Lec 18 (Chapter 13): Sampling Distributions

Corinne Riddell/Alan Hubbard

October 7, 2022

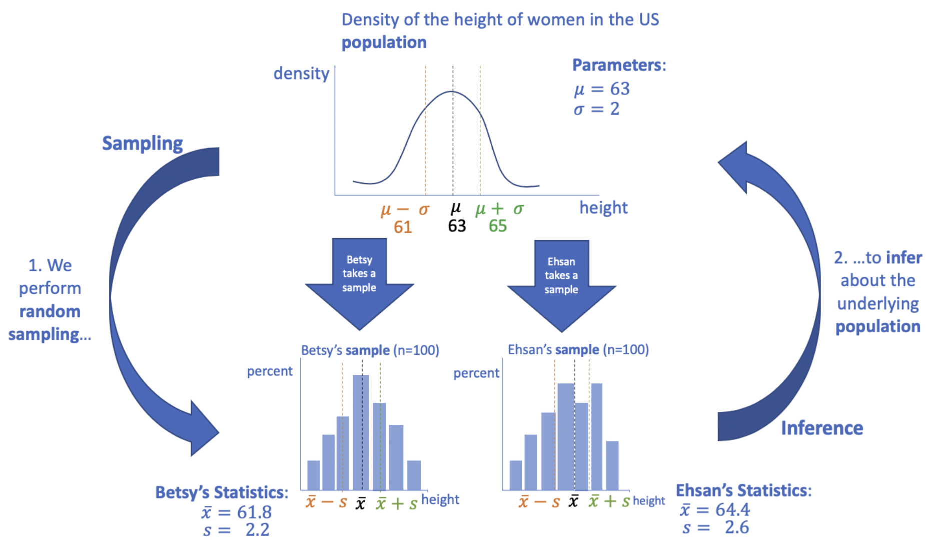
Learning objectives for today

- Define what a sampling distribution for the mean is.
- Investigate the properties of the sampling distribution of the mean and proportion
  - What is the mean of the sampling distribution of the mean (or proportion)?
  - What is the standard deviation for the sampling distribution of the mean (or proportion)?
  - What distribution does it follow?
- Learn about the Central Limit Theorem
- Learn about the Law of Large Numbers

Readings

- Chapter 13 of Baldi and Moore
- Online resource: Sample distribution of a mean, see 10.3.1; 322 of pdf; 309 of textbook, Central Limit Theorem Visual Representation, see page 327 of pdf; page 314 of textbook)

Recall from Chapter 9 the conceptual diagram linking sampling and inference



Parameter and statistic

Parameter: A number that describes the population. Generally the parameter value is unknown, because we cannot often examine the entire population. Our goal is to estimate its value.

Statistic (estimate): A number that can be computed from a sample. We using a sample statistic to estimate a parameter value.

Estimator: An algorithm that reads in the data and spits out the estimate.

Parameter and statistic

μ and p are **population parameters** for the mean and proportion. There is one unique value for each of μ and p in the underlying population.

Note that the p is also a mean (of a 0-1 random variable).

$\bar{x}$ and $\hat{p}$ are **statistics** computed using **samples**. We refer to them as the sample mean and sample proportion, respectively. If we change the sample our statistics will likely also change. Statistics vary across samples. So, estimators produce a random variable when operated on data.

In practice we only take one sample. If our variable of interest is continuous (say, annual income), then we use our sample estimate $\bar{x}$ to estimate the population parameter μ. Very likely, our estimate will deviate from the true value.

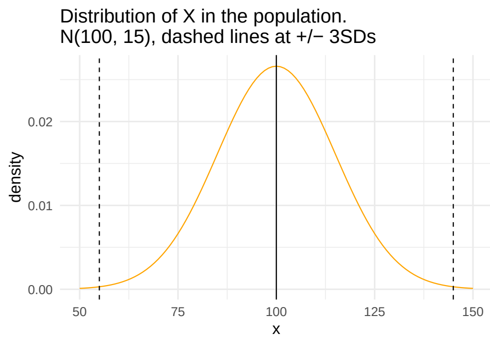
To understand how good our estimate is, we study how much its value varies if we were to take multiple, repeated samples. If its value does not vary much and if it's value is centered on the true value, then we can say that our estimate is probably close to the true population parameter value.

Statistics are random variables

- When we choose a sample **randomly** then the value of the statistic (say the mean) will vary from one sample to the next. E.g., if you and I both choose a random sample, the mean from our samples will be different.
- In practice we only choose one sample and compute one sample mean. However, it is useful to understand the **sampling distribution** of the mean so we can better quantify how much our statistic might differ from the population mean.
- The **sampling distribution** is the histogram based of the sample means (or proportions) for across every possible sample we could take from the distribution of a specified size.
- When we studied the Binomial distribution, we made a histogram of an approximate sampling distribution for X, the number of successes. We will do similar exercises here for the sample mean $\bar{x}$ and the sample proportion $\hat{p}$.

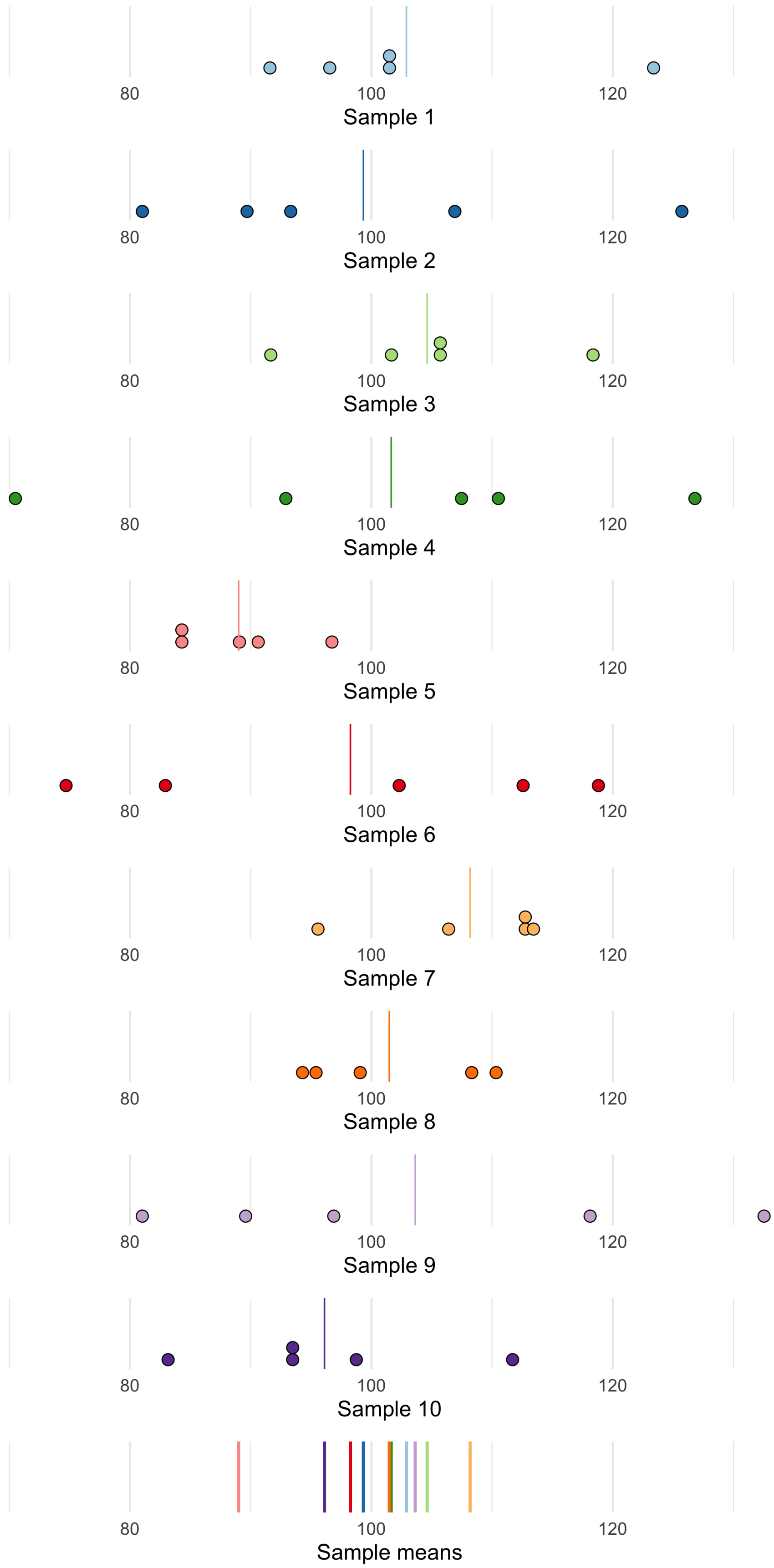
Repeated sampling

- The concept of a sampling distribution is rooted in the idea of repeated experiments (of the type done which generate the data of interest).
- Suppose we have X, a Normally distributed random variable where $X \sim N(100, 15)$.



- We could take a sample from X, say of 5 people, and based on this sample's mean $\bar{x}$ try and estimate μ (which we know to equal 100 in this example).
- The next slide shows the sampled data (using filled circles) and the sample mean using a vertical line. Pretend that 10 different students each took their own samples, independently of one another.
- The bottom panel shows **only** the sampled means.

Repeated sampling

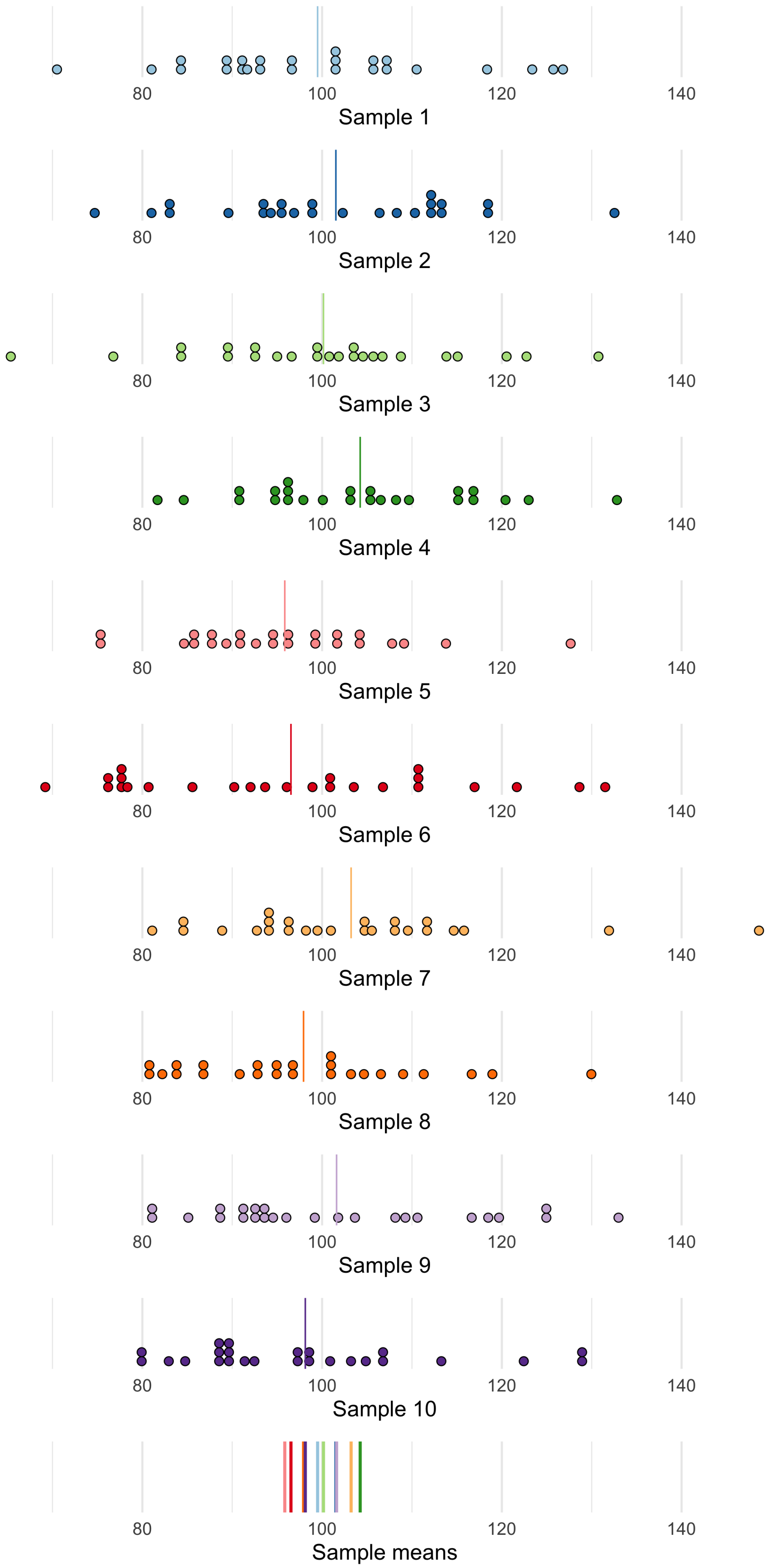


Repeated sampling

What do you notice about the spread of the sample means (see the bottom panel on previous slide) vs. the spread of the sampled data (see all the other panels)?

Let's repeat this process but for $n = 25$. That is, let's increase the sample size.

Repeated sampling



Repeated sampling

What is different about the last plot (i.e., the plot of the sample means) for $n = 25$ vs. $n = 5$?

When the sample size is larger, the spread of the sample means is smaller.

We only had 10 sample means. But imagine if we had the sample mean for every possible sample of size n from the population. This distribution is called the **sampling distribution**.

Sampling distribution of the mean $\bar{x}$

The sampling distribution of the mean is the distribution of values taken by the statistics **in all possible samples** of the same size from the population.

For a simple random sample of size n, the sampling distribution of the mean $\bar{x}$ is centered at μ the population mean and has a standard deviation of $\frac{\sigma}{\sqrt{n}}$

This is true for *any* population, provided that the population is much larger than the sample. A rule of thumb is that the population should be at least 20 times larger than the sample.

Sampling distribution of the mean $\bar{x}$

- Because the average of the $\bar{x}$ across all possible samples equals μ we say that $\bar{x}$ is an unbiased estimator of the parameter μ.
- How close any individual estimate falls to the parameter is quantified by the spread of the sampling distribution. The standard deviation of $\bar{x}$ is called the **standard error**. The standard error is equal to $\frac{\sigma}{\sqrt{n}}$ for the sampling distribution of the mean.
- The standard error of the sample mean is smaller than the standard deviation of the distribution of sampled values. That is, averages are less variable than individual observations.

Sampling distribution of a sample mean for a Normal population

- If individual observations have a $N(\mu, \sigma)$ distribution, then the sample mean $\bar{x}$ of a simple random sample of size n has a $N(\mu, \frac{\sigma}{\sqrt{n}})$
- What does the standard deviation of the sample mean tell us:
  - As n increases, the standard deviation of the sample means ____________.
  - As σ increases, the standard deviation of the sample means ___________.
  - If n increases by a factor of 100, the standard deviation of the sample ___________ by a factor of ________.

Example of the sampling distribution when the underlying population is Normally distributed

Suppose that IQ scores follow a N(100, 15) distribution. This is the population distribution of IQ scores.

What is the mean and standard deviation for the sampling distribution if n = 25? What distribution does it follow?

1. The mean of the sampling distribution is 100, because it is an unbiased estimator
2. The standard deviation of the sampling distribution is $\frac{\sigma}{\sqrt{n}} = 15/5 = 3$. Thus, the sample means are much less variable than the individuals observations.
3. Because the underlying data are Normal, the sampling distribution follows a Normal distribution, here $\bar{x} \sim N(\mu = 100, \sigma/\sqrt{n} = 3)$.

The Central Limit Theorem

- From the previous slide, we know that the shape of the sampling distribution of a sample mean is Normally distributed when the data are Normally distributed to begin with.
- What about for skewed data? Or bimodal data? Or even a binary outcome? What is the shape of the sampling distribution of the average (sample mean)?
- As n increases, the shape of the sampling distribution becomes more and more Normal looking, no matter what the shape of the underlying distribution looked like, so long as the standard deviation is a finite value.

The Central Limit Theorem (CLT)

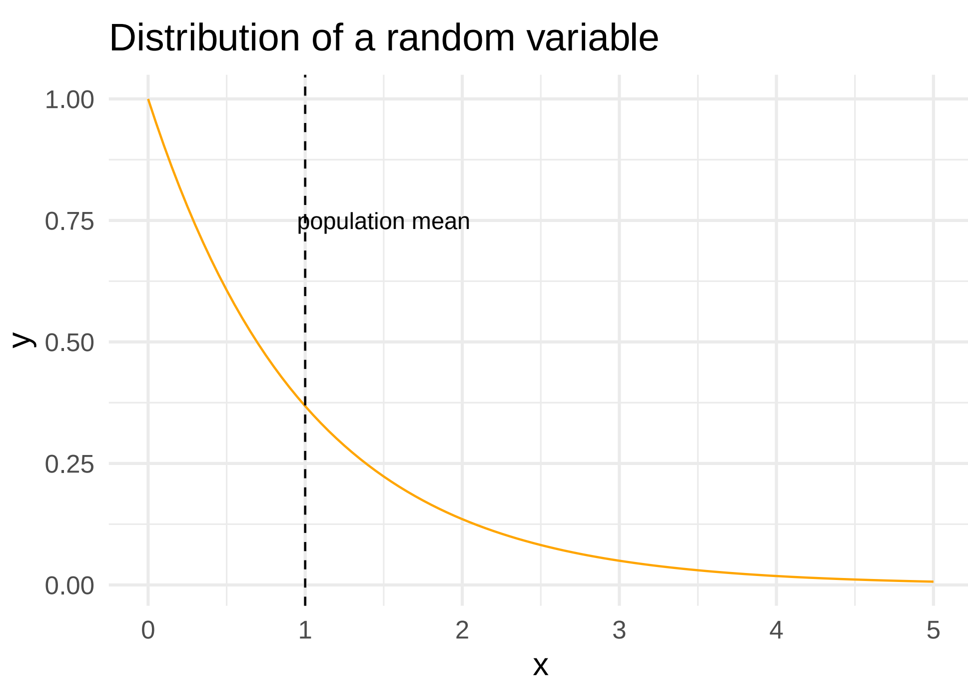
Draw a simple random sample of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \dot{\sim} N(\mu, \frac{\sigma}{\sqrt{n}})$$

The CLT allows us to use Normal probability calculations to answer questions about sample means from many observations (questions relying on the sampling distribution of the sample mean) even when the population distribution is not Normal.

CLT example

Suppose you had a variable whose probability distribution function looked like this. It is strongly skewed right.:

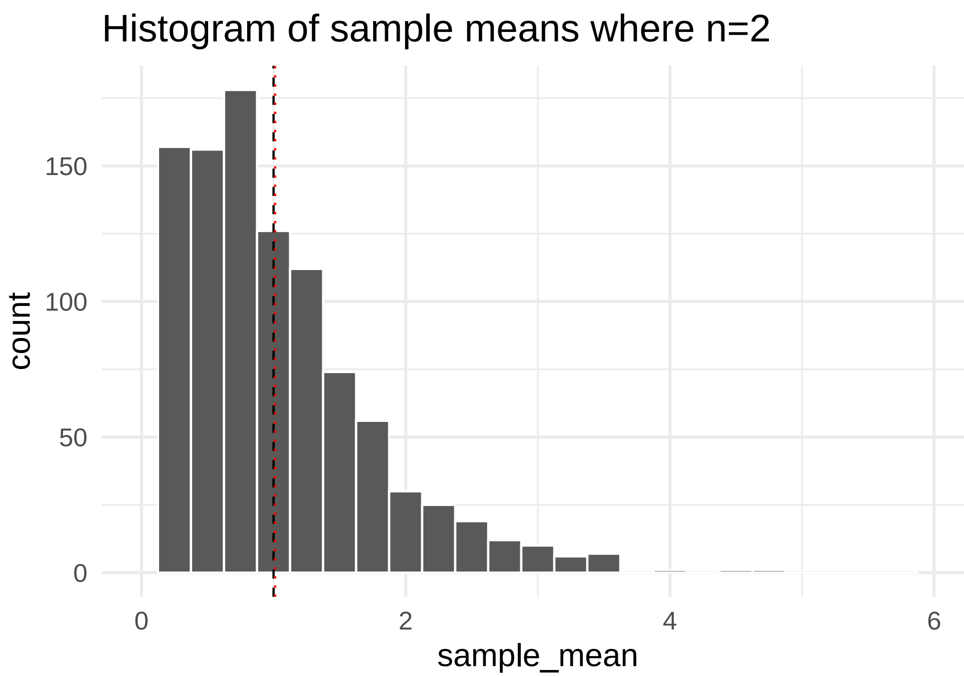


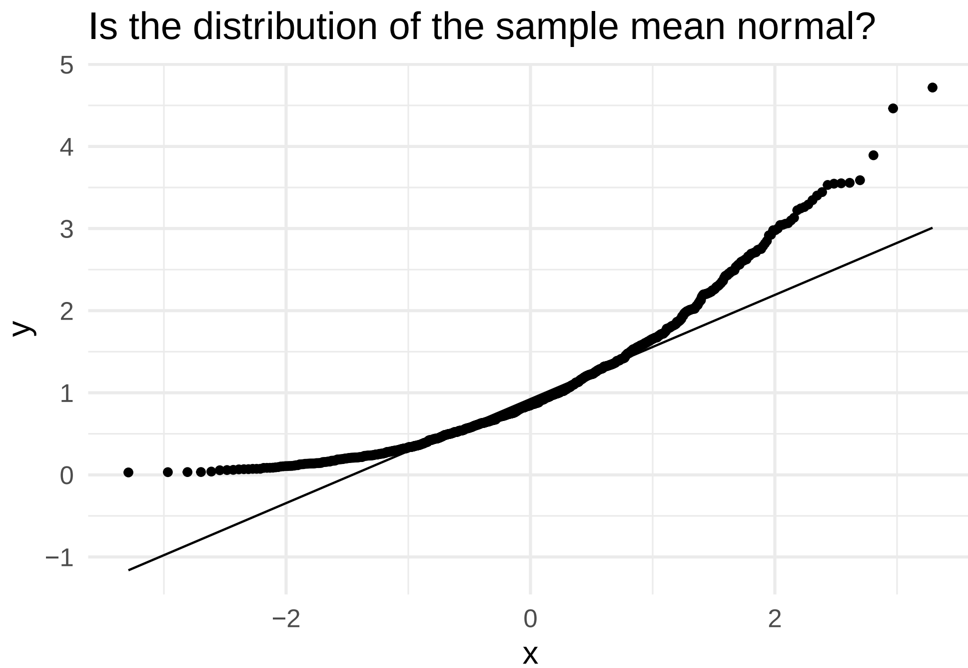
CLT example

1) Take a sample from the distribution of size 2. This is very small!
2) For your sample calculate the mean.
3) Repeat steps 1 and 2 1,000 times.
4) Plot the sampling distribution of the mean. That is, plot the distribution of the 1000 means using a histogram.

Compare the population mean (black dashed line) with the mean of the sampled means (red dotted line)

CLT example



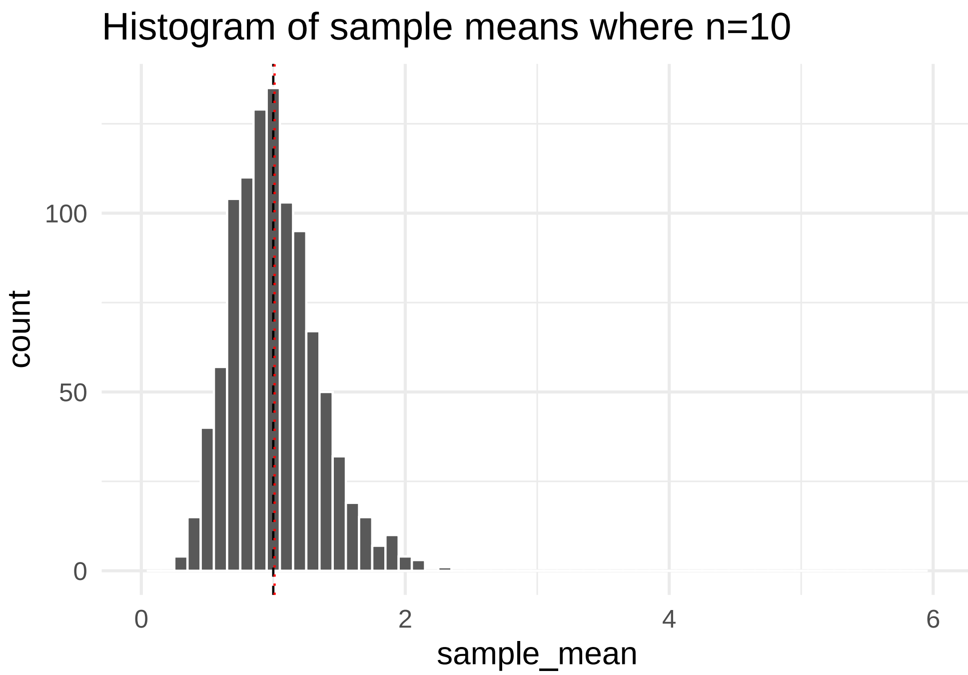


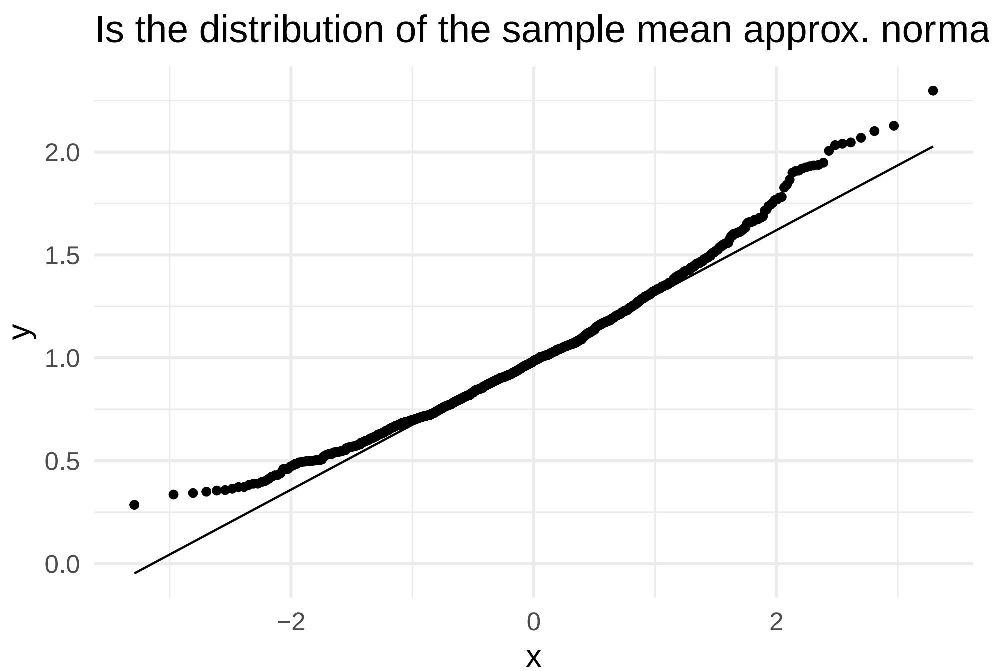
CLT example

1) Take a sample from the distribution of size 10. This is very small!
2) For your sample calculate the mean.
3) Repeat steps 1 and 2 1,000 times.
4) Plot the sampling distribution of the mean. That is, plot the distribution of the 1000 means using a histogram.

Compare the population mean (black dashed line) with the mean of the sampled means (red dotted line)

CLT example





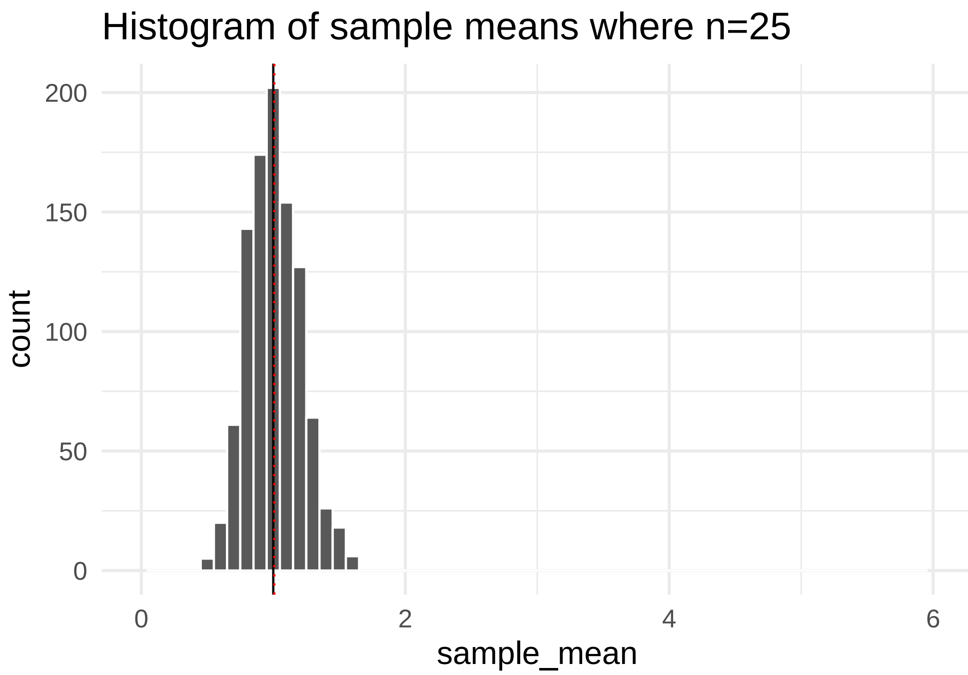
CLT example

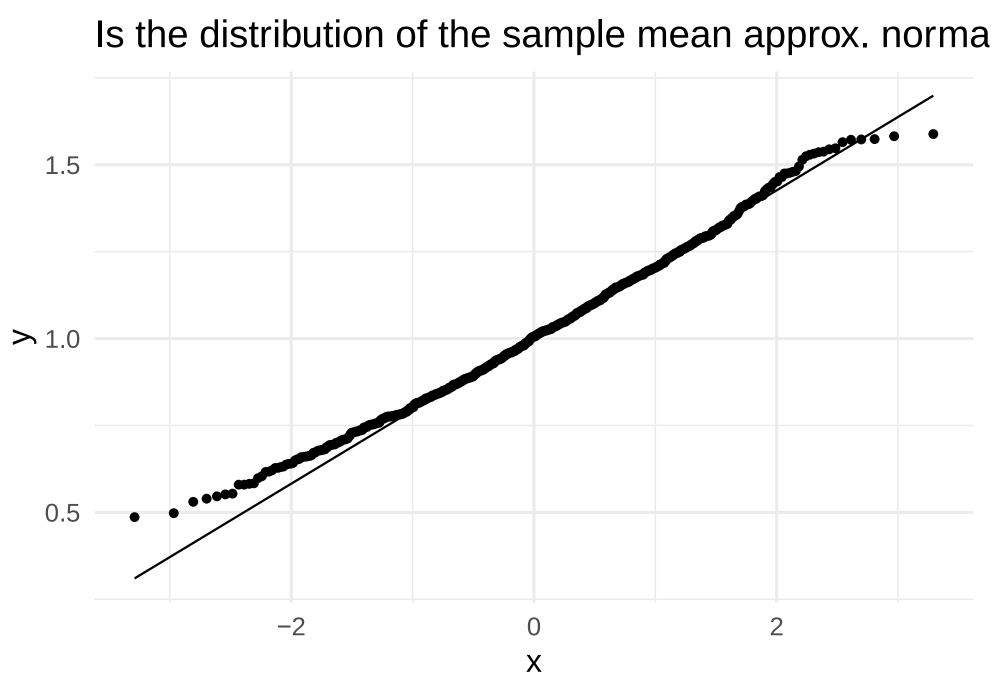
1) Take a sample from the distribution of size 25. This is very small!
2) For your sample calculate the mean.
3) Repeat steps 1 and 2 1,000 times.

4) Plot the sampling distribution of the mean. That is, plot the distribution of the 1000 means using a histogram.

Compare the population mean (black dashed line) with the mean of the sampled means (red dotted line)

CLT example

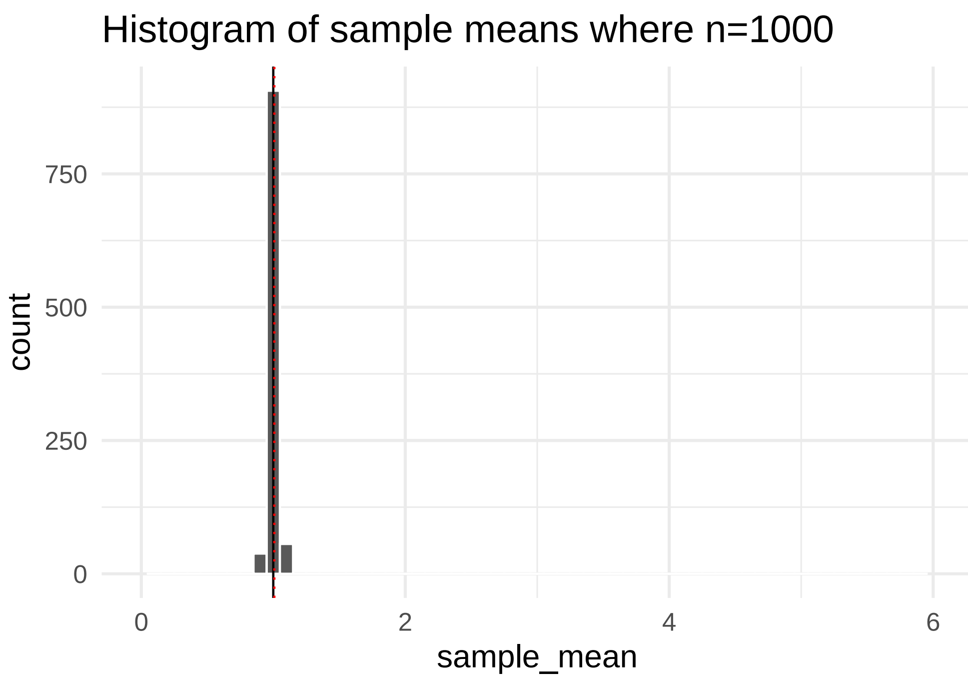


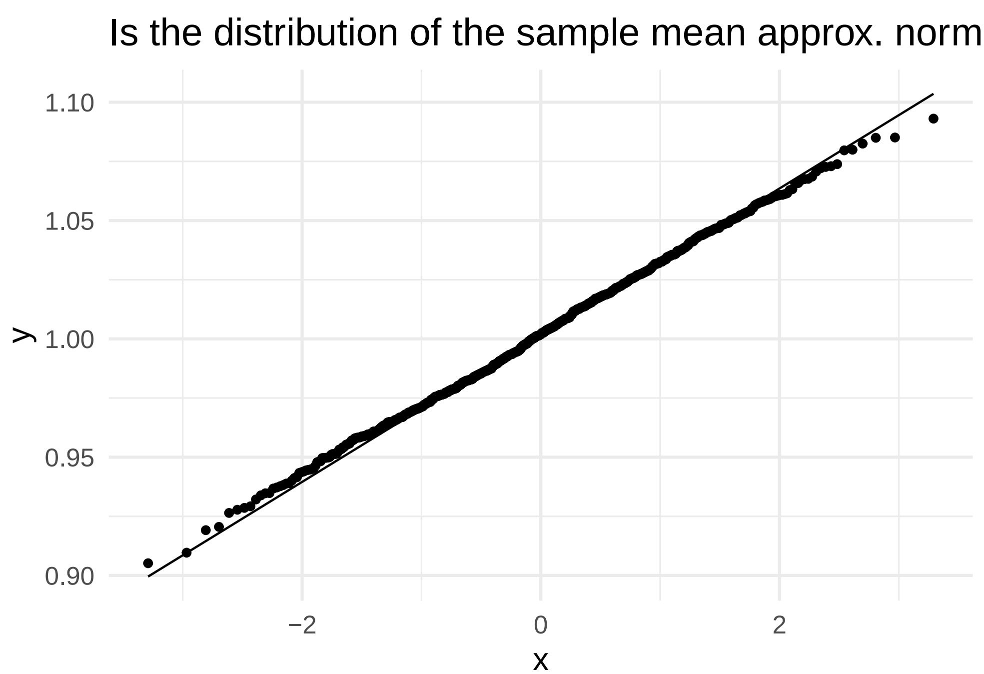


CLT example

1) Take a sample from the distribution of size 1000.
2) For your sample calculate the mean.
3) Repeat steps 1 and 2 1,000 times.
4) Plot the sampling distribution of the mean. That is, plot the distribution of the 1000 means using a histogram.

Compare the population mean (black dashed line) with the mean of the sampled means (red dotted line)





Recap

- In the CLT example, we looked at the sampling distribution for the sample mean, $\bar{x}$, when the underlying data was continuous and skewed right.

- The underlying data did not have to be Normally distributed for the distribution of the sample means to approach a Normal distribution as the sample size n became larger.

Estimating the proportion from a sample

- We can also examine the sampling distribution of the sample proportion $\hat{p}$. For a proportion, recall that the underlying data is categorical and commonly coded using 0/1 coding.
- A proportion is just a special type of mean where the underlying variable is binary (i.e., 0/1, TRUE/FALSE, success/failure are all coding schemes we might use for binary data).
- If our data consists of repeated observations of a random binary variable, $Y_i, i = 1, \ldots, n$ across n independent observations (say $Y_i = 1$ if person i has the event, 0 if not), then the proportion is simply the average of these random variables, or:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Estimating the proportion from a sample

- We have estimated the proportion from a sample a couple times during lecture, lab, and assignment. Let's review.

- Suppose an experiment finds 6 of 20 birds exposed to an avian flu strain develop flu symptoms. Let X represent the number of birds that develop flu symptoms. Based on these data, the estimated proportion, $\hat{p}$ of the number of birds in a larger population that would develop flu symptoms is $\hat{p} = 6/20 = 30\%$.

More generally:

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n},$$

where X represents the total number of "successes" or $Y_i = 1$ from the previous slide: $X = \sum_{i=1}^{n} Y_i$

Recall from Chapter 12...

- The avian flu example might remind you of the Binomial distribution: We have a finite sample size (here n=20) and are counting X, the number of "successes".

- Recall from our lecture on the Binomial distribution that when n is "large enough", X can be approximated by a Normal distribution when $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ That is, $X \dot{\sim} N(np, \sqrt{np(1-p)})$.

- Our rule of thumb for large enough was $np \geq 10$ and $n(1-p) \geq 10$.

- While this tells us the approximate distribution for X, we would like to know the sampling distribution for $\hat{p}$, the sample proportion.

- The sampling distribution will help us evaluate how good our sample estimate $\hat{p}$ is at estimating the population parameter p. Again, we need to ask, what happens when we take many samples?

Sampling distribution of the proportion $\hat{p}$

Choose a simple random sample of size n from a large population that contains population proportion p of successes. Let $\hat{p}$ be the sample proportion of successes. Then:

$$\hat{p} = \frac{\text{count of successes in the sample}}{n} = \frac{X}{n}$$

- The mean of the sampling distribution of $\hat{p}$ is p, the population parameter
- The standard deviation of the sampling distribution of $\hat{p}$ is $\sqrt{\frac{p(1-p)}{n}}$
- As the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately Normal. This is the **Central Limit Theorem** for a proportion!
- For this to apply, we require:
  - the population is at least 20 times as large as the sample
  - both np and n(1-p) are larger than 10.

Sampling distribution of the proportion $\hat{p}$

- Remember the true sampling distribution is the histogram we would make if we looked at **all possible samples of size** n from the population and calculated their sample proportions.
- If we consider repeated sampling (say taking 1000 samples) from a distribution, this gives us an approximation of the sampling distribution for the proportion.
- If we took each estimate of $\hat{p}$ from the 1000 samples and made a histogram it would have a Normal distribution centered at p and its sd would be approximately equal to $\sqrt{\frac{p(1-p)}{n}}$

Related Concept: The Law of Large Numbers (TLLN)

As the sample size increases, $\bar{x}$ is guaranteed to approach μ, and $\hat{p}$ is guaranteed to approach p, the true population parameter.

The most applicable examples of TLLN in the real world are gambling casinos and insurance companies. They rely on TLLN which describes the long-run regularity of gambling (which can be described using probability) and of insurance claims (where most people make only small claims, yet some will make large claims) to make sure their businesses can make a profit.

- Four-minute video on the Law of Large Numbers

The Central Limit theorem summarized

- Applies to both the sample mean $\bar{x}$ and the sample proportion $\hat{p}$.
- When the sample size is large, the sampling distribution is approximately Normal, no matter what the underlying distribution looked like.
- The larger the sample, the better the approximation.
- The mean of the sampling distribution of $\bar{x}$ is μ, and of $\hat{p}$ is p, no matter the sample size. These estimates are called **unbiased estimators**
- The standard deviation of the sampling distribution gets smaller as n increases. In fact, we know the formula for the standard deviation of $\bar{x}$ and $\hat{p}$ so can calculate it for any n.
- The CLT does not "kick in" for small values of n. It kicks in faster for symmetric distributions than for skewed distributions.