Gov 50: 17. Sums and Means in Large Samples

Matthew Blackwell

Harvard University

Fall 2018

1. Today's agenda
2. Sample means
3. Normal distribution
4. Central limit theorem

1/ Today's agenda

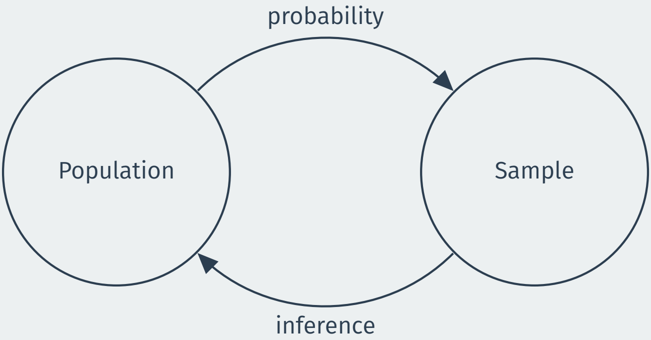
Logistics

- HW 4 due Thursday.
- Groups have been determined for Harvard College students.
    - Paragraph describing data and research question due 11/21
- Midterm 2 next Thursday
    - Review session on Tuesday.

Where are we? Where are we going?

- Last time: defined random variables.
- This time: connect them to data more carefully.
- What happens to our sample means as our samples get big?
  - Law of large numbers
  - Central limit theorem

Learning about populations



- **Probability**: formalize the uncertainty about how our data came to be.
- **Inference**: learning about the population from a set of data.

2/ Sample means

Fulton county data

- `fulton.csv`: data on **all** registered voters in Fulton County, GA in 1994.
- Data on the entire population is a **census**

Name	Description
`turnout`	did person vote (1) or not (0) in 1994?
`black`	is this person black (1) or not (0)?
`sex`	is this person a woman (1) or not (0)?
`age`	age
`dem`	is this person registered as a Democrat (1) or not (0)?
`rep`	is this person registered as a Republican (1) or not (0)?
`urban`	registered in a city (1) or not (0)?

Load Fulton county data

```r
fulton <- read.csv("data/fulton.csv")
head(fulton)
```

```
##   turnout black sex age dem rep urban
## 1       0     0   1  19   0   0     0
## 2       0     0   0  35   0   0     0
## 3       0     1   0  36   0   0     1
## 4       1     0   0  27   0   0     1
## 5       1     1   1  79   1   0     1
## 6       1     0   1  42   1   0     0
```

Large random samples

- In real data, we will have a set of n measurements on a variable:

$$X_1, X_2, \dots, X_n$$

  - X_1 is the age of the first randomly selected registered voter.
  - X_2 is the age of the second randomly selected registered voter, etc.
- Empirical analyses: sums or means of these n measurements
  - Almost all statistical procedures involve a sum/mean.
  - What are the properties of these sums and means?
  - Can the sample mean of age tell us anything about the population distribution of age?
- **Asymptotics**: what can we learn as n gets big?

Sums and means are random variables

- If X_1 and X_2 are r.v.s, then $X_1 + X_2$ is a r.v.
  - Has a mean $\mathbb{E}[X_1 + X_2]$ and a variance $\mathbb{V}[X_1 + X_2]$
- The **sample mean** is a function of sums and so it is a r.v. too:

$$\overline{X} = \frac{X_1 + X_2}{2}$$

- This is the average age of two randomly selected respondents.

Distribution of sums/means

	X_1	X_2	$X_1 + X_2$	$\bar{X}$
draw 1	61	29	90	45
draw 2	23	63	86	43
draw 3	24	47	71	35.5
draw 4	52	46	98	49
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			distribution of the sum	distribution of the mean

Independent and identical r.v.s

- Often work with **independent and identically distributed** r.v.s, $X_1, \ldots, X_n$
  - Random sample of n respondents on a survey question.
  - Written "i.i.d."
- **Independent**: value that X_i takes doesn't affect distribution of X_j
- **Identically distributed**: distribution of X_i is the same for all i
  - $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots = \mathbb{E}(X_n) = \mu$
  - $\mathbb{V}(X_1) = \mathbb{V}(X_2) = \cdots = \mathbb{V}(X_n) = \sigma^2$

Distribution of the sample mean

- **Sample mean** of i.i.d. random variables:

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- $\overline{X}_n$ is a random variable, what is its distribution?
    - What is the expectation of this distribution, $\mathbb{E}[\overline{X}_n]$?
    - What is the variance of this distribution, $\mathbb{V}[\overline{X}_n]$?
    - These will help us know where we should expect the sample mean to be.
- Fulton County data:
    - The average age in a one random sample is different than the average age in another random sample.
    - Will the average age in the sample be close to the population age?

Properties of the sample mean

Mean and variance of the sample mean

Suppose that $X_1, \dots, X_n$ are i.i.d. r.v.s with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2$. Then:

$$\mathbb{E}[\overline{X}_n] = \mu \qquad \mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}$$

- Key insights:
  - Sample mean is on average equal to the population mean
  - Variance of $\overline{X}_n$ depends on the population variance of X_i and the sample size
- Standard deviation of the sample mean is called its **standard error**:

$$SE = \sqrt{\mathbb{V}[\overline{X}_n]} = \frac{\sigma}{\sqrt{n}}$$

Law of large numbers

Law of Large Numbers

Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ and finite variance σ^2. Then, $\overline{X}_n$ converges to μ as n gets large.

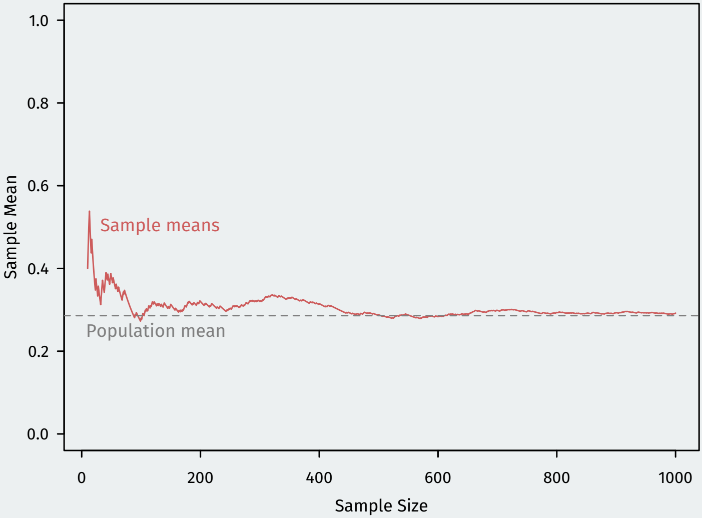
- Intuition: The probability of $\overline{X}_n$ being "far away" from μ goes to 0 as n gets big.
- The distribution of sample mean "collapses" to population mean.

LLN by simulation in R

- Draw a random sample of 1000 from Fulton County data.
- Compare the sample average of Democratic registration as we include more of this sample.
- Like drawing random samples of size 1, 2, 3, 5, . . ., 999, 1000.

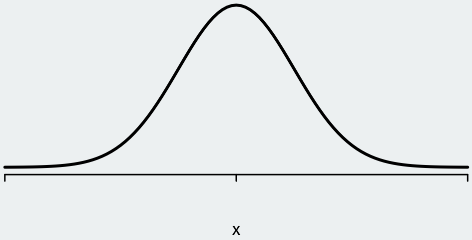
```r
dem.mean <- mean(fulton$dem)
sims <- 1000

# draw a random sample of row numbers (with replacement)
samp <- sample(1:nrow(fulton), size = sims, replace = TRUE)
dem.samp <- fulton$dem[samp]

# calculate the mean of the first i values
samp.means <- rep(NA, times = sims)
for (i in 1:sims) {
  samp.means[i] <- sum(dem.samp[1:i]) / i
}
```

LLN in action



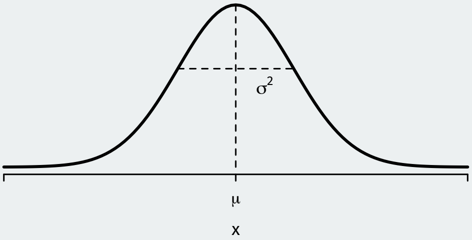
3/ Normal distribution

Normal r.v.



- The **normal distribution** is the classic “bell-shaped” curve.
  - Extremely ubiquitous in statistics.
  - “Sums and means of random variables tend to follow a normal distribution”
- Three key properties:
  - **Unimodal**: one peak at the mean.
  - **Symmetric** around the mean.
  - **Everywhere positive**: any real value can possibly occur.

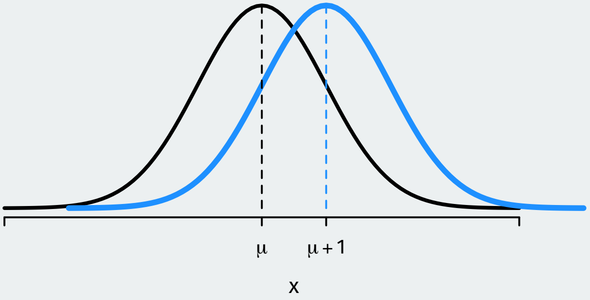
Normal distribution



- A normal distribution can be affect by two values:
  - **mean/expected value** usually written as μ
  - **variance** written as σ^2 (standard deviation is σ)
  - Written $X \sim N(\mu, \sigma^2)$.
- **Standard normal distribution**: mean 0 and standard deviation 1.

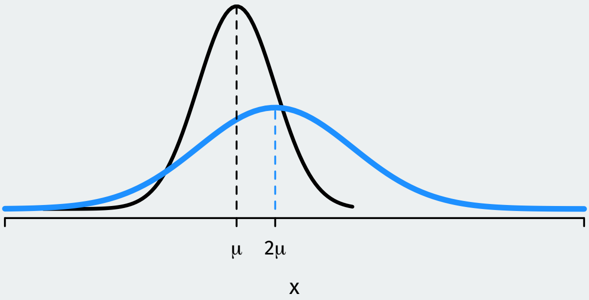
Reentering and scaling the normal

- How do transformations of a normal work?
- Let $X \sim N(\mu, \sigma^2)$ and c be a constant.
- If $Z = X + c$, then $Z \sim N(\mu + c, \sigma^2)$.
- Intuition: adding a constant to a normal shifts the distribution by that constant.



Recentering and scaling the normal

- Let $X \sim N(\mu, \sigma^2)$ and c be a constant.
- If $Z = cX$, then $Z \sim N(c\mu, (c\sigma)^2)$.
- Intuition: multiplying a normal by a constant scales the mean and the variance.



Z-scores of normals

- These two facts imply the **z-score** of a normal variable is a standard normal:

$$z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- Subtract the mean and divide by the SD $\rightsquigarrow$ standard normal.
- z-score measures how many SDs away from the mean a value of X is.

4/ Central limit theorem

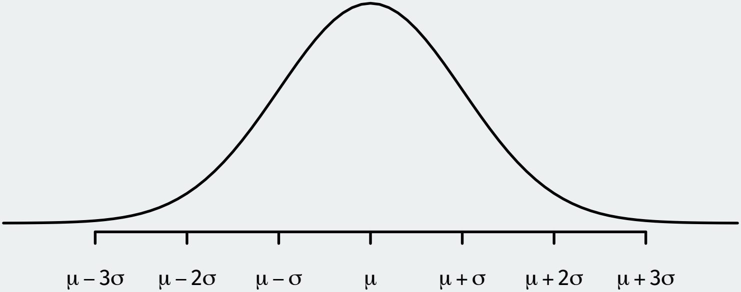
Central limit theorem

Central limit theorem

Let $X_1, \ldots, X_n$ be i.i.d. r.v.s from a distribution with mean μ and variance σ^2. Then, $\overline{X}_n$ will be approximately distributed $N(\mu, \sigma^2/n)$ in large samples.

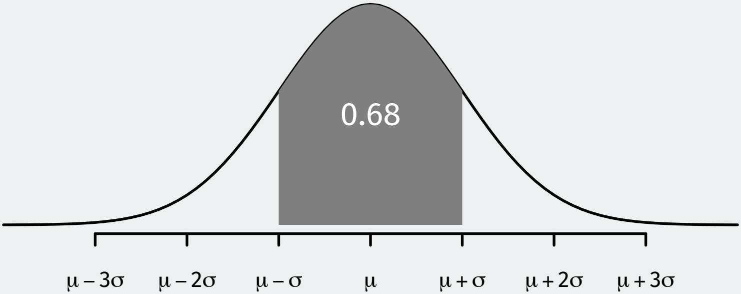
- Approximation is better as n goes up.
- "Sample means tend to be normally distributed as samples get large."
- $\rightsquigarrow$ we know how far away $\overline{X}_n$ will be from its mean.

Empirical Rule for the Normal Distribution



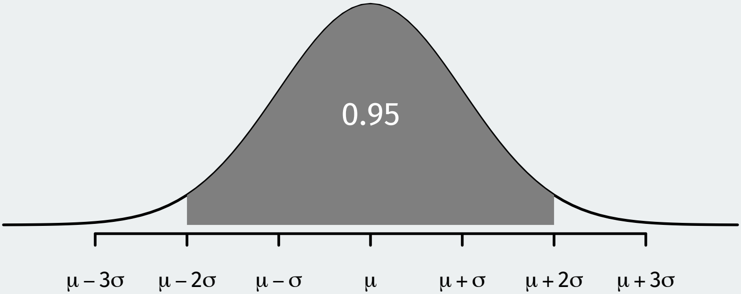
- If $X \sim N(\mu, \sigma^2)$, then:

Empirical Rule for the Normal Distribution



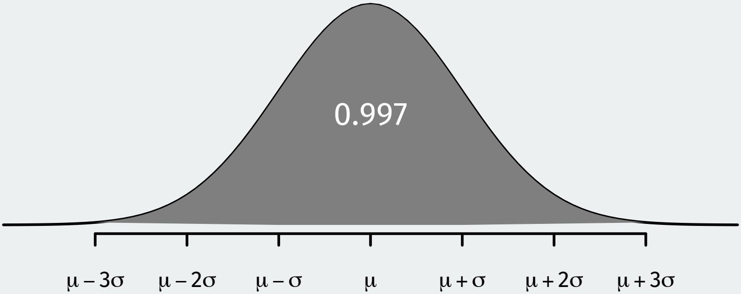
- If $X \sim N(\mu, \sigma^2)$, then:
  - $\approx$ 68% of the distribution of X is within 1 SD of the mean.

Empirical Rule for the Normal Distribution



- If $X \sim N(\mu, \sigma^2)$, then:
  - $\approx$ 68% of the distribution of X is within 1 SD of the mean.
  - $\approx$ 95% of the distribution of X is within 2 SDs of the mean.

Empirical Rule for the Normal Distribution

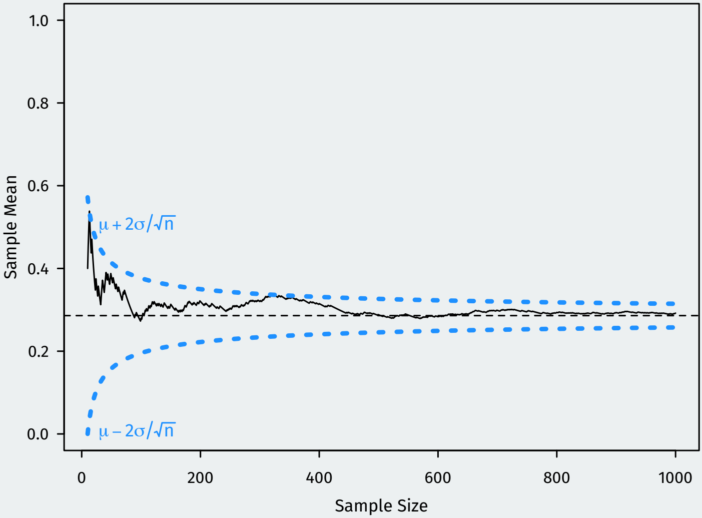


- If $X \sim N(\mu, \sigma^2)$, then:
  - $\approx$ 68% of the distribution of X is within 1 SD of the mean.
  - $\approx$ 95% of the distribution of X is within 2 SDs of the mean.
  - $\approx$ 99.7% of the distribution of X is within 3 SDs of the mean.

Why the CLT?

- By CLT, sample mean $\approx$ normal with mean μ and SD $\frac{\sigma}{\sqrt{n}}$.
- By empirical rule, sample mean will be within $2 \times \frac{\sigma}{\sqrt{n}}$ of the population mean 95% of the time.

CLT in action



CLT simulation

1. Draw a sample of size 1000 from the Fulton county population.
2. Calculate the sample mean of Democratic registration (`dem`) for that sample.
3. Save the sample mean.
4. Repeat steps 1-3 a large number of times.

CLT in action

```r
dem.sigma <- sd(fulton$dem)
n <- 1000
sims <- 5000
dem.means <- rep(NA, times = sims)
for (i in 1:sims) {
  ## take i.i.d. sample of row numbers
  samp.ind <- sample(1:nrow(fulton), size = n,
                     replace = TRUE)

  ## get the values of "dem" for sample
  dem.sample <- fulton$dem[samp.ind]

  ## record mean of this sample
  dem.means[i] <- mean(dem.sample)
}
```

```r
## mean and sd of the sample means from each
## repeated sample
mean(dem.means)
```

```
## [1] 0.286
```

```r
sd(dem.means)
```

```
## [1] 0.0142
```

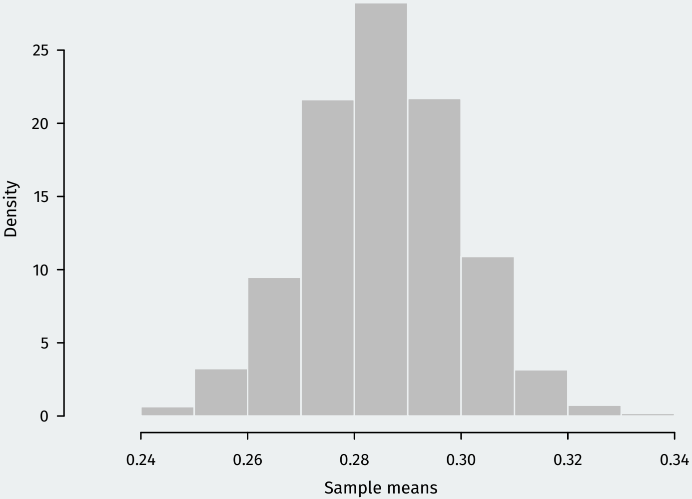
```r
## compare to what the CLT predicts from population
mean(fulton$dem)
```

```
## [1] 0.286
```

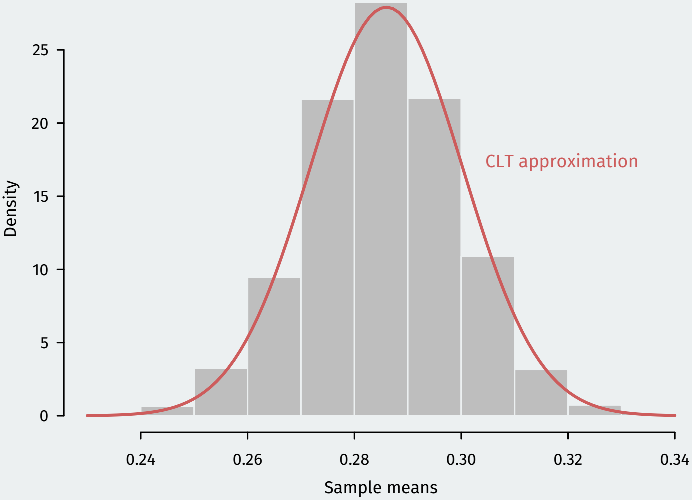
```r
sd(fulton$dem)/sqrt(n)
```

```
## [1] 0.0143
```

Histogram of sample means



Histogram of sample means



Last points

- We usually only 1 sample, so we'll only get 1 sample mean.
- Why do we care about LLN/CLT?
  - CLT gives us assurances our 1 sample mean will won't be too far from population mean.
  - CLT will also help us create measure of uncertainty for our estimates.

Next time

- Today: learning about samples given population information
- Next: Learning about population values from the sample.