GOV 2000 Section 4: Probability Distributions & Sampling Distributions

Konstantin Kashin[1]
Harvard University

September 26, 2012

[1] These notes and accompanying code draw on the notes from Molly Roberts, Maya Sen, Iain Osgood, Brandon Stewart, and TF's from previous years

Outline

Problem Set Expectations

- Third problem set distributed yesterday.
- Corrections for first problem set due next Tuesday.
- Must be typeset using LaTeX or Word and submitted as one document containing graphics and explanation electronically in **pdf** form
- Must be accompanied by source-able, commented code

Midterm Exam

- Window of exam: Tuesday, October 9th (after class) - Sunday, October 14th at 11.59pm
- Needs to be completed in 5 hours
- Open note / book, but **no collaboration** allowed

Outline

Sampling from Common Probability Distributions

How do we sample from the normal distribution with $\mu = 0$ and $\sigma = 2$ in R?

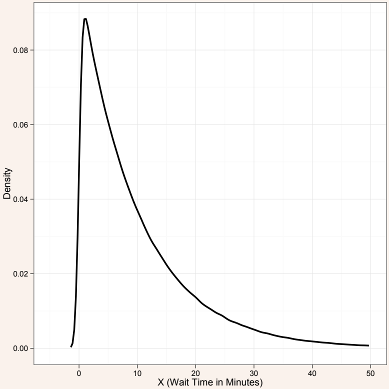
```
set.seed(12345)
rnorm(10, mean=0, sd=2)
```

How do we sample from the uniform distribution on the interval $[0, 10]$ in R?

```
set.seed(12345)
runif(10, min=0, max=10)
```

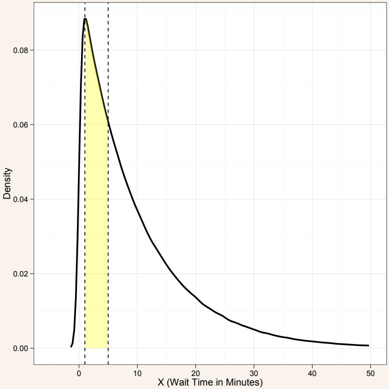
What is the probability that a random variable lies in a particular subdomain?

X is the wait time (in minutes) for the red line in the morning. Let $X \sim \text{Expo}(0.1)$. What is the probability that $X \in [1, 5]$, that one has to wait less than 5 minutes but more than a minute?



What is the probability that a random variable lies in a particular subdomain?

X is the wait time (in minutes) for the red line in the morning. Let $X \sim \text{Expo}(0.1)$. What is the probability that $X \in [1, 5]$, that one has to wait less than 5 minutes but more than a minute?



Analytic Approach: CDFs

$$P(1 \leq X \leq 5) = P(X \leq 5) - P(X \leq 1) = F_X(5) - F_X(1)$$

In R:

```
pexp(q=5, rate=0.1)-pexp(q=1, rate=0.1)
```

$\therefore P(1 \leq X \leq 5) \approx 0.298$

Simulation Approach

1. Sample from distribution of interest to approximate it
2. Calculate proportion of observations in sample that fall in subdomain of interest

In R:

```
set.seed(12345)
exp.vec <- rexp(n=10000,rate=0.1)
mean(1 <= exp.vec & exp.vec <= 5)
```

$\therefore P(1 \leq X \leq 5) \approx 0.3016$

Outline

Our Data

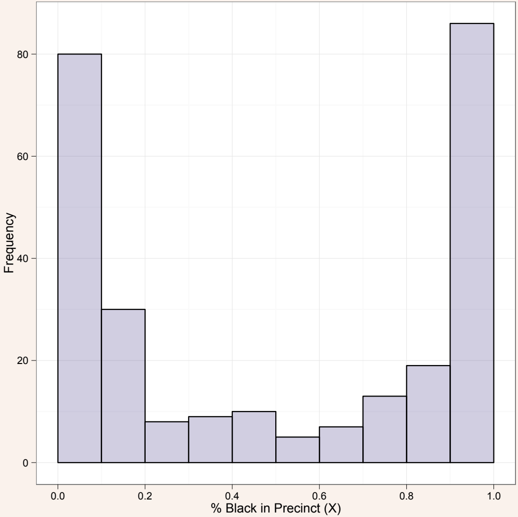
We are going to work with `precincts` dataset from `fulton.RData`.

- Election data at precinct level for Fulton County, Georgia.
- Population: 268 precincts
- Variables: turnout rate, % black, % female, mean age, turnout in Dem. primary, turnout in Rep. primary, dummy variable for whether precinct is in Atlanta, and location dummies for polling stations

Let's define X = % black

Visualizing the Population

What does the population of X = % black look like?



Calculate True (Population) Mean

```
mean(precincts$black)
```

$\mu = 0.506$

Sampling of Precincts

Suppose we can only sample (SRS without replacement) $n = 40$ precincts. How do we do this in R?

First, note that we can subset the dataset as:

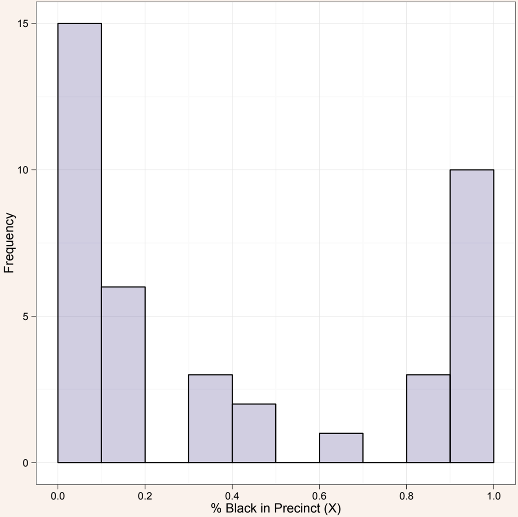
```
precincts[c(10,3,200),]
```

To sample $n = 40$ rows / precincts then:

```
set.seed(12345)
n.sample <- 40
N <- nrow(precincts)
rand.rows <- sample(N, size=n.sample, replace=FALSE)
mypoll <- precincts[rand.rows,]
```

Visualizing the Sample

What does the sample of X = % black look like?



Calculate Sample Mean ($\bar{X}$)

```
mean(mypoll$black)
```

$\bar{X} = 0.412$

Connecting this to Resampling

We can think of our sample as one of many possible samples we can draw from our population:

Samples from Pop.

	1	**2**	**3**	$\cdots$	**10000**
X_1	0.496	$x_{1,2}$	$x_{1,3}$	$\cdots$	$x_{1,10000}$
X_2	0.320	$x_{2,2}$	$x_{2,3}$	$\cdots$	$x_{2,10000}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_{40}	0.172	$x_{40,2}$	$x_{40,3}$	$\cdots$	$x_{40,10000}$
$\bar{X}_{40}$	0.412	$\bar{x}_2$	$\bar{x}_3$	$\cdots$	$\bar{x}_{10000}$
$\bar{S}^2$	0.160	s_2^2	s_3^2	$\cdots$	s_{10000}^2

Connecting this to Resampling

We can think of our sample as one of many possible samples we can draw from our population:

Samples from Pop.

		1	2	3	$\cdots$	10000
	X_1	0.496	$x_{1,2}$	$x_{1,3}$	$\cdots$	$x_{1,10000}$
	X_2	0.320	$x_{2,2}$	$x_{2,3}$	$\cdots$	$x_{2,10000}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	X_{40}	0.172	$x_{40,2}$	$x_{40,3}$	$\cdots$	$x_{40,10000}$
$\hat{\mu} =$	$\bar{X}_{40}$	0.412	$\bar{x}_2$	$\bar{x}_3$	$\cdots$	$\bar{x}_{10000}$
$\hat{\sigma}^2 =$	$\bar{S}^2$	0.160	s_2^2	s_3^2	$\cdots$	s_{10000}^2

Sampling Distribution of $\bar{X}$

The sampling distribution of $\bar{X}$ is the distribution of the following vector:

		Samples from Pop.				
		1	**2**	**3**	$\cdots$	**10000**
	X_1	0.496	$x_{1,2}$	$x_{1,3}$	$\cdots$	$x_{1,10000}$
	X_2	0.320	$x_{2,2}$	$x_{2,3}$	$\cdots$	$x_{2,10000}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	X_{40}	0.172	$x_{40,2}$	$x_{40,3}$	$\cdots$	$x_{40,10000}$
$\hat{\mu} =$	$\bar{X}_{40}$	0.412	$\bar{x}_2$	$\bar{x}_3$	$\cdots$	$\bar{x}_{10000}$
$\hat{\sigma}^2 =$	$\bar{S}^2$	0.160	s_2^2	s_3^2	$\cdots$	s_{10000}^2

Calculating Sampling Distribution with Known Population

1. Start with complete population.
2. Define a quantity of interest (the parameter). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}$.
4. Draw a sample from the population and calculate the estimate using the estimator.
5. Repeat step 4 many times (we will do 10,000).

Calculating Sampling Distribution with Known Population

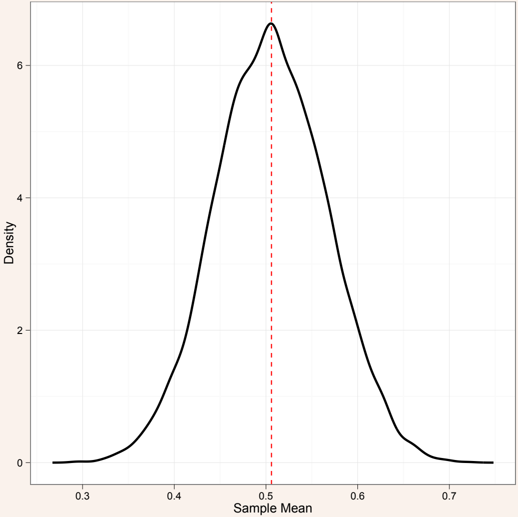
We already saw how to take one sample in R, but let's now repeat it 10,000 times and store $\bar{X}$ for each sample:

```
set.seed(12345)
n.sample <- 40
N <- nrow(precincts)
xbar.vec <- replicate(n=10000, mean(precincts[sample(N
, size=n.sample, replace=FALSE),]$black))

plot(density(xbar.vec), col =  "navy", lwd=2,
main = "Sampling Distribution of Sample Mean",
xlab="Sample Mean")
```

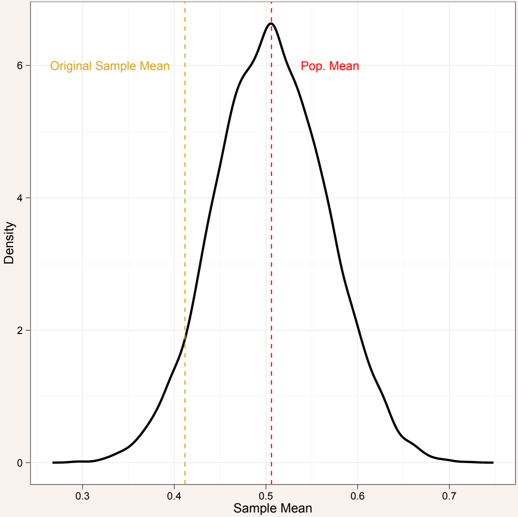
Visualizing Sampling Distributions

Here is the sampling distribution for $\bar{X}$:



Visualizing Sampling Distributions

Here is the sampling distribution for $\bar{X}$:



Characterizing Sampling Distribution of Sample Mean

We know that in large sample (large enough for Central Limit Theorem to kick in), the sample mean will be distributed as:

$$\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$$

Sampling Distributions for Other Statistics

We can calculate infinitely many statistics from a sample. If we have the population, we can simulate a sampling distribution for those! For example, we can look at the sampling distribution of S - the sample standard deviation:

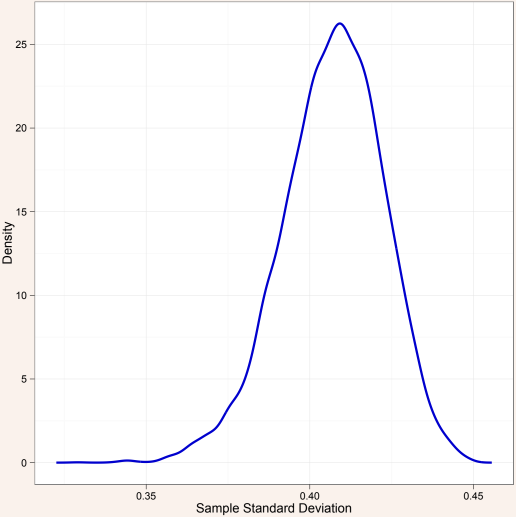
```
set.seed(12345)
n.sample <- 40
N <- nrow(precincts)

sample.fxn <- function(){
poll.i <- precincts[sample(N
, size=n.sample, replace=FALSE),]$black
return(c(mean(poll.i), sd(poll.i)))
}

out.df <- replicate(n=10000, sample.fxn())
head(out.df)
```

Visualizing Sampling Distributions

Here is the sampling distribution for *S*:



Outline

Sampling Distribution of $\bar{X}$ with Unknown Population

In reality, we only draw one sample and cannot directly observe the sampling distribution of $\bar{X}$:

Samples from Pop.

		1	**2**	**3**	$\cdots$	**10000**
	X_1	0.496	?	?	$\cdots$	?
	X_2	0.320	?	?	$\cdots$	?
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	X_{40}	0.172	?	?	$\cdots$	?
$\hat{\mu} =$	$\bar{X}_{40}$	0.412	?	?	$\cdots$	?
$\hat{\sigma}^2 =$	$\bar{S}^2$	0.160	?	?	$\cdots$	?

How do we estimate sampling distribution?

Characterizing Sampling Distribution of Sample Mean

We know that in large sample (large enough for Central Limit Theorem to kick in), the sample mean will be distributed as:

$$\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$$

Estimating the Sampling Distribution of Sample Mean

We can estimate the equation in the previous slide using estimators for the population mean μ and the population variance σ^2. We will use the following estimators:

- $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $\hat{\sigma^2} = S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

The estimated sampling distribution is thus:

$$\bar{X} \sim \mathcal{N}\left(\bar{X}, S_X^2/n\right)$$

Bootstrapping the Sampling Distribution of Sample Mean

1. Start with sample.
2. Define a quantity of interest (the parameter). For us, it's μ.
3. Choose a plausible estimator. For us, it's $\hat{\mu} = \bar{X}$.
4. Take a resample of size n (with replacement) from the sample and calculate the estimate using the estimator.
5. Repeat step 4 many times (we will do 10,000).

Bootstrapping the Sampling Distribution of Sample Mean

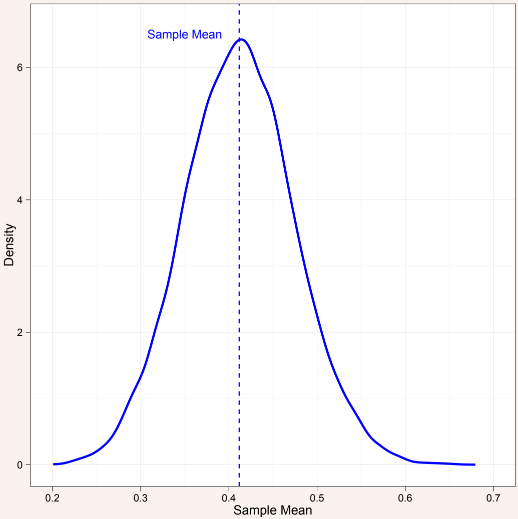
Recall at the beginning of lecture, we already took a sample of size $n = 40$ from the population and stored the resultant dataframe as the object `mypoll`.

```
set.seed(12345)
n.sample <- 40
xbar.vec.bs <- replicate(n=10000, mean(sample(mypoll$
    black
, size=n.sample, replace=TRUE)))

plot(density(xbar.vec.bs), col =  "navy", lwd=2,
main = "Bootstrapped Sampling Distribution of Sample
    Mean",
xlab="Sample Mean")
```

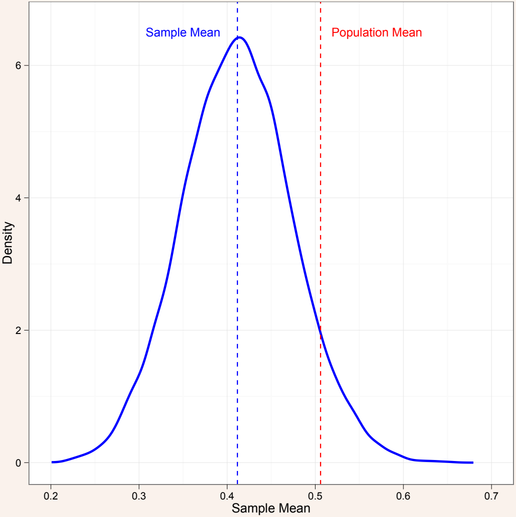
Visualizing the Bootstrapped Sampling Distribution

Here is the bootstrapped sampling distribution:



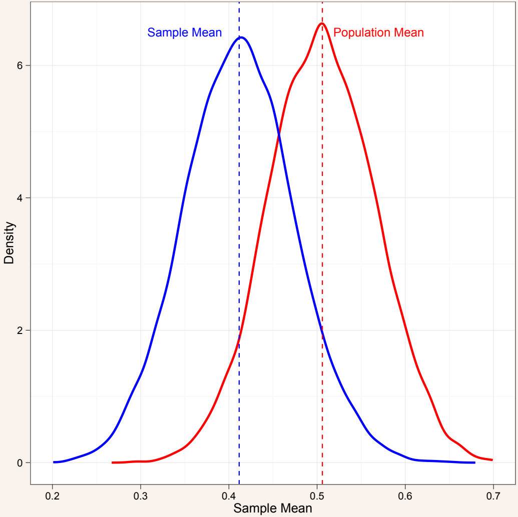
Visualizing the Bootstrapped Sampling Distribution

However, any given sample may be far away from the truth!



Visualizing the Bootstrapped Sampling Distribution

However, any given sample may be far away from the truth!



Conclusion

ANY QUESTIONS?