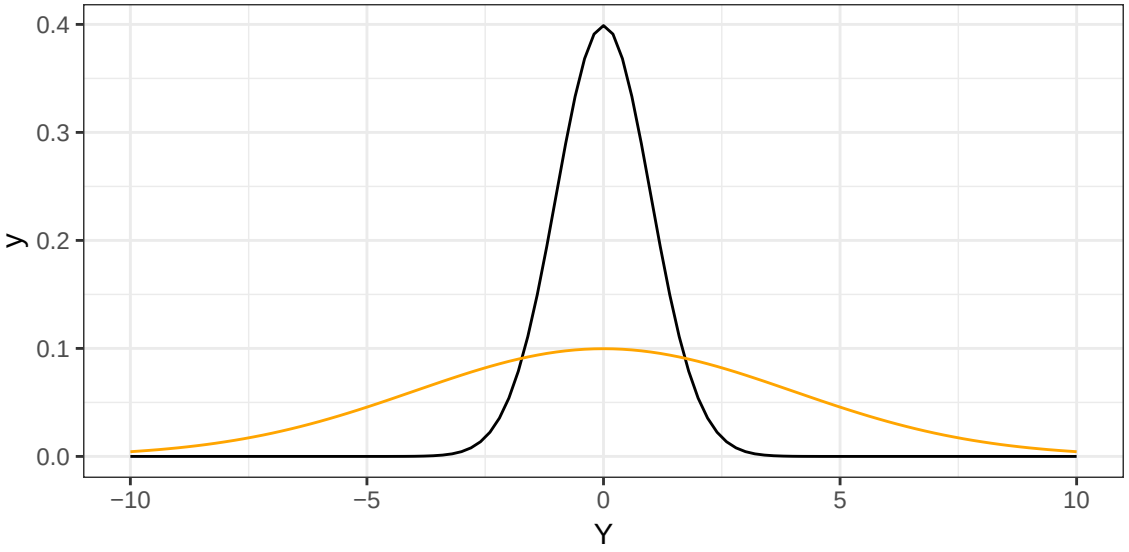
Degrees of Freedom for Tests about a Single Mean

Population Variance

- The variance the average squared difference from the mean.
- Imagine a population of N people, each has a height Y_i
- Mean height is $\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$
- Variance of heights is $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2$



- Both distributions have mean 0
- The distribution shown in orange has larger variance than the distribution shown in black
  - A typical value is farther from 0
  - A typical value has larger squared distance from 0
  - Across all values, the average squared difference from 0 is larger

Sample Variance

- The *sample variance* s^2 is used as an estimate of the population variance σ^2.
- Suppose we tried to use

$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

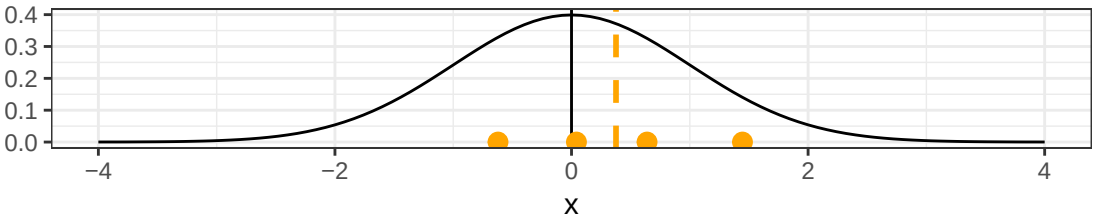
- Doesn't work because:
  - Values in sample tend to be a little closer to $\bar{Y}$ than to μ
  - Squared differences from $\bar{Y}$ are a little smaller than squared differences from μ
  - $\tilde{s}^2$ tends to be less than σ^2
- Dividing by $n-1$ instead of n is just the right adjustment:

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Demonstration by simulation

- "Population": Normal$(0, 1)$ Variance is $\sigma^2 = 1$
- Sample of size $n = 4$

In black: Population distribution, population mean
In orange: Sample observations, sample mean



Here's our data frame with sample data:

```
sample_df
```

```
##             x
## 1  1.44521048
## 2  0.63852357
## 3  0.04167507
## 4 -0.62136233
```

Sample mean is:

```
sample_df %>%
  summarize(
    mean = mean(x)
  )
```

```
##        mean
## 1 0.3760117
```

Average squared difference from population mean of 0:

```
sample_df %>%
  summarize(
    mean_squared_difference_from_0 = mean((x - 0)^2)
  )
```

```
##   mean_squared_difference_from_0
## 1                      0.7210434
```

Average squared difference from sample mean of 0.376:

```
sample_df %>%
  summarize(
    mean_squared_difference_from_0 = mean((x - 0.376)^2)
  )
```

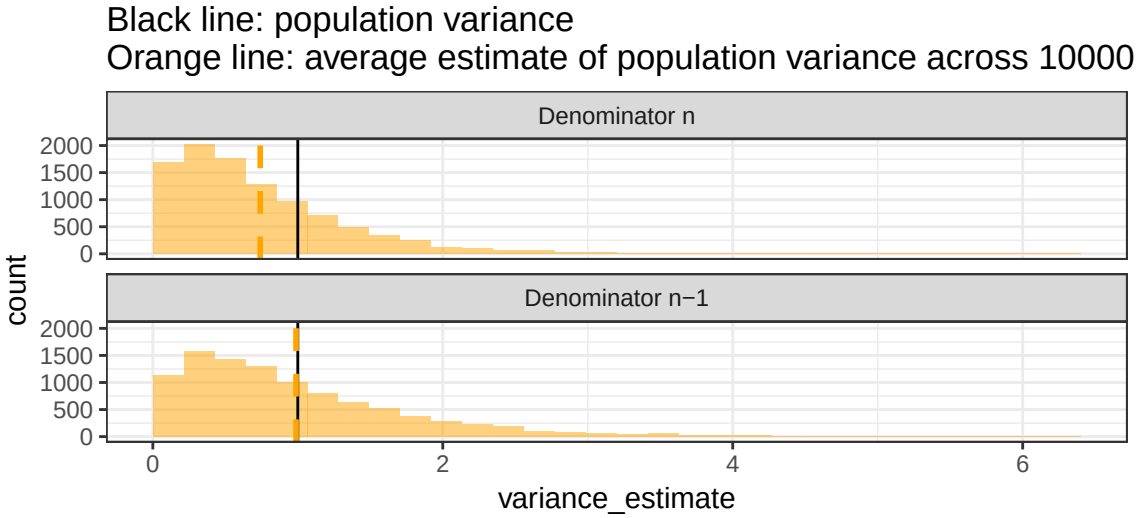
```
##   mean_squared_difference_from_0
## 1                      0.5796586
```

Corrected by dividing by $n - 1$ instead of n:

```
sample_df %>%
  summarize(
    almost_mean_squared_difference_from_0 = sum((x - 0.376)^2) / (4 - 1)
  )
```

```
##   almost_mean_squared_difference_from_0
## 1                             0.7728781
```

- For this particular sample, all three are below the population variance of 1. What about across 10000 samples?



Connection to t tests

- Our t statistic for a test about the value of a mean μ is

$t = \frac{\bar{Y}-\mu}{SE(\bar{Y})} = \frac{\bar{Y}-\mu}{s/\sqrt{n}}$

- s shows up in this calculation, with associated degrees of freedom $n-1$.

What if we had more than one mean (e.g. two or three groups)?

The degrees of freedom is always the sample size minus the number of parameters for the mean.