

جامعة نيويورك أبوظبي



PSYCH-UH 1004Q: Statistics for Psychology

Class 19: Calculating an ANOVA (one-way, independent samples)

Prof. Jon Sprouse
Psychology

Quick review of the logic of ANOVA

The logic of ANOVA

We run an ANOVA when we have 3 or more conditions. It is very simple — we estimate the variance of the population using two distinct methods:

$$F = \frac{MS_B}{MS_W} = \frac{n \frac{\sum (\bar{x}_i - \bar{x}_G)^2}{k-1}}{\frac{\sum (n_i - 1) s_i^2}{n_{\text{total}} - k}} = \frac{\text{variance from condition means}}{\text{variance from raw scores}}$$

MS_B (the numerator) is a good estimate of the variance of the population when **H₀ is true**, and a bad estimate when **H₀ is false**.

MS_W (the denominator) is always a good estimate of the variance of the population.

This means that F will be 1 when **H₀ is true**, and F will be larger than 1 when **H₀ is false**.

So we can use F as a test statistic just like we did t . We just need to find the distribution of F , find critical F values, and calculate p -values from F .

The F distribution

Two dfs for F

The calculation of F involves two different formulae for variance. Therefore we have two different degrees of freedom:

$$F = \frac{MS_B}{MS_W} = \frac{n \frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k-1}}{\frac{\sum (n_i-1) s_i^2}{n_{\text{total}}-k}}$$

$df_B = k-1$
 $df_W = n_{\text{total}}-k$

The degrees of freedom are always the number of scores minus the number of parameters that have been estimated. You can also see them in equations!

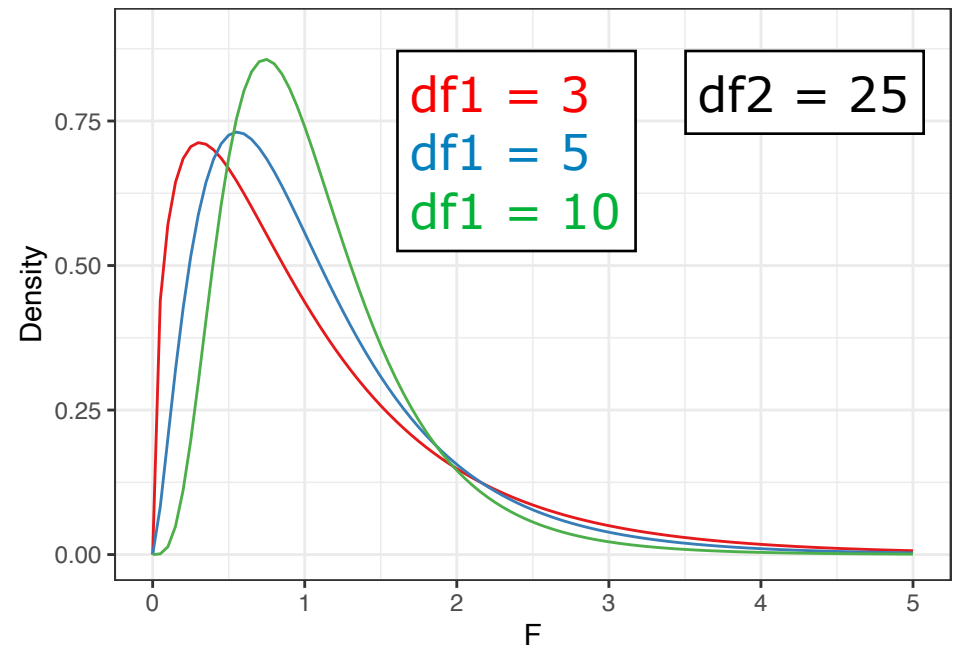
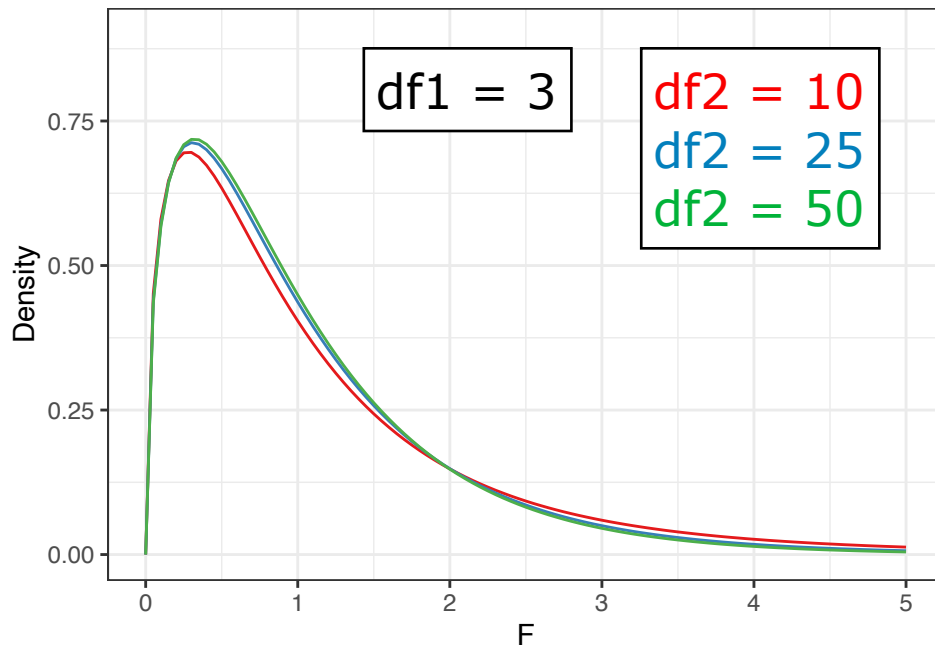
For **MS_B** (the numerator), there are k scores, where k is the number of conditions or groups. Only one mean is estimated, the grand mean of the groups (\bar{x}_G), so the df is $k-1$.

MS_W (the denominator) is just the pooled variance. We know that its degrees of freedom are $(n-1)$ summed for each group in the pool. The number of groups is k , so it is $(n-1) + (n-1) + (n-1)$ up to k times, which is $n_{\text{total}}-k$.

F is a family of distributions using df_B and df_W

It will not surprise you to learn that there is not just one F distribution. F is a family of distributions, just like we saw with t .

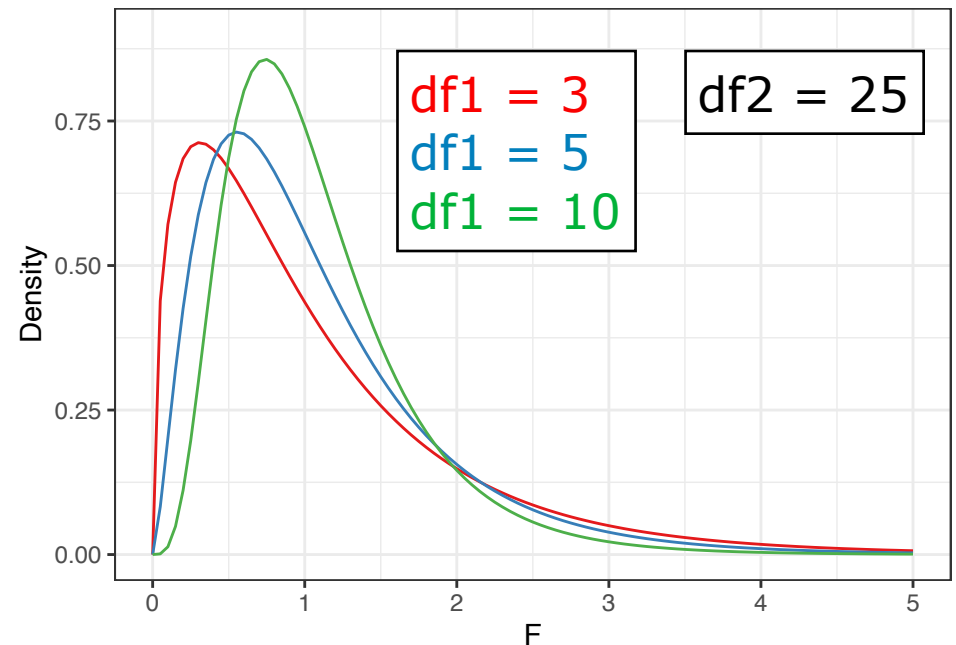
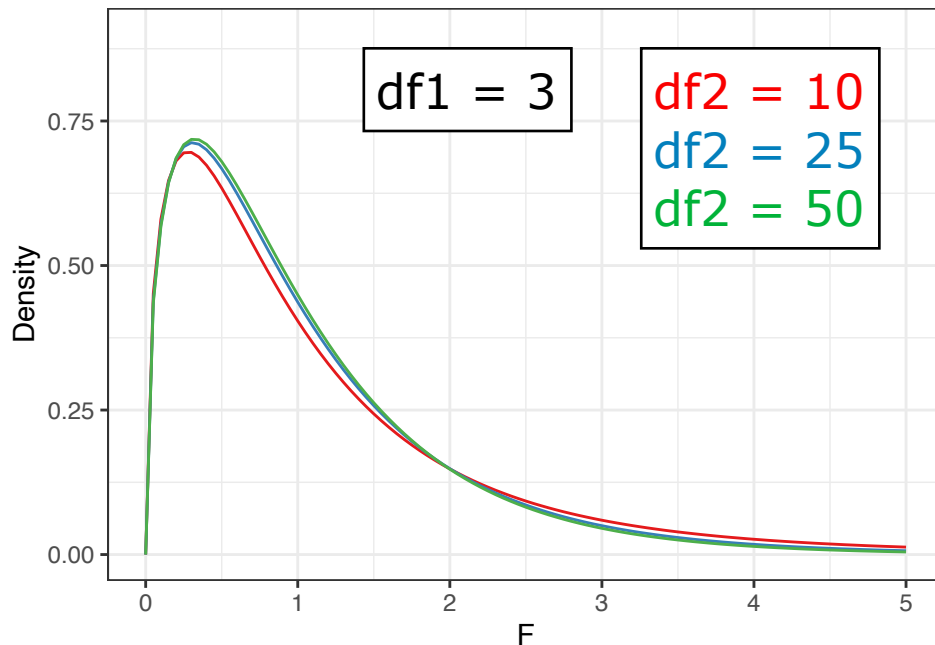
The shape of the F distribution depends on the degrees of freedom in the analysis, again, just like t . In the case of F , there are two degrees of freedom. They are called **df1** and **df2** in R (and by some people when they talk). $df1$ is df_B and $df2$ is df_W .



The F distribution is only positive, and it is skewed for realistic dfs

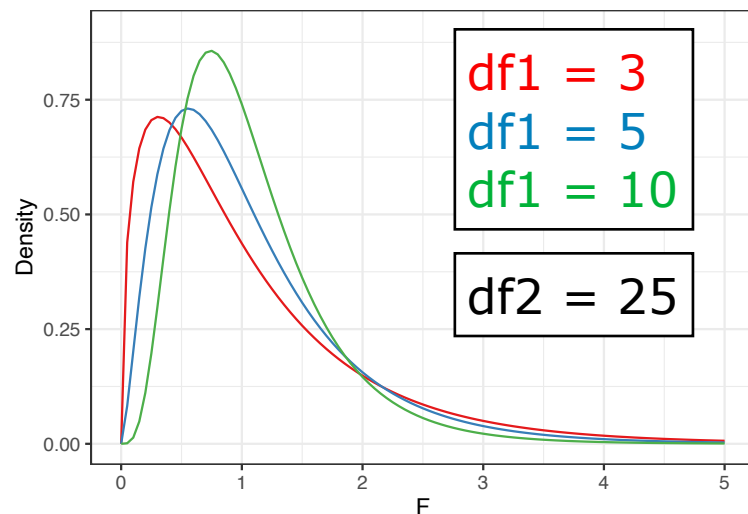
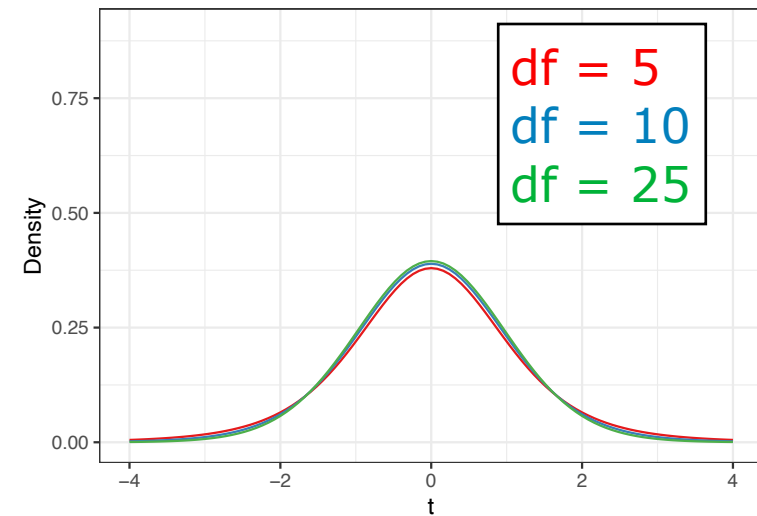
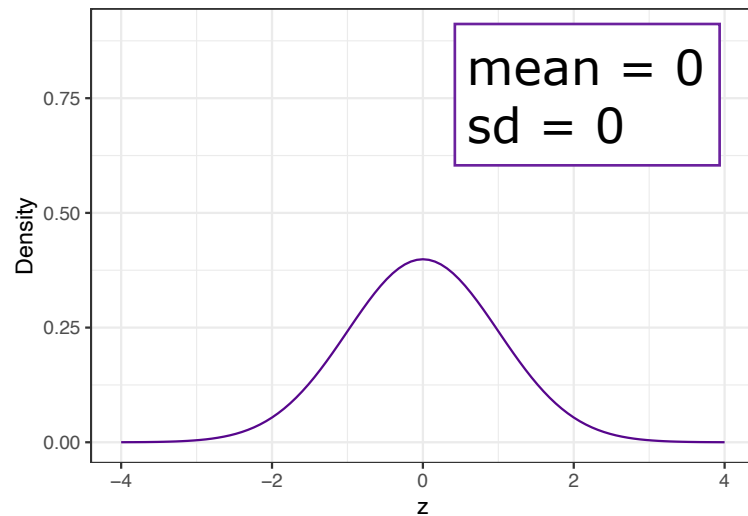
The first thing you will notice is that F can never be negative. This is because variance can never be negative. Since F is just the ratio of two estimates of the variance, it will always be positive.

The second thing you will notice is that the F distribution is not symmetric. It has a positive skew (as expected given that it is bounded to the left by 0). This has no impact on you in your work, it is just something to notice. This is our first asymmetric distribution for a test statistic!



Test statistics can have different distributions

We have now seen 3 test statistics in this class: z , t , and F . They each have different distributions. And they each depend on different parameters.

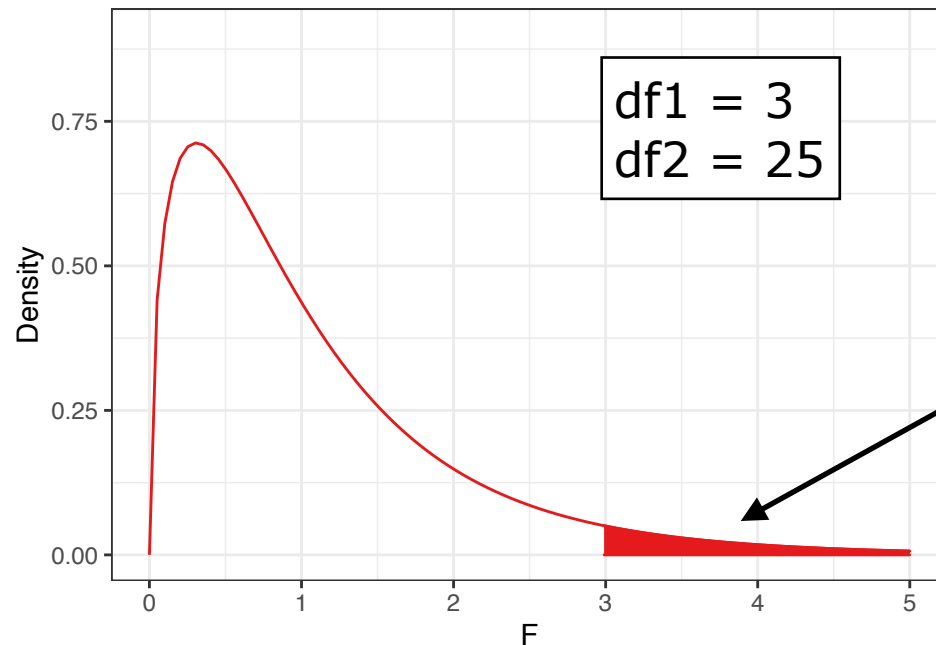


These are the only 3 test statistics that we will see in this course. But there are many others in the field (for other experimental designs).

Identifying a critical region for F

F is just like any other test statistic in null hypothesis testing. Its distribution represents the distribution of possible outcomes for your experiment under the null hypothesis.

Therefore we can identify a critical region that identifies the 5% of results that are the most extreme. This region can be identified by the critical F that cuts 5% (the tail) off from the rest of the distribution.

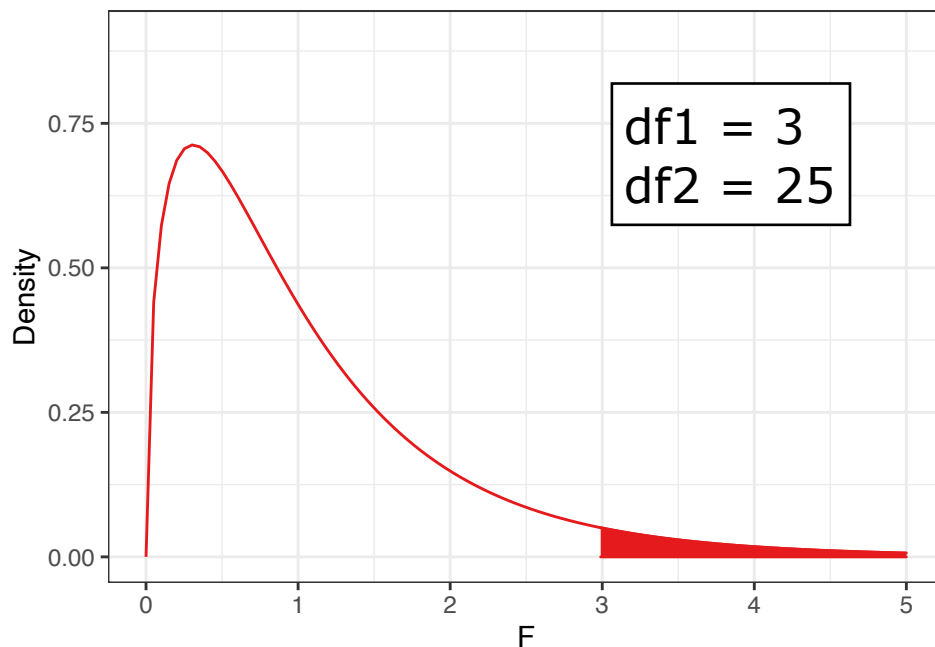


The critical F for 5% ($p < .05$) is 2.99 for $df1=3$, and $df2=25$.

The critical region is always in the right tail

When we compare two means, we can have a directional hypothesis. We can ask whether one mean is above (to the right, right-tailed) or below (to the left, left-tailed) the other mean. In other words, there are two possible alternative hypotheses - one to the right of the null distribution or one to the left.

But when we look at the variance, we aren't comparing values like that. We are just asking "Is this a good estimate of the variance, or is it a bad (=large) estimate"? There is no directionality to the question. There is only one alternative hypothesis, and it is to the right of the null distribution (larger variance).

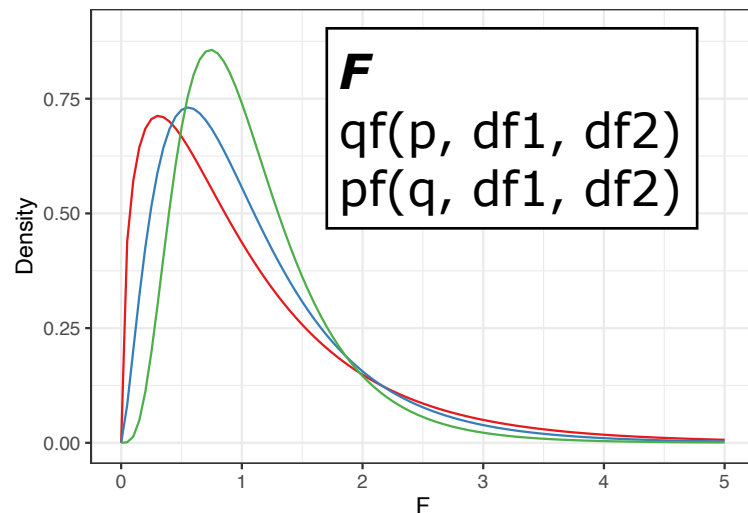
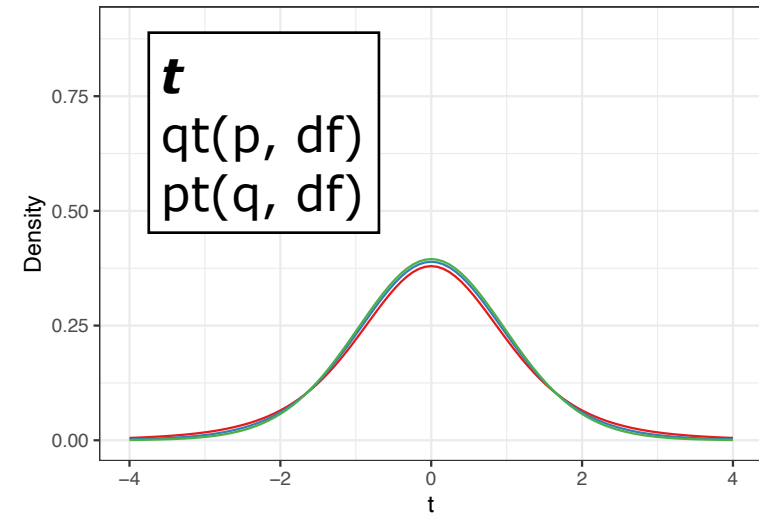
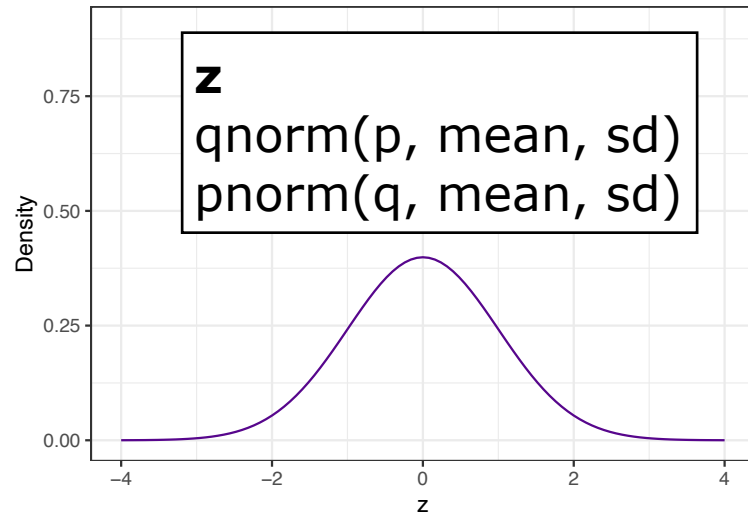


The critical region for F is always in the right tail because there is no directionality in how variance works for our question:

The question is: Do these conditions come from populations with the same mean, or not? No only goes one way - larger.

Using R to find critical Fs and p -values

Once again, we use the same family of R functions to find critical values and p -values for test statistics. You will notice a pattern in the function names!



```
Console Terminal x Jobs x  
~/Desktop/Statistics/jon's R notebooks/ ↗  
> qf(.95, df1=3, df2=25)  
[1] 2.991241  
> pf(2.99, df1=3, df2=25)  
[1] 0.9499374  
> |
```

Let's do it by hand
(This is just calculating variance in two ways!)

This is problem B5 in the book, edited

A psychologist is interested in the relationship between color of food and appetite. To explore this relationship, the researcher bakes small cookies with icing of one of three different colors (red, blue, or green). The researcher offers cookies to participants while they are performing a boring task.

Each participant is run individually under the same conditions, except for the color of the icing on the cookies that are available. Six participants are randomly assigned to each color. The number of cookies consumed by each participant during the 30-minute session is shown in the following table:

red	blue	green
3	2	3
4	0	7
5	4	1
6	6	0
4	4	9
6	1	2

Steps to solve:

1. Calculate the F ratio.
2. (Find the critical F for $\alpha=.05$.)
3. Find the precise p -value for our F .
4. What is our statistical decision?

This is problem B5 in the book, edited

Here are our equations for the ANOVA. Let's try to figure out which quantities we need:

$$F = \frac{MS_B}{MS_W} = \frac{n \frac{\sum (\bar{x}_i - \bar{x}_G)^2}{k-1}}{\frac{\sum (n_i - 1) s_i^2}{n_{\text{total}} - k}}$$

red	blue	green
3	2	3
4	0	7
5	4	1
6	6	0
4	4	9
6	1	2

	red	blue	green
\bar{x}	4.67	2.83	3.67
s	1.21	2.23	3.56
s^2	1.46	4.97	12.67
n	6	6	6
\bar{x}_G	3.72		
n_{total}	18		

This is problem B5 in the book, edited

Next we work on each component separately:

$$MS_B = n \frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k-1} = 6 \frac{(4.67 - 3.72)^2 + (2.83 - 3.72)^2 + (3.67 - 3.72)^2}{3-1}$$

$$MS_B = 5.09$$

	red	blue	green
3	2	3	
4	0	7	
5	4	1	
6	6	0	
4	4	9	
6	1	2	

	red	blue	green
\bar{x}	4.67	2.83	3.67
s	1.21	2.23	3.56
s ²	1.46	4.97	12.67
n	6	6	6
\bar{x}_G	3.72		
n _{total}	18		

This is problem B5 in the book, edited

Next we work on each component separately:

$$MS_w = \frac{\sum (n_i - 1) s_i^2}{n_{\text{total}} - k} = \frac{(6 - 1)1.46 + (6 - 1)4.97 + (6 - 1)12.67}{18 - 3}$$

$$MS_w = 6.37$$

	red	blue	green
3	2	3	
4	0	7	
5	4	1	
6	6	0	
4	4	9	
6	1	2	

	red	blue	green
\bar{x}	4.67	2.83	3.67
s	1.21	2.23	3.56
s^2	1.46	4.97	12.67
n	6	6	6
\bar{x}_G	3.72		
n_{total}	18		

This is problem B5 in the book, edited

Now we can assemble the F ratio:

$$F = \frac{MS_B}{MS_W} = \frac{5.09}{6.37} = 0.80$$

red	blue	green
3	2	3
4	0	7
5	4	1
6	6	0
4	4	9
6	1	2

	red	blue	green
\bar{x}	4.67	2.83	3.67
s	1.21	2.23	3.56
s^2	1.46	4.97	12.67
n	6	6	6
\bar{x}_G	3.72		
n_{total}	18		

This is problem B5 in the book, edited

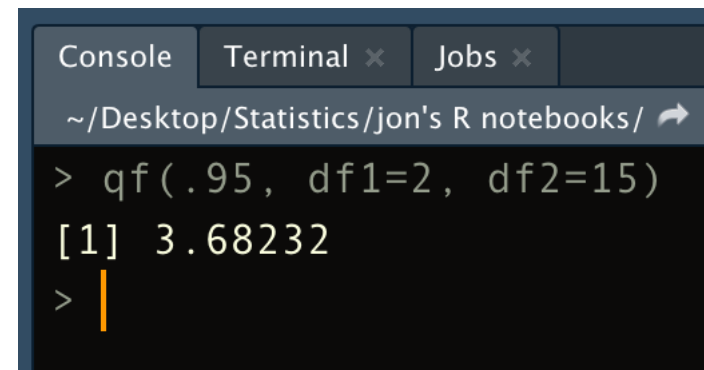
Next, we need to find the critical F for an alpha of .05. Remember, the F distribution depends on two dfs: df_B (or df_1) and df_W (or df_2).

$$df_B = k-1 = 3-1 = 2$$

$$df_W = n_{\text{total}}-k = 18-3 = 15$$

There is a table in your book to look up critical F statistics. But I prefer to use R for this. I am more likely to have a computer in front of me (or the internet) than a textbook.

I use the `qf()` function. This function takes a quantile, like .95, and returns the value at that quantile. Remember that we need to enter both dfs in the correct arguments!



```
Console Terminal x Jobs x  
~/Desktop/Statistics/jon's R notebooks/ ↵  
> qf(.95, df1=2, df2=15)  
[1] 3.68232  
> |
```

This is problem B5 in the book, edited

Next, we can find our precise p -value for the F value that we obtained. Again, remember that there are two dfs that we need to know:

$$df_B = k - 1 = 3 - 1 = 2$$

$$df_W = n_{\text{total}} - k = 18 - 3 = 15$$

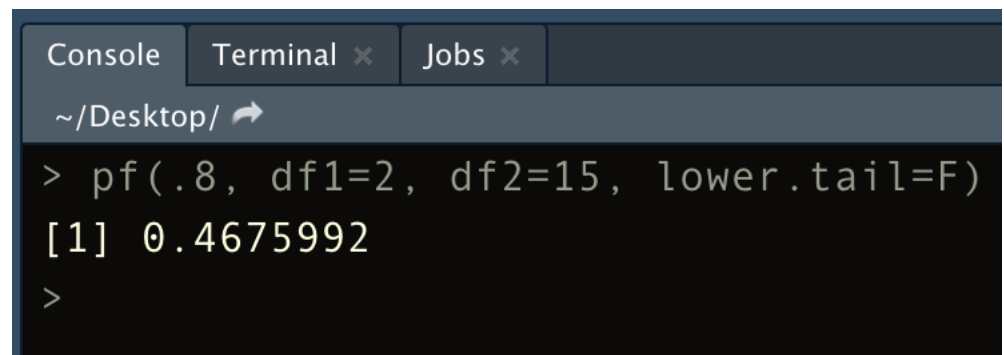
We can write our F in a way that makes the two dfs clear:

$$F(2, 15) = 0.80$$

This is just like we did with t s, except there are two dfs.

To find the precise p -value, we use the `pf()` function. This function takes an F , and two dfs, and returns the p -value for that F .

$$p = .47$$



```
Console Terminal x Jobs x  
~/Desktop/ ↵  
> pf(.8, df1=2, df2=15, lower.tail=F)  
[1] 0.4675992  
>
```

We use `lower.tail=F` to get the upper tail, which is where significant F s are.

This is problem B5 in the book, edited

Finally, we make a statistical decision.

$$F(2,15) = 0.80, p = .47$$

(critical $F = 3.68$)

So, we **fail to reject the null hypothesis** that the sample means come from identical populations.

red	blue	green
3	2	3
4	0	7
5	4	1
6	6	0
4	4	9
6	1	2

	red	blue	green
\bar{x}	4.67	2.83	3.67
s	1.21	2.23	3.56
s^2	1.46	4.97	12.67
n	6	6	6
\bar{x}_G	3.72		
n_{total}	18		

The ANOVA table

The ANOVA table

As you have seen, there are a number of pieces of information that go into calculating an ANOVA. Because of this, it is fairly common to see those quantities reported in a table. This is called an ANOVA table:

	df	SS	MS	<i>F</i>	<i>p</i>
between	2	10.18	5.09	0.80	0.47
within	15	95.5	6.37		

If you write a journal article about your results, you will often put a table like this either directly in the text or in an appendix, so that people can see all of the values in the ANOVA. This is also the output that R will give you, so it is worth taking the time to make sure you see how each of the components maps to your equations.

The ANOVA table

Here I will color each part of the F equation to show how it maps to the table.

$$F = \frac{MS_B}{MS_W} = \frac{n \frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k-1}}{\frac{\sum(n_i-1) s_i^2}{n_{total}-k}}$$

	df	SS	MS	F	p
between	2	10.18	5.09	0.8	0.47
within	15	95.5	6.37		

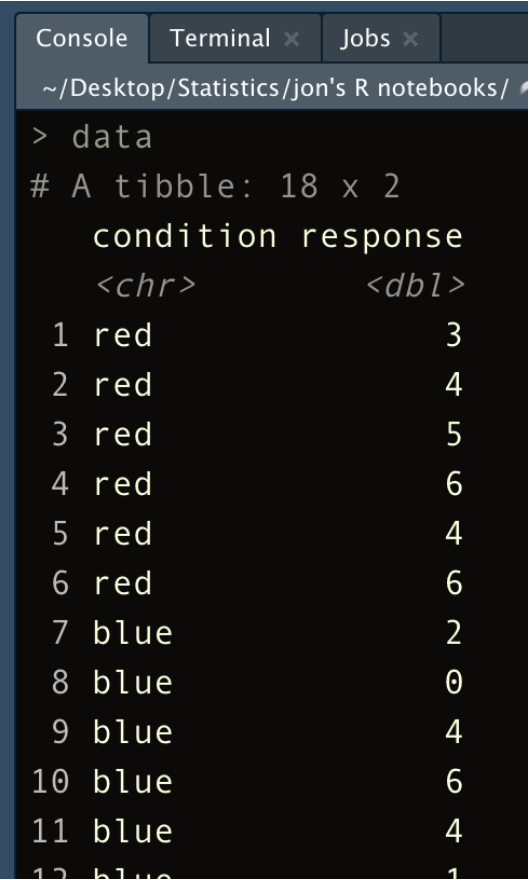
Let's use R - the aov() function

We can also use the `aov()` function in R

It is important to do a few ANOVAs by hand to see how they work. They are simple - they are just two different ways of calculating variance.

But after you understand it, you will primarily use R to calculate the ANOVA for your research. We do this with the `aov()` function and the `summary()` function.

The first thing to note is that the `aov()` function requires you to have your data in **long format**. So, you have to become comfortable getting your data into R. If you aren't comfortable with that yet, please let us know. You will need to do both by-hand calculations and R calculations in your homework and exam in this unit.



```
Console Terminal x Jobs x
~/Desktop/Statistics/jon's R notebooks/
> data
# A tibble: 18 x 2
  condition response
  <chr>      <dbl>
1 red        3
2 red        4
3 red        5
4 red        6
5 red        4
6 red        6
7 blue       2
8 blue       0
9 blue       4
10 blue      6
11 blue      4
12 blue      1
```

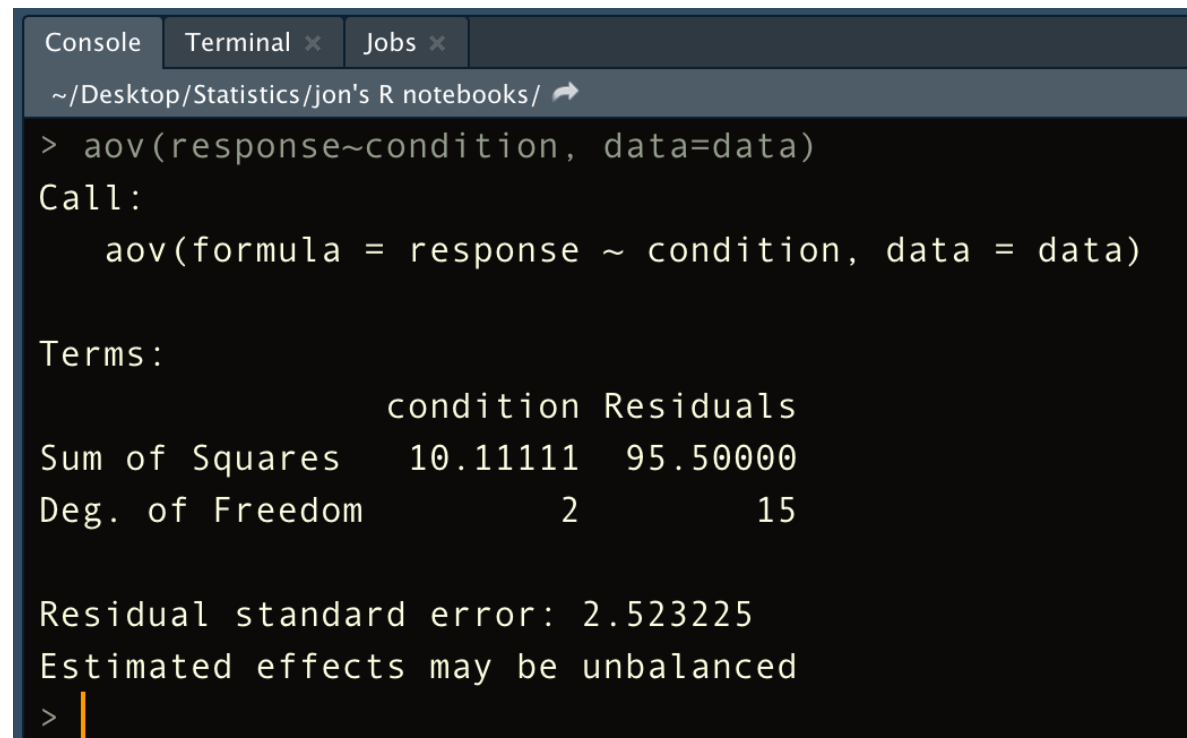
We can also use the `aov()` function in R

The `aov()` function uses formula notation that is identical to the formula notation that we saw for `lm()` in the linear regression chapter.

The formula notation is very common in R. You will see it over and over again. So it is important to become familiar with it.

For the formula, you place the dependent variable to the left of the tilde, and independent variable on the right. You also tell it the name of your data set.

Just like we saw with the `lm()` function, if we run the `aov()` function it simply spits out the SS and df components of MS_B and MS_W . The MS_B is named after the independent variable; MS_W is named “residuals”.



```
Console Terminal x Jobs x
~/Desktop/Statistics/jon's R notebooks/ ↵
> aov(response~condition, data=data)
Call:
  aov(formula = response ~ condition, data = data)

Terms:
              condition Residuals
Sum of Squares   10.11111  95.50000
Deg. of Freedom         2      15

Residual standard error: 2.523225
Estimated effects may be unbalanced
> |
```

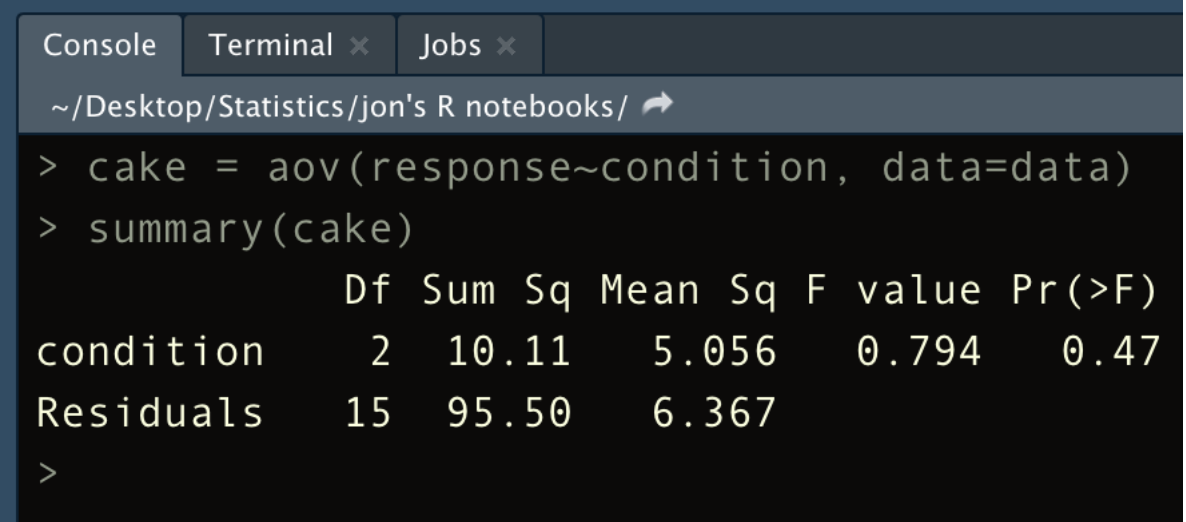
We can also use the `aov()` function in R

To get a hypothesis test out of the `aov()` function, we need to save the model built by the `aov()` function, and then use the `summary()` function to run the hypothesis test.

Again, this is exactly the procedure that we followed with the `lm()` function. This is not an accident. R procedures are typically standardized. (And, we will see later that an ANOVA is just a linear model!)

Notice that when we save the `aov()` model and run `summary()` on it, R gives us an ANOVA table!

The only change from our by-hand table is that MS_B is named after the independent variable; MS_W is named "residuals". This is how R labels ANOVAs.



```
Console Terminal x Jobs x
~/Desktop/Statistics/jon's R notebooks/
> cake = aov(response~condition, data=data)
> summary(cake)
              Df Sum Sq Mean Sq F value Pr(>F)
condition      2  10.11   5.056   0.794   0.47
Residuals     15  95.50   6.367
>
```

(The precise values here are a little different from ours because we were rounding, and R does not round until the very end.)

R's ANOVA table and our by-hand table

```
Console Terminal x Jobs x  
~/Desktop/Statistics/jon's R notebooks/ ↩  
> cake = aov(response~condition, data=data)  
> summary(cake)  
              Df Sum Sq Mean Sq F value Pr(>F)  
condition     2  10.11   5.056   0.794   0.47  
Residuals    15  95.50   6.367  
>
```

	df	SS	MS	<i>F</i>	<i>p</i>
between	2	10.18	5.09	0.8	0.47
within	15	95.5	6.37		

Is ANOVA a completely new test?
(No! It is a generalization of t -tests!)

F is related to t : $t^2 = F$

ANOVA can seem like a completely different test than a t -test. After all, it is about variances, not means. And uses a different statistic, F , rather than t .

But Fisher created ANOVAs as a generalization of Gosset's t -test.

His insight was that even though the t -test looks at the differences between means, if we square the equation, the numerator becomes a measure of variance, and the denominator becomes a measure of variance. I am serious!

Your book walks you through this at the very beginning of the chapter. I like to show it at the end.

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{n}}} \xrightarrow{\text{square}} t^2 = \frac{(\bar{x}_1 - \bar{x}_2)^2}{\frac{s_p^2}{n} + \frac{s_p^2}{n}} = \frac{(\bar{x}_1 - \bar{x}_2)^2}{\frac{2s_p^2}{n}} = \frac{n(\bar{x}_1 - \bar{x}_2)^2}{2s_p^2}$$

F is related to t : $t^2 = F$

Now we just need to see that the numerator of t^2 is equal to the numerator of F , and the denominator of t^2 is equal to the denominator of F .

$$t^2 = \frac{\frac{n(\bar{x}_1 - \bar{x}_2)^2}{2}}{s_p^2} \qquad F = \frac{n \frac{\sum (\bar{x}_i - \bar{x}_G)^2}{k-1}}{\frac{\sum (n_i - 1) s_i^2}{n_{\text{total}} - k}}$$

I am not going to work through this, because it involves some flashbacks to FOIL from algebra class. But I think you can see that these are going to be equivalent.

The numerators will be equivalent because the full difference can be thought of as 1, squared, divided by 2 is .5 The difference with the grand mean can be thought of 1/2 for each mean (grand means are the middle!), squared is 1/4, summed together is 1/2!

And the denominator just is a pooled variance. We already know that. So the two denominators are definitely equal.


We can also prove it by running both tests

We can run an independent samples t -test and an ANOVA with two groups, and see that $t^2=F$. To do this, I'll create some fake data for two conditions with different means:

```
placebo=round(rnorm(10, mean=3, sd=.75), 1)
```

```
medicine=round(rnorm(10, mean=5, sd=.75), 1)
```

```
data = tibble(group = rep(c("placebo", "medicine"),  
each=10), wellbeing = c(placebo, medicine))
```

$$4.58^2 = 21$$


t.test() → $t=4.58$

```
Console Terminal x Jobs x
~/Desktop/Statistics/jon's R notebooks/
> t.test(wellbeing~group, data=data2, var.equal=T)

Two Sample t-test

data: wellbeing by group
t = 4.5828, df = 18, p-value = 0.0002308
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.028969 2.771031
sample estimates:
mean in group medicine mean in group placebo
          5.15           3.25
```

aov() → $F=21$

```
Console Terminal x Jobs x
~/Desktop/Statistics/jon's R notebooks/
> summary(aov(wellbeing~group, data=data2))

          Df Sum Sq Mean Sq F value    Pr(>F)
group      1  18.05  18.050      21 0.000231 ***
Residuals 18  15.47   0.859
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
>
```