Lecture 2: Sampling Distibution

Kushal K Dey

04.05.2016

Game of Words

Loading the Data

```
library(devtools)
install_github("kkdey/GOTnames")
```

```
Downloading GitHub repo kkdey/GOTnames@master
Installing GOTnames
'/Library/Frameworks/R.framework/Resources/bin/R'  \
  --no-site-file --no-environ --no-save --no-restore CMD
\
  INSTALL  \

'/private/var/folders/0f/v6kp3_hj7rd2wrhms9mf4h500000gn/T/RtmplJ
\

--library='/Library/Frameworks/R.framework/Versions/3.2/Resource
\
  --install-tests
```

```
library(GOTnames)
data(GOTnames)
```

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

My personal sample of $n = 8$ words

```
mySample
```

```
[1] "Eddard Stark"  "Sansa Stark"   "Robb Stark"
[4] "Arya Stark"    "Benjen Stark"  "Catelyn Stark"
[7] "Bran Stark"    "Jon Snow"
```

The lengths of my $n = 8$ words:

```
 Eddard Stark   Sansa Stark    Robb Stark    Arya Stark
           12             11            10            10
 Benjen Stark Catelyn Stark    Bran Stark      Jon Snow
           12             13            10             8
```

Average length of my sample of $n = 8$ words:

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

```
[1] 10.75
```

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

My personal sample of $n = 8$ words

```
mySample
```

```
[1] "Eddard Stark"  "Sansa Stark"   "Robb Stark"
[4] "Arya Stark"    "Benjen Stark"  "Catelyn Stark"
[7] "Bran Stark"    "Jon Snow"
```

How many of my words contain the letter "a"?

```
[1] "Eddard Stark"  "Sansa Stark"   "Robb Stark"
[4] "Arya Stark"    "Benjen Stark"  "Catelyn Stark"
[7] "Bran Stark"
```

```
[1] 7
```

What proportion of my words contain the letter "a"?

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

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

```
[1] 0.875
```

```
humanSampleMeans <- c(6.9, 12.4, 13.4, 13.7, 14.6, 7.8, 15.5,
    12.1, 18.9, 16.3, 7.1, 8.2, 8.1, 10.1, 10.6, 7.7, 10.3, 9.6,
    6.7, 11.1, 16.1, 10.2, 12)
```

How many sample means (xbars)?

```
[1] 23
```

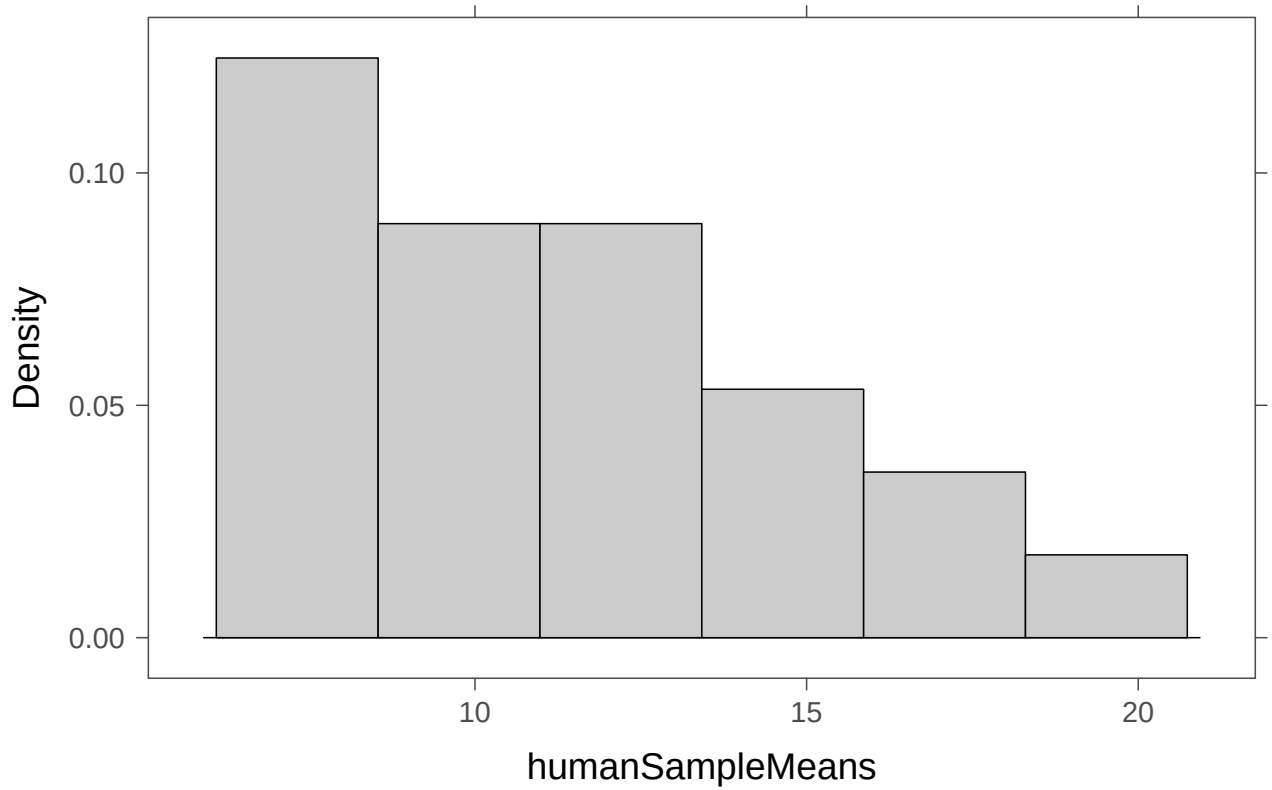
Sample Statistics ($\overline{x}$'s) have a Distribution Too IV

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

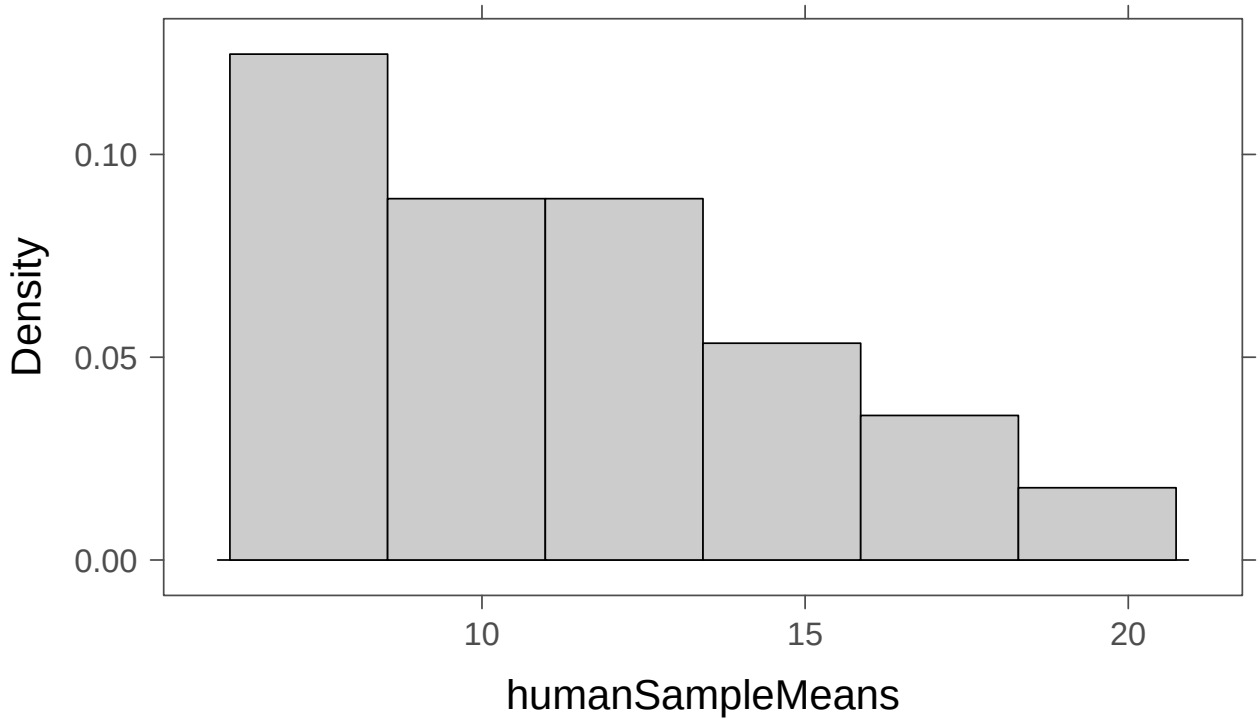
```
  The decimal point is at the |

   6 | 79178
   8 | 126
  10 | 12361
  12 | 01447
  14 | 65
  16 | 13
  18 | 9
```

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

```
histogram(~ humanSampleMeans)
```



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




```
mean(humanSampleMeans)
```

```
[1] 11.28
```

```
worddata <- as.data.frame(GOTnames);
```

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

How many words are in the Game of Thrones characters?

```
glimpse(worddata)
```

```
Observations: 100
Variables: 3
$ x          (fctr)  Tyrion Lannister,  Cersei Lannister...
$ wordlen    (int) 17, 17, 19, 9, 11, 16, 12, 14, 16, 13...
$ A.present (chr) "Yes", "Yes", "Yes", "No", "Yes", "Ye...
```

What is actual average length of all 100 characters in Game of Thrones?

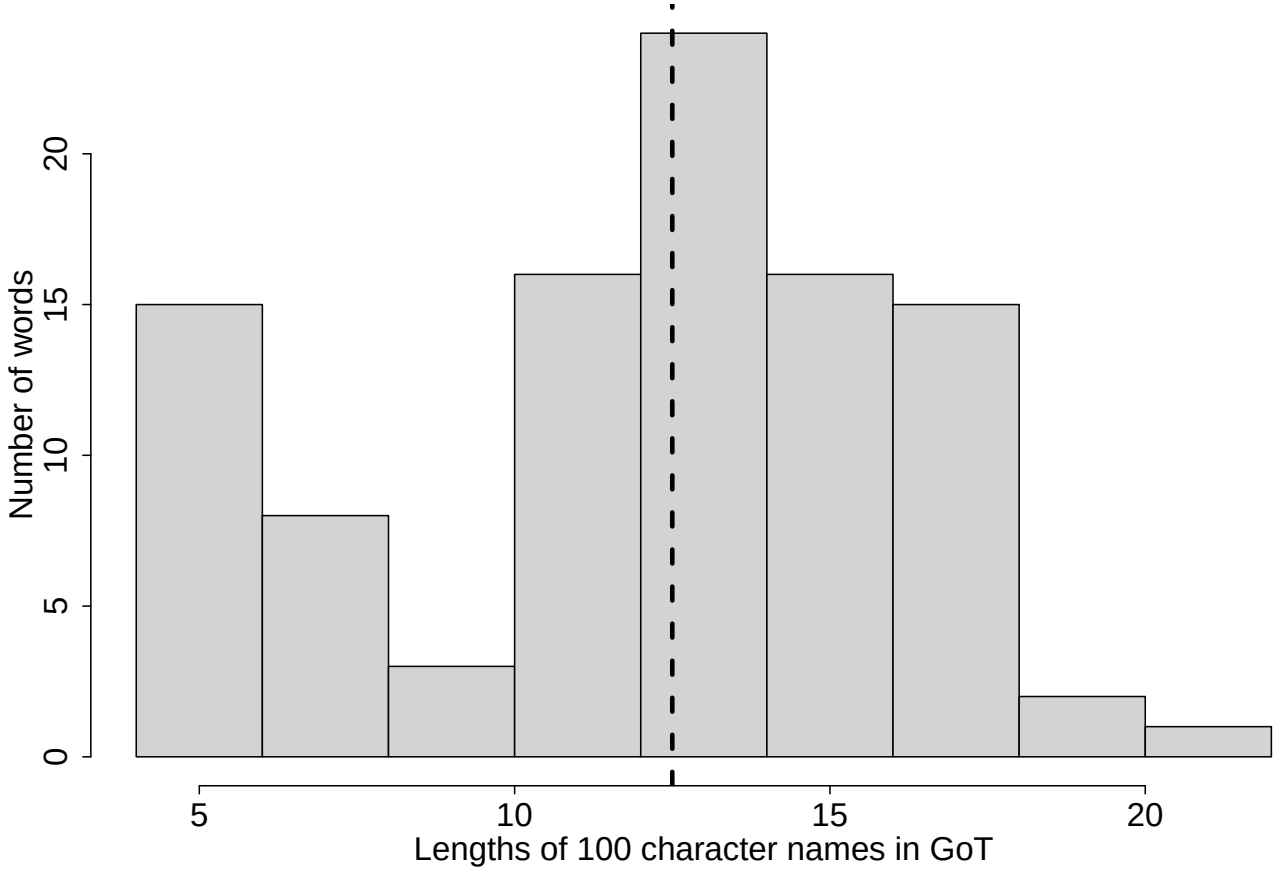
```
mean(wordlen, data=worddata)
```

```
[1] 12.5
```

What symbol do we use to denote this mean?

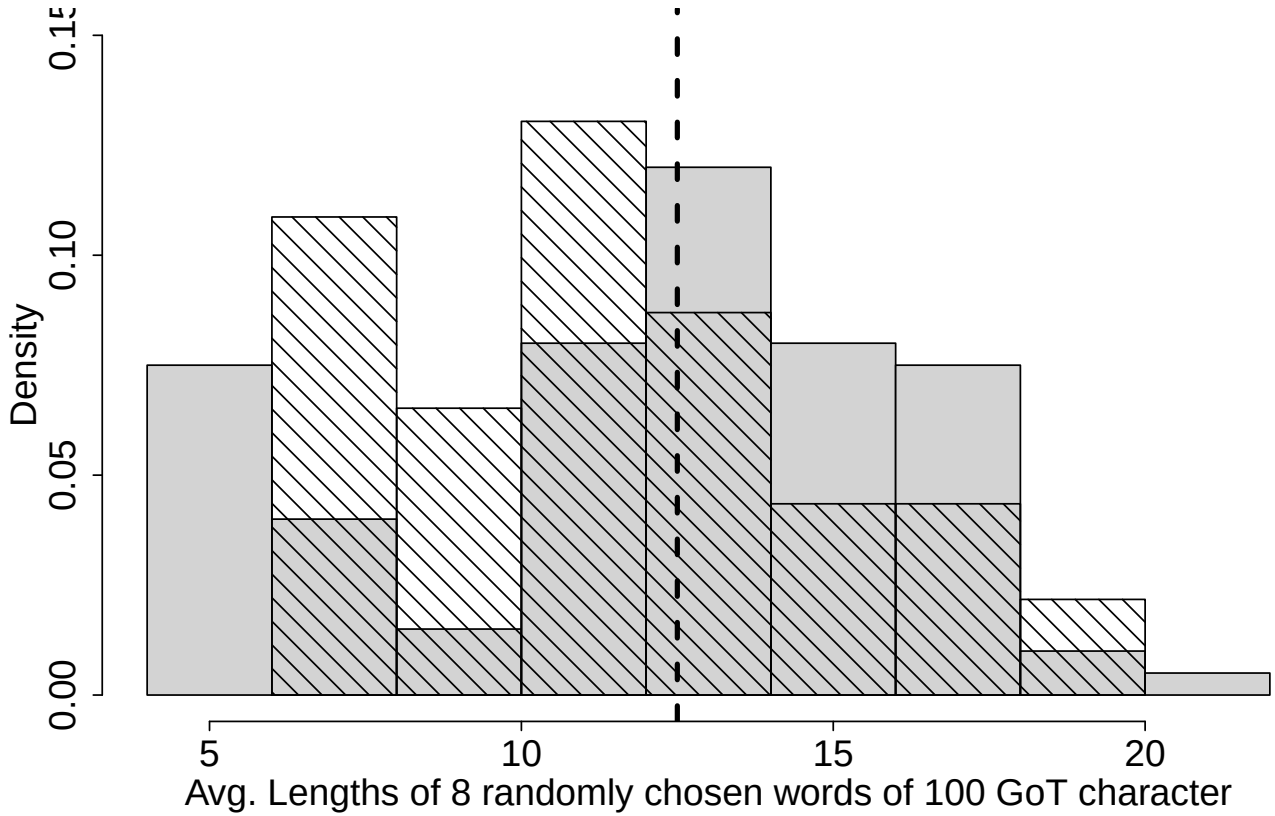
Sample Statistics ($\overline{x}$'s) have a Distribution Too VIII

A histogram of the lengths of all 100 words.



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

How does the original population of word lengths compare with the 23 average lengths (xbars) of 8 human-chosen words?



Our sample averages (xbars) tend to underestimate the true average μ. This is evidence of **bias** in our estimation method.

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

What is the actual proportion of all 100 words that contain an "a"?

```
p
```

```
[1] 0.73
```

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

How many of you had a sample proportion ($\widehat{p}$)
higher than the true value?

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

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

Now, randomly sample just 8 words from the list

Pick a random point in the list and start drawing next 8 characters.

For example, start reading from Grey Worm and go down...

These random numbers correspond to the words...

`Grey Worm, Anguy, Orell, Irri, Craster`
`Mirri Maz Duur, Syrio Forrel, Rakharo`

With word lengths... `10 6 6 5 8 15 13 8`

and average $= \overline{x} = 8.875$ and proportion with "a" $=$ $\widehat{p} = 4/8 = 0.50$.

Oops! My estimate is too low since Did I do something wrong? Is random sampling also biased?
Your averages (xbars) from 8 randomly-chosen words

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

```
humanRandomMeans <- c(11.1, 10.2, 11.8, 12.6, 11.5, 14.6, 11.4,
    13.1, 13.2, 11.5, 12, 11.5, 14.1, 11.5, 11.2, 10.2, 10.9,
    12.5, 13.3, 12.3, 12, 13, 15.1)
```

How many sample means (xbars)?

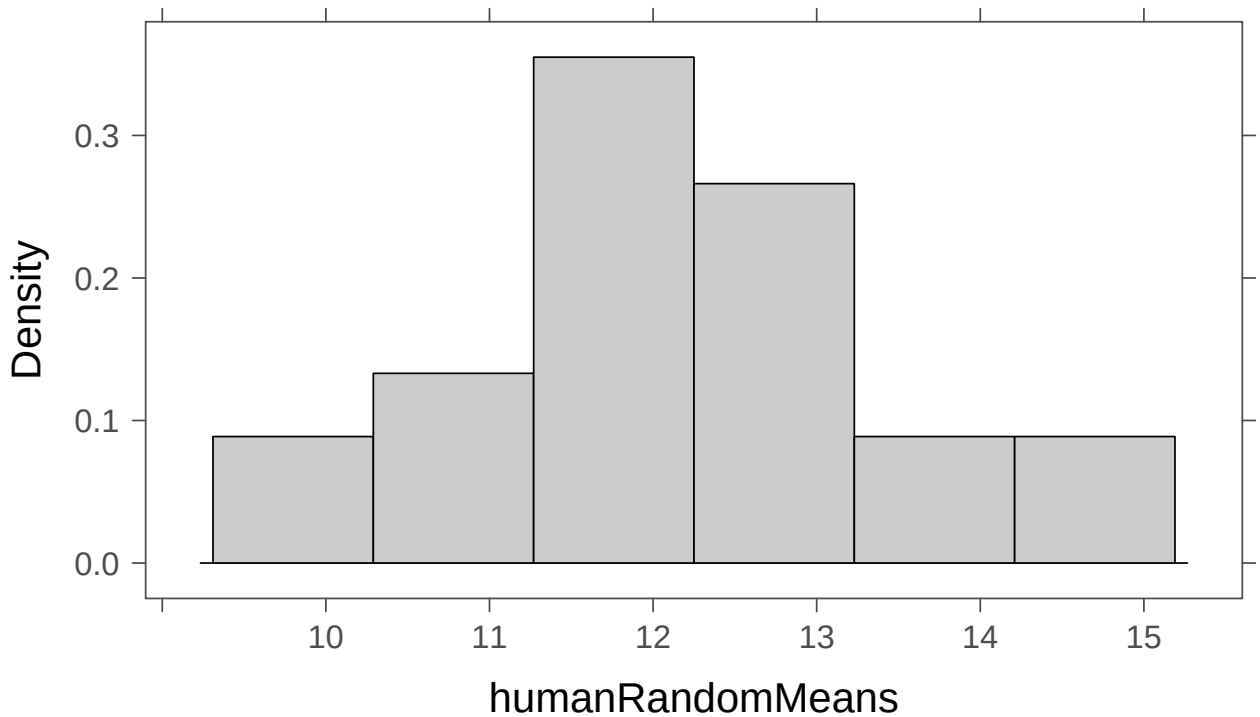
```
[1] 23
```

```
stem(humanRandomMeans)
```

```
  The decimal point is at the |

  10 | 229
  11 | 12455558
  12 | 00356
  13 | 0123
  14 | 16
  15 | 1
```

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

```
histogram(~ humanRandomMeans)
```



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

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

What is the mean length "on average" for your samples of 8 "random" words vs. 8 "representative" words?

```
mean(humanRandomMeans)
```

```
[1] 12.2
```

```
mean(humanSampleMeans)
```

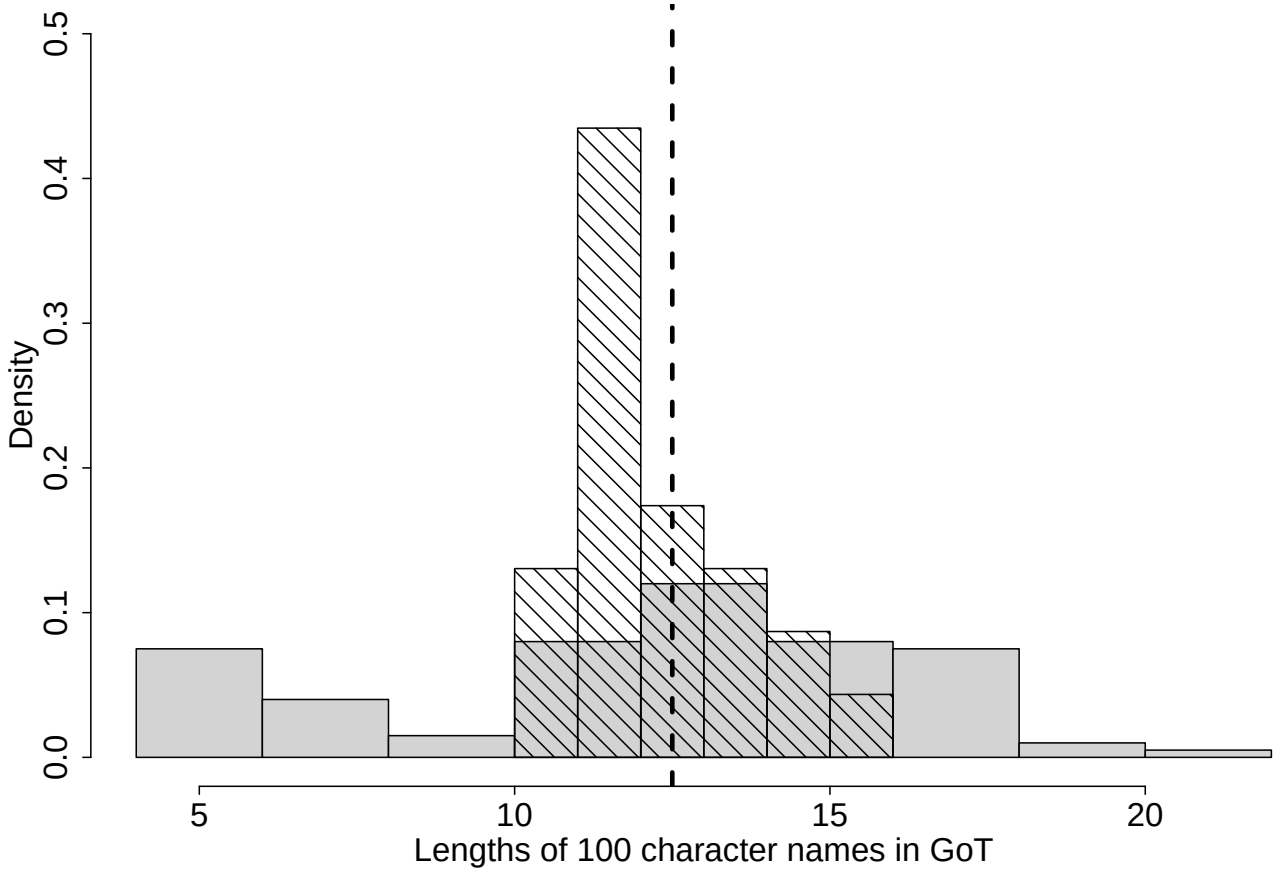
```
[1] 11.28
```

```
mu
```

```
[1] 12.5
```

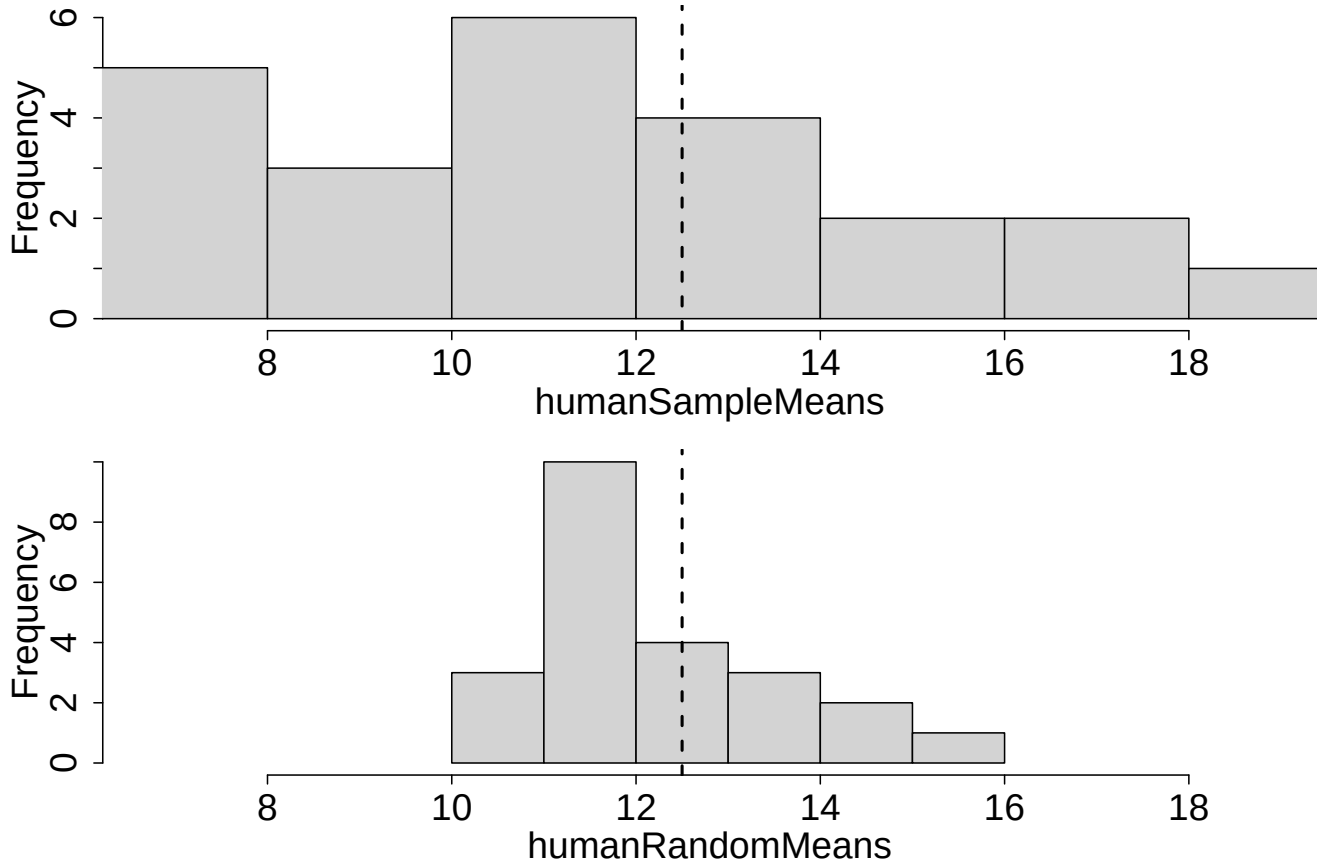
Sample Statistics ($\overline{x}$'s) have a Distribution Too XV

How does the original population of word lengths compare with the 23 average lengths (xbars) from $n = 8$ randomly-chosen words?



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

How do the averages from the "representative" samples of $n = 8$ compare with the random samples of n=8?



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

Let's let R randomly sample 8 words from the list of character names in GoT 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 XVIII

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

```
sample1 <- sample(1:100,8);   sample1

[1] 80 75 39 34 35 19 51  9

x[sample1]

[1]  Spice King         Xaro Xhoan Daxos  Jeor Mormont
[4]  Grenn              Ramsay Snow       Davos Seaworth
[7]  Eddison Tollett    Tywin Lannister
100 Levels:  Alliser Thorne  Alton Lannister ...  Yoren

wordlen[sample1]

[1] 11 17 13  6 12 15 16 16

mean(wordlen[sample1])

[1] 13.25
```

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

```
sample2 <- sample(1:100,8);   sample2

[1] 99 17 79 58 87 84 94 95

x[sample2]

[1]  Balon Greyjoy      Joffrey Baratheon  Qyburn
[4]  Hot Pie            Alton Lannister    Selyse Baratheon
[7]  Syrio Forrel       Rakharo
100 Levels:  Alliser Thorne  Alton Lannister ...  Yoren

wordlen[sample2]

[1] 14 18  7  8 16 17 13  8

mean(wordlen[sample2])

[1] 12.62
```

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

```
mean(wordlen[sample1])

[1] 13.25

mean(wordlen[sample2])

[1] 12.62

mu

[1] 12.5
```

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

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

```
replicate(10, wordlen[sample(1:100,8)] )
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]   16   16   13    6   15    7   16   16   17     6
[2,]   16   17   14   13   13   15   11   14   14     6
[3,]   17    9    6   15   14    5    8   15   15    11
[4,]   14   14   15    7   14   12    5    6   14    17
[5,]    4   12    4   14   17    8   17   12    8    15
[6,]   11   17   10   11   14   17   15    8    6    18
[7,]   17   14    5   17    6   16   15   15    6    19
[8,]   19    4   14   17   14   11   11   12   19    14
```

```
replicate(10, mean(wordlen[sample(1:100,8)]) )
```

```
 [1] 14.25 12.88 10.12 12.50 13.38 11.38 12.25 12.25 12.38
[10] 13.25
```

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

Let's repeat the random sampling 500 times

```
randomSampleMeans = replicate(500, mean(wordlen[sample(1:100,8)]))
sort(randomSampleMeans[1:20])
```

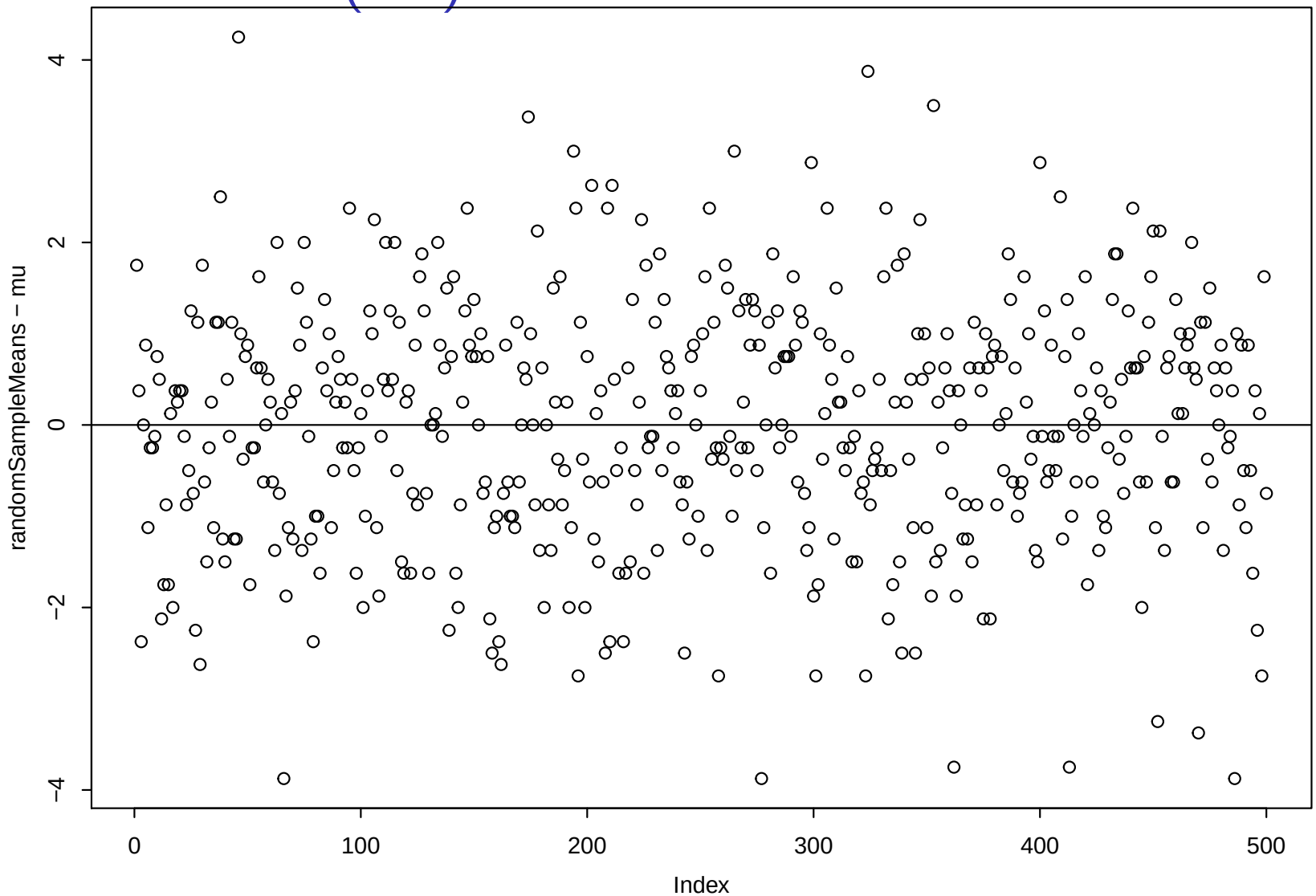
```
 [1] 10.12 10.38 10.50 10.75 10.75 11.38 11.62 12.25 12.25
[10] 12.38 12.50 12.62 12.75 12.88 12.88 12.88 13.00 13.25
[19] 13.38 14.25
```

```
mu
```

```
[1] 12.5
```

```
plot(randomSampleMeans - mu)
abline(h=0)
```

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



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

What is the average lenght (xbar) "on average" for many, many ($M = 500$) samples each with $n = 8$ randomly chosen words?

```
mean(randomSampleMeans)
```

```
[1] 12.48
```

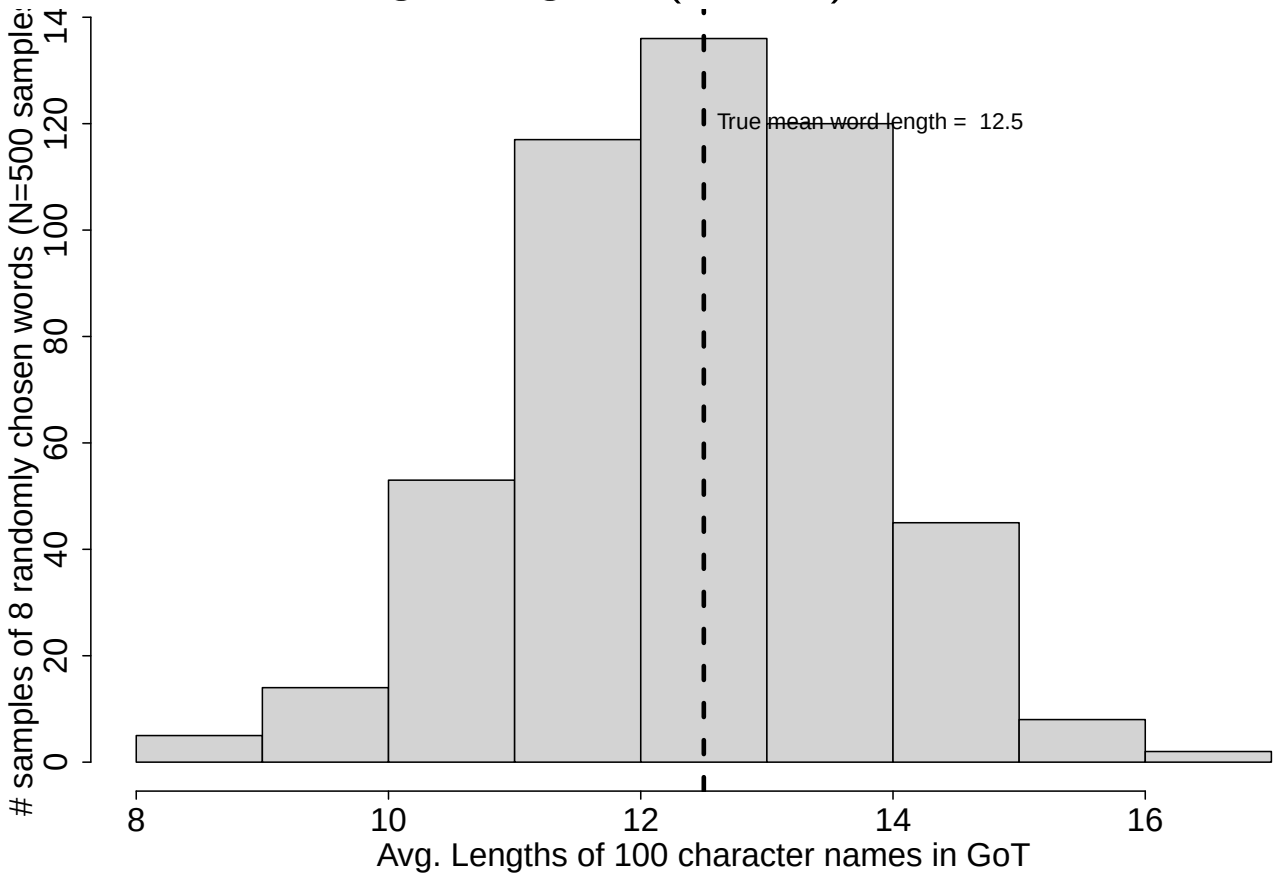
If this "mean of the averages" is close to the true mean we say that the statistic ($\overline{x}$) is an **unbiased** statistic (estimator) for the parameter (μ).

```
mu
```

```
[1] 12.5
```

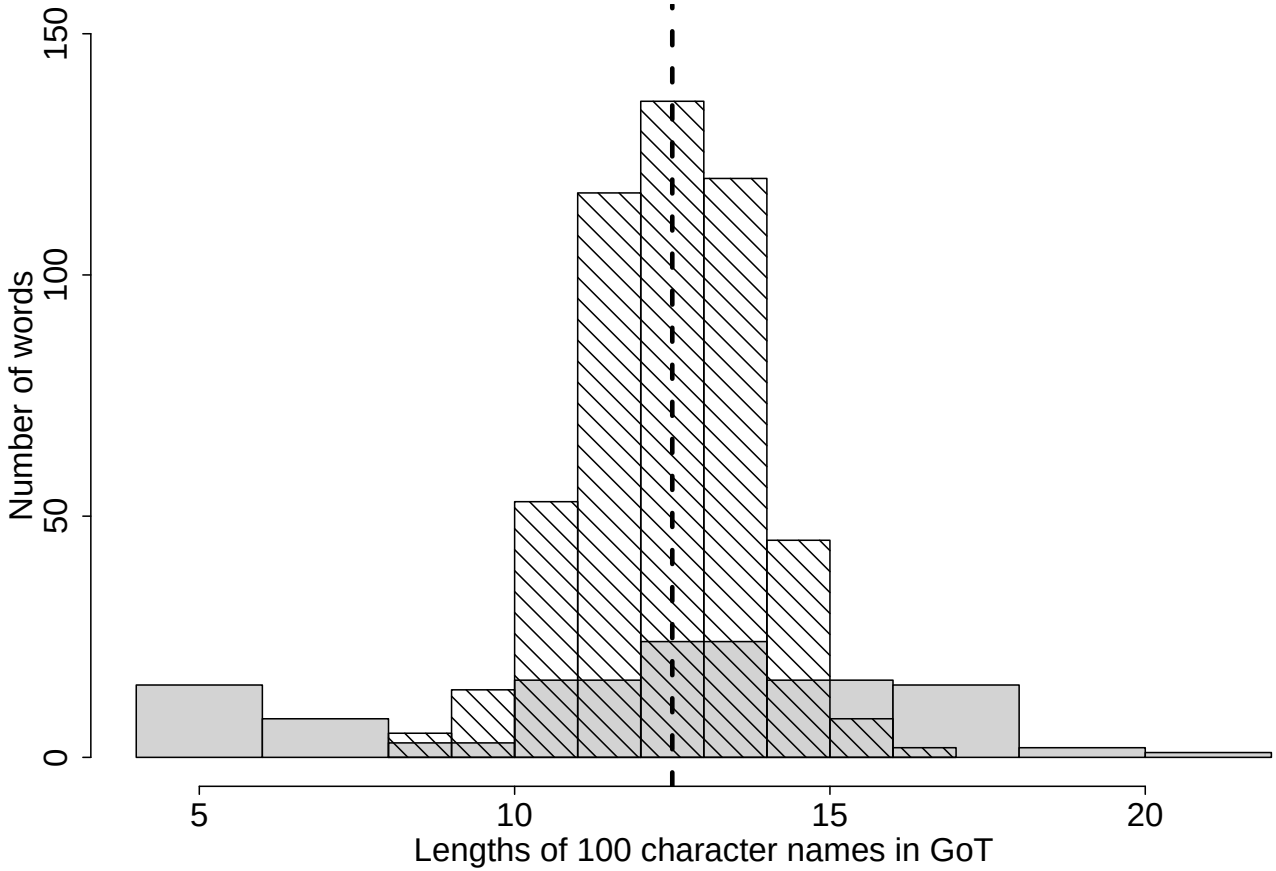
Sample Statistics ($\overline{x}$'s) have a Distribution Too XXV

Histogram of the average lengths ($n = 8$)



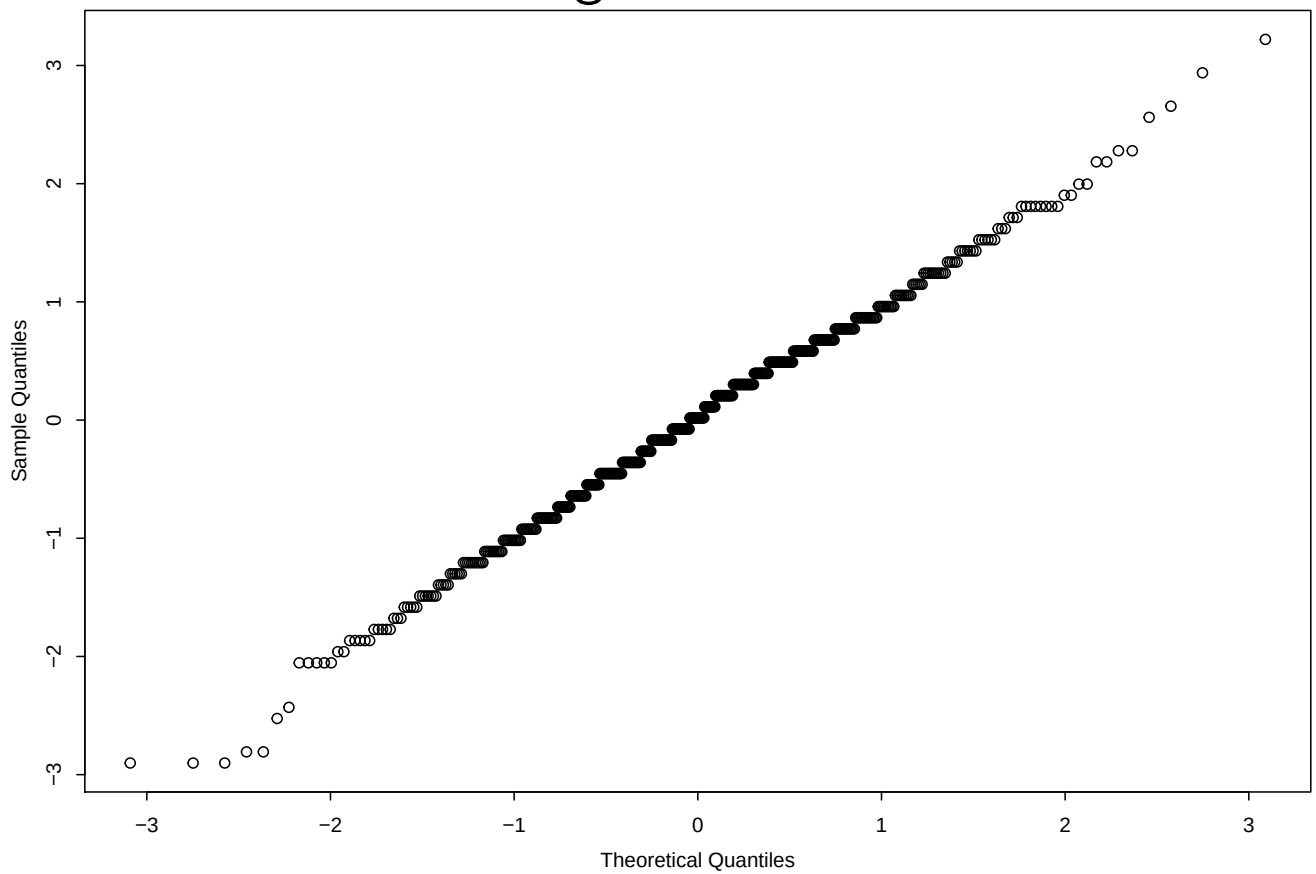
Sample Statistics ($\overline{x}$'s) have a Distribution Too XXVI

How does the original population of word lengths compare with the $M = 500$ avreage lengths (xbars) of $n = 8$ randomly chosen words?



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

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



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

Let's randomly sample $n = 15$ words instead of 8

Let's repeat the random sampling 500 times

```
randomSampleMeans.15 = replicate(500, mean(wordlen[sample(1:100,
sort(randomSampleMeans.15[1:20])
```

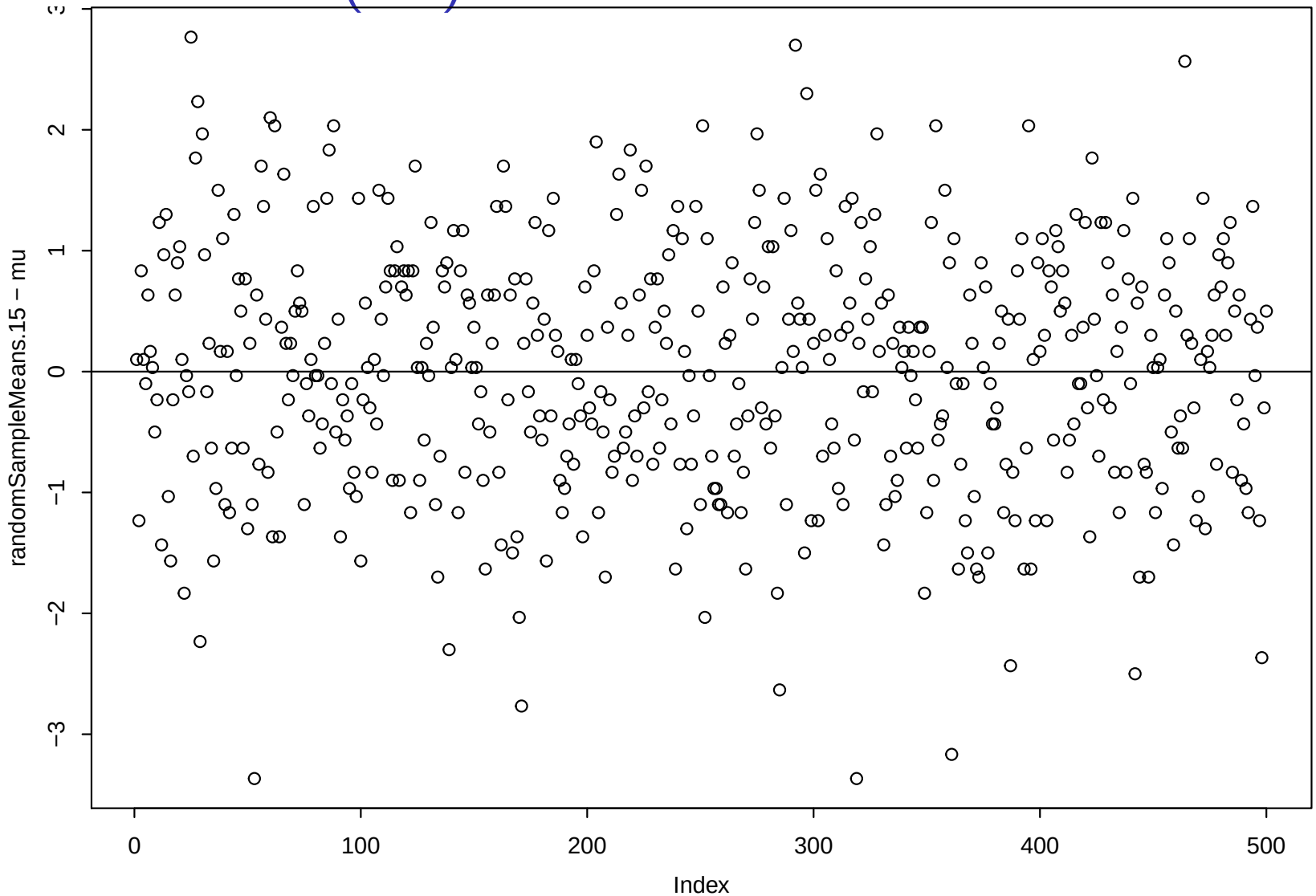
```
 [1] 10.93 11.07 11.27 11.47 12.00 12.27 12.27 12.40 12.53
[10] 12.60 12.60 12.67 13.13 13.13 13.33 13.40 13.47 13.53
[19] 13.73 13.80
```

```
mu
```

```
[1] 12.5
```

```
plot(randomSampleMeans.15 - mu)
abline(h=0)
```

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



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

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

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

```
[1] 12.51
```

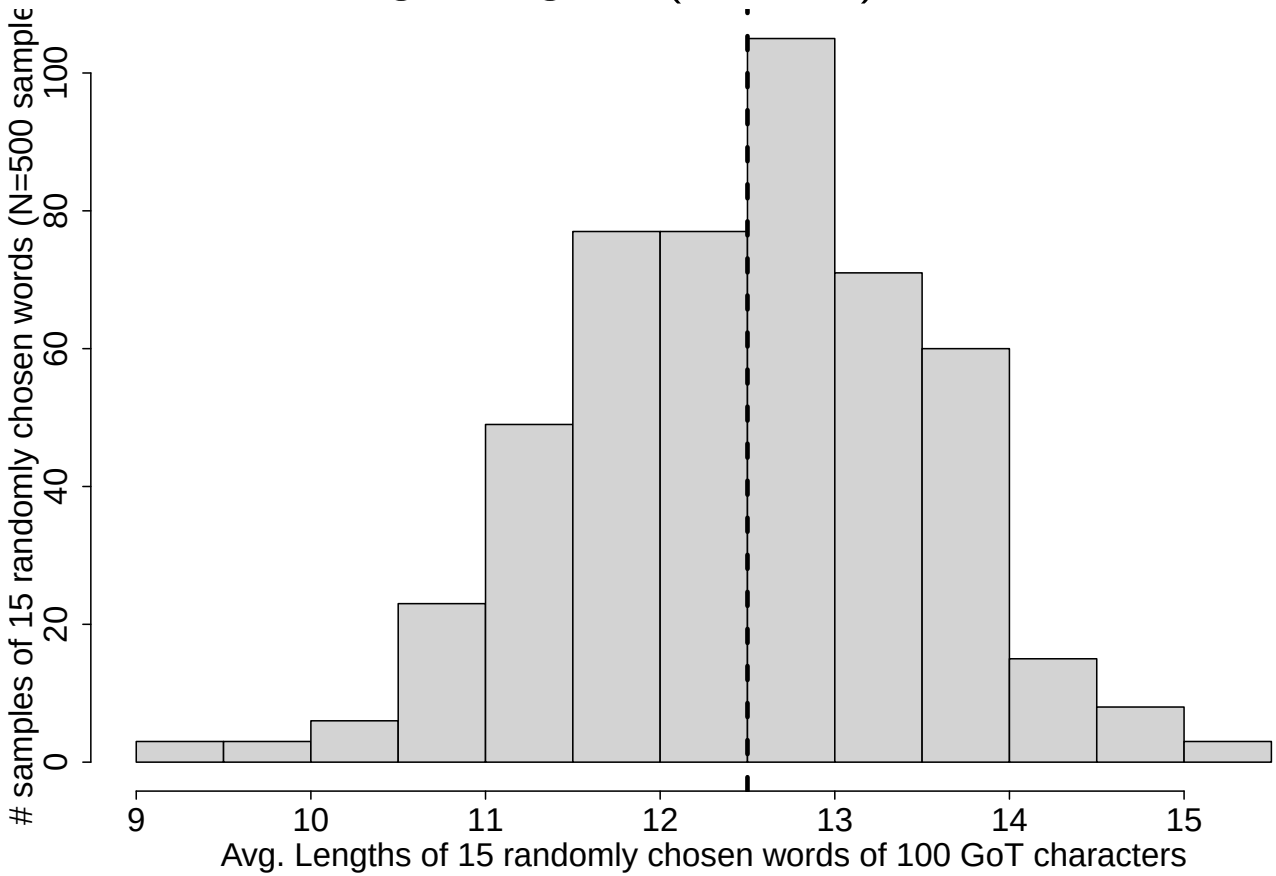
If this "mean of the averages" is close to the true mean we say that the statistic ($\overline{x}$) is an **unbiased** statistic (estimator) for the parameter (μ).

```
mu
```

```
[1] 12.5
```

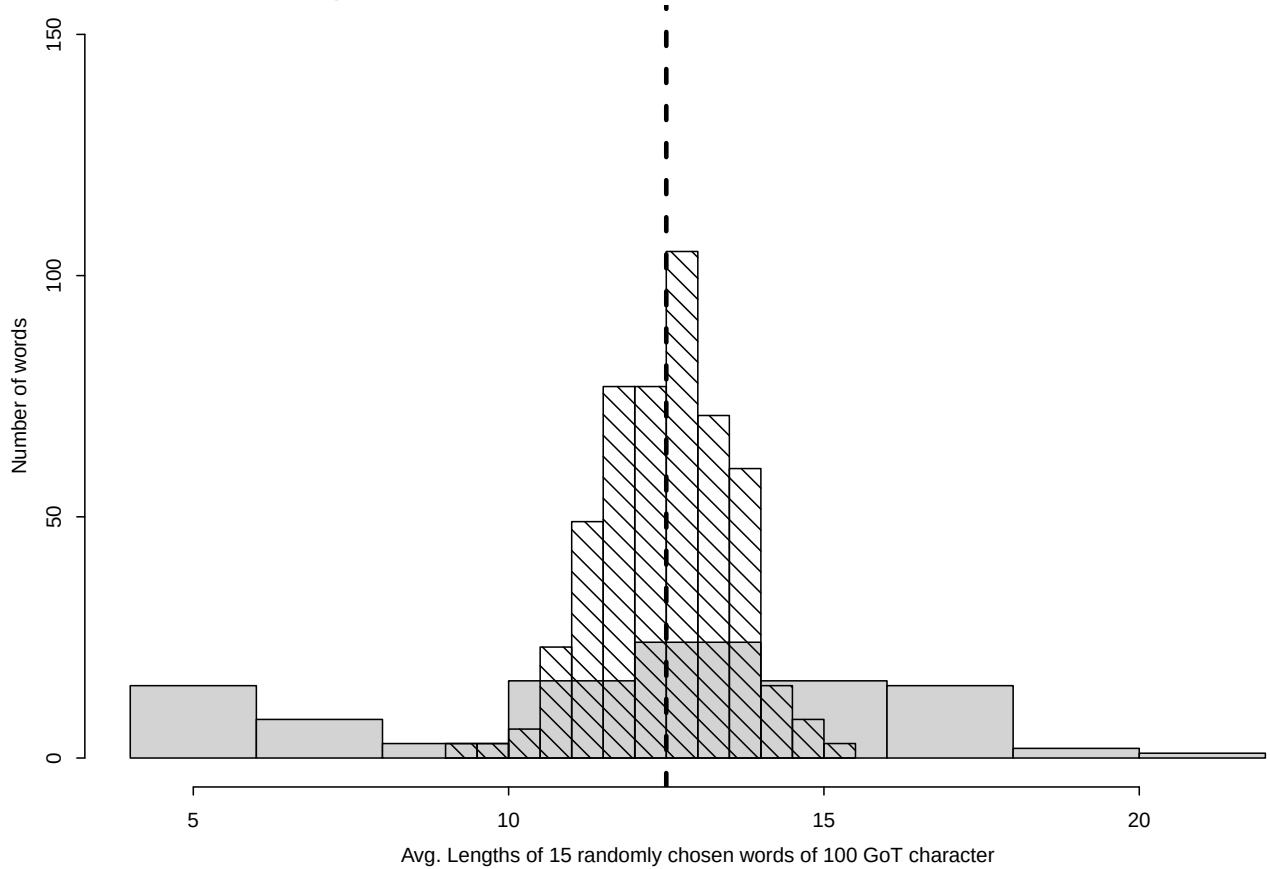
Sample Statistics ($\overline{x}$'s) have a Distribution Too XXXI

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



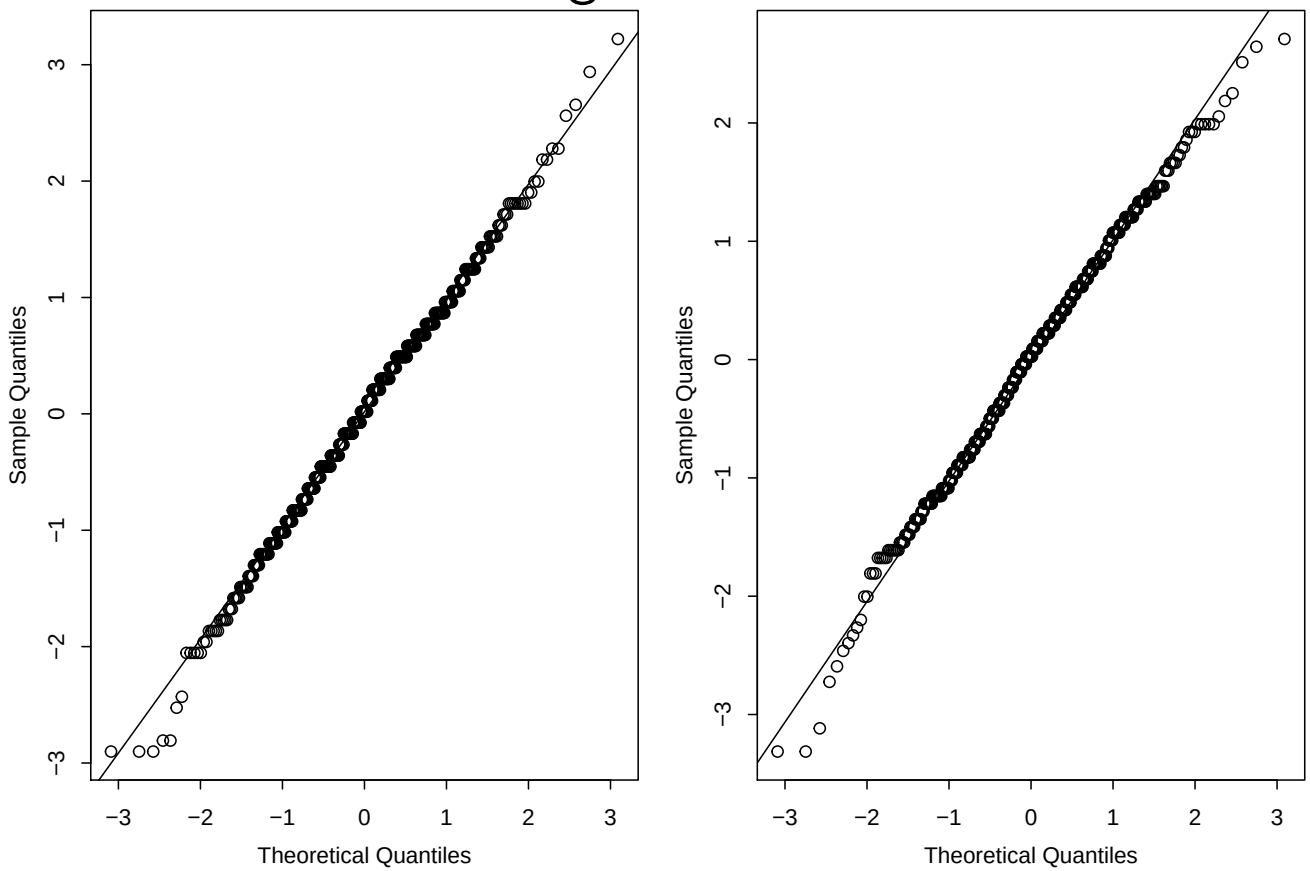
Sample Statistics ($\overline{x}$'s) have a Distribution Too XXXII

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



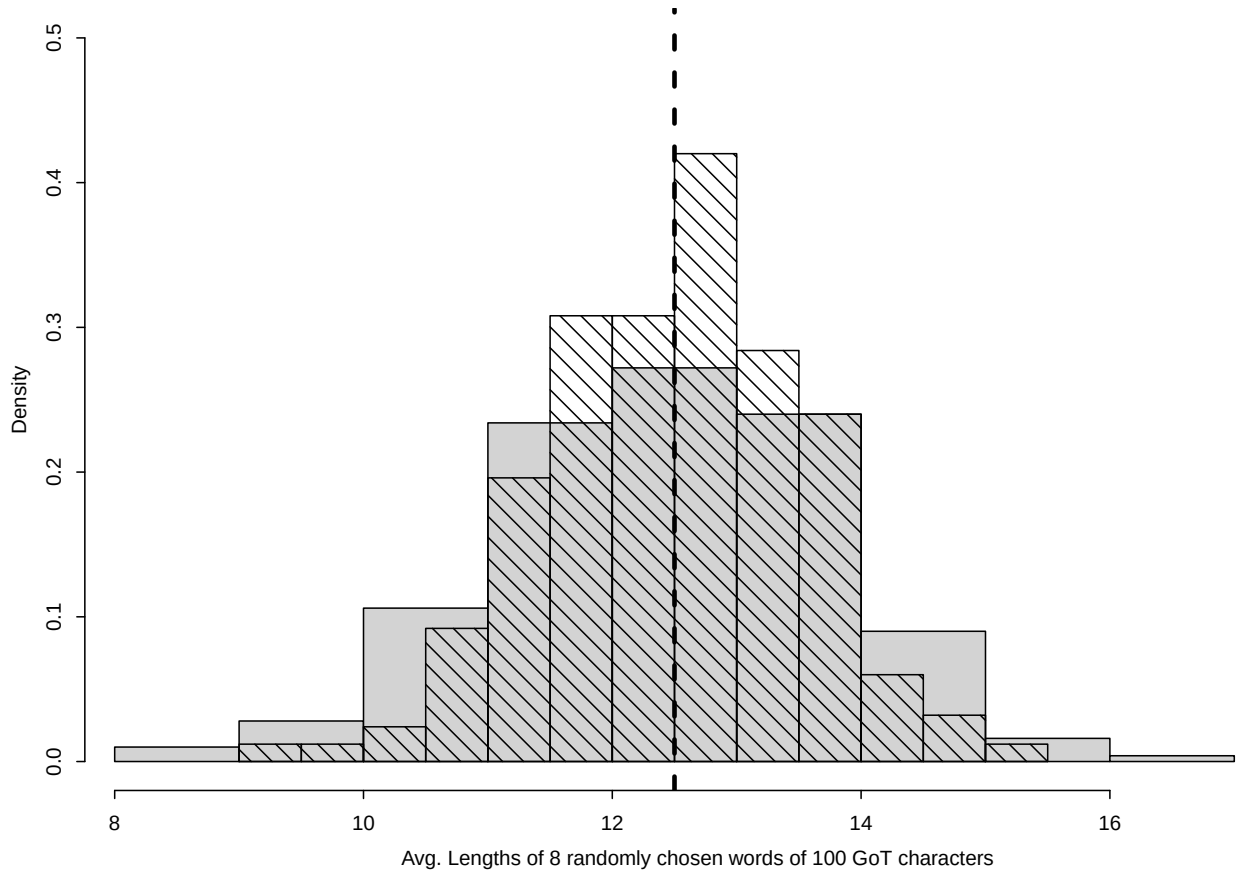
Sample Statistics ($\overline{x}$'s) have a Distribution Too XXXIII

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



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

How do the means from the random samples with 15 words compare with the $M = 500$ mean lengths of 8 randomly chosen words?



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

```
favstats(randomSampleMeans)

   min    Q1 median    Q3   max  mean    sd   n missing
 8.625 11.62   12.5 13.38 16.75 12.48 1.327 500       0

favstats(randomSampleMeans.15)

   min   Q1 median   Q3   max  mean    sd   n missing
 9.133 11.8  12.53 13.2 15.27 12.51 1.018 500       0

mu

[1] 12.5

sd(wordlen, data=worddata)

[1] 4.118
```