GEA cheat sheet

Quantitative reasoning with data (National University of Singapore)

❖ Chapter 1 Summary statistics

➢ Mean

- Adding a constant value c to all the data points changes the mean by that constant value.
- Multiplying a constant value of c to all the data points will result in the mean being changed by the same factor of c
- It tells that it is not possible for all data points to be below / above the mean
- It does NOT tell that half of the data points are above the mean and half are below

➢ Standard deviation

- Adding a constant c to all data points does not change the standard deviation.
- Multiplying all the data points by a constant c results in the standard deviation being multiplied by |c|, the absolute value of c

➢ Median

- If we add a positive constant c to all the data points, the median value will increase by c
- When a constant c is multiplied to all the data points, then median is also multiplied by c, w/o modulus
- 50% of the data is less than or equal to this value, 50% of the data is more than or equal to this value

➢ Quartiles and IQR

- If we add a positive constant c to all the data points, Q1 and Q3 are increased by c. 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|

➢ Choice of summary statistics

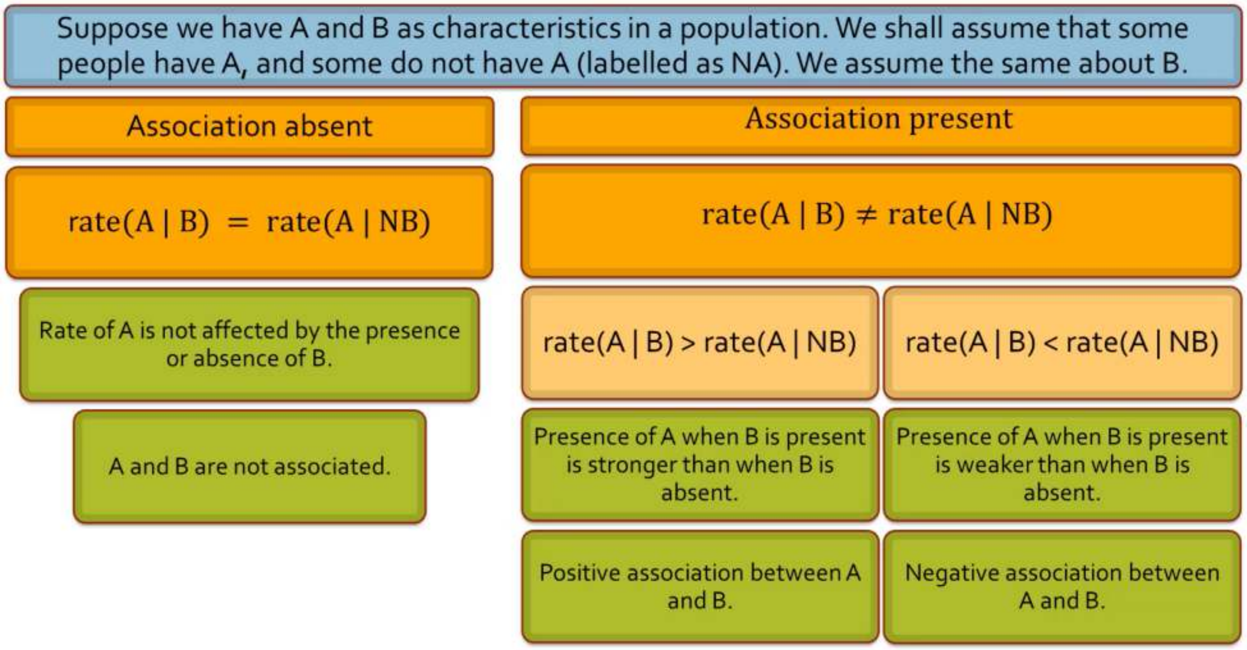
- For a numerical variable,we can always use the **mean** and **standard deviation** as a pair of summary statistics to describe the central tendency as well as the dispersion and spread of the data. Similarly, the **median** and **IQR** can also be used.
- The choice depends on the distribution of the data. Generally speaking, the **median** and **IQR** is preferred if the distribution of the data is not symmetrical or when there are outliers.
- Outliers influence **mean** and **standard deviation** by a great deal but have little to no effect on **IQR**, **median** and **mode** which are thus known as robust statistics

❖ Chapter 2 Dealing with categorical data

➢ Basic rules on rate

- Basic rule: The overall rate(A) will always lie between rate(A | B) and rate(A | NB).
  - rate(A) = rate(A|B) * rate(B) + rate(A|NB) * rate(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) = 12 [rate(A |B) + rate(A | NB)].
  - Consequence 3: If rate(A | B) = rate(A | NB), then rate(A) = rate(A | B) = rate(A | NB).

➢ Association



➢ Symmetric rule

<u>Symmetry Rule Part 1:</u>

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

<u>Symmetry Rule Part 2:</u>

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

<u>Symmetry Rule Part 3:</u>

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

➢ Simpson's paradox

	Large stones			Small stones			Total (Large + Small)		
	Successful treatments	Total number of treatments	rate(Success) in %	Successful treatments	Total number of treatments	rate(Success) in %	Successful treatments	Total number of treatments	rate(Success) 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%

❖ Chapter 3 Dealing with numerical data

➢ Univariate

■ Histogram

- Histogram can provide a better sense of the shape of the distribution, especially when there are great differences between the frequencies of the data points
- Histogram provides a clearer sense of the frequency distribution of data points

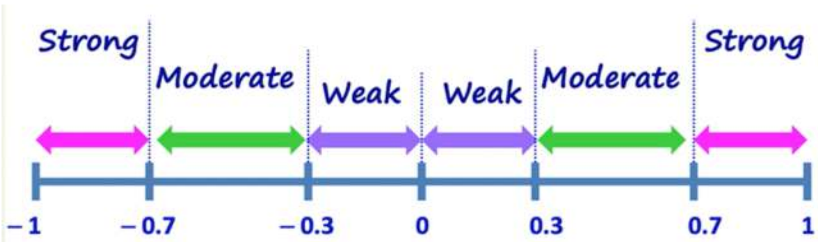
■ Box Plots

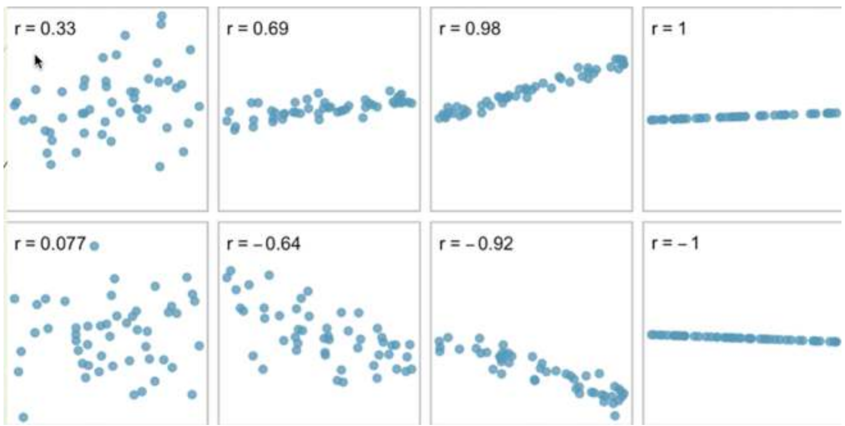
- **Merits**
  - Box-plots are more useful when comparing between different data sets
  - Box-plot allow us to identify outliers quite easily
- **Disadvantage**
  - Box-plot does not give any information about how many data points we are working with
  - Box-plot does not tell anything about the frequency distribution e.g. we cannot compute passing / failure rates based on a boxplot of the scores of an exam. As shown above, datasets with disparate distribution can produce the exact same boxplot.
  - Box-plot does not provide information on standard deviation and mean

➢ Bivariate

■ Correlation coefficient

- **Correlation coefficient *r* is a measure of linear association**, range is between -1 and 1. It summaries the direction and strength of linear association. Correlation coefficient is often denoted by ***r***.
- $r > 0$: position association
- $r < 0$: negative association
- $r = 0$: no linear association, not necessarily imply no association
- $r = 1$: perfect positive association
- $r = -1$: perfect negative association

- 

- 

■ ***r*** is ***NOT*** affected by
  - Interchanging the two variables ie. x-axis variable ⟵⟶ y-axis variable
  - Adding a constant value to all data points of a variable
  - Multiplying all data points of a variable by a scalar

■ The **sign of r will change** when we multiply a negative number to one of the variables.

■ Limitation of correlation coefficient
  - Correlation does not imply causation
  - **r** does not reflect non-linear association so we should always look at the scatter plot
  - Outliers may decrease / increase / not change the strength of the correlation

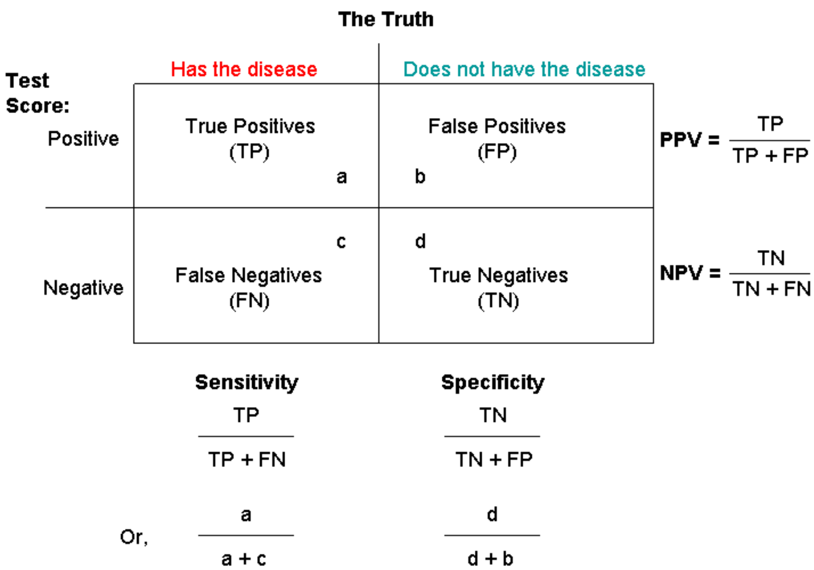
■ **Linear regression**
  - The correlation coefficient r between the variables X and Y is closely related to the regression line Y = mX + b obtained using the method of least squares. More precisely, we have $m = \frac{S_y}{S_x} r$.
  - Where S_y is the standard deviation for y and S_x is the standard deviation for x.
  - With this relationship, we see that if the **correlation coefficient** r is **positive**, then the **gradient** of the regression line is also **positive**. Similarly, if the **correlation coefficient** is **negative**, then the **gradient** of the regression line will also be **negative**. However, it is important to remember that the correlation coefficient is not necessarily equal to the gradient of the regression line.

❖ Chapter 4 Statistical inference

➢ Conditional probability

■ P(E | F) : the probability of E given F. This is interpreted as how likely the outcome is in E if we know that it is in F, which is P(EF)P(F). If P(F) = 0, then we stipulate P(E F)=0

- Conditional probability is equivalent to conditional rates. ie. P(A | B) = rate(A | B)
- **Sensitivity and specificity**
  - Sensitivity: P(test positive|has the disease)
  - Specificity:P(test negative|do not have the disease)



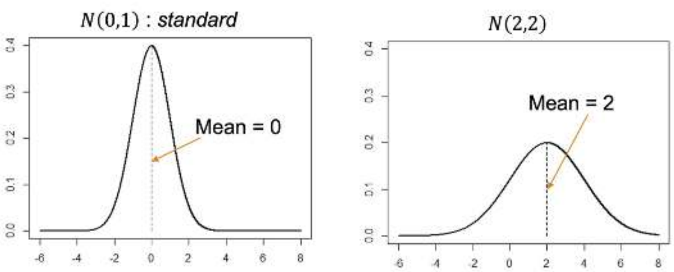
  -

Independence & mutually exclusive

- Two events A and B are said to be independent if
  - P(A) = P(A|B); **OR**
  - P(A)P(B) = $P(A \cap B)$
    - Derived from $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - To put in words: what is the chance of A happening if B has already happened
- If rate(A) = rate(A|B), then events A and B are not associated
  - Two events are said to be independent if they are not associated.
- Two events are mutually exclusive or disjoint if they cannot both occur at the same time.

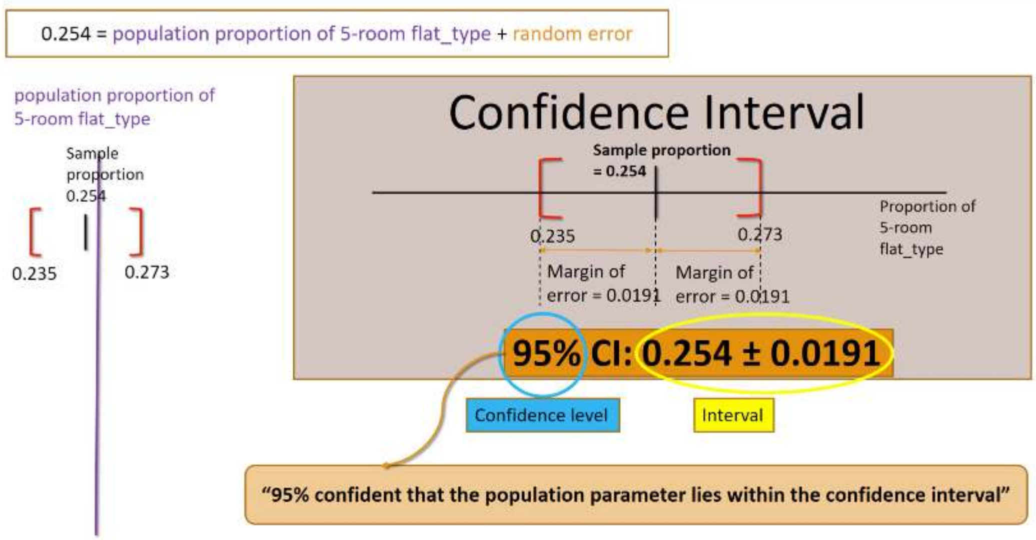
Random variables

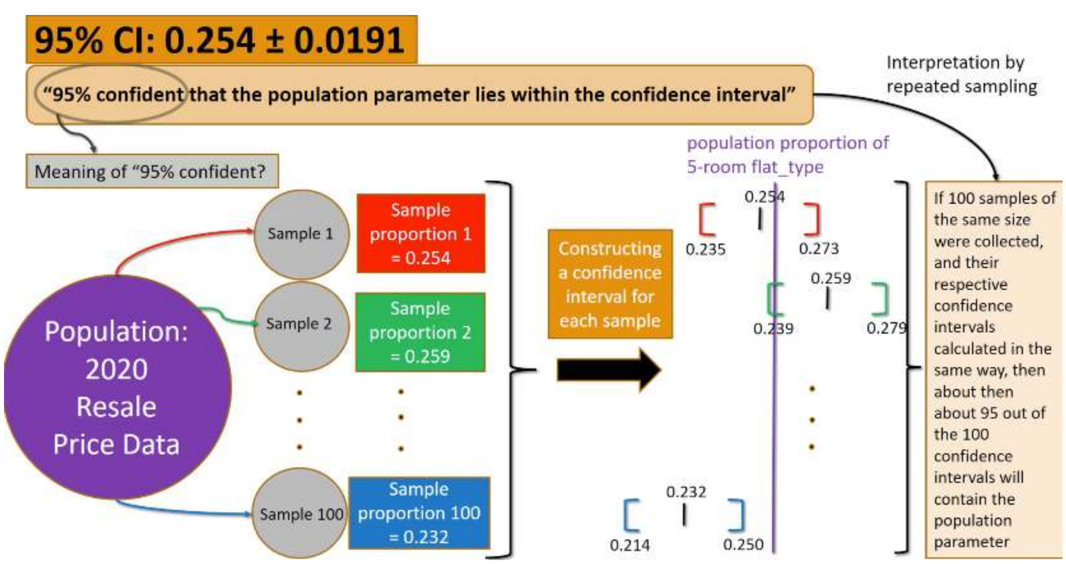
- The normal distribution is a **symmetrical**, bell-shaped distribution in which the **mean, median and mode are all equal.** The density curve of any normal distribution is symmetrical about its mean, and its mean is equal to its mode, it is symmetrical about its mode. A symmetrical distribution cannot be left-skewed or right-skewed.
- Normal distribution is a distinguished class of continuous random variables. We use N(x, y) to denote the normal distribution with *mean* x and *variance* y.



- A **smaller variance** corresponds to a **thinner** bell shape, while a **greater variance** corresponds to a **fatter** bell shape
- The mean value determines where the peak of the graph occurs

➢ Confidence interval

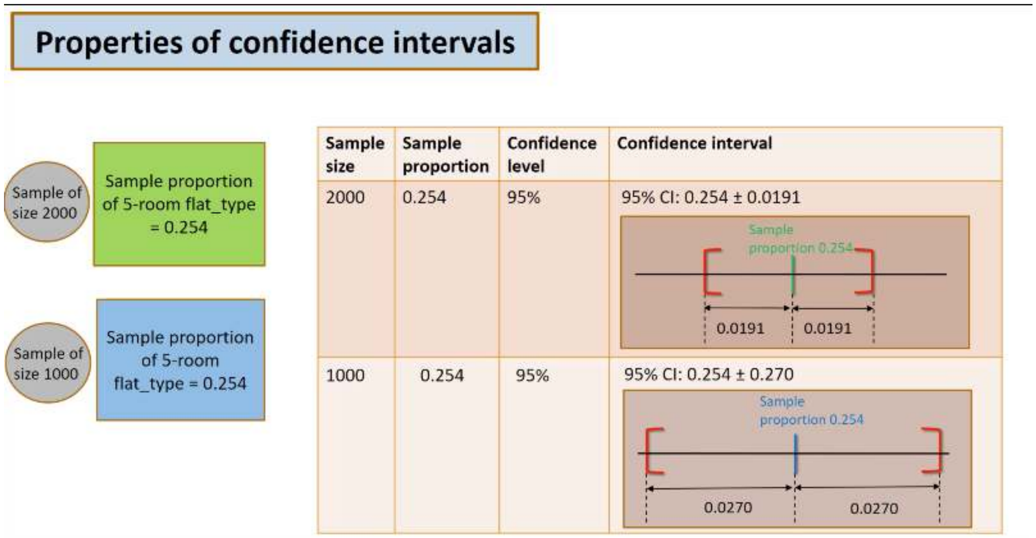


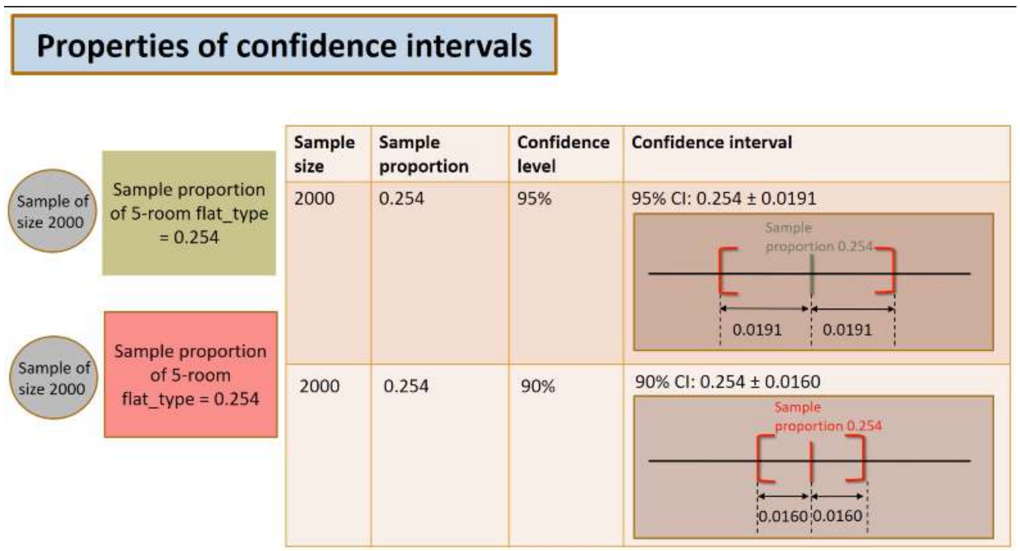


  - 95% confidence level does not mean there is a 95% chance that the population parameter lies within our confidence interval since the population parameter is a fixed value, it either lies within the interval, or it does not. Rather, when we say 95% confidence level, we refer to

- 95% of the sample statistics we collect will contain the population parameter
- We are 95% confident that the population parameter lies within our confidence interval

➢ **Factors affecting confidence interval**

- 

- 

➢ Hypothesis testing

- General procedure
  - Step 1: Identify the question and state the null hypothesis and alternative hypothesis
    - H_0 : takes a stance of no stance or no effect. This hypothesis assumes that any differences seena re due to variability inherent in the population and occurred by chance
    - H_1 : which we wish to confirm and pit against the null hypothesis
  - Step 2: Collect relevant data. Decide on the relevant test statistic.
    - A test statistic is a random variable that is to be calculated from sample data and used in hypothesis testing.
    - $z = \dfrac{\overline{x} - \mu}{\sigma / \sqrt{n}}$

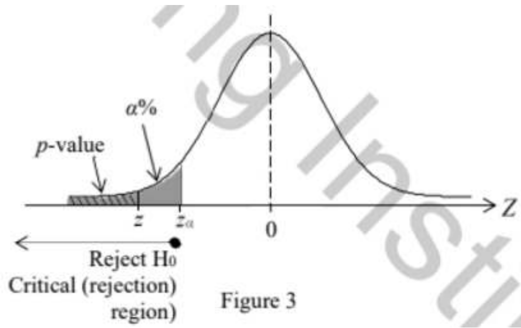
- Step 3: Determining the level of significance and computing the p-value
  - The lower the level of significance, the more evidence we need to reject the null hypothesis. Commonly used level of significance are 1%, 5% and 10%
  - Level of significance is defined as a% = P(rejecting H_0 | H_0 is true)
  - p-value is a probability value, defined as P(Z < z)
  - 

Figure 3
- Step 4: Making conclusion about the null hypothesis
  - Reject null hypothesis in favour of the alternative if p-value < significance level
  - Otherwise, do not reject the null hypothesis if p-value > significance level. Our result is inconclusive. We also cannot conclude that the null hypothesis is correct.

➢ One sample t-test

- when population is normally distributed if sample size < 30.
- The sample should be random
- Hypotheses
  - $H_0 : \mu = \mu_0$
  - $H_1: \mu > / < / =/= \mu_0$
  - μ_0 is the assumed population mean

➢ Chi-squared test

- Commonly used to check whether two categorical variables, A and B are associated at the population level
- The data must be counts for the categories of a categorical variable
- The sample should be random
- Hypotheses
  - H_0 : A and B are not associated. ie rate(A|B) = rate(A|NB)
  - H_1 : A and B are associated ie rate(A|B) =/= rate(A|NB)

➢ P-value

- In the context of p-value computation, “at least as extreme” is interpreted as “at least as favourable to the alternative hypothesis”.