جامعة نيويورك أبوظبي

NYU | ABU DHABI

PSYCH-UH 1004Q: Statistics for Psychology

Class 18: The logic of ANOVA

Prof. Jon Sprouse
Psychology

The challenge for experiments with 3 or more conditions

The problem with differences between means

When you have 2 conditions, it is really straightforward to use the difference between means of the two conditions as the effect. You simply subtract them and ask whether that difference is different from your null hypothesis. If your null hypothesis is 0, as it usually is, it might look like this example. *t*-tests simply take variability into account when they formalize this!

condition 1

$\bar{y} = 2$

condition 2

$\bar{y} = 2$

$\bar{y} - \bar{y} = 2 - 2 = 0$

But what happens if you have 3 conditions? Here are three conditions that all have the same mean. If we try to use the difference between means, something odd happens:

condition 1

$\bar{y} = 2$

condition 2

$\bar{y} = 2$

condition 3

$\bar{y} = 2$

$\bar{y} - \bar{y} - \bar{y}$

$2 - 2 - 2 = -2$

These are all the same, but the difference between means is not zero!

The problem with differences between means

It can also be the case that 3 conditions have very different means, but still yield a difference of 0:

condition 1

$\bar{y} = 6$

condition 2

$\bar{y} = 4$

condition 3

$\bar{y} = 2$

$\bar{y} - \bar{y} - \bar{y}$

$6 - 4 - 2 = 0$

These are all different, but the difference is zero!

So what we see is that for 3 (or more) conditions, the simple approach of looking at the difference between means stops serving us well.

What we need is a method for determining if the conditions come from populations with the same parameters or not that works for 3 (or more) conditions. That is what ANOVA is for.

Note: There are other advantages to ANOVA that the book mentions briefly, but we won't be able to see those in detail until next week, so I am setting those aside until then.

We can use variance!

(I know, this seems weird. Bear with me.)

Analysis of Variance

ANOVA stands for Analysis of Variance. The big idea is that there are two methods for estimating the variance of a population from samples. We can compare those two variance estimates to each other.

We already know **one method** for estimating the variance of a population from two or more sample means. We start by calculating the variance for a single sample:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} \quad \text{or} \quad s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

And then we can combine their variances together into the pooled variance (a weighted mean of two variances):

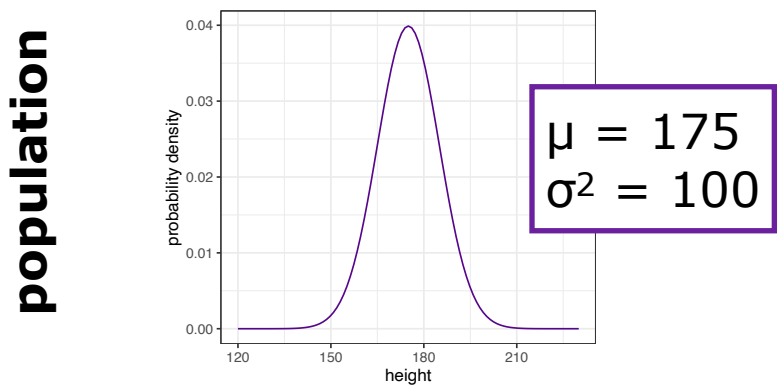
$$s_p^2 = \frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{(n_1-1) + (n_2-1)} \quad \text{or} \quad s_p^2 = \frac{\Sigma\,(n_i-1)\,s_i^2}{n_{total}-k}$$

There is a **second method** to calculate the variance from two (or more samples) - we can use the sample means directly! (Really!) Let's see it now.

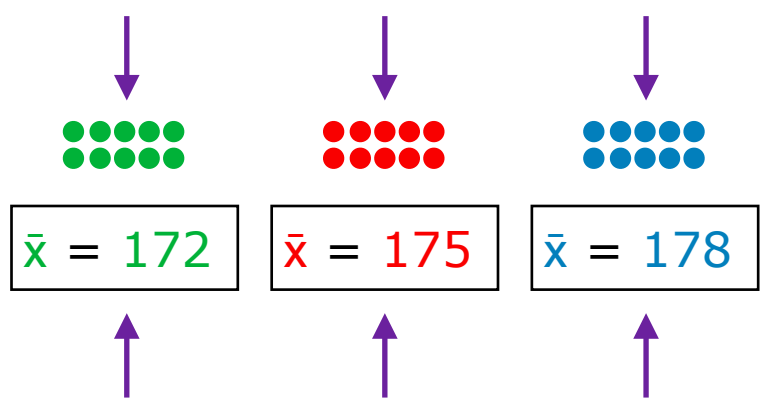
Estimating variance from sample means

To see that we can use sample means to estimate the variance of a population, we need to see the full picture of where sample means come from:

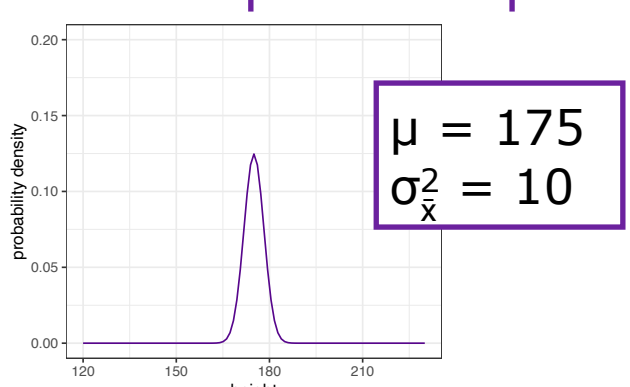
We start with a population. I'll use our old friend height:



We draw 3 samples of 10 from this population. Like an experiment with 3 conditions!



The means of these samples came from the distribution of sample means - the sampling distribution of the mean!

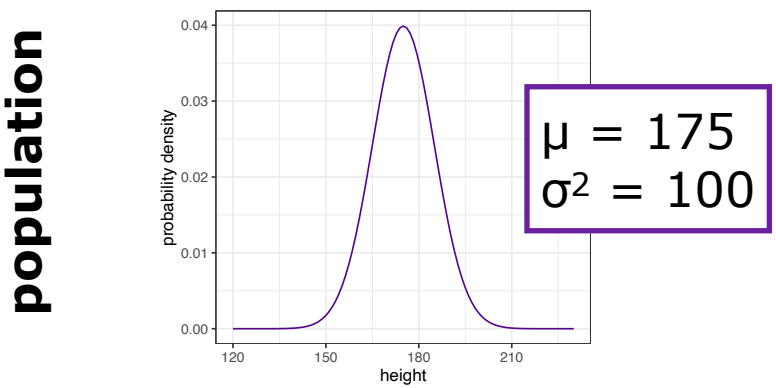


Take a moment to make sure you see how the samples come from the population, but the sample means come from the sampling distribution of the mean. This is critical!

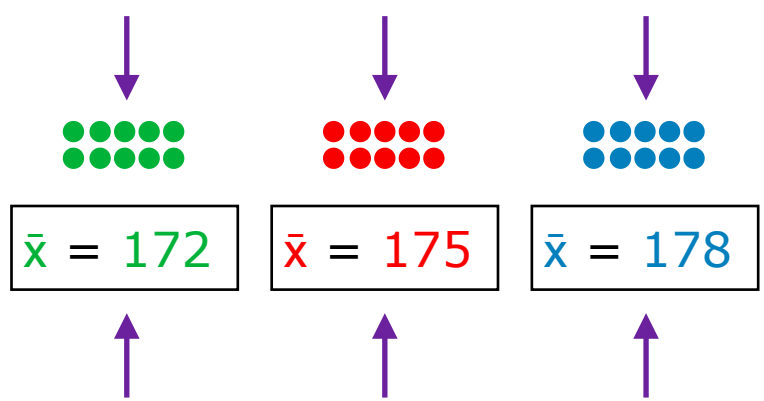
Estimating variance from sample means

To see that we can use sample means to estimate the variance of a population, we need to see the full picture of where sample means come from:

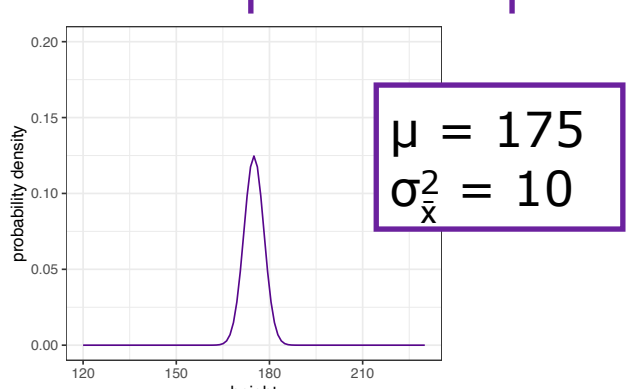
We start with a population. I'll use our old friend height:



We draw 3 samples of 10 from this population. Like an experiment with 3 conditions!



The means of these samples came from the distribution of sample means - the sampling distribution of the mean!



We can calculate a variance score using these 3 numbers:

$$s^2_{\bar{x}} = \frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k\text{-}1}$$

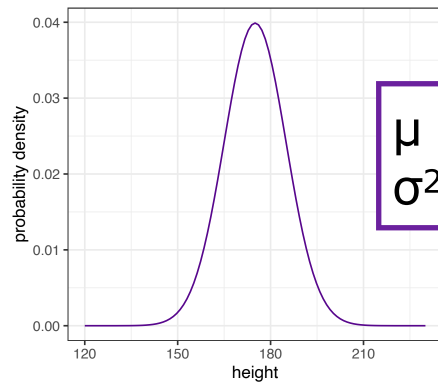
But notice that it is the variance of the sampling distribution of the mean, not the population, because these are means!

Estimating variance from sample means

To see that we can use sample means to estimate the variance of a population, we need to see the full picture of where sample means come from:

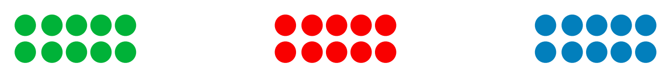
We start with a population. I'll use our old friend height:

population



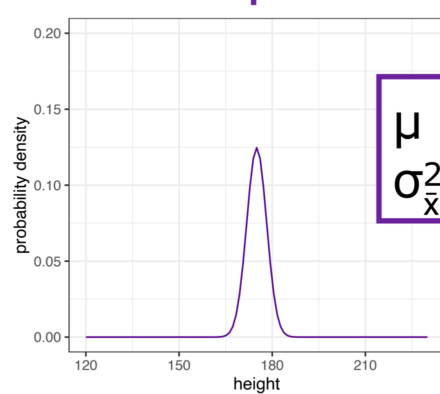
$\mu = 175$
$\sigma^2 = 100$

We draw 3 samples of 10 from this population. Like an experiment with 3 conditions!



$\bar{x} = 172$ | $\bar{x} = 175$ | $\bar{x} = 178$

The means of these samples came from the distribution of sample means - the sampling distribution of the mean!



$\mu = 175$
$\sigma^2_{\bar{x}} = 10$

Finally, remember that $\sigma_{\bar{x}}$ is related to the population σ using this equation:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(But you have to square both sides to make this about variances.)

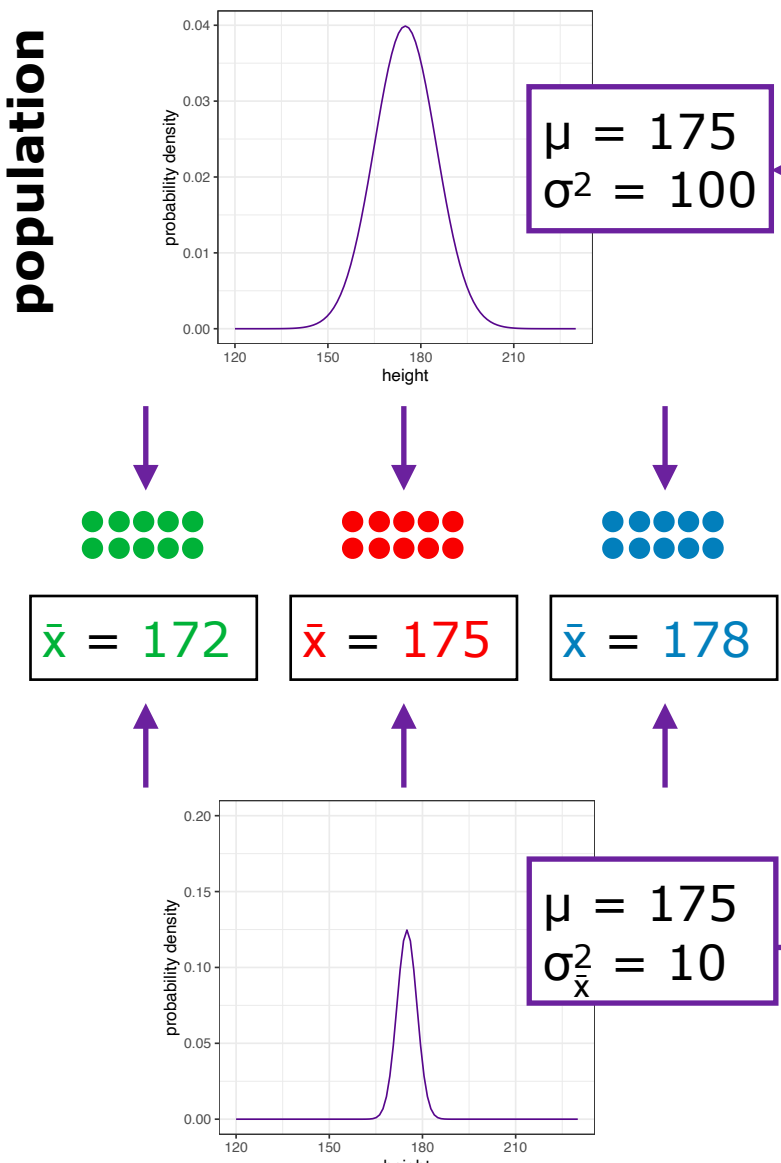
Estimating variance from sample means

To see that we can use sample means to estimate the variance of a population, we need to see the full picture of where sample means come from:

We start with a population. I'll use our old friend height:

We draw 3 samples of 10 from this population. Like an experiment with 3 conditions!

The means of these samples came from the distribution of sample means - the sampling distribution of the mean!



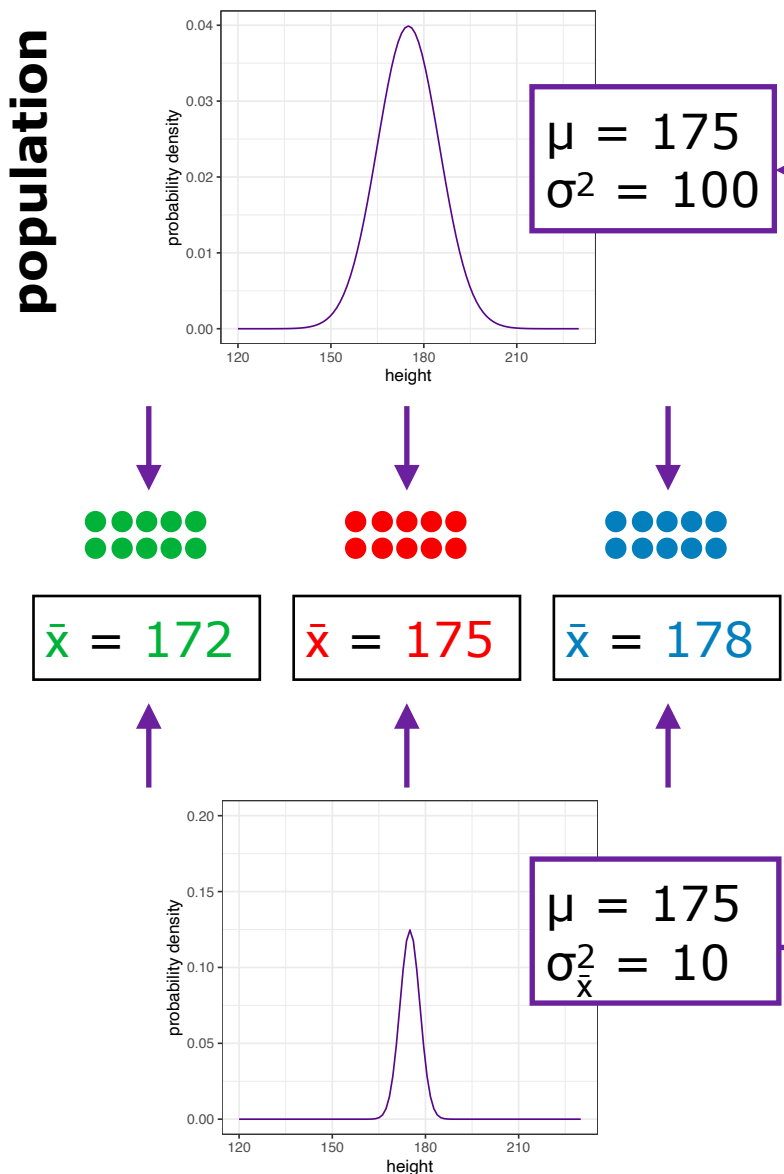
Estimating variance from sample means

To see that we can use sample means to estimate the variance of a population, we need to see the full picture of where sample means come from:

We start with a population. I'll use our old friend height:

We draw 3 samples of 10 from this population. Like an experiment with 3 conditions!

The means of these samples came from the distribution of sample means - the sampling distribution of the mean!



So, we estimate the variance of the sampling distribution of the mean using the sample means:

$$s^2_{\bar{x}} = \frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k-1}$$

Then plug it into the relationship:

$$\sigma^2 = n\,\sigma^2_{\bar{x}}$$

The result is an estimate of σ^2 from the sample means!

Two ways to estimate variance for two or more samples

Method 1: For two or more samples, we can first calculate the sample variances:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} \quad \text{or} \quad s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Then pool the variance of each sample to get an even better estimate (this is a weighted mean):

$$s_p^2 = \frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{(n_1-1) + (n_2-1)} \quad \text{or} \quad s_p^2 = \frac{\sum (n_i-1)\, s_i^2}{n_{total}-k}$$

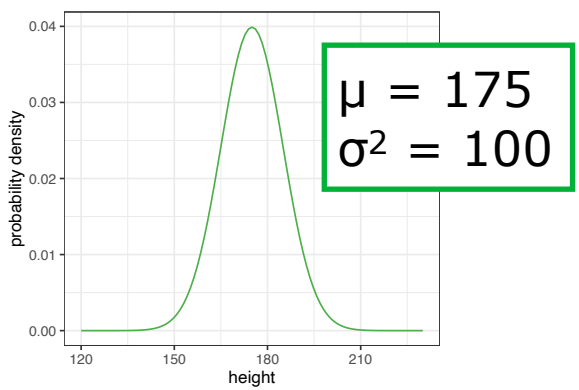
Method 2: For two or more samples, we can use the sample means to calculate the population variance based on the relationship between the variance of the sampling distribution of the mean and the population variance:

$$s_{\bar{x}}^2 = \frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k-1} \quad \text{then:} \quad s^2 = n\, s_{\bar{x}}^2$$

Crucial fact: This second method only yields a good estimate if H_0 is true!

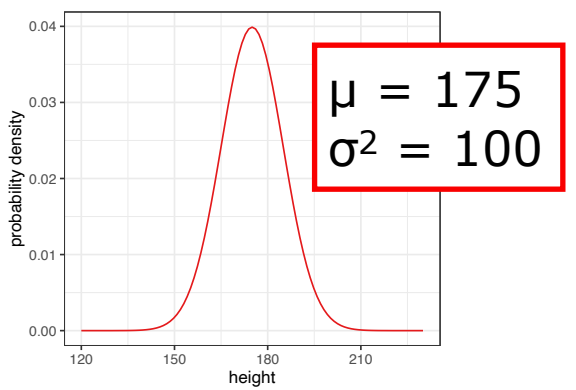
What is the H_0 for three conditions?

To create a hypothesis test around variance, we need to be clear about what our null hypothesis is. Our null hypothesis is that all three conditions come from populations with the same means and variances.



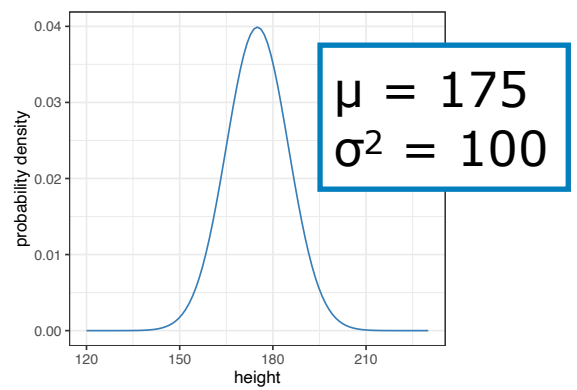
$\mu = 175$
$\sigma^2 = 100$

$\bar{x} = 172$



$\mu = 175$
$\sigma^2 = 100$

$\bar{x} = 175$



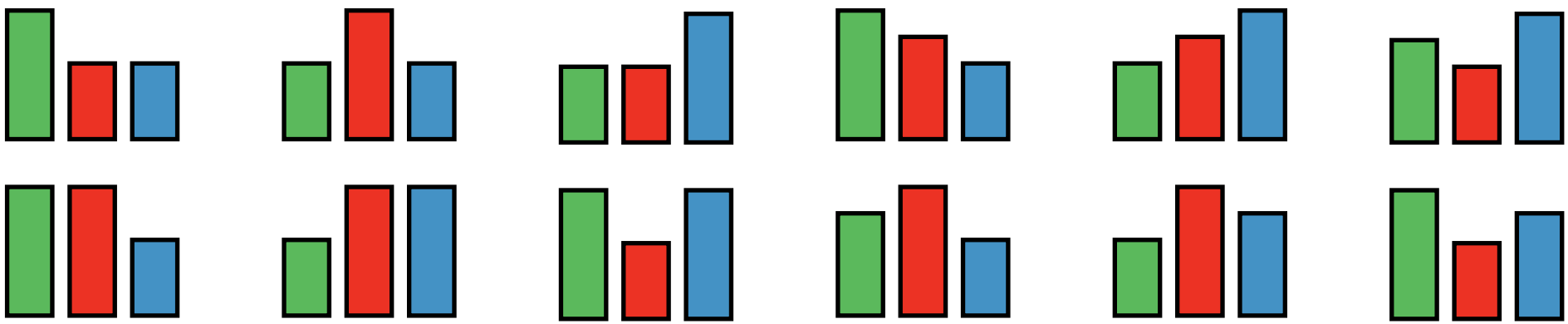
$\mu = 175$
$\sigma^2 = 100$

$\bar{x} = 178$

We often take the thinking shortcut of saying "all 3 conditions come from the same population". But we know they don't come from the same population, because they are different conditions! So, our real H_0 is that the three populations that they come from all have the same mean and variance.

What is the H_0 for three conditions?

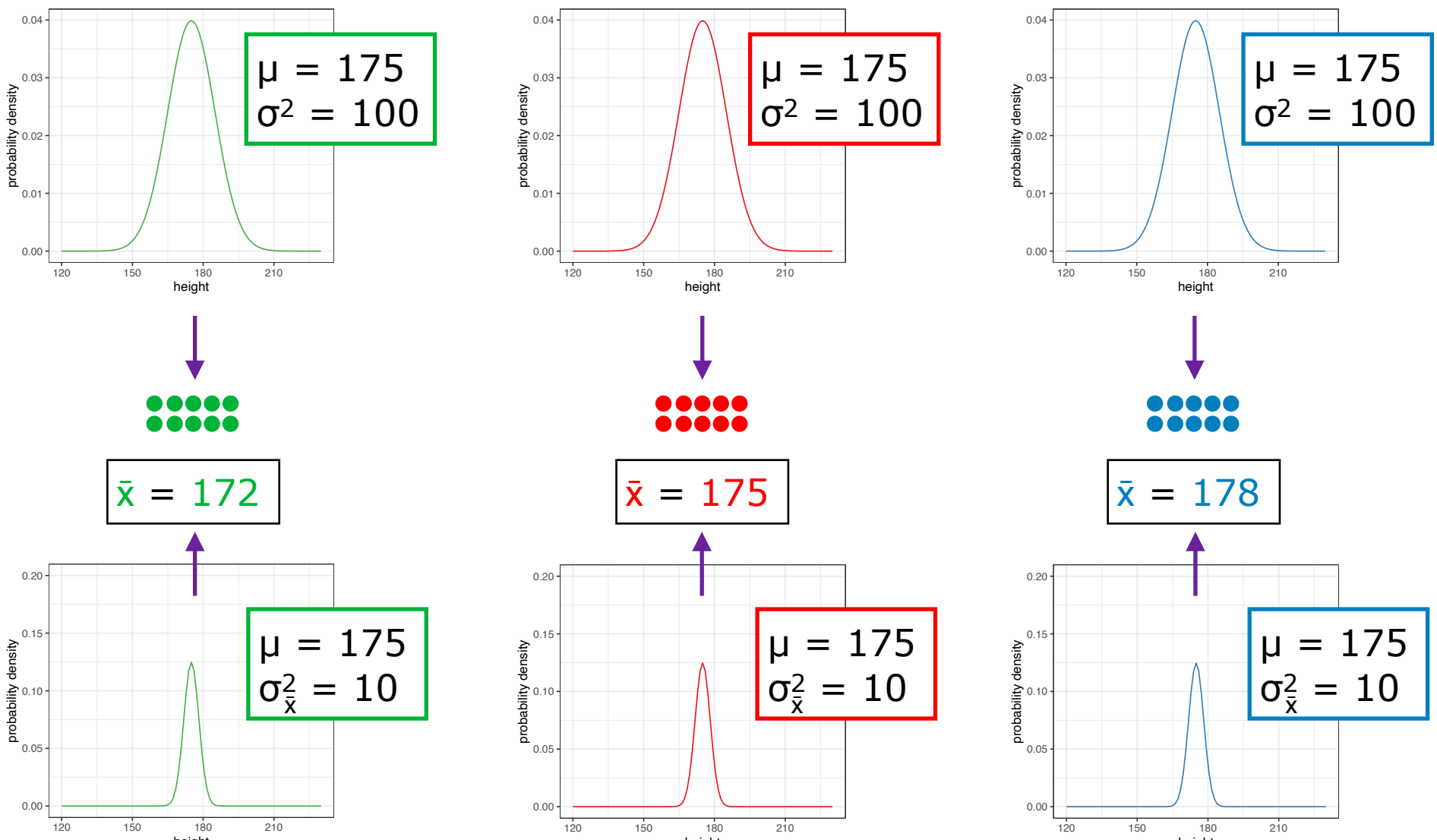
Here are all (I think?) of the patterns of results that will yield a significant result in an ANOVA with three conditions. As you can see, at least one condition needs to be different from the other three.



And here is the null hypothesis:

If **H_0 is true**, our new/second method of estimating variance will be a good estimate

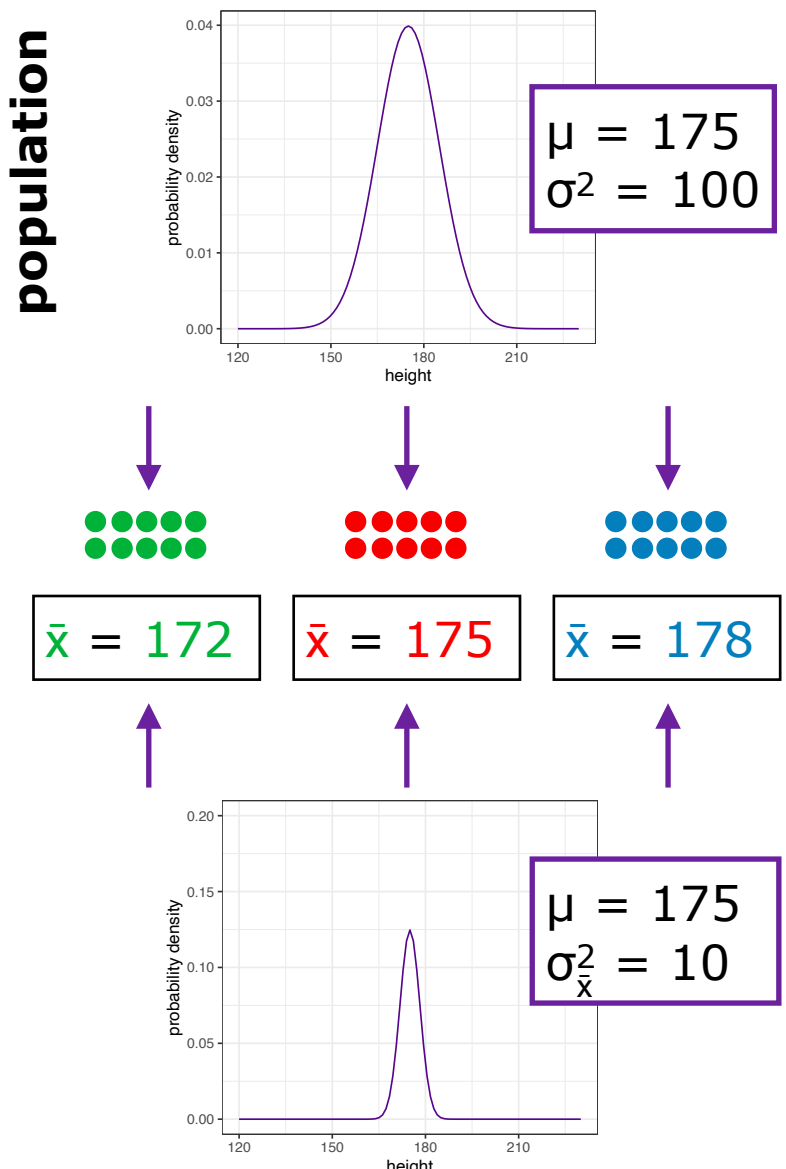
Our new estimate will be a good estimate if H_0 is true because the sampling distributions for each population will also be identical:



If **H_0 is true**, our new method of estimating variance will be a good estimate

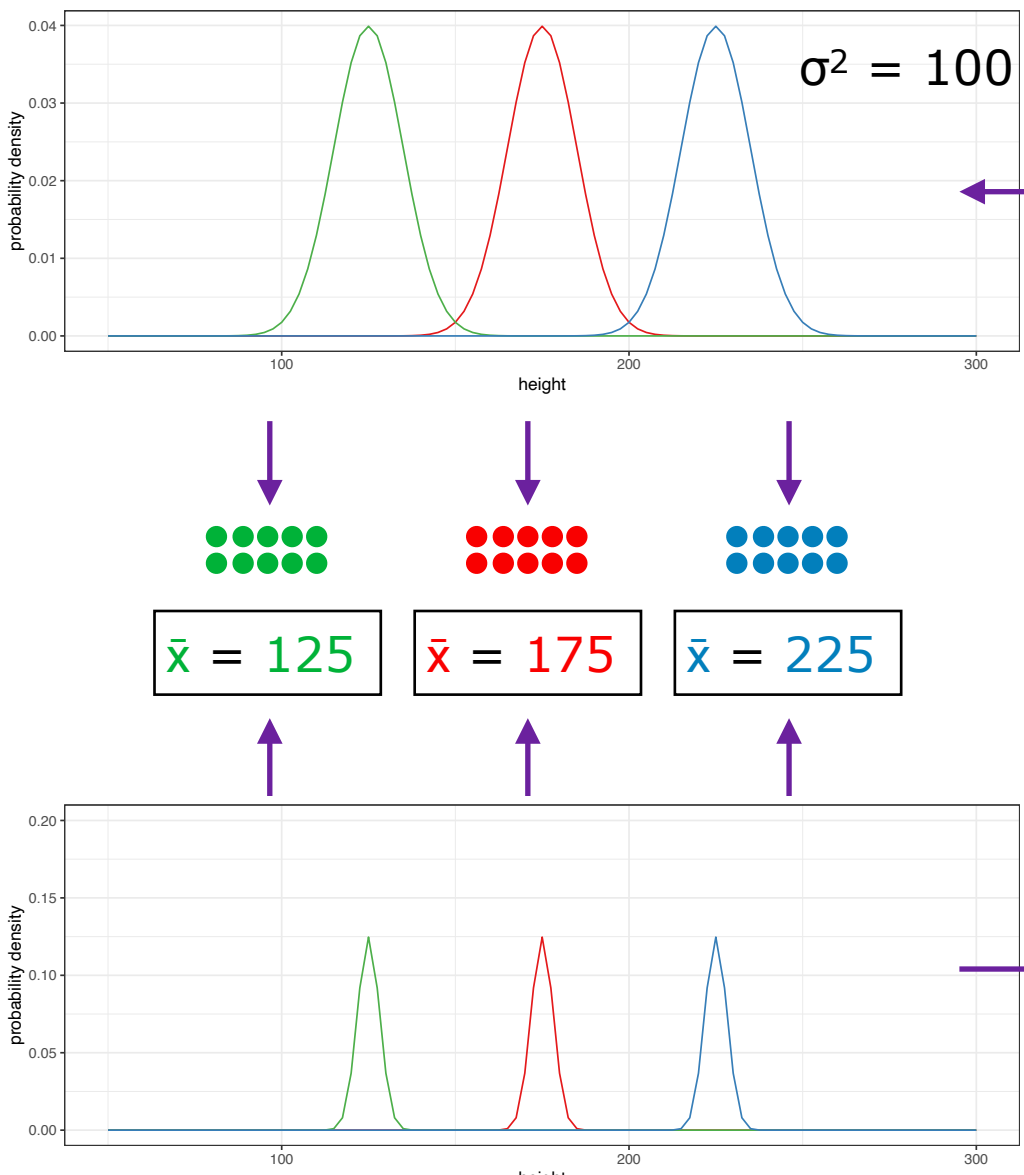
The thinking shortcut for this that it is as if they all come from the same population (same mean and variance), and therefore the same sampling distribution:

I like to think of this single purple distribution as the three populations (green, red, blue) overlapping perfectly.



If **H_0 is false**, our new method of estimating variance will be a bad estimate

Crucially, if the three samples come from populations with different means, then the estimated variance will be **larger than the variance of the population:**



This is because the samples come from three different sampling distributions of the mean. So they will be spread out!

So, when we calculate the variance:

$$s^2_{\bar{x}} = \frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k-1}$$

Then plug it into the relationship:

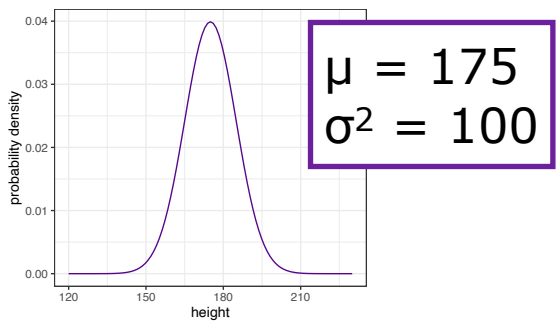
$$\sigma^2 = n\,\sigma^2_{\bar{x}}$$

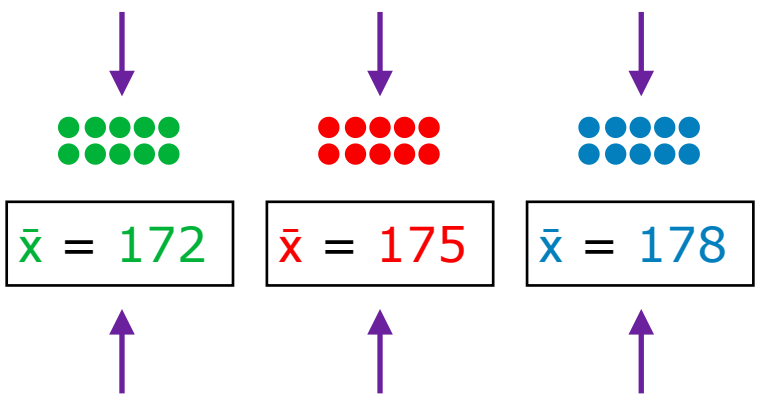
The result will be a much larger estimate of σ^2. Here it is 25,000!

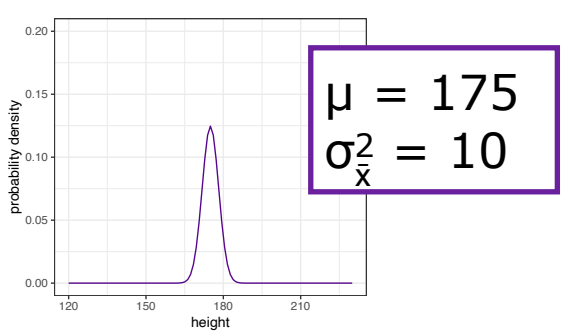
So, two different universes lead to two different outcomes!

If H_0 is true, we get a **good estimate**

populations

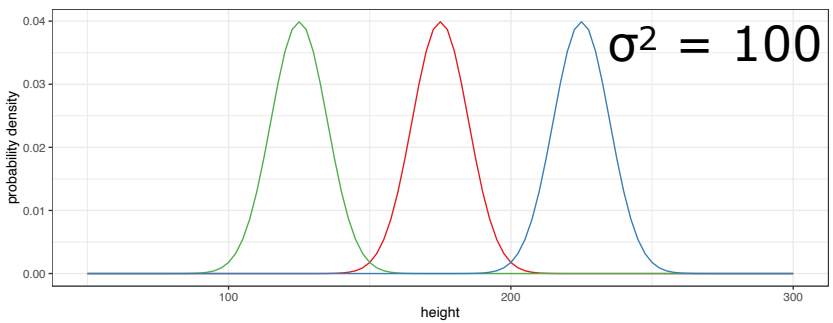


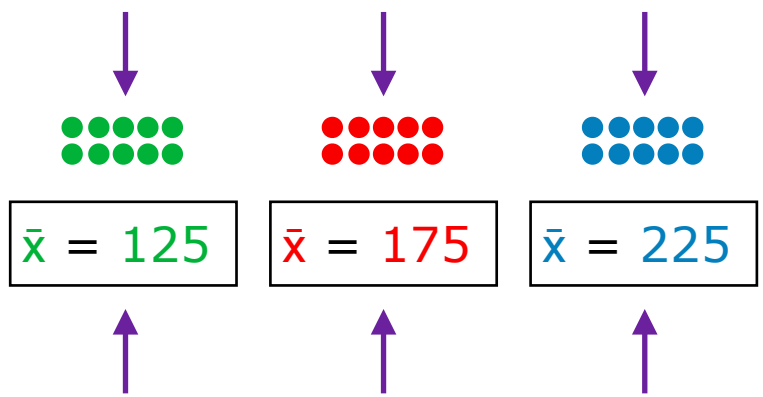


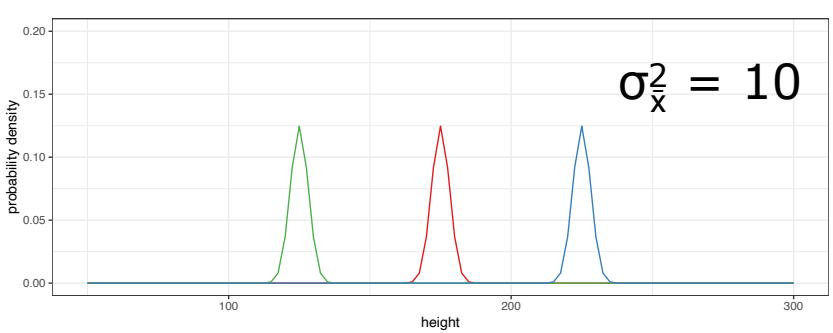


estimate = 90

If H_0 is false, we get a **large estimate**







estimate = 25,000

We can use this to turn variance into a test of the null hypothesis!

We can call variance "mean squares"

Statistics likes to have lots of terms for the same thing. This is because of the history of the field. Here's a new term for variance: **mean squares**

$$s^2 = MS = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

You can see where this name comes from. Variance is the sum of squares, divided by the degrees of freedom… so it is like a mean of the sum of squares.

Our new method is "between groups"

Our new method of calculating variance uses sample means from three different groups. It is looking at the variability between groups, and using that variability to estimate the population variability. Therefore we call it **mean-squares-between:**

$$MS_B: \quad n\frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k-1}$$

Notice that the **MS_B** equation has "n" in it. This is because it is a combination of the two equations we use in this method: the equation for the variance of the sampling distribution of the mean is plugged it into the equation that relates the variance of the sampling distribution of the mean to the variance of the population:

$$s^2_{\bar{x}} = \frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k-1} \qquad s^2 = n\,s^2_{\bar{x}}$$

In this equation, **k** is the **number of groups**, and **n** is the **sample size**.

Equal sample sizes vs unequal sample sizes

The formula we just built assumes that all of the groups (the conditions) have the same sample size. But this is not strictly necessary. We can use a specific n for each group like this:

equal sample size

MS_B: $$n\frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k-1}$$

unequal sample size

MS_B: $$\frac{\sum n_i(\bar{x}_i - \bar{x}_G)^2}{k-1}$$

Notice that one change is in the position of the n. This follows from the laws of summation — the equal sample size version could have had the n inside the summation too. I kept it outside to make the connection to the relationship between standard error and standard deviation very clear.

I will try to keep the sample sizes equal for all of the ANOVAs that we run in the homework and exam to keep things simple. But I want you to know that you don’t have to do that. You can have unequal sample sizes.

Our previous method is "within groups"

Our pooled variance method calculates the variability within each group by looking at the raw scores in the samples. In other words, it is an estimate of the population variance based on the variability within each sample. So we call it **mean-squares-within**. Here is the equation for two samples:

$$MS_W = \frac{(n_1-1)\,s_1^2 \;+\; (n_2-1)\,s_2^2}{(n_1-1) + (n_2-1)}$$

We can use summation notation to expand this for any number of samples:

$$MS_W = \frac{\sum (n_i-1)\; s_i^2}{n_{total}-k}$$

Here, n_{total} is the sum of the sample sizes for all of the groups. And k is the number of groups. (And notice that we already have a different n for each group, so this equation works for both equal and unequal sample sizes.)

What counts as a "good" estimate?

Remember, MS_B is the interesting one. It will be a good estimate when H_0 is true, and a bad estimate when H_0 is false:

$$MS_B = n\frac{\sum(\bar{x}_i - \bar{x}_G)^2}{k-1}$$

But to know whether it is "good" or "bad", we need to know the variance of the population. How do we do that? **We use MS_W!**

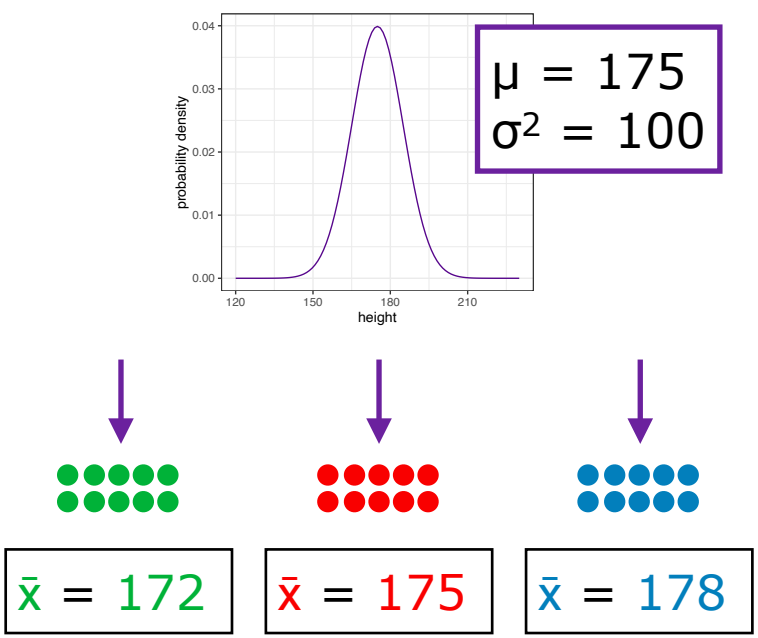
$$MS_W = \frac{\sum (n_i-1)\ s_i^2}{n_{total}-k}$$

MS_W is always a good estimate of the variance. This is because we require that the conditions come from populations with equal variances when we run ANOVAs. We call this **homogeneity of variance**. It is an assumption of the test. That means it is a requirement.

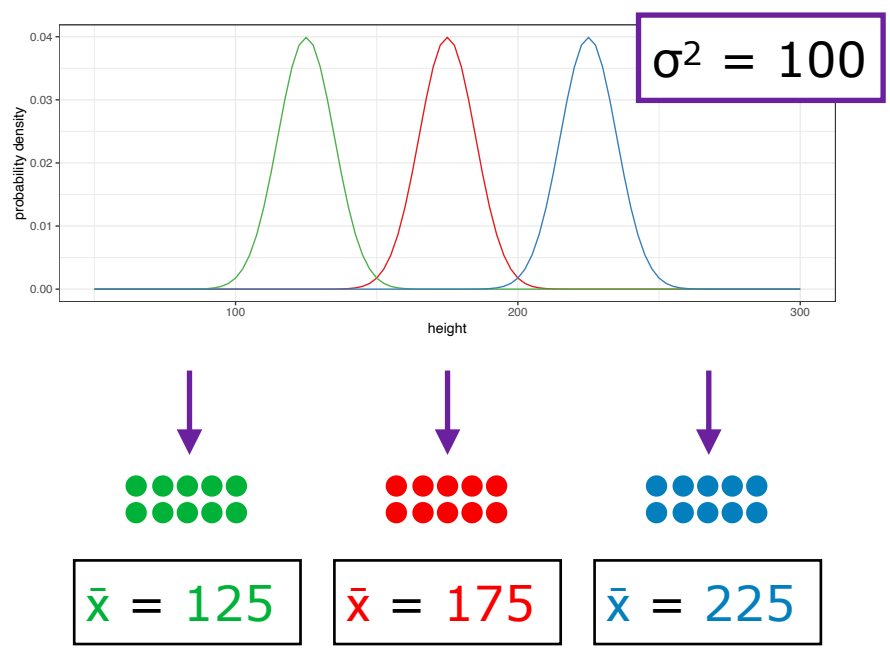
Homogeneity of variance

Homogeneity comes from the word homogenous, which is a fancy way of saying identical or the same kind. ANOVA assumes (and therefore requires) that the conditions all have homogenous variance.

If H_0 is true, variance is equal.



If H_0 is false, variance is equal.



If all of the samples come from populations with the same variance, then pooling them together in MS_W will give a really good estimate of the population variance!

The ANOVA equation

ANOVA puts these two ideas together. MS_B varies based on H_0. It is the core of our test. MS_W is always a good estimate, so we use it to scale MS_B - to tell us what a good estimate of the variance is. That means we divide MS_B by MS_W.

$$F = \frac{MS_B}{MS_W}$$

We call the resulting statistic "*F*" in honor of Ronald Fisher, who developed the ANOVA method.

Even before we look at *F* in detail, you can already see what it does:

If H_0 is true, **MS_B** will be a good estimate, and **MS_W** will be a good estimate (because it always is), therefore *F* will be roughly 1 because the numerator and denominator will be roughly equal.

If H_0 is false, **MS_B** will be large, while **MS_W** will remain a good estimate. So *F* will be larger than 1 because the numerator will be larger than the denominator.

Putting the logic together

ANOVA is a way to test experiments with 3 or more conditions.

It tests the H_0 that all conditions come from populations with the same means and variances.

It works because one way of estimating variance, called MS_B, is only a good estimate of the population variance when H_0 is true. When H_0 is false, it gets large. But the other way of estimating variance, called MS_W, is always a good estimate.

So, if we put these two estimates in a ratio like this, called an *F*-ratio or *F*-statistic, we get a test of H_0. F will be 1 when H_0 is true, and larger than 1 when H_0 is false.

$$F = \frac{MS_B}{MS_W}$$

From here, we can do what we always do in statistics! We can figure out what the distribution of F looks like. We can find critical *F*-statistics for any given experiment. We can calculate an *F* for a real experiment. We can calculate a *p*-value from that F. We can plan experiments by figuring out the sample size for good statistical power. And much more.

But we will do that next time. Today was just about the logic.