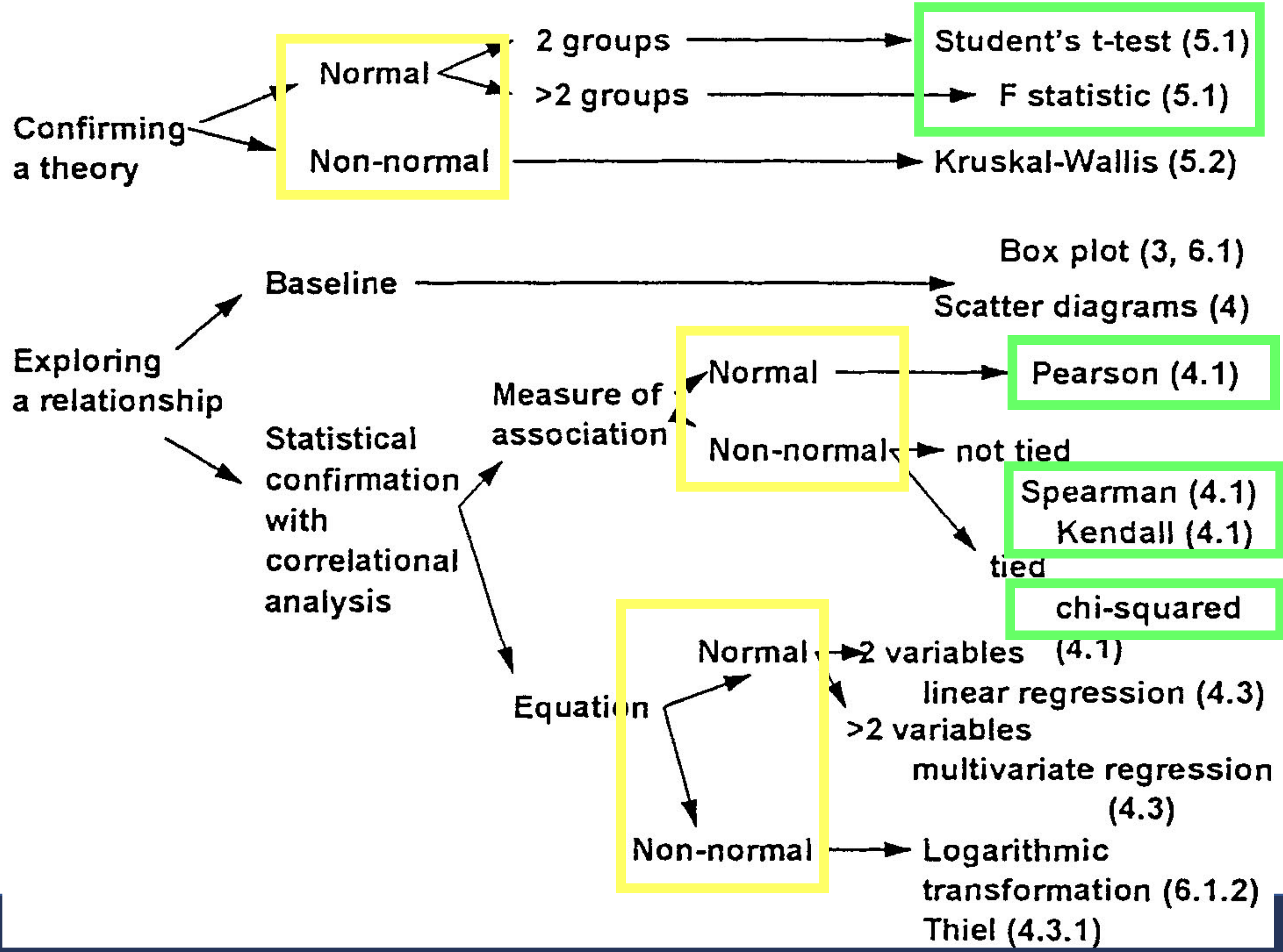
ECE1724H S2: Empirical Software Engineering

Statistical Analysis 2

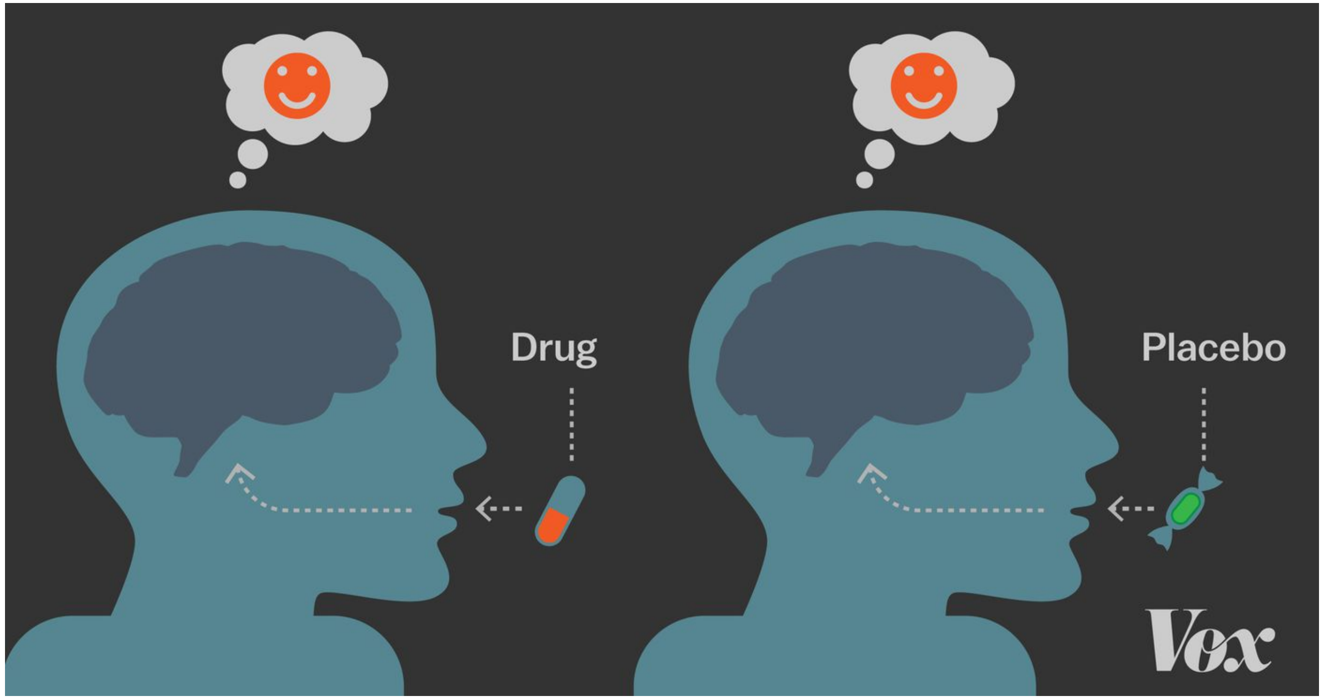
The Edward S. Rogers Sr. Department
of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

Which Statistical Test?



Student’s t test

- For testing whether two samples really are different
  - given: two experimental treatments, one dependent variable
  - Assumes:
    - the variables are normally distributed in each treatment
    - the variances for the treatments are similar
    - the sample sizes for the treatments do not differ hugely
  - Basis: difference between the means of samples from two normal distributions is itself normally distributed.



Student's t test

- Procedure:
  - H_0: "There is no difference in the population means from which the samples are drawn"
  - P-Value: Choose a significance level (e.g. 0.05)
  - T score: a ratio between the difference between two groups and the difference within the groups. The larger the t score, the more difference there is between groups.

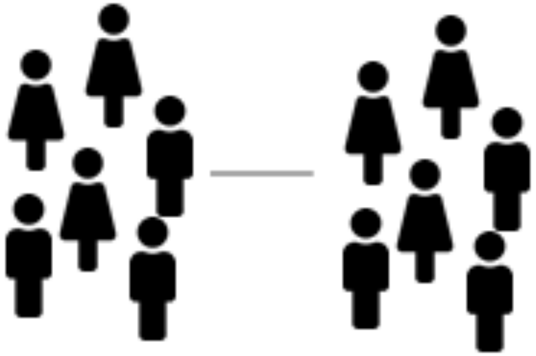
  - Calculate t as $t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{(SE_A)^2 + (SE_B)^2}}$ where $SE = \frac{SD}{\sqrt{N}}$

Student’s t test

One sample t-test

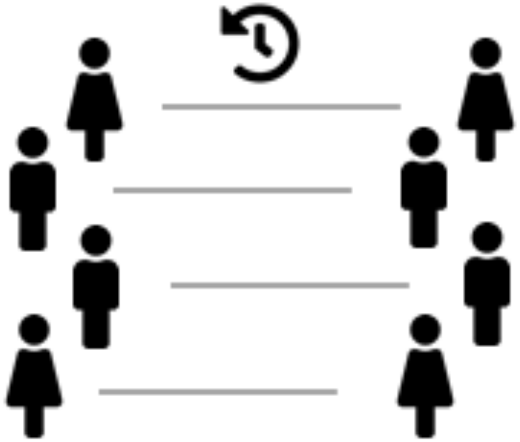
Is there a difference between a group and the population

Unpaired t-test

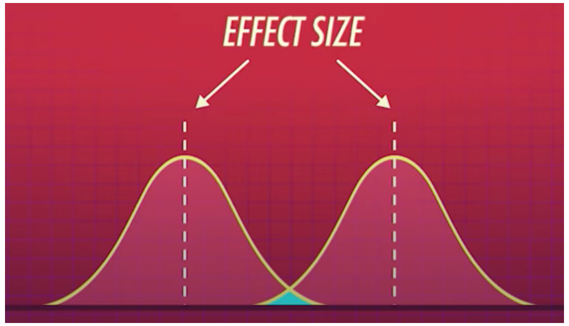


Is there a difference between two groups

Paired t-test



Is there a difference in a group between two points in time



- **Cohen's d** (Cohen 1998).

Cohen's $d = (M_2 - M_1) / SD_{pooled}$

$SD_{pooled} = \sqrt{((SD_1^2 + SD_2^2)/2)}$

Relative size	Effect size	% of control group below the mean of experimental group
	0.0	50%
Small	0.2	58%
Medium	0.5	69%
Large	0.8	79%
	1.4	92%

Group 1

Mean (*M*):

Standard deviation (*s*):

Sample size (*n*):

Group 2

Mean (*M*):

Standard deviation (*s*):

Sample size (*n*):

Calculate

Analysis of Variance (ANOVA)

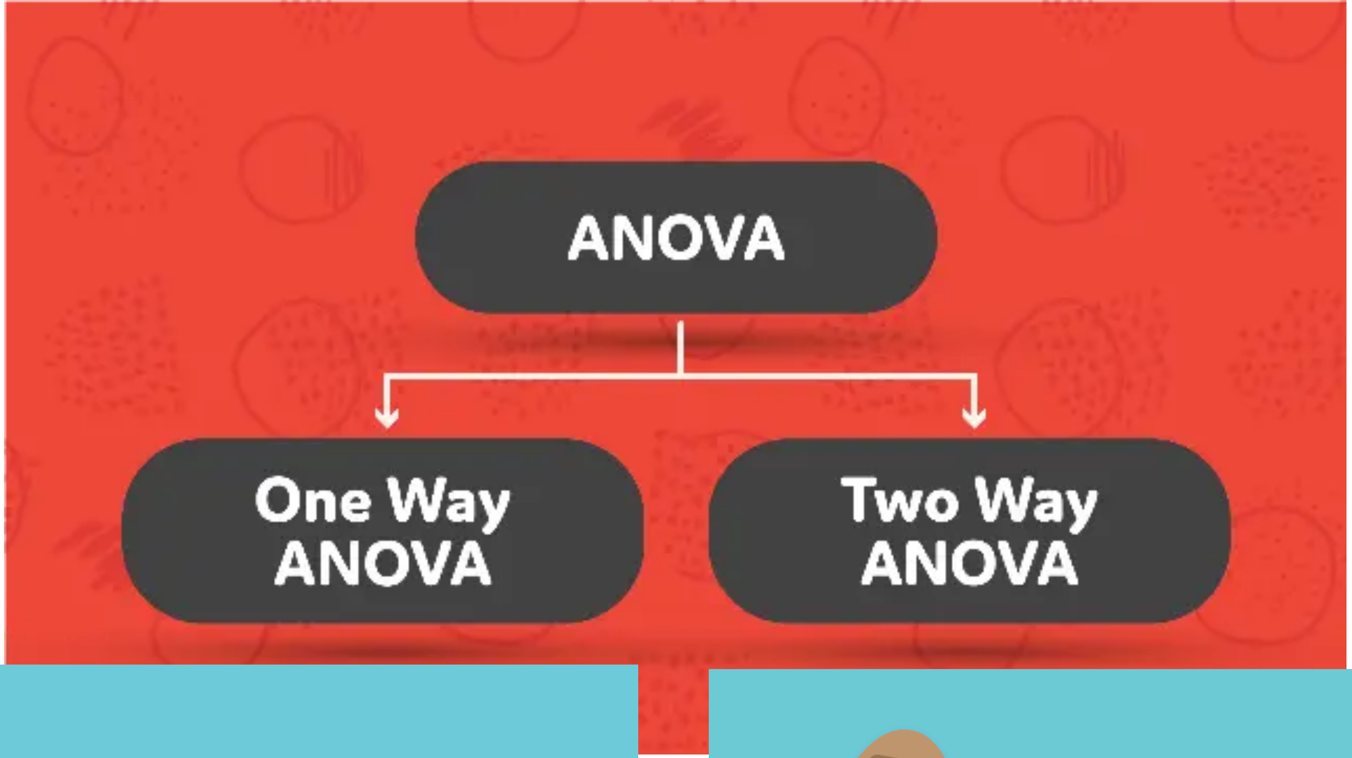
- or F-test
- Generalization of t-test for >2 treatments
  - given: n experimental treatments, one dependent variable
  - Assumes:
    - the variables are normally distributed in each treatment
    - the variances for the treatments are similar
    - the sample sizes for the treatments do not differ hugely
  - Works by analyzing how much of the total variance is due to differences <u>within</u> groups, and how much is due to differences <u>across</u> groups.

Sir
Ronald Fisher
FRS

Fisher in 1913

Analysis of Variance (ANOVA) / F-Statistics

- F statistics (F critical value)
- F value
- F statistic **must be used in combination with the p value**


ANOVA
One Way ANOVA
Two Way ANOVA


DECEMBER
MARCH


DECEMBER
MARCH

Analysis of Variance (ANOVA)

EXAMPLE

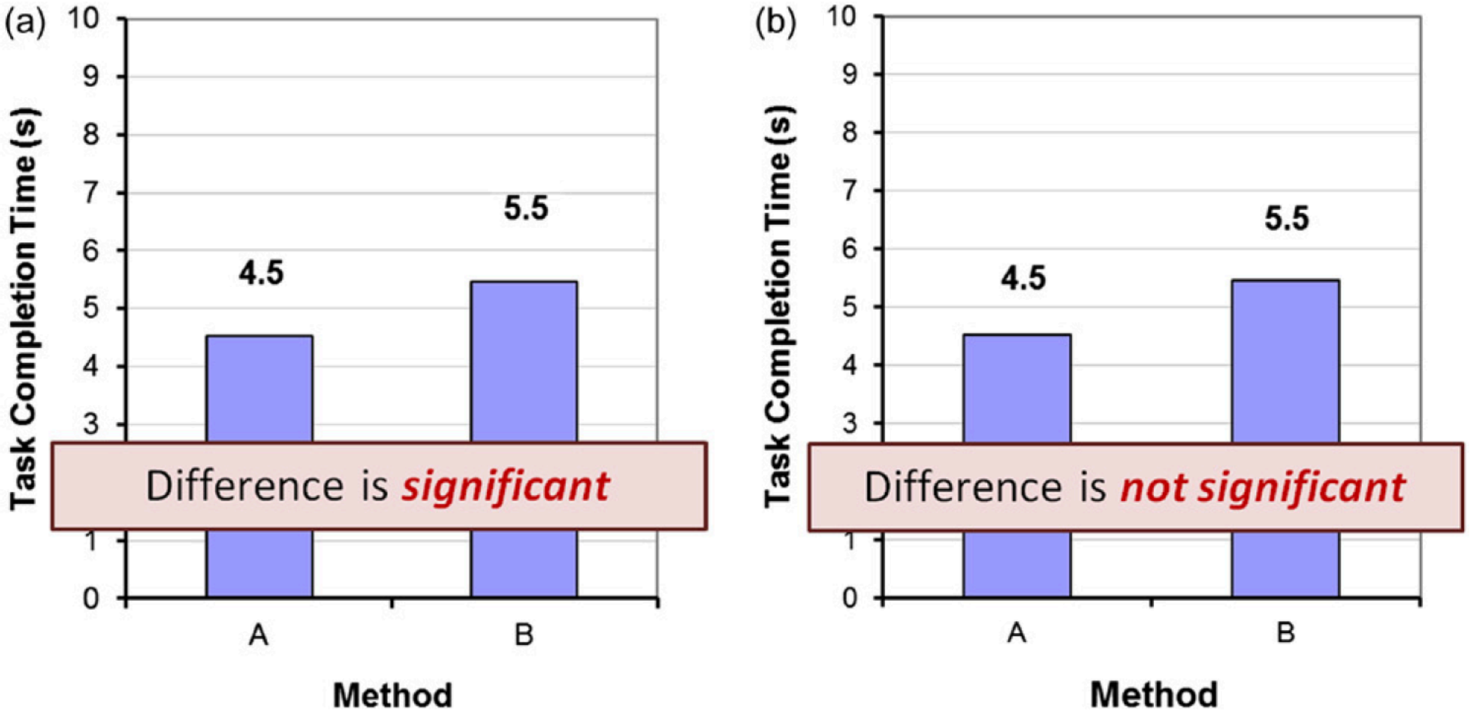


FIGURE 6.2

Difference in task completion time (in seconds) across two test conditions, Method A and Method B. Two hypothetical outcomes are shown: (a) The difference is statistically significant. (b) The difference is not statistically significant.

variance could explain

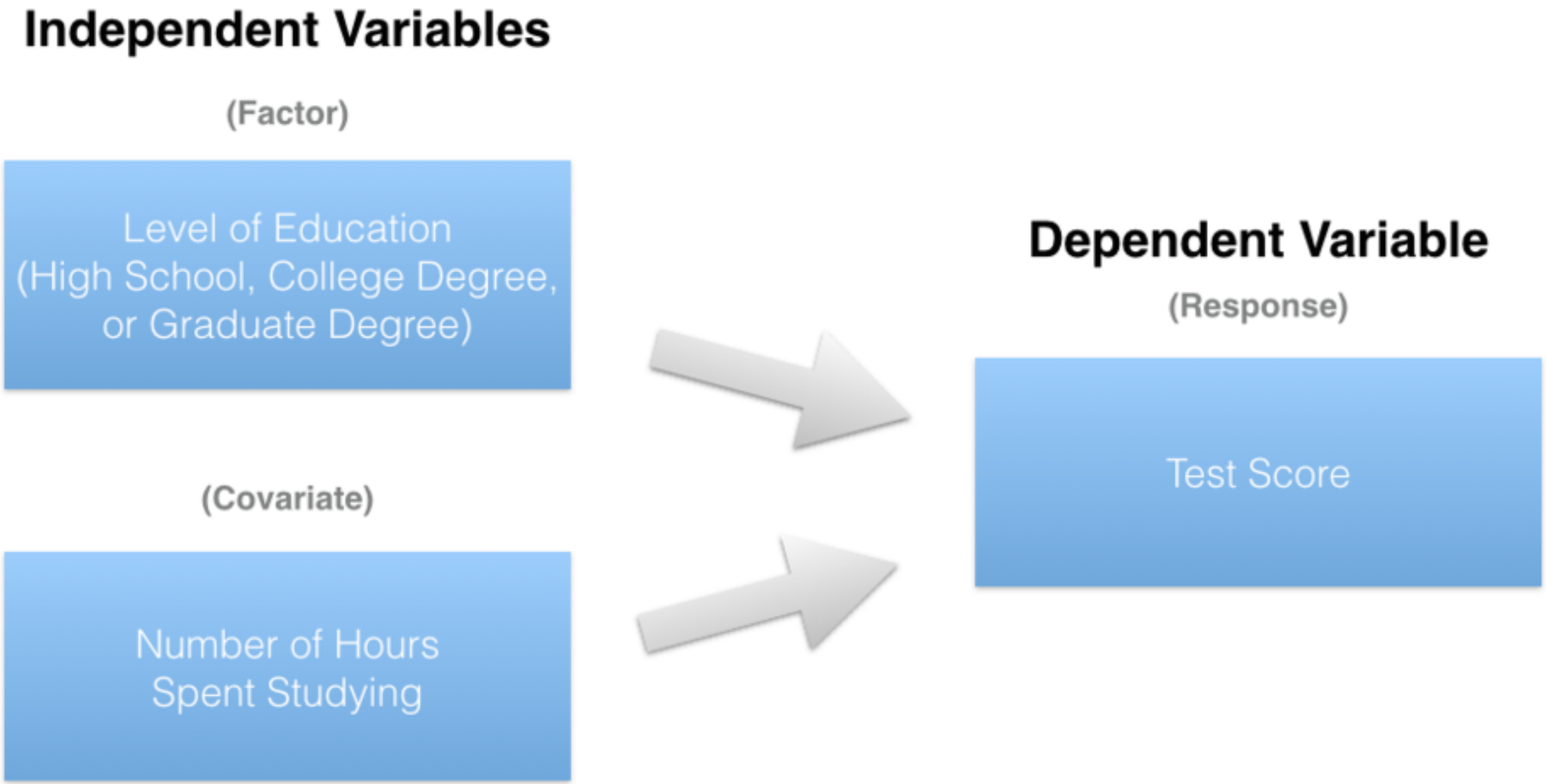
(Ch 6) Hypothesis Testing. from MacKenzie. Human-Computer Interaction. Elsevier 2013

ANOVA Table

Source of variation	Sum of squares (SS)	Degrees of freedom(DF)	Mean Square (MS)	F-statistic
Treatments	$SS_{between}$ (SS_b)	k-1	$MS_b = SS_b/(k-1)$	$F=MS_b/MS_w$
Error (or Residual)	SS_{Within} (SS_W)	N-k	$MS_W = SS_W/(N-k)$	
Total	SS_{Total} (SS_T)	N-1		

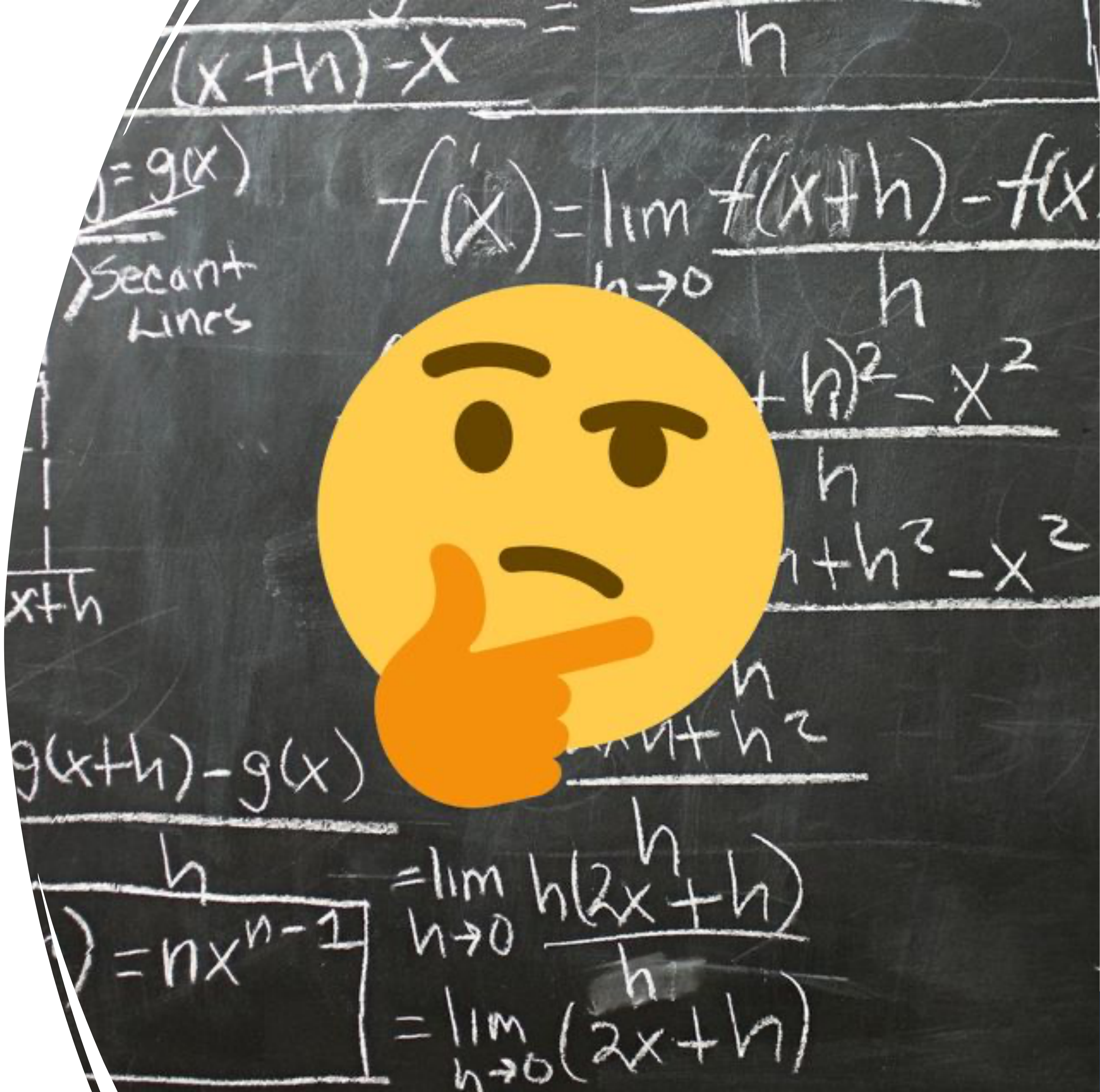
https://sixsigmastudyguide.com/anova-analysis-of-variation/

ANCOVA EXAMPLE

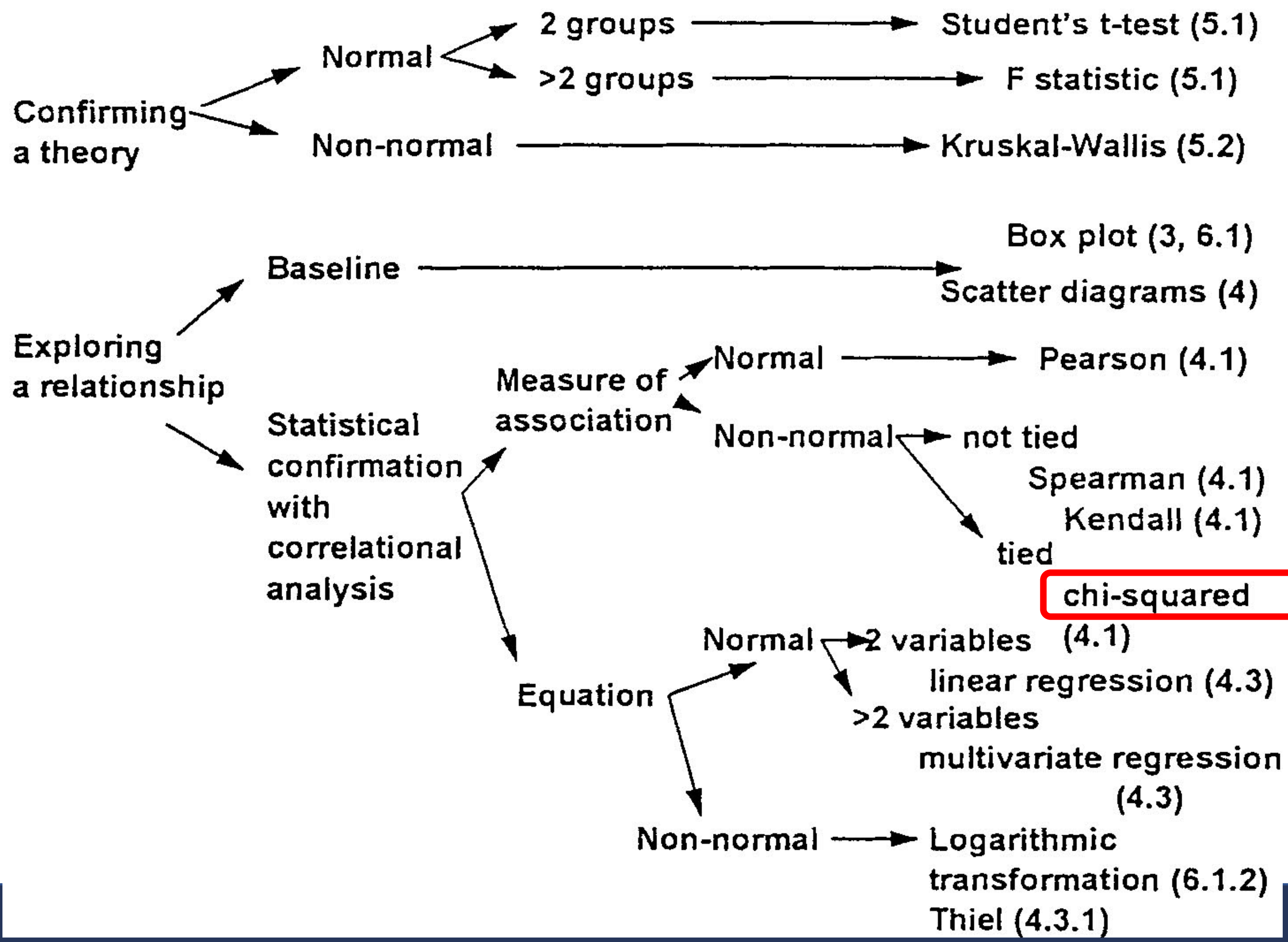


Question:

Correlation between categorical data?



Which Statistical Test?

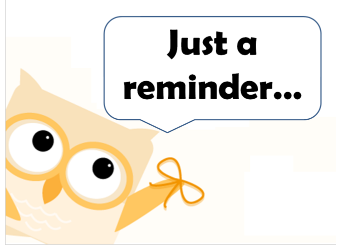


Chi-squared test of Independence

- The **chi-square test** is a hypothesis test used for categorical variables with **nominal or ordinal** scales of measurement.
- The chi-square test checks **whether the frequencies occurring in the sample differ significantly from the frequencies one would expect**. Thus, the observed frequencies are compared with the expected frequencies and their deviations are examined.

- **Null**: Variable A and Variable B are independent
- **Alternate**: Variable A and Variable B are not independent.

How will you measure things?



Type	Meaning	Example	Admissible Operations
Nominal Scale	Unordered categories	Gender, Political preferences, Place of residence	=
Ordinal Scale	Ranking of objects into ordered categories (intervals between the values are not necessarily of the same size)	Satisfaction, Happiness, grades	=, <, >
Interval Scale	Differences between points on the scale are meaningful (equal intervals)	Celsius, Fahrenheit Temperature, IQ (intelligence scale), SAT scores	=, <, >, difference, mean
Ratio Scale	Ratios between points on the scale are meaningful (ordered, equal intervals with a zero point)	weight, height, sales figures, ruler measurements, number of children	=, <, >, difference, mean, ratio

Chi-squared test

- Procedure:
  - Calculate an expected value (mean) for each column
  - Calculate χ^2:

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

    - Where O_i is an observed frequency
    - E_i is the expected frequency asserted by the null hypothesis
  - Compare with lookup value for a given significance level and degree of freedom.

EXAMPLE

P Value from Chi-Square Calculator

This calculator is designed to generate a *p*-value from a chi-squ
score from raw data, you should use our chi-square calculator
for you).

The calculator below should be self-explanatory, but just in cas
chi-square score box, you stick your degrees of freedom in the
square test for independence), select your significance level, th

Report a Chi-Square Result (APA)

Chi-square score: 58

DF: 1

Significance Level:

- ○ 0.01
- ◉ 0.05
- ○ 0.10

The P-Value is < .00001. The result is significant at p < .05.

$$\frac{(O-E)^2}{E} + \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E}$$

$$\frac{(527-484)^2}{484} + \frac{(72-115)^2}{115} + \frac{(206-249)^2}{249} + \frac{(102-59)^2}{59}$$

①Chi-Squared = **58.4**

②Degree of freedom (rows - 1) * (column – 1) = (2-1) * (2-1) = 1

Observed

	Male	Female	Total	
Do not purchase grocery online	484 527	115 72	599	66%
purchase grocery online	249 206	59 102	308	34%
Total	733	174	907	

Chi-squared test

EXAMPLE

	Female	Male
Without graduation	6	7
College	13	16
Bachelor's degree	16	15
Master's degree	8	11
Total	**43**	**49**

Is there a relationship between gender and the highest level of education?

↓

Chi² - Test

Chi-squared test --- strength of association

- Cramér's V

Cramer's V is a measure of the strength of association between two nominal variables.

It ranges from 0 to 1 where:

- **0** indicates no association between the two variables.
- **1** indicates a strong association between the two variables.

$$\text{Cramer's V} = \sqrt{(X^2/n) / \min(c-1, r-1)}$$

Describing Strength of Association

Characterizations	
>.5	high association
.3 to .5	moderate association
.1 to .3	low association
0 to .1	little if any association

Commonly used statistical tests

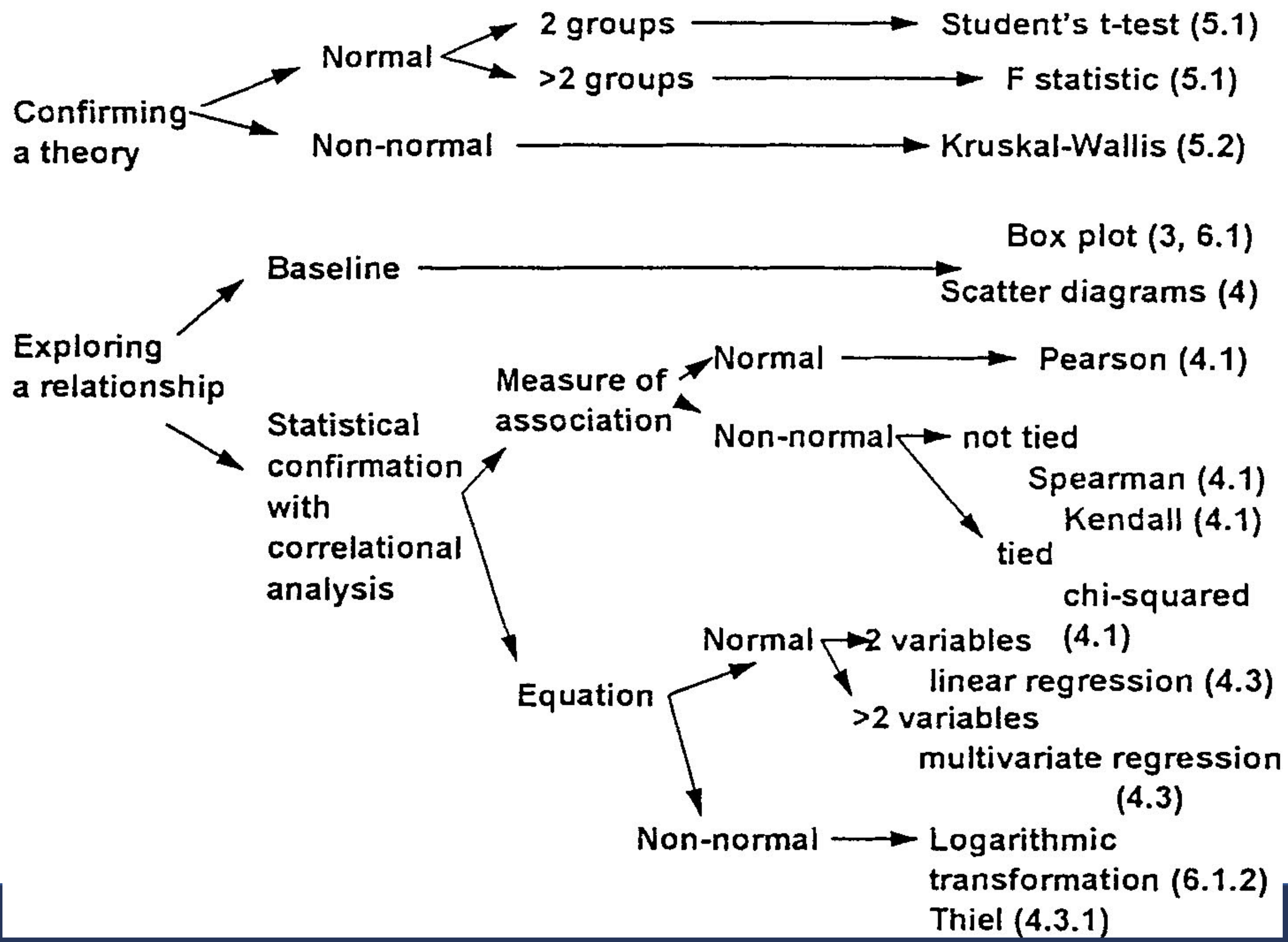
- **Pearson's r:** two continuous variables are correlated or related to each other
- **chi-square:** whether or not there is a relationship between two categorical variables
- **t-test:** whether there is a difference between two groups on a continuous dependent variable
- **ANOVA:** test differences between three or more groups
- **Spearman:** employ calculations based on ranks.

Tests for Ranked Data

English (mark)	Maths (mark)	Rank (English)	Rank (maths)
56	66	9	4
75	70	3	2
45	40	10	10
71	60	4	7
61	65	6.5	5
64	56	5	9
58	59	8	8
80	77	1	1
76	67	2	3
61	63	6.5	6

- Friedman test
- Kruskal-Wallis test
- Rank products
- Spearman's rank correlation coefficient
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test

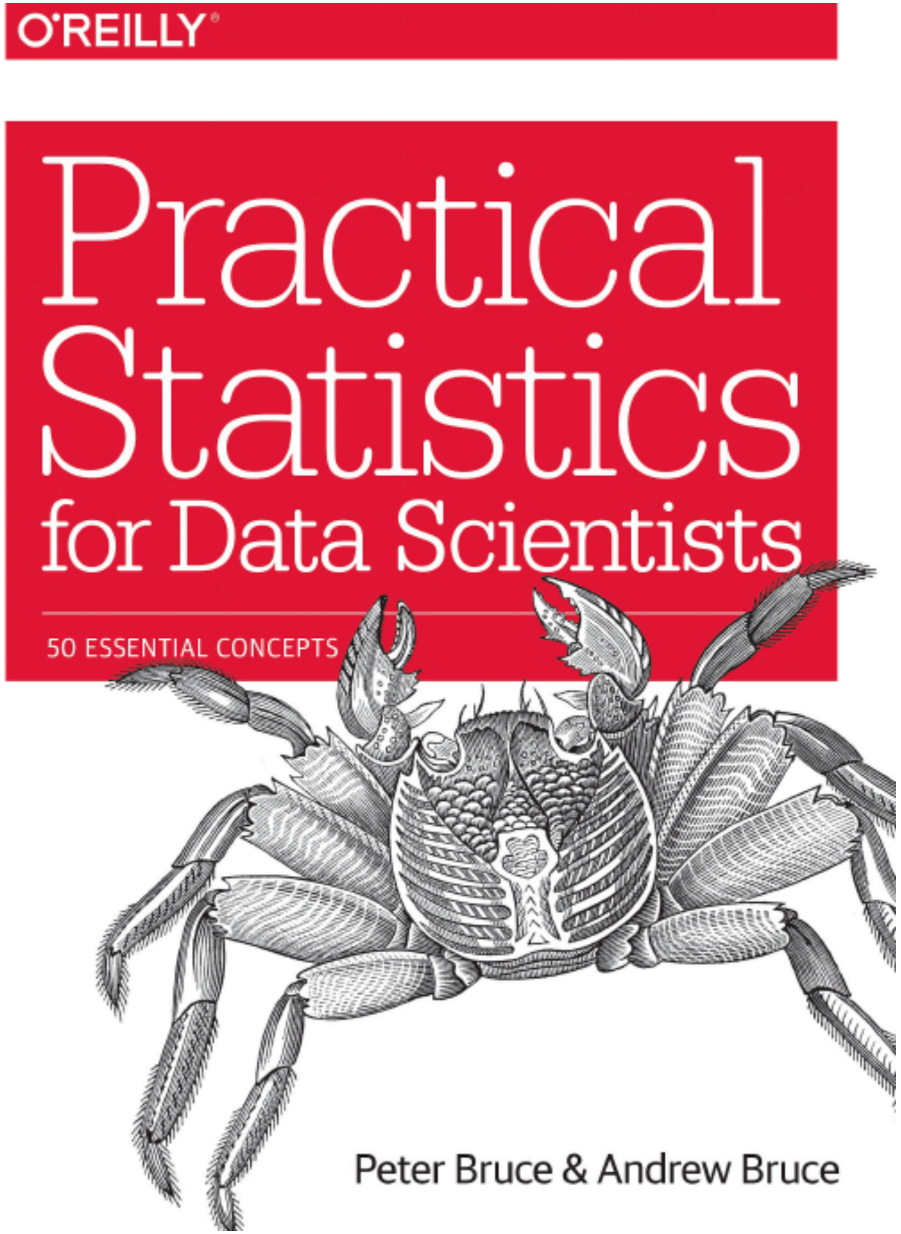
Which Statistical Test?



Which correlation metric should you use?

	Categorical	Continuous
Categorical	Lambda, Corrected Cramer's V	Point Biserial, Logistic Regression
Continuous	Point Biserial, Logistic Regression	Spearman, Kendall, Pearson

https://medium.com/@outside2SDs/an-overview-of-correlation-measures-between-categorical-and-continuous-variables-4c7f85610365


O'REILLY
Practical Statistics for Data Scientists
50 ESSENTIAL CONCEPTS
Peter Bruce & Andrew Bruce