Exploratory Data Analysis

Systematic process to explore dataset

1. Generate research questions about data
2. Search for answers using data visualization and modelling
3. Evaluate how well the data answers the questions
4. Refine existing questions or generate new questions

Population: entire group (of individuals or objects) that we wish to know something about

Research question: seeks to investigate some characteristic of a population

- Make an estimate about the population
- Test a claim about the population
- Compare two sub-populations
- Investigate relationship between two variables in the population

Considerations for research questions:

Narrow	• Lacks further context • Answerable with simple statistic
Unfocused	• Hard to identify research methodology
Simple	• Answers easily found online without analysis required

Sampling

Population of interest: group in which there is an interest in drawing conclusions on in a study

Population parameter: numerical fact about a population

- Must align exactly with what is being studied

Estimate: inference about the population parameter based on the information obtained from a sample

Sampling frame: list from which the sample was obtained

Characteristics of sampling frame

1. Includes all available sampling units from the population
2. No irrelevant/extraneous sampling units from another population
3. No duplicated sampling units
4. No clustered sampling units

Census: attempt to reach out to the entire population of interest; not always possible due to a) high cost, b) long time to complete, c) not guaranteed 100% response rate

Bias: skews conclusion from sample as it means that it is not generalizable

1. Selection bias: researcher's biased selection of units into the sample; caused by imperfect sampling frame or non-probability sampling
2. Non-response bias: participants' non-disclosure or non-participation in research study; caused by inconvenience or unwillingness; occurs with probabilistic and non-probabilistic sampling

Probability sampling: sampling method where selection done using known randomized mechanism; every unit has a non-zero probability of being selected (unequal probability is fine); element of chance to eliminate bias; does not guarantee equal group size; random experiments often use this; distribution of characteristics across random samples likely similar

Sampling	Process	Remarks
Simple Random 	Units randomly selected with equal probability using RNG and without replacement	Variability entirely up to chance; requires knowledge of total samples present
Systematic 	Units selected at regular k intervals and random starting point r $r, r+\frac{n}{k}, r+\frac{2n}{k}, \ldots, r+\frac{(k-1)n}{k}$	If n is not known, just keep sampling till no more samples are available; subject to inherent grouping or ordering of units so might not be entirely representative
Stratified 	Divide population into groups (strata) with similar characteristics and randomly sample from each stratum	Each group does not need to be the same size; requires information about characteristics; difficult to pinpoint the characteristics of all data points
Cluster 	Divide population into clusters and randomly select n clusters to take all samples from	Simpler, less costly, and less resource intensive as clusters are naturally defined; high variability if dissimilar clusters are chosen; low cluster count chosen not representative of the population

Advantages/Disadvantages:

Sampling	Advantages	Disadvantages
Simple Random	Good representation of population	Time-consuming; accessibility of information and sampling frame
Systematic	Simple selection process	Potentially under-representing the population
Stratified	Good representation of the sample by stratum	Require sampling frame and classification criteria of the population into stratum
Cluster	Less time-consuming and less costly	Requires clusters to be reasonably heterogeneous and not have cluster-specific characteristics

Non-probability sampling: selection of units is not done by randomization

Sampling	Process	Remarks
Convenience	Researchers choose subjects to form a sample among those that are easily available to participate	Introduces selection bias; non-response bias prevalent
Volunteer	Subjects volunteer themselves	Contains subjects with a strong opinion about the question which is unrepresentative

Generalizability: must be able to generalize the findings from the sample to the population

1. Sampling frame that is equal to or greater population of interest
2. Using probability sampling
3. Having large sample size
4. Minimize non-response

Representative: sampling frame used contains similar distribution of characteristics in people sampled; i.e. representative of whole population

Sampling approach:

1. Design a sampling frame
2. Decide on the most appropriate sampling method to generate a sample from the sampling frame; prefer probability sampling
3. Remove unwanted units (those not from the population) from the sample

Variables

Variable: attribute that can be measured or labelled

Dataset: collection of individuals (objects or people) and variables pertaining to the individuals

Independent variables: subjected to adjustments (deliberate or sudden)

Dependent variables: changes depending on independent (hypothesized)

Categorical		Numerical	
Variables that taken on categories or label values that are mutually exclusive (countably infinite)		Variables that taken on numerical values and meaningful arithmetic operations (addition/average) can be performed (uncountable)	
Ordinal	*Nominal*	*Discrete*	*Continuous*
Natural ordering present	No intrinsic ordering	Gaps in the set of possible numbers	Can take on all possible values in a range/interval

Strings/dates: are usually treated as separate categories of variables

Whole numbers in range: are usually treated as ordinal categorical data

Fractional numbers: are usually treated as numerical data

Summary statistics

Five-number summary: min, Q1, median, Q3, max

Mean: average value of a numerical variable x

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

- Adding/multiplying constant c to all x will change the mean by that constant
- Does not provide information about spread of data
- Skews if extreme values are present
- Adding more data points that are less than the previous mean will lower the overall mean

Weighted average: computing the overall mean from a set of means given the total population per group

$$\frac{p_1}{n} \times \overline{a_1} + \frac{p_2}{n} \times \overline{a_2} + \cdots + \frac{p_k}{n} \times \overline{a_k}$$

where p_m is the number of elements in the subgroup and a_m is the average of the subgroup

- Always between the smallest and largest means

Proportions: used if the distribution of data is very uneven

Variance and Standard deviation: measure the spread of data about the mean

$$v = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

$$s_x = \sqrt{v}$$

- Always non-negative; only 0 if all x is the same
- Same unit as x
- Adding constant c to all x does not change the standard deviation
- Multiplying constant c to all x multiplies the standard deviation by $|c|$
- $\bar{x} + s_x \neq \max(x)$
- Visually, if more points are closer to the mean then there is less standard deviation (look at count and closeness to centre)

Range: difference between largest and smallest data point in the distribution

Coefficient of variation: degree of spread relative to mean

$$c_x = \frac{s_x}{x}$$

- No units
- Compares degree of variation across different means even if they are drastically different
- Higher values imply more dispersion about the mean

Median: middle value of sorted data (if even number of values, then take average of middle values)

- Adding/multiplying constant c to all x changes the median by c
- Cannot use the compute the weighted median like weighted mean
- Represents a numerical value where 50% of the data is less than or equal to the median, aka 50th percentile
- Mean is not the same as median and cannot be used interchangeably with one another

Quartiles:

- First quartile (Q_1): 25th percentile
  - Median of lower half of data points
- Third quartile (Q_3): 75th percentile
  - Median of upper half of data points

Interquartile range (IQR):

$$IQR = Q_3 - Q_1$$

- Shares similar properties to standard deviation
- Adding constant c to all x does not change IQR but changes Q_3 and Q_1

Preference of metric:
- Prefer median and IQR if distribution of data is not symmetrical or when there are outliers

Mode: value that appears most often in the data; peak of the distribution (i.e. highest probability to occur when randomly sampled)
- Applicable to both numerical and categorical data; other statistics are only for numerical

Balancing point: distributes the points into two equal groups; equals to the mean if present in the dataset; not always the mode/median

$$(y - x_1) + (y - x_2) + \cdots + (y - x_k) = (x_{k+1} - y) + (x_{k+2} - y) + \cdots + (x_n - y)$$
$$\Leftrightarrow ny - \sum_{i=1}^{n} x_i = 0$$
$$\Leftrightarrow ny = \sum_{i=1}^{n} x_i$$
$$\Leftrightarrow y = \frac{\sum_{i=1}^{n} x_i}{n} = mean$$

Experimental study

Intentionally manipulate the independent variable to observe the effects it has on the dependent variable; used to prove cause-and-effect relationship between variables
- Often involves a control group and treatment group to highlight difference caused by manipulating independent variable
- Control group does not imply NO treatment, it just implies different treatment; depends on what relationship is being observed
- Must ensure that the independent variable is the only factor that impacts the dependent variable (no other confounders)
  - Achieved using random assignment

Random assignment: impartial assignment of subjects to treatment and control groups; used to control confounders
- Assures that the subjects in both groups should generally be similar in all aspects with less confounders

Overstating: subjects in the control group performing poorly due to the knowledge that they are at a disadvantage

Understating: subjects in the control group performing better due to working harder to make up for disadvantages

Placebo: inactive substance/intervention that looks the same as the actual treatment
- Used to balance the effects of over/understating in control group

Blinding: not informing subjects which group they're in (single blinding) and not informing researchers of the groups to avoid bias (double blinding); cannot control confounders that are naturally assigned like gender

Observational study

Observing individuals and measures the variables of interest, usually without any direct/deliberate manipulation of variables by researchers
- Used if there are ethical issues with using controlled experiments (like human experiments)
- Lacks convincing evidence of a cause-and-effect relationship between variables; offers association, not causation
- Includes non-exposure and exposure groups
- Subjects are automatically assigned to either group

Slicing: used to control confounders in observational studies

Rates

$$\frac{\text{Number of values in sub-category}}{\text{Total across all categories}} = \text{rate}$$
$$0\% \leq \text{rate}(X) \leq 100\%$$
$$0 \leq \text{rate}(X) \leq 1$$

Contingency table:

	B	Not B	Row total
A	w	x	$w+x$
Not A	y	z	$y+z$
Column total	$w+y$	$x+z$	$w+x+y+z$

Conditional rate: rate X given Y – rate(X | Y)
- Read by column or row (using column/row total)
- rate(A | B) = w / (w + y)

Joint rate: rate X and Y – rate(X and Y)
- Read by single cell (using overall total)
- rate(A and B) = w / (w + x + y + z)

Association

Positive association: presence of A when B is present is stronger compared to when B is absent

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

Negative association: presence of A when B is present is weaker compared to when B is absent

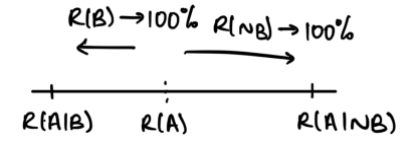
rate(A | B) < rate(A | NB)

No association: presence of A is not affected by presence of B

rate(A | B) = rate(A | NB)

Symmetry rule: rate(A | B) > rate(A | NB) ⇔ rate(B | A) > rate(B | NA)

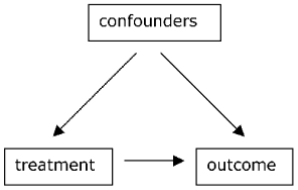
Basic rule on rates: the overall rate(A) will always lie between rate(A | B) and rate(A | NB)
- The closer rate(B) is to 100%, the closer rate(A) is to rate(A | B)
- If rate(B) = 50%, then rate(A) = ½ [rate(A | B) + rate(A | NB)]
- If rate(A | B) = rate(A | NB), then rate(A) = rate(A | B) = rate(A | NB)
  - rate(B) ≠ rate(NB) is fine



Simpson's Paradox

Phenomenon where a trend appears in more than half of the groups of data but disappears (no longer associated) or reverses when the groups are combined (inverse of slicing the data)

Confounder: (third) variable that is **must be** associated with both the independent and dependent variables whose relationship is being investigated
- Countered with random assignment



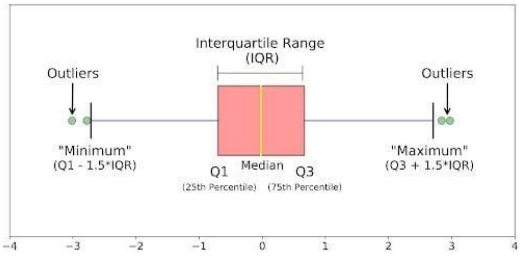
- Simpson's Paradox implies confounders, but the converse is not true

Distribution (Univariate)

Orientation of data points broken down by their observed number or frequency of occurrence; only for numerical data points, categorical does not work

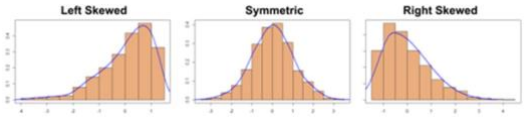
Histogram: graphical representation that organizes numerical data points into ranges or bins; particularly useful with large data sets
- No best way to derive number of bins
- Avoid too large or too small bin widths

Boxplot: distribution of numerical data points using five-number summary (min, Q1, median, Q3, max)
- Used to visualize outliers better
- Whiskers extend to the smallest/largest value that is not an outlier
- Distance between max/min and median tell us how much variability there is; larger distance => more variability => skewed in that direction => long tail
- Provide information about median, not mean or proportion per quartile or standard deviation

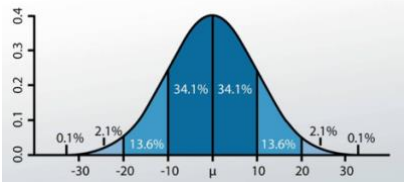


Outlier: value > *Q3 + 1.5 x IQR* OR < *Q1 - 1.5 x IQR*
- Useful when identifying any strong skewness in a distribution
- Should not be removed unnecessarily unless they have minimal effect on conclusion and no reason to be there
- Affects the mean the most; median and mode are robust statistics

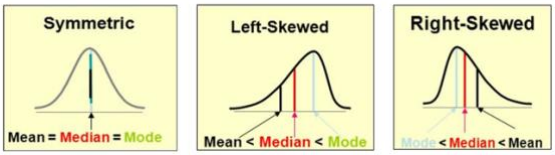
Shape:
1. Peaks: unimodal means one distinct peak; bimodal means two distinct peaks; multimodal means many distinct peaks
2. Skew: look at where the tail is, not where the peak is



E.g. normal distribution/bell curve: symmetrical distribution



Central tendency: mean, median, mode (tend to follow the pattern but now always)



Spread: how data varies around the central tendency; measured using standard deviation and range

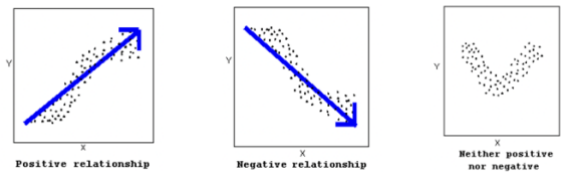
Bivariate

Relationship between two variables

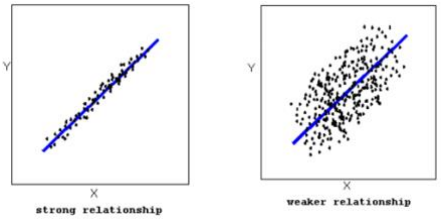
Deterministic: relationship is exact so one can derive the other and vice versa

Non-deterministic/Statistical: relationship/association is not exact; find an average of another variable given one; not a causal relationship

Direction: positive (one increase, the other increase), negative (one increase, the other decrease), neither

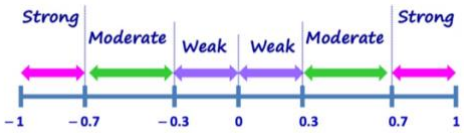


Form: general shape of scatter plot; linear (straight line) or non-linear (smooth curve)
Strength: closeness of points; strong (tight clustering of points); weak (spread out points)



Trendline: show relationship of scatter plot (see blue line)
Correlation coefficient (*r*): measure of linear association between two numerical variables; ranging between -1 and 1 which is used to summarise the direction and strength of a linear association

- Direction: r > 0 => positive; r < 0 => negative
- Strength: r -> 1 or r -> -1 => straight line; r = -1 or r = 1 => perfective positive/negative association; r = 0 => no association
- No linear association =/=> no association
- Unaffected by interchanging x and y variables, or adding/multiplying a constant to all numbers
- Outliers can strengthen or weaken correlations or not at all
- Not equals to gradient
- Deterministic linear relationship (straight line) => r = 1 or r = -1
- Points that form an ellipse have a weak correlation
- Points that lie above y=x =/> any association needed



Computation:
1. Compute the mean and standard deviation of both variables, x and y
2. Convert all data from each variable to standard unit:

$$\frac{x - \bar{x}}{s_x}$$

3. Compute xy in standard units
4. Sum xy and divide by n - 1

Linear regression
Modeling the relationship between two variables by fitting a straight line to the observed data

$$y = mx + c$$

where

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

- Always passes through that point of averages for the dataset
- **x can measure y but not the other way around**
- Predictions are only applicable to within the range of the independent variables; should not be used to make predictions outside of the range
  - Check carefully for range in exam
- Gradient changes if the values are all altered by some constant (multiply or add)
- Gradient => increase in y when x increases by 1
- Gradient not equal in X vs Y and Y vs X but follows same sign
- Predictions are not definitive; check carefully

Error (e_i): how far the predicted value is from the observed value
Method of least squares: finding a trendline that minimizes the overall sum of squares of errors (square used to avoid sum evening out to 0)
Exponential relationships:

$$y = cb^t \Rightarrow \ln y = \ln c + t \ln b$$

- Apply natural logarithm on y and plot against t to find a LINEAR relationship between ln y and t (typically exponential relationship)
  - Solve for ln c and ln b to find the linear regression line for ln y

Probability
Probability experiment: must be repeatable and allows for all possible outcomes which is known as the *sample space*
Event: sub-collection within sample space

- Probability is the total probability that the outcome of the experiment is an element of the event
- $A \subseteq S \wedge |A| \neq 1 \Rightarrow$ event where A is a sub-collection within a sample space
- $A \subseteq S \wedge |A| = 1 \Rightarrow$ outcome = event (outcomes are events)

Mutually exclusive events: events cannot occur simultaneously; does not imply that must have equal probability of occurring
Probability rules:

1. 0 <= P(E) <= 1
2. P(S) = 1 if S is the entire sample space
3. P(E ∪ F) = P(E) + P(F) if E and F are mutually exclusive
4. $P(E \cap F) = 0$ if E and F are mutually exclusive
5. If E and F are mutually exclusive, then P(E) + P(F) must be <= 1

Uniform probability: way of assigning probabilities to outcomes such that equal probability is assigned to every outcome in the finite sample space; 1/N probability for every outcome
Conditional probability: P(E | F) or probability of E given F; also known as conditional rates

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

- If there is no overlap between E and F, then P(E | F) = 0
- If P(F) = 0, then P(E | F) = 0
- P(A | B) = 0 if A and B are mutually exclusive

Law of total probability: if E, F, G are events from the same sample space S such that E and F are mutually exclusive and E ∪ F = S

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

True positive rate: aka sensitivity of test

$$P(\text{testing positive} \mid \text{actually positive})$$

- Among those with a disease, sensitivity% will test positive
- Does not mean that if the test is used, then sensitive% will be actually positive

True negative rate: aka specificity of test

$$P(\text{testing negative} \mid \text{actually negative})$$

- Same attributes as sensitivity

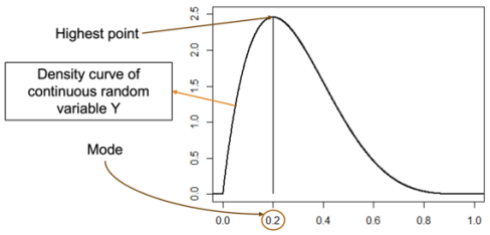
Base rate: basic probability without any conditional probability
Independent events: P(A) = P(A | B) meaning that B does not affect A; e.g. tossing coin and rolling a die; P(A | B) = P(B | A)

$$P(A) \times P(B) = P(A \cap B)$$

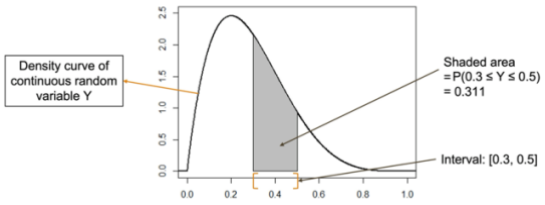
Random variable: numerical variable with probabilities assigned to each of the possible numerical values taken by the variable; e.g. rolling a dice

- Applicable to both discrete and continuous numerical variables
- Mode of a discrete random variable is the value of x that attains the highest y
- P(condition) = P(event A) + P(event B) + … where event A, B are events that satisfy the condition

Density curve: continuous series of points given a continuous random variable



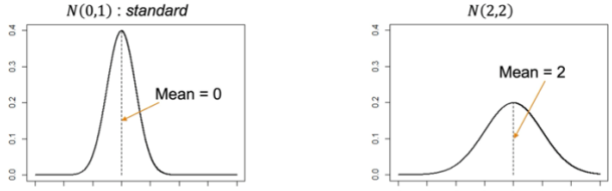
- Y-axis is the probability density, not probability
  - Area under the curve is the probability y takes on the values between range of x



- Area under density curve is 1
- Mode is the value of x with the highest point on the density curve

Normal distributions
Class of continuous random variables, denoted by N(x, y) where x is the mean and y is the variance

- Completely described by mean and variance
- Density curve of continuous random variable that is normally distributed is always bell-shaped
- Peak of the curve is the mean implying that mode = mean
- Density curve is symmetrical about the mean implying that median = mean
- If spread is larger, curve is flatter with lower peak



Statistical inference/sample statistic
Use of samples to draw inferences or conclusions about the population in question

- Sample statistic = population parameter + bias + random error
  - Becomes population parameter/proportion + random error if random sampling is used with perfect sampling frame with 100% response rate

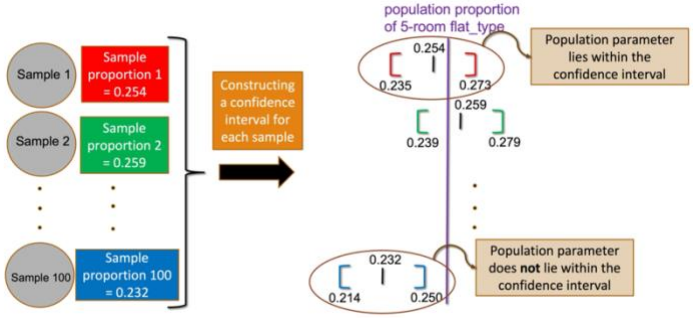
Confidence interval

Range of values that is likely to contain a population parameter based on a certain degree of confidence also known as the *confidence level*

- Used to quantify random error in every sample
- Population parameters to use: proportion, mean, standard deviation

Sample proportion = population proportion + random error assuming no bias

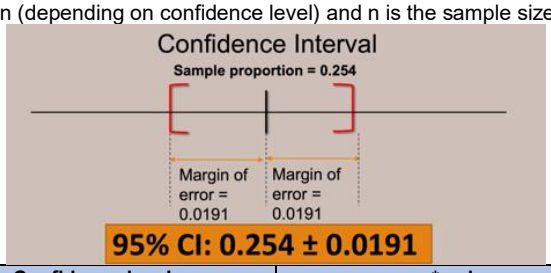
- Interpretation: margin of error (number after +/-) affects the width of the confidence interval
- Confidence of n% (not the same as n% chance) implies that if many simple random samples of the same size are taken and a confidence interval is constructed for each of them, then about n% of them will contain the population parameter
  - Population parameter remains unknown and fixed
  - Either in or not in confidence interval; uncertainty arises from sampling



Confidence interval for population proportion:

$$p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$$

where p* is the sample proportion, z* is the z-value from standard normal distribution (depending on confidence level) and n is the sample size

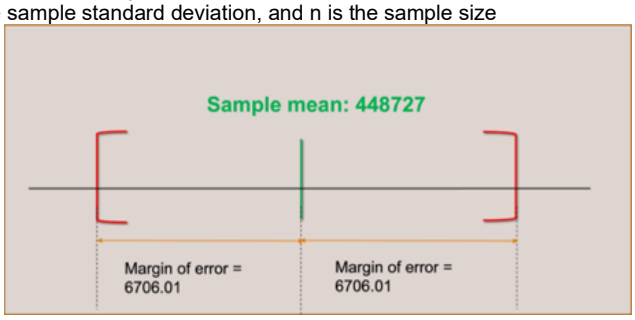


Confidence level	z* value
50%	0.67
75%	1.15
90%	1.645
95%	1.96
97%	2.17
99%	2.58
99.9%	3.29

Confidence interval for population mean:

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

where x̄ is the sample mean, t* is the t-value from the standard t-distribution, s is the sample standard deviation, and n is the sample size



Properties:

1. If the same sampling frame and method is used on a smaller sample size, then the confidence interval widens
2. The larger the sample size, the smaller the random error, which narrows the confidence interval
3. The higher the confidence level, the wider the confidence interval

Hypothesis testing

Only performed when only sample data and not information on the entire population is available; irrelevant if consensus or full population present

1. Identify the question and state the null hypothesis and alternative hypothesis (mutually exclusive states)
2. Collect the relevant data that is necessary for the test
   a. Introduce random variable and its probability distribution
   b. Include test statistic: the value computed/observed from the sample data that will be used to determine whether the null hypothesis is to be rejected or not
3. Set the significance level of the test
4. Make a conclusion on whether to reject or not reject the null hypothesis
   a. Reject the null hypothesis in favor of the alternative if p-value < significance level, else do not reject
   b. Cannot accept the null hypothesis or reject the alternative hypothesis
   c. Not rejecting != accepting

Null hypothesis (H_0): asserts the stand of no effect or no difference (maintains the truth to be disproven)

- Rejecting the null hypothesis at 5% => rejected at >5% but =/> rejected at < 5%

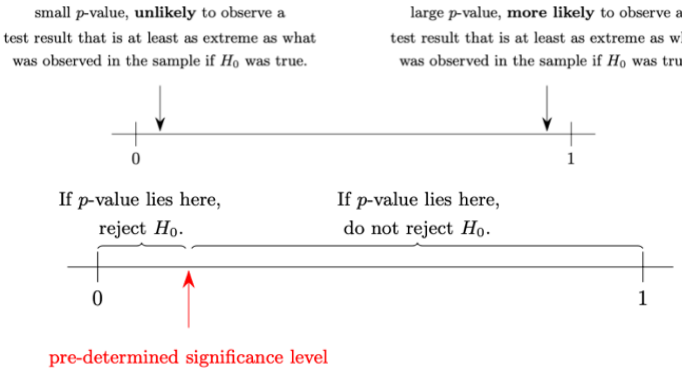
Alternative hypothesis (H_1): what is to be confirmed and pit against the null hypothesis

Significance level: how "convincing" the evidence needs to be before we can reject the null hypothesis in favor of the alternative hypothesis

- Between 0 and 1
- Lower the significance level, the "greater" the evidence needs to be before rejecting the null hypothesis
- Common value: 0.05 or 5% level of significance
- If p-value >= significance, do not reject null hypothesis as test is inconclusive
- If p-value < significance, reject the null hypothesis

p-value: probability of obtaining a test result at least as extreme as the result observed assuming the null hypothesis is true

- i.e. the probability of observing a test result that favors the alternative hypothesis at least as much as what is observed in the current sample while assuming that the null hypothesis is true
- Changes based on what's the null and alternative hypothesis



One-sample t-test: test difference between sample mean and a known or hypothesized mean;

- Used when sample size is < 30 (assume that population distribution of the variable is approximately normal) and when the sample was randomly produced

Chi-squared test: test whether two categorical variables are associated at the population level using their counts; data should be randomly produced

General GEA1000 Tips

1. If information is not present or something cannot be definitively concluded, then questions that say "must be true" will be false instead
2. Predictions using linear regression are not indicative of the actual value
3. Be very careful with indicating how linear association changes
   a. -ve to 0 => increase
   b. +ve to 0 => decrease
4. Any options about using chance or probability for confidence intervals is wrong
5. Any options using y to find x when linear regression is y = mx + b is wrong
6. For questions about sensitivity/specificity, draw the contingency table to assert
7. For questions about probability, list out all possible outcomes in the sample space and compute the probability from there
8. If events occur independently, do not use the above method, compute the probabilities separately and add