Distributions for sample mean and sample sd

STA 032: Gateway to data science Lecture 18

Jingwei Xiong

May 12, 2023

Recall: What is statistical inference?

- **Descriptive statistics** are numbers that are used to summarize and describe data. Descriptive statistics do not necessarily generalize beyond the data, because of sampling variability
- **Inferential statistics** or **statistical inference** gives us an idea about how sample means (for example) from different samples are likely to vary from each other and from the population mean
- **Statistical inference** allows us to draw conclusions about the larger population. It is the process of using sample data to make conclusions about the underlying population the sample came from.
- If the sample is **representative**, then we can use the tools of probability and statistical inference to make generalizable conclusions to the broader population of interest.

Similar to tasting a spoonful of soup while cooking to make an inference about the entire pot.

Recall: Many Topics in Statistical Inference

- Fundamentals: probability, distributions, random variables, ...
- **Sampling**
- Hypothesis testing
- Point estimates and confidence intervals
- Modeling: Linear regression, analysis of variance, nonparametric models, machine learning, ...

Sampling Distribution of the Sample Mean

Recall our shoe size example, where wearers of men's shoe sizes follow a $N(11, 1.5^2)$ distribution.

Say we are interested in the sample mean of shoe sizes. We have a sample of 1000 observations.

```
set.seed(0)
sampled1000_1 <- rnorm(1000, 11, 1.5)
head(sampled1000_1, 20)
```

```
 [1] 12.894431 10.510650 12.994699 12.908644 11.621962  8.690075  9.607149
 [8] 10.557919 10.991349 14.606980 12.145390  9.801486  9.278514 10.565808
[15] 10.551177 10.382734 11.378335  9.662118 11.653525  9.143692
```

```
mean(sampled1000_1)
```

```
[1] 10.97626
```

Sampling Distribution of the Sample Mean

Now we repeat the experiment, i.e., get a different sample of 1000 observations. You will find the sample mean changed a little bit.

```
set.seed(10)
sampled1000_2 <- rnorm(1000, 11, 1.5)
head(sampled1000_2, 20)
```

```
 [1] 11.028119 10.723621  8.943004 10.101248 11.441818 11.584691  9.187886
 [8] 10.454486  8.559991 10.615282 12.652669 12.133672 10.642650 12.481167
[15] 12.112085 11.134021  9.567584 10.707274 12.388282 11.724468
```

```
mean(sampled1000_2)
```

```
[1] 11.01706
```

```
all.equal(mean(sampled1000_1), mean(sampled1000_2))
```

```
[1] "Mean relative difference: 0.003717704"
```

Sampling Distribution of the Sample Mean

If we repeat the experiment an infinite number of times, what distribution of sample means would we get? This is known as the **sampling distribution**.

1. Take a sample of size n and calculate its mean $\overline{x}_1$
2. Take a second sample of the same size and calculate its mean $\overline{x}_2$
3. Repeat this many times to get a dataset of sample means $\overline{x}_1, \overline{x}_2, \ldots$

What is the distribution of the *statistics* $\overline{x}_1, \overline{x}_2, \overline{x}_3, \ldots$?

The sample mean $\overline{X}$, is defined as $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. Each realization above is a draw from $\overline{X}$, denoted with a small letter $\overline{x}$.

Here we consider $X_1, \ldots, X_n$ that are **independent and identically distributed**. (E.g., $X_1, \ldots X_n \sim N(11, 1.5^2)$ for the shoe size distribution.)

Sampling Distribution of the Sample Mean

1. Take a sample of size n and calculate its mean $\overline{x}_1$
2. Take a second sample of the same size and calculate its mean $\overline{x}_2$
3. Repeat this many times to get a dataset of sample means $\overline{x}_1, \overline{x}_2, \ldots$

We cannot repeat an infinite number of times, but we do this 10,000 times in R.

```
set.seed(0)
# Here replicate generate a matrix of 10000 rows, each row we have a 1000 sam
repeat10000 <- t(replicate(n = 10000, rnorm(1000, 11, 1.5)))
str(repeat10000)
```

```
 num [1:10000, 1:1000] 12.9 10.6 11.7 10.2 12.2 ...
```

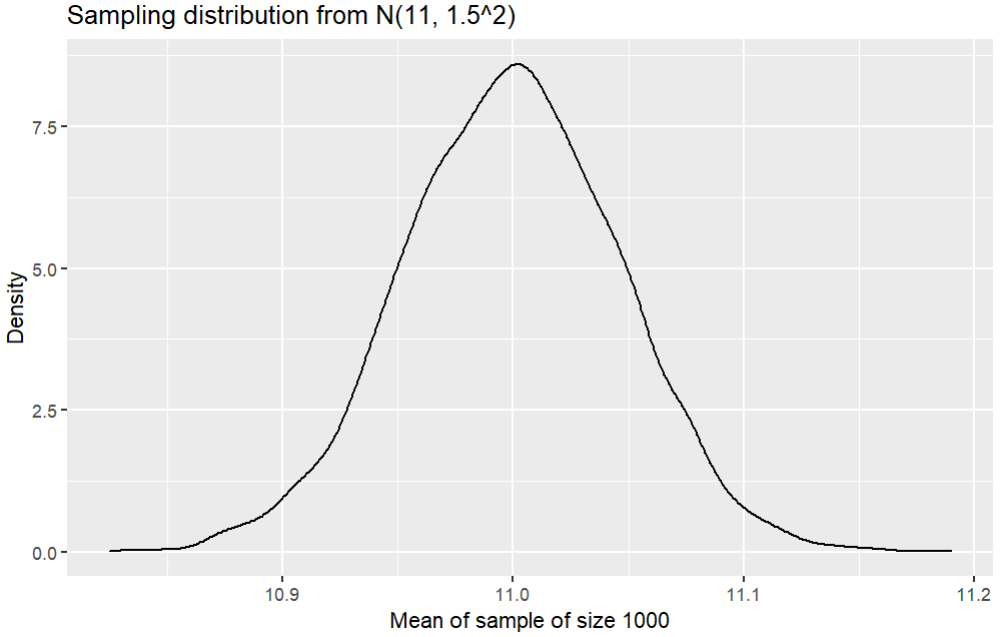
```
# Use rowMeans to obtain the sample mean for each row
head(rowMeans(repeat10000), 20)
```

```
 [1] 10.97626 10.96282 11.10221 11.00373 11.00616 11.02959 10.99695 11.06309
 [9] 10.92670 11.08920 10.97525 10.95720 10.95437 11.07828 11.02202 11.05240
[17] 11.00455 10.95824 10.96431 10.96413
```

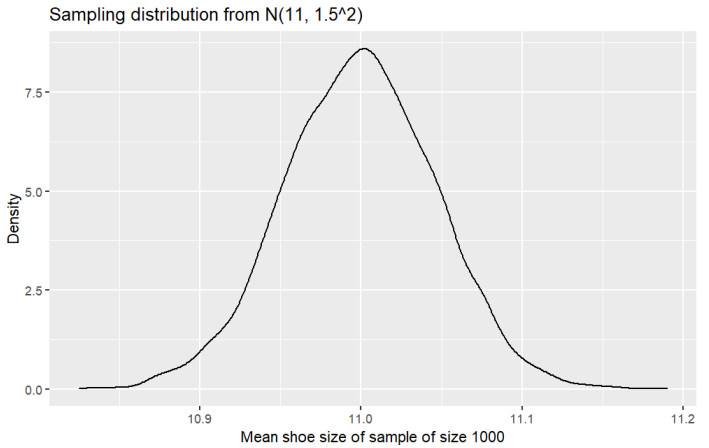
```
means10000 <- rowMeans(repeat10000)
```

Sampling Distribution of the Sample Mean

```
# ggplot has to be applied on a dataframe.
data.frame(shoesMean = means10000) %>%
  ggplot(aes(x = shoesMean)) +
  geom_density() +
  labs(x = "Mean of sample of size 1000",
       y = "Density",
       title = "Sampling distribution from N(11, 1.5^2)")
```



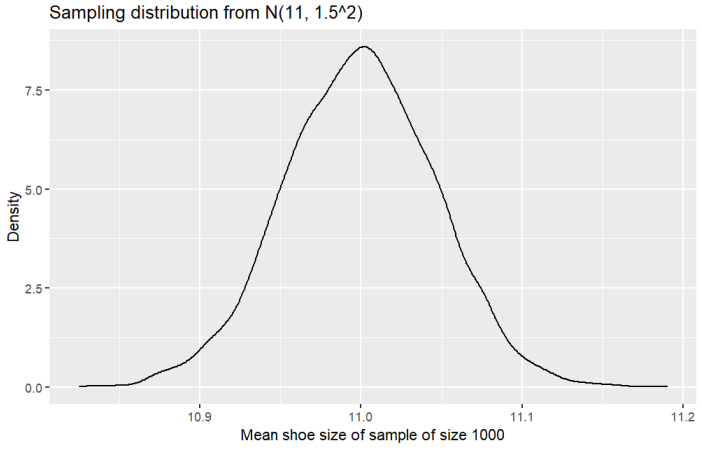
Sampling Distribution of the Sample Mean



How would we describe this distribution?

- Center:
- Spread:
- Shape:

Sampling Distribution of the Sample Mean



How would we describe this distribution?

- Center
  - The distribution is centered at 11, which is the same as the population parameter
- Spread
  - The variability looks to be much smaller than the original distribution (the original distribution has standard deviation 1.5)
- Shape
  - The distribution is symmetric and bell-shaped, and it resembles a normal distribution.

Effect of changing sample size

- Earlier we used a sample size of 1000. What if we used a sample size of 50?

```
set.seed(0)
repeat10000_n50 <- t(replicate(n = 10000, rnorm(50, 11, 1.5)))
str(repeat10000_n50)
```

```
 num [1:10000, 1:50] 12.89 11.4 12.17 13.21 9.43 ...
```

```
head(rowMeans(repeat10000_n50), 20)
```

```
 [1] 11.03590 11.03211 10.68366 11.17968 11.06924 11.13242 11.03710 10.97258
 [9] 10.96971 10.88460 10.93739 10.80427 11.03709 10.90858 10.84133 11.32991
[17] 11.04654 10.74770 10.98847 10.88685
```

```
means10000_n50 <- rowMeans(repeat10000_n50)
```

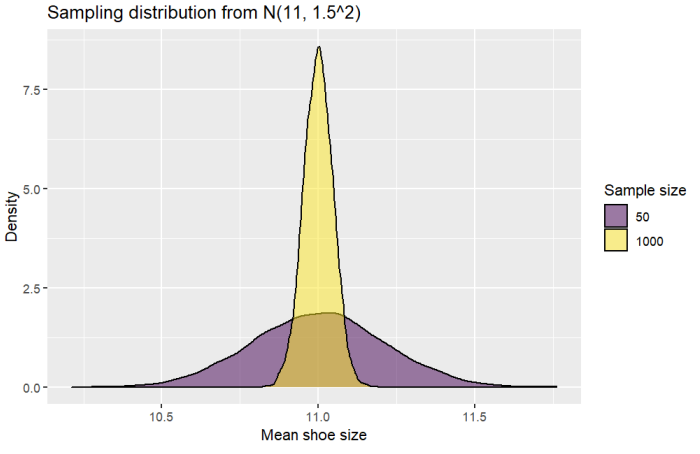
To study this, we first construct the data frame for ggplot2 using bind_rows:

```
data.frame(shoesMean = means10000, sampleSize = 1000) %>%
  bind_rows(
    data.frame(means10000_n50, sampleSize = 50) %>%
      rename(shoesMean = means10000_n50)
  )  %>%head(3)
```

```
  shoesMean sampleSize
1  10.97626       1000
2  10.96282       1000
3  11.10221       1000
```

Effect of changing sample size

```
data.frame(shoesMean = means10000, sampleSize =
  bind_rows(
    data.frame(means10000_n50, sampleSize = 50)
      rename(shoesMean = means10000_n50)
  )  %>%
  ggplot(aes(x = shoesMean,
             fill = as.factor(sampleSize),
             alpha = 0.7)) +
  geom_density() +
  labs(x = "Mean shoe size",
       y = "Density",
       title = "Sampling distribution from N(11,
       fill = "Sample size")  +
  scale_fill_viridis_d()
 # guides(fill = "none")
```



- What do you notice about the spread?
  - A larger sample size produces more precise estimates
  - We will formalize this intuition using the Central Limit Theorem

Recall: Sum of independent normal random variables

- Important property: Any linear combination of normal random variables is a normal random variable with expectation and variance given by the formulas for expected value and variance of linear combinations (Lecture 15)
- Recall: A linear combination of two random variables, X and Y, is of the form $aX + bY$, where a and b are constants
- Recall:
  - $E(aX + bY) = aE(X) + bE(Y)$
  - For a linear combination of **independent** random variables $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$
- Hence if $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ are independent, $W = X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$
- Extends to more than two random variables in the linear combination. Note also that b can be negative, e.g., $E(X - Y) = E(X) - E(Y)$ and $Var(X - Y) = Var(X) + Var(Y)$.

Distribution of sample mean when the population is a normal distribution:

Because we know the linear combination of normal random variables are still normal RV, so we can determine the distribution of sample mean by calculating its expectation and variation.

> This is because the expectation μ and variation σ^2 are the only 2 parameters of normal distribution.

The sample mean $\overline{X}$, is defined as $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. $X_i \sim N(\mu, \sigma)$

$$E(\overline{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{E\left(\sum_{i=1}^{n} X_i\right)}{n} = \frac{\sum_{i=1}^{n} EX_i}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \mu$$

$$Var(\overline{X}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{Var\left(\sum_{i=1}^{n} X_i\right)}{n^2} = \frac{\sum_{i=1}^{n} VarX_i}{n^2} = \frac{\sum_{i=1}^{n} \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Where $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} VarX_i$ because they are independent.

> Note this is only true when the population is normally distributed!

Note on sampling distributions

- From Open Intro Statistics: "Sampling distributions are never observed, but we keep them in mind."
- In real-world applications, we never actually observe the sampling distribution
- Even in simulations, we cannot run experiments an infinite number of times to generate the sampling distribution
- Yet it is useful to think of a sample statistic as coming from such a hypothetical distribution
- Understanding the sampling distribution will help us characterize and make sense of the sample statistics that we do observe.

Sampling distributions, confidence intervals and hypothesis testing

Knowing the sampling distribution can help us

- Estimate a population parameter as point estimate $\pm$ margin of error, where the margin of error is comprised of a measure of how confident we want to be and the sample statistic's variability. (Coming soon: confidence intervals)
- Test whether a population parameter is equal to some value, by evaluating how likely it is that we have obtained the observed sample statistic, if the population parameter is indeed that value. (Coming soon: hypothesis testing)

  > You may ask: why? The previous conclusion only work when we know the population is normally distrbuted. But what if they don't?

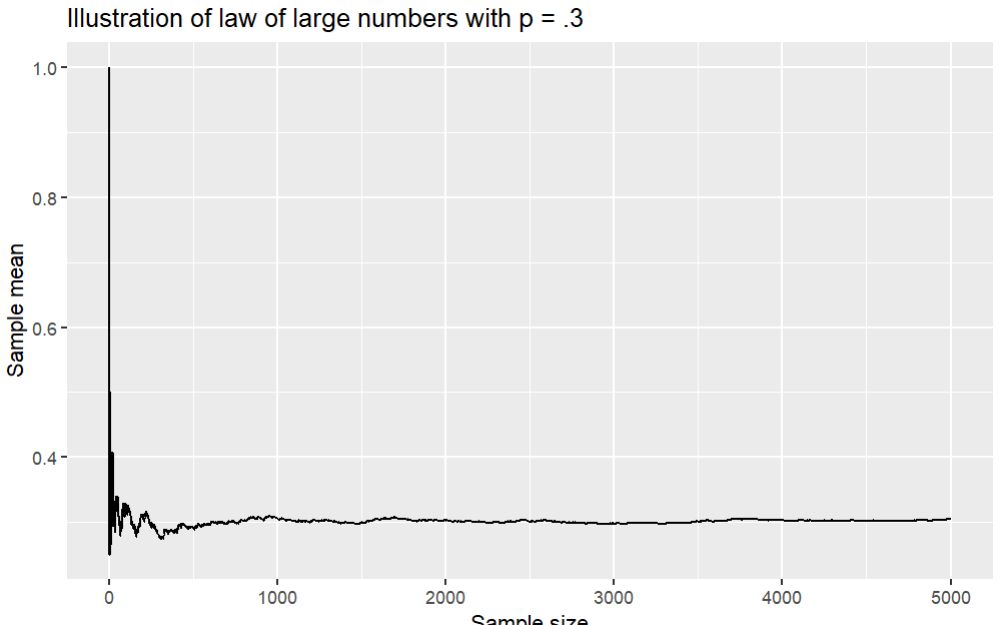
- Solution: Central Limit Theorem (CLT)

Central Limit Theorem

- The **Central Limit Theorem** says that for **any** distribution with a well-defined mean and variance, the distribution of **sample means** for a sample of size n is approximately normal.
- Formally: for a population with mean μ and standard deviation σ, taking independent samples $X_1, \ldots, X_n$, the following three important properties of the distribution of the sample mean $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ hold:
  - The mean of the sampling distribution is identical to the population mean μ, i.e., $E(\overline{X}) = \mu$
  - The standard deviation of the sampling distribution is $\frac{\sigma}{\sqrt{n}}$, i.e., $Var(\overline{X}) = \frac{\sigma^2}{n}$
  - For n large enough (in the limit as $n \to \infty$), the shape of the sampling distribution is approximately normal (Gaussian), i.e., $\overline{X} \approx N(\mu, \frac{\sigma^2}{n})$
- This result has strong implications in many areas of statistics, including in construction of confidence intervals and in hypothesis testing.

Intuition

- The average of many measurements of the same unknown quantity tends to give a better estimate than a single measurement
    - If we want to know the population mean test score of the class, getting information from a sample of 10 students is better than asking a single student
- Recall the law of large numbers: as the sample size grows, the sample mean gets closer to the expected value, or population mean



```r
# code to generate the sample:
bernoulliDraws <- rbinom(n = 5000, size = 1, prob = inputP)
myMeans <- data.frame(sampleSize = 1:5000, myMean = NA)
meanFun <- function(inputSampSize, outcomes) {
  return(mean(outcomes[1:inputSampSize]))
}
myMeans$myMean <- sapply(myMeans$sampleSize, meanFun, bernoulliDraws)
#code to generate the plot
myMeans %>%
  ggplot(aes(x = sampleSize, y = myMean)) +
  geom_line() +
  labs(x = "Sample size",
       y = "Sample mean",
       title = "Illustration of law of large numbers with p = .3")
```

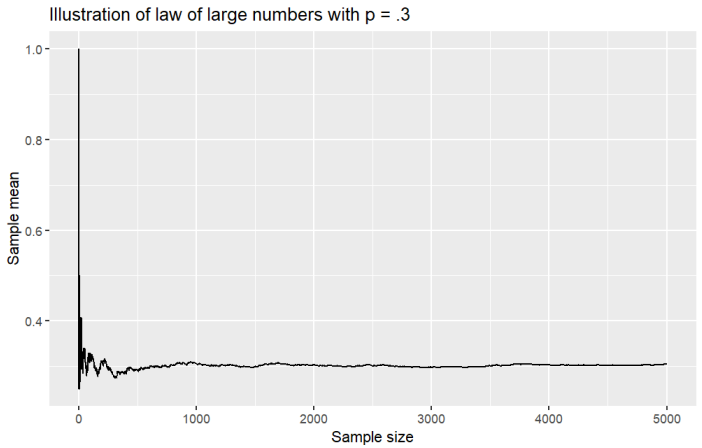
Here this sapply serves the same as the following code:

```r
for(i in myMeans$sampleSize){
  myMeans$myMean[i] = meanFun(i, bernoulliDraws)
}
```

Which means it will use each element of the myMeans$sampleSize in turn as the first argument of meanFun

And bernoulliDraws as the second argument of meanFun. Then construct all of these result as a vector, and then assign to myMeans$myMean

Intuition

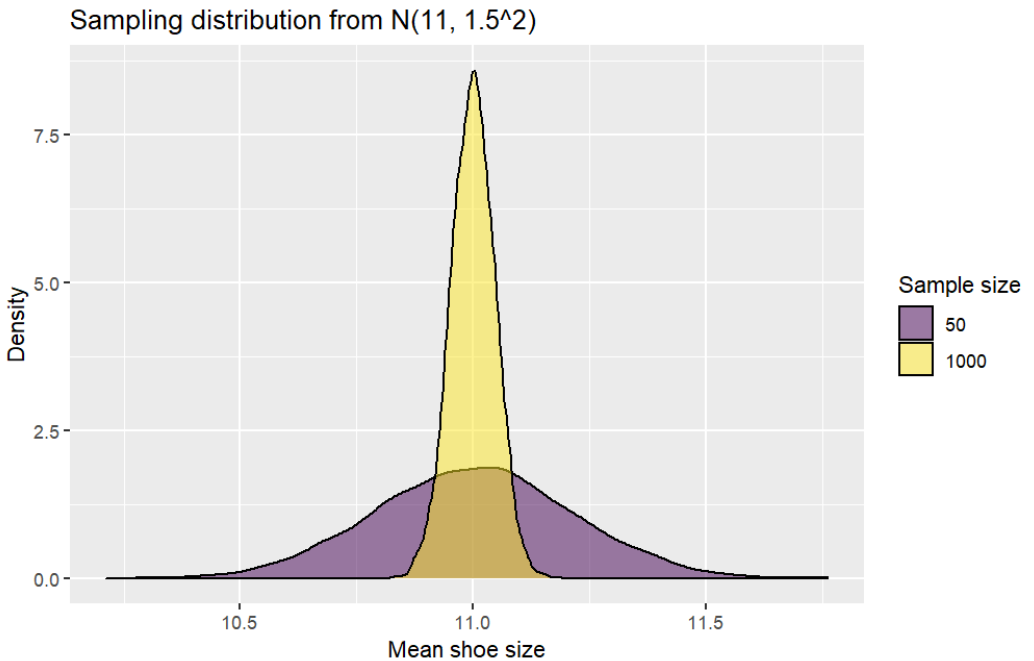


- Note that here we are using Bernoulli(.3), so $\mu = p = .3$ and $\sigma^2 = p(1-p) = .3(.7)$

- In this illustration, you can think of each point plotted as a single draw from $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$, as we vary n, the sample size, plotted on the horizontal axis.
- The Central Limit Theorem tells us the distribution of $\overline{X}$, at each value of n.
- For large values of n, we see that $Var(\overline{X}) = \frac{\sigma^2}{n}$ gets very small, so any draw from this distribution will be very close to $E(\overline{X}) = \mu$.

Intuition

- We can see the same narrowing distribution (smaller variance) with our shoes example:



- Nice applet where you can adjust the sample size and other parameters and see the impact on the sampling distribution: http://demonstrations.wolfram.com/SamplingDistributionOfTheSampleMean/

Central Limit Theorem

For a population with mean μ and standard deviation σ, taking independent samples $X_1, \ldots, X_n$, the following three important properties of the distribution of the sample mean $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ hold:

- The mean of the sampling distribution is identical to the population mean μ, i.e., $E(\overline{X}) = \mu$
- The standard deviation of the sampling distribution is $\frac{\sigma}{\sqrt{n}}$, i.e., $Var(\overline{X}) = \frac{\sigma^2}{n}$
- For n large enough (in the limit as $n \to \infty$), the shape of the sampling distribution is **approximately** normal (Gaussian), i.e., $\overline{X} \approx N(\mu, \frac{\sigma^2}{n})$

Notice that this does not **restrict** the distribution of the underlying $X_1, \ldots, X_n$ in any way. These can be normal, binomial, Poisson, ...

- (When the population is normal, the sample mean distribution will **alaways be normal**, regardless of the sample size.)

Central Limit Theorem with different underlying distributions

- All we need to know is $E(X_i)$ or μ, and $Var(X_i)$ or σ^2
- For normally distributed random variables with mean μ and variance σ^2, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ (actually, we don't need CLT for this. Why?)
- For Bernoulli distributed random variables with probability of success p, $Var(X_i) = p(1-p)$, $\overline{X} = \hat{P} \approx N(p, \frac{p(1-p)}{n})$
- For Poisson(λ) distributed random variables with $E(X_i) = \lambda$ and $Var(X_i) = \lambda$, $\overline{X} \approx N(\lambda, \frac{\lambda}{n})$

How Large is Large Enough for n?

- A commonly used rule of thumb is $n > 50$
- For Bernoulli data, one rule of thumb is that n should be large enough that $np > 5$ and $n(1-p) > 5$. Sometimes you also see $np > 10$ and $n(1-p) > 10$. These are just rules of thumb and approximations!
    - So if p is around a half, you need a smaller sample size than if p is close to the boundaries of 0 or 1.
    - We will see this again when we talk about the normal approximation for the binomial

Example: Normal data

In the shoe size example, we have $X_1, \ldots, X_n \sim N(11, 1.5^2)$. Say we collected 1000 samples, so the sample size $n = 1000$. What distribution does the sampling distribution of the sample mean follow? What is $P(\overline{X} < 10.9)$? Calculate this in two ways: using the original distribution, and using the standard normal distribution.

By the Central Limit Theorem, $\overline{X} \approx N(\mu, \frac{\sigma^2}{n})$, in this case $N(11, \frac{1.5^2}{1000})$.

Standardizing, we have $P(\overline{X} < 10.9) = P(\frac{\overline{X} - 11}{1.5/\sqrt{1000}} < \frac{10.9 - 11}{1.5/\sqrt{1000}}) = P(Z < \frac{-.1}{1.5/\sqrt{1000}})$

```
pnorm(10.9, mean = 11, sd = sqrt(1.5^2/1000))
```

```
[1] 0.01750749
```

```
pnorm((10.9 - 11)/(1.5/sqrt(1000)) ) # standardized version
```

```
[1] 0.01750749
```

Example: Bernoulli data

Assume that 67% of the population in India has had a prior COVID infection. Define the variable X_i to take value 1 if the ith randomly sampled Indian resident has been infected, and let it be 0 otherwise. Assume the samples are independent.

What distribution does X_i follow? What are the parameters? What is the mean? What is the variance?

- The random variable X_i follows a Bernoulli distribution. A Bernoulli random variable has mean p and variance $p(1-p)$. Here $p = 0.67$.
- If we take a random sample, the sample mean $\overline{x} = \hat{p}$ is an *estimate* of this probability (it's just the fraction of 1's)
- If we take repeated samples of Indian residents and compute the proportion with prior infection in each, what values will we see? Will we get 0.67 each time?

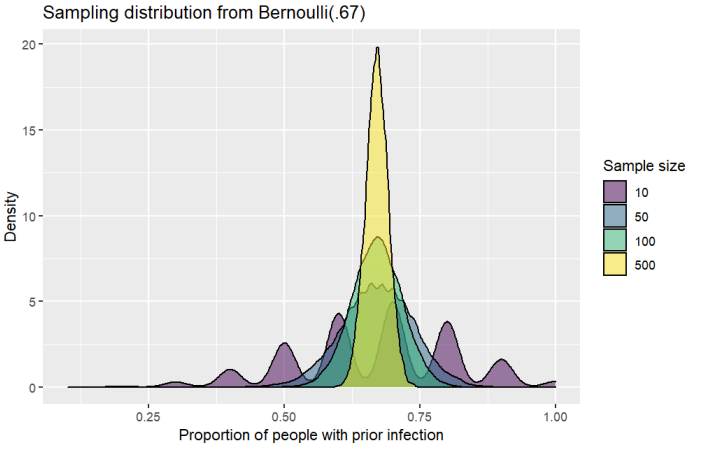
Simulation: Prior Infections

Again, we perform 10000 experiments where each time we sample from a Bernoulli distribution with `prob = 0.67` and different values of `n`.

```
set.seed(0)
n10 <- t(replicate(n = 10000, rbinom(10, size = 1, prob = .67)))
n10 <- rowMeans(n10)

set.seed(0)
n50 <- t(replicate(n = 10000, rbinom(50, size = 1, prob = .67)))
n50 <- rowMeans(n50)

set.seed(0)
n100 <- t(replicate(n = 10000, rbinom(100, size = 1, prob = .67)))
n100 <- rowMeans(n100)

set.seed(0)
n500 <- t(replicate(n = 10000, rbinom(500, size = 1, prob = .67)))
n500 <- rowMeans(n500)
```

Simulation: Prior Infections

```
data.frame(propCovid = n10, sampleSi
  bind_rows(
    data.frame(n50, sampleSize = 50)
      rename(propCovid = n50)
  ) %>%
  bind_rows(
    data.frame(n100, sampleSize = 10
      rename(propCovid = n100)
  )%>%
  bind_rows(
    data.frame(n500, sampleSize = 50
      rename(propCovid = n500)
  ) %>%
  ggplot(aes(x = propCovid,
             fill = as.factor(sample
  geom_density(alpha = .5) +
  labs(x = "Proportion of people wit
       y = "Density",
       title = "Sampling distributio
       fill = "Sample size")  +
  scale_fill_viridis_d()
```



Sampling distribution of the sample proportion

- The sample proportion is the same as the sample mean, when the distribution is Bernoulli
- The sample proportion is $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim$ Bernoulli(p).
- Earlier example:

Assume that 67% of the population in India has had a prior COVID infection. Define the variable X_i to take value 1 if the ith randomly sampled Indian resident has been infected, and let it be 0 otherwise. Assume the samples are independent.

- The random variable X is Bernoulli. A Bernoulli random variable has mean p and variance $p(1-p)$. Here $p = 0.67$.
- The sample proportion that had a prior infection is just the fraction of 1's
- The sample mean is defined as $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. When X_i is Bernoulli, it can only take the values 0 and 1, and the sample mean is the same as the sample proportion. $\overline{X} = \hat{P}$.

Sampling distribution of the sample proportion

- Recall: the Central Limit Theorem does not restrict the distribution of the underlying $X_1, \ldots, X_n$ in any way. They can be normal, binomial, Poisson, ...
- $\overline{X} \approx N(\mu, \frac{\sigma^2}{n})$
- We get the sampling distribution of the sample proportion for free!
- When $X_i \sim$ Bernoulli(p), $\mu = p$ and $\sigma^2 = p(1-p)$.
- By the Central Limit Theorem, $\overline{X} = \hat{P} \approx N(p, \frac{p(1-p)}{n})$

Example: cats

We ask a random sample of 435 people if they like cats. Assume that the probability that a randomly selected person likes cats is 75%. Let X_i be the Bernoulli variable representing whether or not the ith person likes cats. What is the sampling distribution of the sample proportion? What is the probability that the sample proportion is smaller than .7?

By the Central Limit Theorem, $\overline{X} = \hat{P} = \frac{\sum_{i=1}^{n} X_i}{n} \approx N(p, \frac{p(1-p)}{n})$.

In this case, $\overline{X} = \hat{P} = \frac{\sum_{i=1}^{435} X_i}{435} \approx N(.75, \frac{.75(1-.75)}{435})$.

```
pnorm(.7, mean = .75, sd = sqrt(.75*(1-.75)/435))
```

```
[1] 0.008013087
```

Summary

- Central Limit Theorem: $\overline{X} \approx N(\mu, \frac{\sigma^2}{n})$
- Sampling distribution of the sample mean
- Sampling distribution of the sample proportion: $\overline{X} = \hat{P} \approx N(p, \frac{p(1-p)}{n})$