Sample Statistics ($\overline{x}$'s) have a Distribution Too I

My personal sample of $n = 10$ words

```
mySample
```

```
 [1] "endure"  "have"    "which"   "testing" "world"
 [6] "we"      "perish"  "poor"    "never"   "detract"
```

The lengths of my $n = 10$ words:

```
 endure    have   which testing   world      we  perish
      6       4       5       7       5       2       6
   poor   never detract
      4       5       7
```

Average length of my sample of $n = 10$ words:

```
myxbar <- mean(mySampleWordLen)
myxbar
```

```
[1] 5.1
```

Sample Statistics ($\overline{x}$'s) have a Distribution Too II

My personal sample of $n = 10$ words

```
mySample
```

```
 [1] "endure"  "have"    "which"   "testing" "world"
 [6] "we"      "perish"  "poor"    "never"   "detract"
```

How many of my words contain the letter e?

```
[1] "endure"  "have"    "testing" "we"      "perish"
[6] "never"   "detract"
```

```
[1] 7
```

What proportion of my words contain the letter e?

```
myphat <- length(grep("e", mySample)) / length(mySample)
myphat
```

```
[1] 0.7
```

Sample Statistics ($\overline{x}$'s) have a Distribution Too III

```
humanSampleMeans <- c(6.9, 8.4, 6.4, 6.7, 6.6, 6.8, 5.5, 5.1,
    5.9, 8.3, 6.1, 8.2, 5.1, 6.1, 7.6, 6.7, 7.3, 5.6, 8.7, 7.1,
    6.1, 6.2, 6)
```

How many sample means (xbars)?

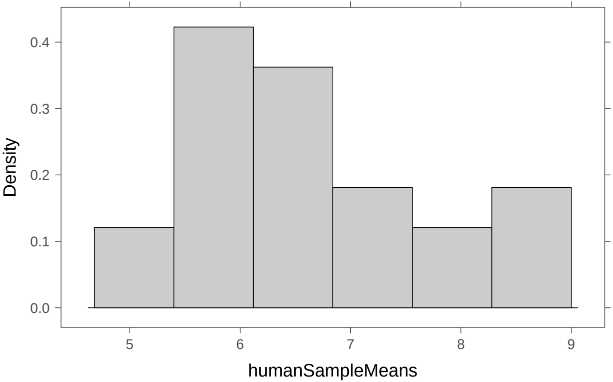
```
[1] 23
```

```
stem(humanSampleMeans, scale=2)
```

```
  The decimal point is at the |

  5 | 11
  5 | 569
  6 | 011124
  6 | 67789
  7 | 13
  7 | 6
  8 | 234
  8 | 7
```

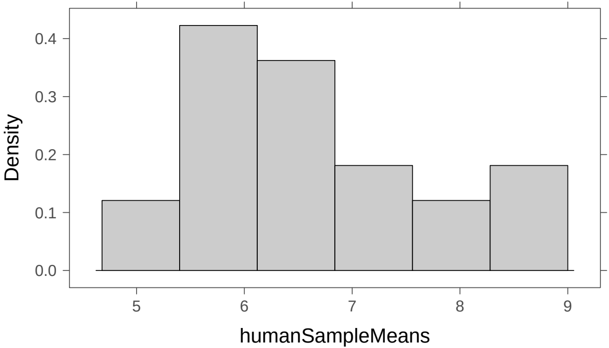
Sample Statistics ($\overline{x}$'s) have a Distribution Too IV

```
histogram(~ humanSampleMeans)
```



Sample Statistics ($\overline{x}$'s) have a Distribution Too V

What is the mean length (xbar) "on average" for your 23 samples?



```
mean(humanSampleMeans)
```

```
[1] 6.67
```

Sample Statistics ($\overline{x}$'s) have a Distribution Too VI

```
worddata <- read.csv("../data/address.csv")
```

What is stored in the data did we just read in?

How many words are in the Gettysburg Address?

```
glimpse(worddata)
```

```
Observations: 268
Variables: 4
$ id       (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,...
$ word     (fctr) Four, score, and, seven, years, ago, ...
$ wordlen  (int) 4, 5, 3, 5, 5, 3, 3, 7, 7, 5, 4, 4, 9,...
$ containE (fctr) No, Yes, No, Yes, Yes, No, No, Yes, N...
```

What is actual average length of all 268 words in the Address?

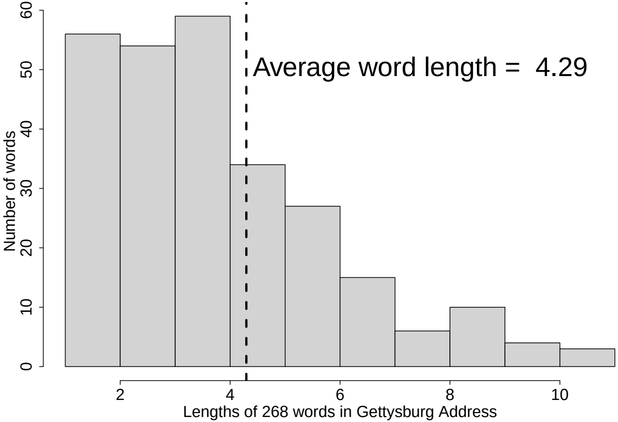
```
mean(wordlen, data=worddata)
```

```
[1] 4.295
```

What symbol do we use to denote this mean?

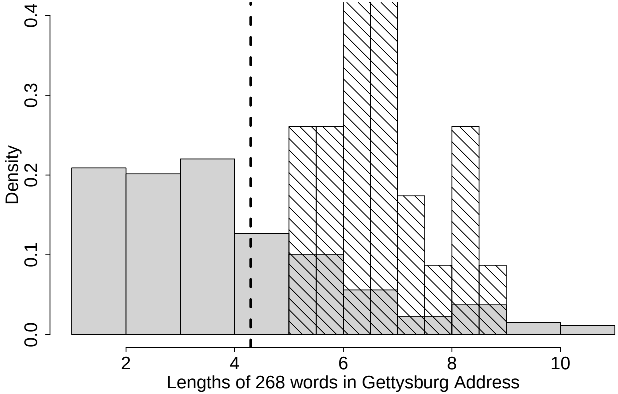
Sample Statistics ($\overline{x}$'s) have a Distribution Too VII

A histogram of the lengths of all 268 words



Sample Statistics ($\overline{x}$'s) have a Distribution Too VIII

How does the original population of word lengths compare with the 23 average lengths (xbars) of 10 human-chosen words?



Our sample averages (xbars) tend to overestimate the true average ($\mu = 4.29$). This is evidence of **bias** in our estimation method.

Sample Statistics ($\overline{x}$'s) have a Distribution Too IX

What is the actual proportion of all 268 words that contain an "e"?

```
    No    Yes
0.5336 0.4664
```

```
p
```

```
[1] 0.4664
```

This is a population parameter labeled p (sometimes π).

How many of you had a sample proportion ($\widehat{p}$)
higher than the true value?
The population parameter $p = 0.466$.

Is this evidence of **bias** in our estimation method?

Sample Statistics ($\overline{x}$'s) have a Distribution Too X

Now, randomly sample just 5 words from the Address using the table of random digits on the back side of the handout.

Just pick any spot to start reading in the table. Read upwards, to the right, left, diagonally, whatever.

For example, Row 7, column 8 reading left to right...

```
 59   136   85   175   258
```

These random numbers correspond to the words...

```
 a   to   their   work   the
```

With word lengths... `1 2 5 4 3`

and average $= \overline{x} = 3$ and proportion with "e" $= \widehat{p} = 2/5 = 0.40$.

Oops! My estimate is too low since $\mu = 4.29$
Did I do something wrong? Is random sampling also biased?

Sample Statistics ($\bar{x}$'s) have a Distribution Too XI

Your averages (xbars) from 5 randomly-chosen words

```
humanRandomMeans <- c(4.8, 4.8, 4.6, 3.8, 3.2, 4.2, 3.4, 4.2,
    2.4, 4.2, 5.1, 5.2, 4.8, 4.8, 6, 3, 3.2, 4.6, 5, 4)
```

How many sample means (xbars)?

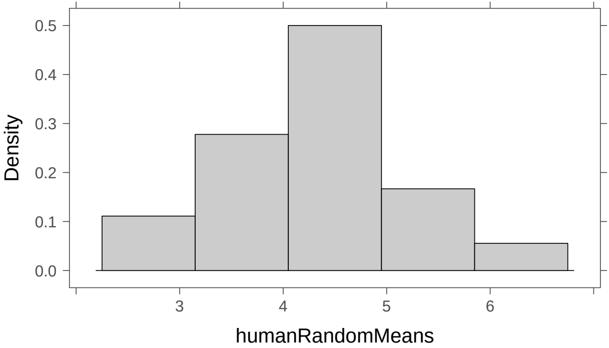
```
[1] 20
```

```
stem(humanRandomMeans)
```

```

  The decimal point is at the |

  2 | 4
  3 | 02248
  4 | 0222668888
  5 | 012
  6 | 0
```

Sample Statistics ($\overline{x}$'s) have a Distribution Too XII

```
histogram(~ humanRandomMeans)
```



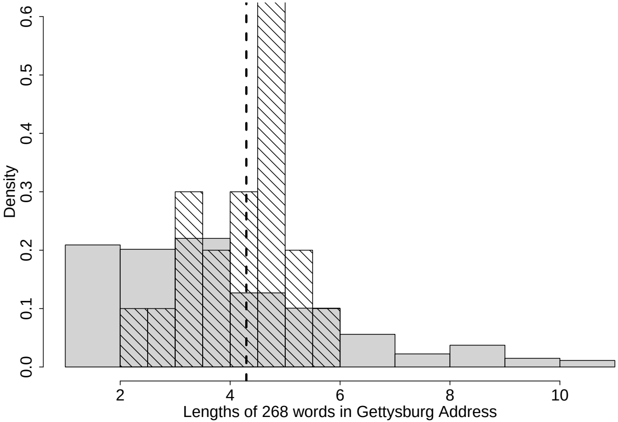
Sample Statistics ($\overline{x}$'s) have a Distribution Too XIII

What is the mean length (xbar) “on average” for your 20 samples?
What is the mean length ”on average” for your samples of 5 “random” words vs. 10 “representative” words?

```
mean(humanRandomMeans)

[1] 4.265

mean(humanSampleMeans)

[1] 6.67

mu

[1] 4.295
```

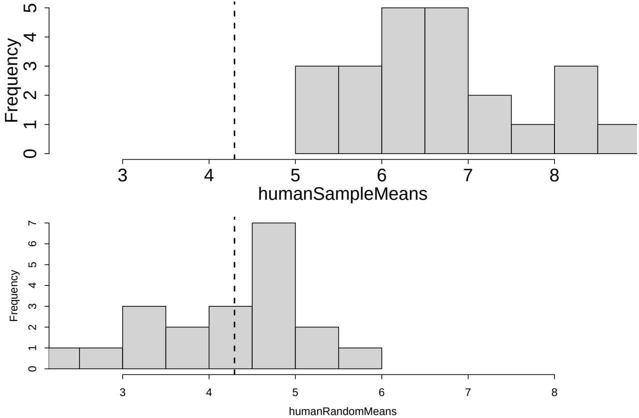
Sample Statistics ($\overline{x}$'s) have a Distribution Too XIV

How does the original population of word lengths compare with the 20 average lengths (xbars) from $n = 5$ randomly-chosen words?



Sample Statistics ($\overline{x}$'s) have a Distribution Too XV

How do the averages from the "representative" samples of $n = 10$ compare with the random samples of n=5?



Sample Statistics ($\overline{x}$'s) have a Distribution Too XVI

Let's let R randomly sample 5 words from the Gettysburg Address and record their average length (xbar).

Repeat this 500 times.

Will all of the 500 sample averages be the same?

Sample Statistics ($\overline{x}$'s) have a Distribution Too XVII

To get started, look at a couple of samples and their means

```
sample1 <- sample(1:268,5);   sample1

[1] 214 202 105  91  96

word[sample1]

[1] cause us    a     live  and
144 Levels: a above add advanced ago all altogether ... years

wordlen[sample1]

[1] 5 2 1 4 3

mean(wordlen[sample1])

[1] 3
```

Sample Statistics ($\overline{x}$'s) have a Distribution Too XVIII

```
sample2 <- sample(1:268,5);   sample2

[1]  54 143  26 262  45

word[sample2]

[1] endure little all    people any
144 Levels: a above add advanced ago all altogether ... years

wordlen[sample2]

[1] 6 6 3 6 3

mean(wordlen[sample2])

[1] 4.8
```

Sample Statistics ($\overline{x}$'s) have a Distribution Too XIX

```
mean(wordlen[sample1])

[1] 3

mean(wordlen[sample2])

[1] 4.8

mu

[1] 4.295
```

Sample Statistics ($\overline{x}$'s) have a Distribution Too XX

Now, let's repeat the random sampling a few times

```
replicate(10, wordlen[sample(1:268,5)] )
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    4    3    5    4    6    1    4    6    9     4
[2,]    3    4    7    2    4    2    5    4    6     9
[3,]    4    2    2    2    2    3    2    6    3     4
[4,]    4    2    4    9    2   11    4    2    5     3
[5,]    3    6    2    4    4    4    4    1    8     6
```

```
replicate(10, mean(wordlen[sample(1:268,5)]) )
```

```
 [1] 3.6 3.4 4.0 4.2 3.6 4.2 3.8 3.8 6.2 5.2
```

Sample Statistics ($\overline{x}$'s) have a Distribution Too XXI

Let's repeat the random sampling 500 times

```
randomSampleMeans = replicate(500, mean(wordlen[sample(1:268,5)]
sort(randomSampleMeans[1:20])
```

```
 [1] 2.8 3.4 3.6 3.6 3.8 3.8 4.0 4.2 4.2 4.2 4.2 4.2 4.4 4.4
[15] 4.8 5.0 5.2 5.2 6.2 6.2
```

```
mu
```

```
[1] 4.295
```

```
sort(randomSampleMeans[1:20] - mu)
```

```
 [1] -1.49478 -0.89478 -0.69478 -0.69478 -0.49478 -0.49478
 [7] -0.29478 -0.09478 -0.09478 -0.09478 -0.09478 -0.09478
[13]  0.10522  0.10522  0.50522  0.70522  0.90522  0.90522
[19]  1.90522  1.90522
```

Sample Statistics ($\overline{x}$'s) have a Distribution Too XXII

What is the average lenght (xbar) "on average" for many, many ($M = 500$) samples each with $n = 5$ randomly chosen words?

```
mean(randomSampleMeans)
```

```
[1] 4.251
```

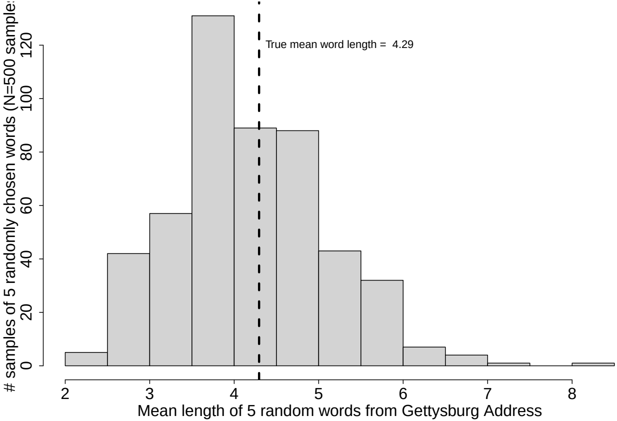
If this "mean of the averages" is close to the true mean we say that the statistic ($\overline{x}$) is an **unbiased** statistic (estimator) for the parameter (μ).

```
mu
```

```
[1] 4.295
```

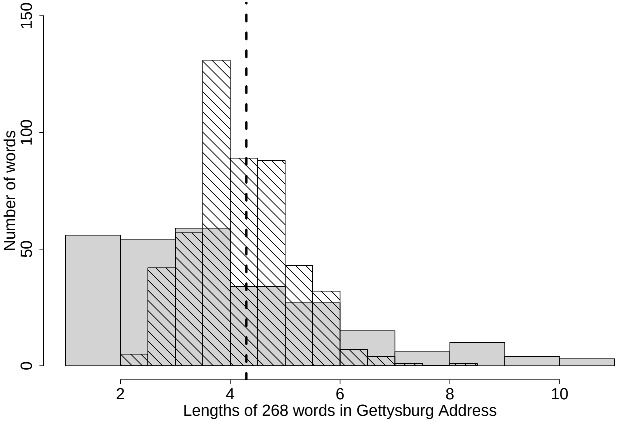
Sample Statistics ($\overline{x}$'s) have a Distribution Too XXIII

Histogram of the average lengths ($n = 5$)



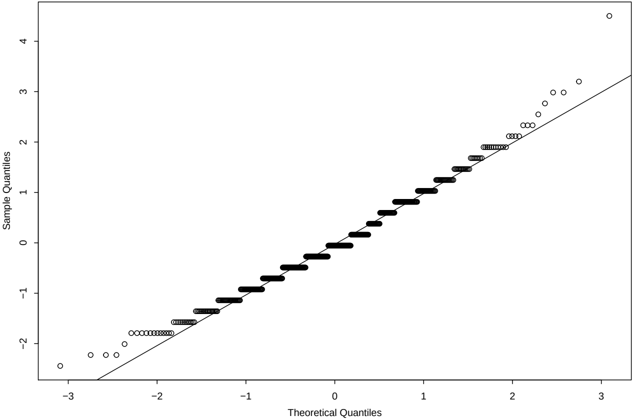
Sample Statistics ($\overline{x}$'s) have a Distribution Too XXIV

How does the original population of word lengths compare with the $M = 500$ avreage lengths (xbars) of $n = 5$ randomly chosen words?



Sample Statistics ($\overline{x}$'s) have a Distribution Too XXV

Can the distribution of xbars be well-approximated by a normal density? Standardize the averages



Sample Statistics ($\overline{x}$'s) have a Distribution Too XXVI

Let's randomly sample $n = 15$ words instead of 5

Let's repeat the random sampling 500 times

```
randomSampleMeans.15 = replicate(500, mean(wordlen[sample(1:268,
sort(randomSampleMeans.15[1:20])
```

```
 [1] 3.533 3.600 3.800 3.800 3.867 3.933 4.000 4.067 4.067
[10] 4.133 4.133 4.200 4.267 4.267 4.267 4.733 5.000 5.000
[19] 5.133 5.400
```

```
mu
```

```
[1] 4.295
```

```
sort(randomSampleMeans.15[1:20] - mu)
```

```
 [1] -0.76144 -0.69478 -0.49478 -0.49478 -0.42811 -0.36144
 [7] -0.29478 -0.22811 -0.22811 -0.16144 -0.16144 -0.09478
[13] -0.02811 -0.02811 -0.02811  0.43856  0.70522  0.70522
[19]  0.83856  1.10522
```

Sample Statistics ($\overline{x}$'s) have a Distribution Too XXVII

What is the mean length "on average" for many, many ($M = 500$) samples of $n = 15$ randomly chosen words?

```
mean(randomSampleMeans.15)
```

```
[1] 4.287
```

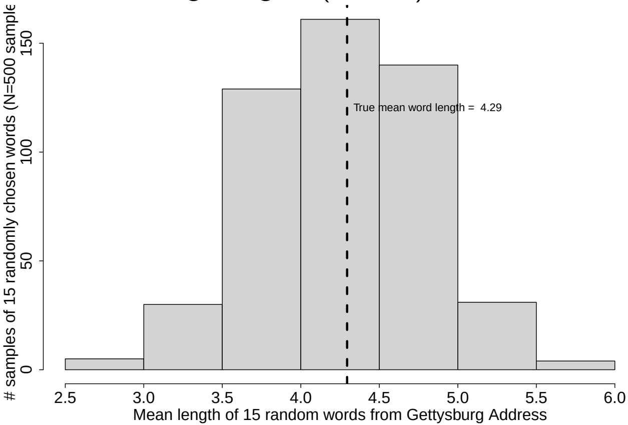
If this "mean of the averages" is close to the true mean we say that the statistic ($\overline{x}$) is an **unbiased** statistic (estimator) for the parameter (μ).

```
mu
```

```
[1] 4.295
```

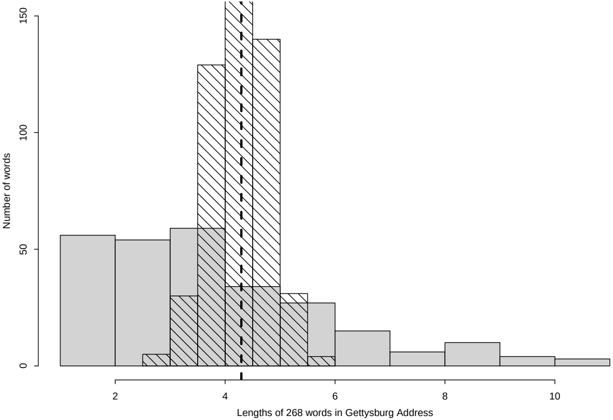
Sample Statistics ($\overline{x}$'s) have a Distribution Too XXVIII

Histogram of the average lengths ($n = 15$)



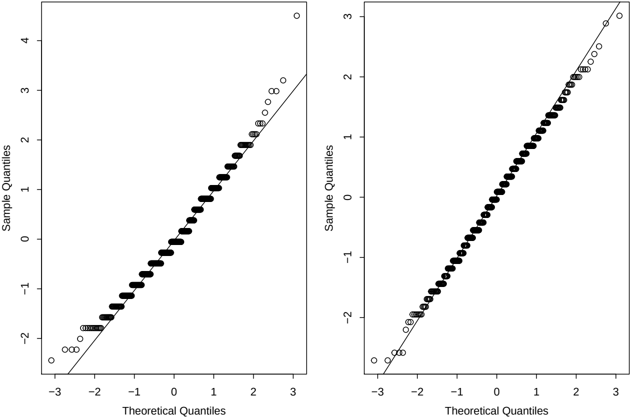
Sample Statistics ($\overline{x}$'s) have a Distribution Too XXIX

How does the original population of word lengths compare with the $M == 500$ mean lengths of $n = 15$ randomly chosen words?



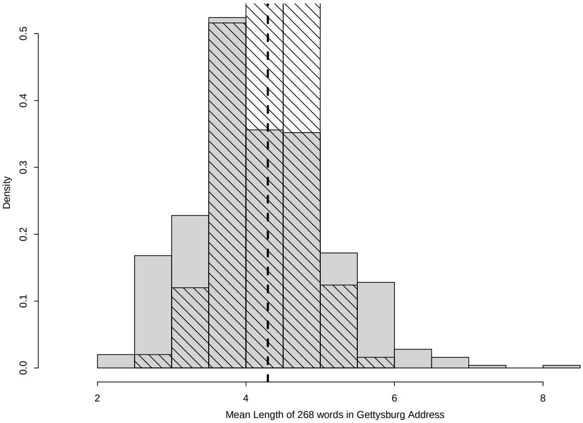
Sample Statistics ($\overline{x}$'s) have a Distribution Too XXX

Can the distribution of xbars be well-approximated by a normal density? Standardize the averages



Sample Statistics ($\overline{x}$'s) have a Distribution Too XXXI

How do the means from the random samples with 15 words compare with the $M = 500$ mean lengths of 5 randomly chosen words?



Sample Statistics ($\overline{x}$'s) have a Distribution Too XXXII

```
favstats(randomSampleMeans)
```

```
 min  Q1 median   Q3 max  mean     sd   n missing
   2 3.6    4.2 4.85 8.4 4.251 0.9216 500       0
```

```
favstats(randomSampleMeans.15)
```

```
   min    Q1 median    Q3   max  mean     sd   n missing
 2.867 3.933    4.3 4.667 5.867 4.287 0.5239 500       0
```

```
mu
```

```
[1] 4.295
```

```
sd(wordlen, data=worddata)
```

```
[1] 2.123
```