Degrees of Freedom Exploration

Motivation:

Degrees of freedom can be a confusing concept for students and, often, instructors (including me) hand-wave them in an introductory course so as to not "sew confusion." I wanted to make a document that covers the basics of this concept so it's not just one more thing to memorize for you when you utilize the concept. There are numerous sources on the web for those who want to know more and/or need a different explanation than I can provide here.

Estimates:

If you've spent any time in this class (or any sort of statistics class), hopefully you've grasped the concept that a lot of our effort is aimed toward learning something about a population based on a sample. Usually that involved estimating a parameter using a statistic or testing how much evidence a sample statistic provides reference a hypothesis concerning a parameter value.

Degrees of freedom are the number of pieces of information we have to estimate these population values using our statistics. So let's look at a couple statistics that you are familiar with.

Mean:

Let's say that we have a set of values that represent the amount that ten random Firsties spent on their class rings: $[1523, 1452, 1324, 1789, 2310, 1123, 957, 1231, 1134, 1423]$ We could use this sample to estimate the average that Firsties spend on their rings.

```
library(tidyverse)

costs = c(1523, 1452, 1324, 1789, 2310, 1123, 957, 1231, 1134, 1423)

mean(costs)
```

```
## [1] 1426.6
```

The equation for to calculate this mean is:

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

This equation is otherwise known as "add up all the observations and divide by the sample size." How many pieces of information did we have to compute this estimate? We had ten values that could take on any value they want (more to follow with this), so we had ten degrees of freedom for this calculation.

Standard Deviation

Let's calculate the standard deviation of this sample:

```
sd(costs)
```

```
## [1] 389.9827
```

The equation for sample standard deviation is:

$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

You can see from this equation that we require the sample mean ($\bar{x}$) in order to calculate this value. Because we had to calculate this value in order to calculate the standard deviation, we have lost one degree of freedom. Why? We calculated the mean which, as it is a statistic, is a numerical summary of the data. The numerical summary is calculated to tell us something about the sample of data without having to provide the whole sample.

If we only had this summary of the data, how much information do we have about the data? Put another way: what happens if I try to develop another sample with the same summary value? We are free to pick whatever values we want for the first 9 values but we are restricted on the last if we wish to have the same mean. We have lost one degree of freedom.

$[1250, 1433, 1130, 2304, 1523, 1442, 1130, 1328, 1782, ???]$

Because the sample mean is involved in the calculation of the standard deviation, we no longer have ten (n) degrees of freedom and are now down to nine ($n-1$). You'll see this fact taken into account in the denominator of our equation. There is another reason for dividing by one that I'll cover in the *Bessel's Correction* session in the next section.

Bessel's Correction

There is another aspect to using $n-1$ as opposed to n in the denominator of our sample standard deviation (s) and that relates to the fact that our equation for s is a derivative of the equation for our population standard deviation (σ):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$

As you can see these equations are very similar with two important differences: n in the denominator and the use of μ instead of $\bar{x}$. By substituting the sample mean in place of the population mean, we are adding uncertainty to our estimate of the population standard deviation. Often we don't have a choice because we don't know the value of μ.

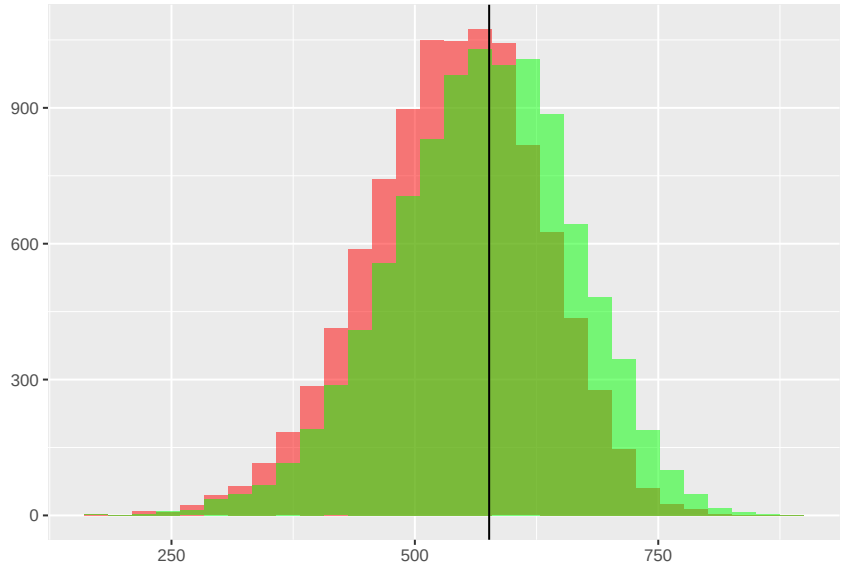
By substituting $\bar{x}$ for μ and not changing anything else, we have created a *biased* estimator of the population standard deviation. We create the *unbiased* estimator by using $n-1$ in place of n. Here is a small simulation below that might help convince you.

```
#Create a "population" so I can know the standard deviation
pop = runif(5000, min = 0, max = 2000)

pop_sd = sd(pop)

pop_sd
```

```
## [1] 576.304
```

```
replications_dataframe = NULL

num_reps = 10000

sample_size = 10

for (i in 1:num_reps){
  trial_sample = sample(pop, sample_size, replace = FALSE)

  biased_sd = sqrt(sum((trial_sample - mean(trial_sample))^2) / sample_size)

  unbiased_sd = sqrt(sum((trial_sample - mean(trial_sample))^2) / (sample_size - 1))

  trial_stats = cbind(biased_sd, unbiased_sd)
```

```
  replications_dataframe = rbind(replications_dataframe, as.data.frame(trial_stats))
}

replications_dataframe %>%
  ggplot() +
  geom_histogram(aes(x = biased_sd),
                 fill = "red",
                 alpha = 0.5) +
  geom_histogram(aes(x = unbiased_sd),
                 fill = "green",
                 alpha = 0.5) +
  geom_vline(xintercept = sd(pop)) +
  labs(x = "", y = "")
```



Because I have the entire population, I know the population standard deviation (the line on the plot). I draw samples from this population repeatedly, calculate the biased and unbiased standard deviations of these samples and record them for each trial. The histogram demonstates that using $n-1$ provides an unbiased estimator for the population standard deviation.

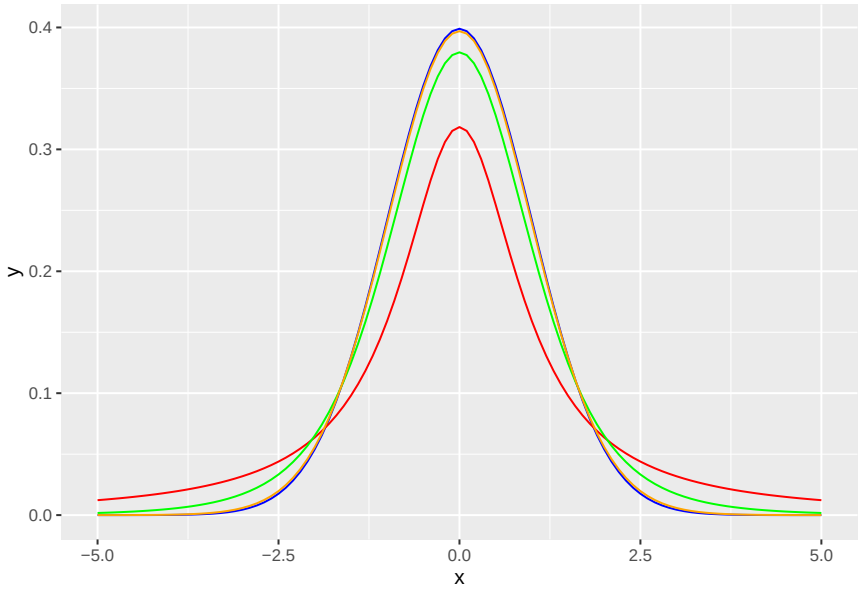
Degrees of Freedom in Hypothesis Tests:

t-distribution

You may recall that we first start talking about this degrees of freedom business in hypothesis tests (at least overtly) when we started using the t-distribution. You may recall that we discussed the fact that we were using s as an estimator for σ in our standardized statistic. In order to account for the uncertainty introduced by doing so, we utilized the t-distribution instead of the normal distribution.

The shape of every distribution that has one or more "degrees of freedom" parameters is determined by these parameters. With the t-distribution, a larger sample size meant more degrees of freedom and more degrees of freedom meant a shape closer to the normal distribution. Hopefully this makes sense as the larger the sample size, the less uncertainty about using s instead of σ, and therefore we get closer to what we expect based on the *central limit theorem*.

```r
ggplot(data.frame(x = c(-5,5)), aes(x = x)) +
  stat_function(fun = dnorm,
                color = "blue") +
  stat_function(fun = dt,
                args = list(df = 1),
                color = "red") +
    stat_function(fun = dt,
                args = list(df = 5),
                color = "green") +
    stat_function(fun = dt,
                args = list(df = 50),
                color = "orange")
```



Pearson's Chi-Squared Test:

The chi-squared test we discuss in this class is used for comparing multiple proportions. As with the one- and two-sample z-tests, a summary table of your sample is useful in providing the necessary proprtions for your calculations. If you are familiar with the steps of calculating the chi-squared statistic (χ^2), you will know that this summary table is a requirement.

The degrees of freedom parameter for the χ^2 distribution is determined by the number of rows (r) and number of columns (c) in your summary table. The relatively simple equation is as follows:

$df = (r-1) \times (c-1)$

Similar to how we lost one degree of freedom when calculating the sample standard deviation, given the row and column totals in your summary table remain the same, only $r-1$ row values and $c-1$ columns values are free to vary with the last being determined by that associated total.

ANOVA:

The F-statistic is is used when you are testing whether there is a difference in means between groups. The the F-distribution has two degrees of freedom parameters:

$df_1 = k-1$ where k is the number of groups you are comparing
$df_2 = n-k$ where n is your sample size

You may recall that the F-statistic is a ratio of two variances: "between group variance" and "within group variance". Respectively, our two degrees of freedom values correspond to these variances.

A very important thing to keep in mind is that the p-value (as determined by the F-statistic) is the probability of the null hypothesis being true. Therefore we are assuming the truth of the null hypothesis when we are calculating these statistics.

Between group variance:
The equation for *treatment mean squares* is:

$MSTr = \frac{\sum_{i=1}^{k} n_i \times (\bar{y}_i - \bar{y})^2}{k-1}$ where:
k is the number of groups
$\bar{y}_i$ is the group sample mean for the response variable
$\bar{y}$ is the overall sample mean for the response variable

You will notice in this equation that we are utilizing the overall (grand) mean. If you have k group sample means that factor into the calculation of this grand mean, then only $k-1$ of them are free to vary. Hence our degrees of freedom for this measurement is $k-1$.

Within group variance:
The equation for *error mean squares* is:

$MSE = \frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{n-k}$ where:
k is the number of groups
n_i is the group sample size
s_i^2 is the group variance

In this case, we are calculating the group sample standard deviation for k groups. The degrees of freedom for each one of the standard deviation calculations is $n_i - 1$ (see top of document for standard deviation explanation). We know that $\sum_{i=1}^{k} n_i = n$ (all the group sample sizes add up to n), so if we have k group when our degrees of freedom is $n-k$. Another way to say this is $\sum_{i=1}^{k} (n_i - 1) = n - k$.