9. Regression

- used to make predictions on the basis of information that one has obtained
- Scatterplots can diagnose outlier points
- X variable → horizontal axis → predictor (independent variable)
- Y variable → vertical axis → criterion (dependent variable)

- The regression line is the best fitting line that can be drawn through the points on the scatterplot.
  → straight line merely approximates the data it represents → it can never pass through every value unless ~~it~~ it's corelation is perfect.
  → Regression line allows us to make predictions.

Regression equation:

$$\hat{Y} = bX + a$$

$\hat{Y}$ = predicted Y value - value on Y axis
b = slope of the line
a = the Y intercept point (point where the line intersect Y axis)
X = predictor variable (value on the x-axis)

→ Regression line minimizes the squared distances of the observed data points from the line.

Computing the values in the regression equation.

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

n = number of observation

ΣXY = total of X times Y (remember do each line (score) & then add together)

ΣX = total of x scores/values

ΣY = total of Y scores/values

ΣX^2 = total of X^2 scores (calculate each line and then total)

$(\Sigma X)^2$ = total of ΣX times by ΣX

- if the computed value is

possitive = possitive slope runs bottom left to top right /

negative = negative slope - top left to bottom right. \

$$a = \bar{Y} - b\bar{X}$$

$\bar{Y}$ = mean of Y values ($\Sigma Y/N$)

$\bar{X}$ = mean of X values ($\Sigma X/n$)

b = slope - (computed above)

∴ can now replace a + b in the equation..

You can now work out the regression line at each point by inserting each X score + calculating.

Drawing the regression line.
- Straight line

2 ways of drawing the line.

Method 1:

* calculate 2 points
* mark on scatter plot
* join 2 points to form a straight line

Method 2:

* use the Y intercept (a) + any other pair of numbers

$$\hat{Y} = bX + a$$

* calculate Y intercept by replacing X with zero in the formula ($X = 0$ on the X axis)
* calculate another point by replacing X with an score.

→ 'within dataset' predictions are called interpolations
→ predictions 'beyond' the range of values in the dataset are called extrapolations
→ predictions from regression equations are more accurate when they are made about points that fall within the range of points covered by original dataset.
→ Standard error of estimate measures the degree to which the regression lines "fits" the observed data. (Do not need to calculate)

→ The degree of scatter around the straight line is called the correlation

→ Correlation is a measure of the strenght and direction of the linear association between 2 variables

→ Casual Inferences should only be drawn from regression analyses with great caution

Standard error of estimate = (know what it is) no need to calculate.

→ if points on the original scatterplot have no clear pattern along with a wide scatter of points, then the line will have very little predictive power + will not be very meaningful.

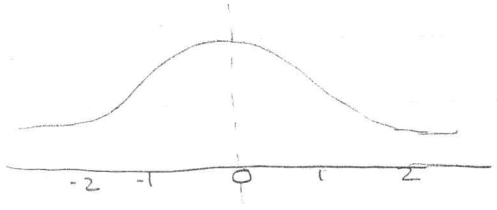
→ if regression line is a good fit - distances should be quite small - shorter distances from the regression line.

→ cannot simply average the distances → it could balance each other out → by using the standard error of estimate we can take account of ~~the~~ this difficulty (error)

10/ Normal distributions

→ normal distributions have the following characteristics

* Bell shape → clearly identifiable bell shape with one distinct peak → unimodal
* Symmetry → figure forms an exact mirror image on either side of the center line
* area or portion → below the normal distribution equals 100%. Since figure is symmetrical it means that the area on the left occupies 50% and the Right the other 50%



Standard normal distribution

→ distribution is defined by its shape, mean and variance

→ normal distributions allow us to determine where an individual score lies relative to other scores in a set of scores

→ Infinite number of normal distributions, each with an unique mean, and variance

→ to determine proportion of cases falling above or below a score in a distribution we need 3 pieces of information about the distribution
→ shape → mean → variance.

→ Standard normal distribution is a normal distribution with a mean equal to 0 and a variance equal to 1

→ z scores indicate the number of standard deviation units a score lies above or below the mean.

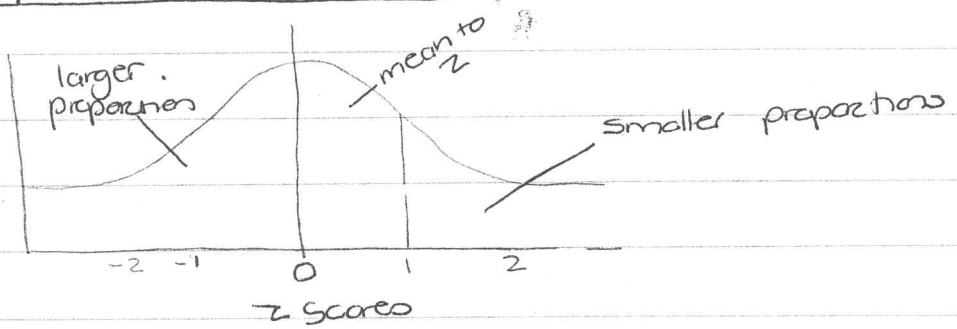
→ Standard normal distribution is defined in terms of standard deviation, which are called z scores

→ z-scores range from 3 standard deviations above it and 3 standard deviations below it.

→ The z-table tells us what proportion of the area under the curve of a z distribution lies above or below a particular z score.

* When working with z scores its always best to draw sketch of the distribution. Make the center z=0 all negative z scores are on the left and positive on the right

Using the z score tables



→ Although the z-table only gives us proportions lying above positive z scores ⊛

1) because the total area under the curve is 1, we can calculate the proportion below a positive z-score by subtracting the area above the z score from 1 and

2) because the normal distribution is symmetrical

⊛ the proportion lying above a positive z-score is the same as the area below the negative z score of the same absolute value.

- The z table allows us to look up proportions, given z-scores as well as determine z-scores given proportions.

- standard normal distribution is very convenient since any set of normally distributed scores can be converted to standard scores to enable us to directly compare data from different sets by using

$$z = \frac{X - \mu}{\sigma}$$

- once converted to standard scores, the standard table of z scores.

→ conversion does not change the distribution of scores only the way they are represented.

⇒ 2 fixed limits commonly used 95% + 99% z values are.

95% → ±1.96 → 99% ±2.58

11/ Basic concepts of probability

Probability → used in the definition of power
→ no simple definition for probability.

→ Probability is defined as $p = {}^{s}/_{n}$
→ number of sucesses divided by the total number of events.
→ events are said to be independent when they do not affect the probability of each other occuring

→ single event have more than 2 outcomes example:-
pull a card from a deck of cards and you have 52 possible outcomes.
→ may bet it will be hearts → only 13 hearts in the pack → the probability of sucess is thus

$$p = {}^{s}/_{n}$$
$$= {}^{13}/_{52}$$
$$= 0.25$$
$$= 1 \text{ in } 4 \text{ or } 25\% \text{ chance/probability.}$$

- Probabilities are based on a past evidence or on the chance of a random event occurring

→ events must be independent or we will not be able to easily estimate the probability of events taking place.

→ probability law of conjunctions (multiplicative law)

→ probability of 2 independent events jointly occurring is the product of their individual probabilities

$$p(a+b) = p(a) \times p(b)$$

Random sampling without replacement undermines independence of events.

∴ when 2 events are not independent from each other we must multiply the probabilities of the outcomes together.

example: tossing a coin twice + getting the same result twice → heads in the first and the second toss will be calculated as follows:

2 possible outcomes: heads
tails

∴ $p = {}^{a}/_{n}$

$= \frac{1}{2} = 0.5$

$= 50\%$ chance

∴ $(pa) \times (pb)$

$= 0.5 \times 0.5$

$= 0.25$

∴ a 25% chance of that happening

The same can be done will drawing a card a second time if card is ~~replac~~ placed back in card deck.

it will be 0.25×0.25

$= 0.0625 = 6\%$

-If the card is NOT placed back into the deck the deck

is reduced to 51 and the conditions has thus changed → the events are NO longer independent to each other and we have to calculate the new probability $^{12}/_{51} = 0.235$

⇒ now we multiply these probabilities

$= p(a) * p(b)$

$= {}^{13}/_{52} \times {}^{12}/_{51}$

$= 0.25 \times 0.235$

$= 0.05875$

$= 0.059$

Probability law of disfunction (additive law)

- the probability of either of 2 independent events occuring is the sum of their individual probalities

$p(a \text{ or } b) = p(a) + p(b)$

Probability of drawing a diamond or a heart

$= {}^{13}/_{52} + {}^{13}/_{52}$

$= 0.25 + 0.25$

$= 0.5$

Independant event → when the occurance of an event has no effect on the probability of an occurance of another

Mutually exclusive → when the occurance of one event precludes the occurance of another

Exhaustiveness → A set of events representing all probable outcomes

Law of disjunctions (Additive law) → events are mutually exclusive

Law of conjunctions (multiplicative law) → when events are independent.

12/ Sampling distributions and hypothesis testing.

Sample → subset of the population → generalize from the sample

It was proven that generalization from a sample is never to different from the result that would be obtained if the entire population average was ~~op~~ obtained ∴. It is justified to generalize from sample to a population

Sampling → the process of identifying + grouping a subset of an entire population

Random sampling means identifying and collecting the sample in such a way that every member of the population has an equal change of being in the sample.

∴ sample is representative of the population

Distributions → graphical representation of the ~~entire~~ entire set of data.

Sampling distributions of the mean → the frequency distribution of sample means, not individual scores.

→ one of the main function of statistics is to make inferences

Drawing inferences involves estimating properties of population from information about samples.

→ many ways to make such estimates → in everyday life people make unscientific inferences → which they treat this hearsay as fact.

→ To draw inferences a representative, random sample is required
→ It is impossible to gain access to the entire population ∴ researc relies on drawing of sample cases

Sampling means

when we sample repeatedly from the same population, we expect the means of these samples to be different.

→ Draw several samples from population + get a mean for each sample
→ now take the mean of all the samples and construct a frequency distribution for these means, we would have a sampling distribution of the mean for the samples

→ Plotting the means of an infinite number of samples of size n, drawn from a population will give us the sampling distribution of the mean.

→ eventhough the distributions look the same at first glance there are differences
  * Scores - different sample values
  * Means - means will be different as sample scores are different
  * Variances - different samples/means have different variances

Sampling distribution of the mean → distribution of an infinite number of sample means of a particular size randomly selected from a population

∴ impossible to compute the average of an population + therefor we compute the sample mean using the notation $\bar{X}$ ⇒ Assume everything that apply to $\bar{X}$ applies to the population μ

∴ we are ultimately interested in the population we use μ rather than $\bar{X}$

Hypothesis Testing

hypothesis → tentative statement of a relationship between 2 variables

Neuman's definition → educated guesses about how the social world works.

hypothesis → an idea → expectation that there is a correlation between 2 or more variables

Research hypothesis → define a statement about the variables
eg: there is a difference between height of women + men in South Africa.

How do hypotheses fit into the research process?
→ have to prove ~~*~~ a statement scientifically.
→ need to follow a process in order to prove or disprove the statement statistically (scientifically)

Sampling distribution of the mean → distribution of an infinite number of sample means of a particular size randomly selected from a population

∴ impossible to compute the average of an population + therefor we compute the sample mean using the notation $\bar{X}$ ⇒ Assume everything that apply to $\bar{X}$ applies to the population μ

∴ we are ultimately interested in the population we use μ rather than $\bar{X}$

Hypothesis Testing

hypothesis → tentative statement of a relationship between 2 variables

Neuman's definition → educated guesses about how the social world works.

hypothesis → an idea → expectation that there is a correlation between 2 or more variables

Research hypothesis → define a statement about the variables
eg: there is a difference between height of women + men in South Africa

How do hypotheses fit into the research process?
→ have to prove ~~it~~ a statement scientifically.
→ need to follow a process in order to prove or disprove the statement statistically (scientifically)

9 steps in statistical / hypothesis testing

1) * formulate a null hypothesis → a statement that maintains that there is no difference between the groups or conditions. It is represented by the symbol H_0

→ negative statement about a positive difference between 2 or more variables

eg: There is no difference between the height of men + the height of woman

or Men and women are equally tall

Formulated as: $H_0 = \mu_m = \mu_w$.

OR $\mu_m - \mu_w = 0$

2) Formulate the alternative hypothesis

→ affirmative statement. ∴ there is a difference between variables.

* non-directional → 2 variables are not equal. => there is a difference between the height of women + men

* directional → one variable is smaller than the other ⇒ men are shorter than women

* directional → one variable is larger than the other ⇒ men are taller than women

example: There is a difference between the job be performance between women + men

$H_1: \mu_m \neq \mu_w$

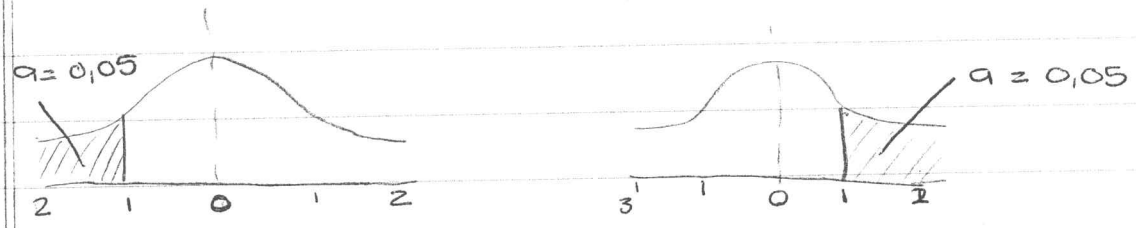
$H_1: \mu_m - \mu_w \neq 0$

3) Determine whether the test is one-tailed or two tailed?

→ one tailed → directional → are taller < or >

→ two tailed → non directional → not sure then =

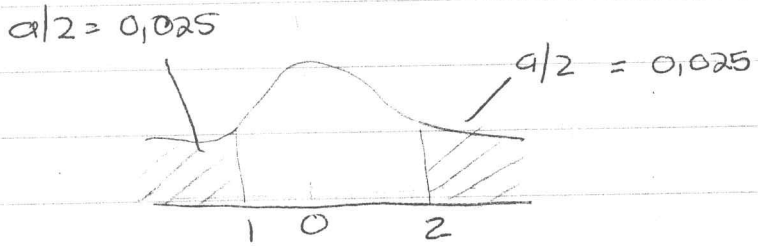
Directional - one tailed < or >



Rejects extreme outcomes at only one end of the distribution

Non-Directional - two tailed

=



4) Determine the level of significance.

work with 5% or a 1% level of significance

type I error ⇒ rejecting a null hypothesis when its true

II error ⇒ not rejecting a null hypothesis when its false

∴ probability with which we ~~reject~~ are willing to reject the null hypothesis when its correct

5) Compute the test statistic.

→ use a few test statistic

T test

f test (chi - square test)

when we want to determine whether there is an significant difference between sample means

→ Test statistics in which the mean plays a role can be classified in 2 categories: those relating to only 2 samples (t-tests) and those relating to more than 2 samples (F test)

2 samples → t test for related samples

t test for unrelated samples

3 or more samples → F test (one way analysis of variance)

1 or more variables with Frequencies - chi-square test

Test Statistics F + Chi-square refer to the answers that you obtain when testing a hypothesis ~~Empe~~ empirically (statistically) and to where the calculating of the test-statistics fit into the 9 steps of hypothesis testing.

6) Determine the degrees of freedom

→ need to determine the degrees of freedom as to interpret, your computations properly.

7) Determine the critical value.

take into account

* whether it is a one tailed or 2 tailed test.
* level of significance
* degrees of freedom

to determine critical value.

critical value is an indication of where rejection region begins + indicates that all scores above or below this point fall into the 5% or 1% level of significance.

8) Reject or do not reject the null hypothesis

* if test statistic > critical value reject Ho
* if test statistic < critical value do not reject Ho

~~Also~~ Always work with absolute values -> ignore the negative sign + read as positive.

9) Interpret the finding

- Interpret the rejection or non rejection of the null hypothesis in relation to the original research problem.

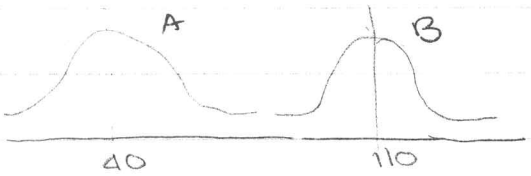
NB Make sure you can write down the 9 steps

13 hypothesis tests applied to means t-tests

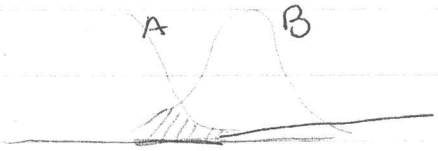
t-test is used to compare 2 (estimated) population means

→ compare normal distributions

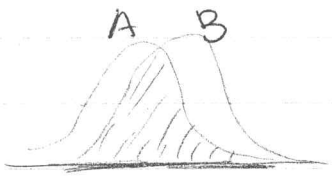
How do you know if the distributions is from the same population.



A - lower scores
B - higher scores
∴ not overlapping - score from A unlikely to appear in B



Distributions overlap.
Scores can belong to A or B
Seperate but similar as they do overlap a number of scores.



- overlap remarkably
- appear the same

T-test allows us to determine the degree to which 2 distributions overlap.

The degree to which 2 distributions overlap is determined by the difference between the means and variance of each sample.

$$t = \frac{\text{difference between the means}}{\text{Standard error}}$$

$$z = \frac{\text{difference between score + population mean}}{\text{Standard deviation of the population}} = \frac{x - \mu}{\sigma}$$

Central difference between t-test and the z-score transformation is that the population parameters (μ + σ) do Not have to be known to perform a t-test.
This is important because in the large majority of cases the population parameters are unknown.

⇒ Unlike a z-test no population parameters are needed to perform a t-test

Standard error is the standard deviation of the sampling distribution of the mean (the distribution created by taking repeated sample means from a population)

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

To analyse data with a t-test, the data need to comply with the assumptions of normality, homogeneity of variance and independence.

purpose of hypothesis testing with 2 independent means is to help you decide whether an observed difference between 2 sample means is accidental or whether it represents a real difference between populations

- Independent samples

* sampling distribution of difference between means is the distribution of an infinite number of a particular size randomly selected from a population

* because we sample each population independently, the sample means will also be independent

* mean of the sampling distribution will be $\mu_1 - \mu_2$

Formula used for t-value of 2 independent Groups.

* requires the use of the variance of both groups
  If you do not have the variance, you must compute it using:

$$S^2x = \frac{\Sigma X^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

Do the same for Y (second variable)

Now compute the t-value
Steps 1-9 Chapter 12

- Formulate the null hypothesis
- Formulate the alternative hypothesis
- Determine if the test is 1 or 2 tailed
- Determine level of significance
- Compute test statistic

$$= t = \frac{\bar{X}_1 - X_2}{\sqrt{\frac{S^2}{N_1} + \frac{S^2}{N_2}}}$$

- Determine degrees of freedom

$$df = N_1 + N_2 - 2$$

- Determine critical value (table A1.2)
- Determine the null hypothesis should be rejected/not rejected
- Interpretat the finding

Pooling variances → when we deal with samples of unequal sizes

- Use the pooled variance as your standard error estimate unless the sample variances are <u>not</u> homogeneous

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

n_1 = sample size of 1st Sample

n_2 = Sample size of 2nd sample

S_1^2 = variance of first sample

S_2^2 = variance of second sample.

T-test independent samples

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} \quad \text{or} \quad t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_2^p \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

X_1 = mean of 1st sample

X_2 = mean of 2nd sample

$S_{\bar{X}_1 - \bar{X}_2}$ = estimate of the standard error.

Use separate variance estimates if the variances of your 2 samples are highly different (one variance is at least 4 times as big as the other)

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

S_1^2 = variance 1st sample

S_2^2 = variance 2nd sample

n_1 = sample size of 1st sample

n_2 = sample size of 2nd sample

heterogeneity variance → when samples are drawn from the populations have different variances

→ we compensate for heterogeneity by computing the t-test seperately for the different variances (not pooling them)

smaller N is then used to read the degrees of freedom

13 Non parametric equivalent of the t-test

↳ statistical tests that do not rely on parameter estimation and/or distributional assumptions

↳ Mann-Whitney test:

↳ tests whether 2 independent samples have the same median

Wilcoxon matched pairs test → tests wheter 2 related samples have the same median

wilcoxon → signed ranks test

→ rank order differences (ignore sign of difference)

Related Samples

↳ samples that are dependent.

↳ 2 sets of data on the participants (same participant)

↳ we expect data to corrdate.

∴ 2 groups of people/participants are matched according to a variable in 1 group. → Group 2 assigned a member that corresponds with the member of group 1 in respect of the particular variable.

∴ Particular pairs are selected in respect of the relevant variables → age, height, eye colr can be used to match the persons/variables

2 ways of obtaining related sample scores

→ by matching

→ when 1 participant contributes 2 score

* Compute the mean $\bar{D}$
* standard diviation (s_D) of different scores

t - test for related groups

$$t = \frac{\bar{D} - 0}{\frac{s_D}{\sqrt{N}}}$$