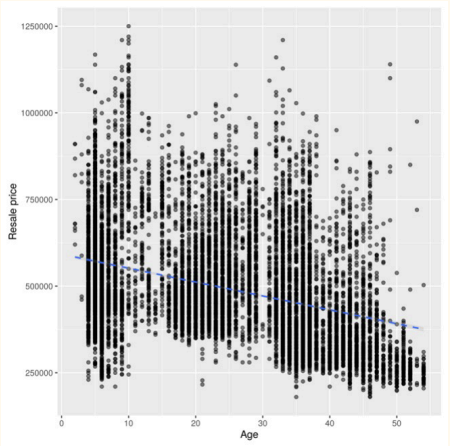
Linear Regression

If we believe that two variables X and Y are linearly associated, we may model the relationship between the two variables by fitting a straight line to the observed data. This approach is known as linear regression.



- In the figure above, the straight line in red is the regression line that is fitted to the observed data, represented by the blue dots.
- Consider the i-th observation (X_i, Y_i). The "?" in the figure represents the residual of the i-th observation, which is the observed value of Y for X_i (that is, Y_i) minus the predicted value of Y for X_i (predicted by the straight line).
- This residual, denoted by e_i, is sometimes also called the error of the i-th observation as it measures how far the predicted value is from the observed value.

How to Use Linear Regression



So the **predicted** resale price of a 40 year old flat is $431,577. It is important to note that we are **not concluding** that

> A 40 year old resale flat will be sold at $431,577.

But instead our linear regression model predicts that

> The average resale price of 40 year old HDB flats is $431,577.

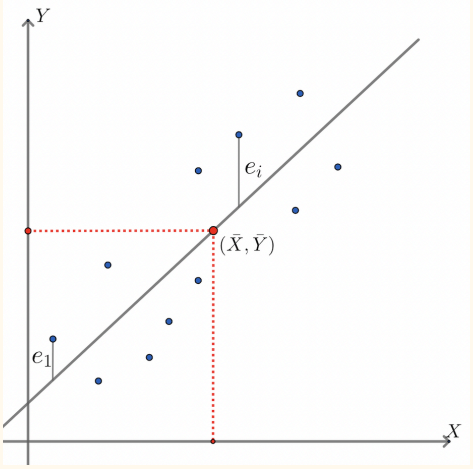
With X representing the age of the resale flat and Y being the resale price, the regression line obtained from the data set is

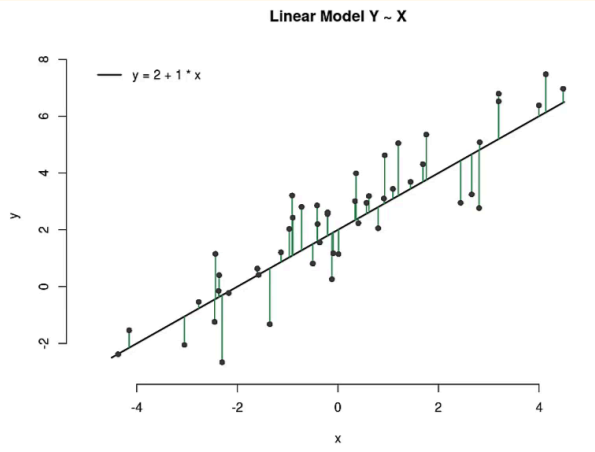
$$Y = -4007X + 591857$$

- The equation can predict the corresponding average Y value for any known X within the range of data used for finding the regression line.
    - cannot be used for extrapolation
    - cannot be used to predict X for known Y
- Furthermore, as the correlation between resale flat price and age of the flat is weak, the prediction obtained from the linear regression above may not be as accurate compared to the scenario where the correlation is stronger.

Finding Linear Regression for Linear Relationship

There are several ways to assess which straight line fits the observed data better. One of the most common ways is the method of least squares.





the method of least squares seek to find a straight line that minimises the overall sum of squares of errors,

$$e_1^2 + e_2^2 + \ldots e_n^2$$

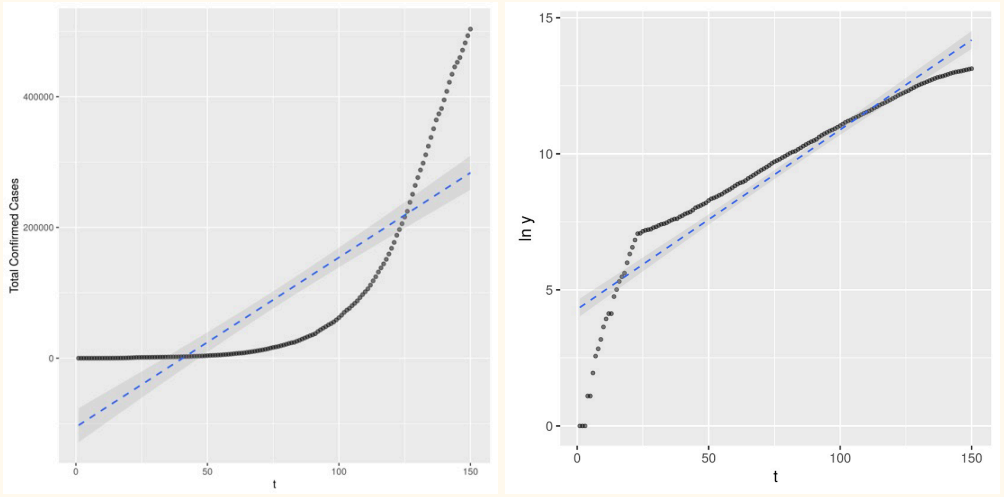
Properties of Linear Regression

- The least squares regression line obtained from a set of observed data points (xi, yi) will always pass through that point of averages for that data set, that is, ($\bar{x}$, $\bar{y}$).
- The least squares regression line can be used to predict the average y for x, the same regression line cannot be used to predict average x for a given y.
    - The reason is essentially because of the way the regression line was obtained.
    - In obtaining the regression line with the independent variable x and the dependent variable y, the line was fitted to minimise the square of error terms between the observed and predicted y.
    - If the intention was to use a given y to predict the average age of x, then have to find another regression line that minimises the square of error terms between the observed and predicted x.
    - The two regression lines are different and thus not interchangeable.
- The correlation coefficient r between the variables X and Y is related to the gradient of the regression line obtained using the method of least squares.

- $m = \frac{s_Y}{s_X} r$

- Another important point to note about the linear regression line obtained using a data set is with regards to the range of the independent variable in the data set. ie. the regression line cannot be used for extrapolation.
  - extrapolation or predicting beyond the data range is dangerous as the best fit regression line may change

Finding Linear Regression for Non-Linear Relationship



if the relationship between y and t is indeed exponential in nature, we can model this relationship using the equation:

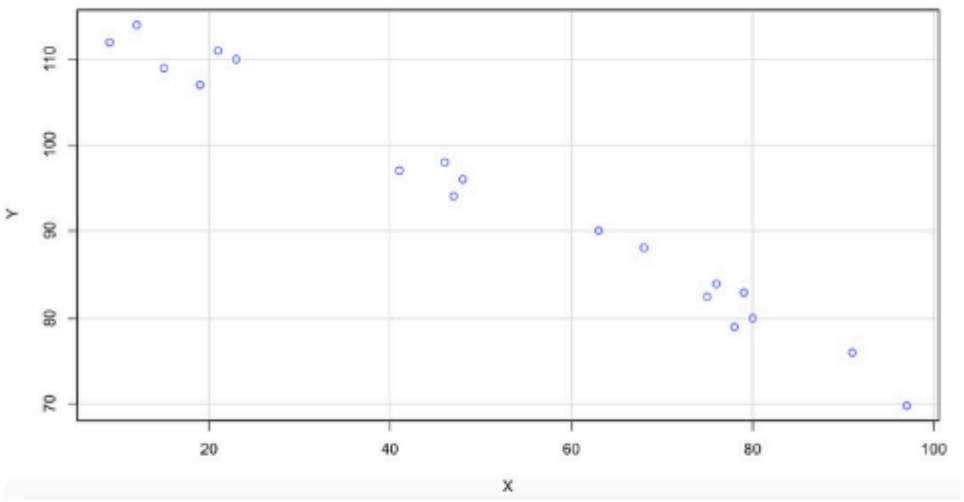
$$y = cb^t$$
$$ln(y) = ln(cb^t)$$
$$ln(y) = ln(c) + tln(b)$$
$$Y = ln(c) + ln(b)X$$

Then scatter plot of ln(y) vs X is plotted and the regression line for ln(y) vs X is found, value of c and b can then be determined:

$$ln(c) = C = y - intercept$$
$$ln(b) = m = gradient$$

Quiz 3

12. What will happen to the correlation coefficient between X and Y if a point with coordinates $(80, 110)$ is added to the scatter plot shown below?



(A) It will increase.

(B) It will decrease.

(C) It will remain the same.

-1 increase to closer to 0, decrease in strength of correlation but increase in the value of correlation coefficient

Beware the magnitude of correlation coefficient gives the strength of correlation. So if the correlation coefficient is negative and the strength of correlation decrease, the correlation coefficient actually increase

25. The relationship between the number of glasses of beer consumed daily (x) and blood alcohol content in percentage (y) was studied in young adults. The equation of the regression line is $y = -0.015 + 0.02x$ for $1 \leq x \leq 10$. The legal limit to drive in Singapore is having a blood alcohol content below 0.08%. Des, a young adult, had just finished 5 glasses of beer. After that, he wanted to take his car out for a drive. Is it legal for him to drive in Singapore?

(A) Yes.
(B) No.
(C) Unable to determine.

need more info on how the study is done? cuz nvr say how long aafter drinking alcohol content is measure, or how fast they drank

the predicted average is higher than the legal limit. but for a particular case for Des, his level might be below the average

Regression line can only be used to predict an average value of y given the x value, so cannot be used to determine with 100% that his blood alcohol content is below the legal limit.

The regression line for **Y** vs **X** is given by **Y** = 0.82**X** + 59.1. The standard deviations for **X** and **Y** are 1.5 and 2.2 respectively. Suppose now we construct a regression line that uses **Y** to predict **X**.

The predicted average increase of **X** when **Y** is increased by 1 unit is _____.

(Give your answer correct to 2 decimal places.)

0.38

given the gradient and standard deviation, the gradient of the other way of prediction can be calculated as the relationship of $m = \frac{s_Y}{s_X} r$ and the fact that r is the same when the axis is swapped.

Chapter 4 - Statistical Inference

- Chapter 4.1 - Probability
- Chapter 4.2 - Conditional Probability and Independence
- Chapter 4.3 - Random Variables
- Chapter 4.4 - Confidence Intervals
- Chapter 4.5 - Hypothesis Testing
- Quiz 4

Statistical inference is the process of drawing conclusions about the population based on the sample data.

- This is used as a census is often impossible.

Chapter 4.1 - Probability

- Terminology
- Probability
- Finite Sample Space
- Uniform probability

Terminology	**Example A: 2 coin tosses**	**Example B: d6 toss**
Probability experiment A probability experiment must be repeatable and allows for the exact listing of all the possible outcomes.	The procedure of tossing the coin twice is called a probability experiment.	The procedure of tossing the die is called a probability experiment.
Sample Space A sample space is the set of all possible outcomes of a probability experiment.	{HH, HT, TH, TT}	{1, 2, 3, 4, 5 ,6}

Event A subset of the sample space is called an event. Note that an outcome can also be considered an event but not all events are outcomes. This is clear as there exist subsets of only one element but not all subsets have only one element.	“two in a row” = {HH, TT} “at least one tail” = {HT, TH, TT} “The first toss is heads, second toss is tails.” = {HT}	“die shows an even number” = {2, 4, 6}
Probability of an event The probability of an event of the sample space is the total probability that the outcome of the experiment is an element of the event. E = event S = sample space where $E \subseteq S$ $P(E) = P(x \in E)$	Probabilities are numerical values between 0 and 1 (both inclusive), so P(E) takes on a numerical value between 0 and 1 and this is the probability assigned to event E.	

Probability

Mathematically, P(E) is defined as the long run proportion of observing E when a large number of repetitions of the experiment is being performed. P(E) = n(E)/N

- The estimate of P(E) we obtain from these N repetitions of the experiment is likely to be different if the experiment is repeated another N times and to get another estimate.
- Such estimates get more accurate and closer to the true value of P(E) as N approaches infinity.
- Thus it is virtually impossible to verify what is the true probability for an event of a probability experiment. But in the analysis of data, it is sufficient to treat the estimates as if it is the true probability.
- It is more important for the to obey the Rules of Probabilities:
    - The probability of each event E, denoted by P(E) is a number between 0 and 1 (inclusive).
    - If the entire sample space is denoted by S, then the probability of S, P(S) is 1.
    - If E and F are mutually exclusive events (meaning both events cannot occur simultaneously), then the probability of E or F occuring is equal to the sum of the probabilities of E and F. That is,
        - if $P(E \cap F) = 0$ then $P(E \cup F) = P(E) + P(F)$

Finite Sample Space

When the sample space contains only a finite number of outcomes, we only need to assign probabilities to the outcomes so that these probabilities sum up to 1. The probabilities of all other events can then be derived from there.

Example 4.1.6 Suppose we have a *biased* six-sided die being rolled once. The following probabilities are assigned to the six possible outcomes.

Outcome	1	2	3	4	5	6
Probability	0.1	0.1	0.1	0.1	0.1	0.5

Check that the probabilities add up to 1. We are now able to derive the probabilities of certain events by applying the third rule of probability as stated above. For example, if E is the event "an odd-numbered face" and F is the event "an even-numbered face", it is easy to see that

1. $P(E)$ is the sum of $P(1)$, $P(3)$ and $P(5)$, so $P(E) = 0.3$. (Here "1", "3", "5" are mutually exclusive events.)
2. $P(F)$ is the sum of $P(2)$, $P(4)$ and $P(6)$, so $P(F) = 0.7$. (Here "2", "4", "6" are mutually exclusive events.)
3. E and F are mutually exclusive events, so $P(E \cup F) = P(E) + P(F) = 0.3 + 0.7 = 1$.

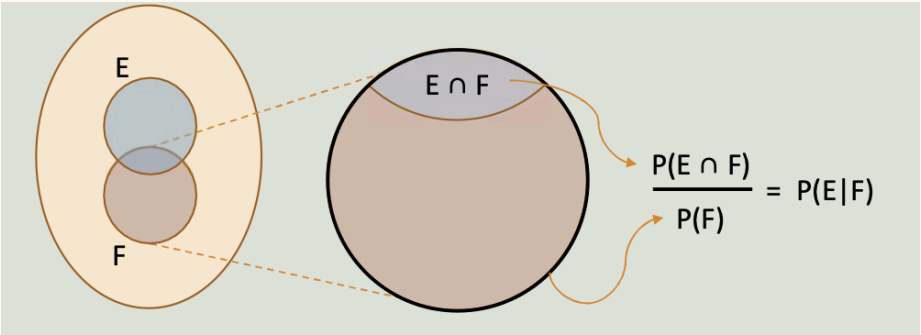
Uniform Probability

Uniform probability is the way of assigning probabilities to outcomes such that equal probability is assigned to every outcome in the finite sample space. Thus, if the sample space contains a total of N different outcomes, then the probability assigned to each outcome is 1/N.

- In SRS a probability experiment with uniform probability is actually being conducted:
  - Where the sample space is the sampling frame that contains all the units that could possibly be selected.
  - The probability of selecting a particular unit at the first draw from the sampling frame is 1/N where N is the size of the sampling frame.
  - For an event denoted by A, the probability of this event, P(A) is interpreted as the likelihood of selecting a unit belonging to A into the sample. This is equal to the rate of A in the sampling frame.
    - P(A) = R(A)

Conditional Probability and Independence

Conditional Probability



The probability of E given F measures how likely the outcome of the probability experiment is an element of E, if we already know that it is an element of F. To compute conditional probabilities, the idea of restricting the sample space based on the condition that event F is known to have occurred is used.

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

- If there is no overlap between events E and F, then given that F occurs, E cannot occur.
  - this agrees with P(E ∩ F) = 0 and thus P(E|F) = 0
- If event F itself cannot occur, P(F) = 0, then by convention P(E|F) = 0
- If event E and F are independent, P(E|F) = P(E), thus P(E ∩ F) = P(E) P(F)

Law of Total Probability

The law of total probability states that if E, F and G are events from the same sample space S such that
1) E and F are mutually exclusive; and

2) E ∪ F = S

Then,

$$P(G) = P(G \mid E) \times P(E) + P(G \mid F) \times P(F)$$

Example 1

500 adults, comprising 280 males and 220 females, as participants in a lucky draw where there is only one prize to be won. What is the probability of Male A being the winner given that the winner is male

$$\begin{aligned} P(Male\ A | Winner\ is\ male) &= P(Male\ A \cap Winner\ is\ male)/P(Winner\ is\ male) \\ &= P(Male\ A)/P(Winner\ is\ male) \\ &= (1/500)/(280/500) \\ &= 1/280 \end{aligned}$$

Example 2

Suppose there are two bags, each containing 10 coloured balls. Bag A contains 7 red balls and 3 green balls while Bag B contains 4 red balls and 6 green balls. One bag is randomly selected and a ball is then randomly selected from the chosen bag. What is the probability that the selected ball chosen is green?

Let us consider the events A, B and G such that

- A is the event that Bag A is selected. B is the event that Bag B is selected.
  - Event A and event B are mutually exclusive
- G is the event that the selected ball is green.

$$P(G) = P(G \cap A) + P(G \cap B)$$

Sub in conditional probability

$$= P(G|A) * P(A) + P(G|B) * P(B)$$

This is the law of total probability

Rates and Probability

Random sampling	Corresponds to	Probability experiment
Sampling frame	Corresponds to	Sample space
A subgroup A of the sampling frame	Corresponds to	An event A of the sample space
The rate of A, rate(A)	Corresponds to	The probability of A, $P(A)$

Conditional rates also correspond to conditional probability.

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} && \text{(by using the idea of restricted sample space)} \\ &= \frac{\text{rate}(A \cap B)}{\text{rate}(B)} && \text{(by the correspondence between probability and rates)} \\ &= \frac{\text{size of } (A \cap B)}{\text{size of sampling frame}} \div \frac{\text{size of } B}{\text{size of sampling frame}} \\ &\quad \text{(by the definition of rates as ratios of two sizes)} \\ &= \frac{\text{size of } (A \cap B)}{\text{size of } B} \\ &= \text{rate}(A \mid B) && \text{(by the definition of rates as ratios of two sizes)} \end{aligned}$$

Covid Test Example

For most medical diagnostic tests, there are four possible scenarios that can happen when the test is administered to an individual to assess if the individual is infected. The possible scenarios are:

1. Scenario 1: Individual is known to be infected and test shows positive.
2. Scenario 2: Individual is known to be infected and test shows negative.
3. Scenario 3: Individual is known to be not infected and test shows positive.
4. Scenario 4: Individual is known to be not infected and test shows negative.

Scenario 1 is concerned with the conditional probability of an individual being tested positive, given that the individual is infected. This is known as the true positive rate. P(Test positive | Individual is infected) is known as the sensitivity of the test.

Scenario 4 is concerned with the conditional probability of an individual being tested negative, given that the individual is not infected. This is known as the true negative rate. P(Test negative | Individual is not infected) is known as the specificity of the test.

The better the test the closer the sensitivity and selection is to 1.

In reality, these two conditional probabilities are not helpful to average users to know whether they are indeed infected or not. However, it is known with certainty whether the individual's test returns positive or negative. Therefore, instead of the conditional probability

P(Test positive | Individual is infected)

which is difficult to ascertain if the "condition" is fulfilled, we look at the conditional probability

P(Individual is infected | Test positive)

It is important to gain insight into this conditional probability as it can cause an individual much distress after being tested positive only to find out later that the person involved is actually not infected.

To determine this conditional probability, having only the sensitivity and specificity of the test is insufficient. One additional piece of information is required, which is the base rate of infection in the population. This is the infection rate in the population and we can interpret this as the probability a person selected at random from the population is infected with COVID-19.

Known values for calculation:

- P(Test positive | Individual is infected) = 0.8
- P(Test negative | Individual is not infected) = 0.99
- P(Individual is infected) = 0.01

Can use a contingency table with a nice, round and big number for total. Highlighted are the numbers needed. All the other ones are chosen/derived from those given. This means using a contingency table, can find the other probabilities.

	Tested Positive	Tested negative	Row total
Infected	1000*0.8 = 800	1000*0.2 = 200	100000*0.01 = 1000
Not infected	99000*0.01 = 990	99000*0.99 = 98010	100000*0.99 = 99000
Column total	1790	98210	100000

Therefore, P(Individual is infected | Test positive) = 800/1790 = 0.447 (rounded to 3 significant figures)

- This conditional probability is rather low so typically, more rigorous tests need to be conducted to ascertain if the individual is indeed infected with COVID-19.

Additionally, P(Individual is infected | Test negative) = 200/98210 = 0.002 (rounded to 3 significant figures)

Independence

When two events A and B are independent, it means that the probability of A is the same as the probability of A given B. So, the fact that event B has occurred does not affect the probability of A occurring.

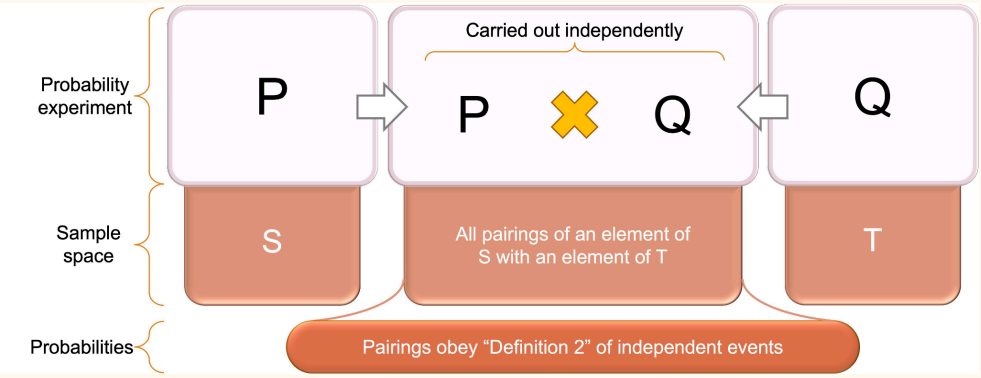
$P(A) = P(A \mid B)$ Definition 1 of independent events

Since $P(A \mid B) = P(A \cap B) / P(B)$

$P(A \cap B) = P(A)\,P(B)$ Definition 2 of independent events

Since rates are corresponding to probabilities, $P(A) = P(A \mid B) \Leftrightarrow R(A) = R(A \mid B)$. Independent events would imply that the events are not associated with each other.

Independent Probability Experiment



- When 2 independent probability experiments are carried out and coupled together, it just means that the sample space of the overall experiment would contain all the pairing of elements in both the sample space.
- Thus obeying the definition 2 of independent events => $P(A \cap B) = P(A)\,P(B)$

Mutually Exclusive

$P(A \cup B) = P(A) + P(B)$

if $P(A) + P(B) > 1$ then there must be overlap in the event A and B therefore not mutually exclusive

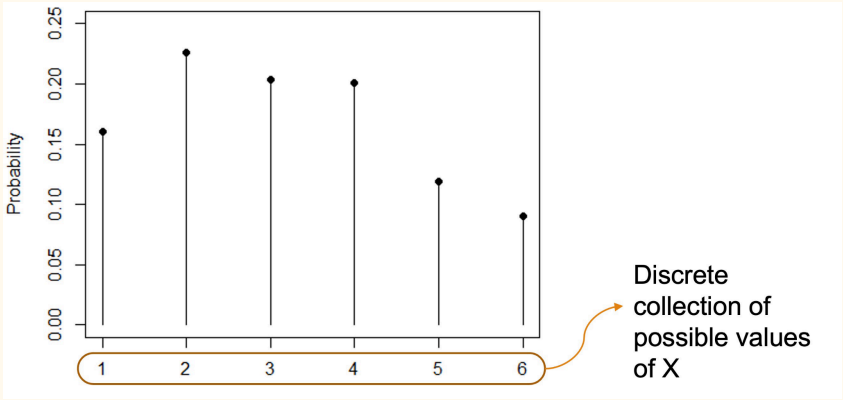
Random Variables

A random variable is a numerical variable with probabilities assigned to each of the possible numerical values taken by the numerical variable.

Random variables were conceived as a mathematical way to model data distributions. if the distributions are similar.

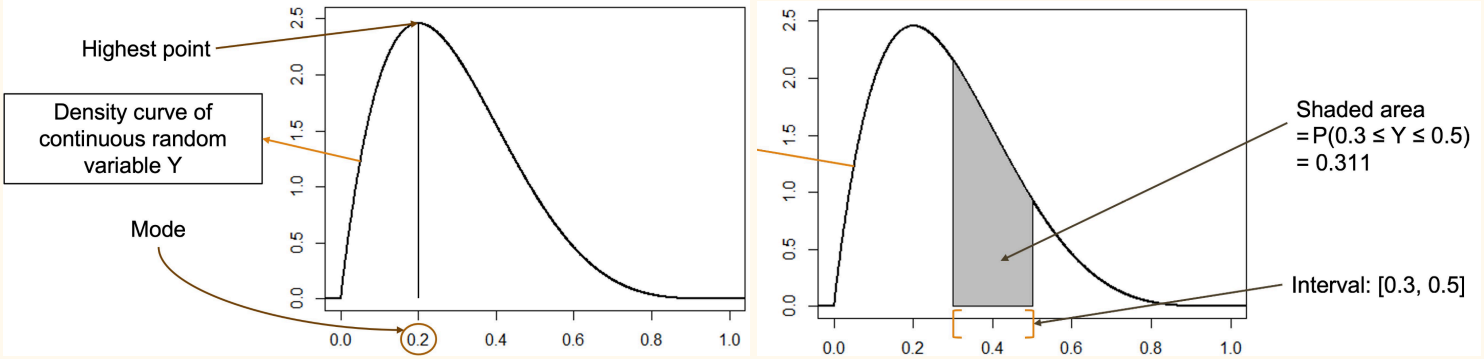
Discrete Random Variable

Household size	1	2	3	4	5	6
Probability	0.16	0.226	0.204	0.201	0.119	0.09



- Each point represents the possible value of x and the height represents the probability of x taking that value.
- The gaps between each point is an indication that the random variable is discrete.
- Sum of the heights = sum of the probability = 1
- Mode = highest point
- Since each point/each result of x is mutually exclusive the probability of x taking on a range of values = sum of the probability of x taking each of the discrete values that falls within the range.
    - $P(X > 4) = P(X=5) + P(X=6) = 0.119 + 0.09 = 0.209$

Continuous Random Variable



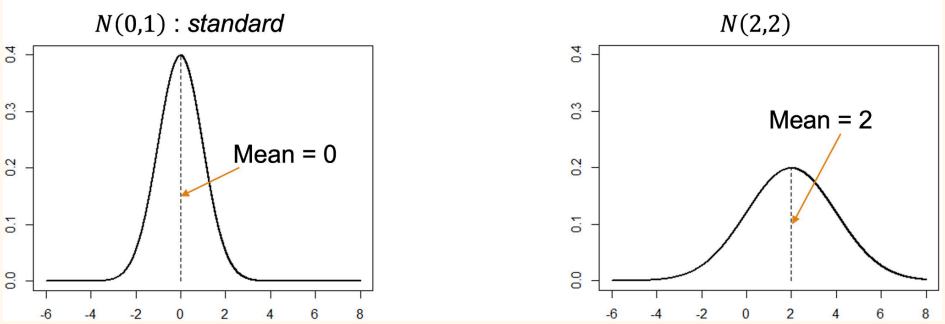
- The curve is continuous, thus the values x can take is continuous
- However, it should be noted that the y-axis in the density curve does not represent probability (unlike for discrete random variables), but instead it is the probability density. For the purpose of this module, we will not go into details of the interpretation of probability density. Nevertheless, it is important to remember that for a continuous random variable, the area under the density curve is always equal to 1.
- Mode = highest point
- $P(X > X1 \cap X < X2)$ = area under the probability density curve between X1 and X2

Normal Distribution

Normal distributions are a class of continuous random variables that is denoted by its mean x and variance y.

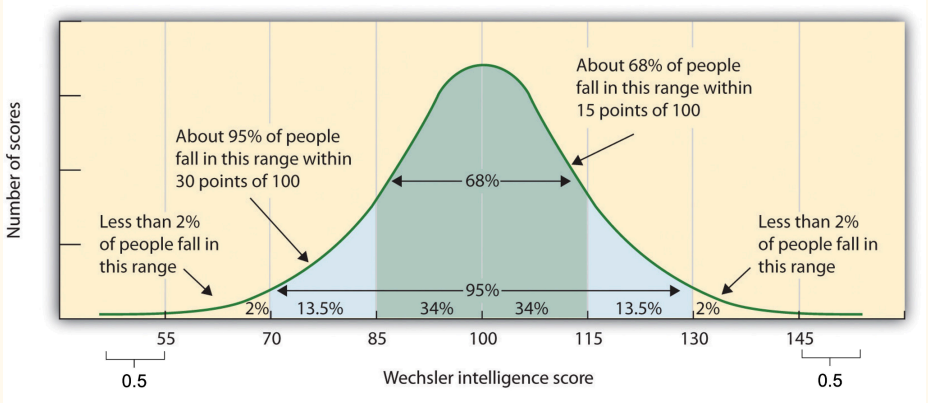
N(x, y)

1. A particular normal distribution is completely described by its mean and variance. Therefore, any two normal distributions can only differ by their means and/or variance.
2. The density curve of a continuous random variable that is normally distributed is always bell shaped.
3. The peak of the curve occurs at the mean. This implies that the mode is equal to the mean.
4. The density curve is symmetrical about the mean. This implies that the median is also equal to the mean.
5. Thus, the mean, mode and median of any normal distribution are the same.



- The density curve on the left is that of the normal distribution N(0, 1), the mean (as well as the mode and median) occurs at 0 and the variance, which measures the spread of the distribution is 1. The normal distribution N(0, 1) is commonly known as the standard normal distribution.
- The density curve on the right is that of the normal distribution N(2, 2). Here, the mean occurs at 2 and the variance is 2. Note that the spread of the distribution is larger than that for N(0, 1) and thus the curve is flatter with a lower peak. This is a consequence of the required property that for any continuous random variable, the area under the density curve must always be equal to 1.

Example of normal distribution: Wechsler Adult Intelligence Scale



Confidence Intervals

- Confidence Intervals
- Confidence Intervals for proportion
- Confidence Intervals for mean
- Interpretation of Confidence Intervals
- Properties of Confidence Intervals

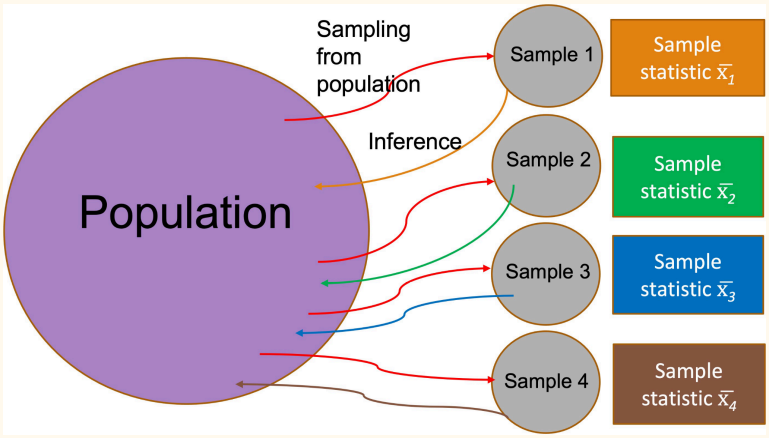
Confidence Intervals

Estimating population parameters from sample statistics will have some inaccuracy that's due to bias and random error.

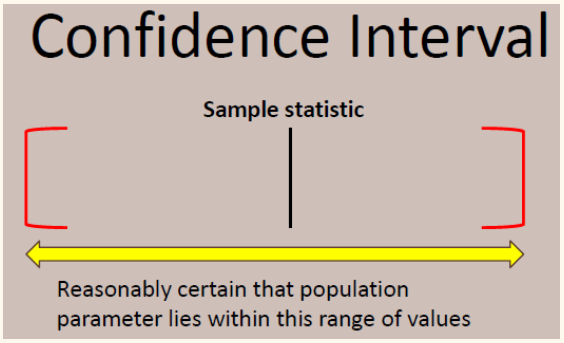
Sample statistic = population parameter + bias + random error

By using a good sampling method, i.e. using random sampling, and practise like having a good sampling frame, bias can be reduced to an insignificant level. For GEA1000, can assume samples are simple random samples taken from a perfect sampling frame with 100% response rate, so no selection bias and nonresponse bias. Thus,

Sample statistic = population parameter + random error
Sample statistic = what we know
population parameter = what we want to estimate
random error = what we don't know



Since the sample statistic will always have random error, different samples will likely have different sample statistics, thus a confidence interval is needed to have a range of values that we are reasonably certain the population parameter lies in.



A confidence interval is a range of values that is likely to contain a population parameter based on a certain degree of confidence. This degree of confidence is called the confidence level and is usually expressed as a percentage (%).

- Confidence intervals can be constructed for proportion, mean and standard deviation. For GEA1000, only CF for proportion and mean is thought.
- Confidence intervals are a way to quantify random error that is present in every sample.

Confidence Intervals for proportion

$$c\,\%\ CI\ for\ p\ = p^*\ \pm\ z^*\ \sqrt{\frac{p^*(1-p^*)}{n}}$$

$c\ =\ confidence\ level\ for\ the\ confidence\ interval$
$p\ =\ population\ proportion$
$p^*\ =\ sample\ proportion$
$z^*\ =\ "z\ value"\ from\ standard\ normal\ distribution\ depends\ on\ the\ confidence\ level,\ c.\ (get\ from\ software)$
$n\ =\ sample\ size$

Stand normal distribution

Standard Normal Probabilities

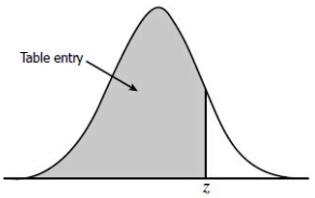


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Confidence Intervals for mean

$$c\,\%\ CI\ for\ \mu = \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

$c = confidence\ level\ for\ the\ confidence\ interval$

$\mu = population\ mean$

$\bar{x} = sample\ mean$

$t^* = "t\ value"\ from\ t\ distribution\ depends\ on\ the\ sample\ size,\ n,\ and\ confidence\ level,\ c.\ (get\ from\ software)$

$s = sample\ standard\ deviation$

$n = sample\ size$

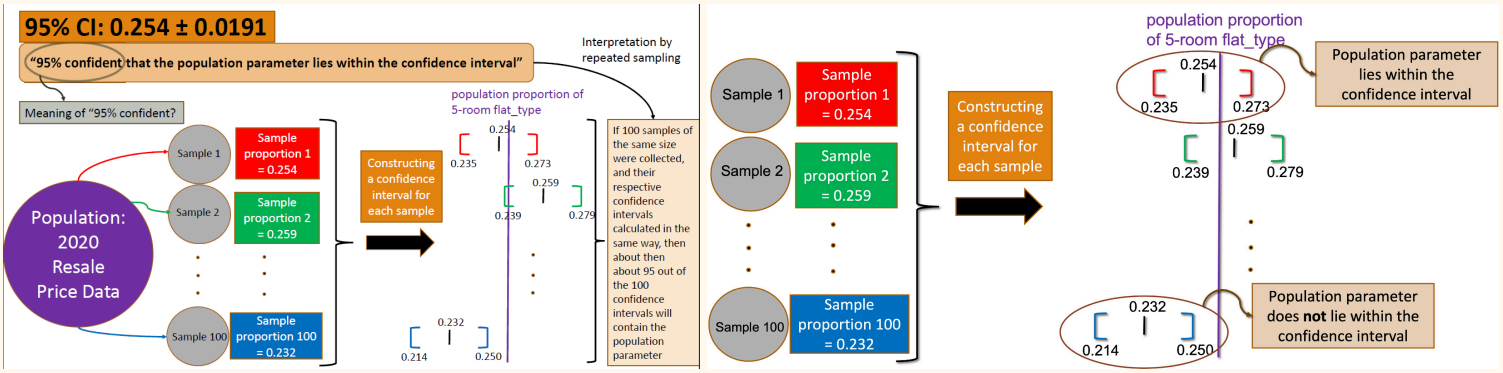
The t-curve has a similar shape as the z-curve but is shorter and wider. t-value changes with sample size.

Interpretation of Confidence Intervals

c % CI: x ± e

- is interpreted as c % confident that the population parameter lies within x - e and x + e.
- e is known as the margin of error which directly impacts the width of the interval.

Meaning of Confidence Level



Confidence level can be explained by "repeated sampling":

- If repeated sampling of the same size is taken and the confidence interval of each sample is constructed in the same way and with the same confidence level, then about 95% of the confidence intervals constructed would contain the population parameter.
- It is important to remember that in actual fact, we do not know what is the exact value of the population parameter. Confidence intervals certainly give us a better idea of where this parameter lies but they can never tell us its exact value.
- For any given sample and its confidence interval, there is no way of knowing whether the parameter is inside the interval or not.

- It is a common mistake to say that there is a 95% chance that the population parameter falls within the interval.
    - The population parameter is a fixed constant.
    - For any given sample, CI only depends on the sample statistics and confidence level chosen, Thus, CI is also fixed.
    - Thus there is no element of chance in the fixed value and range, the value is either in the range or its not.
    - The element of chance comes from the uncertainty in the random sampling rather than the uncertainty in the value of the population parameter. Thus the idea of CI is from the repeated sampling.
    - This why the word “confidence” is used when talking about CI
        - c % confident that the population parameter lies within x - e and x + e.
    - Given any interval there isn't a chance of the population parameter falling within it, either in or out. But any random sample and its interval would have a 95% chance that the population parameter falls within the interval.

21. A 95% confidence interval, constructed from a random sample, for the population mean number of children per household in Country Z is (1.21, 4.67). Which of the following statements is/are true? Select all that apply.

(A) The probability that the population mean number of children per household in Country Z lies between 1.21 and 4.67 is 0.95.

B and E

(B) We are 95% confident that the population mean number of children per household in Country Z lies between 1.21 and 4.67.

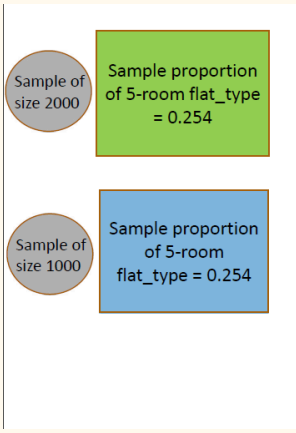
(C) 95% of all samples of the same size and sampling procedure should have sample mean number of children per household between 1.21 and 4.67.

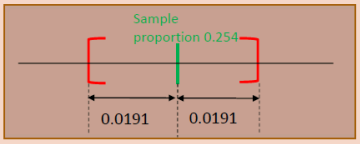
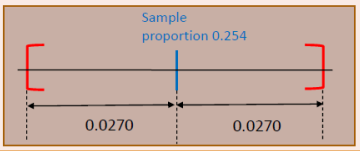
(D) 95% of all households in Country Z have between 1.21 and 4.67 children.

(E) If we take 100 different samples of the same size using the same sampling procedure and compute the confidence interval for each sample in the same way, approximately 95 of the intervals will contain the true population mean.

Properties of Confidence Intervals

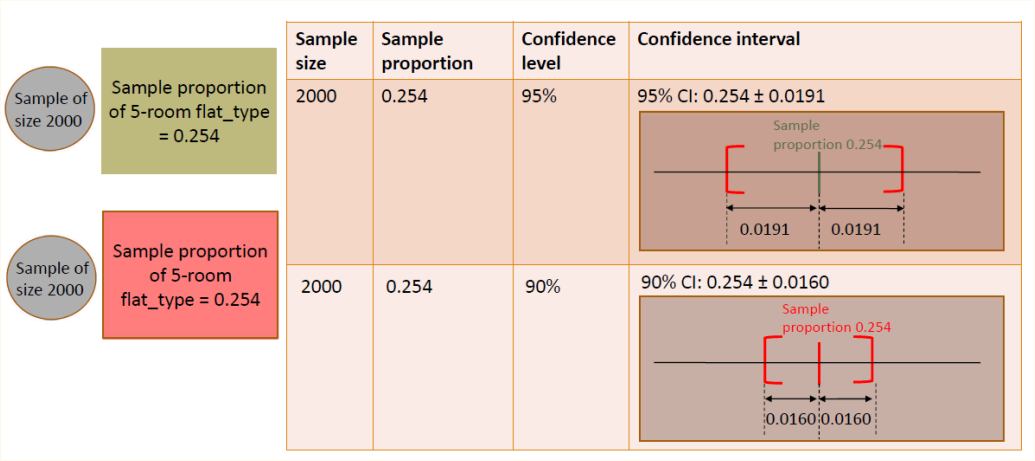
Sample Size and CI



Sample size	Sample proportion	Confidence level	Confidence interval
2000	0.254	95%	95% CI: 0.254 ± 0.0191 
1000	0.254	95%	95% CI: 0.254 ± 0.270 

- Confidence interval has to do with the random error when taking samples and bigger sample has smaller random error thus CI from bigger sample size has a smaller margin of error when compared to CI of smaller sample size but same confidence level.
- This is apparent when looking at the formulas also.

Confidence Level and CI



Sample size	Sample proportion	Confidence level	Confidence interval
2000	0.254	95%	95% CI: 0.254 ± 0.0191
2000	0.254	90%	90% CI: 0.254 ± 0.0160

- CI of lower confidence level has lower z-value/t-value so width of interval is smaller, this is from the point of view of the formula.
- Using the idea of repeated sampling, having a narrower interval would imply that a smaller percentage of repeated samples would contain the population parameter. Thus, less confident that the population parameter falls within this interval is smaller.

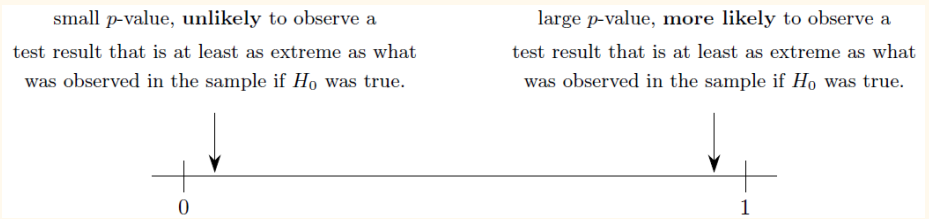
Hypothesis Testing

Used to answer yes/no questions regarding the population property from the sample.

- Hypothesis tests are only done when only the sample data obtained and not information on the entire population.
- In the unlikely event that the entire population data is obtained, all can be determined and there is no need for any hypothesis test.

Four steps of hypothesis testing

1. **Identify the question and state the null hypothesis and alternative hypothesis.**
    a. The null hypothesis usually asserts the stand of no effect or no difference. Indirectly, this means that whatever differences or variances that are observed in the sample data is not inherent in the population and had occurred by random chance when we were choosing the sampling units.
    b. The alternative hypothesis, on the other hand, is typically what we wish to confirm and pit against the null hypothesis. In many research questions, we often hope that the sample data provides sufficient evidence for us to reject the null hypothesis in favour of the alternative hypothesis.
    c. It is important to note that the null hypothesis and the alternative hypothesis must be mutually exclusive, meaning that they cannot be true simultaneously.
2. **Collection of relevant data that is necessary for the test.**
    a. The process of testing a hypothesis usually involves a random variable and its probability distribution.
    b. So the p-value calculated is from a random variable model assuming H_0 is true. The value observed/computed from the sample is called the test statistics.
3. **The significance level of a hypothesis test is chosen and p-value if calculated.**
    a. The significance level of a hypothesis test is always a number between 0 and 1.
    b. The significant level can be thought of as how “convincing” the evidence has to be before the null hypothesis is rejected in favour of the alternative hypothesis.



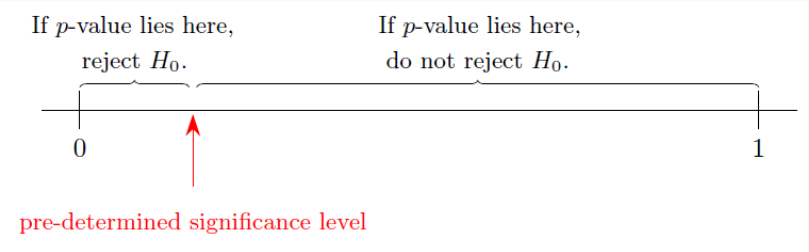
A commonly used level of significance is 0.05 (also referred to as 5% level of significance). Other levels used frequently are 0.10 (10% level of significance) and 0.01 (1% level of significance).

    c. The p-value is the probability of obtaining a test result at least as extreme as the result observed (favours the alternative hypothesis at least as much as what is observed in the current sample), assuming the null hypothesis is true.

p-value = P(Obtaining a result at least as extreme as observed | H_0 is true)

This is finding out how likely the extremity observed will occur when we model the population using a random variable assuming H_0 is true.

4. **Compare p-value and significance level and conclude on whether to reject or not reject the null hypothesis.**



Only two possible conclusion:

a. p-value < significance level
   i. The null hypothesis is rejected in favour of the alternative/Reject H_0
b. p-value ≥ significance level
   i. The null hypothesis is not rejected, Do not reject H_0
   ii. This does not mean that H0 is true.
   iii. This means that the test is inconclusive and we do not know if the observation is due to chance or not and nothing further can be said about the null hypothesis based on this test.

It is important to note that the following are NEVER conclusions from any hypothesis test:

- "Accept H0.": We never accept the null hypothesis. Our test only attempts to reject this hypothesis based on the observation in the sample.
- "Reject H1." We should also avoid saying that the alternative hypothesis is rejected when the p-value is greater than the significance level. It is more appropriate to say that based on the observation in the sample, there is insufficient evidence to support the alternative hypothesis, at the significance level that we have set for our test.

One should remember that the significance level for a hypothesis test is decided before the p-value is calculated and not the other way round.

<u>Example: Checking if coin is biassed</u>

Suppose we believe that a particular coin is "loaded" and thus biassed towards one of the two sides, say heads. This means that any toss of the coin will more likely show heads rather than tails.

Step 1: H_0 & H_1

Null hypothesis H_0 : "The coin is fair" OR P(H) = 0.5 where H is the event of observing heads in a toss.
Alternative hypothesis H_1 : "The coin is biassed towards heads" OR P(H) > 0.5

Step 2: Collection of relevant data

The coin is tossed several times (say 8 times) and the outcomes (heads or tails) of the tosses are observed.
Let X be the discrete random variable 0-8 = number of heads

Step 3: Significance level of a hypothesis test is chosen and p-value if calculated.

For demonstration, let's choose **:**

5% significance level and the test result in 8 tosses 7 are heads

p-value = P(Obtaining a result at least as extreme as observed | H_0 is true)
= P(X ≥ 7 | H_0 is true)
= P(X = 7 | H_0 is true) + P(X = 8 | H_0 is true)
= $8C1\ (1/2)(1/2)^7 + (1/2)^8$
= 9/256
= 0.03515625

Suppose in 8 tosses 3 are heads, then

p-value = P(Obtaining a result at least as extreme as observed | H_0 is true)
= P(X ≥ 3 | H_0 is true)
= P(X = 3 | H_0 is true) + P(X = 4 | H_0 is true) + … + P(X = 8 | H_0 is true)
= 1 - P(X = 0 | H_0 is true) + P(X = 1 | H_0 is true) + P(X = 2 | H_0 is true)

Suppose the alternative hypothesis H_1 is "The coin is biassed towards tails", P(H) < 0.5 instead, and in 8 tosses 3 are heads, then
p-value = P(Obtaining a result at least as extreme as observed | H_0 is true)
= P(Number of tails more than or equal to 5 | H_0 is true)
= P(Number of heads less than or equal to 3 | H_0 is true)
= P(X ≤ 3 | H_0 is true)
= P(X = 0 | H_0 is true) + P(X = 1 | H_0 is true) + P(X = 2 | H_0 is true) + P(X = 2 | H_0 is true)

The p-values are different despite having the same observation of 3 heads out of 8 tosses. The reason for this difference lies in the difference in the alternative hypothesis.

Step 4: Conclusion

Since p-value = 0.03515625 < 0.05 = significance level, there is sufficient evidence supporting H_1 and that the coin is biassed towards heads, therefore H_0 is rejected.

Suppose 3% significance level is chosen instead, then 0.03515625 > 0.03 there is insufficient evidence supporting H_1 and that the coin is biassed towards heads, therefore H_0 is rejected.

One-sample t-test and Chi-squared test

One-sample t-test	Chi-squared test
Mainly used to test difference between sample mean $\bar{x}$ and a known or hypothesised population mean μ. H_0: μ = 600000 H_1: μ > 600000	Mainly used to test for association between two categorical variables at the population level. H_0: Household size is not associated with flat value in Singapore H_1: Household size is associated with flat value in Singapore
Population distribution should be approximately or assumed to be normal if sample size is smaller than 30.	Data required for the test is the count for the categories of a categorical variable.
Data used should be acquired via random sampling.	Data used should be acquired via random sampling.

*technicalities involved in t-test and chi squared test are beyond the scope of this module. In reading material the test is run by radiant.

Quiz 4

21. A 95% confidence interval, constructed from a random sample, for the population mean number of children per household in Country Z is (1.21, 4.67). Which of the following statements is/are true? Select all that apply.

(A) The probability that the population mean number of children per household in Country Z lies between 1.21 and 4.67 is 0.95.

B and E

(B) We are 95% confident that the population mean number of children per household in Country Z lies between 1.21 and 4.67.

(C) 95% of all samples of the same size and sampling procedure should have sample mean number of children per household between 1.21 and 4.67.

(D) 95% of all households in Country Z have between 1.21 and 4.67 children.

(E) If we take 100 different samples of the same size using the same sampling procedure and compute the confidence interval for each sample in the same way, approximately 95 of the intervals will contain the true population mean.

27. A researcher is interested to know if smoking and heart disease are associated with each other in the population of Singapore. The researcher carries out a census of the population with a 100% response rate. The researcher conducts a chi-squared test on the census data at 5% significance level and obtains a p-value of 0.001. Which of the following is a valid conclusion?

(A) Since p-value is less than 0.05, the null hypothesis is **rejected** at the 5% significance level, and the researcher concludes that smoking and heart disease **are associated** with each other in the population.

(B) Since p-value is less than 0.05, the null hypothesis is **rejected** at the 5% significance level, and the researcher concludes that smoking and heart disease **are not associated** with each other in the population.

(C) Since p-value is less than 0.05, the null hypothesis is **not rejected** at the 5% significance level, and the researcher concludes that smoking and heart disease **are associated** with each other in the population.

(D) Since p-value is less than 0.05, the null hypothesis is **not rejected** at the 5% significance level, and the researcher concludes that smoking and heart disease **are not associated** with each other in the population.

(E) None of the other options is a valid conclusion.

Answer is (E). The researcher took a census of the population, not a sample of the population. There is no estimation involved in this scenario, only a measurement of the population, so hypothesis testing is irrelevant. Hypothesis tests regarding statements about the population should be conducted on probability-based samples only.

If the population data is obtained there is no need for doing hypothesis testing

28. A group of students wants to find out if there is any association between staying in a hall and being late for class in NUS in a particular month. If students are late for at least 5 classes, they are considered "late for class" in that month. After collecting a random sample of 1000 students, they found that 200 out of 350 students who stay in a hall are late for class, while 390 out of 650 students who do not stay in a hall are late for class.

A chi-squared test was done to test for association between staying in a hall and being late for class at 5% level of significance. The p-value derived from the chi-squared test is 0.3809.

Which of the following statements is/are true? Select all that apply.

(A) There is a positive association between staying in a hall and being late at the sample level.

(B) There is a negative association between staying in a hall and being late at the sample level.

(C) Since the p-value is more than 0.05, we can conclude that there is an association between staying in a hall and being late at the population level.

(D) Since the p-value is more than 0.05, we cannot conclude that there is an association between staying in a hall and being late at the population level.

(B) and (D) are correct. To check for association between staying in hall and being late at the sample level, we compare

$$\begin{aligned} &\text{rate(Late for class} \mid \text{Staying in hall)} = \frac{200}{350} \\ < \ &\frac{390}{650} = \text{rate(Late for class} \mid \text{Not staying in hall)} \end{aligned}$$

Hence, we see that staying in hall is negatively associated with being late at the sample level. As we are doing a chi-squared test to see if there is an association at the population level, recall that the null hypothesis states that there is no association at the population level, while the alternative hypothesis states that there is an association at the population level. Since the p-value is greater than the level of significance, we do not reject the null hypothesis. Hence, we cannot conclude that there is an association between staying in hall and being late at the population level.

Sample conclusion from doing rates and population conclusion from doing chi-square test are different

There are 5 students in a GEA1000 project group. Assume that birthdays are equally likely to occur on each day of the year, a year consists of 365 days, and that students in the project group have birthdays independent of one another. What is the probability (to 3 significant figures) that at least 2 students have the same birthday?

- 0.0137
- 0.000325
- 0.00274
- 0.0271.

P(>= 2 same) = P(2 same) + P(3 same) + P(4 same) + P(5 same) = 1 - P(All diff)
= 1 - 364/365 - 363/365 - 362/365 - 361/365
= 0.0271

Quiz 1 Stuff

Professor Lim would like to find out if including a peer review component would affect students' final grades. He decided to get a sample of the students in his tutorial classes and place them into 2 groups. He assigned his Monday, Tuesday and Wednesday morning classes into the 'assessment with peer review' group and his Wednesday afternoon, Thursday and Friday classes in the 'assessment without peer review' group.

Which of the following best describes the type of sampling employed?

- None of the other options.
- Systematic sampling.
- Volunteer sampling.
- Cluster sampling.

Answer = none of the other options, because never say about sampling?
Probability sampling will require deliberate use of chance in the sampling process. In this case, the assignment of individuals has been pre-determined by the Professor. Within the types of non-probability sampling methods, this is not an example of volunteer sampling, as all students from both sub-groups were selected by the Professor to do the study and not self-selected.

May, an owner of a tuition center, wishes to find out if using iPads during tuition class improves her students' academic performance. She decided to conduct an experiment as follows:

1. She groups all the students in her center according to the day they come for tuition. For simplicity's sake, we can assume each student only goes for tuition once per week, there is at least one class of tuition every day in her center, and no student drops out halfway.

2. Every student who goes for tuition on weekends will be given an iPad to use during class. The students who go for tuition on weekdays will not be given an iPad.

3. She then keeps track of all her students' academic performance for the next 6 months.

Which of the following statements is/are true?

(I) She used a probability sampling method.

(II) This is a controlled experiment without random assignment.

- ○ Only (I).
- ○ Only (II).
- ○ Neither (I) nor (II).
- ◉ Both (I) and (II).

Statement (I) is incorrect. Probability is not used in the selection of students into treatment/control. In fact, a census, not sampling, is conducted in this case. Statement (II) is correct. The students who go for tuition on weekends will be in the treatment group, and those who go on weekdays will be in the control group. There is no random assignment involved here.

Suppose you are a researcher who is interested in drawing a simple random sample of 200 people from a population of 5000 individuals. Which of the following would be a correct approach? Select all that apply.

- ☐ Sort the names of the entire population by alphabetical order (A to Z) and place the names in a list. Select the people whose names appear at the top 200 of the list.
- ☐ Write all the names of the entire population on equal-sized pieces of paper, mix the papers in a box and draw out 200 pieces of paper at one go. Choose the people whose names appear on the drawn papers.
- ☑ Assign each individual in the population a unique integer from 1 to 5000 by random assignment. Then choose the people assigned numbers 4801 to 5000.

Can take a bunch all at once out also

A multiple-choice mid-term examination was conducted for 2000 students in a General Education module GEB1000. There were 20 questions. Students were awarded 1 mark for each correct answer and received 0 mark for any wrong answer. There was no partial credit awarded for all questions. A teaching assistant helped with the collation of the scores of the paper, and provided the following summary statistics:

- Minimum = 2.0
- 1st Quartile = 7.5
- Median = 11.5
- Mean = 9.0
- Mode = 12.0
- 3rd Quartile = 13.2
- Maximum = 20.0

Which of the following statements is/are true? Select all that apply.

- [ ] None of the other statements is true.
- [ ] The 3rd Quartile is incorrect.
- [ ] Based on the above information, we can conclude that the coefficient of variation is 2.
- [x] Based on the above information, we can conclude that the range is 18.0.

Q3 is wrong because the quartiles of integers cannot have .2, at most is .5

A researcher wants to know the average weight of year 1 students in University A. The researcher does not have access to such information, hence he decided to do a survey.

All University A year 1 students have to take a compulsory module in the first semester of their studies, and hence have to be present for an in-person examination on 24th April at 1pm. The researcher stood outside the examination venue's only exit with a weighing scale and waited for the examination to end.

There were too many students for the researcher to weigh. Hence, to decide whom to weigh, while students were exiting, the researcher used a random integer generator to produce a random integer for each student. If the random integer was even, the researcher will measure the student's weight. If the random integer was odd, the researcher will not measure the student's weight. Assume that all students exited the venue orderly in a line and were compliant with the researcher. There were 800 students exiting the venue, and the random integer generator produced 200 even numbers and thus 200 students were weighed.

What is an/are issue(s) that the study is likely to face? Select all that apply.

- [ ] Bias present due to the random integer generator producing unlikely random numbers.
- [ ] Bias present due to the non-probability sampling method.
- [ ] Bias present due to a poor sampling frame chosen.
- [x] None of the other options.

Each student had a ½ chance of being chosen, it is random sampling.

Question 7 1 / 1 pts

To study the overall satisfaction levels of students staying in a college, a researcher obtained a simple random sample of 100 students from the list of all students staying in the college and sent out a survey form to these 100 students. Which of the following must be true about the study? Select all that apply.

- [x] There will not be any selection bias in this study.
- [ ] The sample is representative of all the students staying in the college.
- [ ] There will not be any non-response bias in this study.
- [ ] The results obtained from the sample cannot be generalised to all the students staying in the college.

Must be true about the study
to be a good representation, must have large enough sample size, but since population variability and size is unknown, it can't be said that sample must be representative of the population

cannot conclude that the result must cannot be generalised to all students, as if sample size is big enough and non response rate is low, then it is generalisable, thus not a must.

Quiz 2

d. Upon completing parts (b) and (c) of this question, your friend Tammy told you that the answer to part (c) could have been deduced from (b). She explains that sex must be a confounder in the study because it is associated with whether a student is enrolled in a STEM first degree, which is then associated with the year of enrolment. Hence, sex is associated with both a student's year of enrolment and whether a student is enrolled in a STEM first degree. Based on Tammy's logic alone, is Tammy's claim necessarily correct?

For sex to be a confounder it must be proven that it is associated with both STEM and year directly, thus Tammy is a dummy. From her statement, it can only be said that STEM is the confounder when studying association between sex and year.

In general, association between categorical variables is not transitive: X is associated with Y and Y is associated with Z does not mean that X is associated with Z.

A study was conducted among 100 subjects to see if Covid-infection status was associated with increased self-reported loneliness (indicated as Yes/No) during lockdown restrictions. The results are summarised in the table below.

	Yes		No	
	Female	Male	Female	Male
Infected	25	25	3	10
Non-infected	10	12	8	7

Which of the following statements is/are correct? Select all that apply.

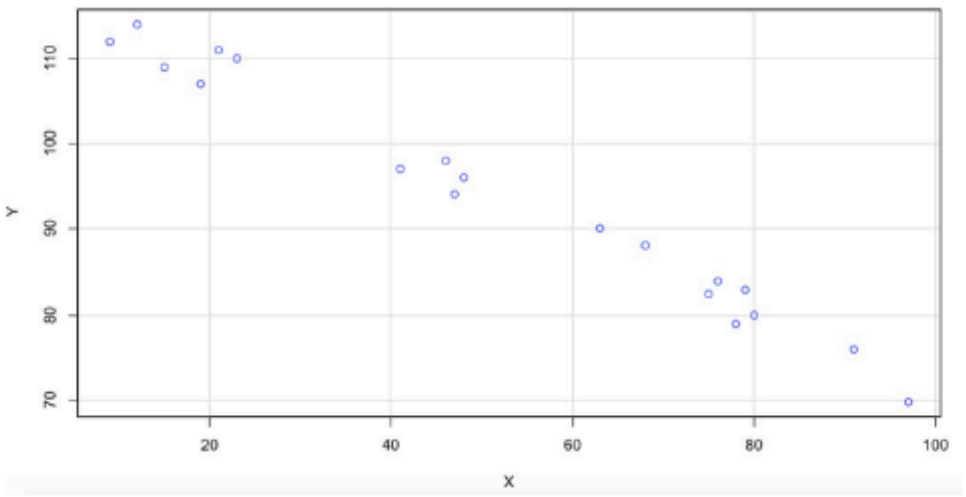
- [ ] This is an experimental study.
- [ ] Simpson's paradox is observed when examining the association between infection status and increased self-reported loneliness.
- [ ] There is a positive association between being female and being Covid-infected.
- [ ] Gender is a confounder in the association between infection status and increased self-reported loneliness.
- [x] Overall, there is a positive association between being Covid-infected and having increased self-reported loneliness.

Just because no simpson's paradox doesn't mean no confounder. gender is a confounder because female and infection is -ve associated and infected and loneliness is positively associated.

Quiz 3

12. What will happen to the correlation coefficient between X and Y if a point with coordinates $(80, 110)$ is added to the scatter plot shown below?

Exercise 3 105



(A) It will increase.
(B) It will decrease.
(C) It will remain the same.

-1 increase to closer to 0, decrease in strength of correlation but increase in the value of correlation coefficient

Beware the magnitude of correlation coefficient gives the strength of correlation. So if the correlation coefficient is negative and the strength of correlation decrease, the correlation coefficient actually increase

25. The relationship between the number of glasses of beer consumed daily (x) and blood alcohol content in percentage (y) was studied in young adults. The equation of the regression line is $y = -0.015 + 0.02x$ for $1 \leq x \leq 10$. The legal limit to drive in Singapore is having a blood alcohol content below 0.08%. Des, a young adult, had just finished 5 glasses of beer. After that, he wanted to take his car out for a drive. Is it legal for him to drive in Singapore?

(A) Yes.
(B) No.
(C) Unable to determine.

need more info on how the study is done? cuz nvr say how long aafter drinking alcohol content is measure, or how fast they drank

the predicted average is higher than the legal limit. but for a particular case for Des, his level might be below the average

Regression line can only be used to predict an average value of y given the x value, so cannot be used to determine with 100% that his blood alcohol content is below the legal limit.

The regression line for *Y* vs *X* is given by *Y* = 0.82*X* + 59.1. The standard deviations for *X* and *Y* are 1.5 and 2.2 respectively. Suppose now we construct a regression line that uses *Y* to predict *X*.

The predicted average increase of *X* when *Y* is increased by 1 unit is _____.

(Give your answer correct to 2 decimal places.)

0.38

given the gradient and standard deviation, the gradient of the other way of prediction can be calculated as the relationship of m = sYsXr and the fact that r is the same when the axis is swapped.

Quiz 4

21. A 95% confidence interval, constructed from a random sample, for the population mean number of children per household in Country Z is (1.21, 4.67). Which of the following statements is/are true? Select all that apply.

(A) The probability that the population mean number of children per household in Country Z lies between 1.21 and 4.67 is 0.95.
(B) We are 95% confident that the population mean number of children per household in Country Z lies between 1.21 and 4.67.
(C) 95% of all samples of the same size and sampling procedure should have sample mean number of children per household between 1.21 and 4.67.
(D) 95% of all households in Country Z have between 1.21 and 4.67 children.
(E) If we take 100 different samples of the same size using the same sampling procedure and compute the confidence interval for each sample in the same way, approximately 95 of the intervals will contain the true population mean.

B and E

27. A researcher is interested to know if smoking and heart disease are associated with each other in the population of Singapore. The researcher carries out a census of the population with a 100% response rate. The researcher conducts a chi-squared test on the census data at 5% significance level and obtains a p-value of 0.001. Which of the following is a valid conclusion?

(A) Since p-value is less than 0.05, the null hypothesis is **rejected** at the 5% significance level, and the researcher concludes that smoking and heart disease **are associated** with each other in the population.

(B) Since p-value is less than 0.05, the null hypothesis is **rejected** at the 5% significance level, and the researcher concludes that smoking and heart disease **are not associated** with each other in the population.

(C) Since p-value is less than 0.05, the null hypothesis is **not rejected** at the 5% significance level, and the researcher concludes that smoking and heart disease **are associated** with each other in the population.

(D) Since p-value is less than 0.05, the null hypothesis is **not rejected** at the 5% significance level, and the researcher concludes that smoking and heart disease **are not associated** with each other in the population.

(E) None of the other options is a valid conclusion.

Answer is (E). The researcher took a census of the population, not a sample of the population. There is no estimation involved in this scenario, only a measurement of the population, so hypothesis testing is irrelevant. Hypothesis tests regarding statements about the population should be conducted on probability-based samples only.

If the population data is obtained there is no need for doing hypothesis testing

28. A group of students wants to find out if there is any association between staying in a hall and being late for class in NUS in a particular month. If students are late for at least 5 classes, they are considered "late for class" in that month. After collecting a random sample of 1000 students, they found that 200 out of 350 students who stay in a hall are late for class, while 390 out of 650 students who do not stay in a hall are late for class.

A chi-squared test was done to test for association between staying in a hall and being late for class at 5% level of significance. The p-value derived from the chi-squared test is 0.3809.

Which of the following statements is/are true? Select all that apply.

(A) There is a positive association between staying in a hall and being late at the sample level.

(B) There is a negative association between staying in a hall and being late at the sample level.

(C) Since the p-value is more than 0.05, we can conclude that there is an association between staying in a hall and being late at the population level.

(D) Since the p-value is more than 0.05, we cannot conclude that there is an association between staying in a hall and being late at the population level.

(B) and (D) are correct. To check for association between staying in hall and being late at the sample level, we compare

$$\begin{aligned} &\text{rate(Late for class} \mid \text{Staying in hall)} = \frac{200}{350} \\ < \ &\frac{390}{650} = \text{rate(Late for class} \mid \text{Not staying in hall)} \end{aligned}$$

Hence, we see that staying in hall is negatively associated with being late at the sample level. As we are doing a chi-squared test to see if there is an association at the population level, recall that the null hypothesis states that there is no association at the population level, while the alternative hypothesis states that there is an association at the population level. Since the p-value is greater than the level of significance, we do not reject the null hypothesis. Hence, we cannot conclude that there is an association between staying in hall and being late at the population level.

Sample conclusion from doing rates and population conclusion from doing chi-square test are different

There are 5 students in a GEA1000 project group. Assume that birthdays are equally likely to occur on each day of the year, a year consists of 365 days, and that students in the project group have birthdays independent of one another. What is the probability (to 3 significant figures) that at least 2 students have the same birthday?

- 0.0137
- 0.000325
- 0.00274
- 0.0271.

P(>= 2 same) = P(2 same) + P(3 same) + P(4 same) + P(5 same) = 1 - P(All diff)
= 1 - 364/365 * 363/365 * 362/365 * 361/365
= 0.0271

---

26. Consider a study that intends to examine whether the colour red makes children act impulsively. A group of 500 children were assigned into two groups by the expert opinion of a child psychologist; group Red if the psychologist pointed to the child, and group Green if the psychologist did not. Each child is then led into a room that has a big button in the colour of their group and labelled "DO NOT PRESS ME!". It is then recorded whether the child presses the button within 10 minutes. All the children were each then given a candy for participating.

    The children were also asked if they like candies. The following table summarises the data. For instance, 22 children from group Red that pressed the button do not like candies.

Exercise 2 93

	Like candies		Does not like candies	
	Red	Green	Red	Green
Pressed button	3	135	22	1
Did not press button	177	60	38	64

Is liking candy a confounder in this study?

(A) Yes.

(B) No.

(C) There is insufficient information given to determine whether liking candy is a confounder in this study.

Answer is (B).

19. A nutritionist wants to estimate the average mass of a particular tortilla chip brand, Naritos. He decides to collect a sample of 100 packets of Naritos chips, using the method described as follows: He went down to the nearest factory that produces Naritos chips and measured the average mass of the first 100 packets of Naritos chips produced on that day. You can assume the mass of the packaging is negligible.

Using the formula

$$\overline{x} \pm t^* \times \frac{s}{\sqrt{n}},$$

he obtained a 95% confidence interval for the population mean as [676.3, 723.7]. Which of the following statements must be true?

(I) If we collect 100 samples of 100 packets of Naritos chips using the same sampling method, about 95% of the samples will have confidence intervals containing the population average mass of Naritos chips.

(II) We can conclude that there is a 95% chance that the population average mass of Naritos chips lies within [676.3, 723.7].

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

Answer is (C).

i is false because sampling is not random thus confidence interval is valid, not random cuz first 100 packet and nearest factory

4. How does "forgiveness" (being forgiving) and empathy go together? The study of Toussaint and Webb on 45 men and 82 women are summarised in the following hypothetical tables:

Distribution of 45 men

	Empathy	No empathy	Row total
Forgiving	10	10	20
Not forgiving	9	16	25
Column total	19	26	45

Distribution of 82 women

	Empathy	No empathy	Row total
Forgiving	30	31	61
Not forgiving	12	9	21
Column total	42	40	82

Which of the following statements is/are true?

(I) Forgiveness and empathy are positively associated among men.

(II) Forgiveness and empathy are positively associated among women.

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

Answer is (A). Among men,

$$\text{rate}(\text{Empathy} \mid \text{Forgiving}) = \frac{10}{20} = 0.5,$$

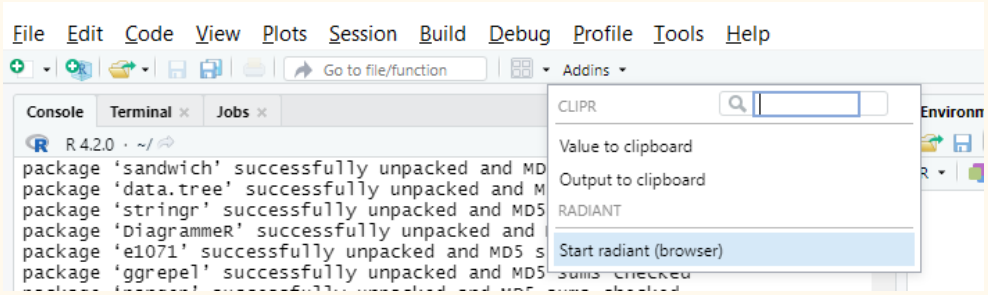
$$\text{rate}(\text{Empathy} \mid \text{Not forgiving}) = \frac{9}{25} = 0.36.$$

So there is positive association between forgiveness and empathy among men. Among women, the corresponding rates are 0.49 and 0.57, so forgiveness and empathy are negatively associated.

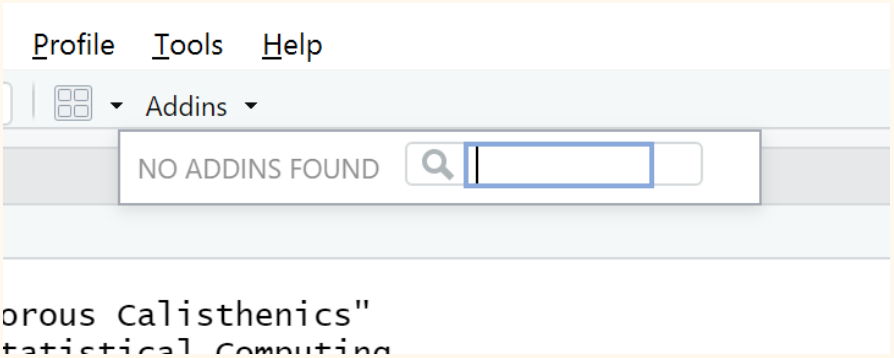
Radiant

Launching Radiant

1. launch r studio
2. then activate radiant in r studio
3. press add-ins then radiant

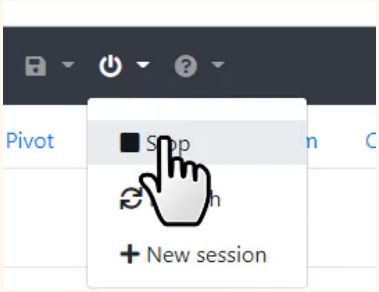


4. If radiant doesn't appear



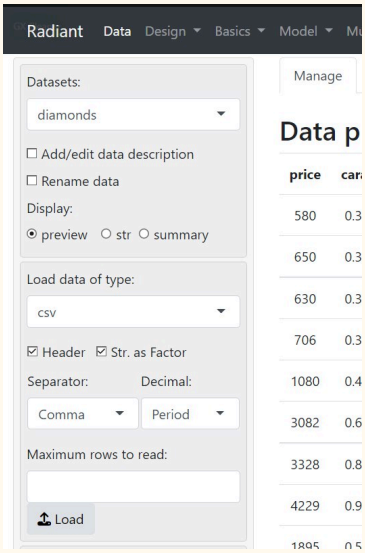
5. type in the command line radiant::radiant()
6. If that still does not work, type this in command line options(repos = c(RSM = "https://radiant-rstats.github.io/minicran", CRAN = "https://cloud.r-project.org")) install.packages("radiant")

Ending Radiant



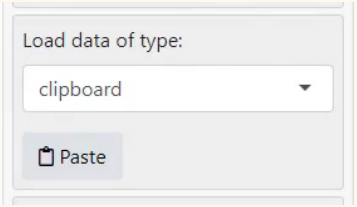
Loading data to radiant

For CSV files:

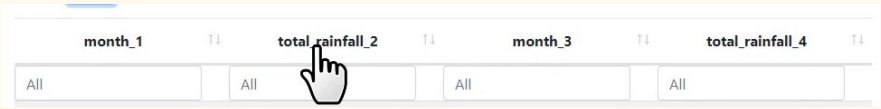


- Under data > manage tab
- Change data type to csv
- Find in folder to upload

For Excel Files:



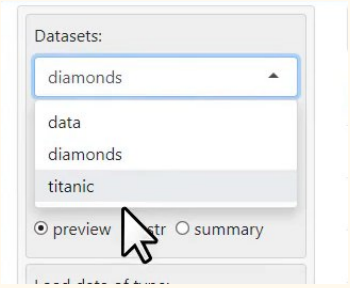
- Go excel copy entire sheet
- Choose clipboard under load data type and paste
- OR
- Save excel in .csv file first, this is more feasible for big data sets



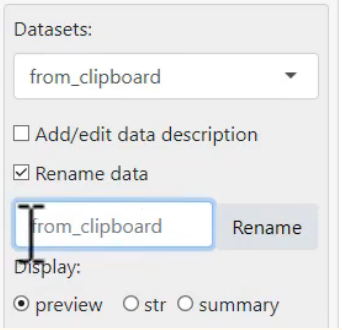
If there are duplicated header/variable names then Radiant will append number behind it

- can always change the variable name under transom and rename

Accessing Datasets

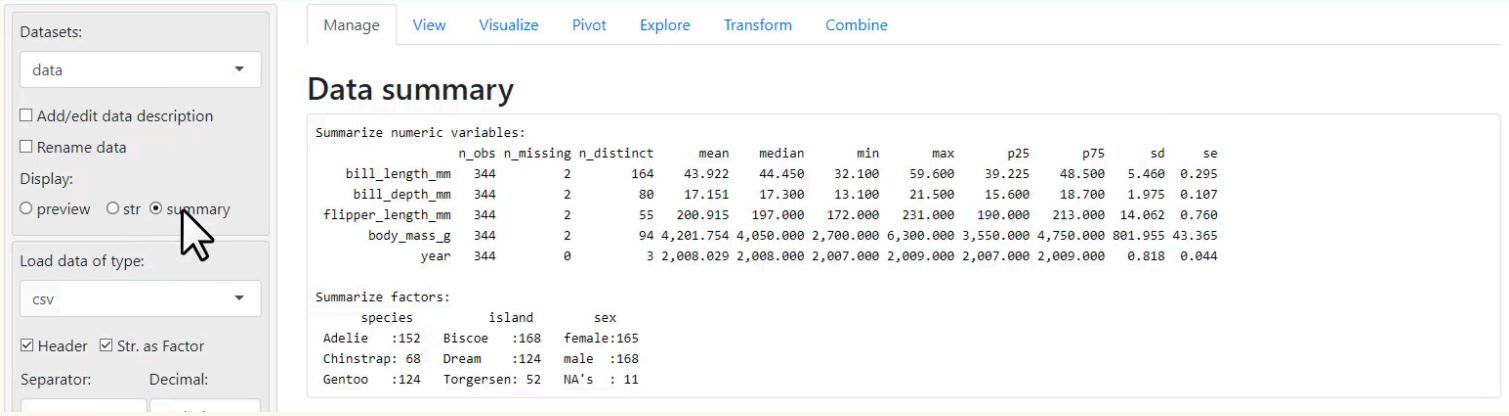


- Datasets uploaded will appear here
- diamonds and titanic are datasets that are included by default



- Datasets can be renamed
- It's a good practise to rename it to avoid confusion

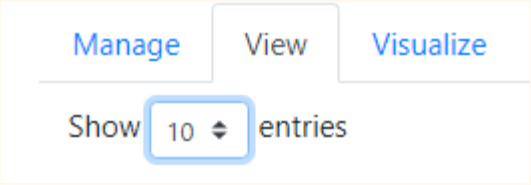
Summary



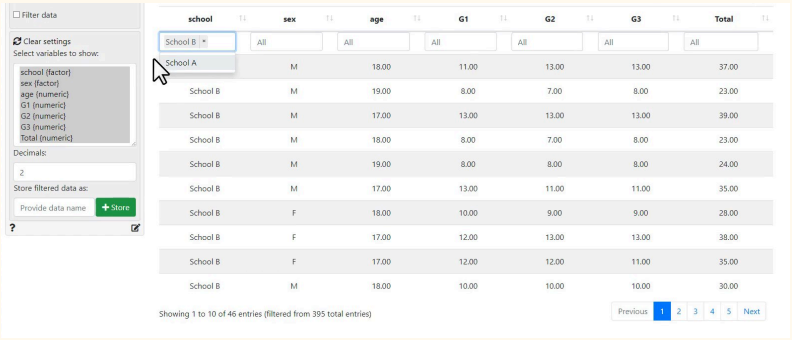
- Clicking summary will give summary statistic for the numerical variable like mean and sd
- And counts for category for categorical data
- Categorical data are known as factors in radiant

View

Under the view tab, users can select how many rows/entries to view. Do not be confused when there are less rows than expected due to this.

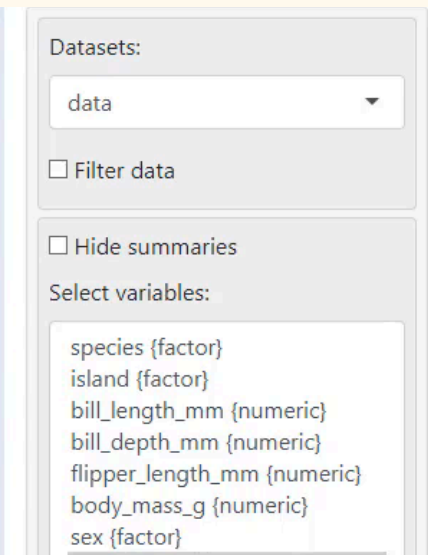


Filtering can be done in the view tab under the headers, save the filtered data in a new dataset. New dataset then can do other stuff like plot graphs

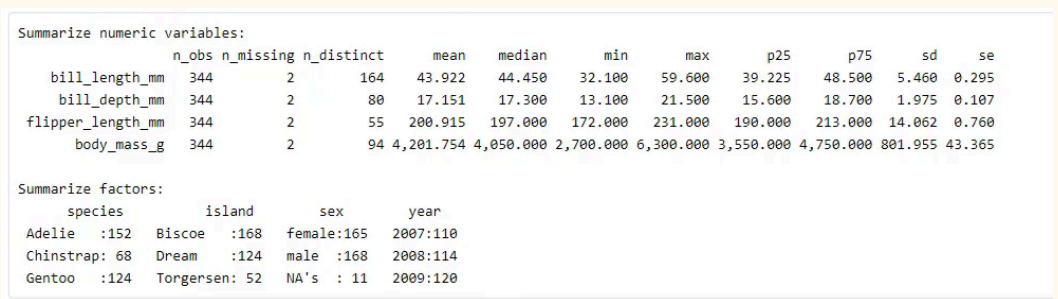


Transform

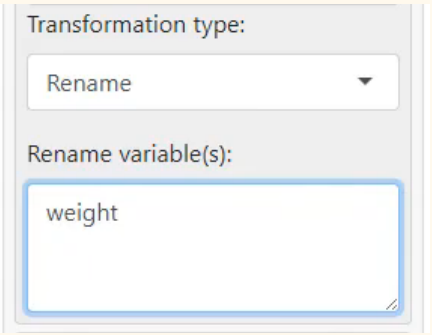
From the screen shot above you can see that year is treated as numerical data type, and mean of year is calculated which dont make sense, it should be treated as categorical data type instead. this can be changed under the transform tab.



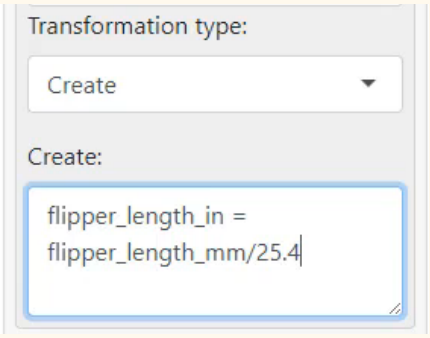
- Transform > select the variable, year
- Transformation type > change type
- Change variable type > change to “As factor” meaning categorical type
- Press store to save changes
- Then it'll change automatically



Similarly can rename the variable by choosing rename

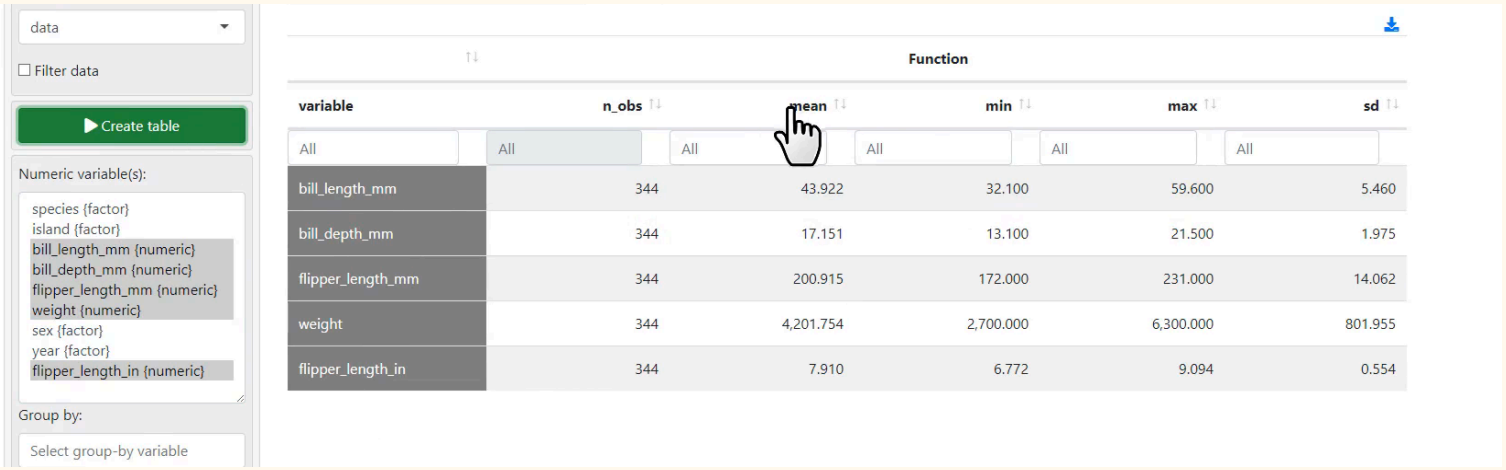


And create new variable, can be used to change unit

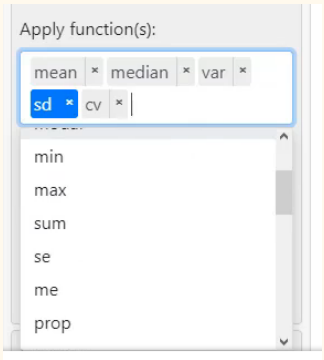


Explore

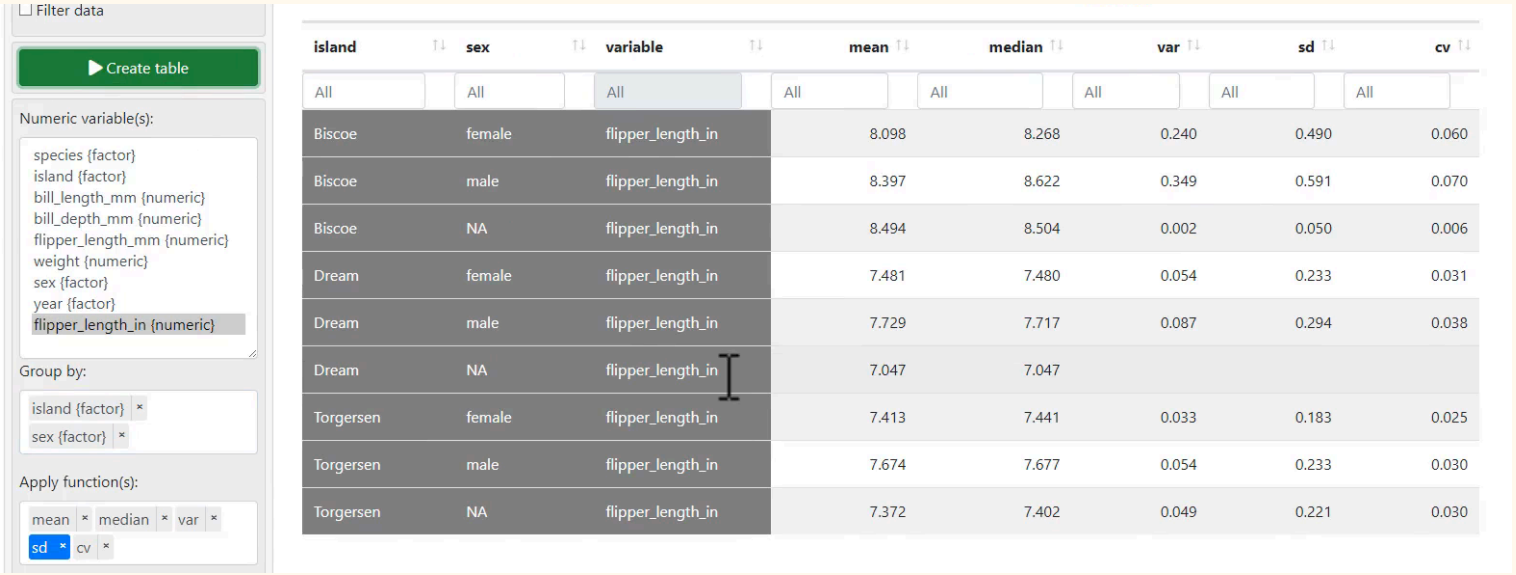
Under the explore tab, summary statistics can be generated.



choose the variable and click create/update table to show table of summary stats



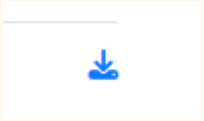
can select different stats to calculate, cv = coefficient of variation



island	sex	variable	mean	median	var	sd	cv
All	All	All	All	All	All	All	All
Biscoe	female	flipper_length_in	8.098	8.268	0.240	0.490	0.060
Biscoe	male	flipper_length_in	8.397	8.622	0.349	0.591	0.070
Biscoe	NA	flipper_length_in	8.494	8.504	0.002	0.050	0.006
Dream	female	flipper_length_in	7.481	7.480	0.054	0.233	0.031
Dream	male	flipper_length_in	7.729	7.717	0.087	0.294	0.038
Dream	NA	flipper_length_in	7.047	7.047			
Torgersen	female	flipper_length_in	7.413	7.441	0.033	0.183	0.025
Torgersen	male	flipper_length_in	7.674	7.677	0.054	0.233	0.030
Torgersen	NA	flipper_length_in	7.372	7.402	0.049	0.221	0.030

can group by many factors, meaning for each island and each gender the summary statistic is calculated by group.

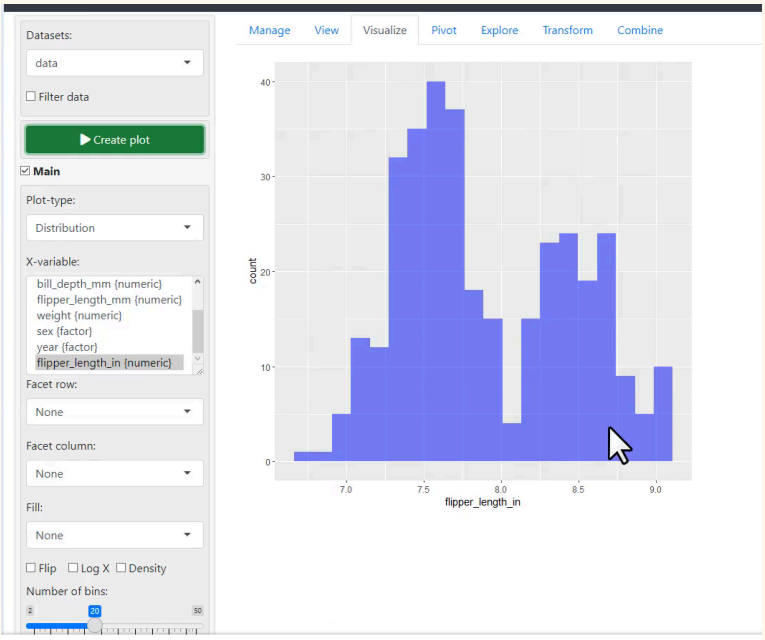
All tables generated can be saved by pressing this button



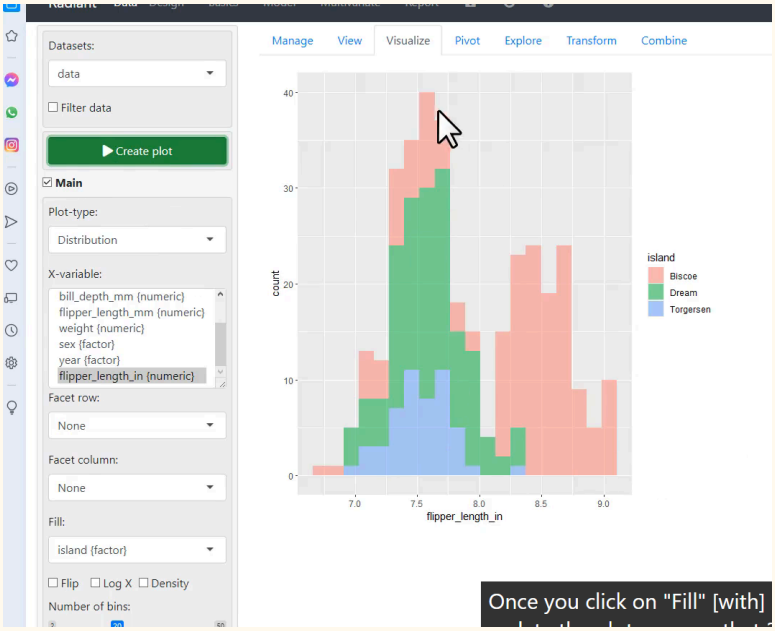
Visualise

Under the visualise tab can create different graphs.

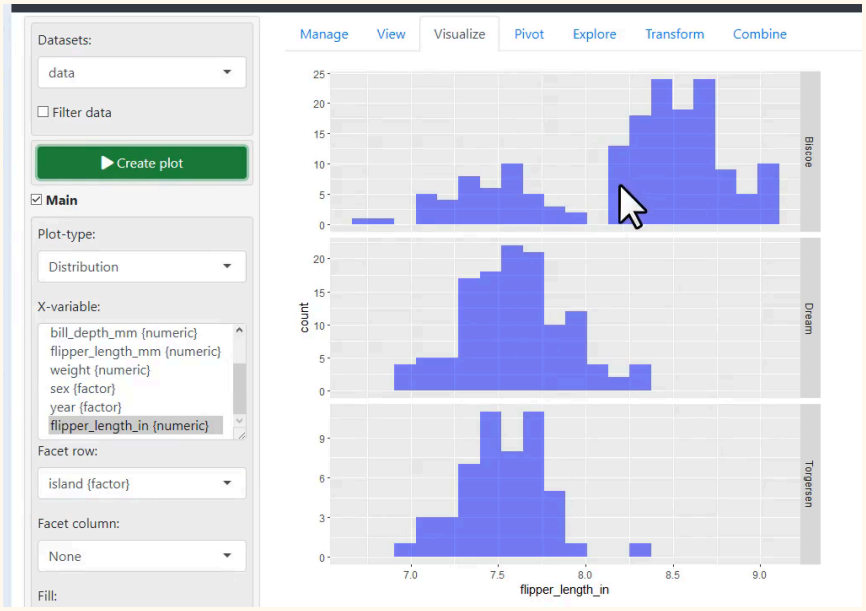
Histogram (Distribution)



Slider at the bottom can change number of bins

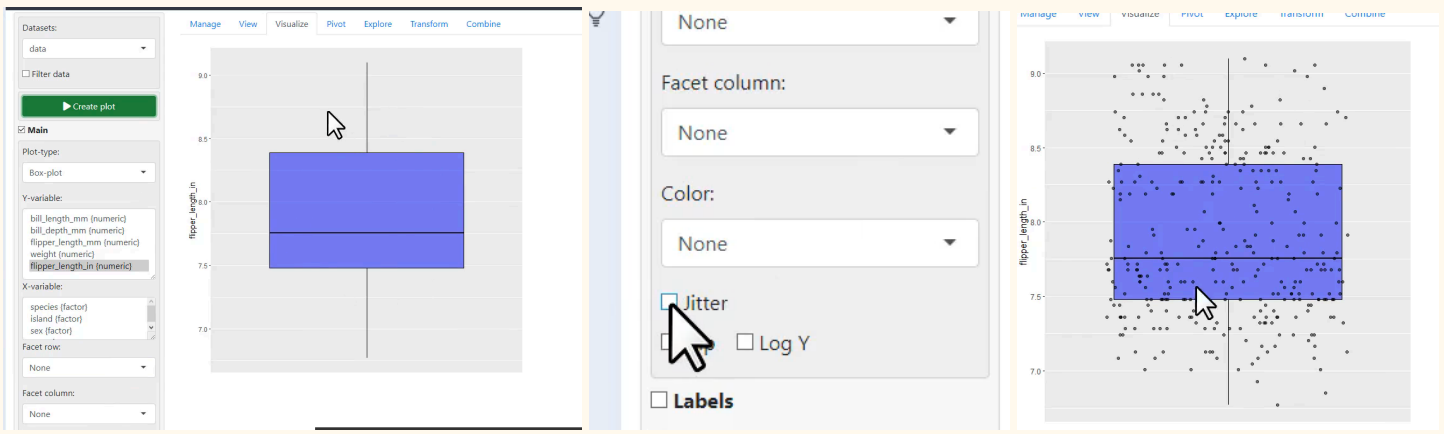


- fill by factor can change the colour by grouping of island
- using this can tell that the first peak is mainly due to green and blue
- while the second peak is due to primary pink island

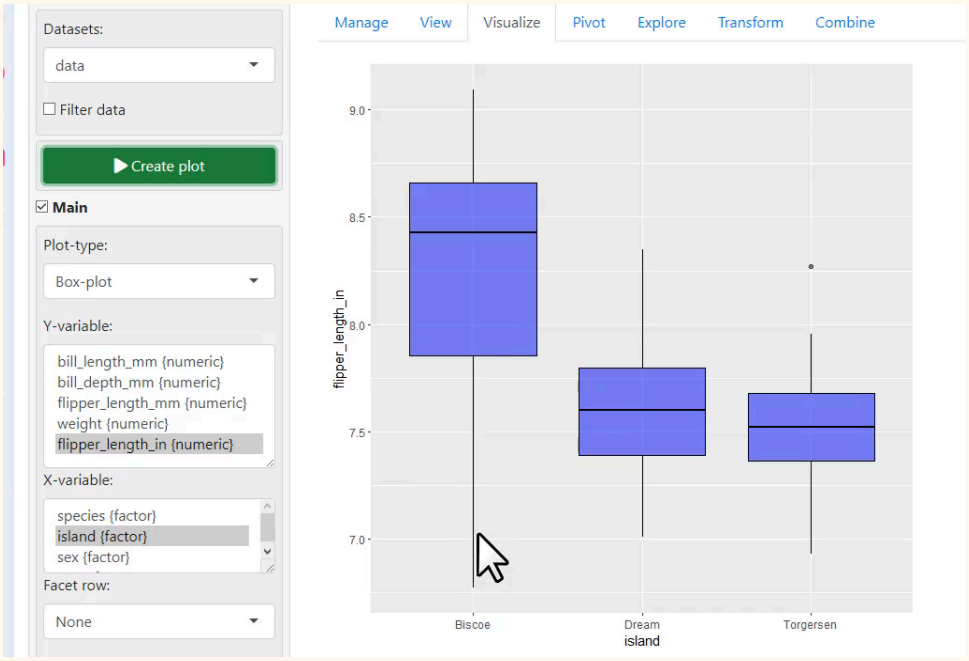


- It's hard to plot on the same graph, histograms of many different factors so can plot them separately by factors
- Under facet row choose the factor and it'll plot three diff histogram for each island

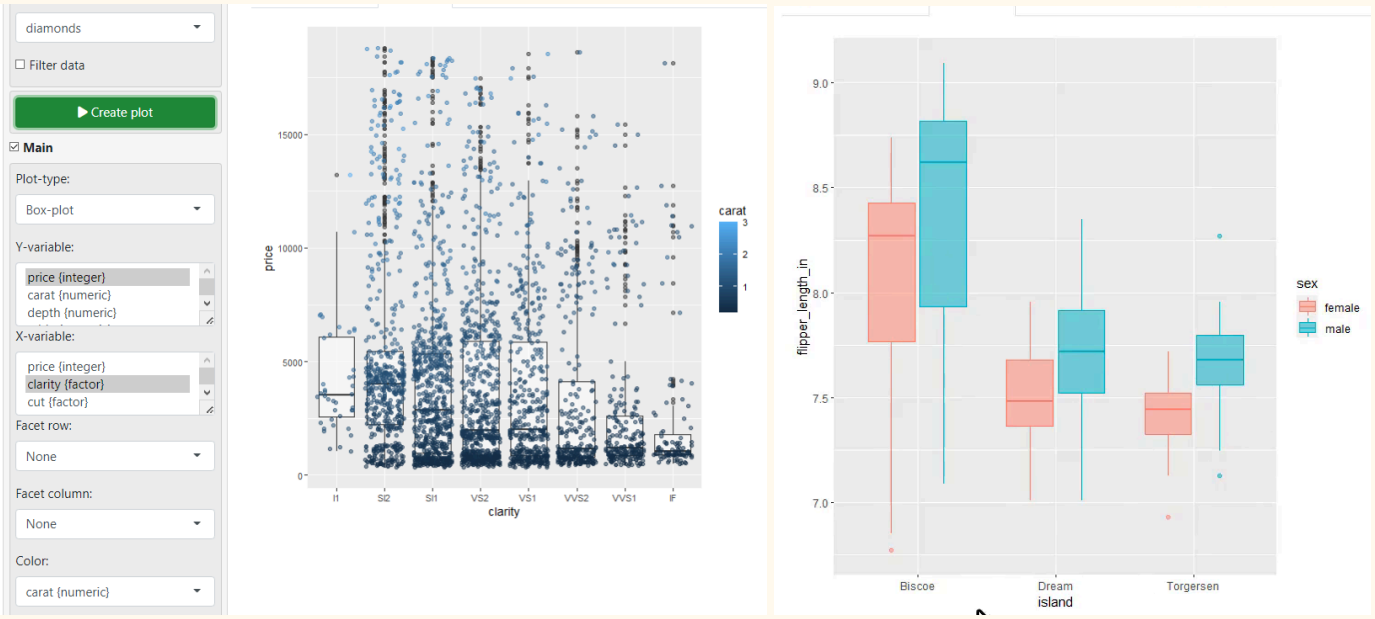
Box Plot



- jitter option will overlay the datapoint that made up this box plots this gives the user idea of the datapoint used to make box plots.



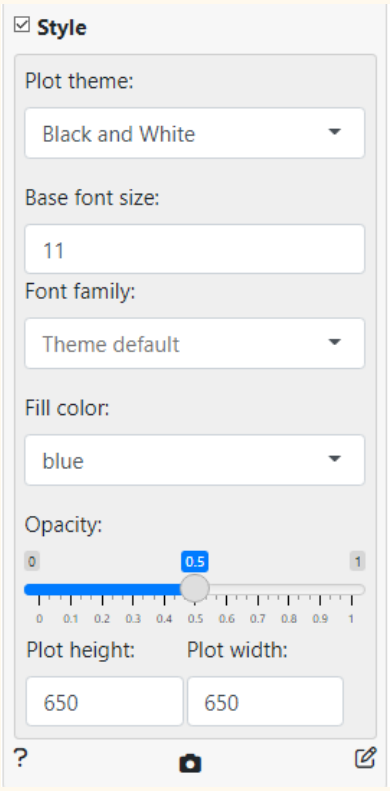
- can change x-variable to island and can see got 3 island box plots
- can tell that the first island have longer flipper length



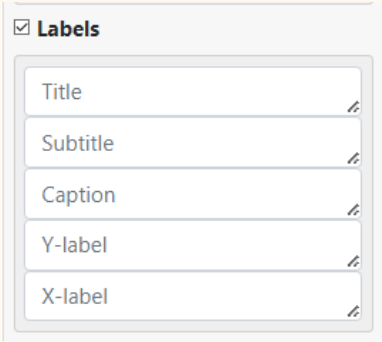
- X-variable to get multiple plots for price
- colour to to see different carat
- so for diamond of each carat the box plot fo price is plotted and the carat is shown
- colour for gender is only 2 so a gradient wasn't used.

Bar Graph

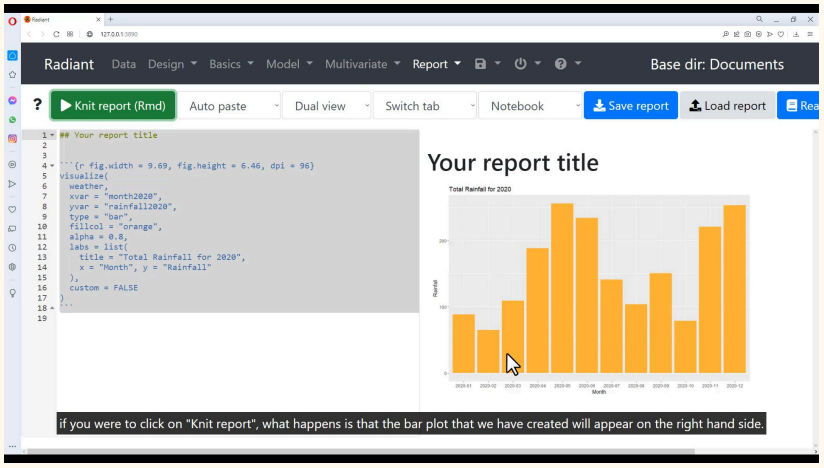
- Under style can



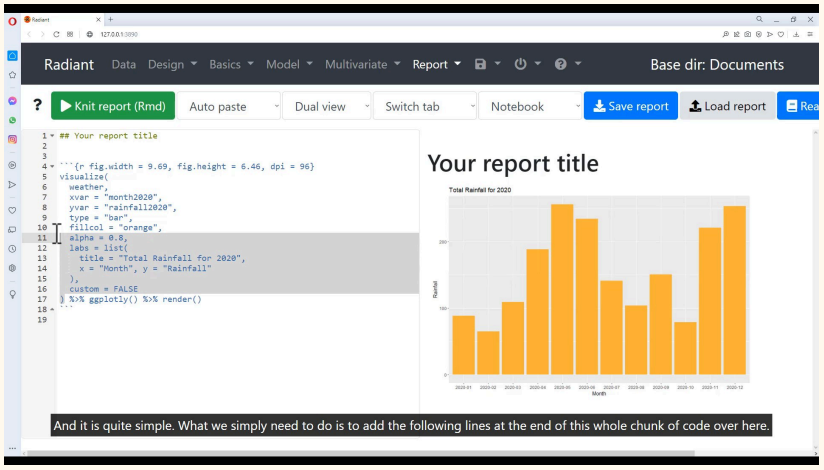
    - change height and width
    - change colour of bars also
- Under labels



    - can change x-axis y-axis labels
- Presseing 
    - the right report result button will bring up a new tab



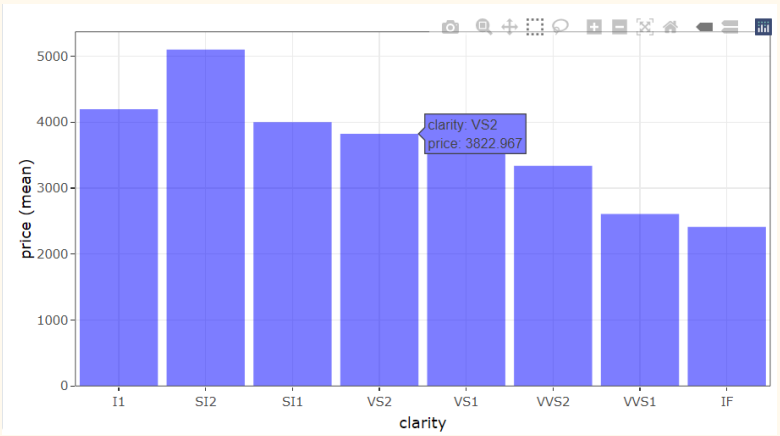
        - Pressing Knit report will show the graph



- typing

  ```
  %>% ggplotly() %>% render()
  ```

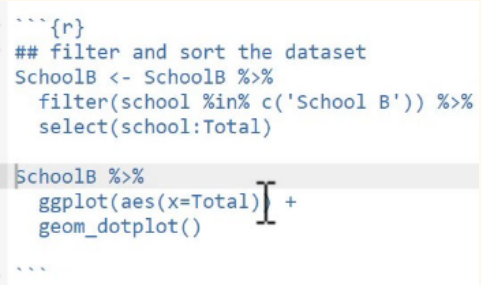
  will generate an intractable plot, can zoom and pan, hovering over the bar gives info.



- graphs can be saved as a html file.

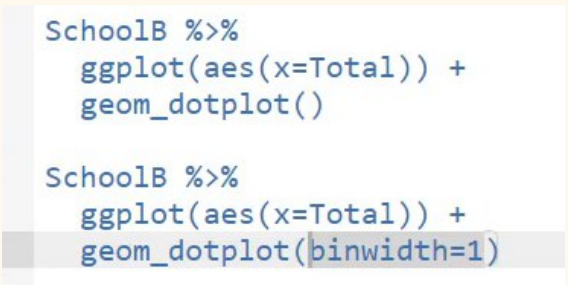
Dot Plot

Radiant don't have dot plot build in but can use some R language command to plot it



```
```{r}
## filter and sort the dataset
SchoolB <- SchoolB %>%
  filter(school %in% c('School B')) %>%
  select(school:Total)

SchoolB %>%
  ggplot(aes(x=Total)) +
  geom_dotplot()
```
```



```
SchoolB %>%
  ggplot(aes(x=Total)) +
  geom_dotplot()

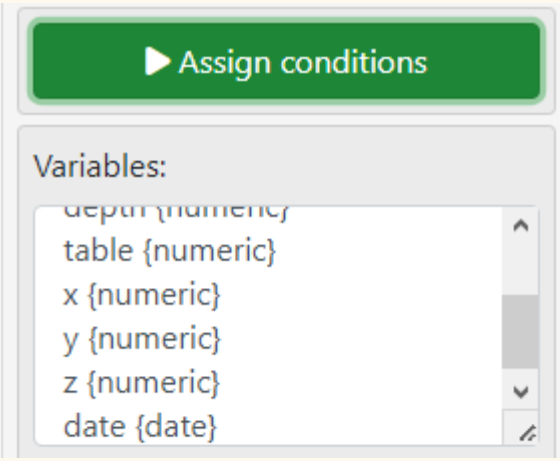
SchoolB %>%
  ggplot(aes(x=Total)) +
  geom_dotplot(binwidth=1)
```

- Extra code should always be added before the ```
- schoolB means using the dataset called schoolB
- %>% means the data set will be passed into the function defined in r called ggplot which is a powerful function in r to plot various graphs
- the aes(x=total) means total is the variable to be plotted
- + geom_dotplot() means telling R to do dot plot
- + geom_dotplot(binwidth = 1) means telling R to do dot plot with binwidth 1, else default bandwidth is calculated by some algorithm.

All plots generated can be saved by pressing this button

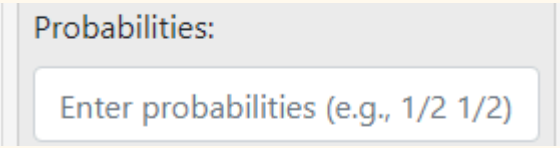
Random Assignment

Design > Random Assignment

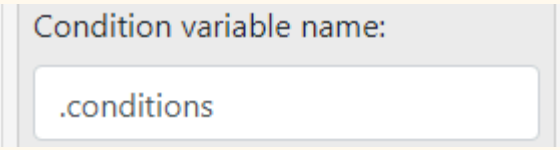


Choose the variable to be assigned into groups

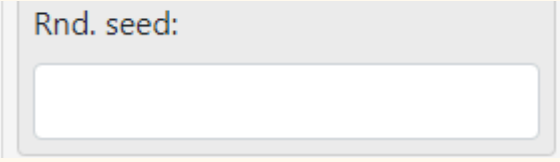
Label of groups, can be more than 2 groups



proportion of entries to be allocated into each group, 0.3 and 0.7 means proportion would be roughly 0.3 and 0.7 as it is a random process. typing 150/400, 250/400 will force 150 and 250 to be allocated.

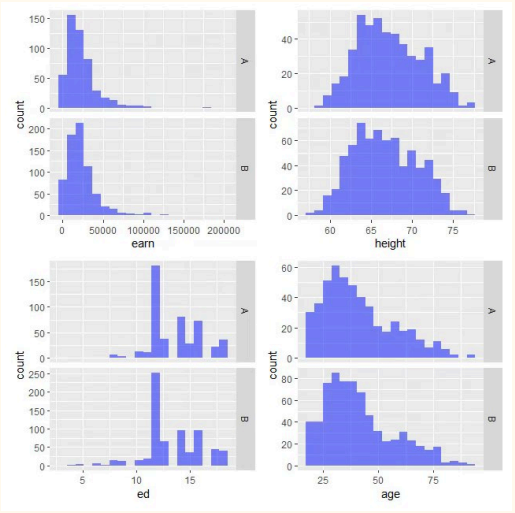


This is just the variable name for the groups, the header.



Leave this blank to get a different seed each time

After random assignment, a new data set can be saved and the graph for numerical and contingency tables (under pivot) for categorical data can be plotted. The similar distribution and proportion shows that random assignment works



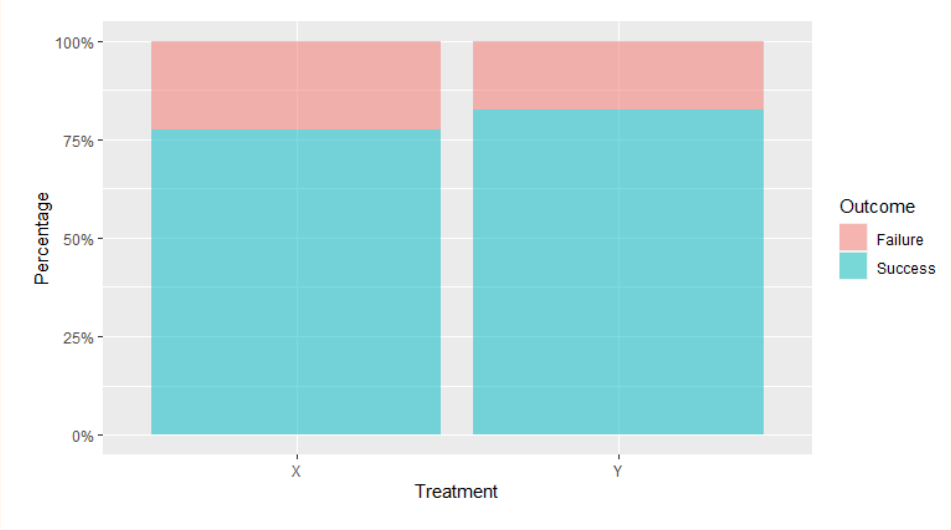
sex	A	B	Total
All	All	All	All
female	0.577	0.576	**0.576**
male	0.423	0.424	**0.424**
Total	**1.000**	**1.000**	**1.000**

* when plotting table, normalised by column to get decimal

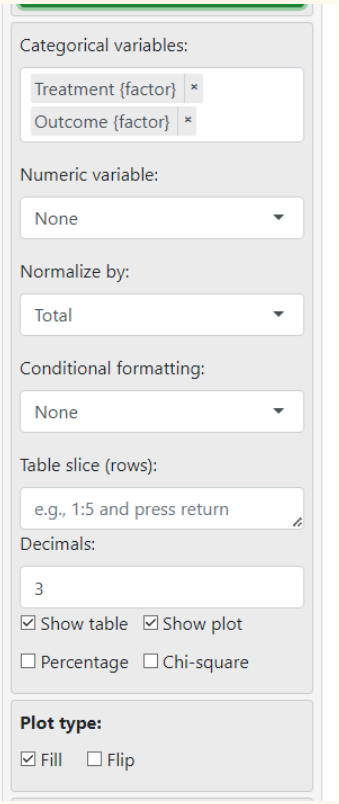
Pivot

	Treatment		
Outcome	**X**	**Y**	**Total**
All	All	All	All
Failure	0.150	0.058	**0.209**
Success	0.516	0.275	**0.791**
Total	**0.667**	**0.333**	**1.000**

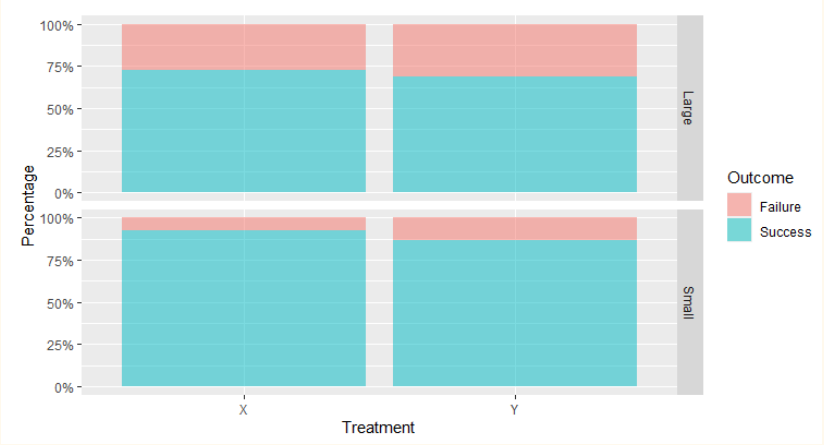
Pivot tab can be used to generate tables like this.



can plot stack or dodged bar chart, select Fill under plot type for stack bar chart

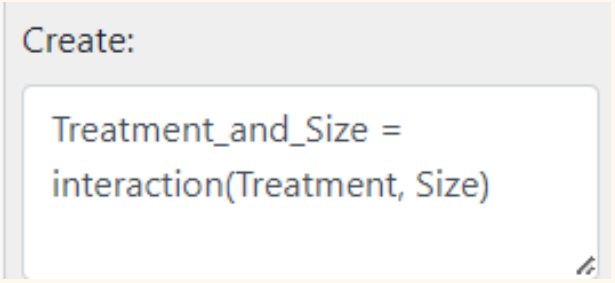


Factoring in size



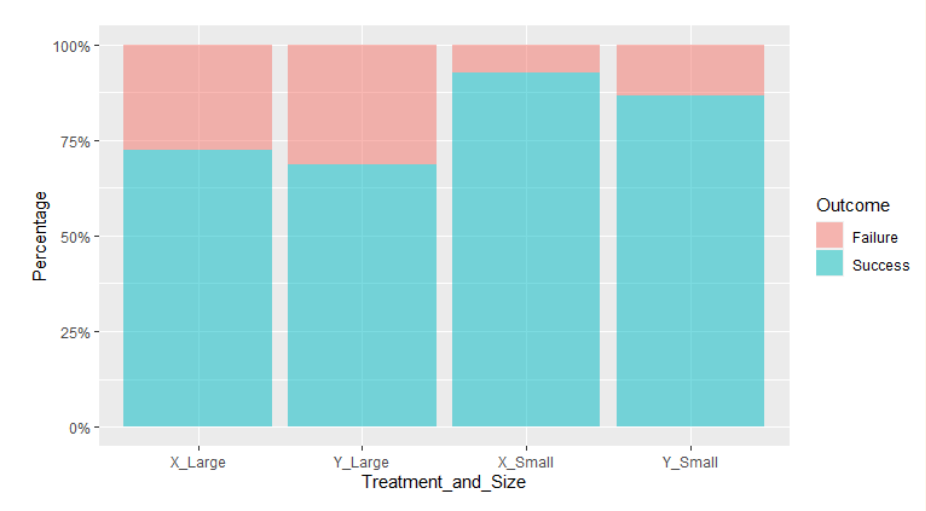
Just add size as a categorical variable and it'll plot this.

To plot in one singular plot, have to combine the size and treatment variable and plot against outcome.



Treatment	Size	Gender	Outcome	Treatment_and_Size
Y	Small	Male	Success	Y.Small
X	Large	Male	Failure	X.Large
X	Large	Male	Failure	X.Large
X	Small	Male	Failure	X.Small

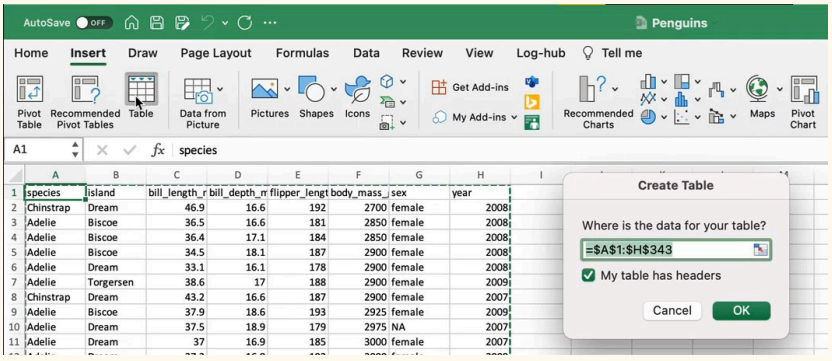
Plotting it will look like this



Can see that X is the better treatment for both large and small stones.

Excel

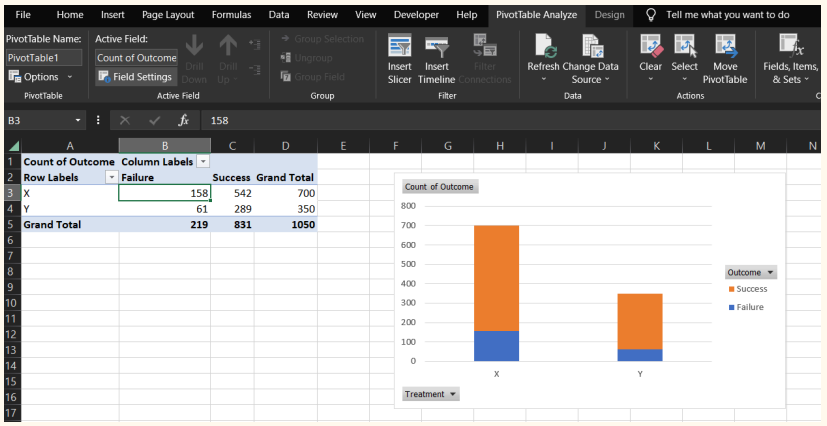
Grouping cells together to form a table: Insert > Table

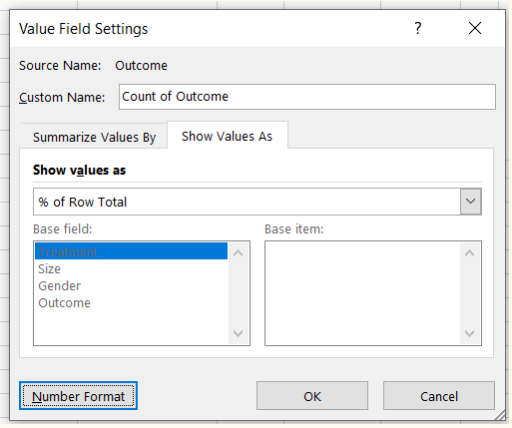


Formulas

AVERAGE - mean
STDEV.S - sample standard deviation

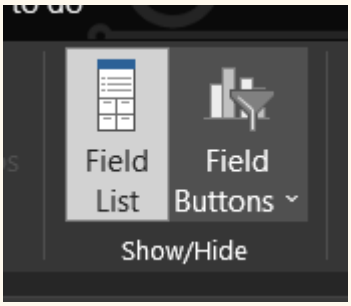
Pivot Chart



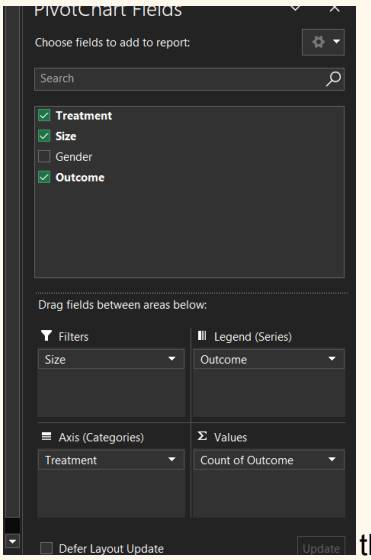


To normalise data to total count so that percentage is plotted. click on the value in pivot table -> Pivot table Analyze -> Field settings -> Show value as -> change to row total, number format can change decimal point.

To factor in the stone size

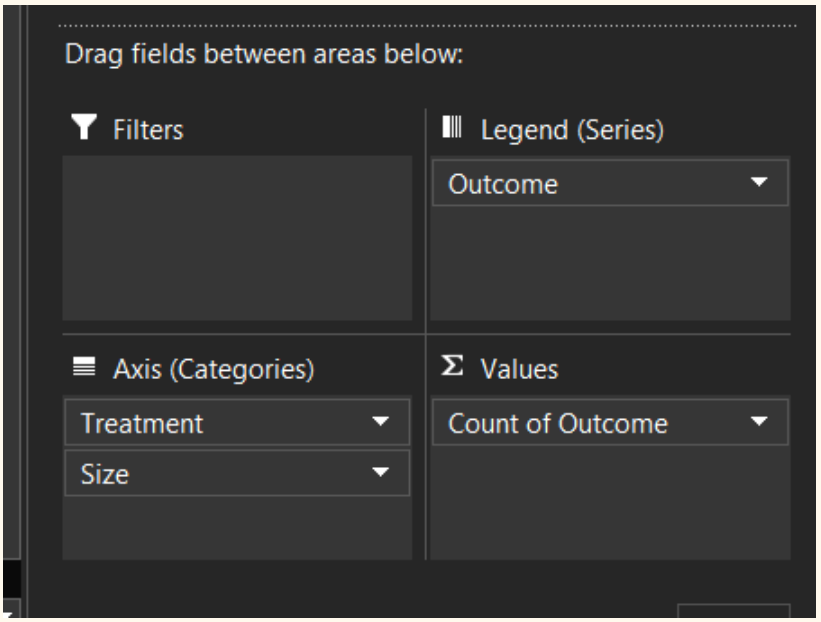


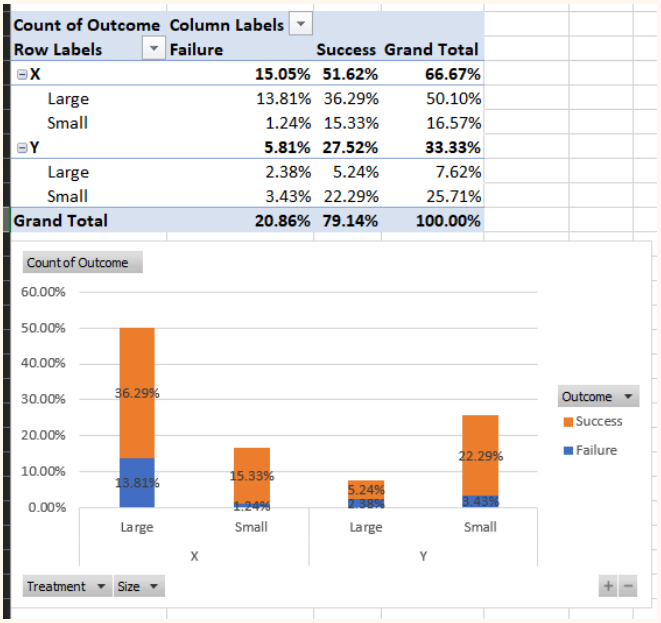
First show the field list which is the setting for pivot table



then drag size into filter, this will allow size to be filtered to plot different bar chart

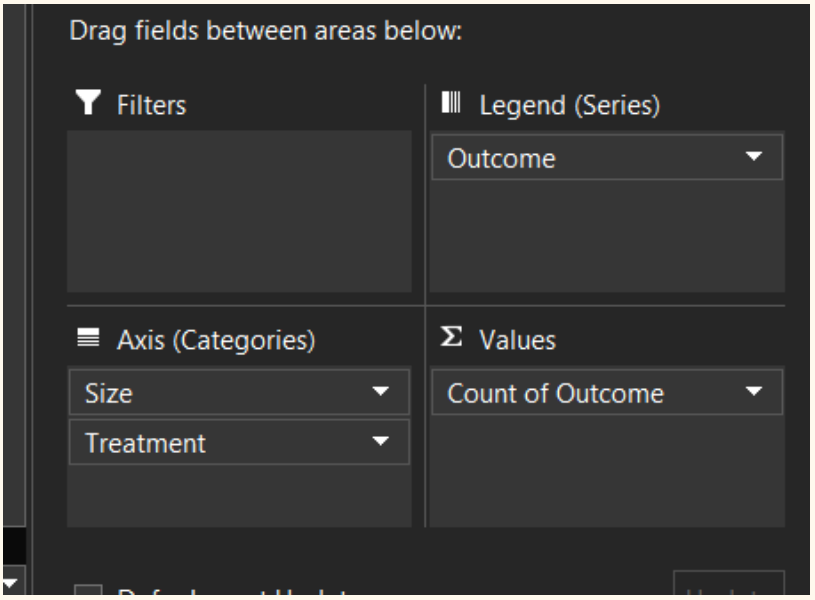
If want to plot size together in the bar chart, then drag size into category axis



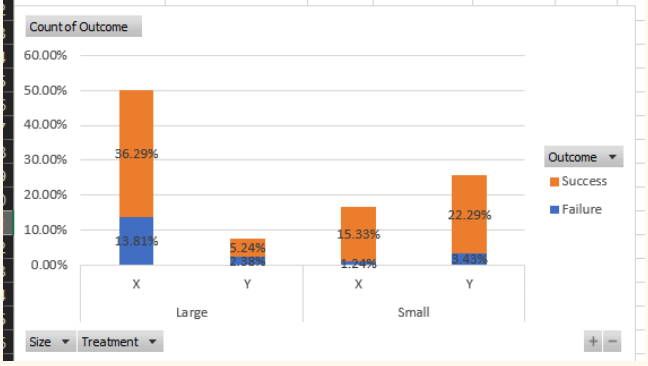


Count of Outcome	Column Labels		
Row Labels	Failure	Success	Grand Total
X	15.05%	51.62%	66.67%
Large	13.81%	36.29%	50.10%
Small	1.24%	15.33%	16.57%
Y	5.81%	27.52%	33.33%
Large	2.38%	5.24%	7.62%
Small	3.43%	22.29%	25.71%
Grand Total	20.86%	79.14%	100.00%

Swap the order of treatment and size to categorise size first

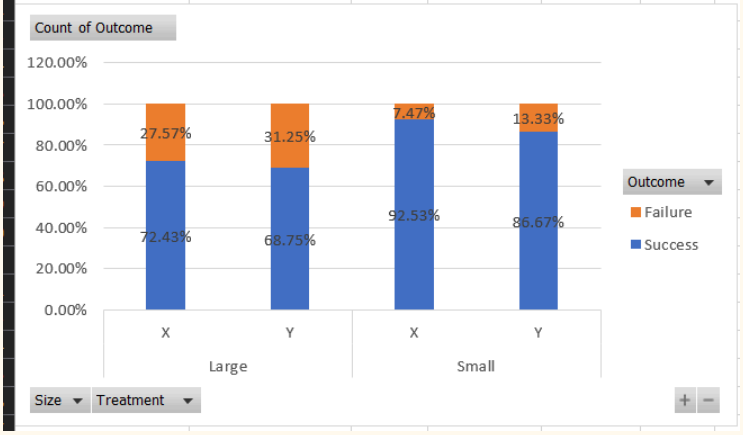


Count of Outcome	Column Labels		
Row Labels	Failure	Success	Grand Total
Large	**16.19%**	**41.52%**	**57.71%**
X	13.81%	36.29%	50.10%
Y	2.38%	5.24%	7.62%
Small	**4.67%**	**37.62%**	**42.29%**
X	1.24%	15.33%	16.57%
Y	3.43%	22.29%	25.71%
Grand Total	**20.86%**	**79.14%**	**100.00%**



The above two charts were drawn normalised to grand total, row total should be used.

Count of Outcome	Column Labels		
Row Labels	Success	Failure	Grand Total
Large	**71.95%**	**28.05%**	**100.00%**
X	72.43%	27.57%	100.00%
Y	68.75%	31.25%	100.00%
Small	**88.96%**	**11.04%**	**100.00%**
X	92.53%	7.47%	100.00%
Y	86.67%	13.33%	100.00%
Grand Total	**79.14%**	**20.86%**	**100.00%**



Can see that X is the better treatment for both large and small stones.

To save, have to save in an excel file not .csv else everything is gone.