Comparing Means

Recap

Normal distributions all have well-characterized properties

- AUC = 1
- ~68% fall within 1 σ, ~95% within 2 σ, and ~99.7% within 3 σ

The standard normal distribution is a particular type of normal distribution

- distribution of z-scores
- $\mu = 0$
- $\sigma = 1$

Using the standard normal & these cool properties, we can make probability statements

- What is the probability of getting a z-score or more extreme?

Sampling Distributions are distributions of statistics

- They also happen to mostly be normally distributed...let's use this!

Example

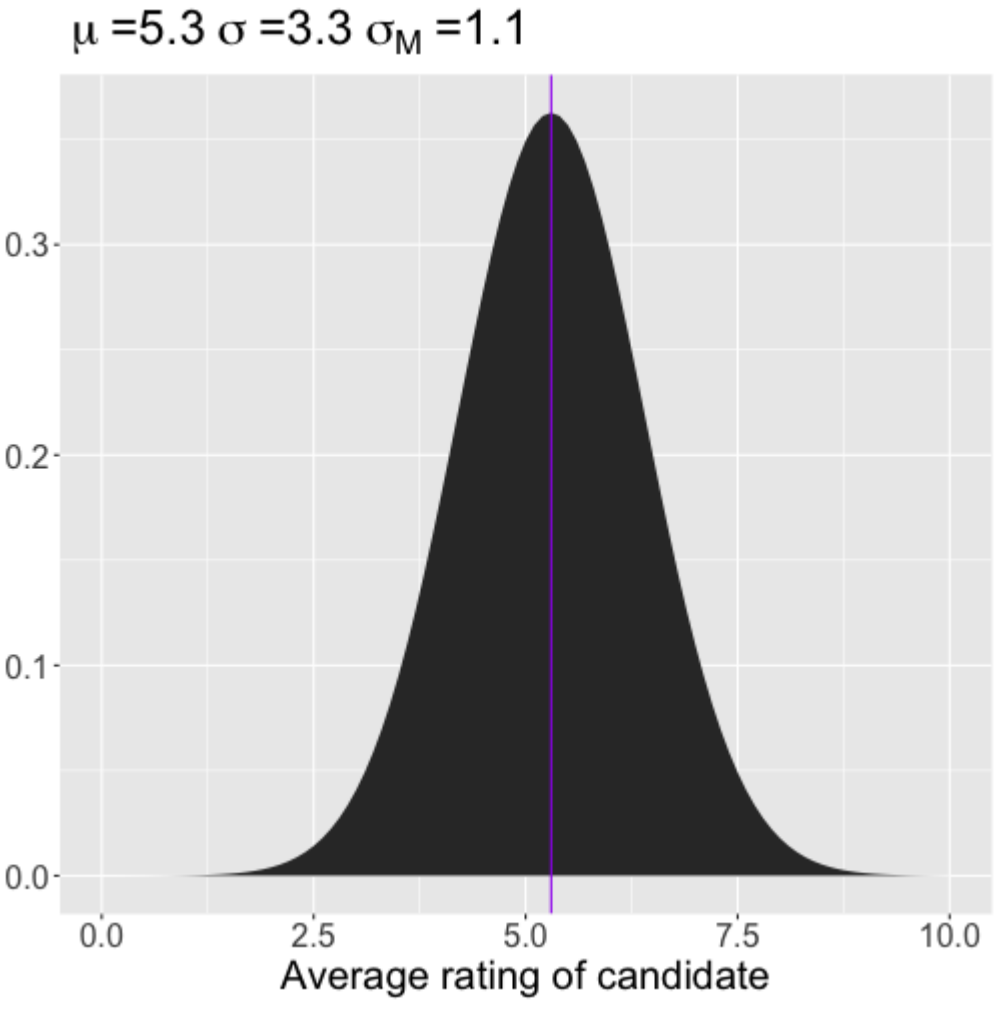
University X has been around for 150 years, and so has 150 years worth of ratings of male applicants. You pay an undergrad dig through all the old university files and calculate the average rating of male applicants (5.3 out of 10) and also the standard deviation of those ratings (3.3).

You then collect the ratings of 9 female applicants in 2018 and calculate their average rating (2.9) and also the standard deviation of their rating (3.1)

How do you generate the sampling distribution around the null?

The mean of the sampling distribution = the mean of the null hypothesis

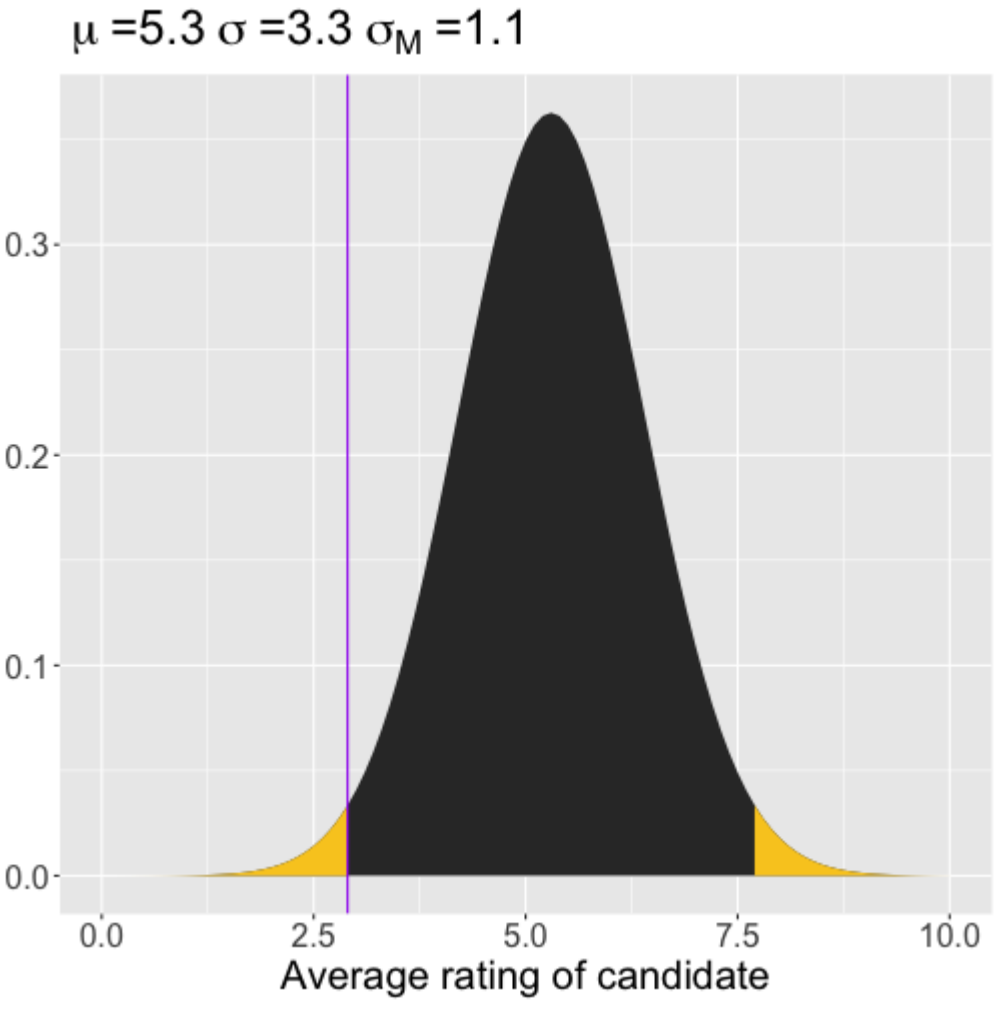
The standard deviation of the sampling distribution:



The mean of the sampling distribution = the mean of the null hypothesis

The standard deviation of the sampling distribution:

$$SEM = \frac{\sigma}{\sqrt{N}}$$



Example Cont...

All well and good.

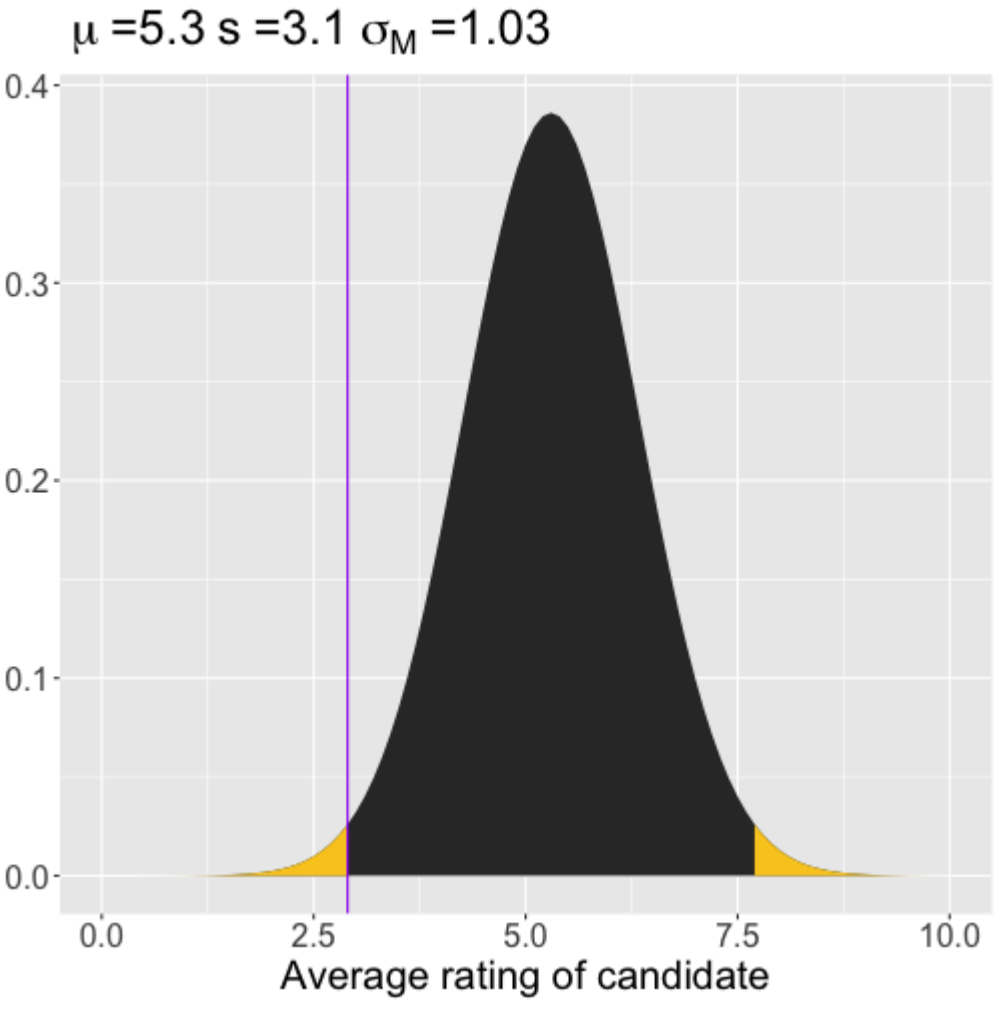
But rarely will you have access to all the data in your population, so you won't be able to calculate the population standard deviation. What ever will you do?

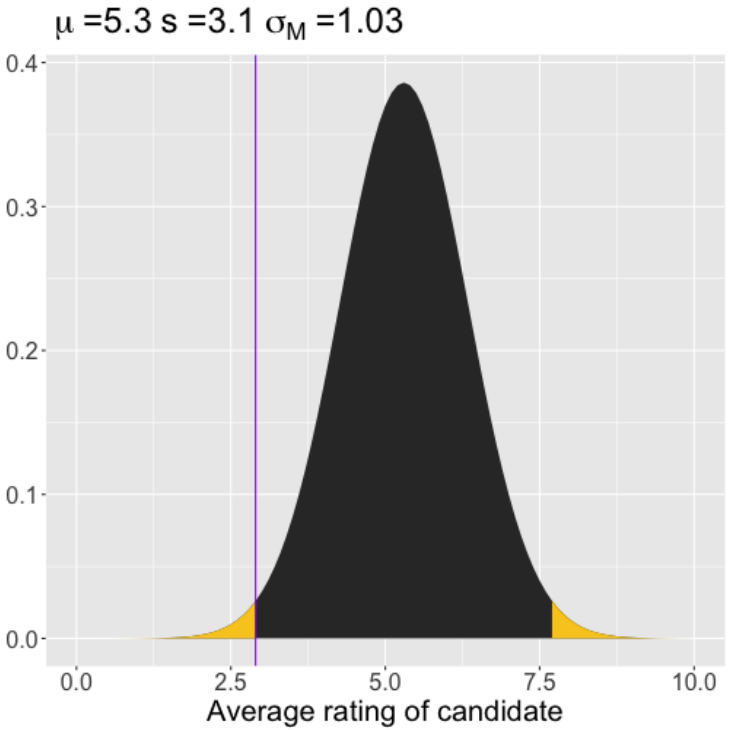
$$SEM = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{s}{\sqrt{N}}$$

So long as your estimate of the standard deviation is already corrected for bias (you've divided by $N - 1$), then you can swap in your sample SD.

If you didn't know the population (male's) standard deviation, you would use the sample of females to estimate the population standard deviation.

$$SEM = \frac{\hat{\sigma}}{\sqrt{N}}$$





We have a normal distribution for which we know the mean (M), the standard deviation (SEM), and a score of interest ($\bar{X}$).

We can use this information to calculate a Z-score; in the context of comparing one mean to a sampling distribution of means, we call this a **Z-statistic**.

$$Z = \frac{\bar{X} - M}{SEM} = \frac{2.9 - 5.3}{1.03} = -2.18$$

$$Z = \frac{\bar{X} - M}{SEM} = \frac{2.9 - 5.3}{1.03} = -2.18$$

And here's where we use the properties of the Standard Normal Distribution to calculate probabilities, specifically the probability of getting a score this far away from μ or more extreme:

```
pnorm(-2.18) + pnorm(2.18, lower.tail = F)
```

```
## [1] 0.02925746
```

```
pnorm(-2.18)*2
```

```
## [1] 0.02925746
```

The probability that the average female applicant's score would be at least 2.32 units away from the average male score is 0.029.

This whole process is called the z-test. It's almost never used IRL, but it's a useful tool in terms of understanding what's happening. We use it when we want to know if our mean is the same or different from a population mean.

NHST Steps

1. Define H_0 and H_1.
2. Choose your α level.
3. Collect data.
4. Define your sampling distribution using your null hypothesis and either the knowns about the population or estimates of the population from your sample.
5. Calculate the probability of your data or more extreme under the null. (To get the probability, you'll need to calculate some kind of standardized score, like a z-statistic.)
6. Compare your probability (p-value) to your α level and decide whether your data are "statistically significant" (reject the null) or not (fail to reject the null).

Who Cares?

Nearly all statistical tests follow this format. The things that are different are which sampling distribution to use (is it normal, a t, a F, a binomial, a poisson etc.)

t-tests

We don't really use z-tests much, but we do use t-tests!

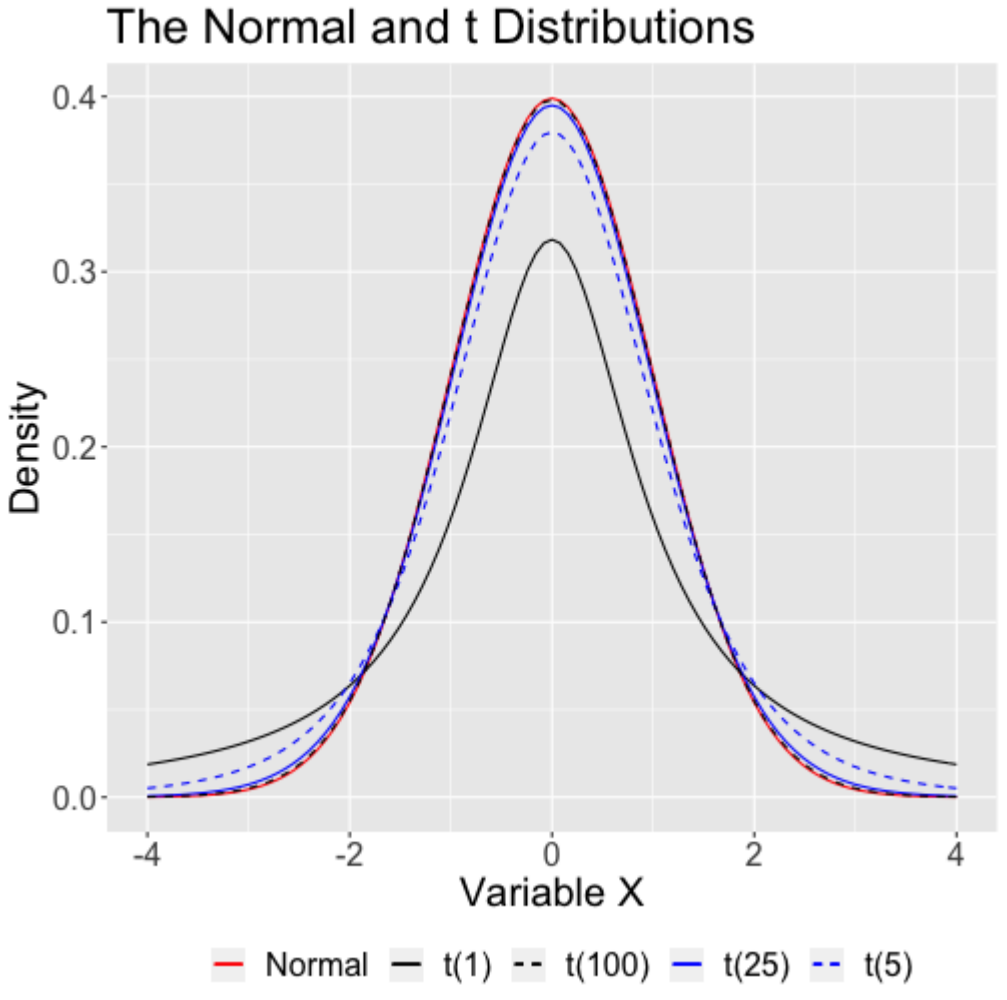
One Sample, Independent Samples (2 kinds), and Paired Samples

- I want you to know...
    - What each of these is testing
    - When we use each one
    - How to interpret the findings
    - How to perform them in R
- I don't care about...
    - Knowing the formula
    - Being able to calculate by hand

The t distribution

The normal distribution assumes we know the population mean and standard deviation. But we don't usually. We only know the *sample* mean and standard deviation, and those have some uncertainty about them.

That uncertainty is reduced with large samples, so that it's "close enough" to the normal. In small samples, the t distribution is better.



t distribution

- The primary difference between the normal distribution and the t distribution is the fatter tails
    - At smaller N, it becomes **harder** to reject the null hypothesis in favor of the alternative
    - The penalty we have to pay for our ignorance about the population
- When we want to do a t-test, we should use the t sampling distribution; not the normal (unless we have a large N, in which case they'll give you the same answers)
- There are different types, so let's work through them

One sample t-test

The question: "Is my sample mean equal to a population mean?" The vast majority of the time, we're asking if it's different from 0.

You've basically already done this! It is the exact same procedure as what we just went through with the z-test. The only differences are:

- We don't know the population standard deviation, so to calculate the SEM, we use our best estimate of sigma ($\hat{\sigma}$), which is our sample standard deviation that has been corrected for bias (s) using that $N - 1$ denominator
- We use the t sampling distribution. If doing it the long way like before, to get the p-values, use the `pt()` function instead of the `pnorm()` function. Or just use the data...

One sample t-test

```
kable(head(iris))
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

One sample t-test

```
# way 1 -- not as recommended unless mu is a number other than 0
t.test(x = iris$Sepal.Length, mu = 0)
```

```
##
##     One Sample t-test
##
## data:  iris$Sepal.Length
## t = 86.425, df = 149, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.709732 5.976934
## sample estimates:
## mean of x
##  5.843333
```

```
# way 2 -- recommended if mu = 0
t.test(Sepal.Length ~ 1, data = iris)
```

```
##
##     One Sample t-test
##
## data:  Sepal.Length
## t = 86.425, df = 149, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.709732 5.976934
## sample estimates:
## mean of x
##  5.843333
```

Accessing Your Results

Ok, you just ran a t-test. And you want to keep that output to use for later, so you store it as an object

```
oneSample <- t.test(Sepal.Length ~ 1, data = iris)
```

If you look in your Environment, you'll notice this is stored as a List object. Lists can be annoying. You can press on the blue arrow to see the different items contained in your list. To actually accsess them we're going to use our old favorite, **indexing**

For instance, to get the p-value, we need to access the 3rd thing in the list

```
oneSample[3]
```

```
## $p.value
## [1] 3.331256e-129
```

See how the name `$p.value` prints out. This makes it hard to actually do math with! It's a *"named number"*. We want to get rid of that name, by going in a little deeper.

Accessing Your Results

The list thing can get obnoxious. We'll revisit it later in the semester. But for now, there's an easier way using `tidyverse`. Specifically, we will use the `broom` package. Even though it's part of the `tidyverse` ecosystem, it does not load when you load `tidyverse`. So you'll need to do that manually.

```
library(broom)
oneSample <- t.test(Sepal.Length ~ 1, data = iris)
tidyOneSample <- tidy(oneSample)

kable(tidyOneSample)
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
5.843333	86.42537	0	149	5.709733	5.976934	One Sample t-test	two.sided

Independent Samples t-test

The question: "Are two means different from one another?" Another way of thinking about this is "is the difference between the means equal to 0?" This is almost always asked in the context of a dichotomous variable.

Two types:

1. Welch's t-test, the default in R. Assumes the variances are unequal
2. Student's t-test. Assumes the variances are equal

```
irisSmall <- iris %>%
  filter(Species != "setosa")

tidyWelch <- tidy(t.test(Sepal.Length ~ Species, data = irisSmall))
tidyStudent <- tidy(t.test(Sepal.Length ~ Species, data = irisSmall, var
```

Independent Samples t-test

```
kable(tidyWelch)
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high
-0.652	5.936	6.588	-5.629165	2e-07	94.02549	-0.8819731	-0.422026

```
kable(tidyStudent)
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high
-0.652	5.936	6.588	-5.629165	2e-07	98	-0.8818516	-0.422148

Paried t-test

In the independent sample t-test, we assume that our data are truly independent. What if they're not? Examples: romantic partners, change in year 1 to year 2 etc. AKA *"repeated measures"*

Let's say we have something like `happiness year 1` and `happiness year 2`. You can't ask if the means are different, because these are very correlated. What we can do is say "is the difference score equal to 0?" That is, "`happiness year 1 - happiness year 2`; is that equal to 0?" This is basically a one-sample t-test but on difference scores now.

Paried t-test

Let's pretend that the the species `versicolor` and `virginica` are actually related (they both start with `v`, right? lol)

```
pairedV1 <- tidy(t.test(Sepal.Length ~ Species, data = irisSmall, paire

kable(pairedV1)
```

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-0.652	-5.275345	3e-06	49	-0.900371	-0.403629	Paired t-test	two.sided

More Means

What if you want to compare more than 2 means? Now you're in ANOVA territory

Oneway ANOVA

You still have a single independent variable, but instead of it being dichotomous, it's trichotomous (or more)

Other ANOVAs

You have more than 1 independent variable, but they are all still factors (not continuous).

ANOVA

To keep things simple, let's say we have a Oneway ANOVA with the original `iris` dataset that has 3 sepcies. The null hypothesis is:

$$\mu_{setosa} = \mu_{versicolor} = \mu_{viriginica}$$

But the alternative hypothesis is that "at least one of these means are different from each other." That could be:

$$\mu_{setosa} \neq \mu_{versicolor} = \mu_{virgininca}$$

$$\mu_{setosa} = \mu_{versicolor} \neq \mu_{virginica}$$

...or any of these combinations.

ANOVA

Instead of using the t or normal distributions, we use the F distribution for ANOVA. The F is a ratio of variances. We take the variance between groups and compare it to the variance within groups (and error).

The idea is that if there is a lot of variance because the means between groups are super different, then the numerator is large while the denominator is small.

If the means aren't that different, then there's not going to be a lot of variance in the numerator. Instead, the variance in the denominator will take over. This would yield a non-significant ANOVA.

Fun things: Your F statistic cannot be negative. There is no such thing as a negative variance, and we're looking at a ratio of variances. 0 is the smallest it gets.

ANOVA

It's the same code from lecture #9. You'll still want to nest `aov()` inside of `summary()`. To store for later, we can still use `tidy()` from the `broom` package. `glance()` from the `broom` package can also help with getting our R^2 (variance explained).

```
summary(aov(Sepal.Length ~ Species, data = iris))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#to tidy it, keep it out of the summary

onewayEx <- tidy(aov(Sepal.Length ~ Species, data = iri
onewayEx
```

```
## # A tibble: 2 x 6
##   term         df sumsq meansq statistic   p.value
##   <chr>     <dbl> <dbl>  <dbl>     <dbl>     <dbl>
## 1 Species       2  63.2 31.6        119.  1.67e-31
## 2 Residuals   147  39.0  0.265       NA  NA
```

Downsides

t-tests and ANOVAs are both just special cases of **regression**. In our regression lecture, we'll talk about this.

Regression is a much more flexible framework for statistical analysis. Generally speaking, t-tests are fine staying as t-tests, but if you ever want to run an ANOVA (especially with 2+ predictors), I **strongly** suggest using regression instead of ANOVA. It's the same thing, but you'll get more for your money with regression, and you can start to include predictors that are continuous and categorical -- no need to choose!

Next time...

- Confidence Intervals
- *p*-values
- Power
- Problems with NHST