Jonathan M Brown
Updated: 2020_08_13

Let's chat about…Degrees of Freedom!

I. Summary

Broadly, what in the heck are "degrees of freedom"?

In physics, this is the idea of "ability to move" or "directions/axes along which something can move".

In statistics "degrees of freedom" translates to AMOUNT of AVAILABLE INFORMATION.

That's it? What kind of "information"?

Generally, it just refers to how much DATA you have! So, when you collect data, you have information.

In most cases, we quantify this "amount of information" as simply the sample size, BEFORE we've run any statistics or made any hypotheses or models.

Once we start making hypotheses, models, or finding different statistics, we TAKE INFORMATION AWAY from our total degrees of freedom. You can't do statistics without information, so our data provides "degrees of freedom" with which to make hypotheses and models.

I'm confused. We start with some amount of "degrees of freedom" and we "use" them with our statistics?

Yes!

Think of "degrees of freedom" as a finite "resource" for your hypotheses/models. The MORE things you try to do with your hypotheses/models, the more information you need. The more information you need, the more "degrees of freedom" you need.

Is there another way to think about this?

Yes.

Think of your data like fresh vegetation in a forest. It's finite, and it's useful. Think of your hypotheses/models as friendly forest monsters that ONLY talk in stats results, like means, t-scores, p-values, etc.

In order for the forest stats monsters to TALK they need to be FED vegetation. If they are fed a LOT of vegetation, then they TALK a lot of stats. If there is only a LITTLE bit of vegetation, then they can only talk a LITTLE bit of stats.

But don't worry, for most introductory statistics cases, you don't need much vegetation! This is just a thought experiment.

II. A technical look

Here's a technical consideration.

To start with:

Each OBSERVATION provides ONE degree of freedom

As you run statistics/models:

Each thing you need to "estimate" TAKES AWAY or EATS ONE degree of freedom.

To end with:

Starting degrees of freedom - number of things you need to estimate

= degrees of freedom leftover

Can you be a little more specific about this relationship?

In the simplest case for a statistical model/hypothesis:

Sample size – number of things you need to estimate

= "remaining" degrees of freedom for your model/hypotheses

Due note: more complicated models and statistical methods employ different kinds of degrees of freedom! But the general idea still applies.

Example:

If I have four data points and I want to calculate their average value, that translates to:

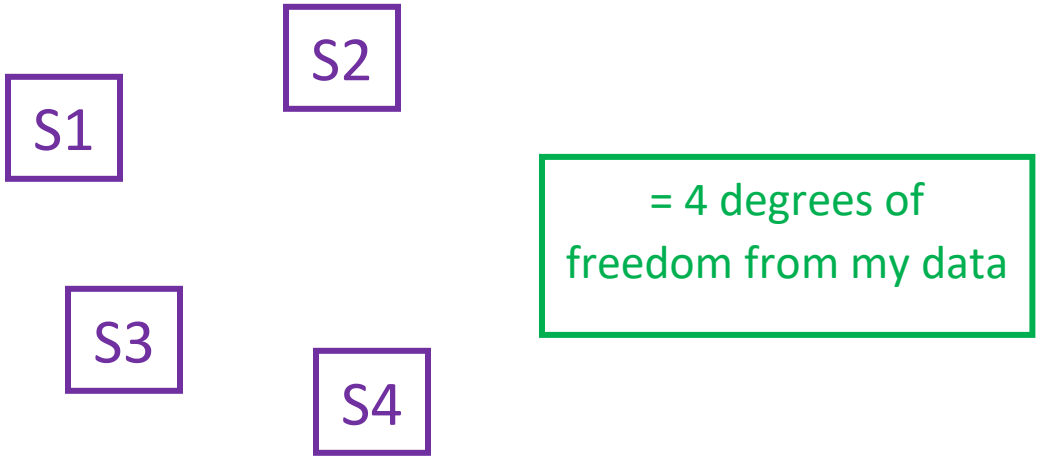
4 data points – **1** thing to estimate = **3** degrees of freedom leftover

Why is this true? I guess I'm not sure why this matters. Would you provide a concrete example of this?

YES!

This will be a little math-y, but it will hopefully illustrate the point.

Suppose, I have four biology class exam scores for four students: S1, S2, S3, and S4. That means I currently have FOUR degrees of freedom, with FOUR pieces of information:



Now, suppose I tell you that I KNOW the average of these four scores to be 90, without actually knowing any of the individual scores:

Now, let's fill in some POSSIBLE individual scores to obtain an average of 90:

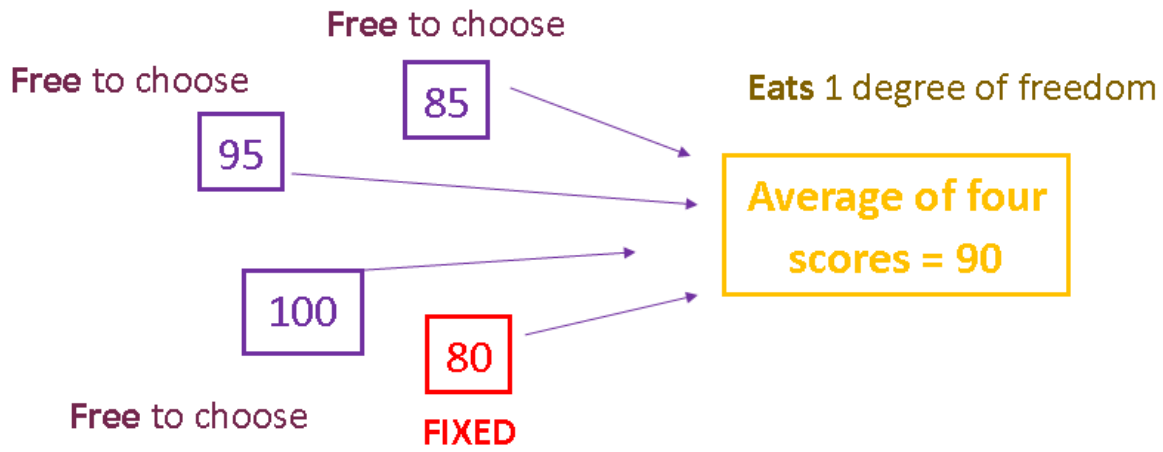
- **Score 1:** What value can I pick to fill in for Score 1 and still mathematically have an average value of 90 across the four scores? ANY SCORE I WANT! I have the "freedom" to choose any value at this point. Let's "freely" pick 95.

- Score 2: Again, what score can I pick that still allows for an average of 90 across the four scores? Again, ANY VALUE I WANT! There's nothing mathematically stopping me from picking anything. Let's "freely" pick 85:

- Score 3: What score can I pick that still allows for an average of 90 across the four scores? Still, ANYTHING! I continue to have "freedom" to do so. Let's "freely" pick 100:

- Score 4: Now it gets interesting. What score can I pick that still preserves an average score of 90 across the four scores? Well, since I have three of the values picked, and I already know the average, there is only ONE VALUE I can choose!!! I do NOT have freedom to choose my score for Score 4! It is thus "constrained". Working backward, there is only ONE value for Score 4 that will produce an average of 90, given the other three exam scores, and that value is 80!

Putting it all together, this summarizes the process we just went through:



We could have done this with any three scores, e.g., pick S2, pick S3, pick S4, and then be constrained to fix S1.

The point: after you "freely choose" the first three scores, the last (fourth, in this case) exam score HAS to be a certain value.

Therefore, when we have a sample size of 4, and we estimate 1 thing (an average, in this case), we have "3 chances to freely pick scores", or "3 degrees of freedom" leftover in our situation.

III. In summary

That's it? Is that how it always works?

In principle, yes!

In most cases, for example, with t-tests, you see the degrees of freedom calculation as the sample size minus "1" or n – 1. This is the exact same idea: the sample provides "n" pieces of information, but since we are estimating some kind of sample mean or sample proportion, that "eats up" one of the degrees of freedom for the t-test.

Does it get more complicated?

Absolutely. For more complicated tests, such as a two-sample t-test, the degrees of freedom are not straightforward, and your software will choose a particular "estimated" degrees of freedom.

For more complicated models or situations, what counts as "information" and what counts as "eating up degrees of freedom" also change. Thus, "n – 1" as a degree of freedom calculation ONLY applies to particular situations.

How will I know what to do in future degrees of freedom calculations?

If you're asked to compute degrees of freedom for something, generally, there will be a formula to follow that captures the ideas above.

In more complicated cases, the software will produce estimated degrees of freedom for you.

What's the main thing I need to keep in mind going forward?

ALL degrees of freedom computations are "starting information" minus "things I want to estimate in my model/situation".

Typically, the larger the sample size (and in some cases, the more variables you have) the MORE starting "information" you have, and the more complicated a model you can run.

Thus, keep in mind, the more complicated a model you want to investigate, the more things you need to "estimate", thus the more your model "eats up" degrees of freedom from your "starting information".