Standard Error

Introduction

Just as the standard deviation describes the distribution of individuals around a population mean, so the standard error of the mean describes the distribution of samples around their sample mean. In fact, the standard error of the mean is simply the standard deviation of the samples around the sample mean.

Case Study: Tomatoes

Once again, we will work with the tomato, barley, cotton, and peanut datasets. Let's load the tomato dataset and look at its top six rows of data.

```
tomato = read.csv("data/tomato_uniformity.csv")
head(tomato)
```

```
##   row col yield
## 1   1   1    48
## 2   2   1    61
## 3   3   1    69
## 4   4   1    59
## 5   5   1    71
## 6   6   1    46
```

Calculating Standard Error

The standard error of the mean is equal to the population standard deviation, divided by the square root of the number of samples. We can calculate these as follows. To get the standard deviation, use the sd() function.

```
yield = tomato$yield  # define the column of interest

tomato_sd = sd(yield)
```

Then we can divide the standard deviation by the square root of n to get the standard error of the mean. What would be the standard error of our sample mean, if we took 4 samples?

```
tomato_se = tomato_sd / sqrt(4)
tomato_se
```

```
## [1] 5.249217
```

We will talk more about the t-distribution in the next exercise, but for now lets assume about 95% of the sample means should be within about two standard errors of the mean of their distribution. Two times the standard error is about 10.6, so we would expect 95% of sample means to be within 50.3 +/-10.6, or (39.7, 60.9).

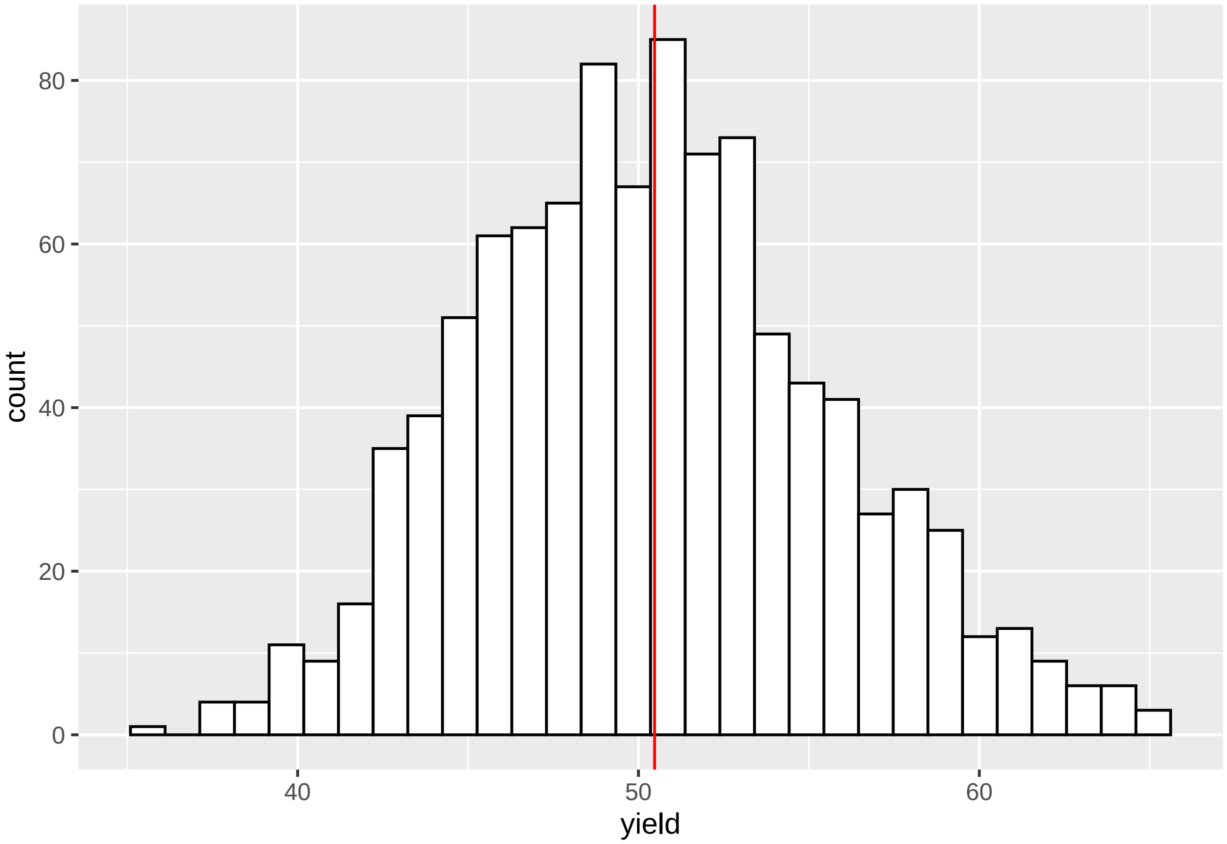
Don't worry about understanding the code below. Just run it and observe the plot. In the plot below, we have simulated 1000 sample means, each based on four samples, from the tomato yield population. We can see the range (39.7, 60.9) does include most of the population, with the exception of the tails.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
set.seed(2003)
sample_list = list()
for(i in c(1:1000)){
  samples = mean(sample(yield, 4))%>%
    as.data.frame()
  sample_list[[i]] = samples
}
sample_list_df = do.call(rbind.data.frame, sample_list)
names(sample_list_df) = "yield"
pop_mean = mean(yield)
ggplot(sample_list_df, aes(x=yield)) +
  geom_histogram(fill="white", color="black") +
  geom_vline(xintercept = pop_mean, color = "red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



What would be the standard error of our sample mean if we took 7 samples?

```
tomato_se = tomato_sd / sqrt(7)
tomato_se
```

```
## [1] 3.968035
```

Practice: Barley

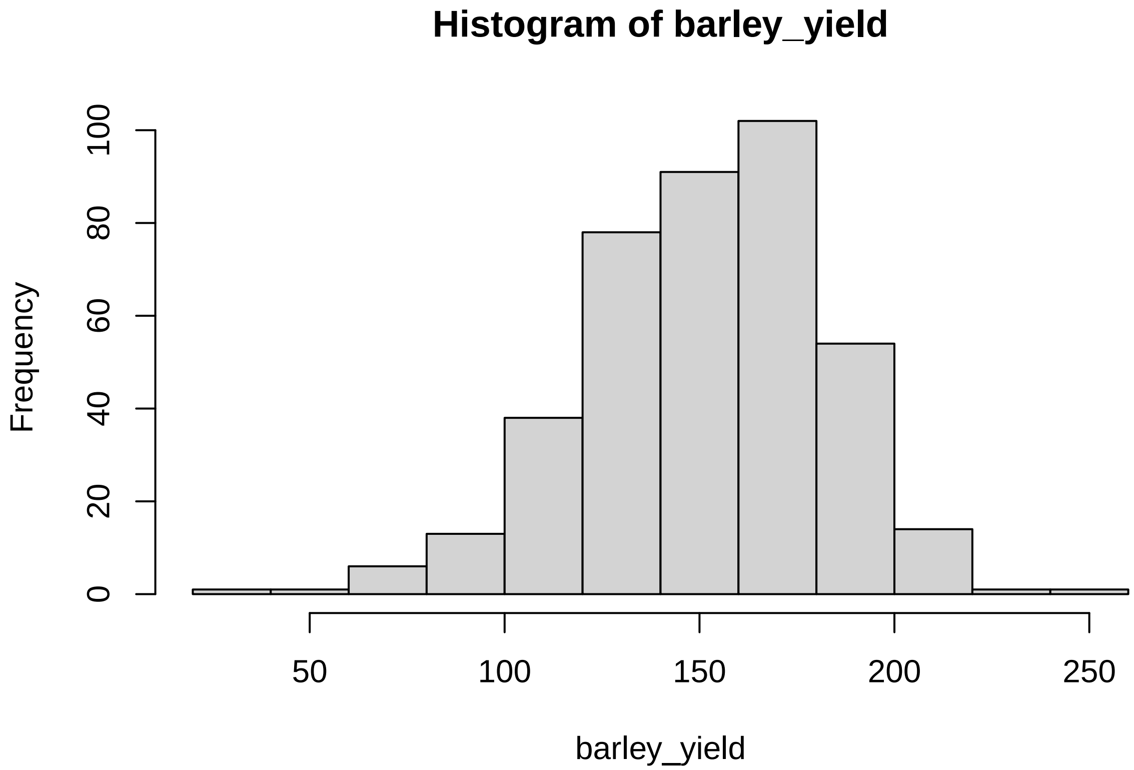
Now let's practice calculating the standard error. What would be the standard error of our sample mean if we took 8 samples? I'll get you started. First, let's load the "barley_uniformity.csv" data and inspect the first six rows.

```
barley = read.csv("data/barley_uniformity.csv")
head(barley)
```

```
##   row col yield
## 1  20   1   185
## 2  19   1   169
## 3  18   1   216
## 4  17   1   165
## 5  16   1   133
## 6  15   1   195
```

Next lets define the column of interest and create a histograma of its distribution.

```
barley_yield = barley$yield
hist(barley_yield)
```



Remember, we need to know the standard deviation in order to calculate the standard error of the sample mean.

```
barley_sd = sd(barley_yield)
```

What would the standard error of the sample mean be if we took two samples?

```
barley_se = barley_sd/sqrt(2)
barley_se
```

```
## [1] 22.01951
```

The standard error woud be about 22.0.

Try calculating additional standard errors. Here are some of the values you should get.

4 samples, se=15.6 5 samples, se=13.9 7 samples, se=11.8

Practice: Cotton

Calculate the standard errors of sample means from the "cotton_uniformity.csv" dataset, for sets of 3, 4, and 6 samples. You should get:

3 samples = 0.13 4 samples = 0.11 6 samples = 0.09