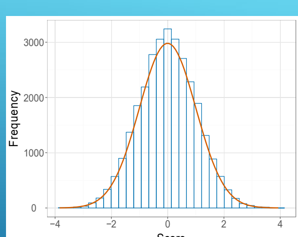
FUNDAMENTALS OF STATISTICS

1

$$\text{Outcome}_i = (\text{Model}) + \text{error}_i$$

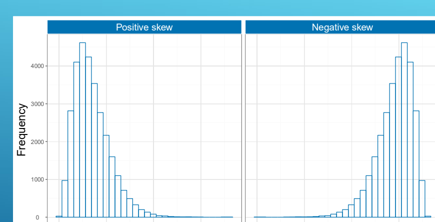
THE ONLY EQUATION YOU WILL EVER NEED

2



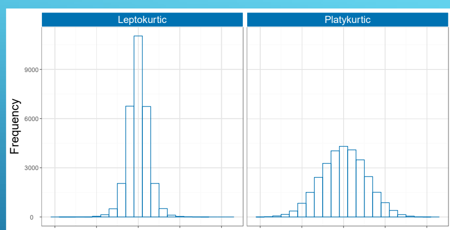
THE NORMAL DISTRIBUTION

3



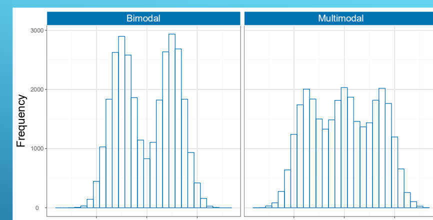
SKEW

4



KURTOSIS

5



BIMODAL AND MULTIMODAL DISTRIBUTIONS

6

THE MEAN

- The mean is the value from which the (squared) scores deviate least (it has the least error).

$$\text{Mean } (\overline{X}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

7

- The mean is a *model* of what happens in the real world: the *typical* score
- It is not a perfect representation of the data
- How can we assess how well the mean represents reality?

MEASURING THE 'FIT' OF THE MODEL

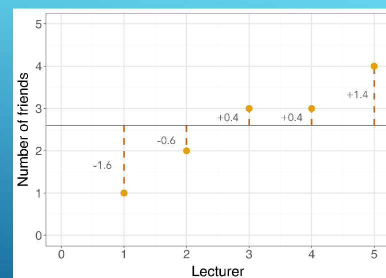
8

- A deviation is the difference between the mean and an actual data point.
- Deviations can be calculated by taking each score and subtracting the mean from it:

$$\text{deviance} = \text{outcome}_i - \text{model}_i$$

CALCULATING 'ERROR'

9



10

USE THE TOTAL ERROR?

- We could just take the error between the mean and the data and add them.

Score	Mean	Deviation
1	2.6	-1.6
2	2.6	-0.6
3	2.6	0.4
3	2.6	0.4
4	2.6	1.4
	Total =	0

11

- We could add the deviations to find out the total error.
- Deviations cancel out because some are positive and others negative.
- Therefore, we square each deviation.
- If we add these squared deviations we get the Sum of Squared Errors (*SS*).

SUM OF SQUARED ERRORS

12

Score	Mean	Deviation	Squared Deviation
1	2.6	-1.6	2.56
2	2.6	-0.6	0.36
3	2.6	0.4	0.16
3	2.6	0.4	0.16
4	2.6	1.4	1.96
		Total	5.20

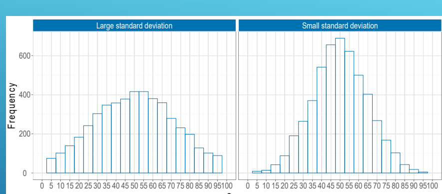
$$SS = \sum (X - \overline{X})^2 = 5.20$$

13

- Variance (s^2) = $\frac{SS}{N-1}$
- Remember the variance is squared so we use the SD.
- s= $\sqrt{s^2}$

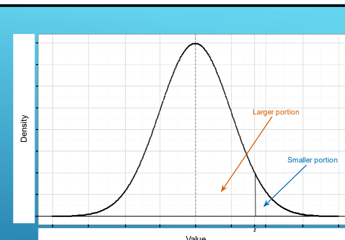
VARIANCE AND STANDARD DEVIATION (S)

14



THE SD AND THE SHAPE OF A DISTRIBUTION

15



THE NORMAL PROBABILITY DISTRIBUTION

16

- Z-scores
  - Standardising a score with respect to the other scores in the group.
  - Expresses a score in terms of how many standard deviations it is away from the mean.
  - The distribution of z-scores has a mean of 0 and *SD* = 1.
    - Mean = 0 requires centering around zero (X- X̄)
    - SD = 1 requires division by *S* to find *Z*

$$z = \frac{X - \overline{X}}{s}$$

GOING BEYOND THE DATA: Z-SCORES

17

- X̄ - Mean
  - $y = \bar{x} + e$
- *b* – OLS
  - $y = a + e$
  - $y = a + b(x) + e$
- *r* – Correlation
- Not perfect, but error is minimized (e.g. OLS)

PARAMETERS AND PARAMETER ESTIMATES

18

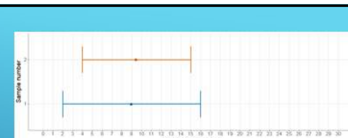
Standard Error

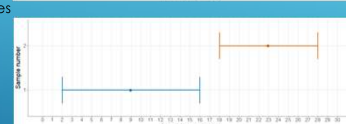
Also called - Standard Error of Sample Means

$$\sigma_{\overline{X}} = \frac{s}{\sqrt{N}}$$

19

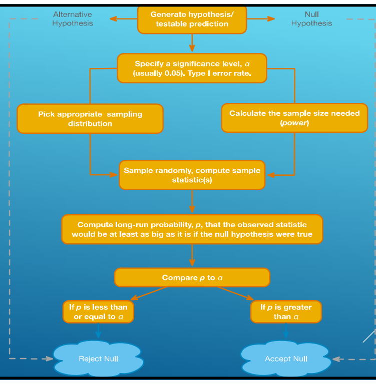
- 95% is most common
- Lower Boundary = X̄ - (CV x *SE*)
- Upper Boundary = X̄ + (CV x *SE*)
- CV (Critical Value) = 95% of Z scores
  Between -/+1.96



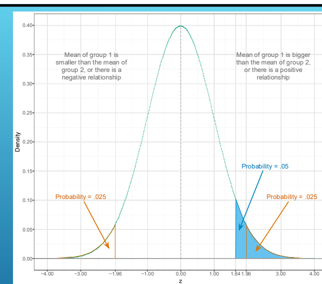


CONFIDENCE INTERVALS

20



21



ONE- AND TWO-TAILED TESTS

22

PROBABILITY VS. ODDS

Probability

- Range 0-1
- P(A) = $\frac{Number\ of\ Event\ A}{Total\ Number\ of\ Events}$
- Conversion
  - P(A) = $\frac{Odds(A)}{1+Odds(A)}$
- P-Values P-Value Video
  - P<.05
  - P<.01
  - P<.001

Odds

- Odds(A) = $\frac{Number\ of\ Event\ A}{Number\ of\ Non\ Event\ A\ (1-A)}$
- Conversion:
  - Odds(A) = $\frac{P(A)}{1-P(A)}$
- Odds Ratios

23

TEST STATISTICS

- A Statistic for which the frequency of particular values is known.
- Observed values can be used to test hypotheses.

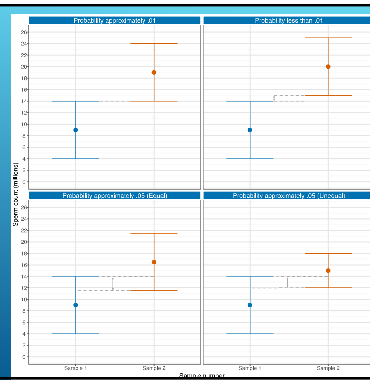
$$\text{Test statistic} = \frac{\text{signal}}{\text{noise}} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}}$$

24

- Type I error
  - Occurs when we believe that there is a genuine effect in our population, when in fact there isn't.
  - The probability is the α-level (usually 0.05)
- Type II error
  - Occurs when we believe that there is no effect in the population when, in reality, there is.
  - The probability is the β-level (often 0.2)

TYPE I AND TYPE II ERRORS

25



CONFIDENCE INTERVALS AND STATISTICAL SIGNIFICANCE

NOTE: TALK ABOUT SAMPLE SIZE AND STATISTICAL SIGNIFICANCE

26

- An effect size is a standardized measure of the size of an effect:
  - Standardized = comparable across studies
  - Not (as) reliant on the sample size
  - Allows people to objectively evaluate the size of observed effect.

EFFECT SIZES

27

- There are several effect size measures that can be used:
  - Cohen's *d*
  - Pearson's *r*
  - Odds Ratio/Risk rates
- Pearson's *r* is a good intuitive measure
  - Oh, apart from when group sizes are different …

EFFECT SIZE MEASURES

28

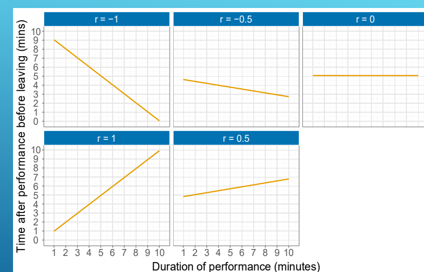
$$\hat{d} = \frac{\bar{X}_1 - \bar{X}_2}{s}$$

$$s_p = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}$$

COHEN'S *D*

29

- We'll cover this measure later in detail, but.....



THE CORRELATION COEFFICIENT, *R*

30

- *r* = .1, *d* = .2 (small effect):
  - the effect explains 1% of the total variance.
- *r* = .3, *d* = .5 (medium effect):
  - the effect accounts for 9% of the total variance.
- *r* = .5, *d* = .8 (large effect):
  - the effect accounts for 25% of the variance.
- Beware of these 'canned' effect sizes though:
  - The size of effect should be placed within the research context.

EFFECT SIZE MEASURES

31

- OR = $\frac{Odds(A)}{Odds(B)}$
- OR = 1 = Same
- OR > 1 = Positive
- OR < 1 = Negative
- Log Odds = LN(OR)
- Conversion
  - OR = exp(Log Odds)

ODDS RATIOS

32