Lecture 5: Sampling Distributions

Jacob M. Montgomery

Quantitative Political Methodology

Lecture 5

Class business

- Problem set 2 will be distributed today via the syllabus
- Groups will be assigned this week.

Roadmap

Last time:

- Understand core concepts of probability
- Understand concept of a "parameter"
- Introduce some probability distributions

This time:

- Understanding the concept of a sampling distribution
- Understand the concept of a standard error
- See how the CLT allows us to know the distribution of certain sample statistics.

Sampling Distributions

A ***sampling distribution*** *is the distribution of a* ***statistic*** *given repeated sampling.*

Sampling Distributions

*A **sampling distribution** is the distribution of a **statistic** given repeated sampling.*

We use probability theory to derive a **distribution for a statistic**

Sampling Distributions

> *A **sampling distribution** is the distribution of a **statistic** given repeated sampling.*

We use probability theory to derive a **distribution for a statistic**, which allows us (eventually) to make inferences about **population parameters**.

Central limit theorem

For random sampling with a **large** sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, where $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Central limit theorem

For random sampling with a **large** sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, where $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

- $\sigma/\sqrt{n}$ is called the standard error.

Central limit theorem

For random sampling with a **large** sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, where $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

- $\sigma/\sqrt{n}$ is called the standard error. - It is the standard deviation of the sampling distribution.
  - Note that the formula includes the population standard deviation σ.
  - Pay attention or you will get them mixed up.

Central limit theorem

For random sampling with a **large** sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, where $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

- $\sigma/\sqrt{n}$ is called the standard error. - It is the standard deviation of the sampling distribution.
  - Note that the formula includes the population standard deviation σ.
  - Pay attention or you will get them mixed up.
- As $n \to \infty$, the standard error is going to get smaller and smaller.

Central limit theorem

For random sampling with a **large** sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, where $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

- $\sigma/\sqrt{n}$ is called the standard error. - It is the standard deviation of the sampling distribution.
  - Note that the formula includes the population standard deviation σ.
  - Pay attention or you will get them mixed up.
- As $n \to \infty$, the standard error is going to get smaller and smaller.
- **This** is why the normal distribution is so very important.

Central limit theorem

For random sampling with a **large** sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, where $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

- $\sigma/\sqrt{n}$ is called the standard error. - It is the standard deviation of the sampling distribution.
    - Note that the formula includes the population standard deviation σ.
    - Pay attention or you will get them mixed up.
- As $n \to \infty$, the standard error is going to get smaller and smaller.
- **This** is why the normal distribution is so very important.
- Usually n=30 is "good enough", but it will depend on the distribution.

Central limit theorem

For random sampling with a **large** sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, where $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

- $\sigma/\sqrt{n}$ is called the standard error. - It is the standard deviation of the sampling distribution.
  - Note that the formula includes the population standard deviation σ.
  - Pay attention or you will get them mixed up.
- As $n \to \infty$, the standard error is going to get smaller and smaller.
- **This** is why the normal distribution is so very important.
- Usually n=30 is "good enough", but it will depend on the distribution.

It works for EVERYTHING

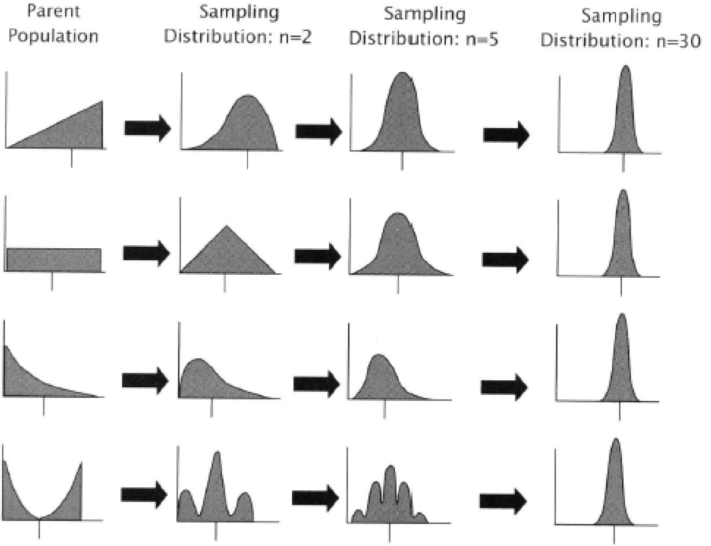
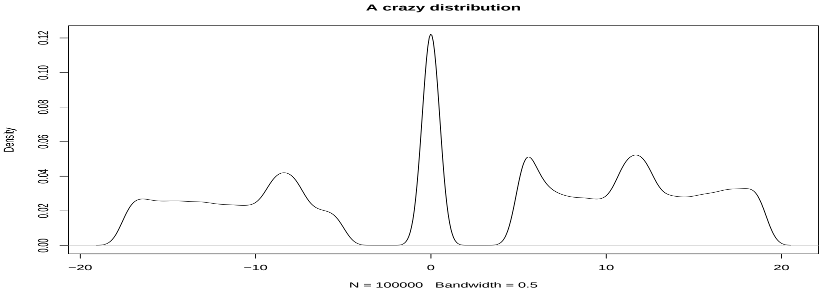


FIGURE 4.33 Given sufficient sample size the sampling distribution of the mean approaches normal shape irrespective of the variable's distributional shape.

Let's do our own experiment in R

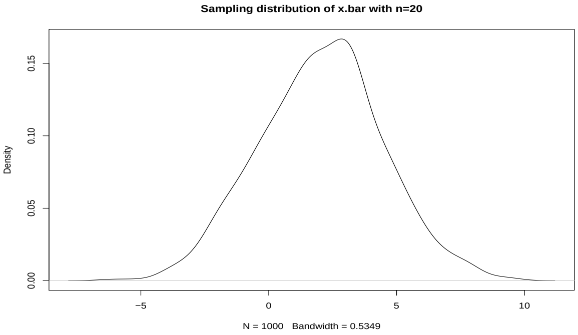
```
x<-runif(100000,min = -1, max=1)
## Crazy transformation of the data
x<-sqrt(1+x)+2*x^3-23*x+abs(log(abs(x)))+2*(x>.5)+-2*(x< -.5)
## Adding a big whole in the middle with a point mass at zero
x[x>-5&x<5]<-0
plot(density(x, bw = .5), main="A crazy distribution")
```



```r
# We are going to take 500 random samples
n.samples<-1000
# The sample size is 20
sample.size<-20
# We are going to use something called a "for loop"
# First we make an empty vector to store all of our
# sample statistics

# Create a vector filled with NA (missing data)
# vector is of length 200
x.bars<-rep(NA, n.samples)
# Now we are going to "loop" over the vector 1, 2, ..., 200
# in each iteration the variable "i" will increment up on value
for(i in 1:n.samples){
  # Draw a random sample
  this.sample<-sample(x, size = sample.size, replace=F)
  # Calculate the mean and add it to the vector
  x.bars[i]<-mean(this.sample)
}
```

```r
plot(density(x.bars),
     main="Sampling distribution of x.bar with n=20")
```

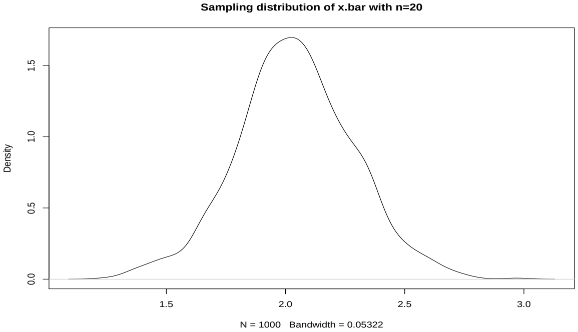


Now with a larger sample size

```
# We are going to take 500 random samples
n.samples<-1000
# The sample size is 2000
sample.size<-2000
# We are going to use something called a "for loop"
# First we make an empty vector to store all of our
# sample statistics

# Create a vector filled with NA (missing data)
# vector is of length 200
x.bars<-rep(NA, n.samples)
# Now we are going to "loop" over the vector 1, 2, ..., 200
# in each iteration the variable "i" will increment up on value
for(i in 1:n.samples){
  # Draw a random sample
  this.sample<-sample(x, size = sample.size, replace=F)
  # Calculate the mean and add it to the vector
  x.bars[i]<-mean(this.sample)
}
```

```r
plot(density(x.bars),
     main="Sampling distribution of x.bar with n=20")
```



Sampling Distribution of $\bar{y}$

A key common statistic is $\bar{y} = \frac{1}{n} \sum y_i$ for a single sample, where n is the sample size. How is this statistic distributed?

Sampling Distribution of $\bar{y}$

A key common statistic is $\bar{y} = \dfrac{1}{n}\sum y_i$ for a single sample, where n is the sample size. How is this statistic distributed? The mean of the distribution is known to be μ (the population mean).

Sampling Distribution of $\bar{y}$

A key common statistic is $\bar{y} = \dfrac{1}{n}\sum y_i$ for a single sample, where n is the sample size. How is this statistic distributed? The mean of the distribution is known to be μ (the population mean). What about the spread?

Standard error

The standard deviation of the sampling distribution of $\bar{y}$, denoted $\sigma_{\bar{y}}$, is called the standard error of $\bar{y}$, and is equal to $\frac{\sigma}{\sqrt{n}}$.

Standard error

The standard deviation of the sampling distribution of $\bar{y}$, denoted $\sigma_{\bar{y}}$, is called the standard error of $\bar{y}$, and is equal to $\dfrac{\sigma}{\sqrt{n}}$.

Under certain circumstances we can safely assume that $\bar{y} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

Class business

- Next class we are going to learn how to calculate probabilities for a known distribution
- Then we pull it all together to make our first true inference