



# RMHI PALS

Friday 26th April 9am-12pm

# Disclaimer

We are **not** professionals.

We're all students here so don't just feel free to chime in every once in a while,

We'd go as far as to encourage you to criticise us as much as you can!

You're all free to discuss with each other in groups if you'd prefer.

(No shouting or screaming though.)

And we would recommend having something to write on or type on open.

# Structure

1. Review of Core concepts
2. Review of Association concepts
3. Review of Prediction concepts
4. Review of Difference concepts

Again, we would like to note that we would prefer everyone to engage with and discuss the material, so feel free to ask as many questions as you'd like!

# Core Concepts: Research Questions

3 parts to a good RQ: Constructs, Population and Relationships.

A research question is not a research hypothesis.

A research hypothesis is not a statistical hypothesis.

# Core Concepts: Constructs, Measures and Score

Constructs are unobservable attributes that we label to describe hypothetical behaviours.

Measures are proven (valid and reliable) methods to obtain a numerical representation of unobservable constructs.

Construct scores are what measures output- the numerical representation of unobservable constructs.

The scaling of these scores can be arbitrary or meaningful.

# Core Concepts: Manipulating scores

4 different types of scores:

- a) Raw/Observed scores: untouched metric from measure,
- b) Deviation score: scaled to the metric mean,
- c) Standardised score: scaled to a specified mean and a specified amount of standard deviations per unit,
- d) Z score: a type of standardised score with mean 0 and 1 standard deviation per unit.

# Sidebar: Standard Deviation (SD)

```
sample1 <- c(1:5) #vector of construct scores
mean(sample1) #mean

#standard deviation
sample1
sample1-mean(sample1) #subtract the mean for the set of deviation scores
(sample1-mean(sample1))^2 #square them to remove the +/-s (sum of any set of
deviation scores is 0)
sum(((sample1-mean(sample1)))^2) #add them all up
sum(((sample1-mean(sample1)))^2)/(length(sample1)-1) #divide them by sample
size minus one.

sqrt(sum(((sample1-mean(sample1)))^2)/(length(sample1)-1)) #square root them
to return them to the sd metric

sd(sample1) #compared to the sd function

(sample1-mean(sample1))/sd(sample1) # a z score with mean 0, 1 unit per 1 sd

(sample1-mean(sample1))/(2*sd(sample1)) # a standardised score with mean 0, 1
unit per 2 sd

#i.e. why z scores are pretty good- 1-to-1 unit/sd metric, easier to
interpret.
```

# Sidebar: SD Exercise- Match the Output

```
## [1] 1 2 3 4 5
```

```
## [1] 2.5
```

```
## [1] -2 -1 0 1 2
```

```
## [1] -1.2649111 -0.6324555 0.0000000 0.6324555 1.2649111
```

```
## [1] 10
```

```
## [1] 3
```

```
## [1] -0.6324555 -0.3162278 0.0000000 0.3162278 0.6324555
```

```
## [1] 1.581139
```

```
## [1] 4 1 0 1 4
```

```
## [1] 1.581139
```



# Side-sidebar: Why $n-1$ ?

Why  $n-1$ ? (Bessel's correction) it allows us to get an unbiased estimate of population variance. It's a long story, but essentially as long as we believe that a population parameter always defines the population distribution- and we don't have the entire population at hand, our sample is always missing something. the "-1" is there to acknowledge that, while parameters define the RVs we are trying to represent through the calculated statistics we make, we subtract the one to account for the one value in the population that is not free to vary. By representing that concept in the sample variance denominator it becomes less biased. don't quote me on that.

# Core Concepts: Popu Paras and Sample Stats

## Similarities:

- Numeric representation of summary characteristics

## Differences:

- Multiple sample statistics (multiple random samples)
- A single theorised population parameter.

# Core Concepts: Random Variables

A variable consisting of the set of all possible outcomes defined by one or more population parameters.

I understand RVs to be the perpetually unknowable variable defined by the existing parameters of the natural world.

If it's easier to grasp as an infinite set ranging from infinity to negative infinity—that may be correct in some situations.

# Core Concepts: Random Variables

What we do know is that they come in two flavours:

- a) Continuous (e.g. height, temperature, etc.) If naturally occurring decimals make sense- it's a continuous variable (13.4 cm, 34.2 celsius)
  
- b) Discrete (or Factor in R) (e.g. measure responses, no. of errors made, etc.)  
Ones where naturally occurring decimals make no sense.

What matters is that we know these numbers behave differently- don't take continuous variables to be discrete and vice versa.

# Sidebar: Statistical framework

The notions of probability we're working with belongs to the Frequentist framework, where probability is considered 'in the long run'.

We do not go into the Bayesian framework where probability is updated.

# Core Concepts: Distributions

The set of different numerical values defined by summary characteristics.

Different distributions:

- a) Sample distribution: the set of different scores collected,
- b) Population distribution: the hypothetical distribution of all scores,
- c) Mathematical probability distribution: theoretical distributions that are defined by population parameters,
- d) Sampling distribution: distributions of sample statistics (one meta-level up).

# Core Concepts: Distributions

The first two refer to distributions of a single data entries.

The other two are a little different-

We know that there can be multiple sample statistics, but only one population parameter.

This means that sample stats have a distribution of their own- they behave like random variables themselves!

Neat applet: [http://onlinestatbook.com/stat\\_sim/sampling\\_dist/](http://onlinestatbook.com/stat_sim/sampling_dist/)

# Core Concepts: Describing Distributions

Features to look out for:

- a) Shape: +ve skew (right skew), -ve skew (left skew), bell-shaped, symmetric
- b) Centre: typically mean, median for skewed data, mode for ordinal/nominal data, etc.
- c) Spread: SD, IQR



# Sidebar: Standard Error

The standard deviation of a sampling distribution is called a standard error-why?

A way to think about this is that a sampling distribution talks of the set of different statistics you can get, so the spread in that distribution is the typical amount the statistic is different by, i.e. the typical amount of error from your computed statistic.

# Sidebar: Observed Test Statistic

An application of the sampling distribution (I believe) is the calculation of the observed test statistic.

If you know what Z-scores are, you know what observed test statistics are.

This is your t-statistic, chi-square statistic, regression coefficient, etc.

Corresponds with a p-value (more on that in a bit).

# Core Concepts: Theoretical Probability Distribution

Sampling distributions are cool, but these are even cooler- because they exist by mathematical functions and don't need an unfeasible amount of resources to make.

A standardised sampling distribution is equivalent to a theoretical probability distribution.

As seen in the applet above, fitted normal curves are equivalent to sampling distributions- so we have the distributions that have the same properties as the sampling distributions without actually needing to construct them manually.

# Sidebar: Estimation

We use sample stats to estimate popu paras.

Estimation is the process of calculating parameter values based on sample data- point estimation is for single values, interval estimation is for a range of values.

Estimation is the process of using an estimator to produce an estimand.

# Sidebar: Properties of Estimators

- a) Unbiased. The notion of bias is distance away from the real population parameter. So we want as unbiased of an estimate as possible.
- b) Consistency. If bias decreases as sample size increases, that estimator is considered to be consistent. So if we have a biased estimator, we'd like it to be consistent.
- c) Efficient. The smaller the standard error, the more efficient. If we encounter two estimators, we'd prefer the more efficient one (provided it isn't biased and inconsistent)

# Core Concepts: NHST

Good ol' null hypothesis significance testing.

Yes, I know there's been less of an emphasis on constructing and interpreting p values, but I do think that knowledge on the p-value (and consequently how it can be manipulated and misinterpreted) is very important.

So what's a p-value? The p-value is a widely used statistical inference method to evaluate statistical hypotheses (e.g. null hypotheses). Note that statistical inference is different from scientific inference, scientific inference may involve statistical inference, but you cannot say that a statistical inference implies a scientific inference. No. The tools we use are not infallible and reflects the researcher's intentions of inferring ideas from data.

# Core Concepts: NHST

Set null and alternative hypotheses,

Decide the alpha (typically .05 and consequently the critical test statistic( $T_{crit}$ )),

Calculate the observed test statistic ( $T_{obs}$ ) from relevant sample stats,

Find the p-value corresponding to the  $T_{obs}$  of the data,

If  $p < \alpha$ , reject (consequently if  $|T_{obs}| > |T_{crit}|$ , reject) otherwise, retain.

# Sidebar: p-value rant.

The tobs (on both sides) represent the distance between the  $h_0$  and the tobs. In other words, the tobs is directly equivalent to the range of RV explained in the theoretical probability distribution (the p value). i.e. probability of less than 1 in a continuous dist = probability of  $0 + -1 + -2 + \dots + -\text{infinity}$

$P(\text{Tobs} | h_0 \text{ is true})$  that is ALL p-value can say.

The probability of observing your observed test statistic given that you assume the null to be true. That's it.

Why do we do this: to declare if something isn't 0 = declaring that it's SOMETHING.

It's kind of an upside-down inference. (blame Popper for this one)



# Sidebar: What the p-value is NOT.

It's NOT:

The probability of the scientific hypothesis being true,

The probability of your alternative statistical hypothesis being false,

The probability of your null hypothesis being true,

The probability of your null hypothesis being false,

The probability of observing your sample statistic.

# Core Concepts: Confidence Intervals (CIs)

CIs are a function of p-values (and of standard error in general)

The margin of error (interval area) is defined by what we don't consider to be statistically significantly different- and therefore 'plausible'.

Remember, the sampling distribution represents multiple samples- and values within the confidence interval are 'plausible' as a function of the preset alpha value and standard error of the sampling distribution to determine what values are 'plausible' in other samples.

In other words, it depends on the standard error value and our preset alpha criterion value.

# Sidebar: Alpha criterion

These serve to control type-1 error rates, and define boundaries of statistical significance.

$P(\text{Rejecting } H_0 | H_0 = T)$  is what the alpha represents.

In the metric of sampling distributions, the alpha criterion is directly equivalent to the critical test statistic (the value that the observed test statistic has to exceed to declare statistical significance).

# Core Concepts: Effect Size

Last one. Hurray!

Kind of an ambiguous term to refer to a method of describing the strength of a relationship.

Examples are Pearson's  $r$ , Cramer's  $V$ , standardised regression coefficients or R-squared.

Effect size measures can also have their own statistical assumptions, and it is important to evaluate the effectiveness of certain measures based on your data (and its possible violated assumptions).

E.g. Bonett's delta, Hedges'  $g$

# Association

Systematic co-occurrence.

'These things occur together' is what association represents.

# Association: Continuous Variables

We use pearson's correlation coefficient (I just call it Pearson's  $r$ )

But first, scatter-plots and covariance!

A neat game: <http://guessthecorrelation.com/>

# Association: Covariance

The covariance of the plots directly refer to how two variables vary together- how much a sample varies on two variables together. In other words, it can tell you the strength (in its own metric) and direction of how much a sample systematically varies on two variables.

But here's a problem- it's in its own metric, so how do we change it to be meaningful?

We standardise it!

And that is what correlation is- standardised covariance. Or to be specific, covariance of z-scored variables.

# Association: Pearson's $r$

$r$  ranges from -1 and +1 and captures the strength and direction of association between two continuous variables.

' $r$ ' is the sample coefficient, the population coefficient is symbolised as rho ( $\rho$ ).



# Association: Categorical Variables

Contingency/Frequency tables!

The Cramer's V statistic is essentially a rescaled chi-square statistic to fit between 0 to 1.

The higher the V the stronger the association.

No direction since the chi-sq stat doesn't have a direction.

# Association: Odds

Odds are the probability of something occurring relative to it not occurring.

i.e. 4:1 odds mean 80% to 20%, 9:1 odds mean 90:10.

20% chance it'll rain = 4:1 odds of not raining. (4 times more likely for it to not rain than it is to rain)

10% chance I'll bring my umbrella = 9:1 odds of not bringing my umbrella (9 times less likely for me to bring my umbrella than me bringing my umbrella)

# Association: Odds Ratios

Odds ratios (OR) take it a step further. The OR represents the ratio of two odds:

$$(9/1 / 4/1) = 2.25$$

It is 2.25x more likely for me to not bring my umbrella when it doesn't rain.

The reciprocal holds true too

$$4/9 = 0.44 (4/1 / 9/1)$$

It is 0.44x more likely for me to bring my umbrella when it doesn't rain.

# Association: Odds Ratios

Provided you don't mess with the raw numerical values (never mess with raw values!!), you can twist your odds ratio around to brain-numbing ways.

1/9 / 4/1 odds ratio of me bringing my umbrella when it doesn't rain

It is 1/36x more likely for me to bring my umbrella when it doesn't rain relative to not bringing my umbrella when it does.

9/1 / 1/4 odds ratio of me not bringing my umbrella when it rains

It is 36x more likely for me to not bring my umbrella when it rains relative to bringing my umbrella when it doesn't rain.

# Cheeky SAQs

- 1) What is the difference between a set of deviation scores and that set's standard deviation?
- 2) Can a p-value of 0.09, equivalent to an observed test statistic of -1.65, be considered statistically significant if we've set the critical test statistic to be 1.6? (Arbitrary values, assuming a two-tailed null hypothesis)
- 3) What is the value of a 0% Confidence Interval?
- 4) Why are values in a confidence interval considered to be 'plausible'?
- 5) Why can covariance and correlation be negative, whereas variance can only be a positive value?

# Answers to said Cheeky SAQs

- 1) A set of deviation scores refer to a set of scores that have been scaled to the mean. Standard deviation is the typical amount of deviation in the set of deviation scores, as computed through the variance (sum of any set of deviation scores equal to 0) and then squared to return the value to its own metric.
- 2) Yes, it is considered statistically significant because the absolute observed test statistic is larger than the critical test statistic.
- 3) A 0% confidence interval would return the sample statistic used to compute the interval. The confidence level is determined by (1-alpha level), and a confidence level of 0 would mean an alpha level of 1- everything would be statistically significant, so there would be no range of plausible values. This is less a practical question and more on the understanding of constructions of CIs.

# Answers to said Cheeky SAQs

4) Values within a confidence interval are considered 'plausible' as an effect of the level of significance and the computed standard error. The notion of 'plausibility' refers to how the range of values cannot be considered statistically significant from the computed statistic- and are therefore possible in other samples

5) Covariance and Correlation can have negative values since they are not squared values. Variance is squared, so it can never have a negative value.

# Prediction

Prediction of a variable usually refers to use of regression, specifically for this course, we'll be considering two types of regression:

- a) Simple linear regression,
- b) Multiple linear regression.



# Prediction: Not Association

Correlation	Linear regression
Linear association between 2 variables	Prediction between dependent and independent variables
Bi-directional	Non bidirectional
<b>Correlation coefficient</b> indicates the strength of association between 2 variables (how 2 variables vary together)	<b>Regression coefficient/Slope of regression line</b> indicates the impact of one unit change in independent variables on dependent variables.
Pearson's $r$ ranges from -1 to +1	$R^2$ measures strength of prediction, ranging from 0 to 1.

# Prediction: Concepts

- Variance: The total amount of variability between observed scores and the mean. Can be referred to as 'sums of squares' (SS), and there are different types of sums of squares in regression models ( $SS_{total} = SS_{reg} + SS_{res}$ )
- Standard deviation: Square root of the variance, back to the scale's metric.
- Correlation Coefficient: Standardised covariance coefficient.
- Covariance: An unstandardised measure of calculating how much two variables vary together.

Standardised measures are typically more informative- since they are interpretable in terms of standard deviation opposed to the raw metric.

# Prediction: Linear Regression

The Full simple linear regression line is expressed as:

$$Y = a + bX + e$$

Y is the dependent variable (predicted)

X is the independent variable (predictor)

a is the intercept of the line

b is the slope of the linear function- and indicates the direction of the relationship between variables.

And e is the 'error' term- also called residual.

# Prediction: Linear Regression

The simplified linear regression line is written as

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

There aren't any error terms in the simplified linear regression line because  $\hat{Y}$  refers to the predicted value of the dependent variable. In other words,  
 $Y = \hat{Y} + \text{error}$

# Prediction: Least Squares Regression

We will be using Ordinary Least Squares (OLS) regression for estimating our regression line.

What OLS regression does is minimise the value of the error terms- or to minimise the sums of squares of residuals (minimising the proportion of variance attributed to residuals)

# Prediction: Multiple Linear Regression

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_nX_n + e$$

The multiple regression line can be considered to be very similar to the simple regression line, but with multiple independent variables- and this results in a very important consideration.

# Prediction: Partialling out variables

Independent variables (IV) like to vary with each other- this can lead to difficulties in inferring a direction of causality. If we say that B predicts A in a model where B and C are very highly correlated, the reality could be that C predicts A, but it appears that B predicts A because we're not considering C's influence on B and A.

The notion of 'partialling' out variables is to hold other variables constant- to ensure that they don't vary while we vary the IV of interest. By doing so, we can infer the predictive effect of that IV, holding constant all other IVs.

# Prediction: R-squared

R-squared (Coefficient of determination) is the proportion of variance explained by the regression model ( $SS_{\text{reg}}/SS_{\text{total}}$ ).

It is a method to assess the strength of the complete regression model (all IVs are taken into consideration).

Another important aspect to consider is that  $SS_{\text{total}}$  is additively decomposed,  $SS_{\text{reg}} + SS_{\text{res}}$  will always equal to  $SS_{\text{total}}$ .



# Prediction: R-squared considerations

R-squared is not the same thing as adjusted R-squared.

The computed R-squared statistic may be biased if:

- a) There are a large amount of predictors,
- b) There is a small sample size.

The adjusted R-squared returns a less biased statistic of the R-squared, so pay attention if the model has a large amount of IVs and a small sample size.

# Prediction: Strength of IVs

- a) Standardised regression coefficients: by using standardised data instead of data in its own metric, we can evaluate our model in terms of standard deviation and can allow you to compare strengths of multiple IVs- since they're now on the same metric: SD.
- b) Semi-partial correlation: the correlation between the chosen IV and the DV, holding all other IVs constant.
- c) Squared semi-partial correlation: The proportion of variance the chosen IV explains in the DV, holding all other IVs constant.

# Prediction: Statistical Assumptions

1. Independence of observations: one participant's scores are observed and recorded independently of other participants.
2. Linearity. Linearity between the predictors and the predicted variable (multiple IVs and the one DV)
3. Homoscedasticity (Constant residual variance). This big chunk of a word ensures that errors vary independently (errors aren't associated to IVs).
4. Normality of residuals. We'd like our residuals to be normally distributed.

# Prediction MCQs

1. For the following regression line, which option is incorrect?

$$Y = 2.6 - 0.41X_1 + 0.51X_2$$

- A. When  $X_1$  increases by 1 and  $X_2$  increases by 1,  $Y$  will increase by 0.10
- B. The value of  $Y$  is 2.6 when scores on all independent variables are zero
- C. Holding the scores on  $X_1$  constant, 1 unit increase in  $X_2$  predicts 0.51 unit increase in  $Y$ .
- D. Holding the scores on  $X_2$  constant, 1 unit increase in  $X_1$  predicts 0.41 unit decrease in  $Y$ .

# Prediction MCQs

2. Which of the following sentences is correct about R-squared values?

- A. When R-squared gets larger, the average size of residuals also gets larger.
- B. R-squared statistics indicates the variation of dependent predicted by each independent variable, holding scores on other independent variables constant.
- C. The huge difference in observed R-squared and adjusted R-squared indicate strong bias in the value of observed R-squared.
- D. R-squared will be less biased when there are more IVs.

# Prediction MCQs

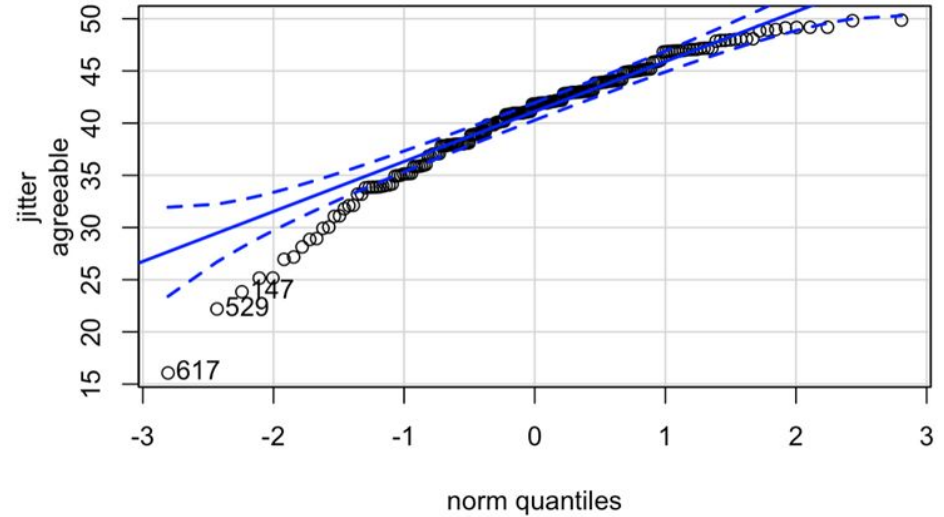
3. Which of the following description is closest to the concept of heteroscedasticity?
- A. Constant residual variance on different predicted values of dependent variables.
  - B. Non-constant residual variance on different values of independent variables
  - C. Residuals are distributed equally above and below zero
  - D. Residuals are not distributed equally above and below zero

# Prediction MCQs

4. Under which conditions is the R-squared value least biased?
- A. Sample size and the number of IVs are large
  - B. sample size and the number of IVs are small
  - C. R-square statistic is computed repeatedly
  - D. Sample size is large but the number of IVs is small

# Prediction MCQs

5. What can be inferred from the following QQplot?
- A. residuals distribution is slightly negatively skewed
  - B. assumption of residual normality is violated
  - C. residuals distribution is slightly positively skewed
  - D. outliers in this distribution are very influential





# Prediction MCQs

6. Which of the following sentences is the most correct interpretation of the following analysis?

- A. Openness is a statistically significantly stronger predictor of "need for cognition" than agreeableness
- B. Agreeableness is a significant predictor of cognition
- C. The evidence is consistent with no prediction by agreeableness
- D. Openness is a strong predictor of cognition

```
Call: lm(formula = cognition ~ agreeable + openness, data = dat.reg)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	16.46652	4.80669	3.426	0.000746
agreeable	0.08754	0.11260	0.777	0.437799
openness	1.09611	0.11333	9.672	< 2e-16

Residual standard deviation: 8.261 on 197 degrees of freedom

Multiple R-squared: 0.3872

F-statistic: 62.23 on 2 and 197 DF, p-value: < 2.2e-16

AIC	BIC
1417.17	1430.36

	Estimate	2.5 %	97.5 %
(Intercept)	16.46652060	6.9873398	25.9457014
agreeable	0.08754218	-0.1345043	0.3095887
openness	1.09610761	0.8726102	1.3196050

# Prediction MCQ Answers

1. For the following regression equation, which option is the incorrect?

$$Y = 2.6 - 0.41X_1 + 0.51X_2$$

- A. When  $X_1$  increases by 1 and  $X_2$  increases by 1, Y will increase by 0.10
- B. The value of Y is 2.6 when scores on all independent variables are zero
- C. Holding the scores on  $X_1$  constant, 1 unit increase in  $X_2$  predicts 0.51 unit increase in Y.
- D. Holding the scores on  $X_2$  constant, 1 unit increase in  $X_1$  predicts 0.41 unit decrease in Y.

# Prediction MCQ Answers

2. Which of the following sentences is correct about R-squared statistics?
- A. When R-squared gets larger, the average size of residuals also gets larger
  - B. R-squared statistics indicates the variation of dependent predicted by each independent variable, holding scores on other independent variables constant.
  - C. The huge difference in observed R-squared and adjusted R-squared indicate strong bias in the value of observed R-squared.
  - D. R-squared will be less biased when there are more IVs.

# Prediction MCQ Answers

3. Which of the following description is closest to the concept of heteroscedasticity?

- A. Constant residual variance on different predicted values of dependent variables.
- B. Non-constant residual variance on different values of independent variables
- C. Residuals are distributed equally above and below zero
- D. Residuals are not distributed equally above and below zero

# Prediction MCQ Answers

4. Under which conditions that R-squared is least biased?
- A. Sample size and the number of IVs are large
  - B. sample size and the number of IVs are small
  - C. R-square statistic is computed repeatedly
  - D. Sample size is large but the number of IVs is small

# Prediction MCQ Answers

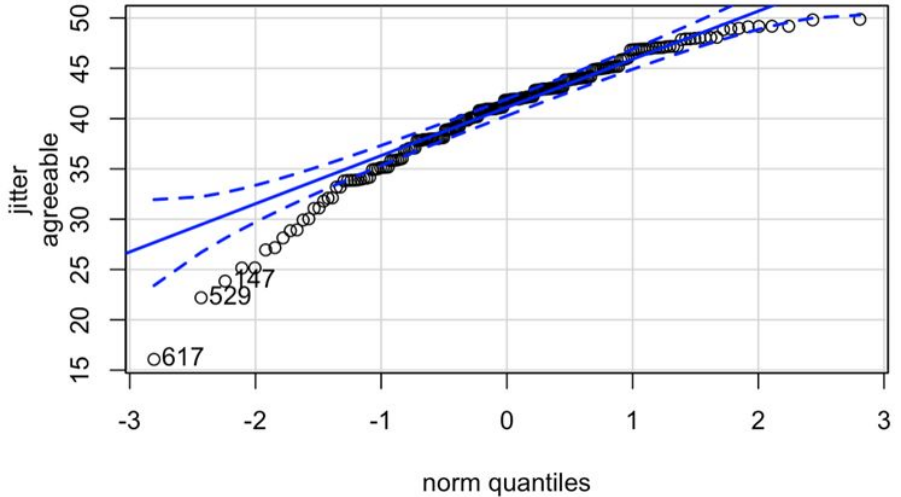
5. What can be inferred from the following QQplot?

A. residuals distribution is slightly negatively skewed

B. assumption of residual normality is violated

C. residuals distribution is slightly positively skewed

D. outliers in this distribution are very influential



# Prediction MCQ Answers

6. Which of the following sentences is the most correct interpretation of the following analysis?

- A. Openness is a statistically significantly stronger predictor of "need for cognition" than agreeableness
- B. Agreeableness is a significant predictor of cognition
- C. The evidence is consistent with zero prediction probability by agreeableness
- D. Openness is a strong predictor of cognition

```
Call: lm(formula = cognition ~ agreeable + openness, data = dat.reg)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	16.46652	4.80669	3.426	0.000746
agreeable	0.08754	0.11260	0.777	0.437799
openness	1.09611	0.11333	9.672	< 2e-16

Residual standard deviation: 8.261 on 197 degrees of freedom

Multiple R-squared: 0.3872

F-statistic: 62.23 on 2 and 197 DF, p-value: < 2.2e-16

	AIC	BIC
	1417.17	1430.36


	Estimate	2.5 %	97.5 %
(Intercept)	16.46652060	6.9873398	25.9457014
agreeable	0.08754218	-0.1345043	0.3095887
openness	1.09610761	0.8726102	1.3196050

# RESEARCH QUESTIONS FOR GROUP DIFFERENCES

A dark blue, diagonal, triangular shape that originates from the bottom-left corner and extends towards the top-right corner, covering the lower half of the slide.



1. Examine **probability functions** encountered in the course



# PROBABILITY FUNCTIONS

- Construct confidence intervals
- Calculate p-values for null hypothesis tests

## DENSITY FUNCTION

- Specify the probability of random variable falling within a particular range of values
  - Bell-shaped curve
- look at probability at one point

## CUMULATIVE DISTRIBUTION FUNCTION

- Specify the probability of a corresponding continuous random variable by calculating
  - confidence interval
  - P-values
- look at the total probability of anything below it

# MCQs

Normal density function is defined by

Normal Probability Density Function

- A. Z-score
- B. Variance
- C. Mean
- D. B and C

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# MCQs

Normal density function is defined by

- A. Z-score
- B. Variance
- C. Mean
- D. B and C

Normal Probability Density Function

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

=> defined by parameters mean and variance


# Core Concepts: Probability functions

Note that the horrible formula above refers to the function that is responsible for determining the normal distribution.

The probability functions create mathematical probability distributions- which we then use to construct p-values and confidence intervals.

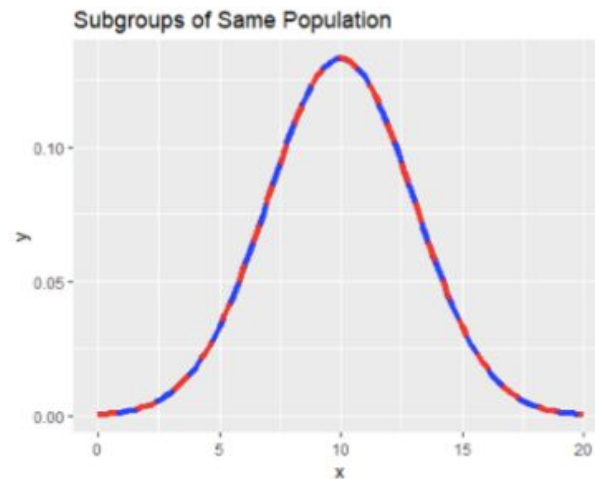
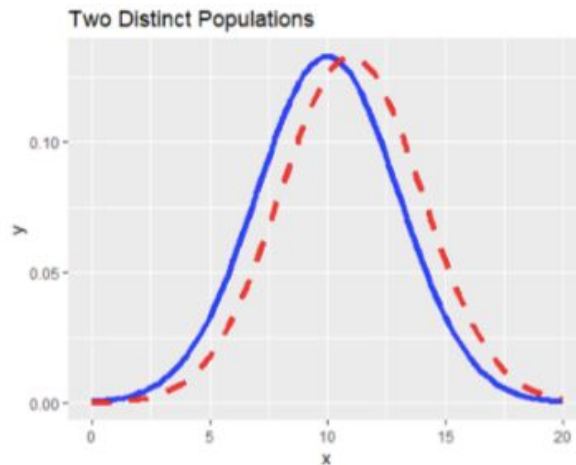
Tl;dr these functions are determined by parameters, and are useful in statistical inference.

## 2. Research questions addressing differences between **TWO GROUPS**



# RQs for differences between two groups

- Distinct populations
- Subgroups within the same population



# Differences

When we talk of differences, we are asking how two groups vary on a construct score.

In this instance, the mean difference between the two scores are being examined.

When the two groups are a subgroup of the same population, we would expect the two means to be equal to each other.



3. Understand how two groups  
can be formed as either  
being independent  
or dependent



# Group Formation

- Mutually-exclusive groups (independent)
- Mutually-Paired Groups (Dependent)

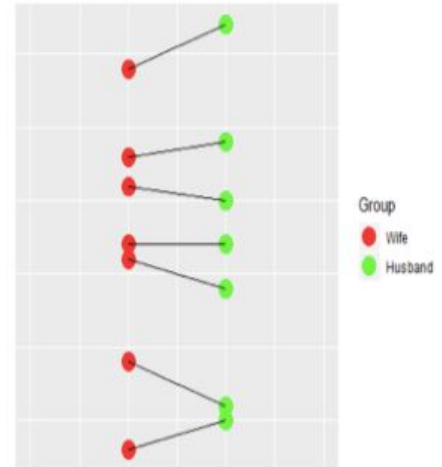
## Two **mutually-exclusive** groups

- each score in one group is independent of all scores in the other group
- participants can only belong to one group
- the size of each group does not necessarily have to be the same



## Two **mutually-paired** groups

- each score in one group is linked to one particular score in the other group
  - same person being measured twice
  - two people having a common dependency
- the size of each group must be the same by design



# MCQs

Which example represents dependent groups?

- A. RMHI and ARMP study
- B. Time 1 and Time 2 study
- C. Female and Male study
- D. Jersey and Friesian study

# MCQs

Which example represents dependent groups?

- A. RMHI and ARMP study
- B. Time 1 and Time 2 study**
- C. Female and Male study
- D. Jersey and Friesian study

# Differences: Independence

When we talk of Independent and Dependent groups- we specifically refer to the notion of independence in terms of the data.

In dependent data, we would expect the data to be related to the other in some way- this can be time series or paired data.

In independent data, we would expect the two groups to be independent of one another- we don't expect group 2 to vary if we varied group 1.

## 4. Know how to undertake investigations of two group differences



# MEAN DIFFERENCES between two INDEPENDENT GROUPS

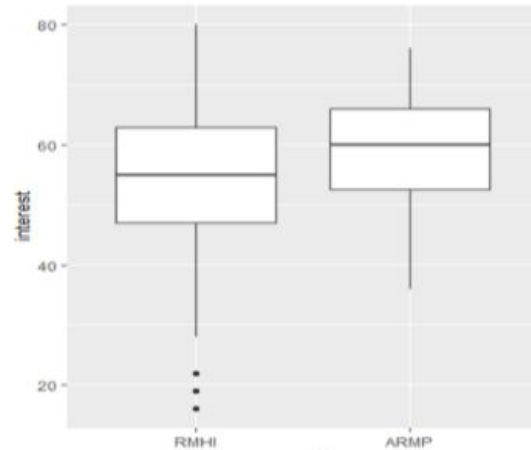
Median

Whiskers

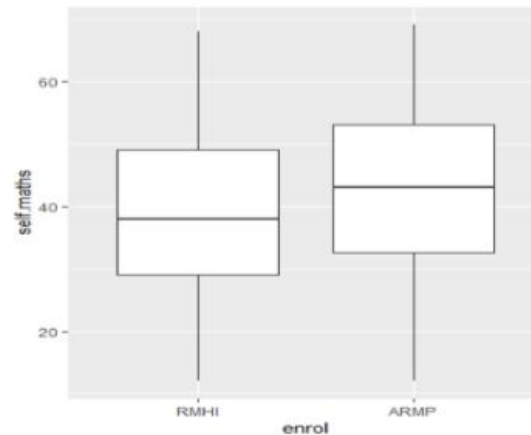
Outliers

Symmetric distribution

Is there a difference in interest in statistics between ARMP and RMHI students in 2019?



Is there a difference in maths self-concept between ARMP and RMHI students in 2019?



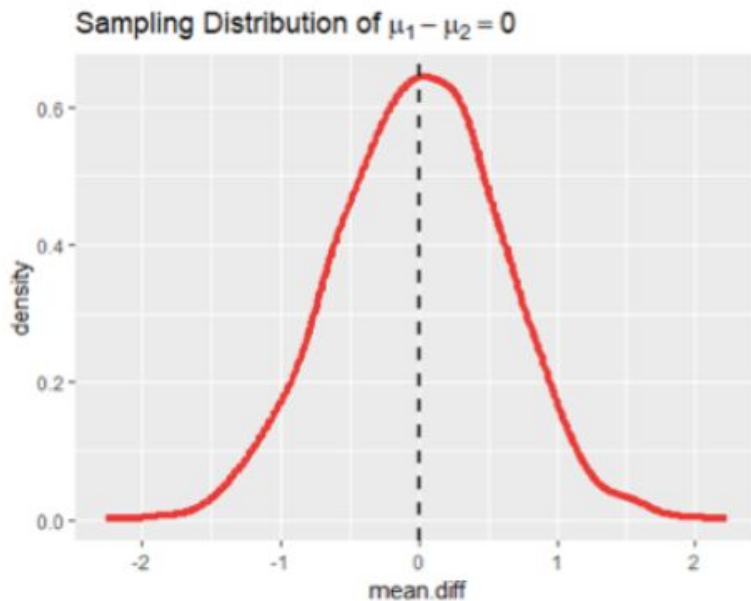
# TRUE MEAN DIFFERENCES vs SAMPLING VARIABILITY

There will always be a DIFFERENCE between sample means of two groups:

- random sampling variability when groups from the same populations
- random sampling variability plus a difference in population means when groups from different populations



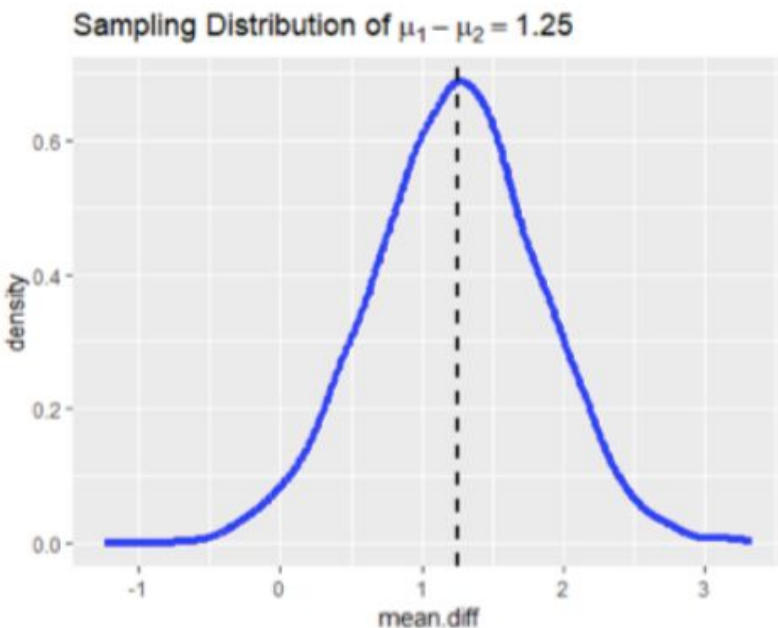
# SAMPLING DISTRIBUTIONS



## SAME POPULATION

- no difference in population means
- sampling distributions center around 0
- Sampling distribution of group mean differences from the same population means will have a mean difference in the sampling distribution = 0

# SAMPLING DISTRIBUTIONS



## DIFFERENT POPULATION

- Sampling distribution of group mean differences from the different populations will have differences in population means

# MCQs

Negative sample mean differences indicate:

- A. two distinct groups from different populations
- B. two subgroups from the same population
- C. the first experimental group has high sample mean than the second one
- D. the second experimental group has high sample mean than the first one

# MCQs

Negative sample mean differences indicate:

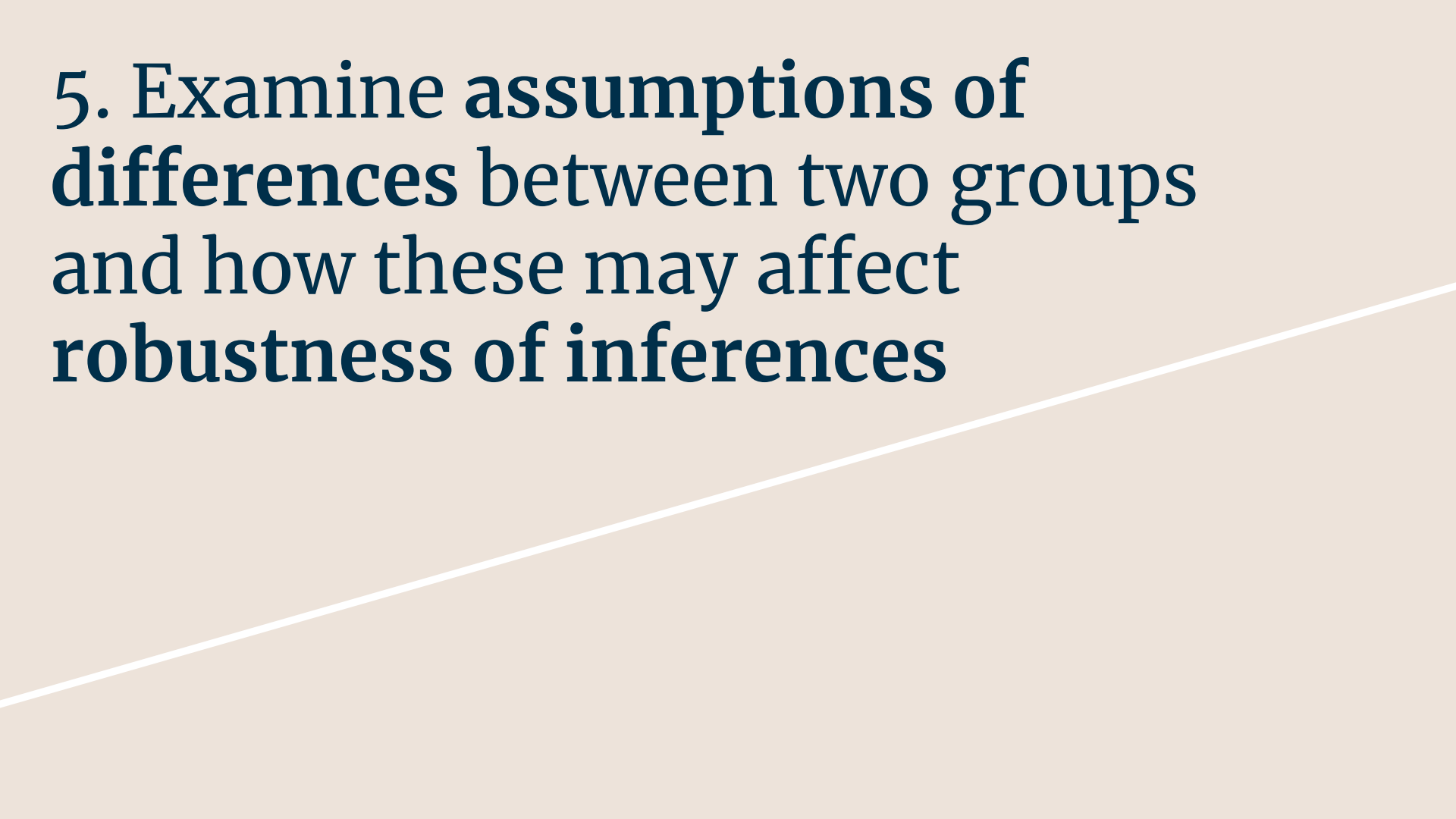
- A. two distinct groups from different populations
- B. two subgroups from the same population
- C. the first experimental group has high sample mean than the second one
- D. the second experimental group has high sample mean than the first one

# Differences: Sampling Distribution in t-tests

The t-test examines the sampling distribution of the mean difference- so the parameter in question is the mean of group 1 - mean of group 2.

If the two groups did not differ, we would assume the mean difference to be 0, and if the two groups did differ, the mean difference would not be 0.

**5. Examine assumptions of differences between two groups and how these may affect robustness of inferences**



# ASSUMPTIONS for MEAN DIFFERENCES between two INDEPENDENT GROUPS

1. Observations are independent
2. Observed scores on the construct measure are normally distributed
3. Variances in two groups are the same

- Levene's Test
- Fligner-Killeen's Test



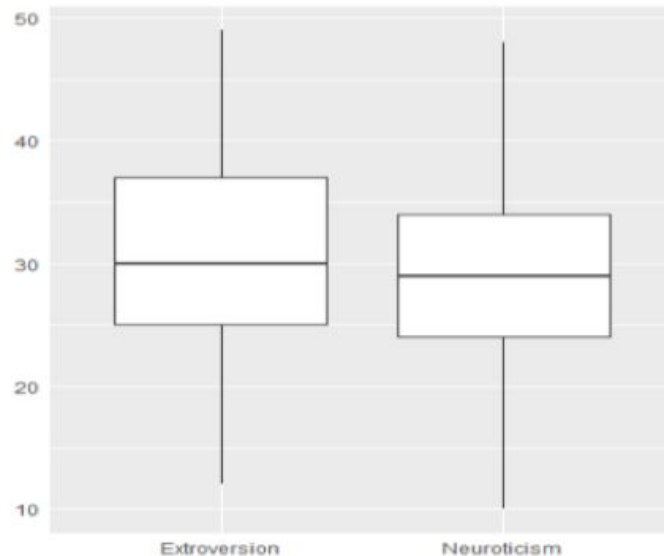
Homogeneity of variance assumption

- Balanced/ Unbalanced design
- Standardized/ Unstandardized confidence intervals against non-normality

# MEAN DIFFERENCES between two DEPENDENT GROUPS

## Difference Score

Is there a difference in level of extraversion and neuroticism between ARMP and RMHI students in 2019?



TWO POSSIBILITIES at the population level:

1. if the population mean difference scores is 0, then there is no difference between the two dependent groups on the construct
2. if the population mean differences scores is not 0, then there is a difference between the two dependent groups on the construct



# ASSUMPTIONS for MEAN DIFFERENCES between two DEPENDENT GROUPS

1. Observations are independent
2. Observed scores on the construct measure are normally distributed
3. Homogeneity of variance assumption is not relevant because the analysis is undertaken on the different scores

## **Possible effects of violation in sample data:**

The robustness of confidence intervals to violation of normality assumptions depends on which type of interval is being used

- Standardized confidence intervals are not robust against mild non-normality in different scores
- Unstandardized confidence intervals are robust against mild-to-moderate non-normality in different scores

# Differences: Statistical Assumptions

For independent t-tests:

- a) Independent observations.
- b) Normality of the two groups.
- c) Homogeneity of Variance- the two groups are approximately normally distributed.

For dependent t-tests:

- a) Independent observations. (this is separate from independent groups)
- b) Normality of the difference data.

# Differences: Statistical Assumptions

Note that homogeneity of variance is not an assumption in the dependent t-test. This is because the dependent t-test works on only the set of different scores, rather than the two sets of scores from different groups.

If a design is 'balanced' (equal sample sizes), then it is robust to violations of homogeneous variance and normality.

When sample size is roughly above 30, it is then somewhat robust to violations of normality (because of the central limit theorem).

# Core Concepts: “Robust”

The notion of “robustness” refers to how good our tools are in capturing what they are set to capture.

If say an alpha is set to 5%, but in reality (or through tons of simulations), the real alpha (type-1 error rate- the rate of false rejections) is 10%, it is called liberal.

The opposite isn't particularly good either- if the real error rate is 2% instead of the set 5%, it is called conservative- and is still an issue since our tools aren't doing what they're supposed to.

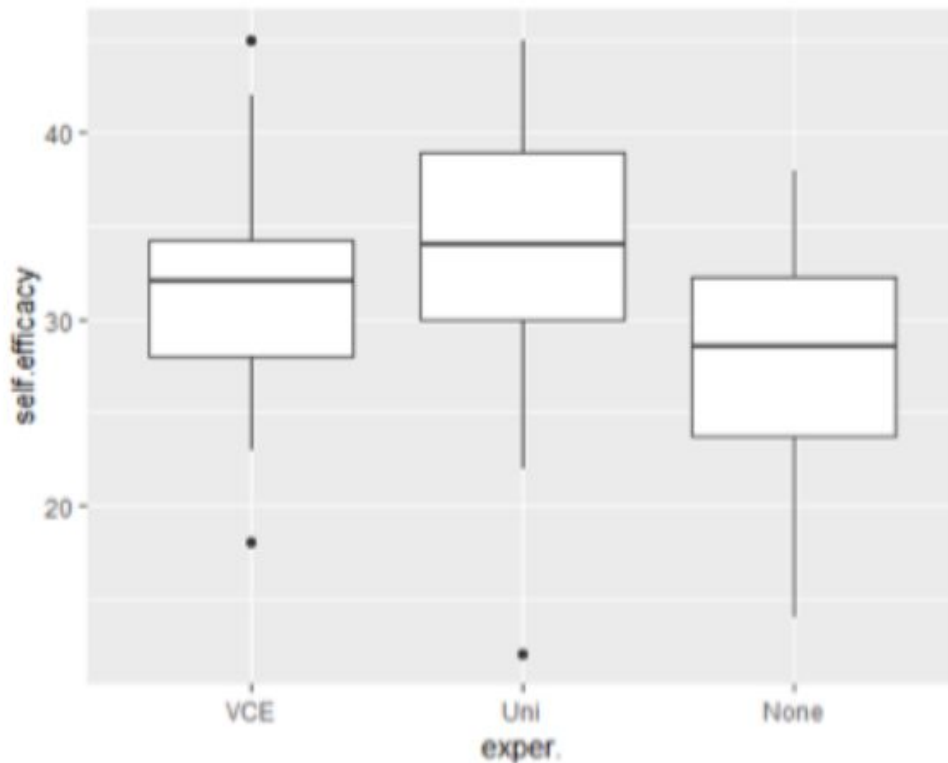
# Research questions addressing differences between THREE or more GROUPS

- examine differences
- only independent groups

# RQs for differences between three groups

Median  
Whiskers  
Outliers  
Symmetric distribution

The differences of statistical self-efficacy between 3 groups: VCE, Uni, and others.



# FACTOR VARIABLES IN R

## **Purpose:**

- to deal easily with categorical variables
- “Levels” of the factor do not operate as numbers

## **Factor variables**

- a variable defined as a factor type in has 2 or more levels
- each level corresponds to one particular category
- each level has an associated label attached it and shown in R

# DUMMY VARIABLE CODING

- Purpose: Dummy variables are used as devices to sort data into mutually exclusive categories (indicate the mutual co-occurrence of different categories)
- **Dummy coding** transforms a categorical variable with  $g$  categories into a meaningful set of  $g - 1$  **dummy variables** that each have values of either 0 or 1

## 2 CATEGORIES

##	ARMP
## RMHI	0
## ARMP	1

## 3 CATEGORIES

##		21-25	26-30	Over 30
## Under 20		0	0	0
## 21-25		1	0	0
## 26-30		0	1	0
## Over 30		0	0	1



# Afterword

We've gone through a **LOT** in a short amount of time.

Don't let anyone tell you otherwise, these concepts take time and practice to understand.

Thank you everyone for listening and give yourselves a pat on the back!

# Afterword

I'd also like to have you all think about these tools a bit in your own time.

Research tools are limited to the extent of the researcher's goals and ability to communicate them.

Believe me or not, I feel that understanding these tools are the easy part-communicating them and describing what they mean in practical terms take even more practice! And I'd argue that proper communication and understanding of these tools are immensely important in any piece of research.

# We'd very much appreciate your feedback!

We'd really appreciate it if you could fill out this google form for feedback regarding the session we've just held.

<https://bit.ly/2IHVd4Z>