Module 3A: ANOVA & Correlation

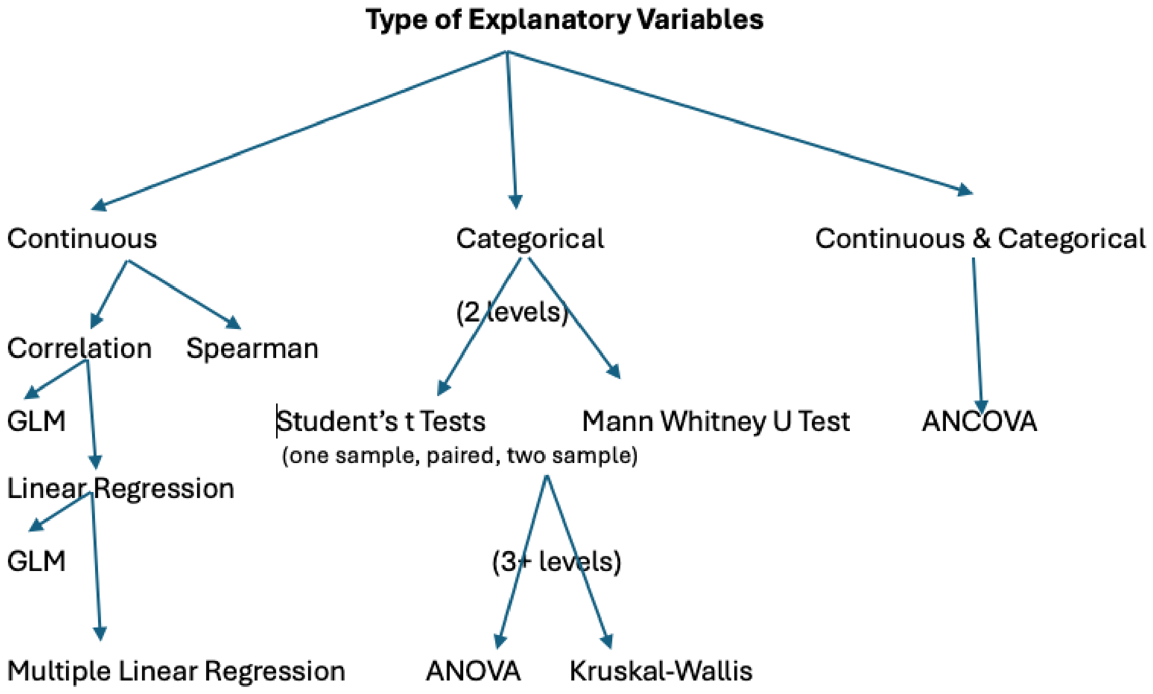
Assigning signal and noise to variation

Agenda:

1. ANOVA: Nuts & Bolts

2. Worked Example
   A. One way ANOVA
   B. Post-hoc tests: Tukey-Kramer
   C. Kruskal-Wallis (nonparametric)

3. Linear Correlation
   A. Spearman’s rank



Type of Explanatory Variables

Continuous

Categorical

Continuous & Categorical

Correlation

Spearman

(2 levels)

GLM

Student’s t Tests
(one sample, paired, two sample)

Mann Whitney U Test

ANCOVA

Linear Regression

GLM

(3+ levels)

Multiple Linear Regression

ANOVA

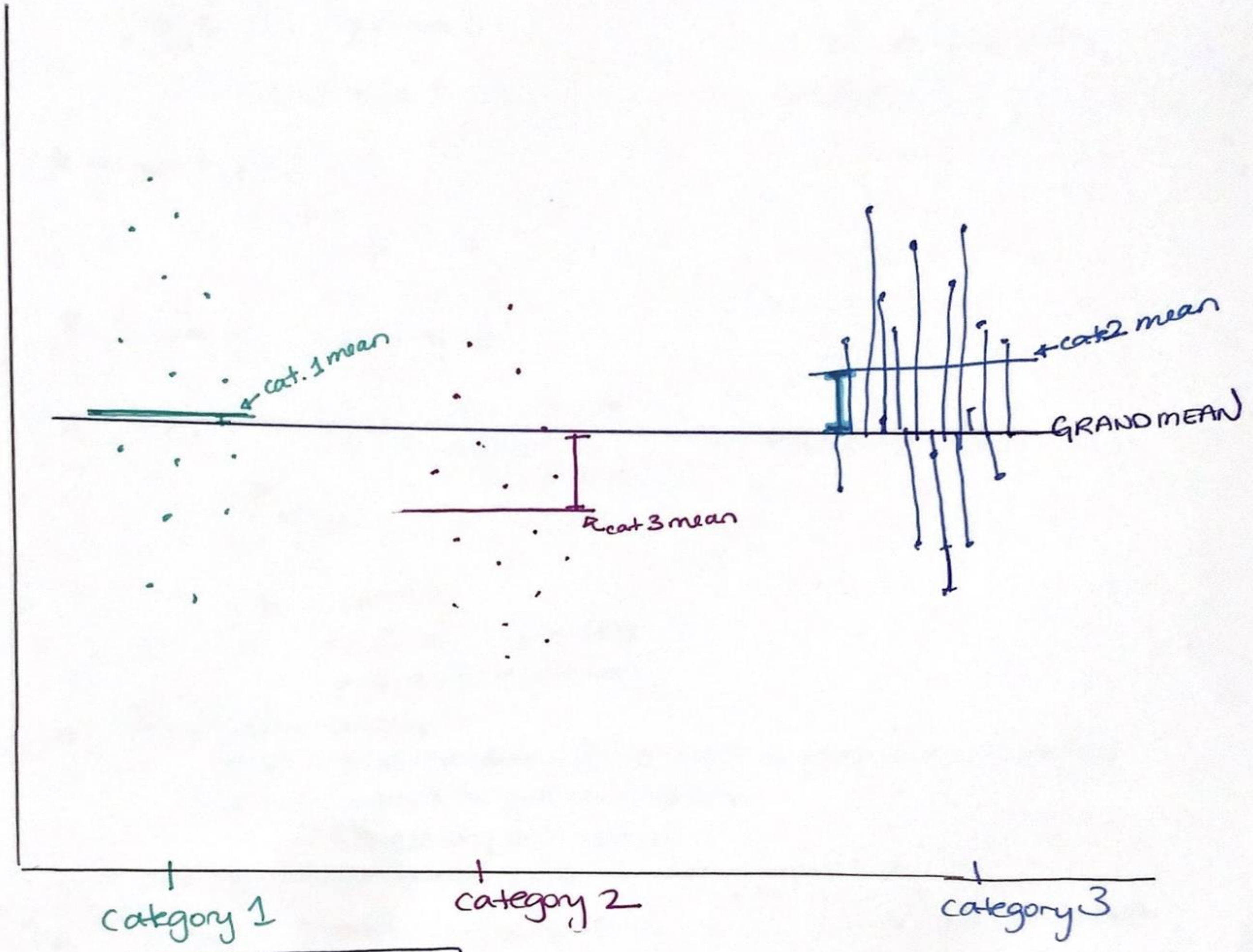
Kruskal-Wallis

Agenda:

1. ANOVA: Nuts & Bolts

2. Worked Example
    - One way ANOVA
    - Post-hoc tests: Tukey-Kramer
    - Kruskal-Wallis

3. Linear Correlation/Regression
    - Spearman's Rank



SIGNAL

$$\boxed{F = \frac{MS_{BETWEEN}}{MS_{ERROR}}}$$

NOISE

- Normally distr^d
- Variance equal (homoscedascity)

Analysis **o**f **Va**riance (ANOVA)

Purpose: compare the means of ≥ 2 groups (independent categorical variable) on 1 dependent continuous variable to see if the groups means are different from each other

- **Question:** Is the variance among groups greater than 0?
  - **Method:** Allocation of the total variability among different sources

Example:

Three independent categories: current best treatment, control, new treatment

Dependent continuous variable: blood pressure

Analysis of Variance

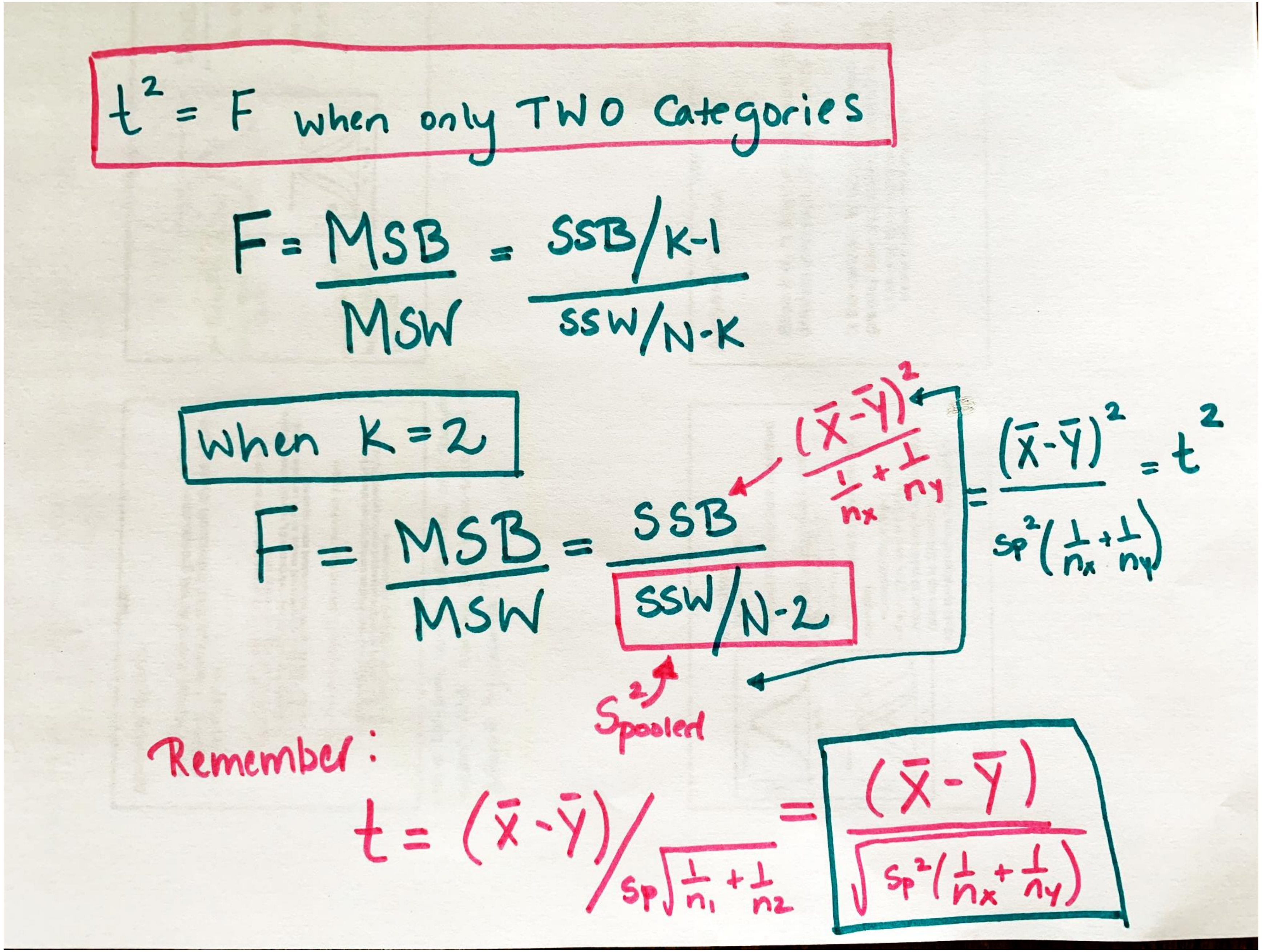
Purpose: compare the means of ≥ 2 groups (independent categorical variable) on 1 dependent continuous variable to see if the groups means are different from each other

Haven’t we already seen a test that compares means?

If there are **≤ 2** groups --> **t-test**

If there are ≥ **2** groups --> **ANOVA**

Why don’t we just use multiple t-tests?



$t^2 = F$ when only TWO Categories

$$F = \frac{MSB}{MSW} = \frac{SSB/K-1}{SSW/N-K}$$

When K=2

$$F = \frac{MSB}{MSW} = \frac{\overset{\frac{(\bar{x}-\bar{y})^2}{\frac{1}{n_x}+\frac{1}{n_y}}}{SSB}}{\underset{S^2_{pooled}}{\boxed{SSW/N-2}}} = \frac{(\bar{x}-\bar{y})^2}{s_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)} = t^2$$

Remember:

$$t = (\bar{x}-\bar{y}) \Big/ s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \boxed{\frac{(\bar{x}-\bar{y})}{\sqrt{s_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}}}$$

Analysis **o**f **Va**riance

- o Is the variance among groups greater than 0?
- o *Same question, different metric:* **Are the group means significantly different from each other and grand mean?**
    - o Allocation of the total variability among different sources

Why don't we just use multiple t-tests?

Answer: Like a *t-test* but can compare the means of > 2 groups *without inflating Type I error*

Analysis **o**f **Va**riance

Are individuals from different groups ***more different***, on average, than individuals chosen from the same group

- H_0: population means are equal, and sample means are only different due to random sampling error (noise)
- H_A: ***at least one mean*** is different from the other groups

$$H_0\text{: Variance among the groups} = 0$$

OR

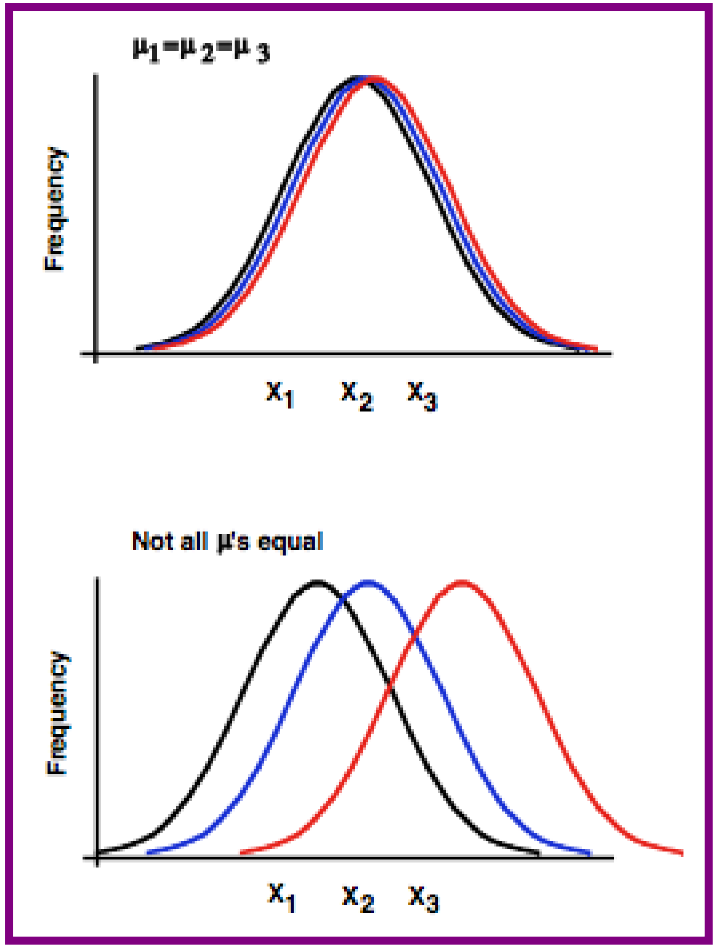
$$H_0\text{:}\ \ \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

Analysis of Variance

Are individuals from different groups ***more different***, on average, than individuals chosen from the same group

- H_0: population means are equal, and sample means only different due to random sampling error
  - Standard error of the null distribution (H_0 is true) is the standard deviation of the group (sample) means so the variance among groups should just be the standard error squared

- H_A: ***at least one mean*** is different from the other groups
  - IF H_0 is NOT true, the variance among groups should be equal to the variance of sample (standard error squared) **PLUS** the real variance among population means



Assumptions:

1. Random samples

2. Normal distribution (each population)

3. Variance among groups is equal

homoscedasticity

- ANOVA is robust to departures from normality
  - especially if n_i is large (Thanks, CTL!)

- If $n_1 = n_2 = n_3$ (and n = large) robust to violations in equal variance (allow up to 10X variance)

- Data transformations can be used if necessary

Analysis **o**f **Va**riance

→ ***Even if H_0 is true***, sample means will be different from each other by chance

Question: **Is the variation among sample means *greater* than expected by chance alone?**

- This is evidence that at least one of the population means is different from the others

Assumptions of ANOVA:

- Measurements are random sample
- Variable is normally distributed
- **Variance is the same in all *k* populations**

How do we handle violations in these assumptions?

1. Robustness (ignore)
   - If data is not normal BUT sample size is large (CLT)
   - variances are not equal, but sample sizes are approximately equal
2. Data Transformation
3. Non-parametric alternative → Kruskal Wallis H test

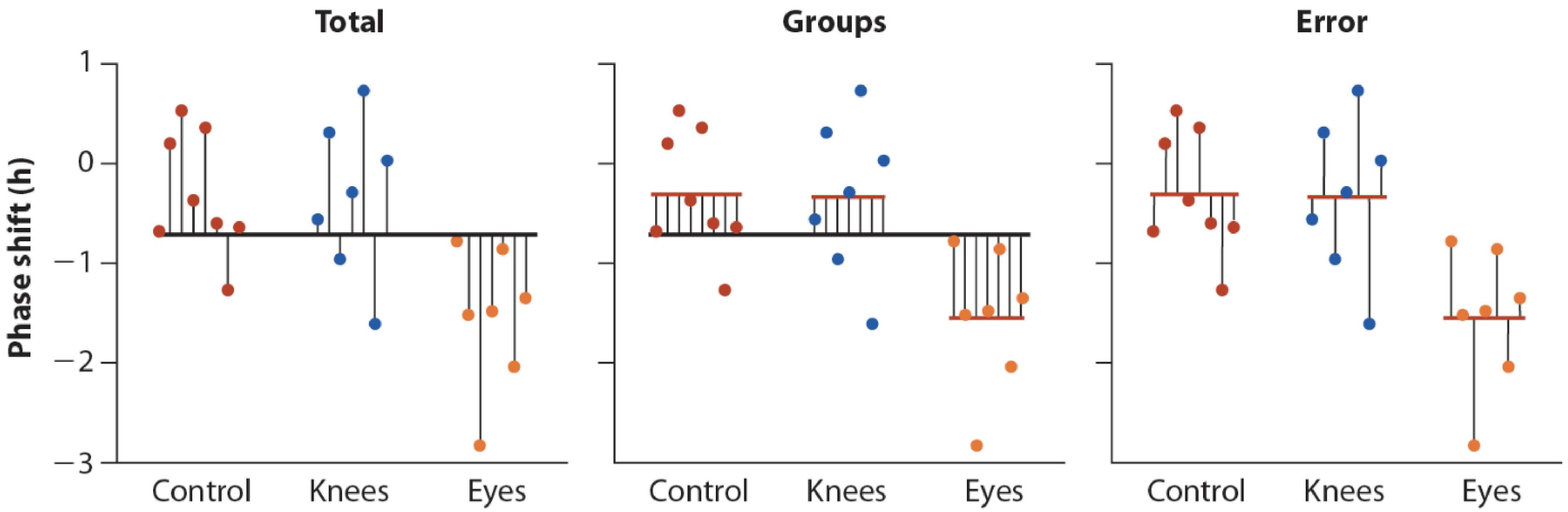


Figure 20.1: Whitock and Schluter, Fig 15.1.2 – Illustrating the partitioning of sum of squares into MS_{group} and MS_{error} components.

- Error Mean Square:
  - A measure of variability within groups

- Group Mean Square:
  - Represents variation among individuals belonging to different groups

Conceptual Crux of ANOVA:

If H_0 is true, then group means should be the same so the two types of mean square should be equal

$$\mathbf{MS_{error} = MS_{groups}}$$

Under H_0, the sample mean of each group ***should only vary*** *because of sampling error*

The standard deviation of sample means, when the true mean is constant, is just the standard error:

$$\boxed{\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}}$$

Squaring the standard error, the variance **among** groups due to sampling error is:

$$\boxed{\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}}$$

If H_0 is **not** true, the variance **among** groups should be equal to the variance due to sampling error ***plus*** the real variance among population means

$$\boxed{\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n} + Variance(\mu_i)}$$

ANOVA tests whether the variance among true group means is significantly greater than zero

We do this by asking whether the observed variance among groups is greater than expected by chance

$$\begin{tikzcd}
\boxed{\sigma^2_{\bar{X}} > \frac{\sigma^2_X}{n}} \arrow[d] \\
\boxed{n\sigma^2_{\bar{X}} > \sigma^2_X}
\end{tikzcd}$$

Population Parameters

Estimates from Samples

$$\boxed{n\sigma^2_{\bar{X}}}$$

Estimated by the “mean square group”

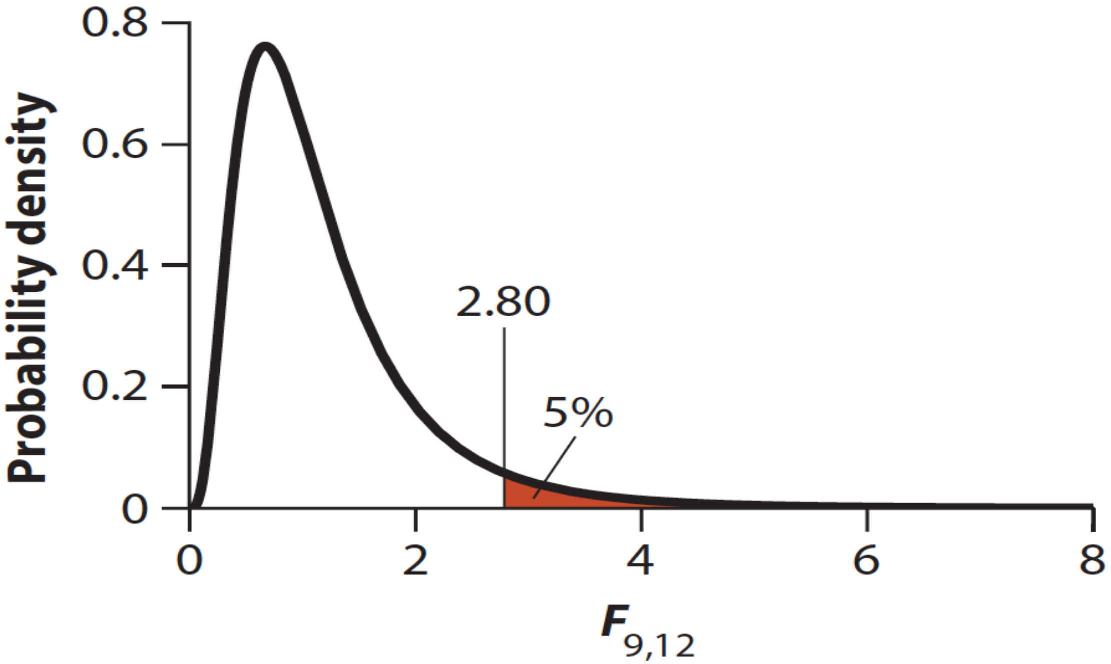
Since it should (almost) **always be the larger value, it is in the NUMERATOR**

MS_{group}

$$\boxed{\sigma^2_X}$$

- The variance within groups
- Estimated by “mean square error”
- One of the assumptions of ANOVA is that this variance is *approximately the same between different groups*

MS_{error}



SIGNAL

$$\text{F-value} = \frac{\boxed{MS_{group}}}{\boxed{MS_{error}}}$$

NOISE

- This is a **one-sided test** which is different from the F test that we used previously to test variances between populations.

- ANOVA F test is one-sided because MS_{group} is ALWAYS in the numerator (there isn’t a 50:50 chance like in the F test for equal variances).

$$\text{F-value} = \frac{\overset{\text{SIGNAL}}{\boxed{MS_{group}}}}{\underset{\text{NOISE}}{\boxed{MS_{error}}}}$$

- reminder: t-tests also involve a ratio
  - numerator in a t-test is the difference between two sample means
  - numerator in ANOVA is *average* difference between means squared

- denominator is equivalent in both:
  - t-test: standard error of difference between means
  - ANOVA: average error within groups squared

summary: *just like in the t-test, in ANOVA we are trying to determine the average difference* ***between*** *group means relative to the average difference* ***within*** *group means*

Conceptual Crux of ANOVA:

If H_0 is true, then group means should be the same so the two types of mean square should be equal

$$MS_{error} = MS_{groups}$$

$$\boxed{F = \frac{MS_{groups}}{MS_{error}} \geq 1}$$

If F≈ 1, we FTR $H_{o.}$ If F >>1, there is enough evidence to reject H_0

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{\square\; s_i^2(n_i-1)}{N-k}$$

$$SS_{error} = \square\; df_i s_i^2 = \square\; s_i^2(n_i-1)$$

$$df_{error} = \square\; df_i = \square\; (n_i-1) = N-k$$

Mean of group i

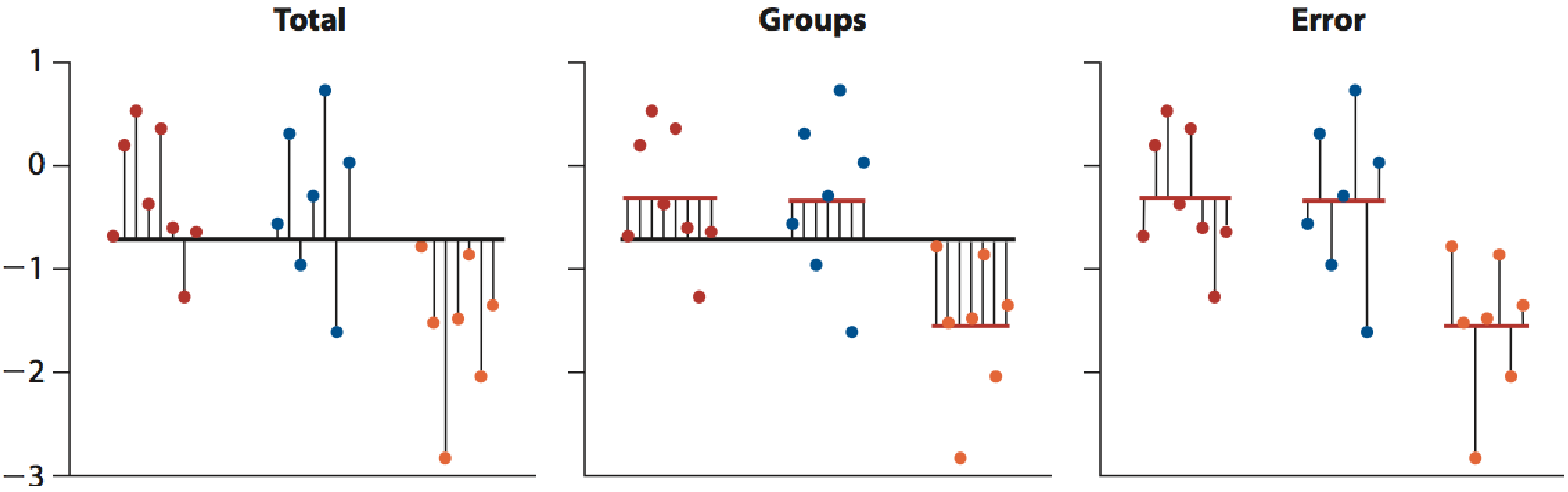
$$MS_{groups} = \frac{SS_{groups}}{df_{groups}} = \frac{\sum n_i(\bar{X}_i - \bar{X}_T)^2}{k-1}$$

$$\bar{X}_T = \frac{\underset{i}{\square}\;\underset{j}{\square}\; X_{ij}}{N}$$

$$\bar{X}_T = \frac{\underset{i}{\square}\; n_i\bar{X}_i}{N}$$

Results are presented in ANOVA Table:

Source of variation	Sum of Squares	df	Mean Squares	F-ratio	P
Groups (treatment)					
Error					
Total					



R^2 value:

- The fraction of variability that is explained by groups
- Measures reduction in scatter around group means compared to the grand mean

$$SS_{Total} = SS_{groups} + SS_{error}$$

$$R^2 = \frac{SS_{groups}}{SS_{Total}} \quad ; 0 < R^2 < 1$$