ANOVA

In a one-way ANOVA with 3 groups, a rejection of the null hypothesis implies that

A. the 3 population means are equal to each other
B. the 3 sample means are equal to each other
C. each population mean differs significantly from all other population means
D. some subset of population means differs from some other subset of population means
E. some subset of sample means differs from some other subset of sample means

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

- 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 $K = 2$

$$F = \frac{MSB}{MSW} = \frac{SSB}{SSW/N-2}$$

$$\frac{(\bar{X}-\bar{Y})^2}{\frac{1}{n_x}+\frac{1}{n_y}} = \frac{(\bar{X}-\bar{Y})^2}{s_p^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)} = t^2$$

s^2_{pooled}

Remember:

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

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

- Asks the question "Are any two (or more) group means significantly different from the other means"?

- Similar to a *t-test* but can compare the means of > 2 groups *without inflating Type I error*

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

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 that sample means only different due to random sampling error
- H_A: ***at least one mean*** is different from the other groups

Null hypothesis for simple ANOVA

H_0: Variance among the groups = 0

OR

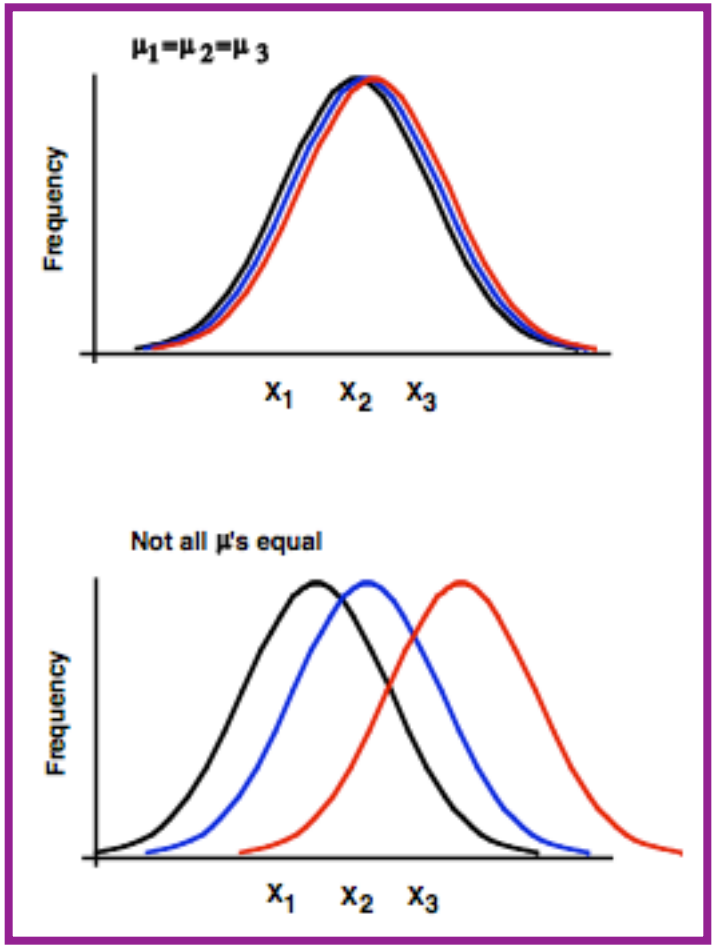
H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

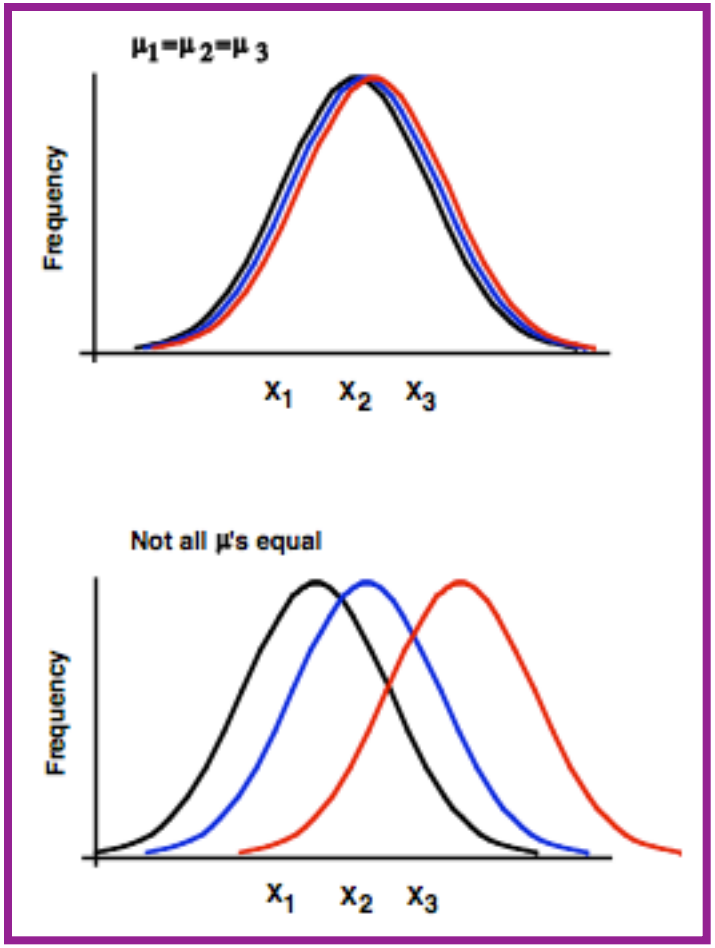
Analysis of Variance

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

- H_0: population means are equal and that 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**

ANOVA


$\mu_1=\mu_2=\mu_3$
Frequency
X_1 X_2 X_3
Not all μ's equal
Frequency
X_1 X_2 X_3



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
- 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 of Variance

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

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
   - If data is not normal BUT sample size is large(CLT)
   - *variances are not equal but only if sample sizes are approximately equal*
2. Data Transformation
3. Non-parametric alternative

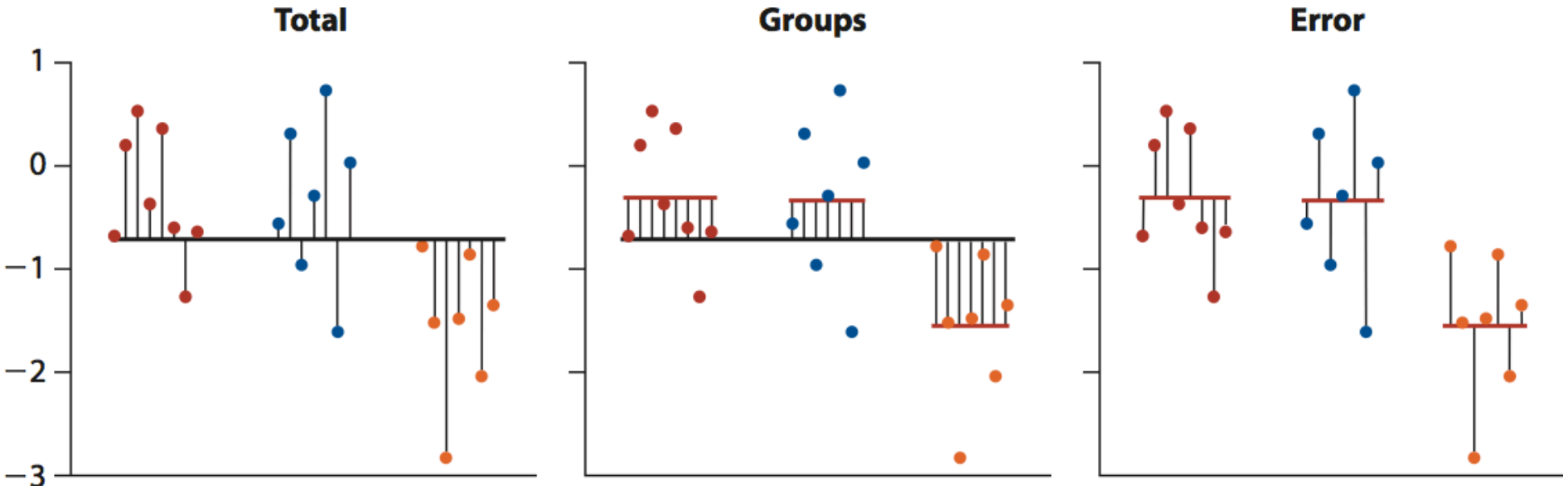


Figure 15.1-2

Demonstrates how to apportion out the variance

Each dot is measurement in a single subject

Horizontal black line is the overall mean, $\overline{Y}$

Horizontal red lines are sample means

- 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 the standard error:

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

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

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

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

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

ANOVA tests whether or not 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:

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

$$n\sigma^2_{\bar{X}} > \sigma^2_X$$

Population Parameters

Estimates from Samples

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

Is estimated by the "mean square group"

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

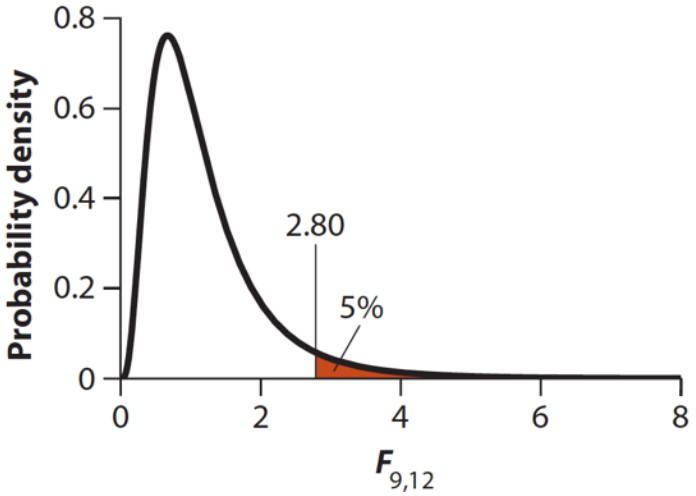
$\mathrm{MS}_{\mathrm{group}}$

σ^2_X

Is the variance within groups estimated by "mean square error" Remember that one of the assumptions of ANOVA is that this variance is approximately the same between different groups

$\mathrm{MS}_{\mathrm{error}}$

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



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 previous F test)

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

- 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

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

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

If F≈ 1, we FTR H_o. If F >1, there is enough evidence to reject H_0

Mean Squares:

1. MS_{error} → Estimates this parameter → σ_x^2

- Measures variance within groups *(the 'noise' part)*

Basic Formula: $$MS_{error} = \frac{SS_{error}}{df_{error}}$$

Mean Squares:

1. MS_{error} — Estimates this parameter → σ_x^2

- Error sum of squares:

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

- Error degrees of freedom:

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

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

- **MS_{error} is like the pooled variance in a 2-sample t-test:**

- $$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}$$

Mean Squares:

Short cut for SS_{Error}

1. MS_{error} → Estimates this parameter → σ_x^2

    1. Subtract the group mean from each individual score in each group

    $$(X - \bar{X})$$

    2. Square each of these deviation scores

    $$(X - \bar{X})^2$$

    3. Add them all up for each group

    $$\sum (X - \bar{X})^2$$

    4. Then add up all of the sums of squares for all of the groups

    $$\sum (X - \bar{X}_1)^2 + \sum (X - \bar{X}_2)^2 + \ldots + \sum (X - \bar{X}_k)^2$$

Mean Squares:

2. MS_{group} → Estimates this parameter → $n(\sigma^2_{\bar{x}} + Variance[\mu_i])$

- Measures variance among groups (is there any actual 'signal')

Basic Formula: $$MS_{groups} = \frac{SS_{groups}}{df_{groups}}$$

Mean Squares:

2. MS_{groups}

Estimates this parameter → $n(\sigma^2_{\bar{x}} + Variance[\mu_i])$

Mean of group i

$$SS_{groups} = \sum n_i(\overline{X}_i - \overline{X}_T)^2$$

$$\overline{X}_T = \frac{\sum_i \sum_j X_{ij}}{N}$$

$$df_{groups} = k - 1$$

$$MS_{groups} = \frac{SS_{groups}}{df_{groups}}$$

$$\overline{X}_T = \frac{\sum_i n_i \overline{X}_i}{N}$$

Mean Squares:

Shortcut for SS_{group}

2. MS_{group} Estimates this parameter → $n(\sigma^2_{\bar{x}} + Variance[\mu_i])$

1. Subtract the ***grand*** mean from the ***group*** mean

$$(\overline{X}_i - \overline{X}_T)$$

2. Square each of these deviation scores

$$(\overline{X}_i - \overline{X}_T)^2$$

3. Multiply each squared deviation by the number of cases in each group:

$$n_i(\overline{X}_i - \overline{X}_T)^2$$

4. Then add up all of the sums of squares for all of the groups

$$\sum n_i(\overline{X}_i - \overline{X}_T)^2$$

Test Statistic:F

If H_0 is true, then: $$n\sigma_{\bar{x}}^2 = \sigma_x^2$$

In other words: $$F_{\alpha(1),k-1,N-k} = \frac{n\sigma_{\bar{x}}^2}{\sigma_x^2} = 1$$

But the above refer to population parameters. We must estimate F from samples with: MS_{group}/MS_{error}

F if H_0 is false:

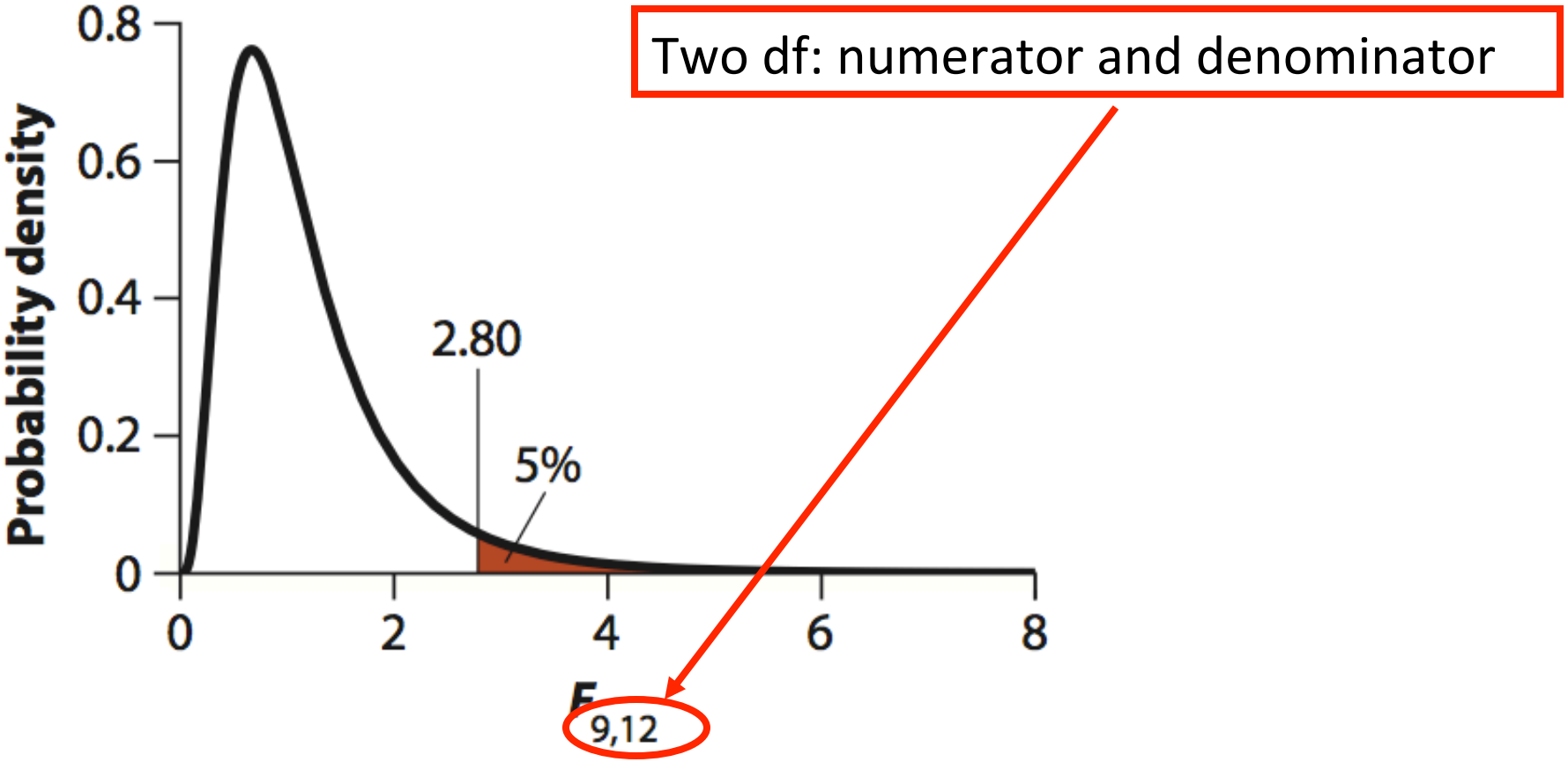
We test whether the F ratio is significantly greater than 1, as it would be if H_0 is false

$$F = \frac{n(\sigma^2_{\bar{x}} + Variance[\mu_i])}{\sigma^2_x} > 1$$

But what about sampling error? F calculated from the data can be > 1 *even when H_0 is true. Compare F to a null distribution.*

The F distribution:

- One sided
- Statistical Table D

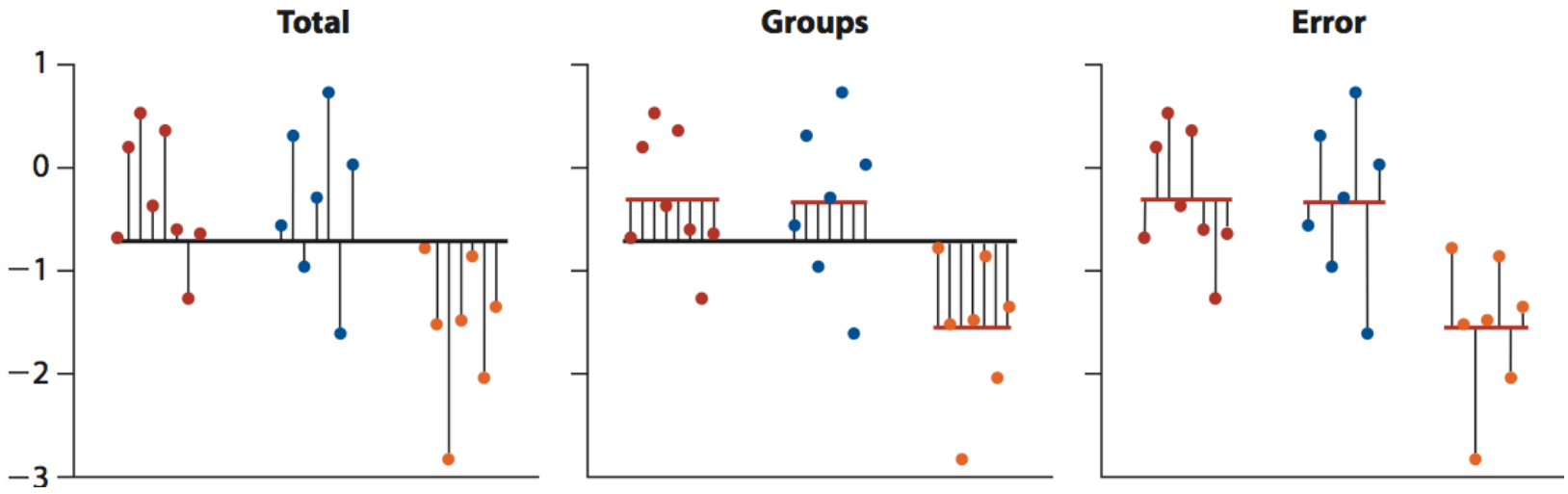


Results are presented in ANOVA Table:

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

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 value:

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

$$R^2 = \frac{SS_{groups}}{SS_{Total}}$$

$$0 < R^2 < 1$$

Means are similar;
Variability is within groups

Little variation left over after different group means are considered;
Variability is BETWEEN groups

- Safest to think of R^2 as a measure of the difference In **scatter** of the different groups.