# Gettysburg simulation

```
library(readxl)
library(infer)
library(tidyverse)
library(broom)

set.seed(43)
gettysburg <- read_excel("./data/gettysburg.xlsx")
glimpse(gettysburg)
```

```
## Rows: 268
## Columns: 6
## $ ID     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
## $ Word   <chr> "Four", "score", "and", "seven", "years", "ago", "our", "father~
## $ Length <dbl> 4, 5, 3, 5, 5, 3, 3, 7, 7, 5, 4, 4, 9, 1, 3, 6, 9, 2, 7, 3, 9, ~
## $ Short  <chr> "no", "no", "yes", "no", "no", "yes", "yes", "no", "no", "no", ~
## $ HasE   <chr> "No", "Yes", "No", "Yes", "Yes", "No", "No", "Yes", "No", "No",~
## $ IsNoun <chr> "No", "Yes", "No", "No", "Yes", "No", "No", "Yes", "No", "No", ~
```

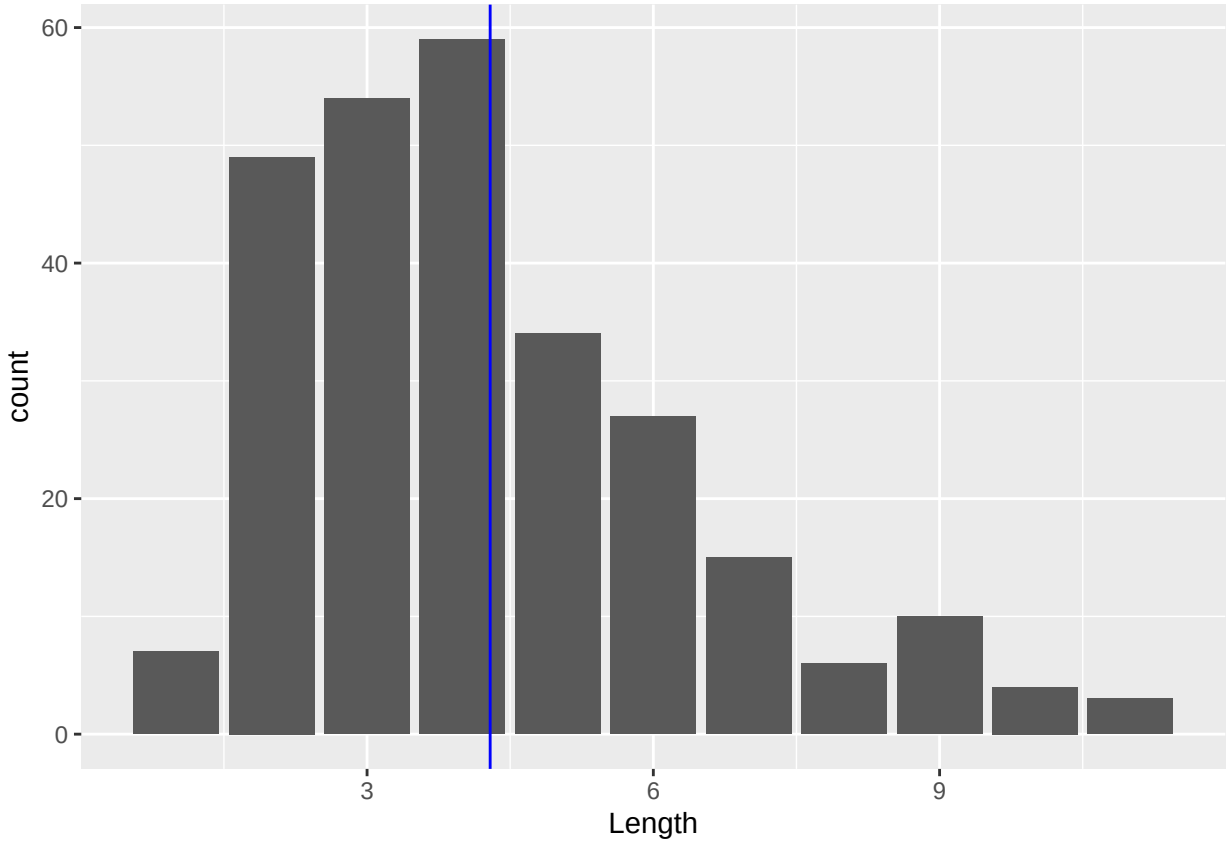
We can confirm that the population mean is 4.29.

```
gettysburg %>%
  summarise(mu = mean(Length))
```

```
## # A tibble: 1 x 1
##      mu
##   <dbl>
## 1  4.29
```

The plot below shows the distribution of length of words in the population. This is the *population distribution*.

```
ggplot(data = gettysburg, aes(x = Length)) +
  geom_bar() +
  geom_vline(xintercept = 4.29, color = "blue")
```

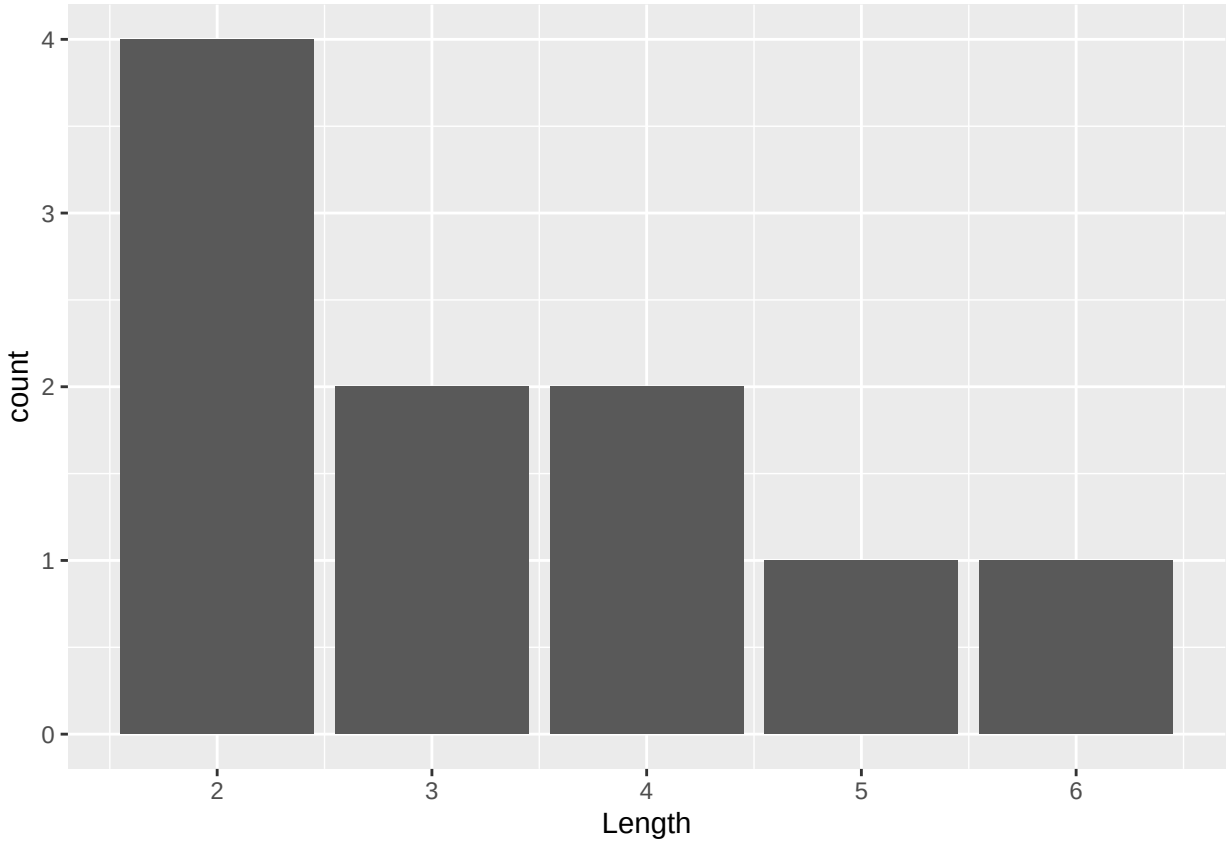



We can select a random sample of 10 words and plot the distribution of word lengths in our sample.

```
sample <- gettysburg %>%
  sample_n(10)
sample
```

```
## # A tibble: 10 x 6
##       ID Word   Length Short HasE  IsNoun
##    <dbl> <chr>   <dbl> <chr> <chr> <chr>
##  1    44 or          2 yes   No    No
##  2   196 to          2 yes   No    No
##  3   149 we          2 yes   Yes   No
##  4    66 have        4 no    Yes   No
##  5   261 the         3 yes   Yes   No
##  6   216 which       5 no    No    No
##  7     7 our         3 yes   No    No
##  8   167 rather      6 no    Yes   No
##  9   130 it          2 yes   No    No
## 10   225 that        4 no    No    No
```

```
ggplot(data = sample, aes(x = Length)) +
  geom_bar()
```




```
sample %>%
  summarise(avg_length = mean(Length))
```

```
## # A tibble: 1 x 1
##   avg_length
##        <dbl>
## 1        3.3
```

We then select 10000 random samples of 10 words...

```
sims <- rep_sample_n(gettysburg, size = 10,
                             reps = 10000, replace = TRUE)
sims %>% print(n = 15)
```

```
## # A tibble: 100,000 x 7
## # Groups:   replicate [10,000]
##    replicate    ID Word     Length Short HasE  IsNoun
##        <int> <dbl> <chr>     <dbl> <chr> <chr> <chr>
##  1         1   180 here          4 no    Yes   No
##  2         1   147 remember      8 no    Yes   No
##  3         1   173 the           3 yes   Yes   No
##  4         1    22 to            2 yes   No    No
##  5         1    78 resting       7 no    Yes   No
##  6         1   228 highly        6 no    No    No
##  7         1   259 people        6 no    Yes   Yes
##  8         1   110 dedicate      8 no    Yes   No
##  9         1   161 It            2 yes   No    No
## 10         1   142 will          4 no    No    No
```

```
## 11         2   216 which         5 no    No    No
## 12         2    72 of            2 yes   No    No
## 13         2   196 to            2 yes   No    No
## 14         2    39 war           3 yes   No    Yes
## 15         2   249 of            2 yes   No    No
## # ... with 99,985 more rows
## # i Use 'print(n = ...)' to see more rows
```

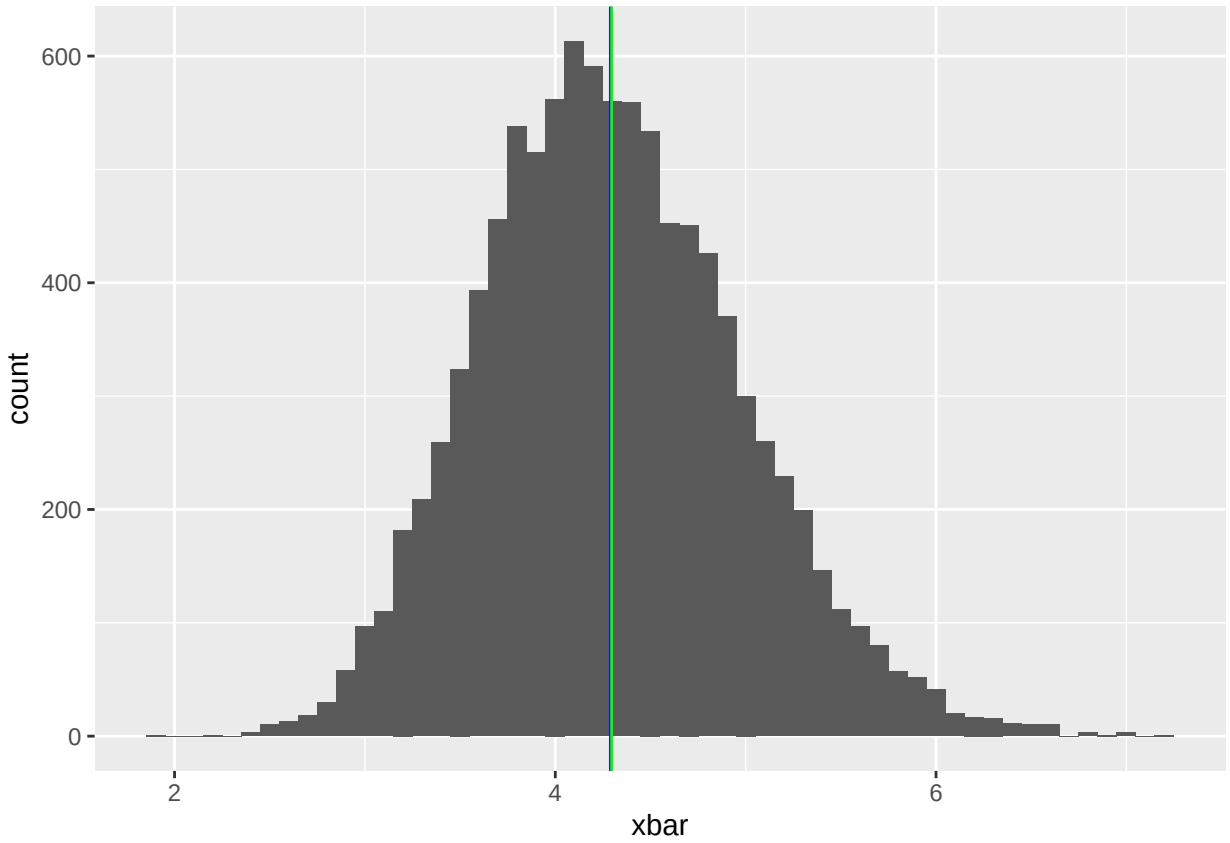
... and compute the mean word length in each sample

```
sample_means <- sims %>%
  group_by(replicate) %>%
  summarise(xbar = mean(Length))

sample_means %>% print(n = 15)
```

```
## # A tibble: 10,000 x 2
##    replicate  xbar
##        <int> <dbl>
##  1         1   5
##  2         2   3.4
##  3         3   4.6
##  4         4   5
##  5         5   4.2
##  6         6   3.8
##  7         7   4.3
##  8         8   3.9
##  9         9   3.8
## 10        10   3.1
## 11        11   5.2
## 12        12   4.6
## 13        13   4.7
## 14        14   4.1
## 15        15   4.6
## # ... with 9,985 more rows
## # i Use 'print(n = ...)' to see more rows
```

We can then plot the *sampling distribution* of the sample mean.

```
Exbar <- mean(sample_means$xbar)

ggplot(data = sample_means, aes(x = xbar)) +
  geom_histogram(binwidth = 0.1) +
  geom_vline(xintercept = 4.29, color = "blue") +
  geom_vline(xintercept = Exbar, color = "green")
```


600
400
200
0
count
2
4
6
xbar