Week 7 Day 2: Hypothesis Testing, Decision Errors, & Multiple Comparisons

General Social Survey – Past Analysis

We've worked with data from the General Social Survey (GSS) for a few different analyses in this class. Specifically, we used these data to test if there was evidence that the true mean number of hours worked in a week for working Americans was different than 40.

Today, we are going to look at a different version of this question. We will investigate if there is evidence that the average number of hours worked per week differs for at least one marital status groups.

Inference for Many Means (ANOVA)

We've seen how to do inference for a single mean and for a difference of means. Now, we'll study how to do inference on **many** means. To do this, we'll use a procedure called ANOVA (**AN**alysis **O**f **VA**riance). With an ANOVA, we're comparing the variability *within* groups (MSE) to the variability *between* groups (MSG).

If we believe that the mean of at least one group is different from the others, ideally in a visualization we'd like to see:

- large differences in the means **between** the groups
- small amounts of variability **within** each group

1. Sketch an example of three boxplots that exhibit the characteristics above.

Hypotheses in an ANOVA

In an ANOVA, we only do hypothesis testing (no confidence intervals until after ANOVA), and the hypotheses are always the same:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_A : \text{At least one of the means is different}$$

Let's consider the variables `marital_status` and `number_of_hours_worked_last_week`. To know how many groups there are in the `marital_status` variable, we can use our friend `favstats()`!

```
favstats(number_of_hours_worked_last_week ~ marital_status,
         data = GSS)
```

```
##   marital_status min   Q1 median   Q3 max     mean       sd   n missing
## 1       Divorced   1 35.0     40 50.0  89 42.04184 14.74340 239     164
## 2        Married   1 37.0     40 50.0  89 41.48220 14.63485 618     380
## 3  Never married   1 35.0     40 48.0  89 40.96744 13.71926 430     240
## 4      Separated  15 37.5     40 54.5  72 44.20930 14.24205  43      32
## 5        Widowed   2 25.0     40 43.5  86 35.47059 16.68455  51     149
```

1. How many groups do we have in our ANOVA?

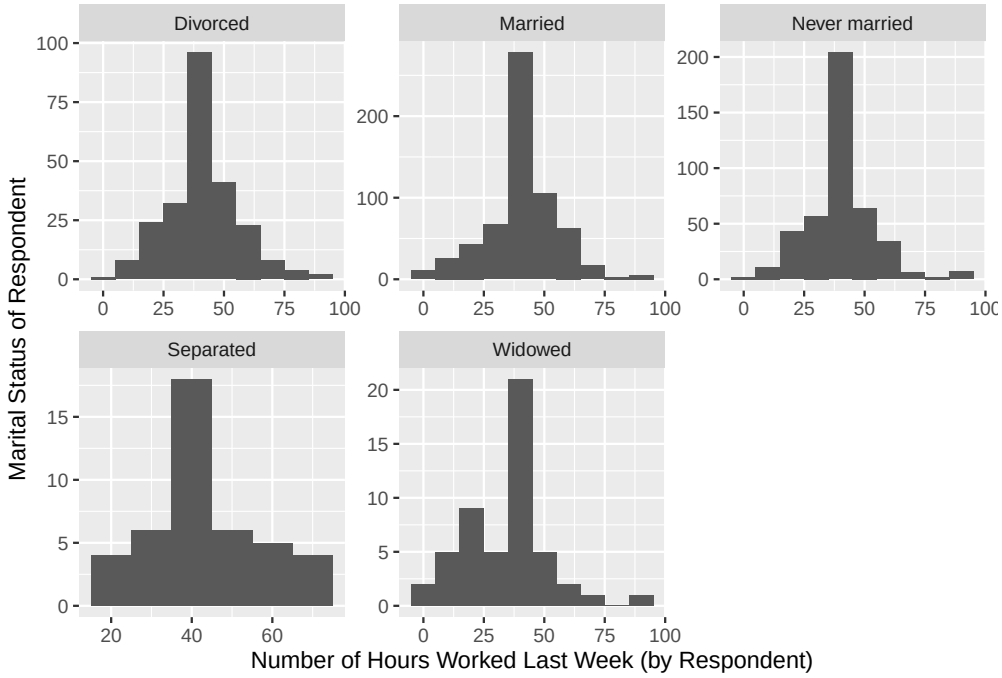
2. Rewrite the null an alternative hypotheses above to reflect the number of groups in our analysis. *It would be nice to know what groups the means correspond with!*

Visualizing an ANOVA

By plotting the data **before** we do a hypothesis test, we get a better understanding of *why* we got a low / medium / high p-value!

Here are faceted histograms visualizing the distribution of hours worked last week across the different marital statuses.

```
ggplot(data = GSS,
       mapping = aes(x = number_of_hours_worked_last_week)
       ) +
  geom_histogram(binwidth = 10) +
  facet_wrap(~ marital_status, scales = "free") +
  labs(x = "Number of Hours Worked Last Week (by Respondent)",
       y = "Marital Status of Respondent")
```



3. How different are the centers of these groups from each other?

4. How different are the spreads of these groups from each other?

5. Do any of the marital groups stand out as **really** different from the others?

Conditions of an ANOVA

Like every statistical analysis we've done in this class, with an ANOVA you have two different types of methods, (1) a simulation-based method or (2) a mathematical theory-based method. The book describes both options, but today we are going to focus on **theory-based** methods.

In an ANOVA there are two conditions that we need to evaluation regardless of which method we use:

- independence of observations within **and** between groups
- equal variance across every group

6. Evaluate if you believe the independence condition is or is not violated. Keep in mind that there are **two** components to this condition you need to discuss!

7. Evaluate if you believe the equal variance condition is or is not violated. Make specific reference to the visualizations and the summary statistics presented on the second page!

Additional Condition for Theory-Based Methods:

As we have seen before, with a theory-based method we have one additional condition:

- nearly normal distributions across every group

8. Why should we use faceted histograms to assess this condition rather than side-by-side boxplots?

9. Evaluate if you believe the normality condition is or is not violated. Make specific reference to the visualization you stated in #8!

Carrying Out an ANOVA in R

Now that we've checked the conditions of an ANOVA, we are ready to perform the analysis! Yesterday, you saw the `aov()` function, which is the tool we use in R to perform an ANOVA.

10. Fill in the necessary components of the code below.

```
aov(_______________________________ ~ ___________________________,
    data = GSS)
```

Alright, if we ran the code you just input, we'd get the following table:

term	df	sumsq	meansq	statistic	p.value
marital_status	4	2296	574	2.752	0.0269
Residuals	1376	287065	208.6	NA	NA

11. What is the mean squares of `marital_status`?

12. What is the mean squares of the errors?

13. How was the `statistic` of 2.752 found? What is the name of that statistic?

14. What is the p-value associated with that statistic?

15. Based on the p-value, at an $\alpha = 0.05$ significance level, what decision would you reach regarding your hypotheses?

16. Based on your decision in #15, what would you conclude regarding the mean hours worked across these marital status groups?

Let's revisit the statistics we first saw. It's entirely possible that in #5 you said that you didn't believe there were "substantial" difference across these groups.

```
##   marital_status min   Q1 median   Q3 max     mean       sd   n missing
## 1       Divorced   1 35.0     40 50.0  89 42.04184 14.74340 239     164
## 2        Married   1 37.0     40 50.0  89 41.48220 14.63485 618     380
## 3  Never married   1 35.0     40 48.0  89 40.96744 13.71926 430     240
## 4      Separated  15 37.5     40 54.5  72 44.20930 14.24205  43      32
## 5        Widowed   2 25.0     40 43.5  86 35.47059 16.68455  51     149
```

17. So, why do you believe we obtained a p-value that suggests at least one of the marital status groups has a different mean number of hours worked?

Hypothesis Testing Errors

In a hypothesis test, there are two competing hypotheses: the null and the alternative. We make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios in a hypothesis test:

	H_0 is True	H_0 is False
Reject H_0	Type I Error	Good Decision!
Fail to Reject H_0	Good Decision!	Type II Error

18. Based on the decision you reached in #15, what type of error could you have made?

19. With an $\alpha = 0.05$, what percent of the time would we expect to make a Type I error?

20. How does α relate to the probability of making a Type II error?

Inference after ANOVA

Because we found a "significant" p-value and concluded that at least one of the means is different. However, an ANOVA **does not** tell us which group(s) is(are) driving the differences.

What we could do is compare all possible combinations of two means. With five groups, that would result in 10 different hypothesis tests for a difference in means (e.g., $\mu_{\text{Divorced}} - \mu_{\text{Married}}$, $\mu_{\text{Divorced}} - \mu_{\text{Widowed}}$, $\mu_{\text{Married}} - \mu_{\text{Never married}}$, etc.).

However, for each hypothesis test we do at an α of 0.05, we risk making a Type I error 5% of the time. In fact, we can make a mathematical equation for the probability of making a Type I Error, based on the number of tests we perform.

$$\text{Probability of Making a Type I Error} = 1 - (0.95)^{\#\text{ of tests}}$$

21. If we do 10 hypothesis tests, what is the probability of us making a Type I Error?

Remedy to Type I Error Inflation

Figure 1: Carlo Emilio Bonferroni (1892 - 1960)

One solution to the problem of multiple comparisons is called the Bonferroni correction. Essentially, you take your α threshold and divide it by the number of tests you are going to perform. This is referred to as α^*.

You then use this α^* value as the new threshold value for **every** pairwise comparison. If a comparison's p-value is less than α^*, then you reject H_0. If a comparison's p-value is greater than α^*, then you fail to reject H_0

22. If our original α was 0.05, what value should we use for α^*?

Post-Hoc Comparisons

Below is a table of all 10 of the hypothesis tests we could do when comparing the means of two groups.

23. Using the α^* you found in #22, circle the hypothesis tests whose p-values are less than α^*.

Group 1	Group 2	p-value
Married	Divorced	0.6111
Never married	Divorced	0.3567
Never married	Married	0.5705
Separated	Divorced	0.3651
Separated	Married	0.2315
Separated	Never married	0.1608
Widowed	Divorced	0.003236
Widowed	Married	0.004344
Widowed	Never married	0.01028
Widowed	Separated	0.003532