Session 12: Sampling

Stats 60/Psych 10
Ismael Lemhadri
Summer 2020

This time

- Sampling from a population
- Estimating population parameters from a sample
- Sampling error and the standard error of the mean
- The central limit theorem
- Confidence intervals

What is the goal of the US Census?

"Representatives and direct Taxes shall be apportioned among the several States ... according to their respective Numbers ... . The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years."

–US Constitution, Article I, Section 2

How is the census performed?

- The Census Bureau develops a comprehensive list of residential dwellings in the United States.
- A census form is mailed to each of those housing units.
- Households are asked to return the completed forms by mail.
- Households that do not return the forms are visited by enumerators.

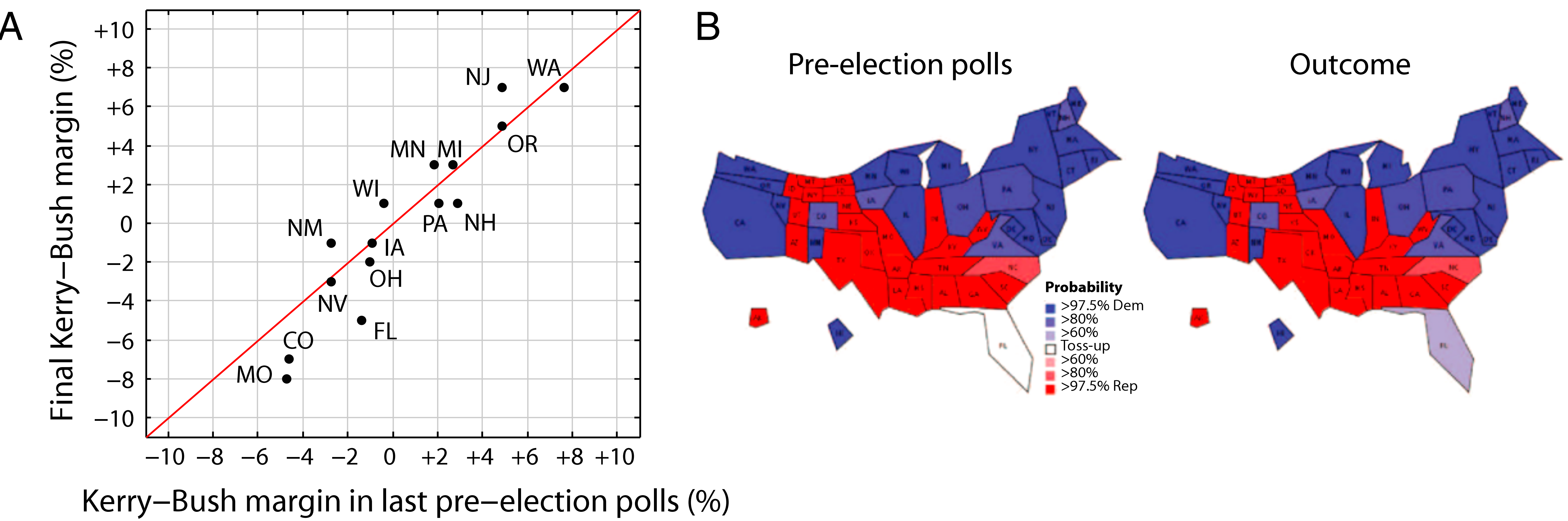
Do you think it's possible to count everyone?

Do you think it's necessary?

Sampling

- Can we estimate parameters of the entire population using just a subset?
- Can you think of examples where this is successful?

The success of election polling: 2004 US Presidential election



Wang, 2015

- Nate Silver correctly predicted outcomes for:
  - 49/50 states in 2008 Presidential election
  - 50/50 states in 2012 Presidential election
- How?

Pollster Accuracy and Bias, 2012 Presidential Election

Likely Voters Polls in Last 21 Days of Campaign

Minimum 5 Polls

Pollster	# Polls	Avg. Error	Bias	Mode	Cell?
IBD / TIPP	11	0.9	R +0.1	Live Phone	Yes
Google Consumer Surveys	12	1.6	R +1.0	Internet	N/A
Mellman	9	1.6	R +0.0	Live Phone	Yes
RAND Corporation	17	1.8	D +1.5	Internet	N/A
CNN / Opinion Research	10	1.9	R +0.6	Live Phone	Yes
Ipsos / Reuters (online)	42	1.9	R +1.4	Internet	N/A
Angus Reid	11	1.9	R +0.8	Internet	N/A
CVOTER International / UPI	13	2.0	R +2.0	Live Phone	??
Grove Insight	18	2.0	R +0.1	Live Phone	Yes
SurveyUSA	17	2.2	R +0.5	Robodial	Yes
Quinnipiac	5	2.3	D +0.3	Live Phone	Yes
Marist	11	2.5	R +1.0	Live Phone	Yes
YouGov	30	2.6	R +1.1	Internet	N/A
We Ask America	9	2.6	D +0.1	Robodial	No
Public Policy Polling	71	2.7	R +1.6	Robodial	No
Gravis Marketing	16	2.7	R +2.7	Robodial	No
JZ Analytics*	17	2.8	R +0.1	Internet	N/A
Washington Post / ABC News	16	2.8	R +2.7	Live Phone	Yes
Pharos Research Group*	14	4.0	D +2.5	Live Phone	No
Rasmussen Reports	60	4.2	R +3.7	Robo + Internet	No
American Research Group	9	4.5	R +4.5	Live Phone	Yes
Mason-Dixon	8	5.4	R +2.2	Live Phone	Yes
Gallup	11	7.2	R +7.2	Live Phone	Yes

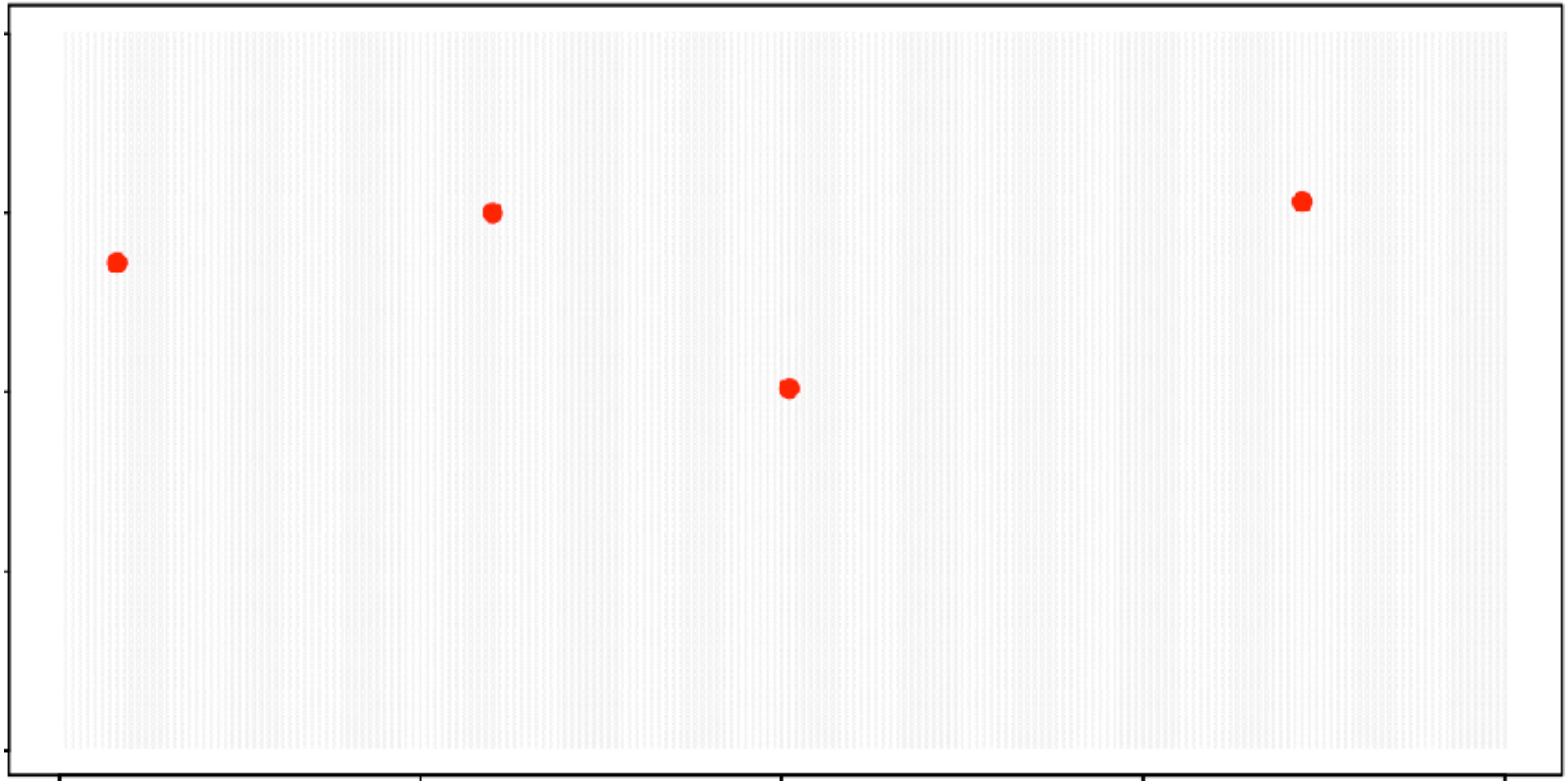
* Not used in FiveThirtyEight forecast.

Each poll includes ~1000 likely voters

Survey of~21,000 voters allows accurate estimation of voting behavior of ~200 million people

https://fivethirtyeight.blogs.nytimes.com/2012/11/10/which-polls-fared-best-and-worst-in-the-2012-presidential-race/

21,000/235 million voters = 0.008% of all voters



40,000 points

each point represents ~5,800 voters

Why doesn't the census use sampling rather than full enumeration?

DEPARTMENT OF COMMERCE ET AL. *v.* UNITED STATES HOUSE OF REPRESENTATIVES ET AL.

APPEAL FROM THE UNITED STATES DISTRICT COURT FOR THE DISTRICT OF COLUMBIA

No. 98–404. Argued November 30, 1998—Decided January 25, 1999*

2. The Census Act prohibits the proposed uses of statistical sampling to determine the population for congressional apportionment purposes.

Sampling error

- When we compute a statistic on a sample, it will have some amount of error
  - Compared to the true value ("population parameter")
- This error varies from sample to sample
  - We refer to the distribution of the statistic computed across samples as its "sampling distribution"

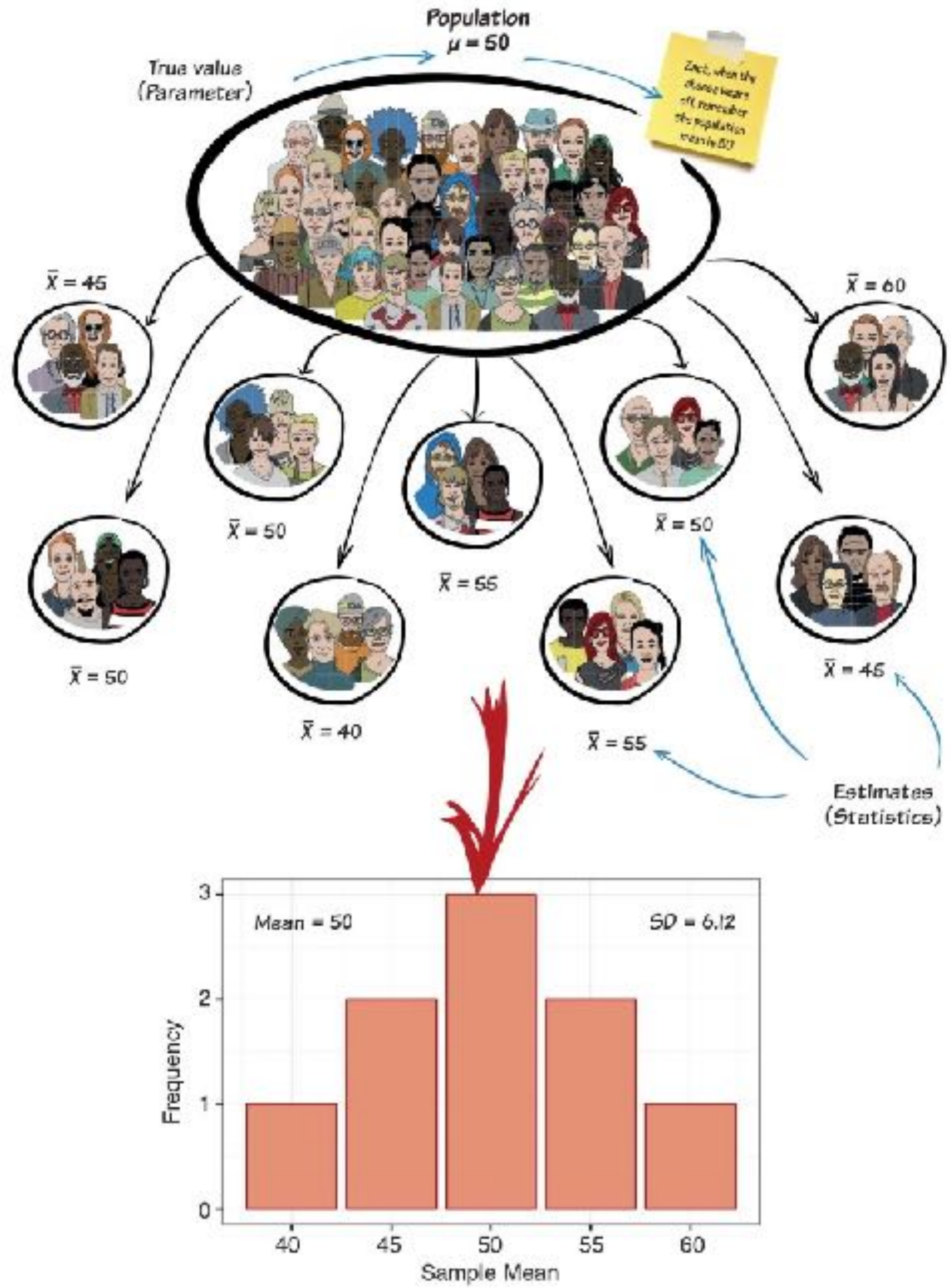


Figure 8.2 Illustration of the standard error (see text for details)

Taking random samples in R

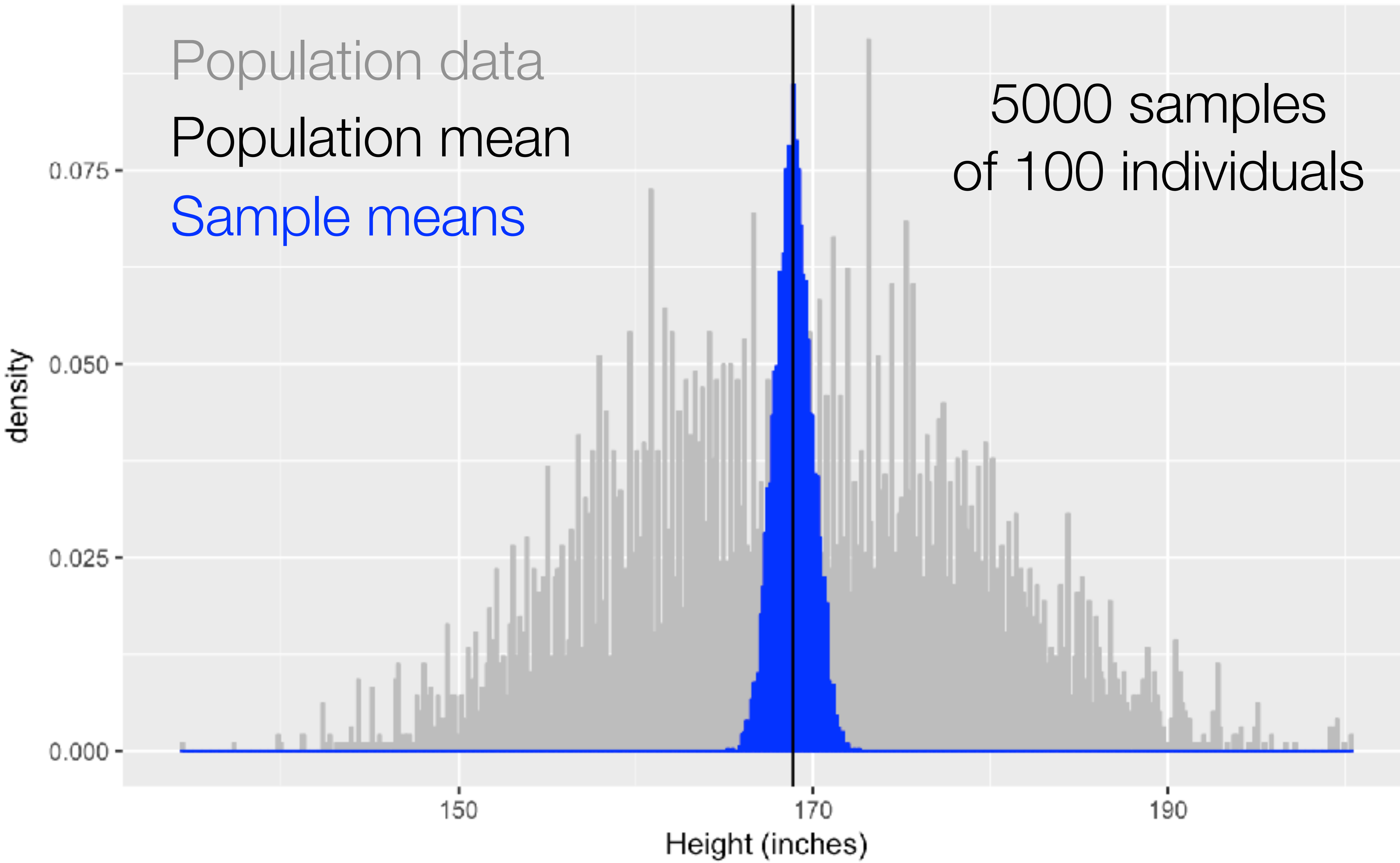
```
exampleSample <- NHANES_adult %>%
         sample_n(10)

dim(NHANES_adult)
## [1] 7424   77
dim(exampleSample)
## [1] 10 77
```

```
print(paste('Population height: mean = ',mean(NHANES_adult$Height)))
## [1] "Population height: mean =  168.86"
```

```
print(paste('Sample height: mean = ',mean(exampleSample$Height)))
## [1] "Sample height: mean =  168.59"
```

Sampling error: NHANES adult height



Standard error of the mean (SEM)

- The standard deviation of the sampling distribution of the mean
  - How variable are our estimates of the mean?

$$SEM = \frac{\sigma}{\sqrt{n}}$$

Population std deviation

Sample size

Computing the standard error of the mean

- We usually do not know the population standard deviation
- Instead we usually "plug in" the sample standard deviation in its place
- This assumes that the sample SD is a good estimate of the population SD
  - With larger samples (>~30) this should be OK

If population SD is known:

$$SEM = \frac{\sigma}{\sqrt{n}}$$

If population SD is unknown:

$$SEM = \frac{SD}{\sqrt{n}}$$

Two samples are obtained from the same population, sample A with 100 subjects and sample B with 200 subjects. What is the relationship between the standard error for the mean in the two samples?

SEM(A) = SEM(B)

SEM(A) = SEM(B)*2

SEM(A)=SEM(B)*1.41

SEM(A)=SEM(B)/2

SEM(A)=SEM(B)/1.41

How to make better measurements

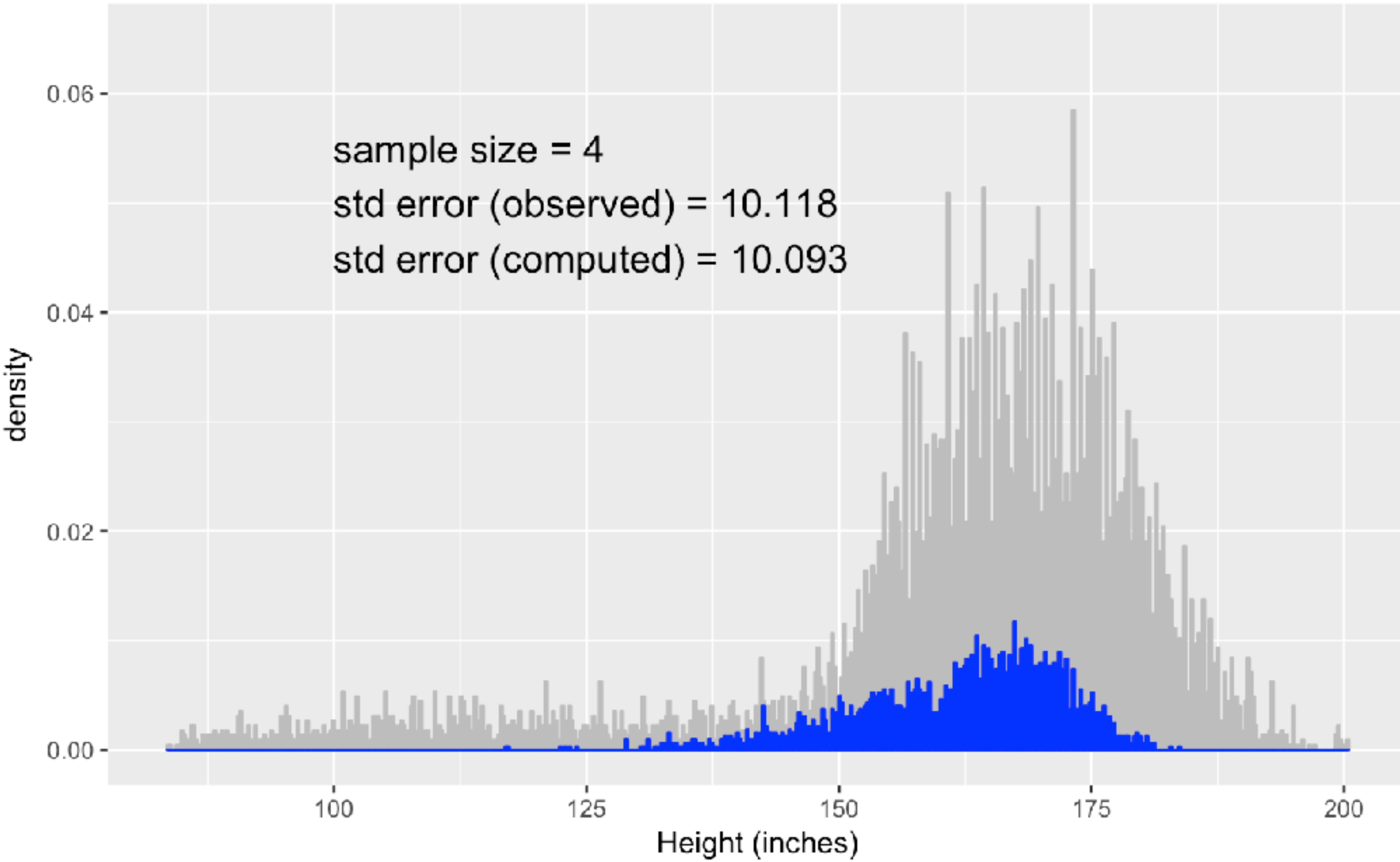
- We don't have control over the population SD
- We usually do have control over the sample size
- Larger samples reduce SEM but give diminishing returns
    - Increasing sample from 16 to 25 (by 9) provides same improvement in SEM as increasing from 100 to 121 (by 21)

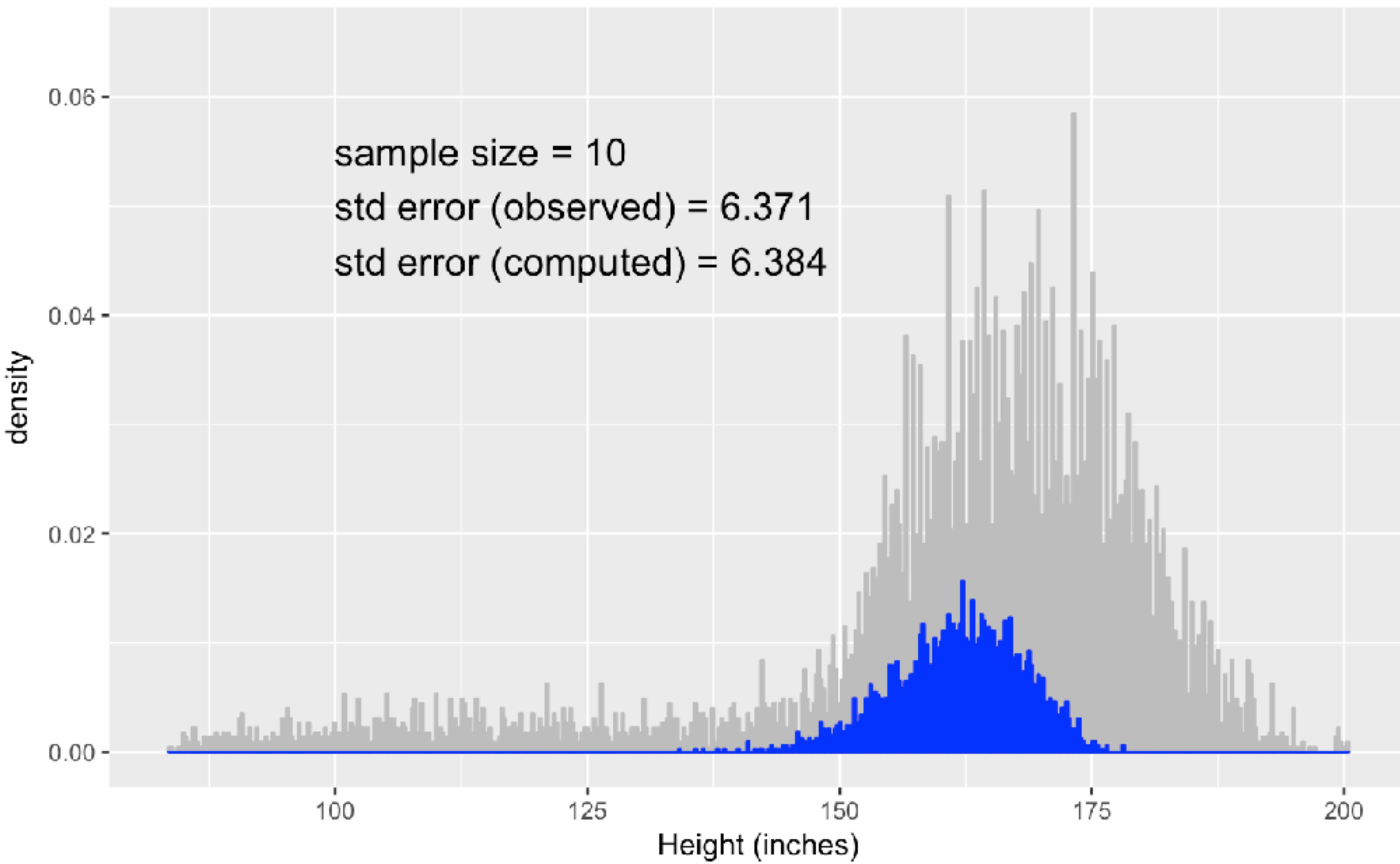
$$SEM = \frac{\sigma}{\sqrt{n}}$$

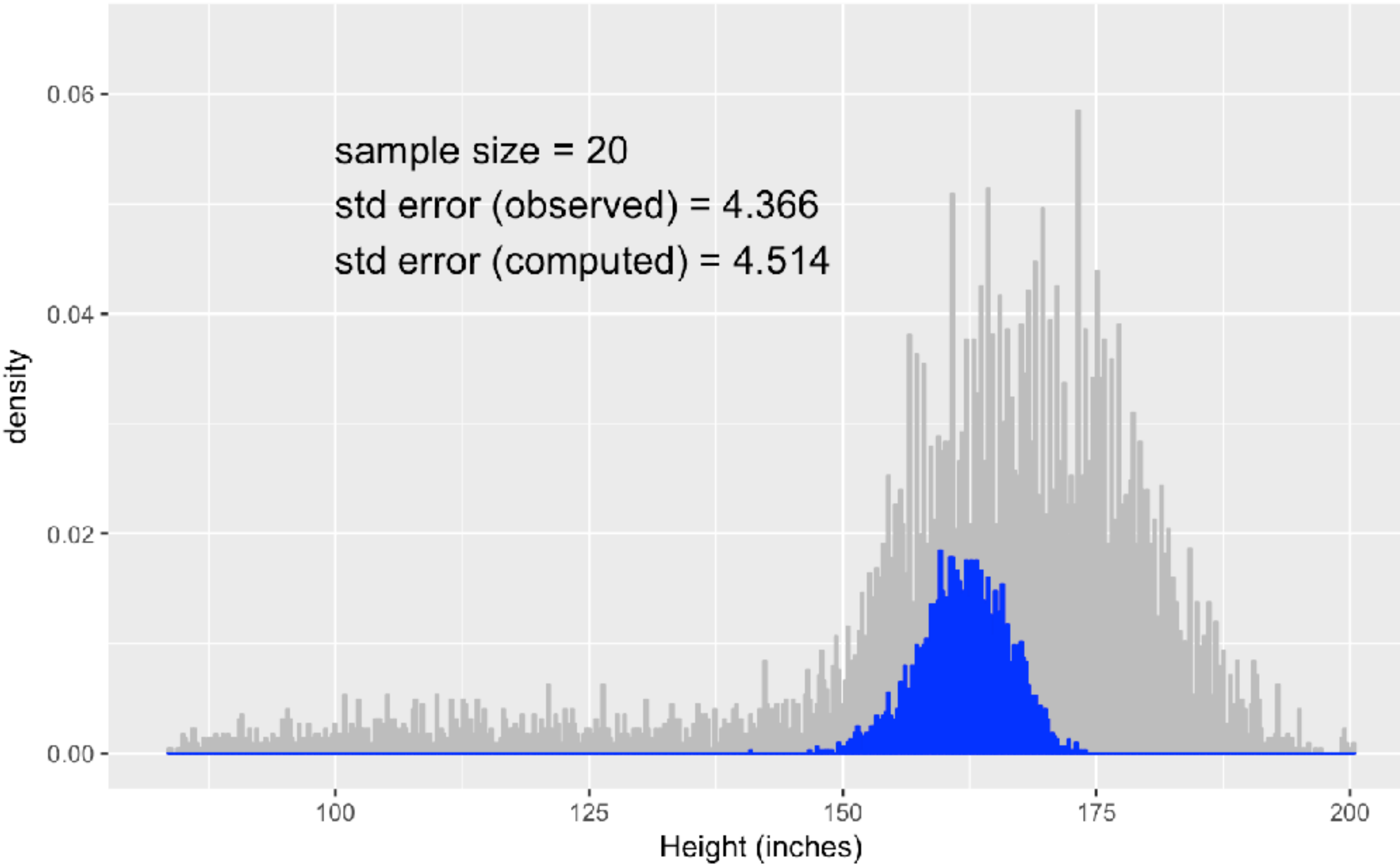
Central limit theorem

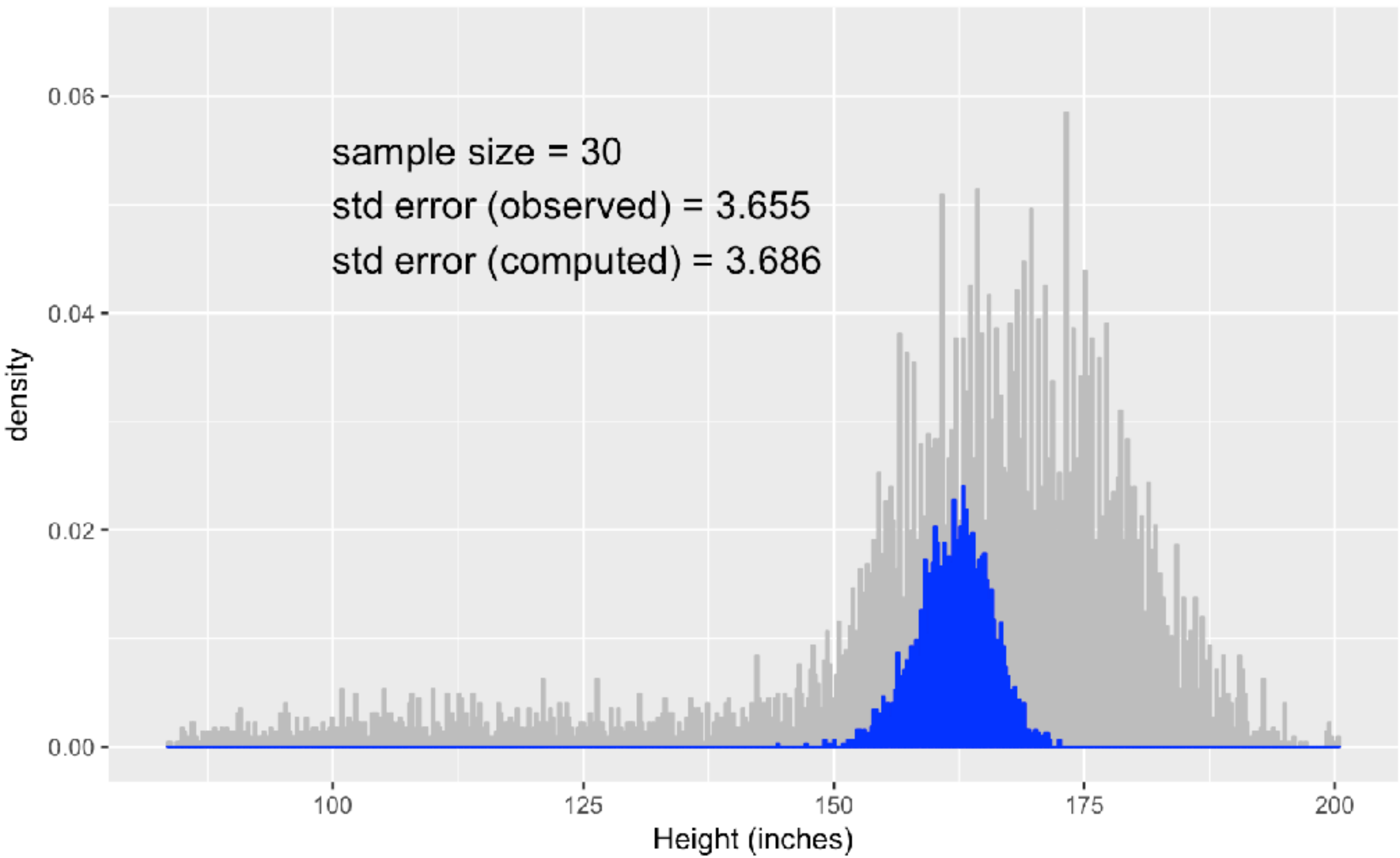
- As the sample size gets large, the sampling distribution of the mean will come to resemble a normal distribution
  - Regardless of the shape of the distribution of the data!
- This probably explains why so many variables in the real world follow a normal distribution

- Let’s take samples from NHANES Height and look at the sampling distribution of the mean

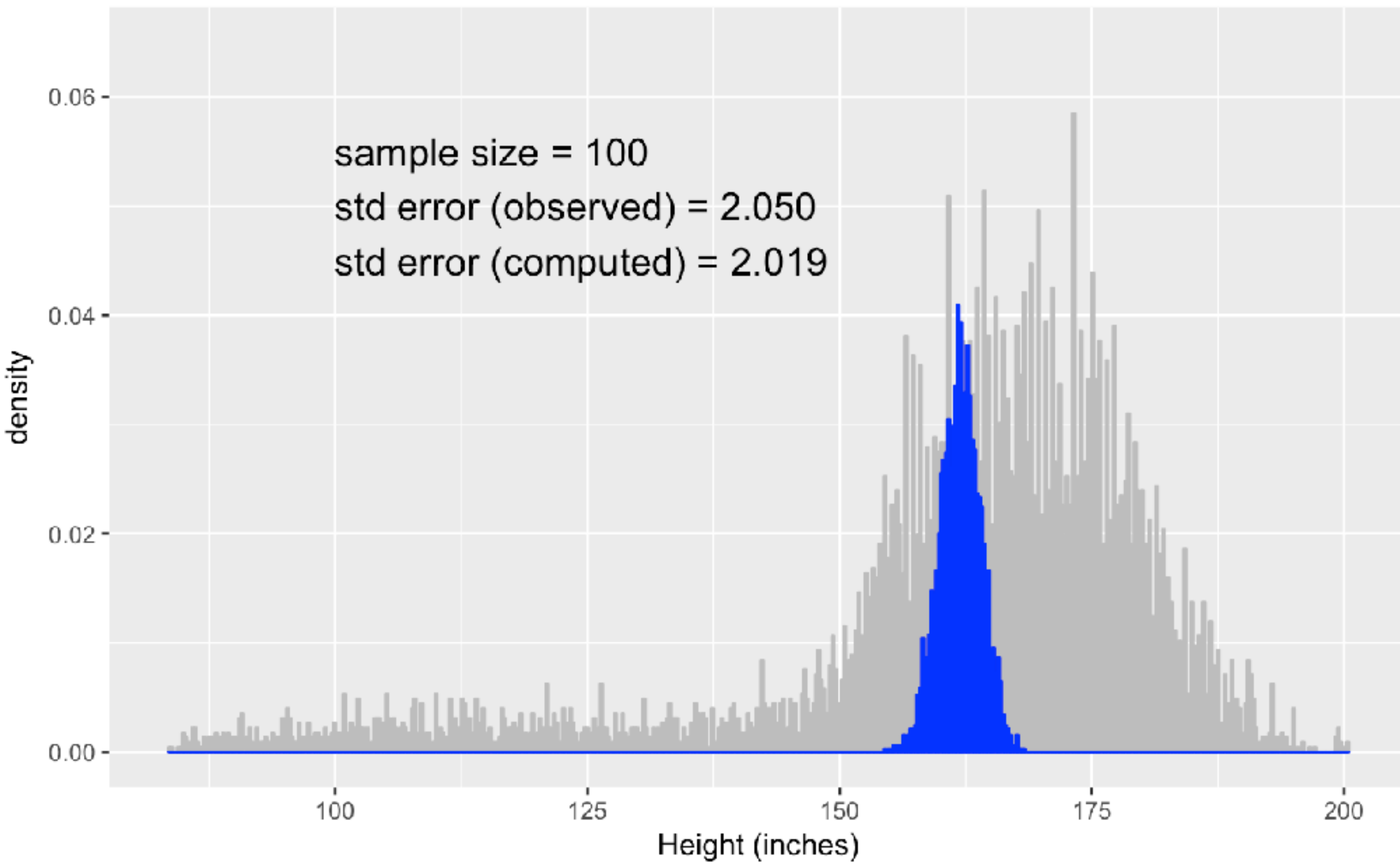

sample size = 4
std error (observed) = 10.118
std error (computed) = 10.093
0.06
0.04
0.02
0.00
density
100
125
150
175
200
Height (inches)

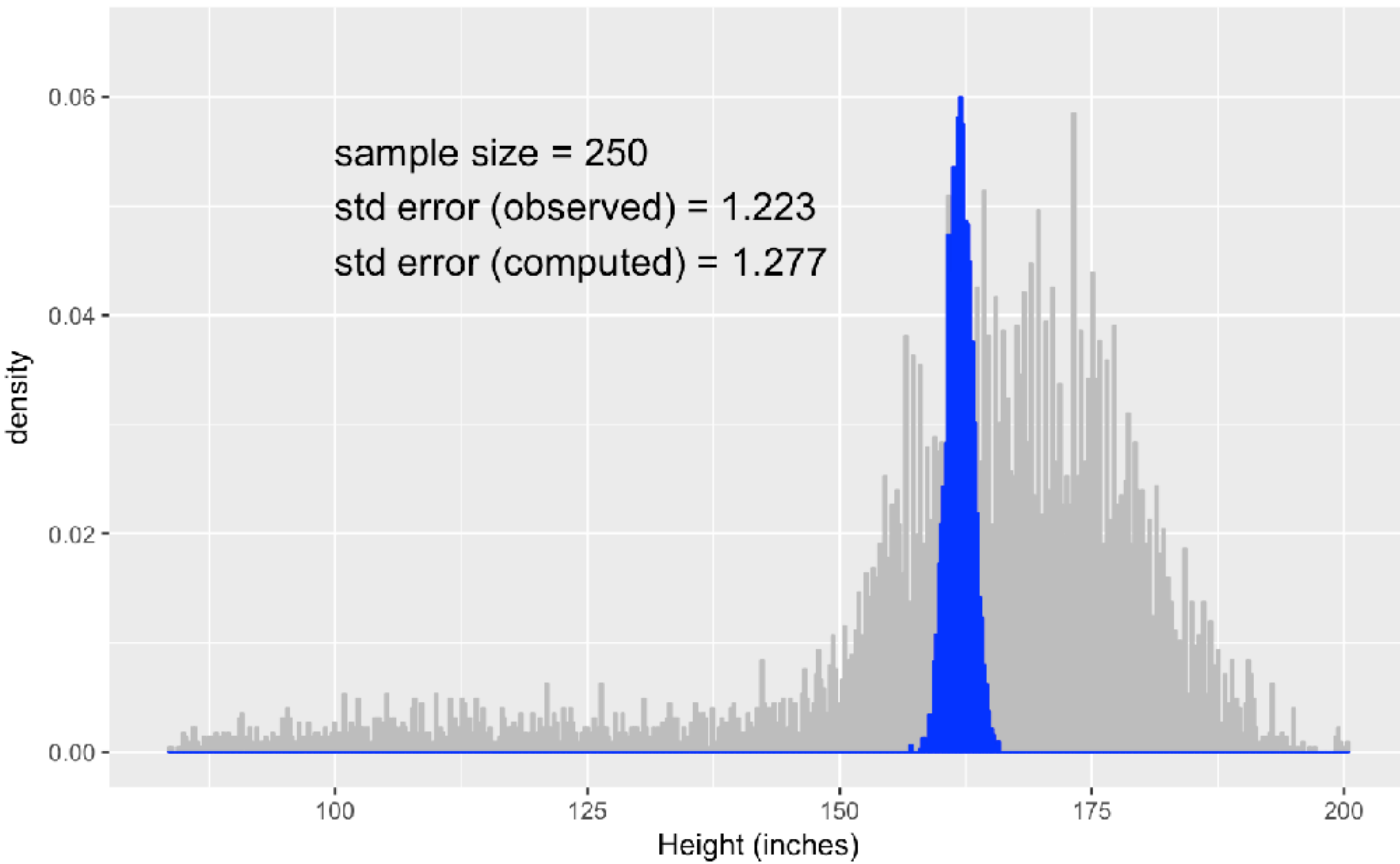

sample size = 10
std error (observed) = 6.371
std error (computed) = 6.384
density
0.06
0.04
0.02
0.00
100
125
150
175
200
Height (inches)


sample size = 20
std error (observed) = 4.366
std error (computed) = 4.514
density
0.06
0.04
0.02
0.00
100
125
150
175
200
Height (inches)



sample size = 30
std error (observed) = 3.655
std error (computed) = 3.686
0.06
0.04
0.02
0.00
density
100
125
150
175
200
Height (inches)

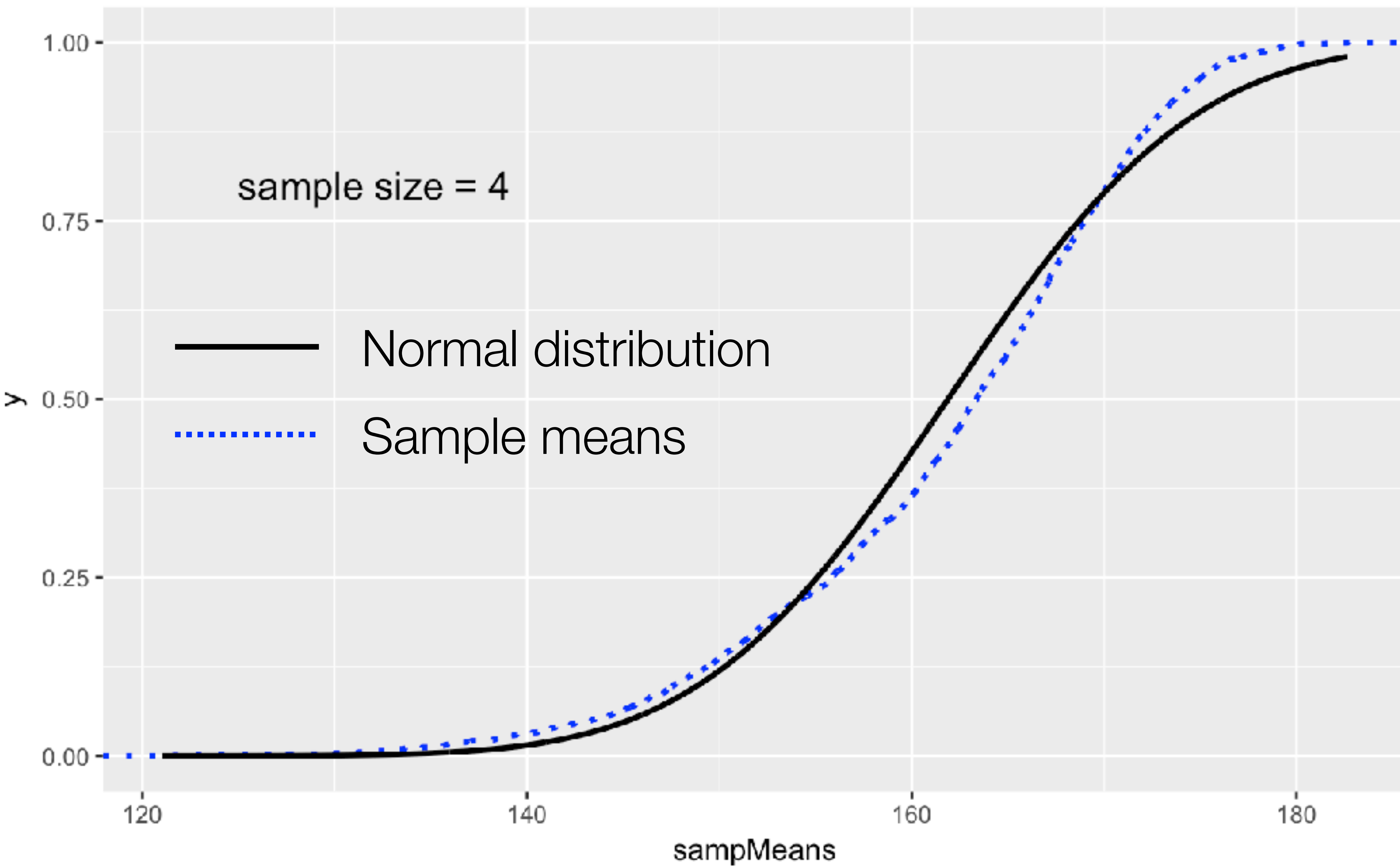

sample size = 100
std error (observed) = 2.050
std error (computed) = 2.019
0.06
0.04
0.02
0.00
density
100
125
150
175
200
Height (inches)

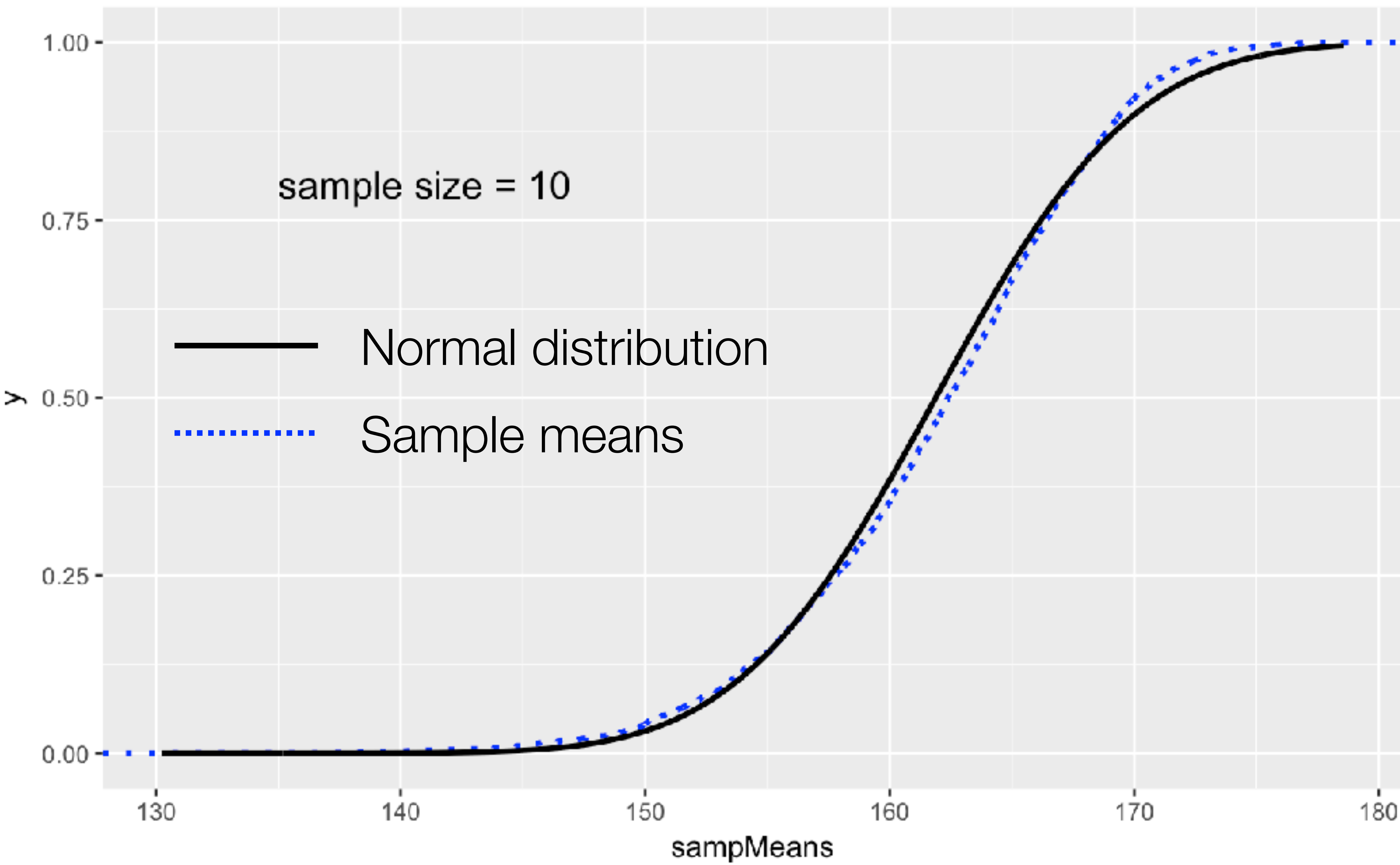


sample size = 250
std error (observed) = 1.223
std error (computed) = 1.277
0.06
0.04
0.02
0.00
density
100
125
150
175
200
Height (inches)

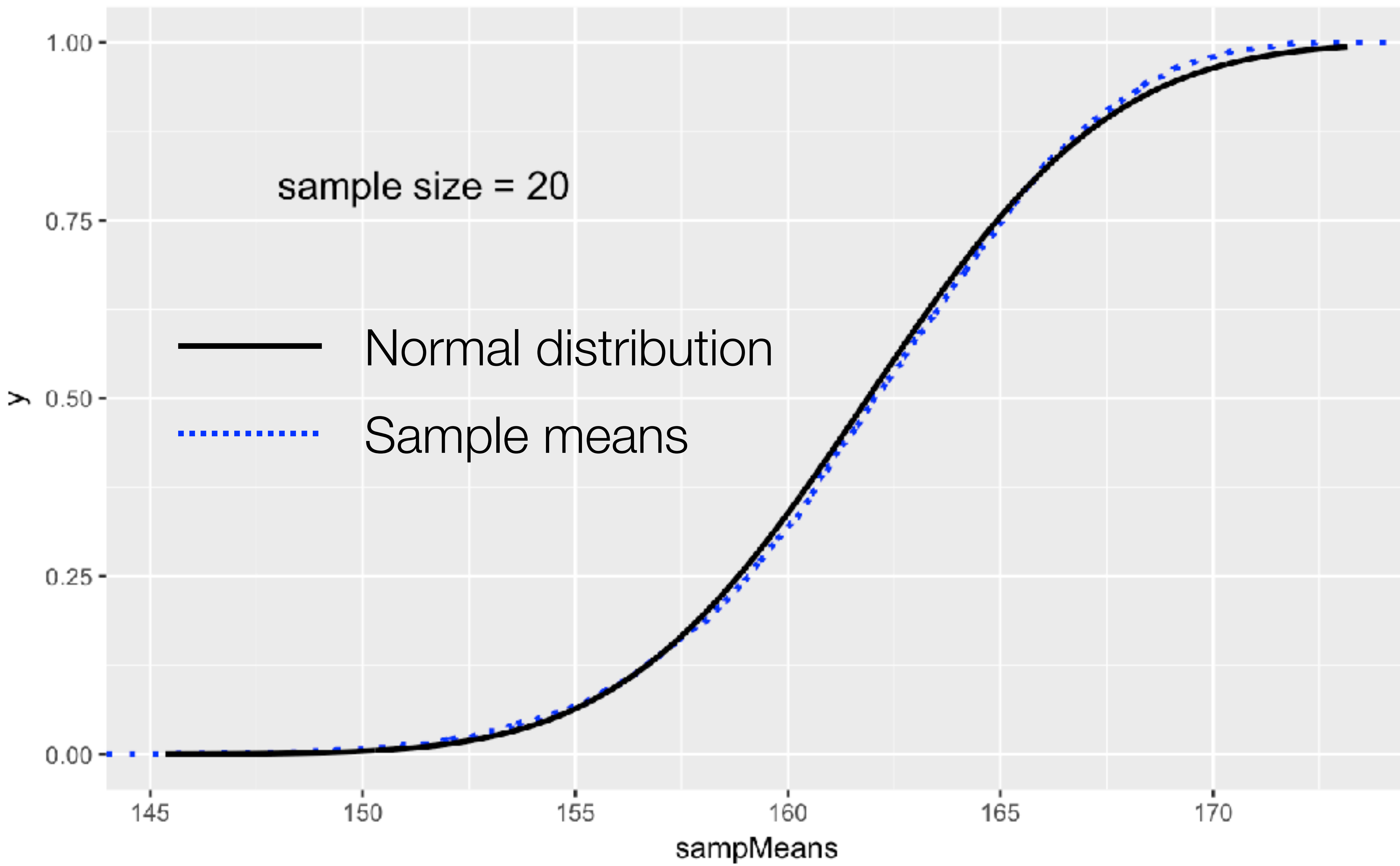
Comparing data to the normal distribution

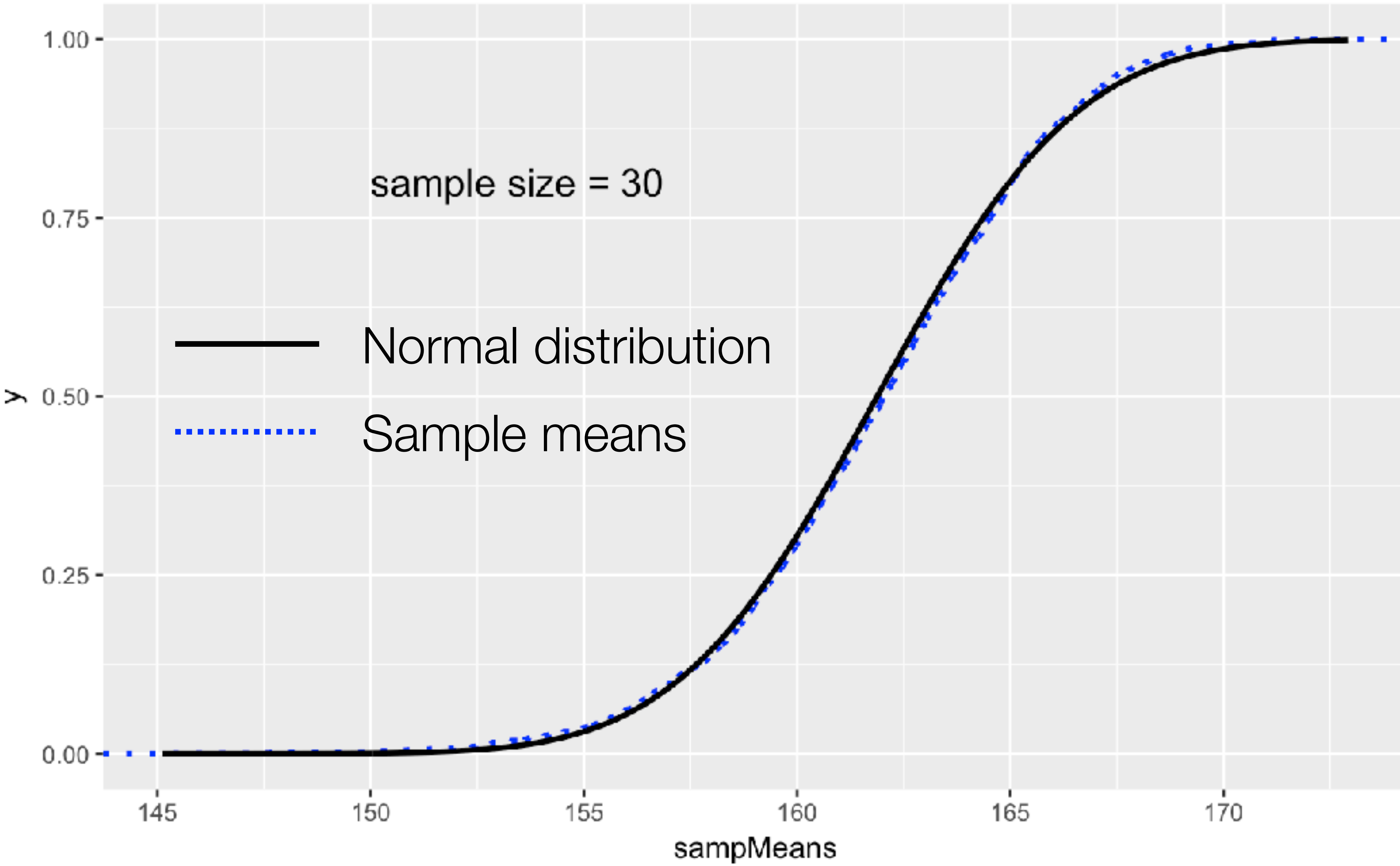
- Plot the cumulative distribution of the data against the cumulative distribution of the normal


sample size = 4
Normal distribution
Sample means
y
1.00
0.75
0.50
0.25
0.00
120
140
160
180
sampMeans

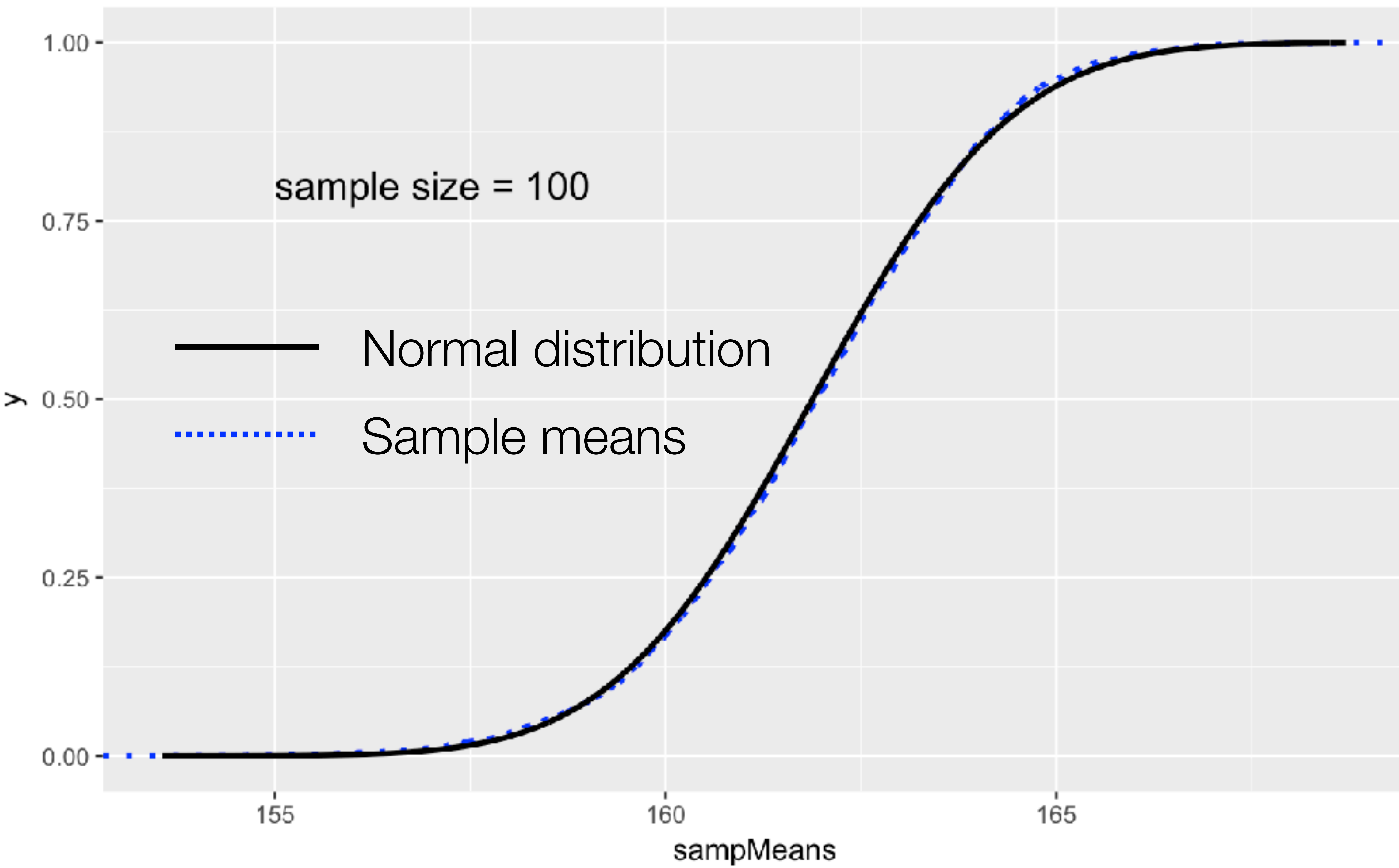


sample size = 10
Normal distribution
Sample means
y
1.00
0.75
0.50
0.25
0.00
130
140
150
160
170
180
sampMeans

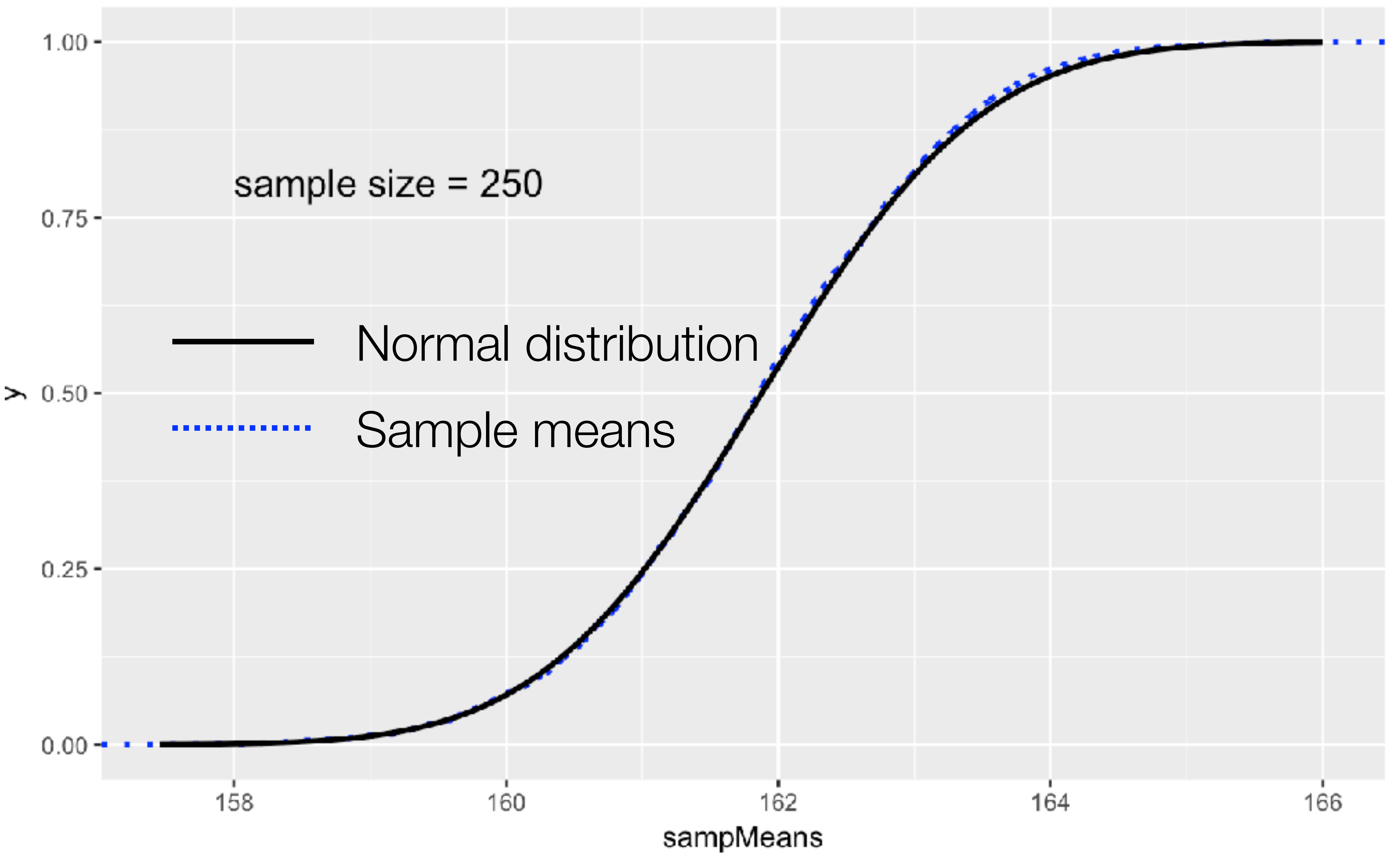

sample size = 20
Normal distribution
Sample means
1.00
0.75
0.50
0.25
0.00
y
145
150
155
160
165
170
sampMeans



sample size = 30
Normal distribution
Sample means
y
1.00
0.75
0.50
0.25
0.00
145
150
155
160
165
170
sampMeans



sample size = 100
Normal distribution
Sample means
y
1.00
0.75
0.50
0.25
0.00
155
160
165
sampMeans



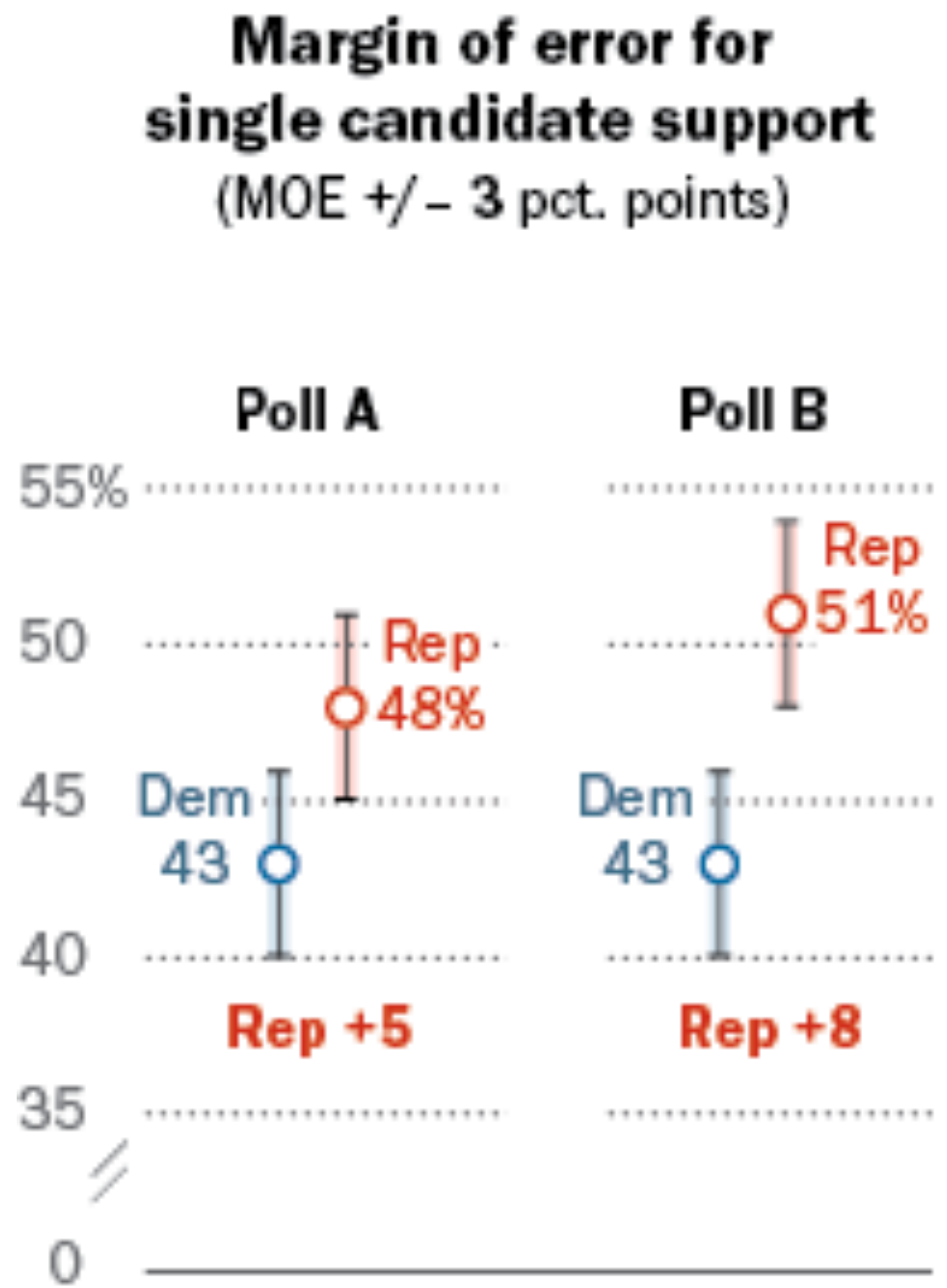
1.00
0.75
0.50
0.25
0.00
y
sample size = 250
Normal distribution
Sample means
158
160
162
164
166
sampMeans

Why a smaller SEM is better

- Larger SEM = more uncertainty about the true value of the parameter in the population
  - Would you believe an election poll if you were told that the results had a margin of error of 15%?
  - And what does “margin of error” mean?

Confidence intervals

- An interval around the mean that expresses our uncertainty about the true value of the parameter

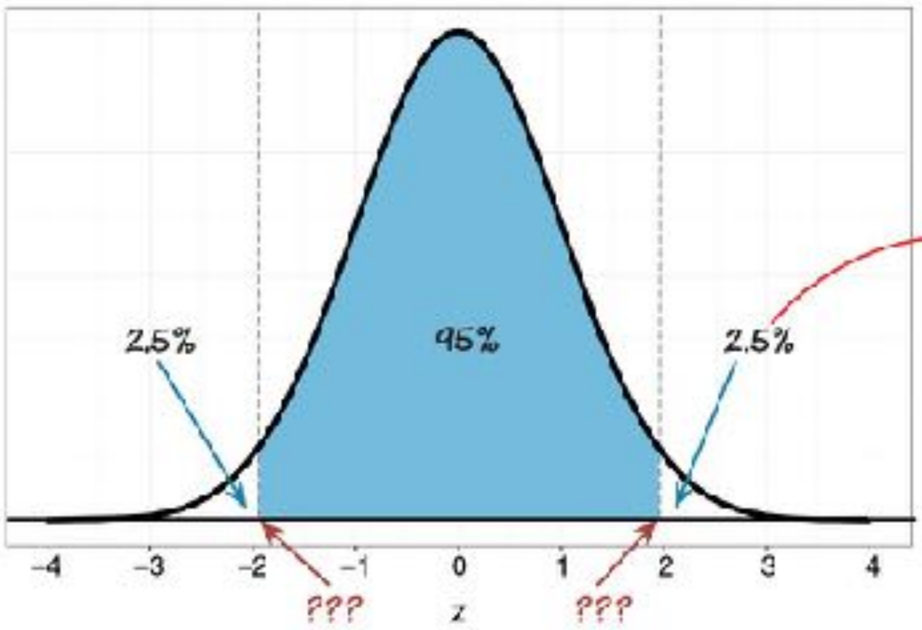


http://www.pewresearch.org/fact-tank/2016/09/08/understanding-the-margin-of-error-in-election-polls/

Computing the confidence interval

- We want to express our uncertainty about the estimate of the mean
- Remember that:
  - the sample means are normally distributed (per the Central Limit Theorem)
  - The standard error is the standard deviation of the sampling distribution
  - What we want to know is: What interval would we expect to capture 95% of values around the mean?

Using the normal distribution to compute the confidence interval around the mean



z	Larger Proportion	Smaller Proportion	Mean to z
0.00	0.50000	0.50000	0.00000
0.01	0.50399	0.49601	0.00399
0.02	0.50798	0.49202	0.00798
0.03	0.51197	0.48803	0.01197
1.95	0.97441	0.02559	0.47441
1.96	0.97500	0.02500	0.47500
1.97	0.97558	0.02442	0.47558
1.98	0.97615	0.02385	0.47615
1.99	0.97670	0.02330	0.47670

Figure 8.7 Finding the limits within which a specific proportion of scores fall

- We want to find the value of the normal distribution such that 5% of responses are excluded
  - 2.5% higher, 2.5% lower
- This value is ~ ±1.96

```
> qnorm(0.025)
[1] -1.959964
> qnorm(0.975)
[1] 1.959964
```

Computing the confidence interval

$$CI(95\%) = mean \pm 1.96 * SEM$$

Sample 250 individuals from NHANES

```
NHANES_sample=sample_n(NHANES,250)
```

Compute sample mean and SD

```
> sampleMean=mean(NHANES_sample$Height)
> sampleMean
[1] 163.3684
> sampleSD=sd(NHANES_sample$Height)
> sampleSD
[1] 19.54375
```

Compute CI

```
> CIupper=sampleMean + 1.96*sampleSD
> CIlower=sampleMean - 1.96*sampleSD
> c(CIlower,CIupper)
[1] 125.0626 201.6742
```

What the confidence interval means

- If we take a large number of samples, the confidence interval will contain the true value 95% of the time

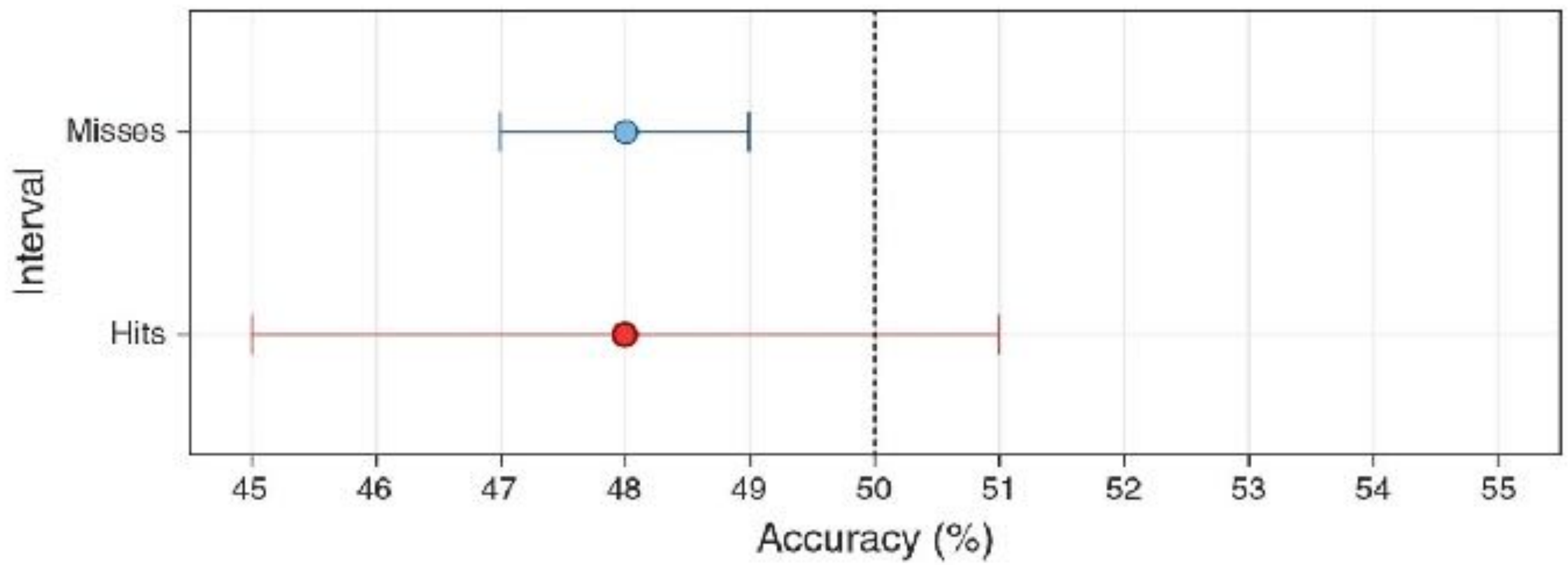


Figure 8.5 Two interval estimates: the red one hits the true population value (dotted line) but the blue one misses it

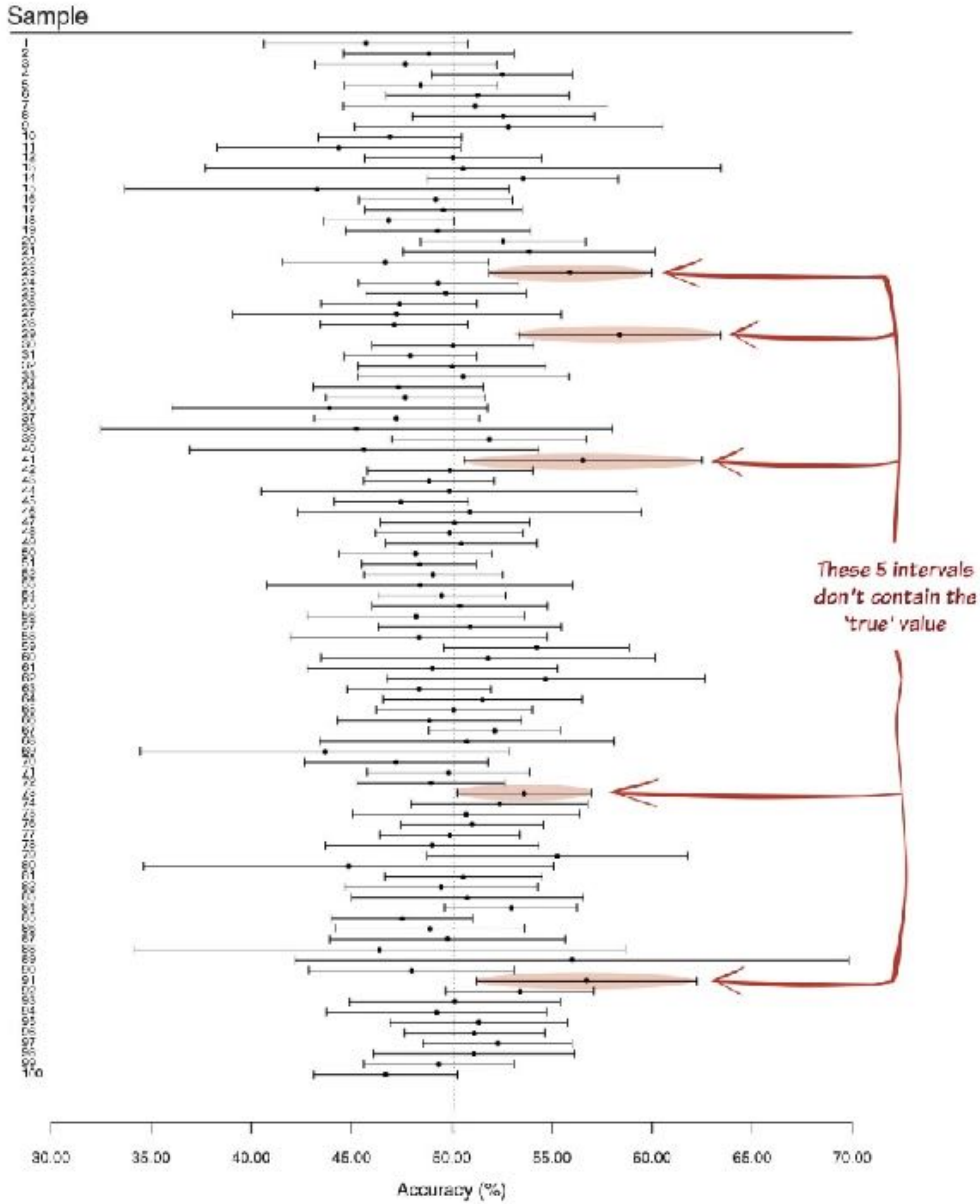
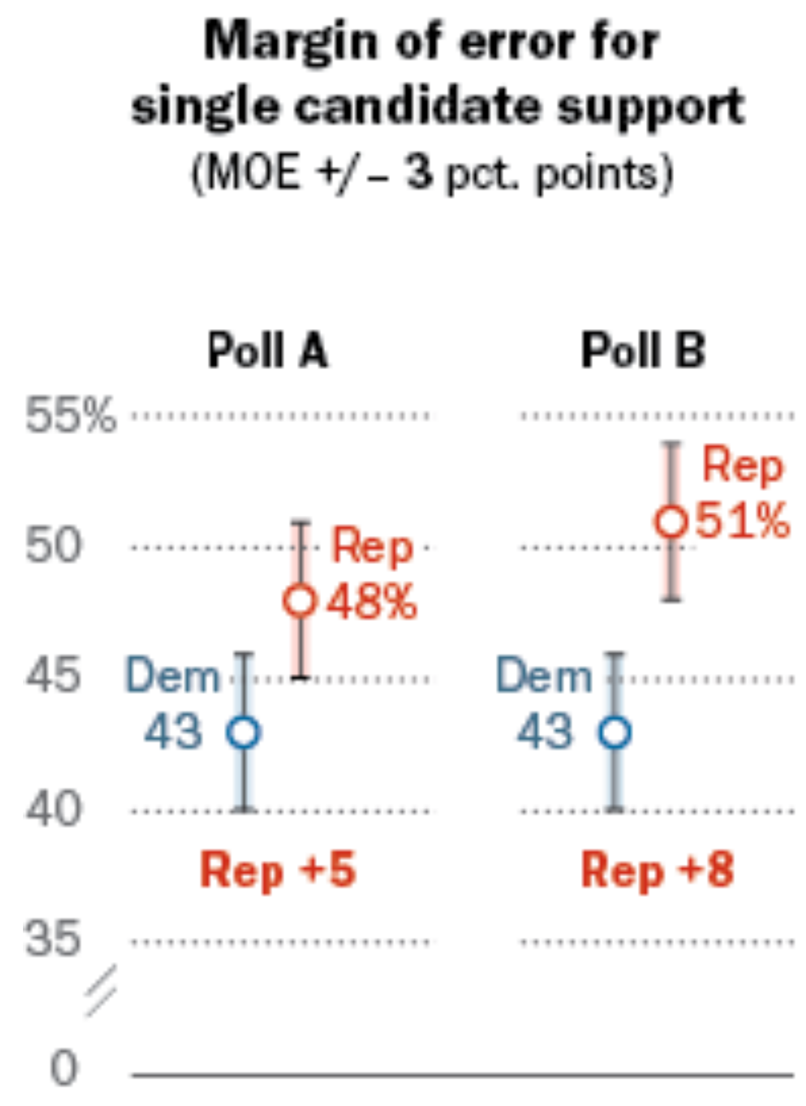


Figure 8.6 The confidence intervals of the accuracy on a statistics test (horizontal axis) for 100 different samples (vertical axis)

What confidence intervals *do not* mean

- Poll A: Republican CI (45% - 51%)
- Does not mean that there is a 95% chance that the true population value falls between 45 and 51
    - It either does or it doesn't!
- We will talk more about the confusing interpretation of confidence intervals when we return to hypothesis testing



Recap

- We can obtain accurate estimates of population parameters through random sampling
- Larger sample sizes give smaller standard error, but with diminishing returns
- The central limit theorem assures us that the sampling distribution of the mean becomes normal with larger N
- Confidence intervals give us a way to express our uncertainty about the mean