ECE1724H S2: Empirical Software Engineering

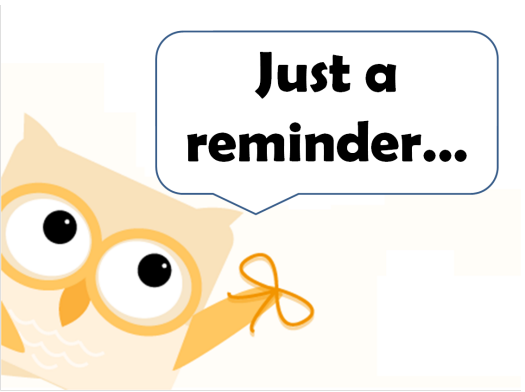
Introduction of Quantitative Study (2)

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

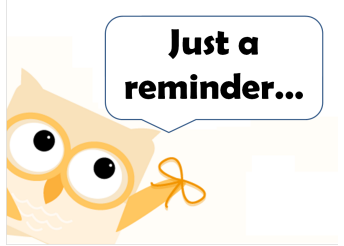
Course Project

Let’s schedule a meeting if you have questions for the comments.


"Measuring programming progress by lines of code is like measuring aircraft building progress by weight."
Microsoft


Just a reminder...

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

Quantitative interpretation

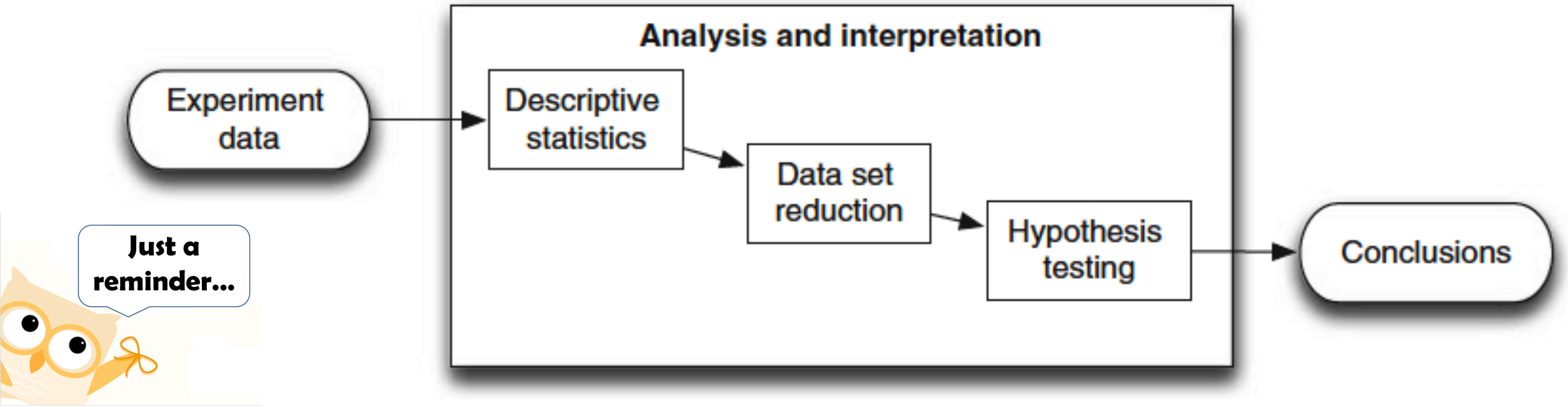


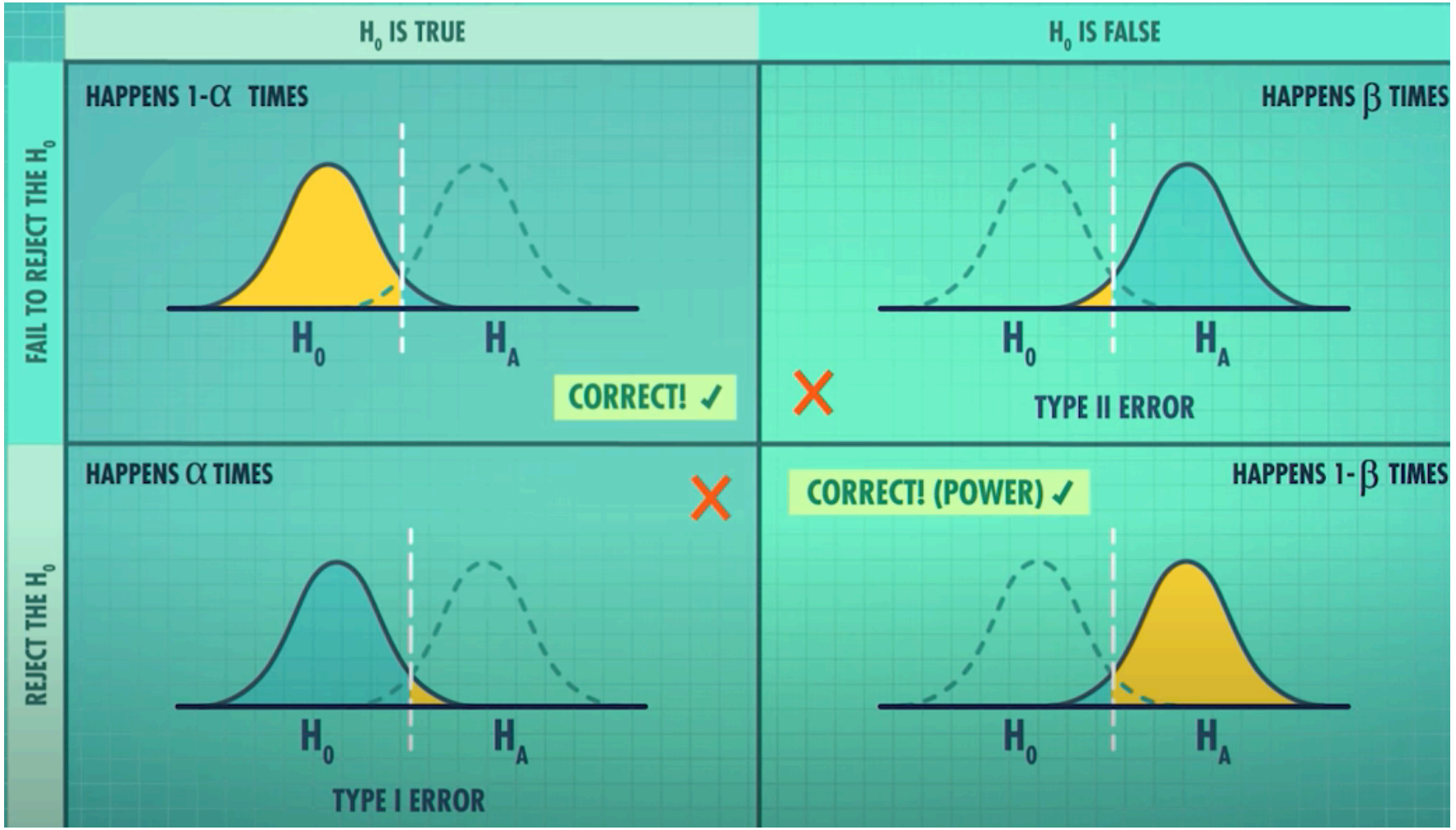
Fig. 10.1 Three steps in quantitative interpretation

C. Wohlin et al., Experimentation in Software Engineering, Springer-Verlag Berlin Heidelberg 2012

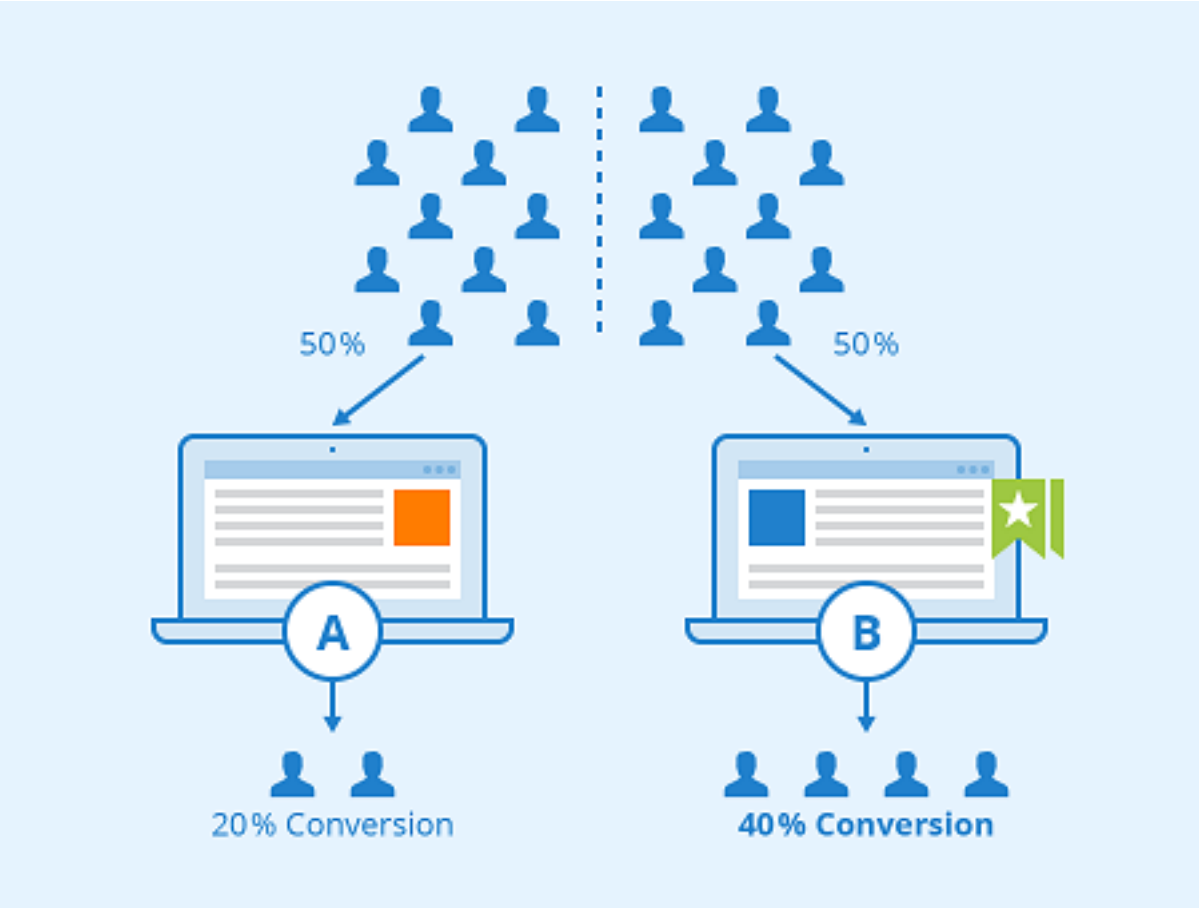
Hypothesis Testing

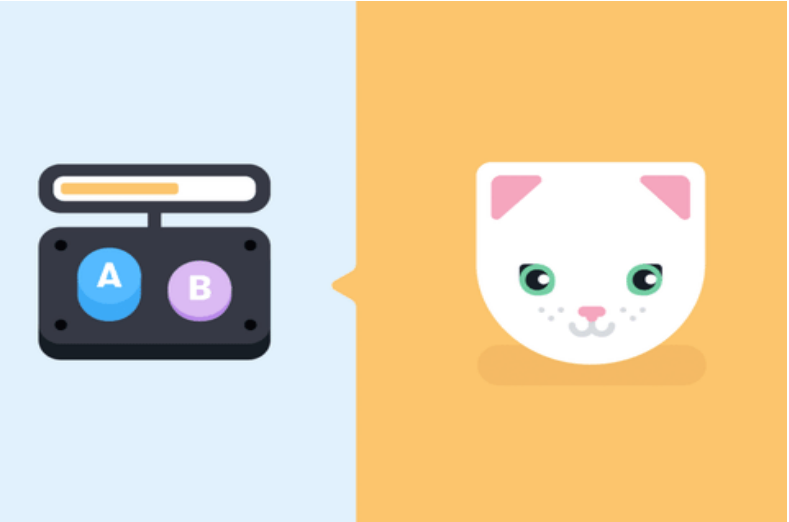
- Set up some hypotheses
  - Null hypothesis (H_0) - asserts that a relationship does not hold
    - In many cases, this is the same as saying there is no difference in the the means of two different treatment groups
  - Alternative hypotheses (H_1, …) - each asserts a specific relationship
  - Type I error: A false positive (rejecting H_0 when it's true)
  - Type II error: A false negative (accepting H_0 when it's false)
- For the statistical tests
  - P value (we calculate this) - probability that a relationship observed in the sample happened by chance
  - Alpha level (selected a priori) - a threshold for p at which we will accept that a relationship did not happen by chance (typically 0.01 or 0.05)
    - This allows us to fix the probability of a type I error in advance
    - If $p < \alpha$, we say the result was significant

Statistical Power



A/B Testing

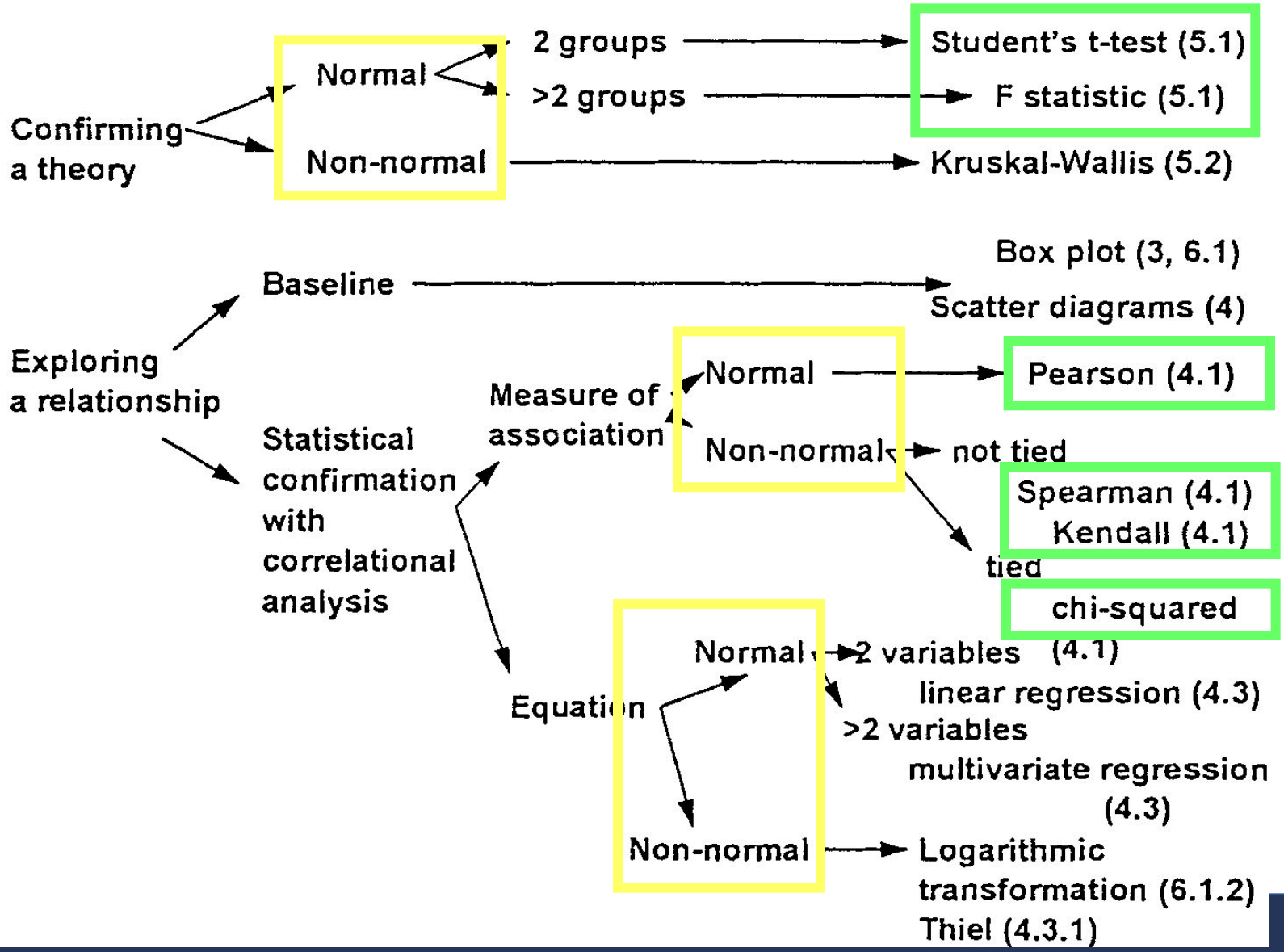




Two flavors of Hypothesis Testing

- Parametric tests -- operate on data from a probability distribution, such as the normal distribution or the t -distribution
- Non-parametric tests: distribution free

Which Statistical Test?



Agenda for today

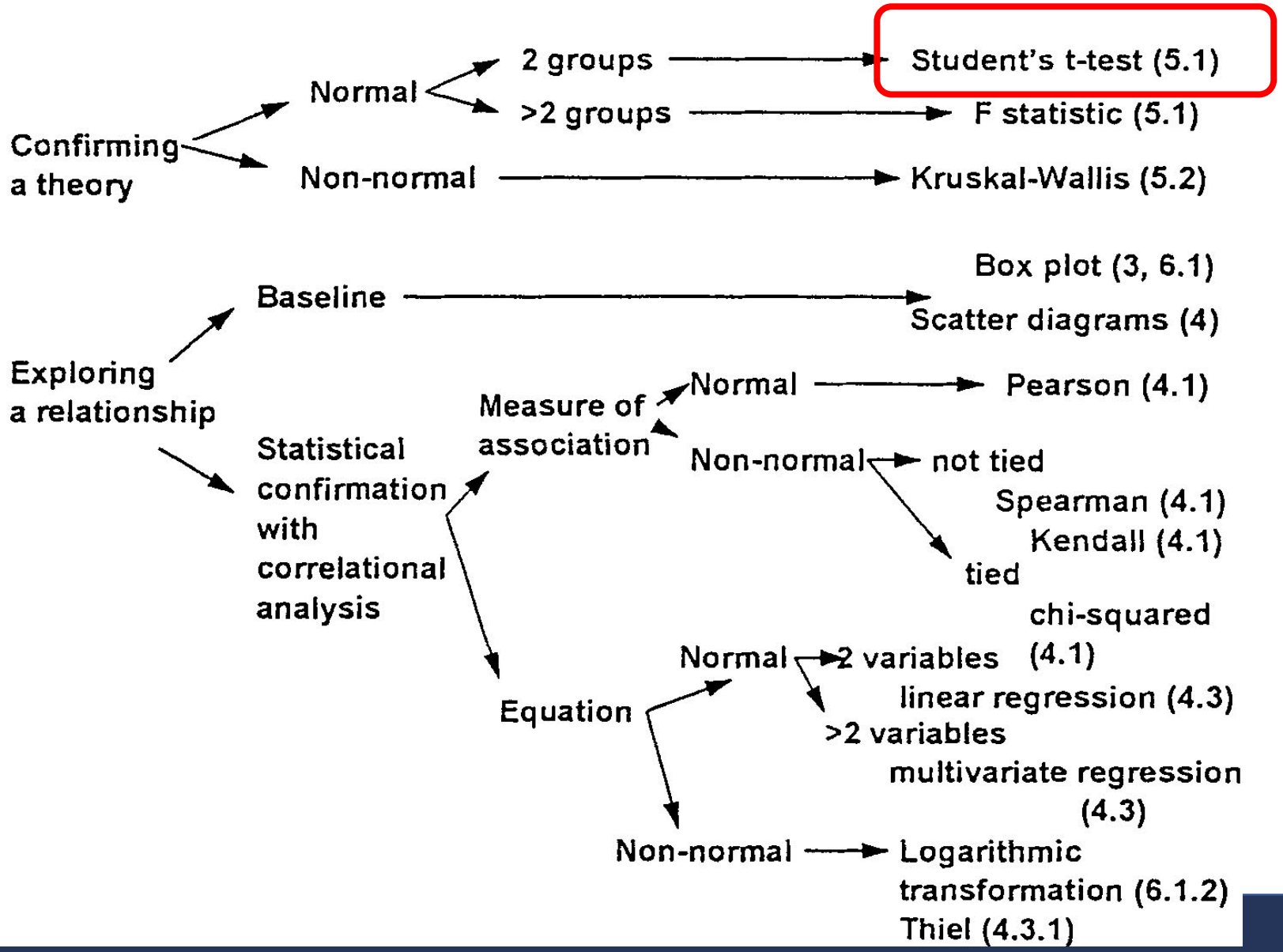
- Paper reading presentation
- Statistical Test
  - Student T-test
  - ANOVA (F-test)
- Experimentation

- Filippova, A., Trainer, E., & Herbsleb, J. D. (2017). From diversity by numbers to diversity as process: supporting inclusiveness in software development teams with brainstorming. ICSE.

Agenda for today

- Paper reading presentation
- Statistical Test
  - Student T-test
  - ANOVA (F-test)
- Experimentation

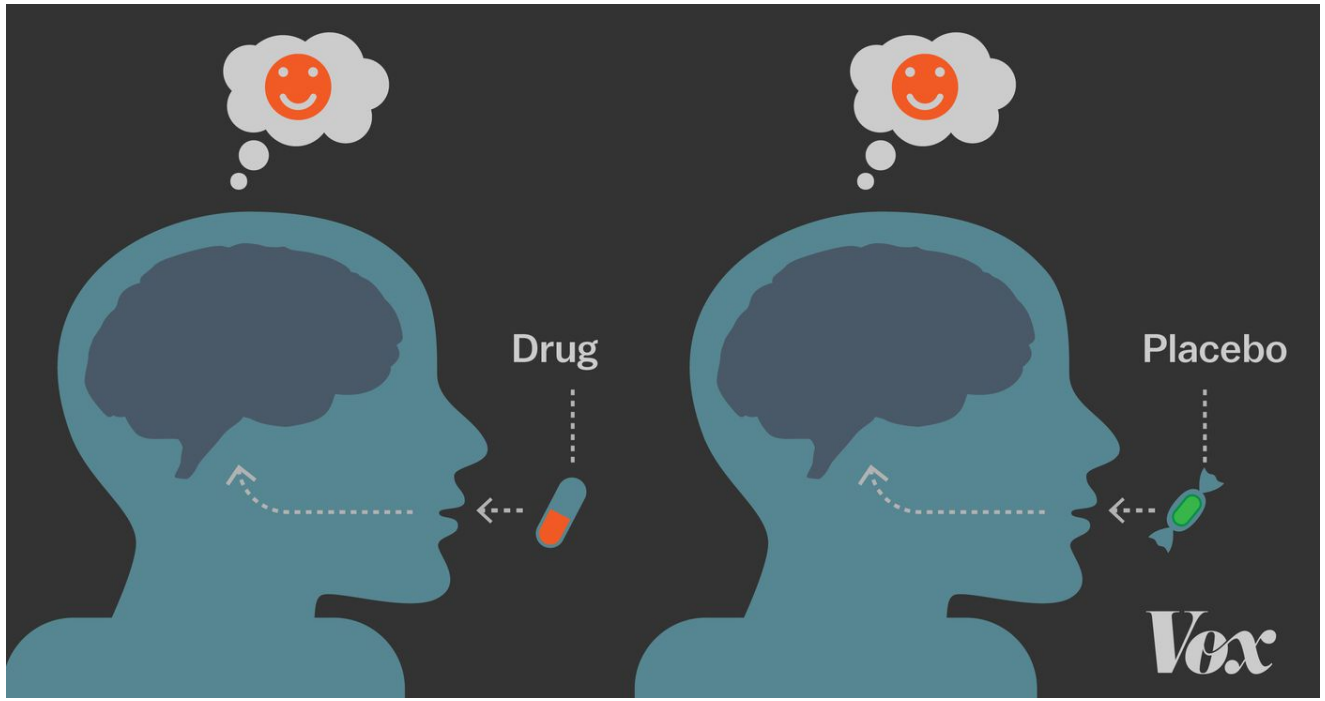
Which Statistical Test?



Student’s t test (T-test) -- History

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

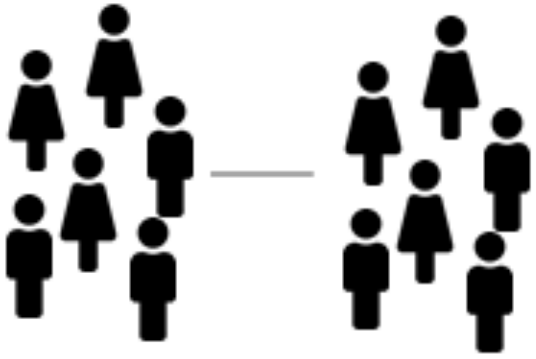
- Procedure:
  - H_0: "There is no difference in the population means from which the samples are drawn"
  - P-Value: Choose a significance level (e.g. 0.05)
  - T score: a ratio between the difference between two groups and the difference within the groups. The larger the t score, the more difference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

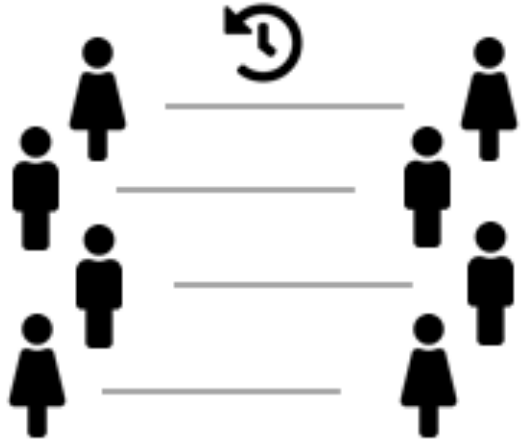
Is there a difference between a group and the population

Unpaired t-test



Is there a difference between two groups

Paired t-test

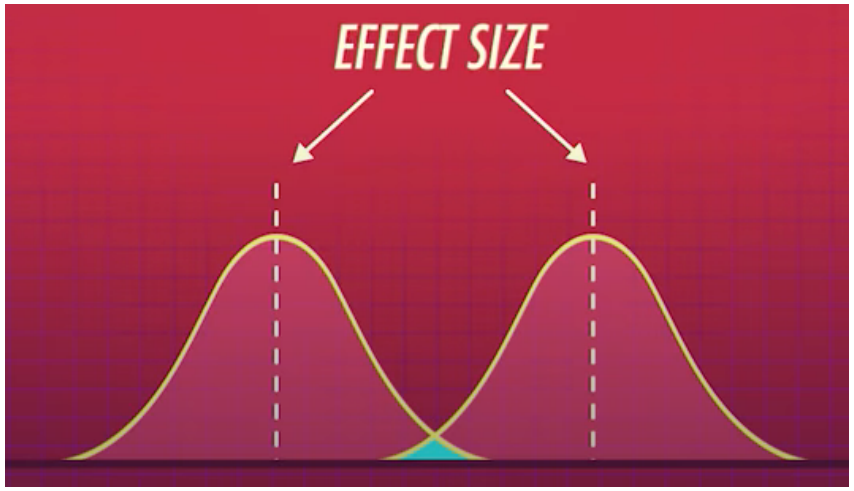


Is there a difference in a group between two points in time

Effect size

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

$$d = \frac{(\text{mean } 1) - (\text{mean } 2)}{\text{std dev}}$$



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%

Effect size

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

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

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

Group 1

Mean (*M*):

Standard deviation (*s*):

Sample size (*n*):

Group 2

Mean (*M*):

Standard deviation (*s*):

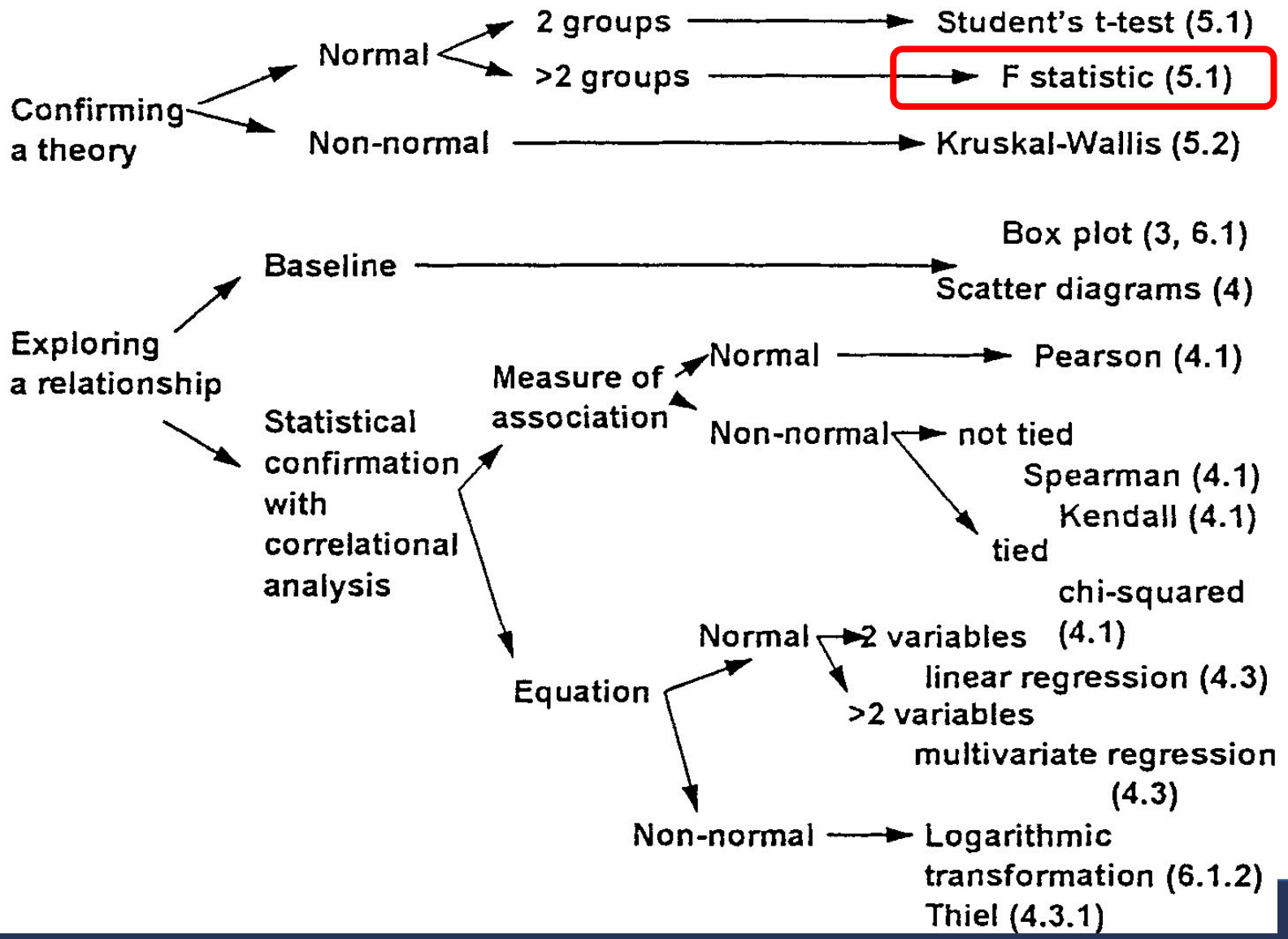
Sample size (*n*):

Calculate

Agenda for today

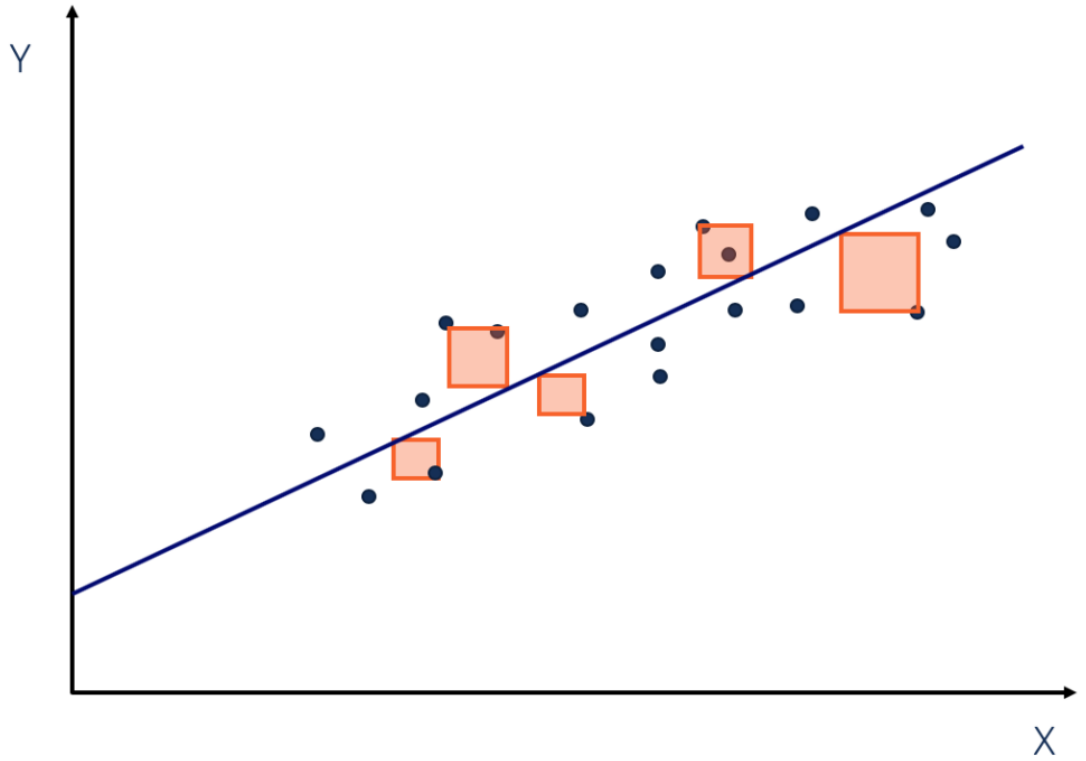
- Paper reading presentation
- Statistical Test
  - Student T-test
  - ANOVA (F-test)
- Experimentation

Which Statistical Test?

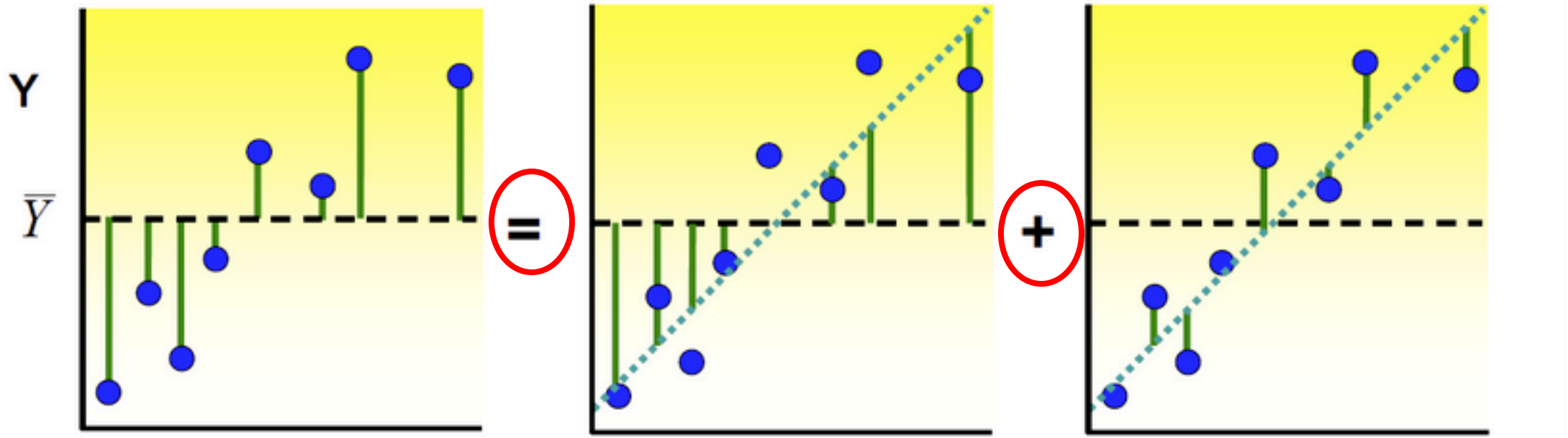


Regression analysis - Sums of Squares (SS)

- **The sum of squares** represents a measure of variation or deviation from the mean.



Regression analysis - Sums of Squares (SS)



Total SS_T

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2$$

Model (Explained) SS_R

$$\sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2$$

Unexplained (Error) SS_R

$$\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

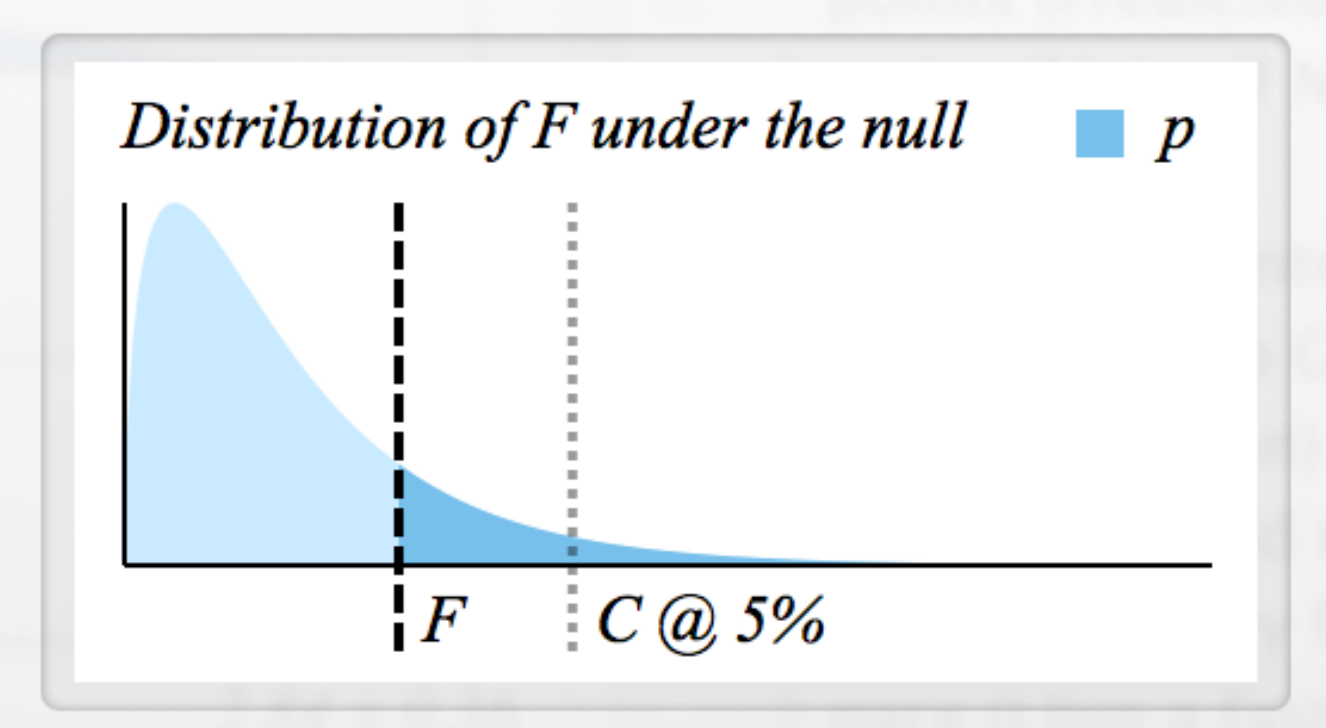
F Statistics

$$F = \frac{SSModel / DF_R}{SSError / DF_E}$$

- **Degree of Freedom (DF):** the number of independent pieces of information we have.

Source of variation	SS	df
Regression	SSR	1
Error	SSE	n-2
Total	SST	n-1

n – sample size



Critical values of F-Statistics

An on-line statistical calculator

This calculator will compute the critical values of *F*-statistics corresponding to n_N (numerator) and n_D (denominator) degrees of freedom, at the desired probability level. See for example Table 3-1 on page 44 of Stanton A. Glanz **"Primer of Biostatistics"**, 3rd Edition, McGraw-Hill, New York, 1992. The numerator and denominator degrees of freedom must be whole numbers correponding to the sample sizes.

Degrees of freedom - numerator:

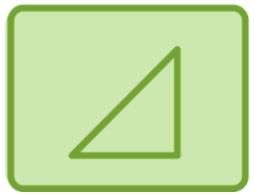
Degrees of freedom - denominator:

Probability level (alpha): 0.05

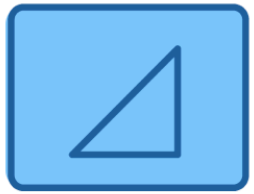
Calculate

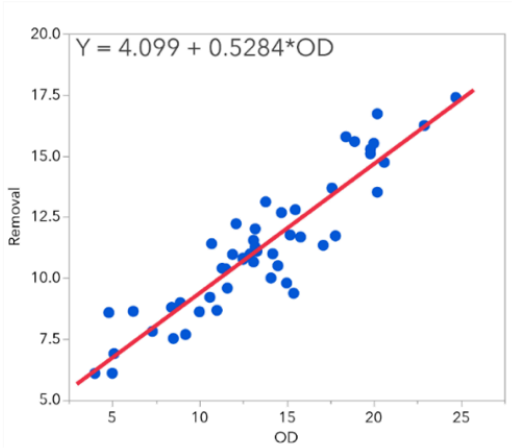
Regression

continuous response



continuous predictor

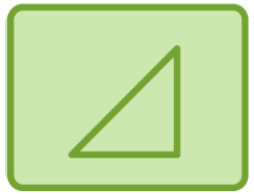




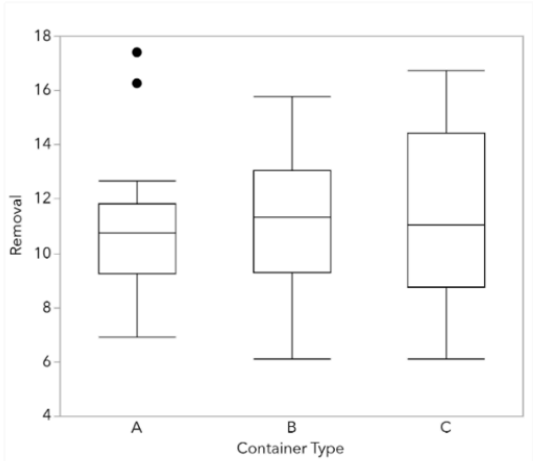
$SSTotal = SSModel + SSError$

ANOVA

continuous response



nominal factor



$SSTotal = SSBetween + SSWithin$

Analysis of Variance (ANOVA) / F-Statistics

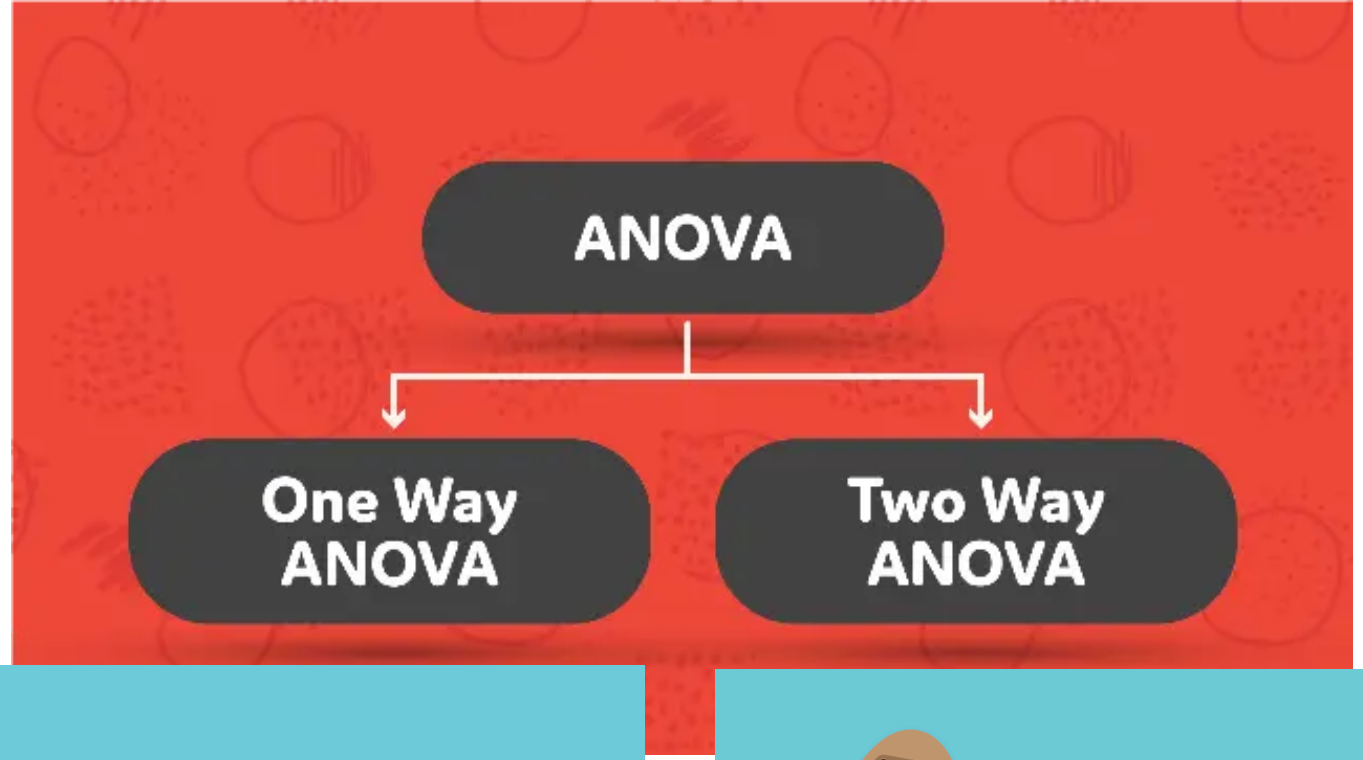
$$F = \frac{SSModel / DF_R}{SSError / DF_E}$$

$$F = \frac{SSBetween / DF_B}{SSWithin / DF_W}$$

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/


ANOVA
One Way ANOVA
Two Way ANOVA


DECEMBER
MARCH


DECEMBER
MARCH

One way ANOVA

- H0: there is no difference between the groups and equality between means. (Walruses weigh the same in different months)
- H1: there is a difference between the means and groups. (Walruses have different weights in different months)

Analysis of Variance (ANOVA)

Does it take less time to complete a task using Method A rather than Method B?

H0: There is no difference in the mean time to complete a task using Method A vs. Method B.

One factor (method), two treatment (A & B)

Dependent variable: task completion time

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

Analysis of Variance (ANOVA)

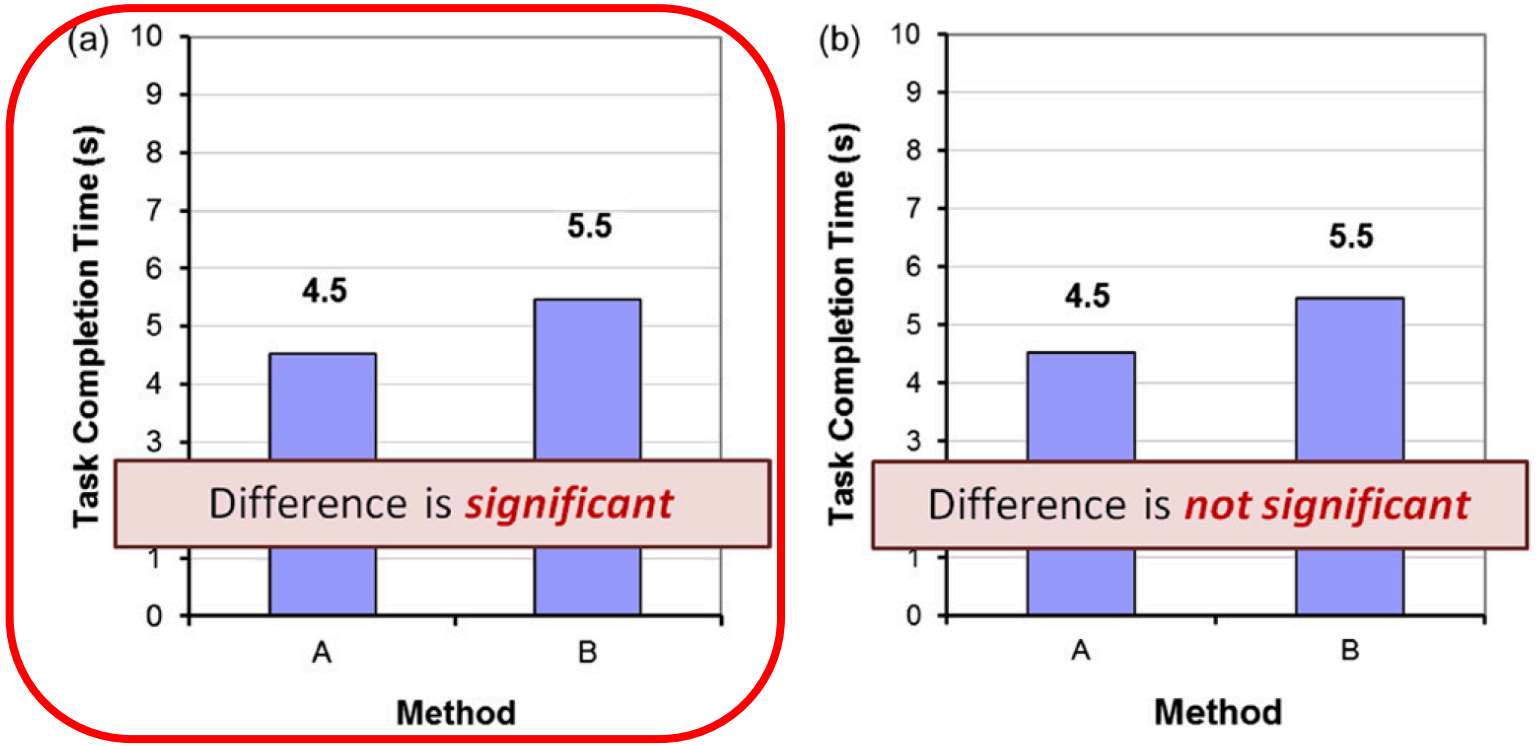


FIGURE 6.2

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.

EXAMPLE

variance could explain

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

Analysis of Variance (ANOVA)

(a)

Participant	Method	
	A	B
1	5.3	5.7
2	3.6	4.8
3	5.2	5.1
4	3.6	4.5
5	4.6	6.0
6	4.1	6.8
7	4.0	6.0
8	4.8	4.6
9	5.2	5.5
10	5.1	5.6
Mean	4.5	5.5
SD	0.68	0.72

(b)

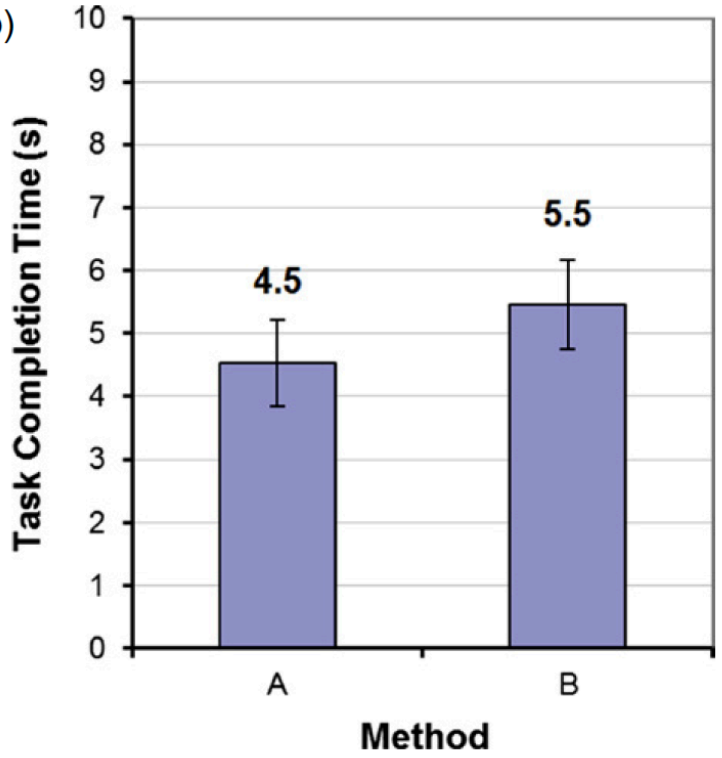


FIGURE 6.3

(a) Data for simulation in Figure 6.2a. (b) Bar chart with error bars showing ±1 standard deviation.

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

Analysis of Variance (ANOVA)

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				

FIGURE 6.4

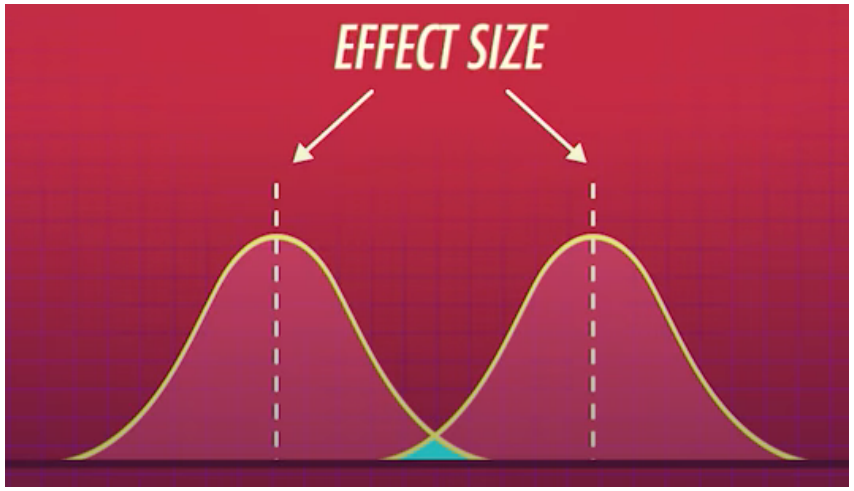
Analysis of variance table for data in Figure 6.3a.

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

Effect size

η^2

- Small: 0.01
- Medium: 0.059
- Large: 0.138



Effect size

(1) for a between groups

(2) for a within subjects

Analysis of Variance (ANOVA)

effect size

The mean task completion time for Method A was 4.5 s. This was 20.1% less than the mean of 5.5 s observed for Method B. The difference was statistically significant ($F_{1,9} = 9.80$, $p < .05$).

FIGURE 6.5

Example of how to report the results of an analysis of variance in a research paper.

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

Two way ANOVA

- H0: The means of all month groups are equal
- H1: The mean of at least one month group is different

- H0: The means of the gender groups are equal
- H1: The means of the gender groups are different

- H0: There is no interaction between the month and gender
- H1: There is interaction between the month and gender

Analysis of Variance (ANOVA)

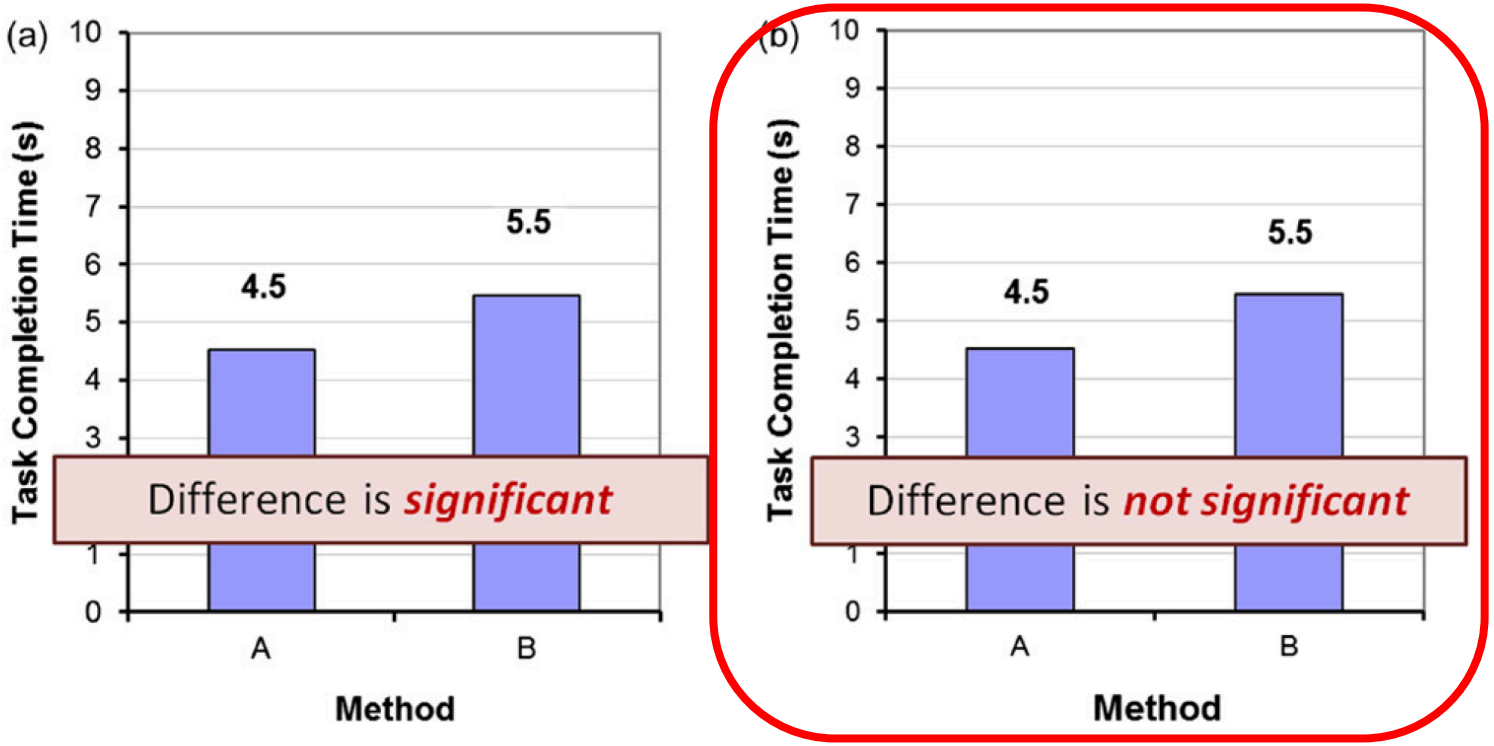
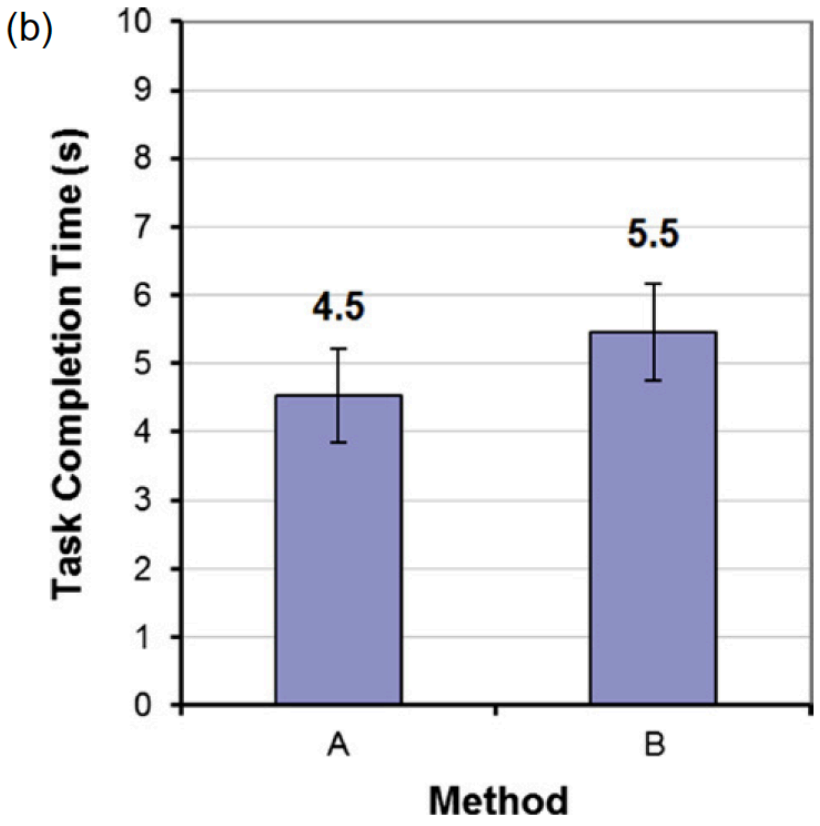


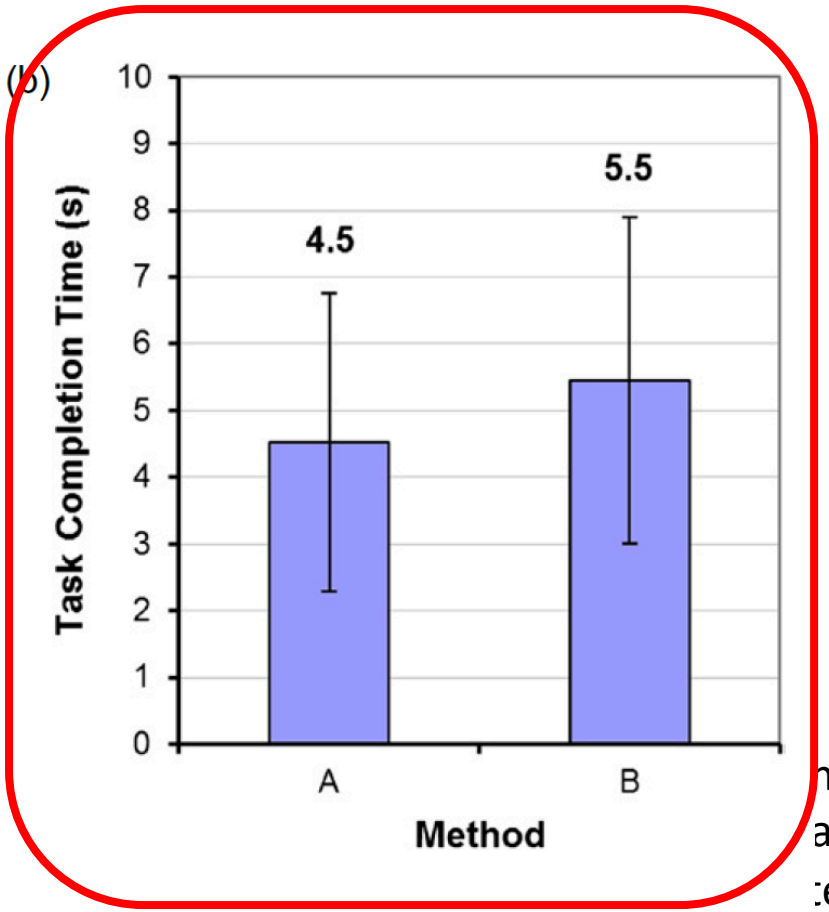
FIGURE 6.2

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.

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

Analysis of Variance (ANOVA)





n 6) Hypothesis Testing. from
acKenzie. Human-Computer
eraction. Elsevier 2013

Analysis of Variance (ANOVA)

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	37.372	4.152				
Method	1	4.324	4.324	.626	.4491	.626	.107
Method * Subject	9	62.140	6.904				

The mean task completion times were 4.5 s for Method A and 5.5 s for Method B. As there was substantial variation in the observations across participants, the difference was not statistically significant as revealed in an analysis of variances ($F_{1,9}$ = 0.626, ns).

FIGURE 6.8

Reporting a non-significant ANOVA result.

Analysis of Variance (ANOVA)

$$F = \frac{SSBetween / DF_B}{SSWithin / DF_W}$$

- Procedure:
  - H_0: "There is no difference in the population means across all treatments"
  - Compute the F-statistic:
    - *F*=(found variation of the group averages)/(expected variation of the group averages)
    - (don't do this by hand!)
  - If H_0 is true, we would expect F=1
  - Note: ANOVA tells you whether there is a significant difference, but does not tell you which treatment(s) are different.

References

- (Ch 10) Analysis and Interpretation . from C. Wohlin et al., Experimentation in Software Engineering, Springer-Verlag Berlin Heidelberg 2012
- (Ch 6) Statistical Methods and Measurement . from F. Shull et al. (eds.), Guide to Advanced Empirical Software Engineering. Springer 2008
- (Ch 6) Hypothesis Testing. from MacKenzie. Human-Computer Interaction. Elsevier 2013