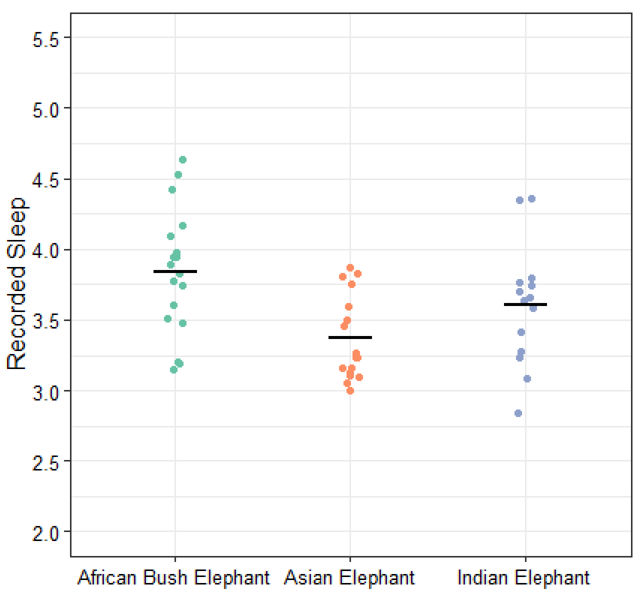
(STAT 200 Only) Chapter 12: Comparing Three or More Means

Comparing Three or More Populations

- Need for a new approach
  - To compare a parameter from three or more populations, we need to change our approach.
    1) We could do pairwise comparisons of all group combinations, but the more combinations we have, the more chances to get a low p-value by random chance.
    2) *Do you remember what type of error that would be?*

  - Instead, we will complete one test with all groups together. This test would consider the null hypothesis that the parameters from all populations are equal.
    1) **H_0: $\mu_1 = \mu_2 = \mu_3$ ...** (for means)
    2) **H_0: $p_1 = p_2 = p_3$ ...** (for proportions)
  - If we reject this null hypothesis, we can then move to pairwise comparisons.

- Focusing on Means
  - If comparing 3+ means, One-Way ANOVA is a good option

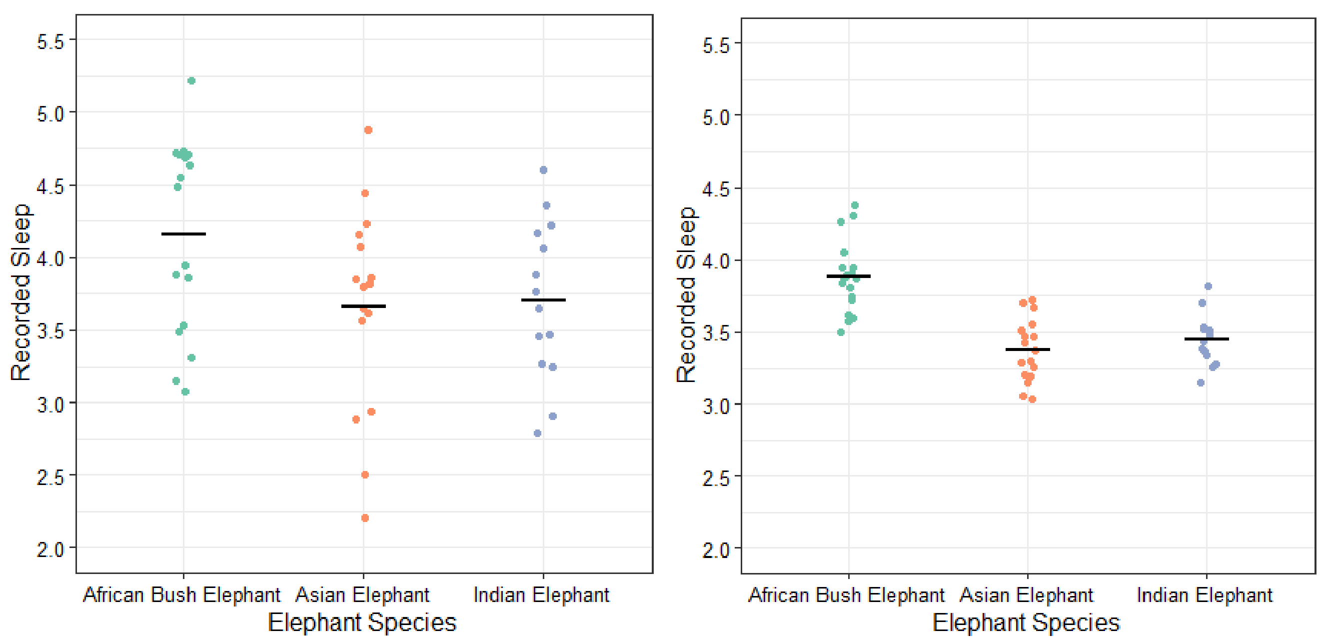


Comparing Elephant Sleep Times

Do these three species of elephants require the same amount of daily sleep on average? The following represents ***fake*** data for sake of this example, recording the total sleep of 17, 14, and 18 elephants of each species. The mean of each sample is plotted as a black bar.

Introducing the ANOVA Method

- What is "ANOVA"?
  - ANOVA stands for ___________________________. It is a way of comparing the means of three or more groups by doing a comparison of different types of variance.
  - **Variance**: Recall that the variance is just the standard deviation squared. It is a way to measure how much data points vary from the mean.
  - When comparing two groups, there are two sources of variation we can identify:
    1) **Variance Within:** Refers to the variation __________________________. In our example, not all elephants of the same species got the same amount of sleep. That is variation from other sources besides the source of interest (species).
    2) **Variance Between:** Refers to the variation _____________________________. In our example, each species does not have the same mean sleep time. This may be because species is a source of variation, or it could just be sampling variability.
  - Let's exaggerate the data now.



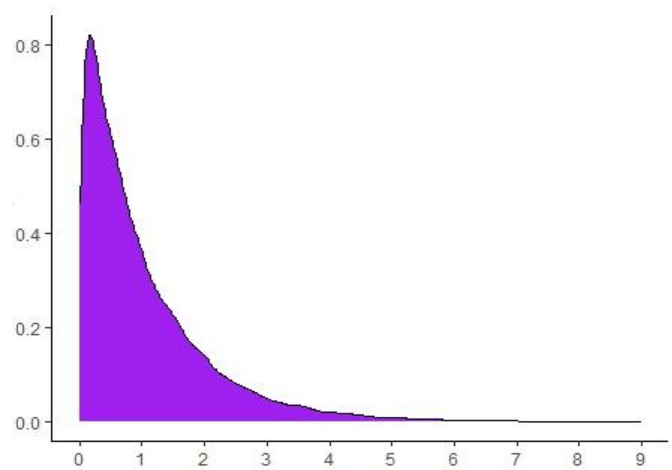
What is different about each situation above? Which set of data provides **stronger evidence** for a **difference** in average sleep between species—The data on the left, or the data on the right?

The Process of ANOVA

- The F distribution and F-test
  - The idea behind ANOVA is to deconstruct the variance in our response variable into two types: The variance from within groups and the variance from between groups.
  - We measure the variance from each source, and then calculate their ratio for comparison.

Informal Representation: **F-ratio** = $\frac{\textbf{Average Between Variance}}{\textbf{Average Within Variance}}$

  - The null hypothesis here is that there is no difference between groups, and this ratio should be relatively low.
  - BUT…even if the null hypothesis is true, we expect to see some variation between the sample means. We expect the ratio to be near ___
  - If the null hypothesis is true, then there is a distribution of possible F ratios we should expect.
  - It will change for different sample sizes and numbers of groups. You can check out how the distribution changes for different situations here: https://istats.shinyapps.io/FDist/



*F Distribution for 3 Groups of size 20 each **if Null is true***

- Calculating an F-Ratio: Variance ____________ groups
  - We start by calculating the sum squared error for each group and adding up across groups.
  - We call this intermediary derivation **The Sum of Squares Error Within**: abbreviated **SS(Within).**

SS(Within) = $\underbrace{\sum_{j=1}^{n_1}(x_{j1}-\bar{x}_1)^2}_{\text{Group 1}} + \underbrace{\sum_{j=1}^{n_2}(x_{j1}-\bar{x}_2)^2}_{\text{Group 2}} + \ldots + \underbrace{\sum_{j=1}^{n_i}(x_{j1}-\bar{x}_i)^2}_{\text{Group i}}$

  - To calculate the Mean Squared Error Within—abbreviated MS(Within)—we would take the SS(Within) and divide by the degrees of freedom for this calculation.
    - *df would be total number of data points minus number of groups.*

MS(Within) = $\frac{\sum_{j=1}^{n_1}(x_{j1}-\bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{j1}-\bar{x}_2)^2 + \ldots + \sum_{j=1}^{n_i}(x_{j1}-\bar{x}_i)^2}{n-i}$ =

You will not need to calculate either of these by hand!

- Calculating an F-Ratio: Variance _______________ groups
  - We now switch to measuring the variation from group to group. Specifically measuring the average variation from each sample mean to the global mean.
    - **Global mean**: The mean measurement from _____ of the data values.
  - Multiplying each squared difference by n_i weights each difference by the group's sample size.

SS(Between) = $\underbrace{n_1(\bar{x}_1 - \bar{x})}_{\text{Group 1}} + \underbrace{n_2(\bar{x}_2 - \bar{x})}_{\text{Group 2}} + \ldots + \underbrace{n_i(\bar{x}_i - \bar{x})}_{\text{Group i}}$

  - Now, we find the average group mean deviation from the global mean by dividing by number of groups. However, we divide by $i - 1$ as the degrees of freedom error correction.

MS(Between) = $\dfrac{n_1(\bar{x}_1 - \bar{x}) + n_2(\bar{x}_2 - \bar{x}) + \ldots + n_i(\bar{x}_i - \bar{x})}{i - 1}$ =

- Creating an ANOVA Table
  - An ANOVA Table is a convenient and helpful way of organizing all of our information
  - The following table illustrates the shape and structure of a standard ANOVA table.

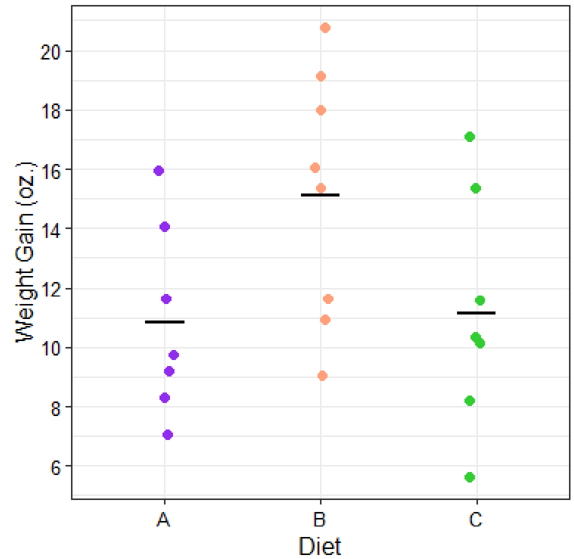
Source of Variation	*Degrees of Freedom (df)*	*Sum of Squares (SS)*	*Mean Square (MS)*	*F-Statistic*	*P-value*
Between Groups	**i - 1**	**SS(between)**	**SS(between)/df(between)**	**MS(between)/MS(within)**	**P(F>f)**
Within Groups	**n - i**	**SS(within)**	**SS(within)/df(within)**	**---**	**---**

- Let's review some terms and notation
  - i = ________________________________
  - n = _________ number of observations
  - n_i = the number of observations in ___________
  - $\bar{x}$ = ________________
  - $\bar{x}_i$ = _________ for group i

Hypothesis Testing using F-Test

- Lamb Diets Example
  - We are investigating the weight gains of lambs on 3 different diets over two weeks. Is there evidence that one diet might produce a different weight gain than another on average? 22 lambs were randomly assigned to each diet, with the results below.

	Weight Gains of Lambs (oz.)		
	Diet A	**Diet B**	**Diet C**
	8	9	15
	16	16	10
	9	21	17
	14	11	6
	10	18	10
	12	15	12
	7	19	8
		12	
Means	**10.86**	**15.13**	**11.14**
Variances	**9.27**	**15.36**	**12.69**
Global Mean = 12.5			



i = ____ n = ____ $\bar{x}$ = ____

$\bar{x}_1$ = ____ $\bar{x}_2$ = ____ $\bar{x}_3$ = ____

- **Stating Assumptions**
  - The groups must come from population distributions that are approximately normally distributed **OR** the sample sizes must be ____________________________.
    1) This can be checked using simulation methods, but in general, group sizes > <u>20</u> are often large enough as long as there isn't a long tail for any of the groups.
  - The **variances** of each group are relatively similar.
    2) *A good benchmark is that no variance is _____________ bigger than another.*

Practice: Do we meet the conditions to complete a reliable F-test for mean comparison?

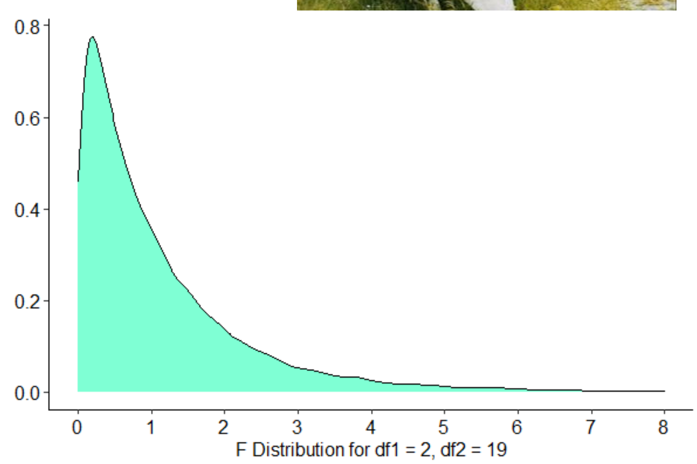
- **Stating Hypotheses:**
  - The null hypothesis: "All three diets result in the same average weight gain for lambs"
    - H_0: ____________________
  - The alternative hypothesis: "At least one of the diets causes an average weight gain that is not equal to the other diets' average weight gain."
    - H_A: ____

- **Completing the F-test for a p-value**
  - To conduct an F-test, we need to calculate our test-statistic: F = $\frac{\text{MS(between)}}{\text{MS(within)}}$
  - Using R, I was able to run an ANOVA to get the following results.

Source	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F-statistic	P-value
Between Groups	2	86.91	43.46		0.075
Within Groups	19	276.57	14.56	---	---

Practice: What is the missing F-ratio for this analysis?

- Making a conclusion:
  - For the lamb data, our p-value is about 7.5%
  - We can identify where this falls on the appropriate F distribution and identify the p-value region.



P-VALUE INTERPRETATION: *If the null hypothesis is true,* there is a 7.5% chance of getting …

What do you think? Do we have evidence that the diet is a source of variation in weight gain, or could differences between the diets be explained by sampling variability?

Pairwise Comparisons

- **Tukey Test (not "Turkey" Test!)**
  - If we discover a difference in means from the ANOVA, we can turn to making pairwise comparisons.
  - One option is the Tukey Honest Significant Difference Test, where we calculate a test statistic called the HSD.
  - The primary difference between an independent samples t-test and a Tukey HSD test is that we use the ______________________________ from all groups as a basis for average variance.
  - **Don't worry about calculating it by hand! We just focus on interpreting results!**

$$\text{HSD} = \frac{M_1 - M_2}{\sqrt{\frac{MS_W}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  - As stated earlier in the chapter, with this many groups, it's not unusual for one sample mean to be unusually high and another unusually low. For this reason, the Tukey Test adjusts what test statistics correspond to what p-values.

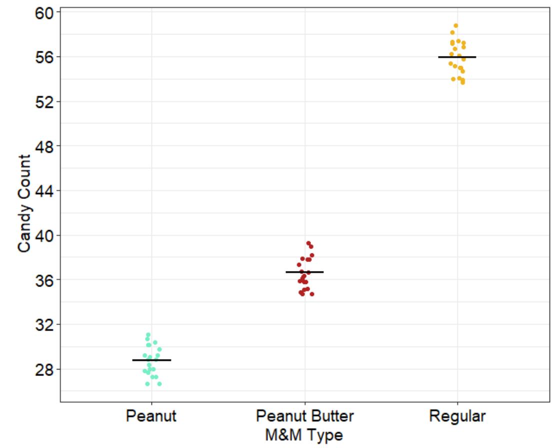
Practice: Is there a relationship between the sport that someone plays and their resting heart rate? One study looked at young adults who played Volleyball, Tennis, Baseball, or Soccer between once and twice a week. After an ANOVA analysis showed a p-value of 0.0206, a Tukey HSD Test was run to identify which group or groups might have populations means that are statistically different from the others:

```
Tukey HSD Post-hoc Test...
     Volleyball vs Baseball: Diff=6.2000, 95%CI=-0.5788 to 12.9788, p=0.0901
     Volleyball vs Tennis: Diff=0.6900, 95%CI=-6.3165 to 7.6965, p=0.9988
     Volleyball vs Soccer: Diff=-1.3100, 95%CI=-7.8531 to 5.2331, p=0.9811
     Baseball vs Tennis: Diff=-5.5100, 95%CI=-12.3724 to 1.3524, p=0.1779
     Baseball vs Soccer: Diff=-7.5100, 95%CI=-13.8986 to -1.1214, p=0.0126
     Tennis vs Soccer: Diff=-2.0000, 95%CI=-8.6297 to 4.6297, p=0.9189
```

Are there any pairwise differences to report?

Practice: We're investigating Regular, Peanut and Peanut Butter M&Ms. Do each of these kinds of M&Ms have the same number of candies in a bag on average in a 1.69 ounce bag? In a previous class I taught, we opened 20 bags of *each* type (60 bags total) and counted how many M&Ms were in each bag.

Are assumptions met to conduct an F-test here?

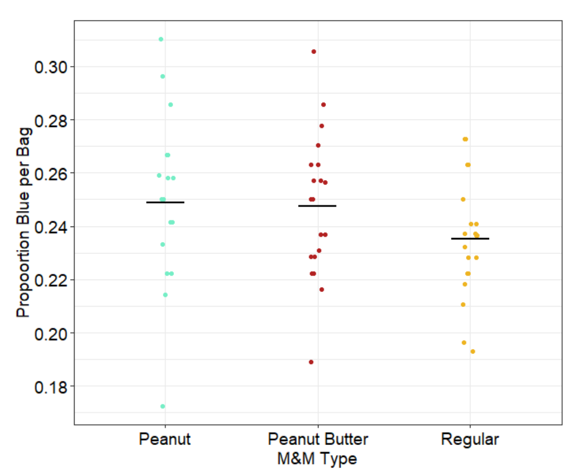


Write out the null and alternative hypotheses for this question.

Based on this plot, make an informal inference. Do you think our sample results provide sufficient evidence to conclude there is <u>really a difference</u> in the average amount of M&Ms in each of these types, or could these differences be explained by sampling variability?

We formally test the results and get a p-value of 2×10^{-16}. Written out, that would be 0.0000000000000002. How would we interpret this?

Practice: Now, let's focus on whether or not Regular, Peanut, and Peanut Butter M&Ms have the same or different proportions of blues on average. Based on the plot, would conditions be met for an F-test?



Make an informal inference. Will we have enough evidence to conclude a difference?

An ANOVA is run and the ANOVA table is presented here:

	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F Statistic	P-value
Between groups	2	0.00225	0.00125		0.215
Within Groups	57	0.04063	0.00071281	--	--

What should the F-statistic come out to be?

Based on our p-value, what might we conclude? Is it plausible to think the proportions of Blue in each type are the same, or do we have evidence that maybe the proportions are different for different types of M&M bags?

Even though the overall result is not significant, let's check the pairwise comparisons for practice. We run a Tukey HSD Test to compare the pairwise differences and got the following results.

```
Tukey HSD Post-hoc Test...
Regular vs Peanut: Diff=0.0136, 95%CI=-0.0067 to 0.0339, p=0.2503
Regular vs Peanut Butter: Diff=0.0123, 95%CI=-0.0080 to 0.0326, p=0.3181
Peanut vs Peanut Butter: Diff=-0.0013, 95%CI=-0.0216 to 0.0191, p=0.9878
```

Do you notice any pairwise differences that are unusual? Or does it seem that all of the pairwise differences can be explained as sampling variability?