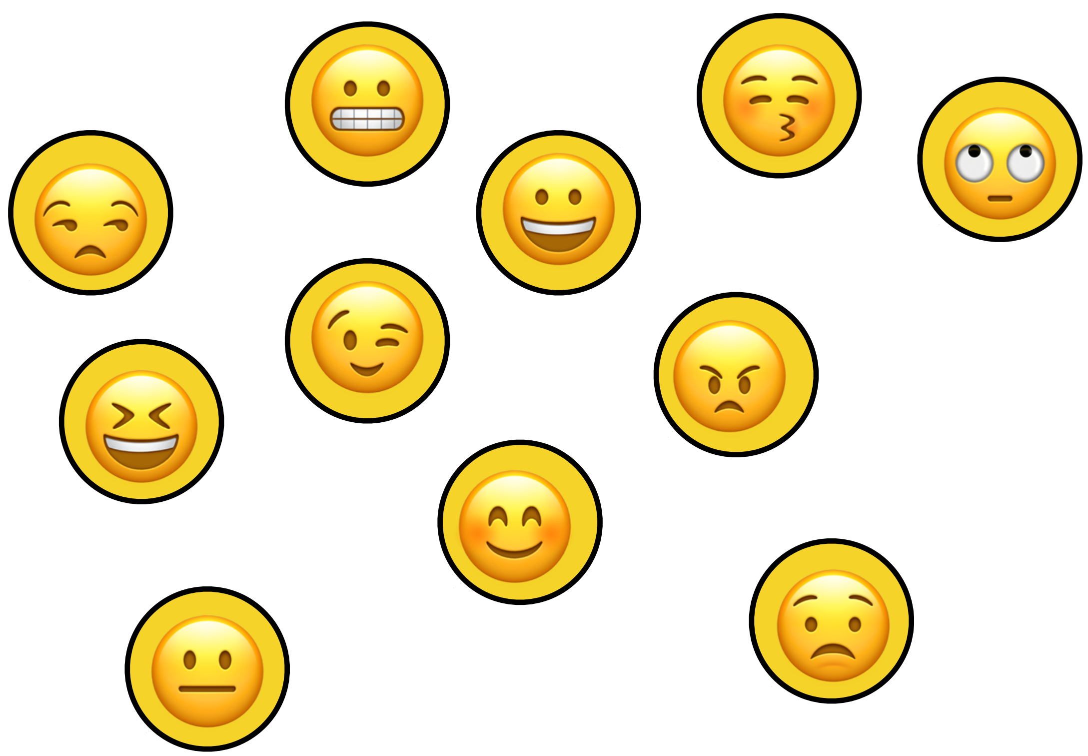
Statistical theory: Sampling distributions

Research Methods for Human Inquiry
Andrew Perfors

Sampling from a population

What are we making inferences about?

- We assume that there exists a **population**
- The population is an abstract concept. It is the people we actually want to know about, like:
  - All people with bipolar disorder over the age of 35
  - All of the people in Australia
  - All native English speakers in the world



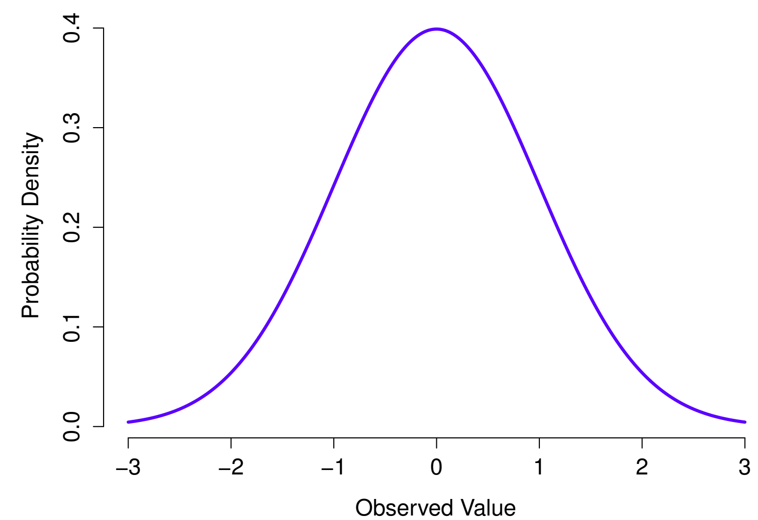
A population is usually very big, much larger than we can realistically study every member of

What are we making inferences about?

average height?

age of learning of first word?

# of episodes before diagnosis?



But luckily all we *really* care about is estimating some property of the population, not measuring everyone for the sake of measuring everyone.

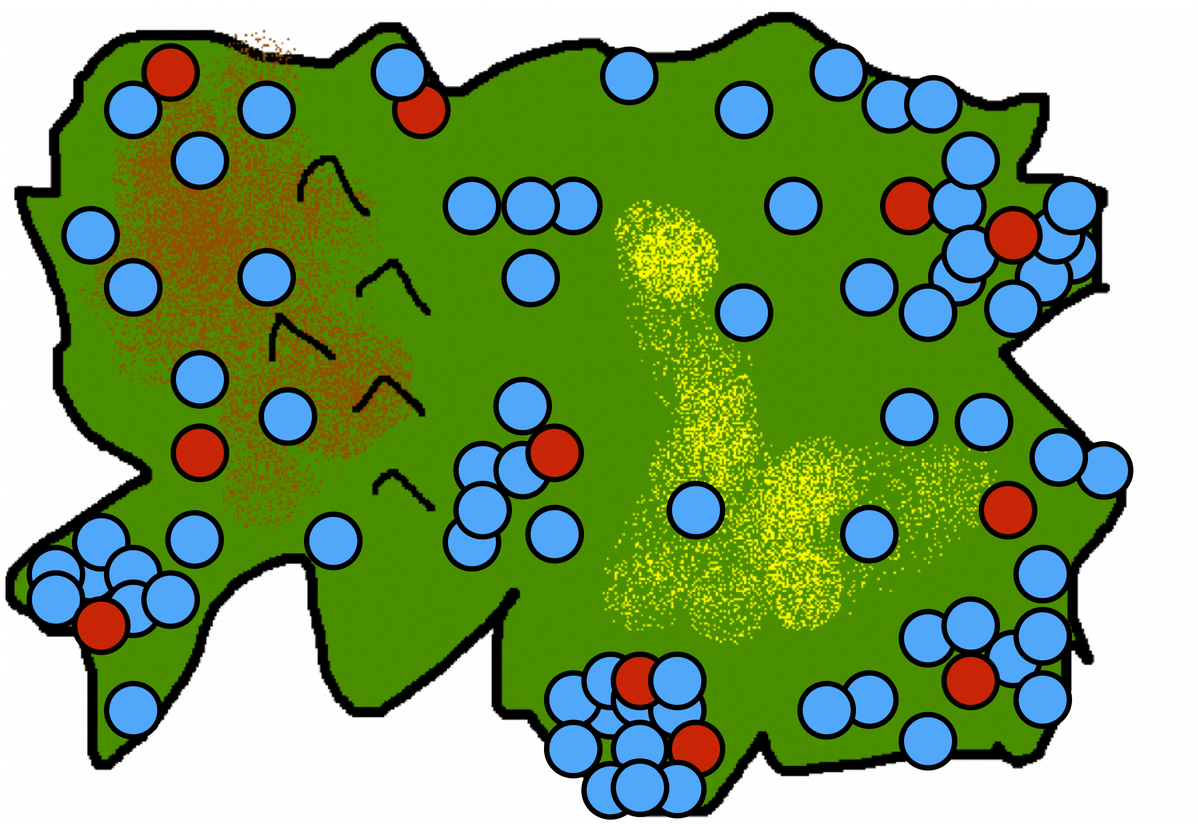
These estimates are called **estimated population parameters.**

Two common ones are mean (μ) and standard deviation (σ)

How do we get good estimates of population parameters?

What are we making inferences <u>about</u>?

- In this case, our dataset was sampled randomly from part of Otherland but we hope it's informative about all of Otherland



- We want to estimate parameters like the quantity of food per person in Otherland (and if it has gone down recently)

Estimating a population parameter

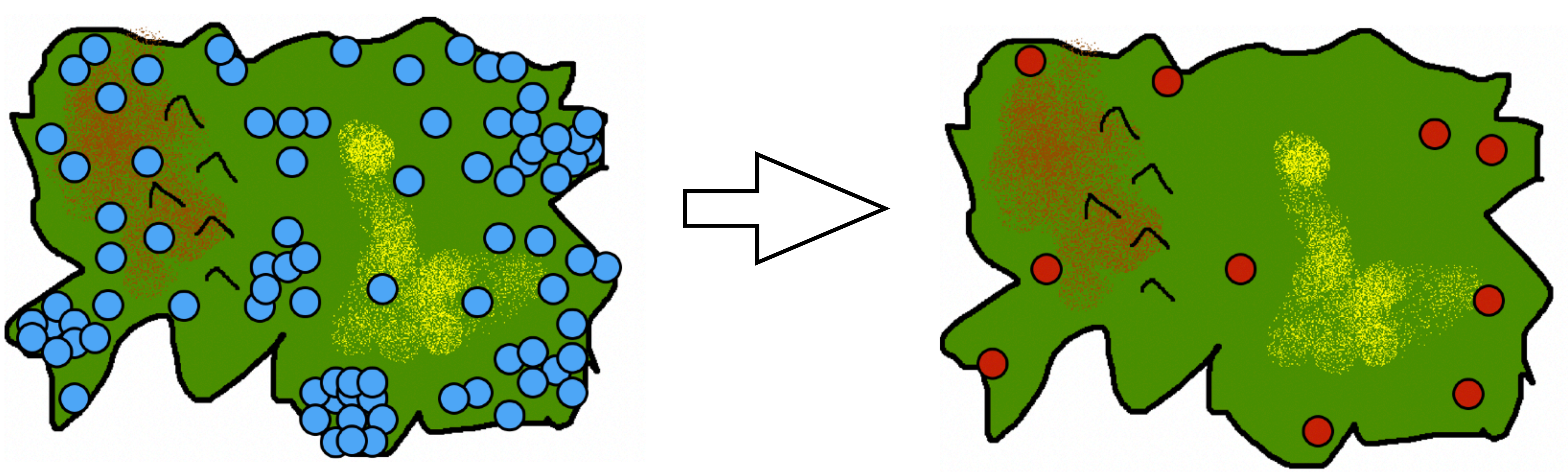
- Our estimate is our **best guess** of the true population parameter **based on our sample** of data
- We estimate this with an "estimator"!
  - (which has a formal definition that the textbook talks about but you don't need to care about)
- Bottom line is: some estimators are better than others
- Good estimators are:
  - **unbiased**: they're too low just as often as they're too high
  - **consistent**: given enough data, they'll eventually give the right answer
  - **low variance**: tend to stay "pretty close" to the true value)

How to estimate a population mean

	"usual" symbol
true population mean	μ
estimated population mean	$\hat{\mu}$

But what is this? How do we calculate it? We don't know the *true* population mean

Answer: we sample from our population and find the mean of that

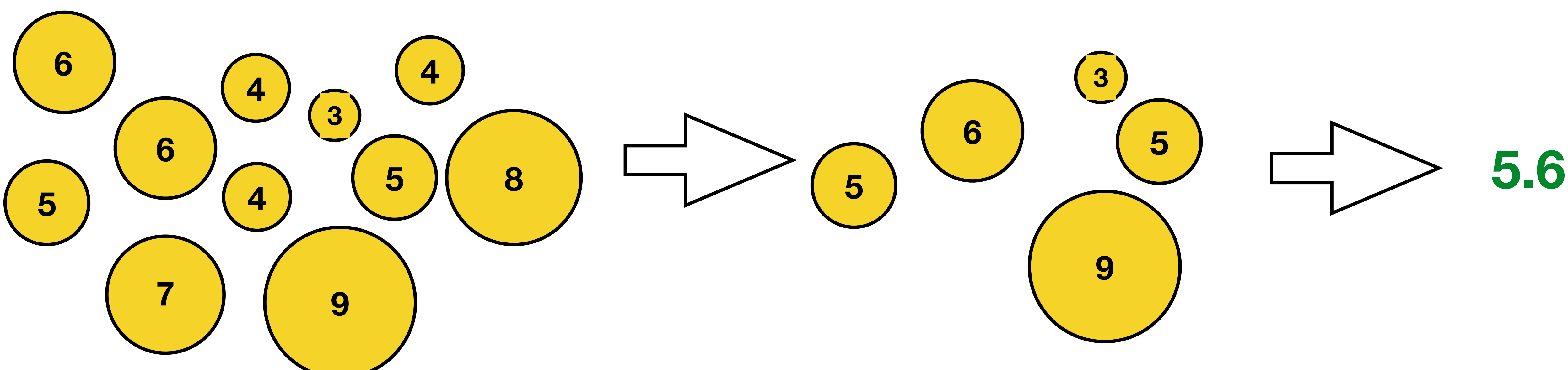


How to estimate a population mean

	"usual" symbol
true population mean	μ
estimated population mean	$\hat{\mu}$

But what is this? How do we calculate it? We don't know the *true* population mean

Answer: we sample from our population and find the mean of that



In numbers…

How to estimate a population mean

	"usual" symbol
true population mean	μ
estimated population mean	$\hat{\mu}$
sample mean	$\bar{X}$ or M

These two are always the same number

This is called the sample mean

Assuming that you have a genuine random sample[*], the sample mean is the best estimator of the true mean.

[*] Usually samples aren't random but we do our best. This is why the validity of your result relies on having a random sample from the population in question. Whether it's "good enough" depends on your study & what you did.

Three things to keep distinct

	"usual" symbol	what is it?	do we know its value?
true population mean	μ	the truth	no
estimated population mean	$\hat{\mu}$	a statistical inference	yes
sample mean	$\bar{X}$ or M	a description of our dataset	yes

Something similar for standard deviation

	"usual" symbol	what is it?	do we know its value?
true population sd	σ	the truth	no
estimated population sd	$\hat{\sigma}$	a statistical inference	yes
sample sd	s	a description of our dataset	yes

This is because the equation for standard deviation is wacky when there is just one data point

Unlike for the mean, these two are NOT the same number

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

x=36
M=36
s=0

Something similar for standard deviation

	"usual" symbol	what is it?	do we know its value?
true population sd	σ	the truth	no
estimated population sd	$\hat{\sigma}$	a statistical inference	yes
sample sd	s	a description of our dataset	yes

This is because the equation for standard deviation is wacky when there is just one data point

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

x=36
M=36
s=0

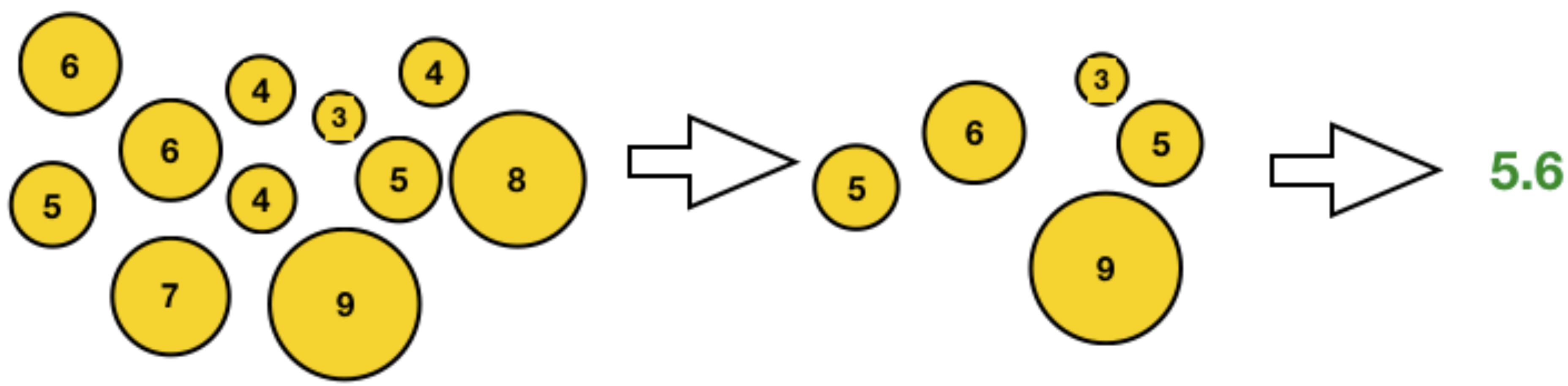
So we usually use the "Bessel-corrected" one as our estimate.

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

x=36
M=36
s=NA

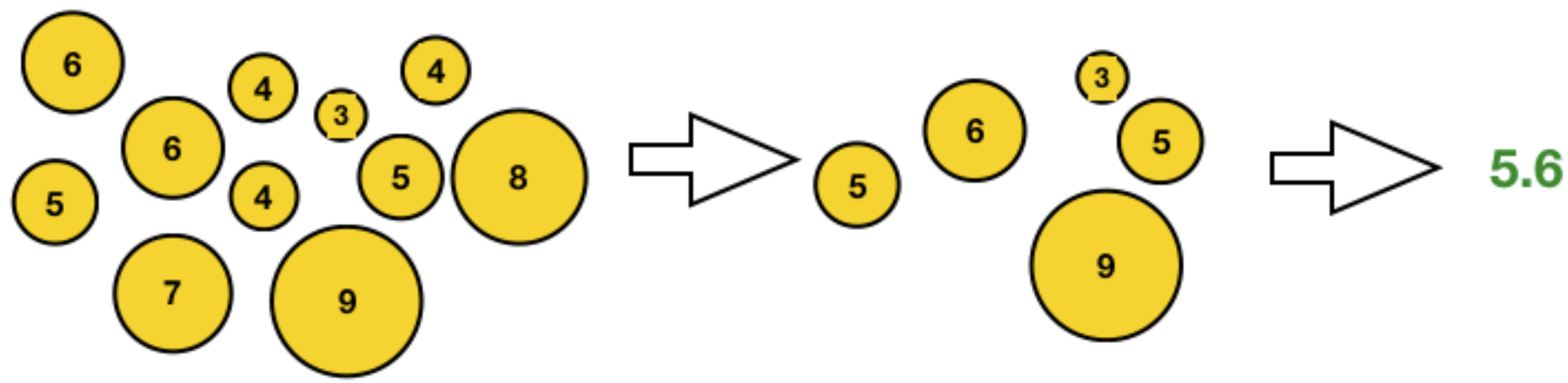
[*] I am not going to ask any exam questions about Bessel corrections. I just wanted to explain so you weren't confused if you tried to take the standard deviation of one data point.

Back to the sample mean…

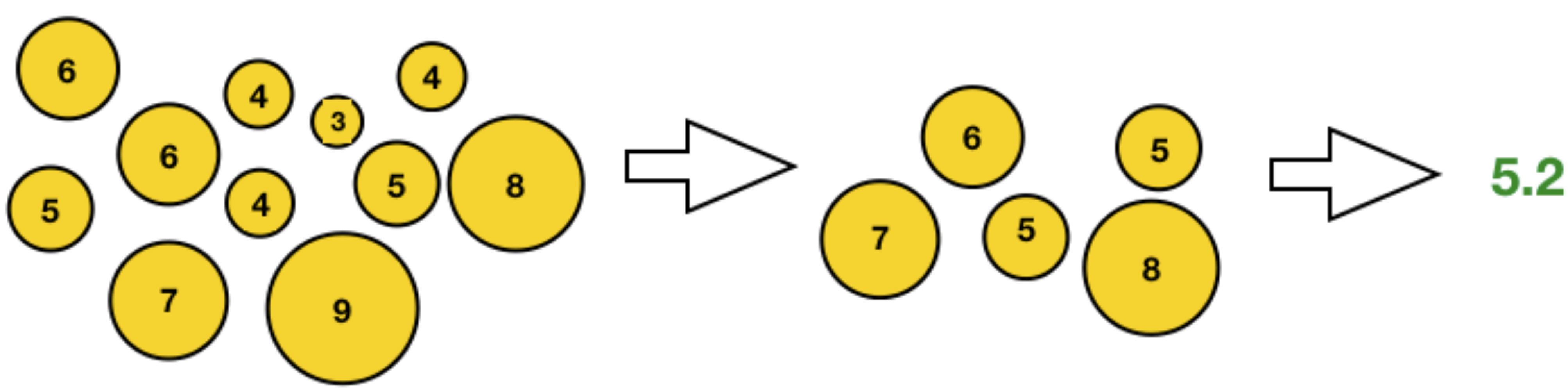


You might have noticed…

The sample mean is just a single best guess given one dataset.



If we had a *different* dataset, we might get a different guess.



You might have noticed…

It would be nice to know what happens if we had lots of different datasets.

Can we say anything about what the means of those datasets would look like?

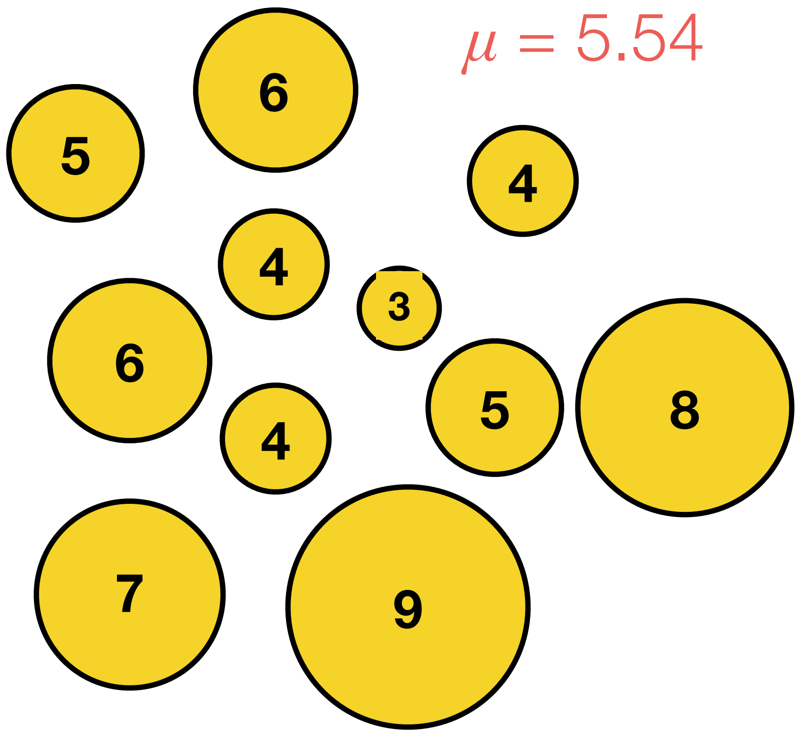
Yes we can! This is the **sampling distribution of the mean**

The sampling distribution of the mean

It is a distribution
of means

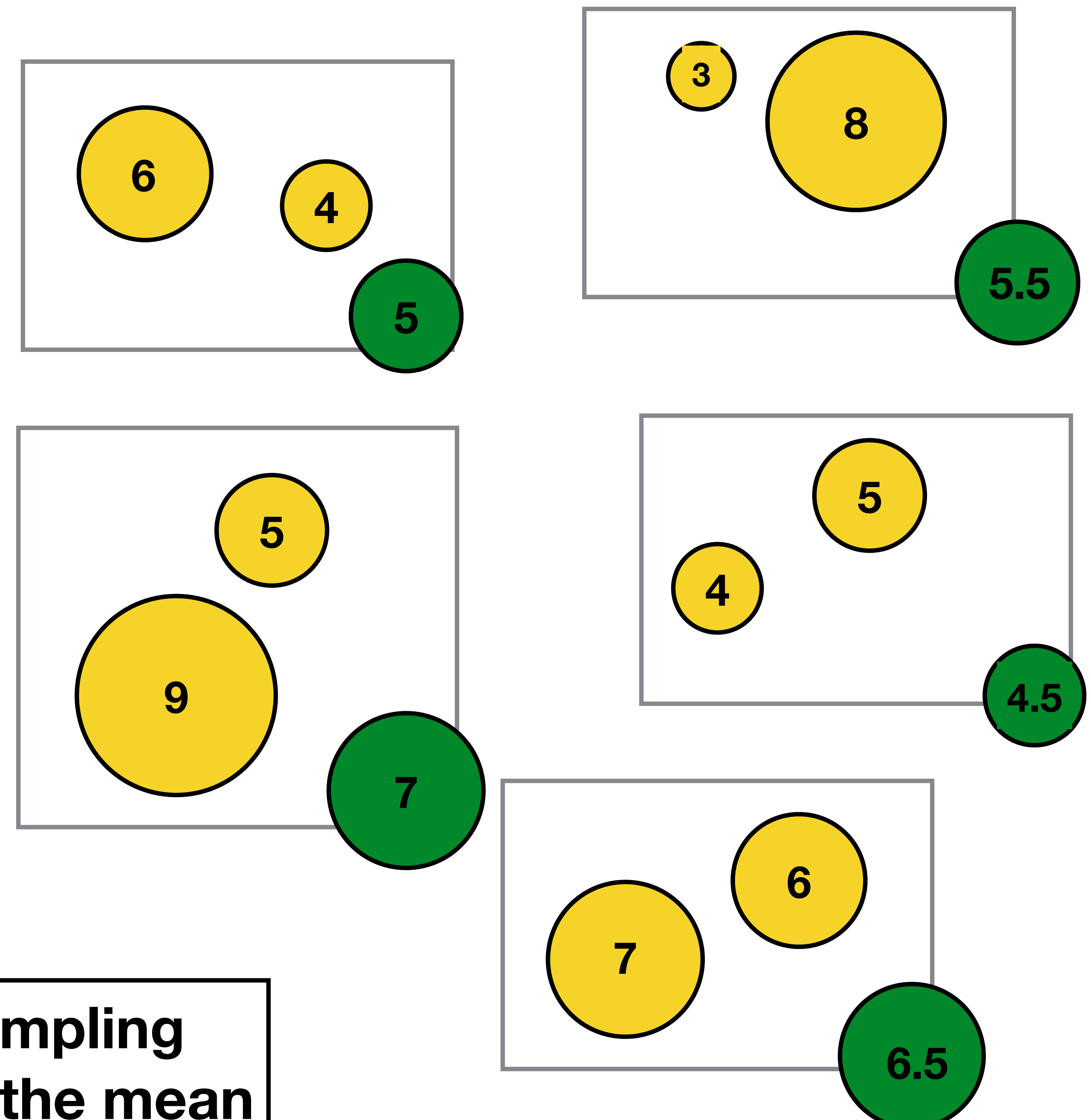
A population of “circles”

$\mu = 5.54$

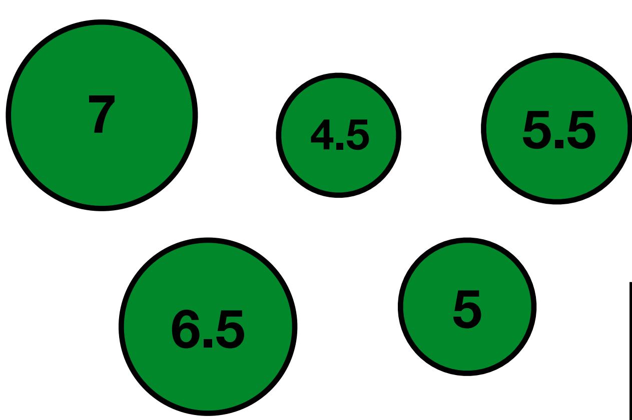


I run a “select two circles experiments”

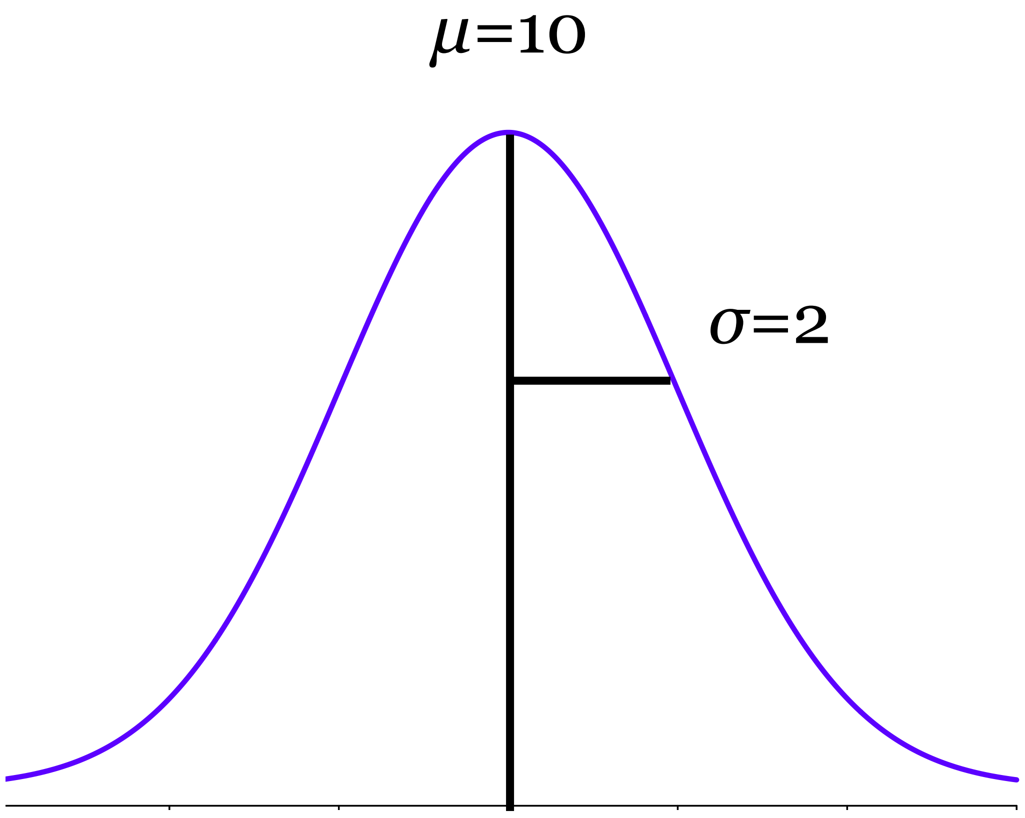
If I run a lot of “select two circles experiments” this is a <u>population</u> of such experiments



These means are the <u>population of means</u> from the set of two-circle experiments



This is the **sampling distribution of the mean**

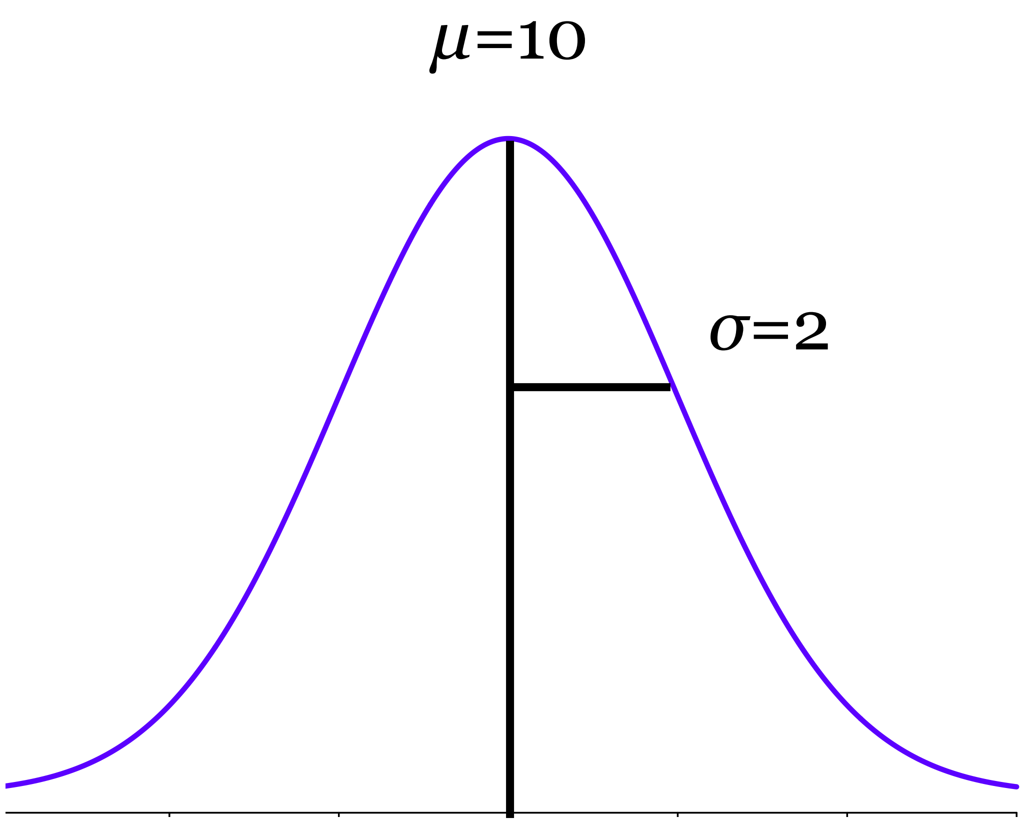


Population distribution

Okay, now let's see an example with real numbers

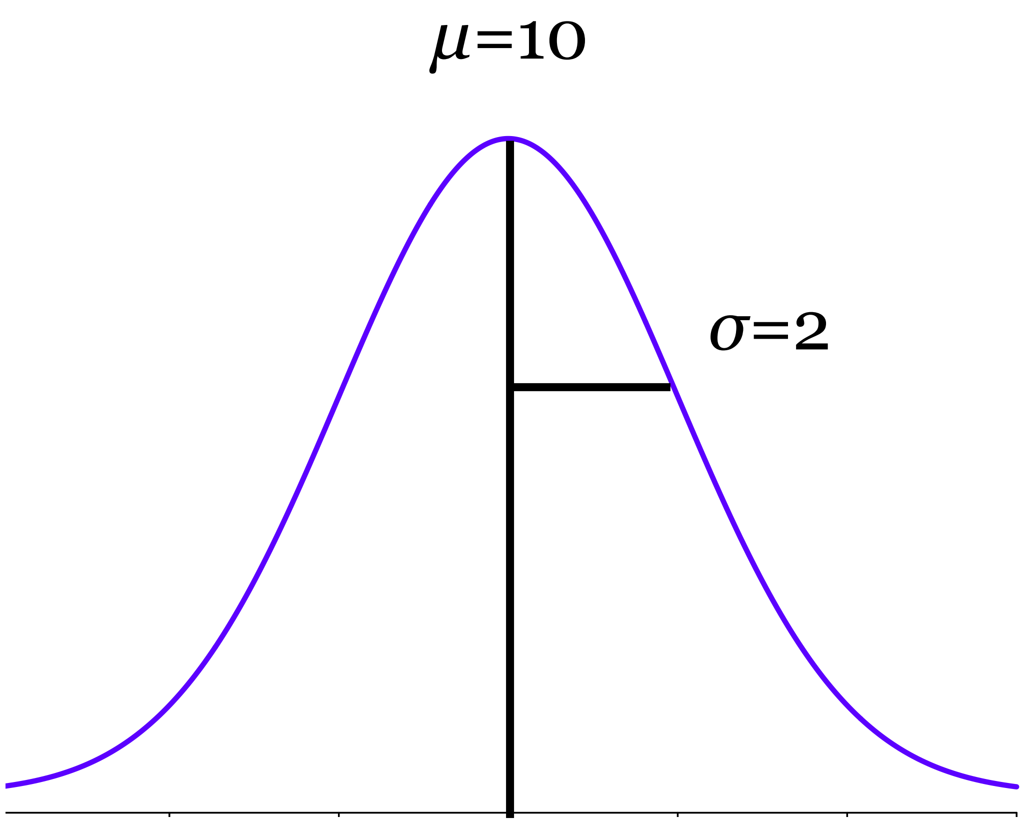
```
round(  rnorm(n=5,mean=10,sd=2), digits=1)
```

X_1	X_2	X_3	X_4	X_5	$\bar{X}$
11.2	12.1	14.3	10.9	8.5	11.4



Population distribution

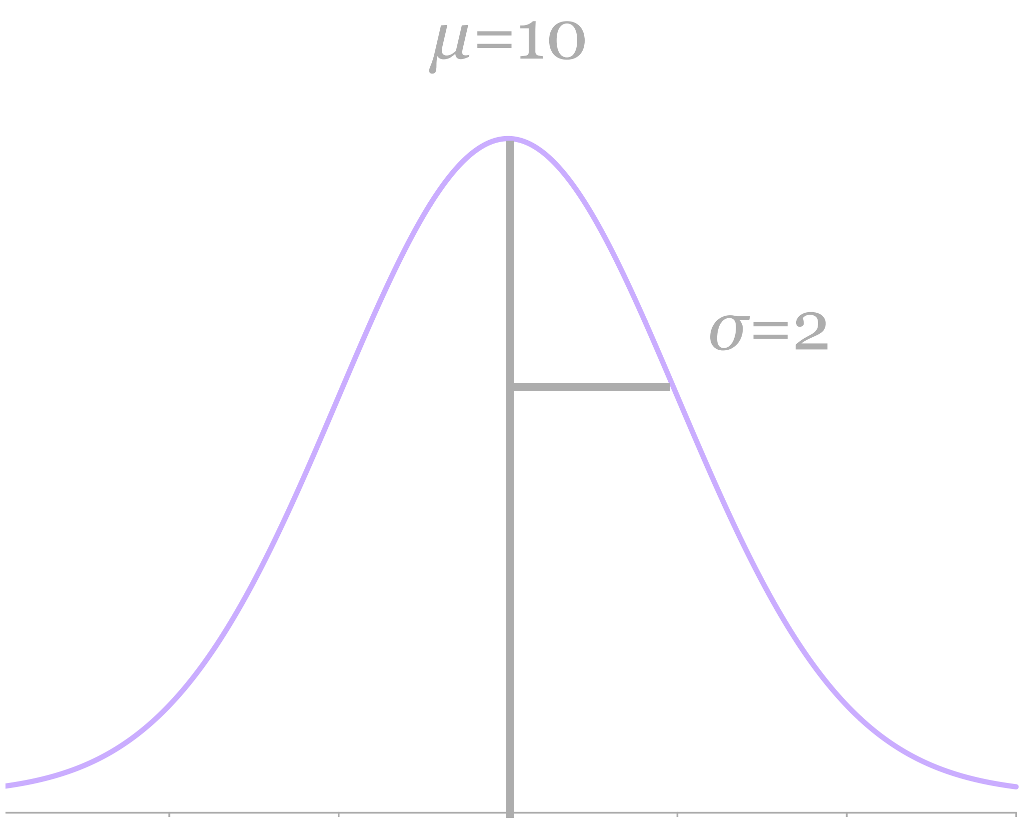
My experiment produces a simple random sample of five observations from this population, and it therefore produces a sample mean



Population distribution

Replication	X_1	X_2	X_3	X_4	X_5	$\bar{X}$
1	11.2	12.1	14.3	10.9	8.5	11.4
2	10.2	6.7	10.8	11.0	7.8	9.3
3	9.6	6.1	12.6	7.5	7.3	8.6
4	9.2	9.8	12.0	7.3	9.8	9.6
5	9.1	10.9	7.6	8.3	14.5	10.1
6	8.3	8.3	10.5	13.2	10.6	10.2
7	10.6	6.2	6.4	10.5	6.8	8.1
8	10.9	10.8	12.4	11.7	9.1	11.0
9	11.9	10.2	13.4	11.4	12.8	12.0
10	6.4	8.1	13.5	11.5	9.4	9.8

Here are 10 replications of the experiment, each with their own sample mean



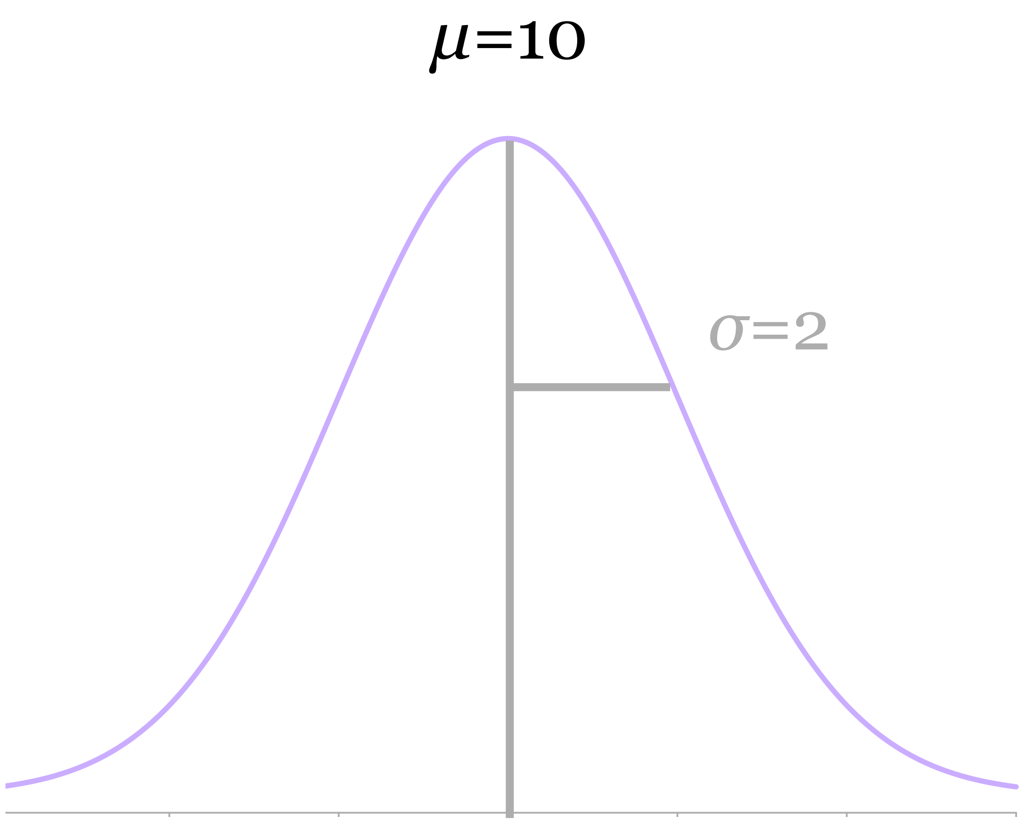
Population distribution

Replication	X_1	X_2	X_3	X_4	X_5	$\bar{X}$
1	11.2	12.1	14.3	10.9	8.5	11.4
2	10.2	6.7	10.8	11.0	7.8	9.3
3	9.6	6.1	12.6	7.5	7.3	8.6
4	9.2	9.8	12.0	7.3	9.8	9.6
5	9.1	10.9	7.6	8.3	14.5	10.1
6	8.3	8.3	10.5	13.2	10.6	10.2
7	10.6	6.2	6.4	10.5	6.8	8.1
8	10.9	10.8	12.4	11.7	9.1	11.0
9	11.9	10.2	13.4	11.4	12.8	12.0
10	6.4	8.1	13.5	11.5	9.4	9.8

These numbers come from the population distribution

These numbers come from the sampling distribution of the mean

The sampling distribution is less variable

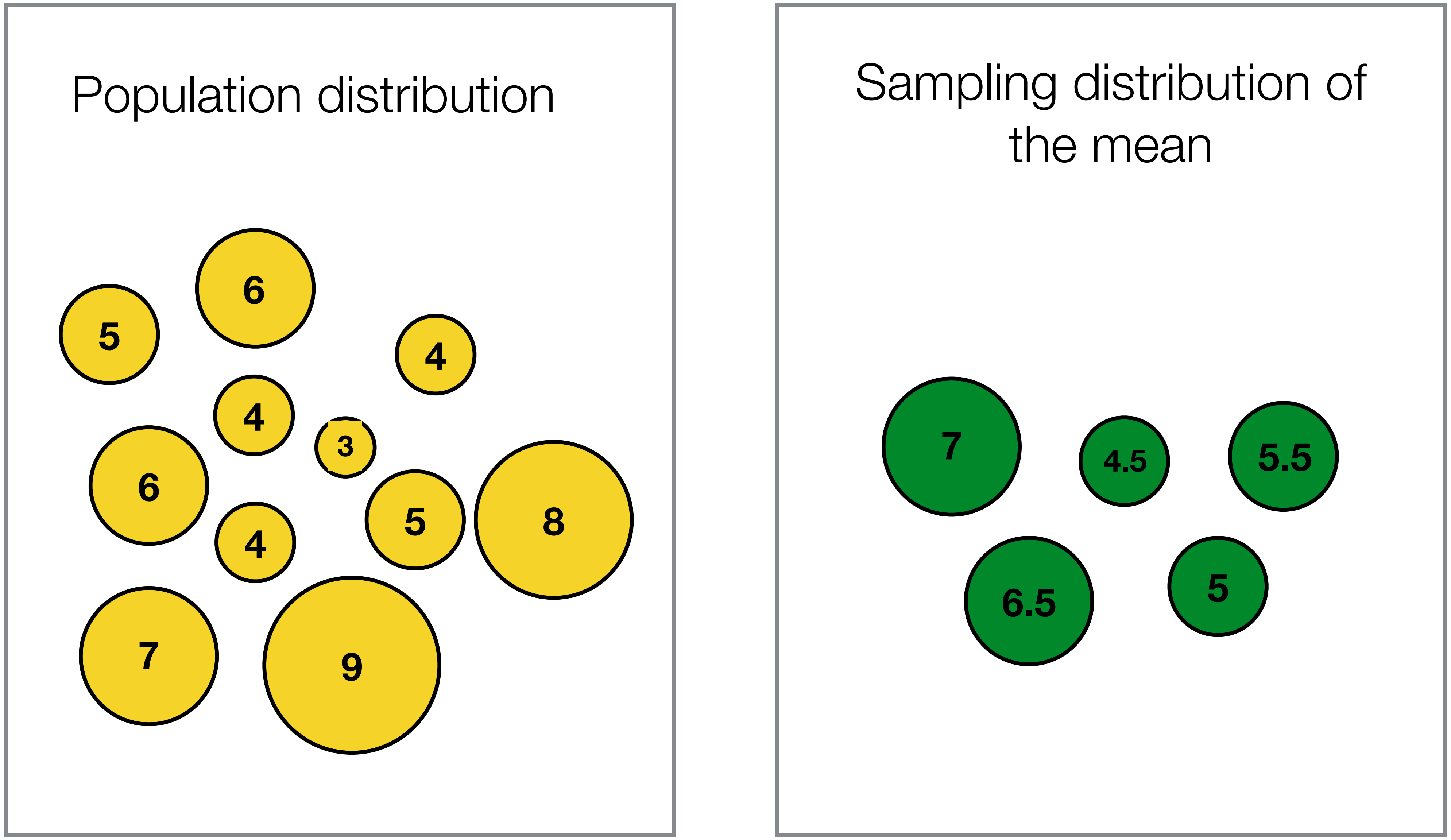


Population distribution

Replication	X_1	X_2	X_3	X_4	X_5	$\bar{X}$
1	11.2	12.1	14.3	**10.9**	8.5	11.4
2	**10.2**	6.7	**10.8**	11.0	7.8	**9.3**
3	**9.6**	6.1	12.6	7.5	7.3	8.6
4	**9.2**	**9.8**	12.0	7.3	**9.8**	**9.6**
5	**9.1**	**10.9**	7.6	8.3	14.5	**10.1**
6	8.3	8.3	**10.5**	13.2	**10.6**	**10.2**
7	**10.6**	6.2	6.4	**10.5**	6.8	8.1
8	**10.9**	**10.8**	12.4	11.7	**9.1**	11.0
9	11.9	**10.2**	13.4	11.4	12.8	12.0
10	6.4	8.1	13.5	11.5	**9.4**	**9.8**

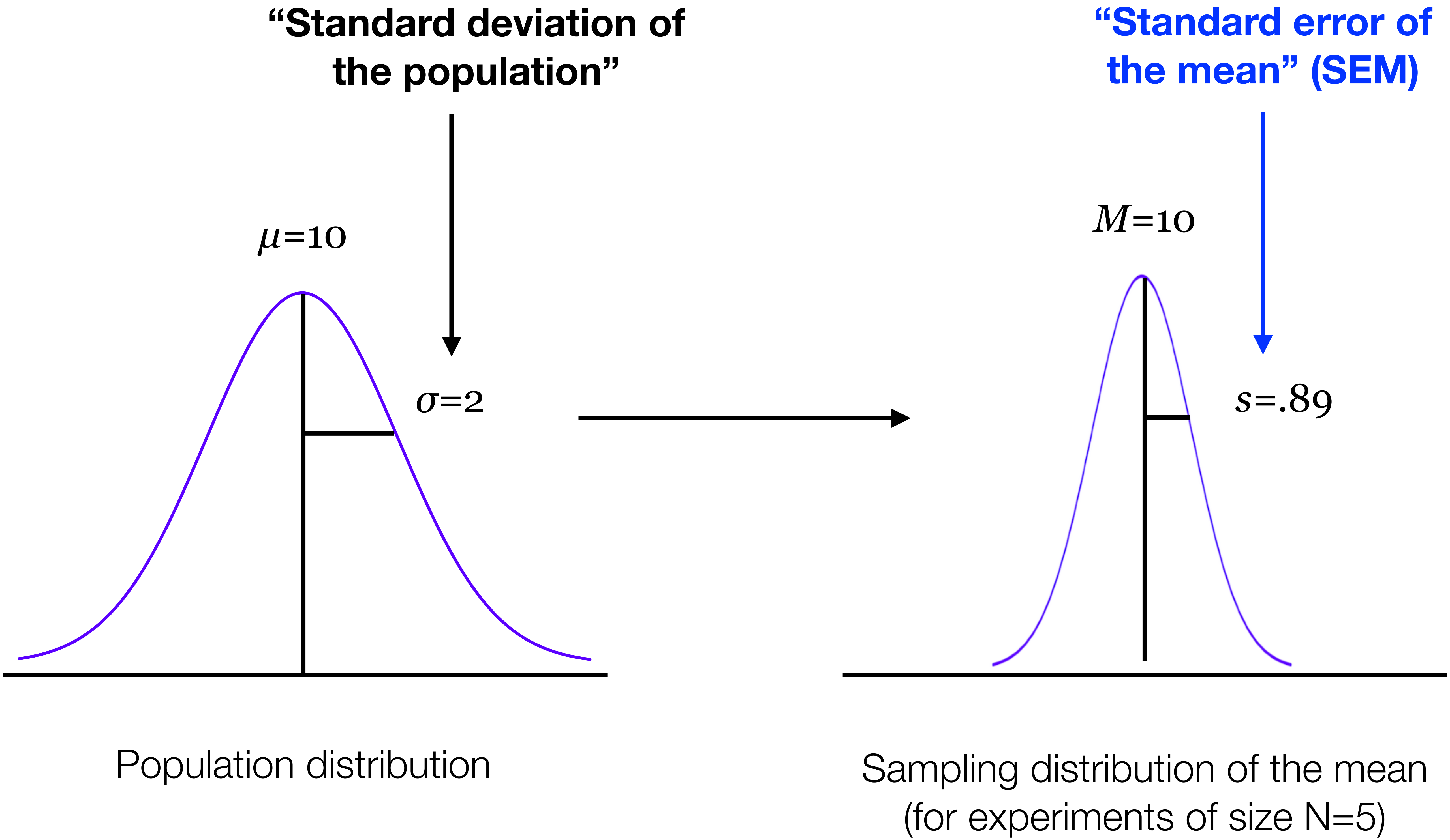
36% of the observations are "close" to the population mean

50% of the sample means are "close" to the population mean



The sampling distribution of the mean is less variable

Intuition: it is the set of means. To get an extreme (very high or very low) mean in an experiment, your experiment had to have had only very high or very low items. This is increasingly unlikely (especially with large sample size) so many sample means tend to be close to the true population mean



The sampling distribution is less variable

The formula is simple

“Standard error of the mean” (SEM)

“Standard deviation of the population”

$$\mathrm{SEM} = \frac{\sigma}{\sqrt{N}}$$

Square root of the sample size N

So as your sample size grows, the variance (i.e., uncertainty about the mean) goes down

There's code for this

In **w5day1analysis.Rmd** file — you don't need to know how to write this, but you can play with it if you want!

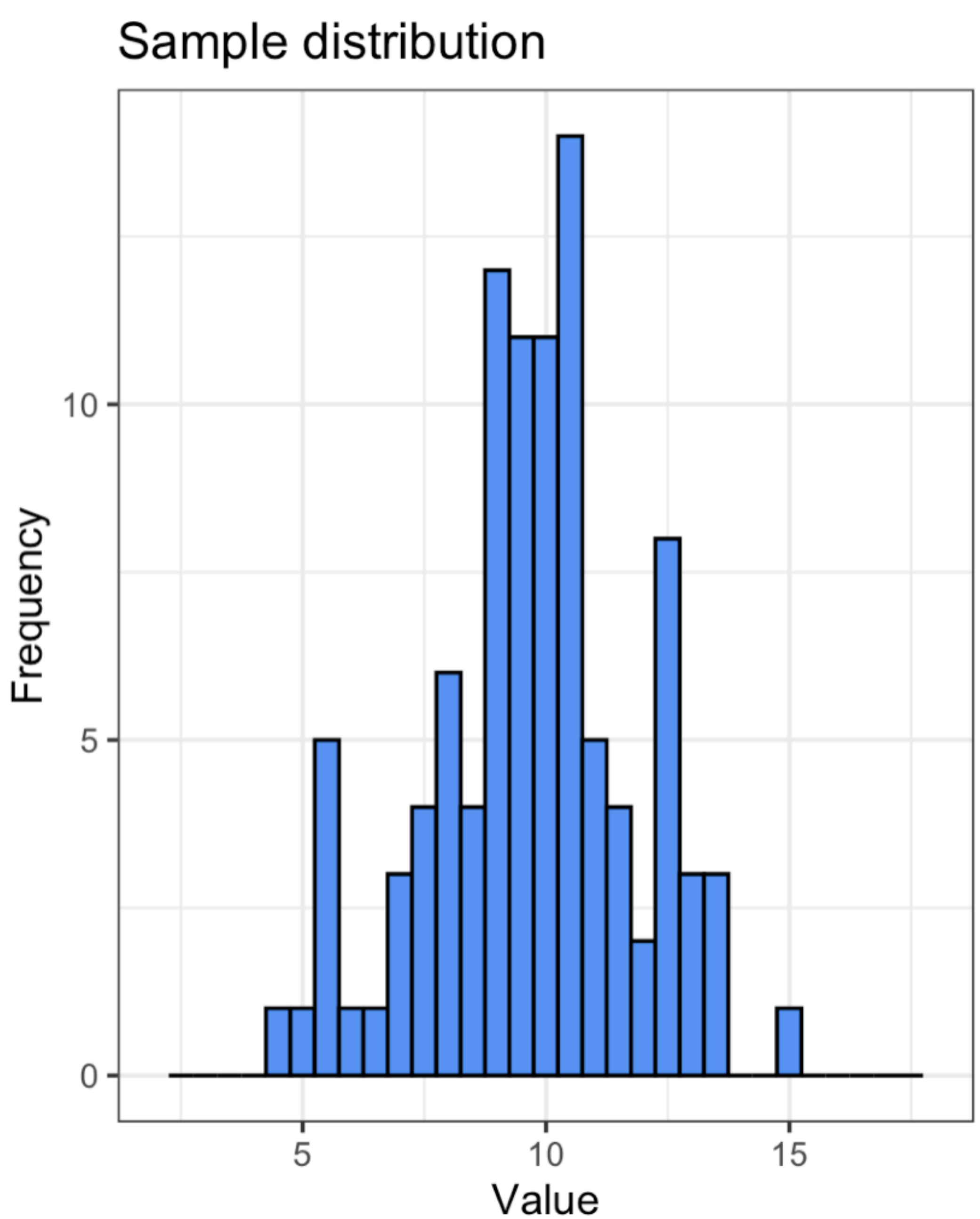
```
```{r showsamplingdists, echo=FALSE, warning=FALSE}
# set trueMean and trueSD (you can play with this if you want by changing them)
trueMean <- 10
trueSD <- 2
# number of experiments
nExperiments <- 100
# number of people per experiment
nPeople <- 200

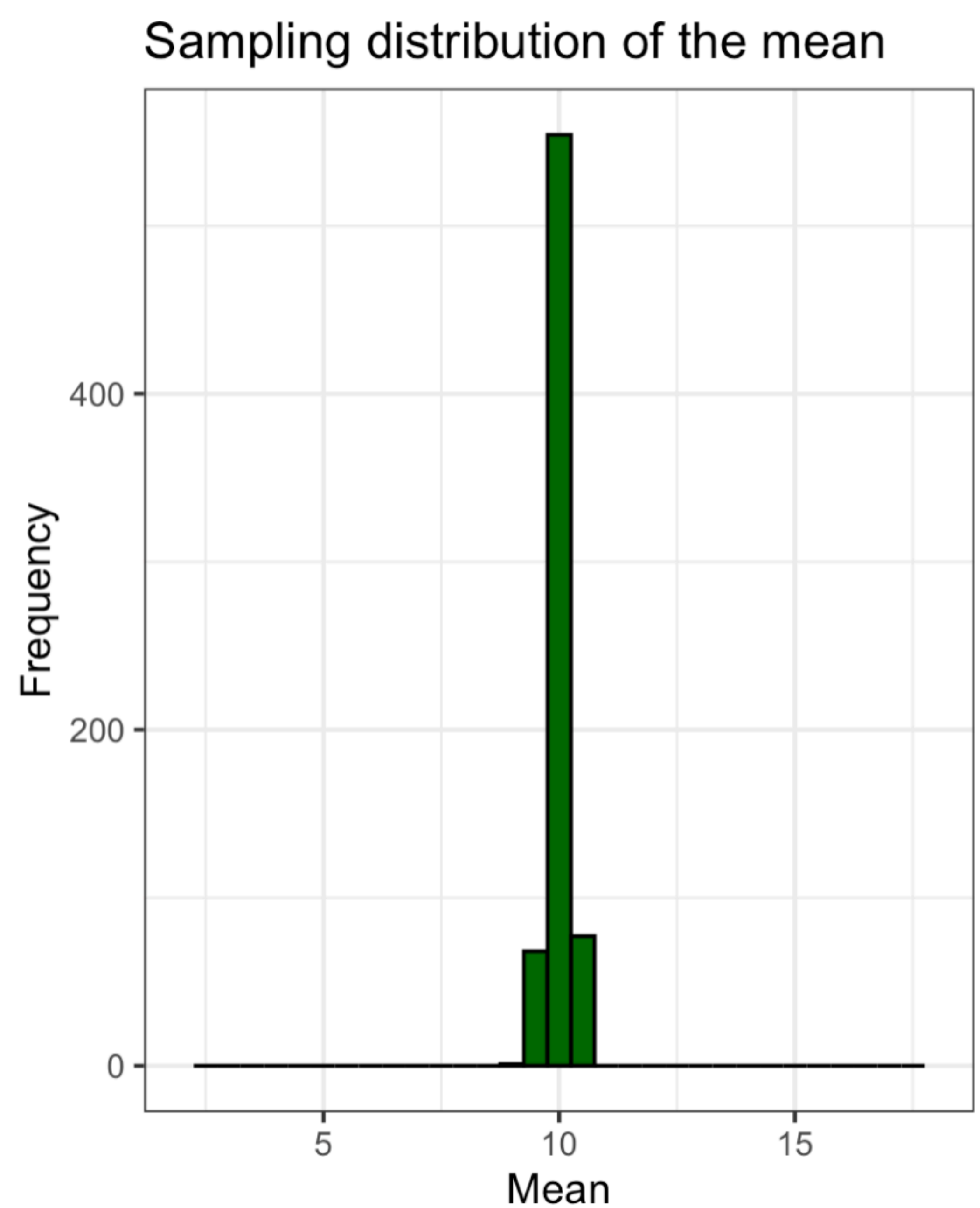
samplingDistMean <- NULL

for (ne in 1:nExperiments) {
  sample <- round(rnorm(n=nPeople,mean=trueMean,sd=trueSD),
                  digits=1)
  samplingDistMean <- c(samplingDistMean,round(mean(sample),digits=1))
}
```
```

There's code for this

In `w5day1analysis.Rmd` file — you don't need to know how to write this, but you can play with it if you want!





See the `w5day1exercises.Rmd` file for the exercises!