Module 3 Agenda: *Hypothesis Testing (long)*

1. What is Hypothesis testing?
    - Overview of the four steps
    - P-values, type I, type II error, Power, Sensitivity, Specificity etc.
    - Binomial distribution
        - What is it
        - How to simulate from scratch, or use built-in functions (dbinom() etc.)
    - Examples:
        - **Binomial test (Bumpus)**
        - **Contingency test (Bumpus)**
        - **Fisher exact test**

2. AUROC

Your pipeline for hypothesis testing in statistics

Step 1

Formulate your null **hypothesis**

- How *unusual* is your data?

Step 2

Identify appropriate **test statistic**

- Assumptions of your test

Step 3

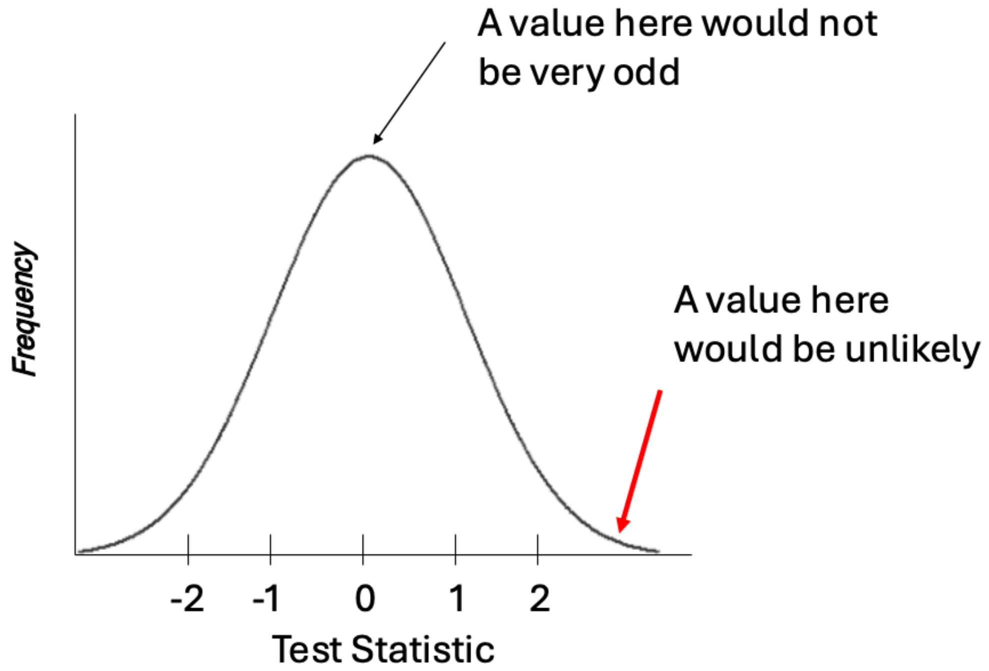
Quantify the results of your test

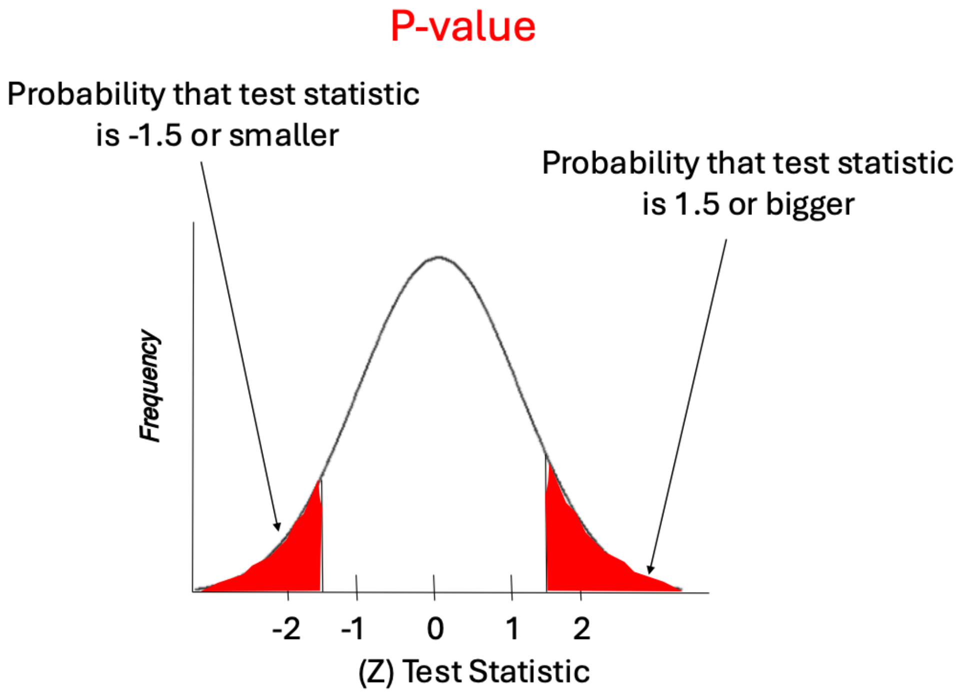
- **P value** or comparison to **critical values**

Step 4

Conclude: reject or fail to reject

- based on alpha value
- if appropriate, confidence interval of the parameter





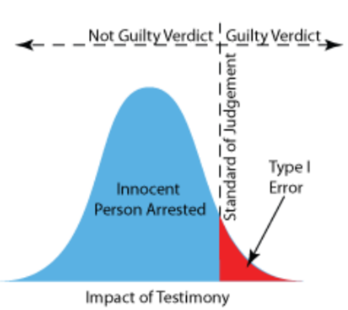
P-Value:

Probability of obtaining data that are equal to or even more extreme than the value assuming the null hypothesis is true

Type I (α) error:

Rejecting a true null hypothesis

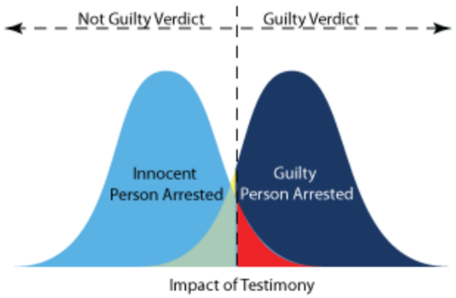
P(reject $H_0|H_0$ = true) = α



Type II (β) error:

Not rejecting a false null hypothesis

P(Fail to reject $H_0|H_0$ is not true) = β



http://www.intuitor.com/statistics/T1T2Errors.html

	No Disease (H_0 true)	Disease (Ho is not true; H_A true)
Fail To Reject H_0	No Error Specificity = P[FTR|Ho is true] True Negative	Type II P[FTR |Ho is not true] (False Negative)
Reject H_0	Type I P[reject|Ho] (False Positive)	No Error Power/Sensitivity P[Reject|Ho is not true] (True Positive)

Sensitivity = $\frac{TP}{TP+FN}$

<u>Types of Errors (Type I, Type II):</u>

Type I = alpha = False Positive = P[rejecting null hypothesis|Ho is actually correct]

Type II - Beta = False Negative = P[NOT rejecting null hypothesis|Ho is not correct]

Power = 1-Type II = True Positive = P[Reject null hypothesis|Ho is not correct]

Sensitivity = True Positive Rate = Power*

<u>wikipedia page that explains Sensitivity, Specificity in more detail</u>

* Power is a type of sensitivity, but not all sensitivity is statistical power.

P-value

The p-value tells you how surprising your data would be if nothing real were going on (i.e., if the null hypothesis were true).

• Small p-value (usually < 0.05): your data are *unlikely* if there's no real effect --> maybe something real is happening.

• Big p-value: your data are *consistent* with chance.

Type I Error (False Positive)

You think you found something real, but you didn't.

- P[rejecting null hypothesis|Ho is actually correct]
- Example: saying a drug works when it actually doesn't.
- Controlled by the **significance level (α)**, often set at 0.05.

Type II Error (False Negative)

You missed something that is real.

- P[NOT rejecting null hypothesis|Ho is not correct]
- Example: saying a drug doesn't work when it actually does

Power

The chance you'll correctly detect a real effect (avoid a Type II error).

• 1-Type II = P[Reject null hypothesis|Ho is not correct]

•Higher power = better chance of spotting true effects.

• Usually aim for **80% or higher**

Sensitivity

If something is real, how often do you correctly catch it?

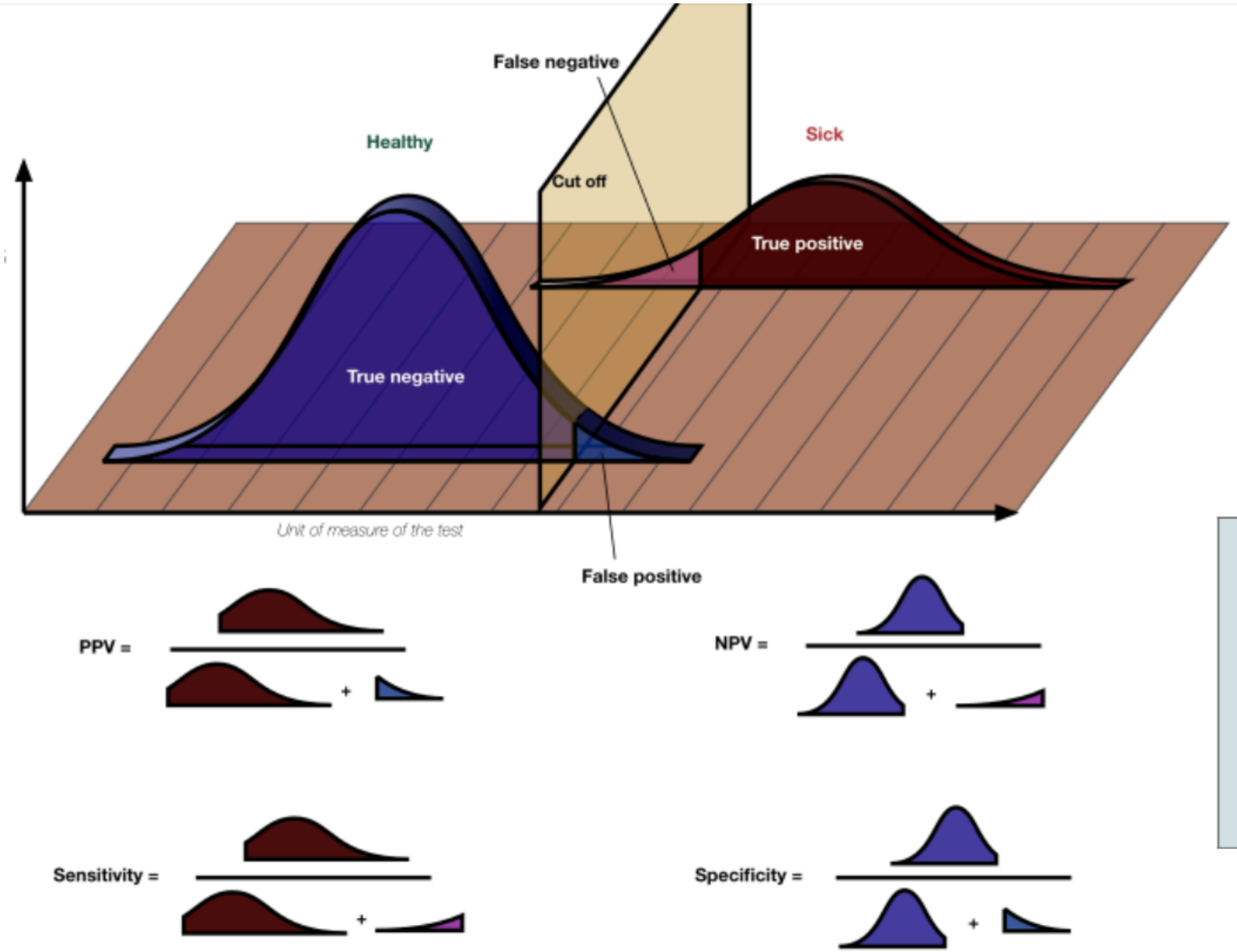
- Example: a COVID test correctly flags people who have the virus.
- "How good the test is at finding the true positives."

Specificity

If something isn't real, how often do you correctly say "no"?

- Example: a COVID test correctly says negative for people who don't have the virus.
- "How good the test is at avoiding false alarms."

categorical variable prediction



		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

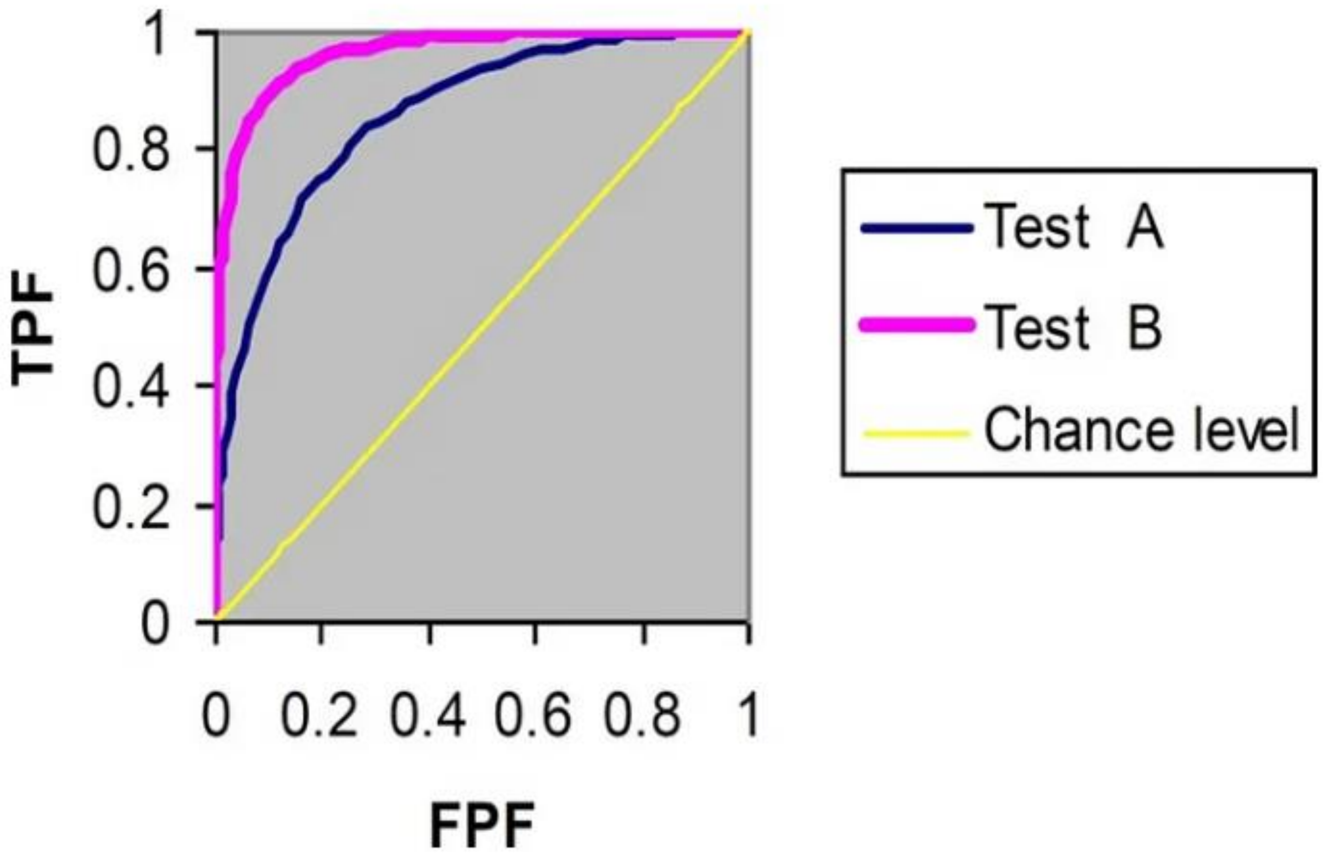
Related ideas:

Accuracy = $\frac{TP + TN}{(TP+TN+FP+FN)}$

Precision = $\frac{TP}{(TP+FP)}$

Recall, Sensitivity, TPR = $\frac{TP}{(TP+FN)}$

FPR = 1- specificity = $\frac{FP}{FP + TN}$



ROC curves of two diagnostic tasks (test A versus test B)(Image source)

https://medium.com/@shaileydash/understanding-the-roc-and-auc-intuitively-31ca96445c02 one sample t-test → two sample t test → ANOVA → (linear) Regression

Correlation

categories → numeric

(1) independent samples

(2) Normal distr^n (A) Visual assessment (boxplot, histograms, qqplot, qqnorm)

(b) Shapiro. Wilk test

BIVARIATE NORMALITY – CONTOUR PLOT

linear – VISUAL

(3) Variances ..... Homoscedascity

– Visual assessment

– Var. test (if normal; n=2)

– Bartlett's test (if normal; n > 2)

– Levene's test (if not normally distr^d)

t.test

t.test

If variances are not equal

Welch's approximate t.test

IF NOT NORMAL

Wilcox.test

aov

IF FTR

Ho: $\mu_1 = \mu_2 = \mu_3$

IF Reject

Tukey HSD test (to identify which group(s))

IF NONPARAMETRIC

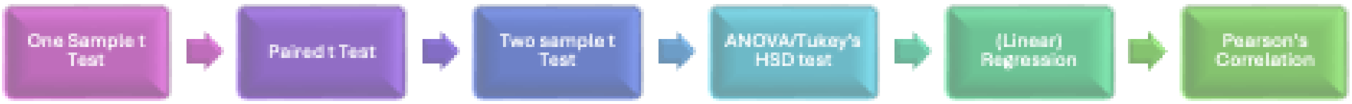
KRUSKALL-WALLIS

IF FTR

Ho: medians of all groups are same

if reject

Dunn's test to identify which group



Independent Random Samples

Normally Distributed

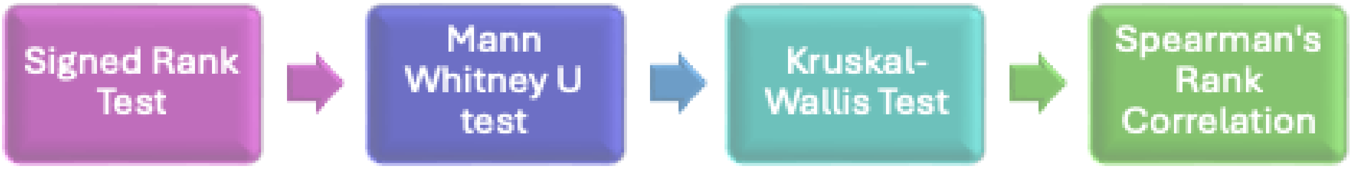
How to test:
- Visual Assessment (Boxplot, histograms, qqplot, qqnorm)
- Shapiro Wilk test

Homoscedasticity (variances are equal)

How to test
- Visual Assessment
- If normal; n=2 → var.test
- If normal; n>2 → Bartlett test
- If not norm → Levene's test
- If n=2 and variances are significantly different:

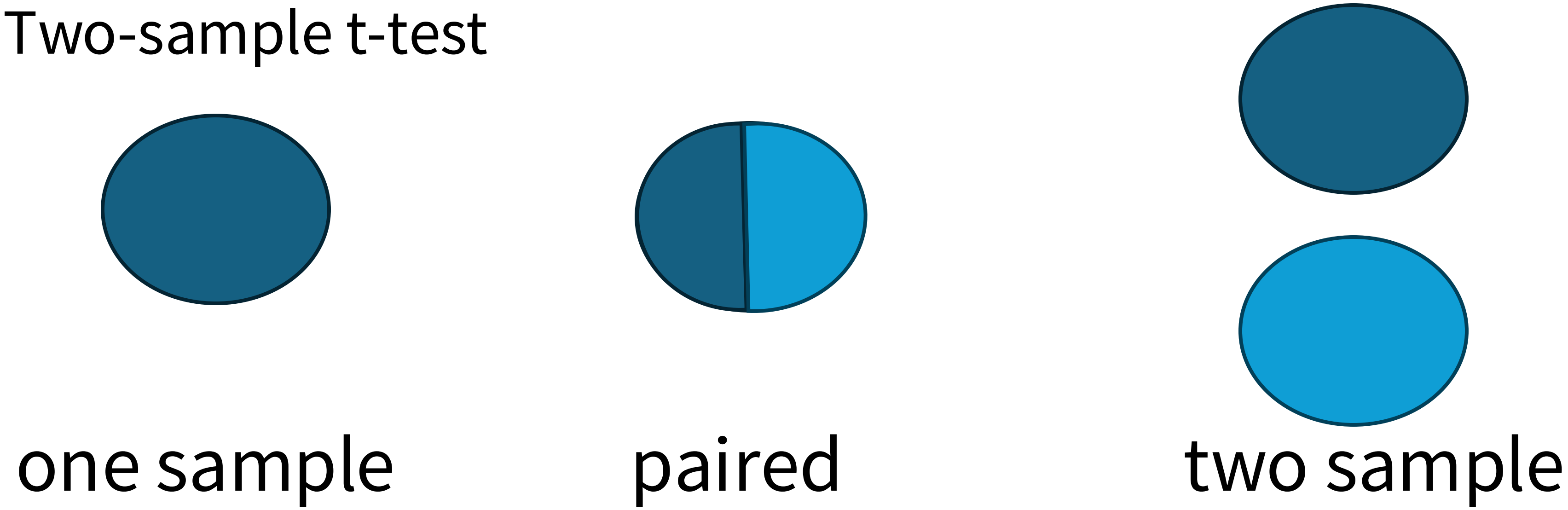
Welch's Approximate t test

Non-parametric analogs of the parametric tests above. Instead of testing population **means**, they test population **medians,** and they use **ranks.**



We will look at the following t-tests:

1. Comparing one mean:
   a. One-sample t-test
2. Comparing two means:
   a. Paired t-test
   b. Two-sample t-test



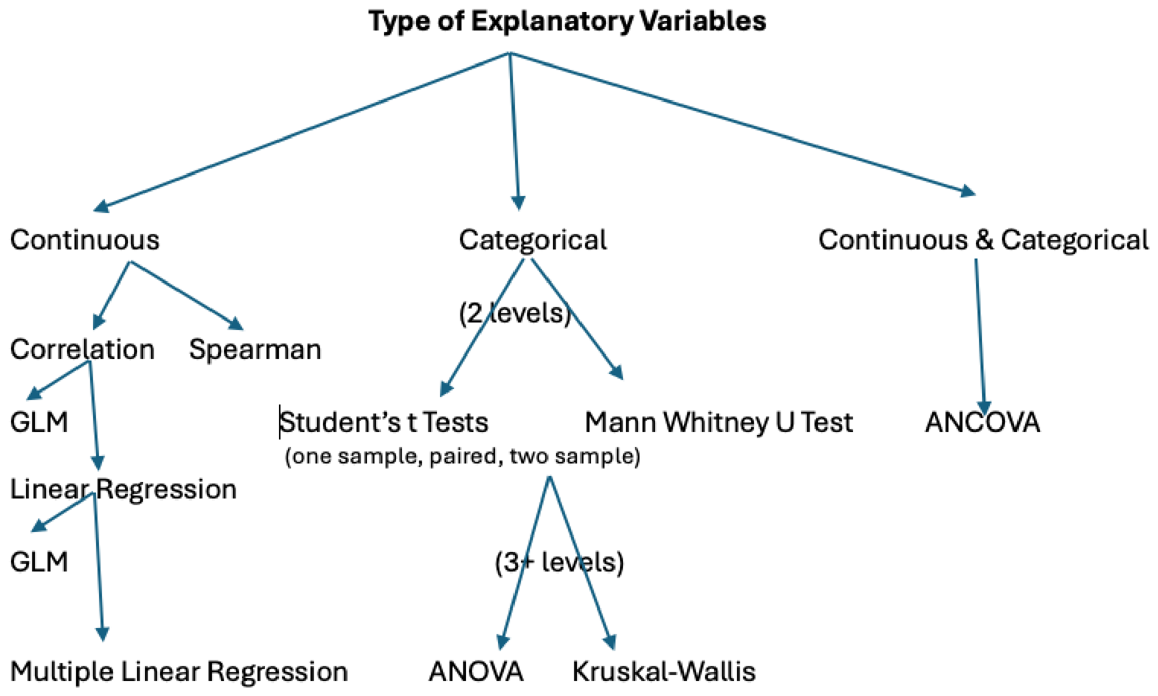
Each of the above tests have ***slightly different assumptions*** *which allow our conclusions to be supported. We will investigate what happens when these assumptions are violated and how robust our various t-tests are to violations.*

Agenda:

1. ANOVA: Nuts & Bolts

2. Worked Example
    1. One way ANOVA
    2. Post-hoc tests: Tukey-Kramer
    3. Kruskal-Wallis (nonparametric)

3. Linear Correlation
    1. Spearman’s rank



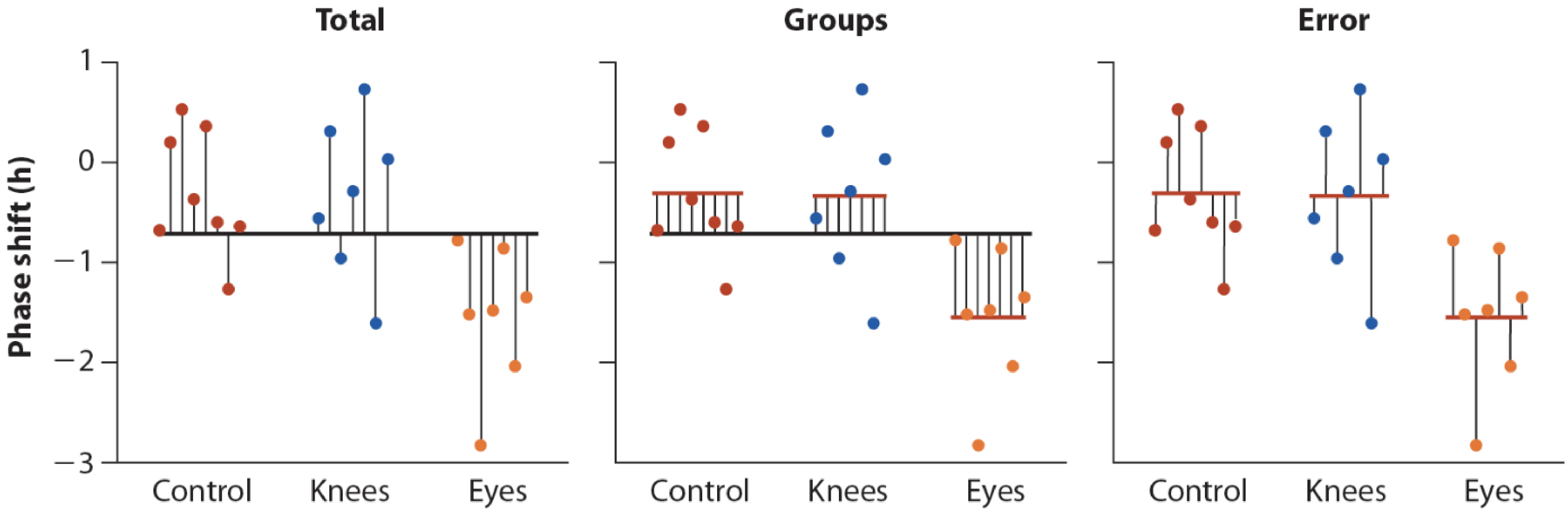
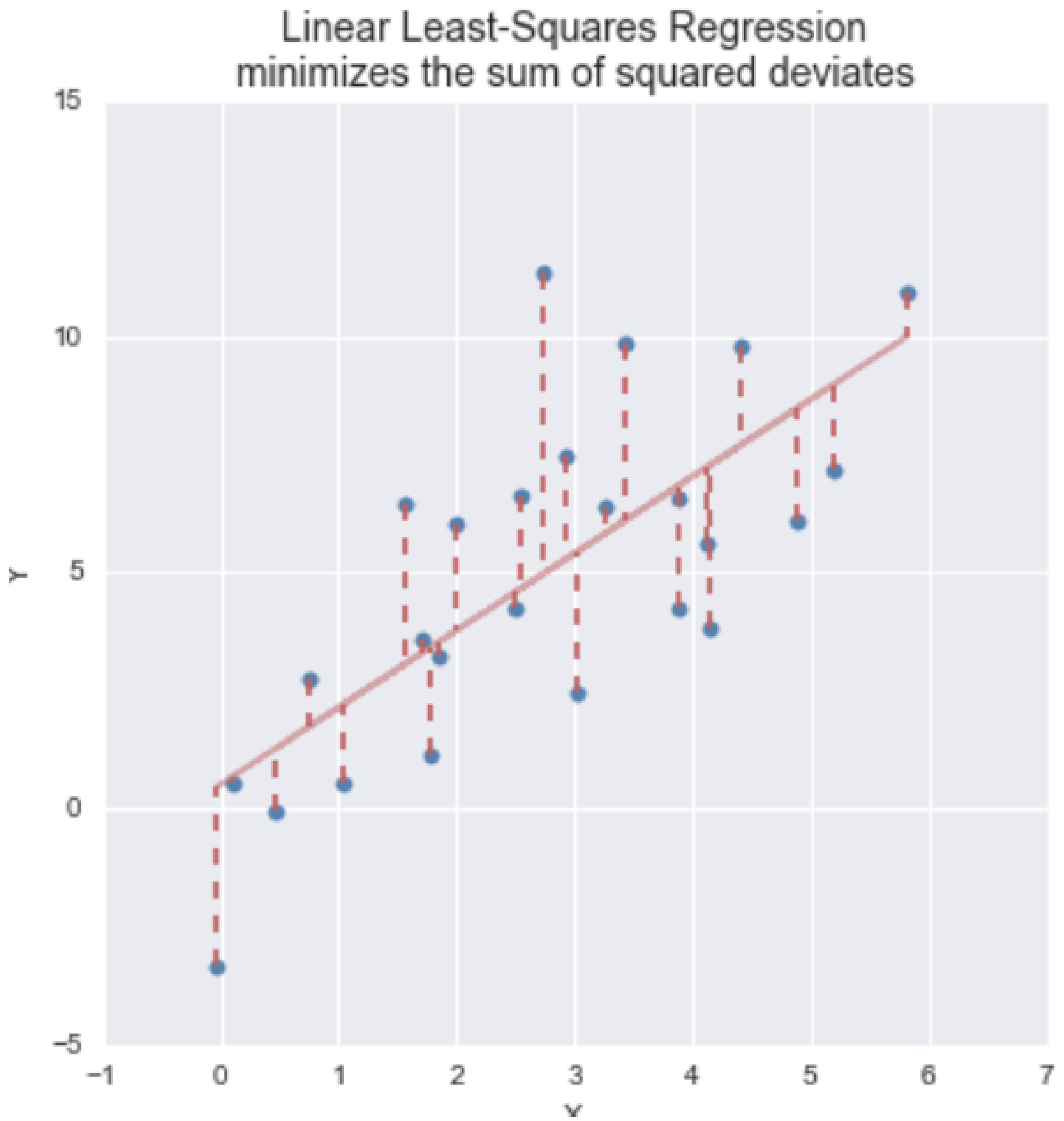


Figure 20.1: Whitock and Schluter, Fig 15.1.2 – Illustrating the partitioning of sum of squares into MS_{group} and MS_{error} components.



Linear Least-Squares Regression
minimizes the sum of squared deviates
15
10
5
0
−5
Y
−1
0
1
2
3
4
5
6
7
X