Gov 51: Large Sample Theorems and the Normal Distribution

Matthew Blackwell

Harvard University

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`	registered as a Democrat (1) or not (0)?
`rep`	registered as a Republican (1) or not (0)?
`urban`	registered in a city (1) or not (0)?

Load Fulton county data

```
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:

$$X_1, X_2, \ldots, X_n$$

  - 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.

- What are the properties of the sample mean of these measurements?
  - Expectation: $\mathbb{E}(\overline{X}) = \mathbb{E}[X_i] = \mu$
  - Variance: $V(\overline{X}) = \mathbb{V}(X_i)/n = \sigma_X^2/n$
  - Valid for any sample size!

- **Asymptotics**: what can we learn as n gets big?

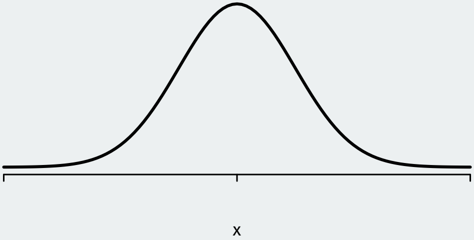
Law of large numbers

Law of Large Numbers

Let $X_1, \dots, X_n$ 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.

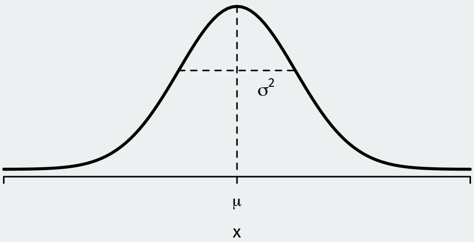
- Probability of $\overline{X}_n$ being “far away” from μ goes to 0 as n gets big.
- The distribution of sample mean “collapses” to population mean.
- Can see this from the variance of $\overline{X}_n$: $\mathbb{V}[X]/n$.

Normal random variable



- A **normal distribution** has a PDF that is the classic “bell-shaped” curve.
    - Extremely ubiquitous in statistics.
    - An r.v. is more likely to be in the center, rather than the tails.
- Three key properties of this PDF:
    - **Unimodal**: one peak at the mean.
    - **Symmetric** around the mean.
    - **Everywhere positive**: any real value can possibly occur.

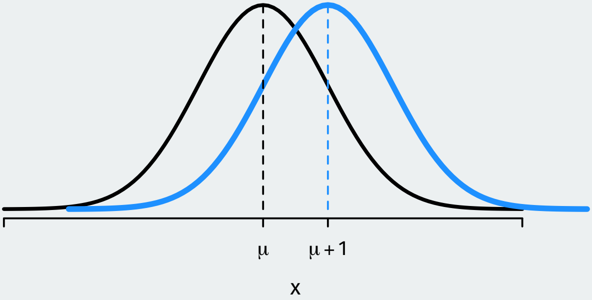
Normal distribution



- 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.

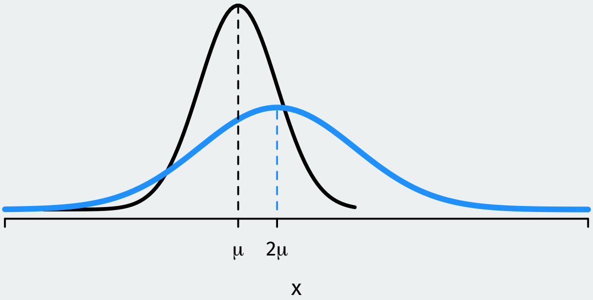
Reentering and scaling the normal

- 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 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.

Central limit theorem

Central limit theorem

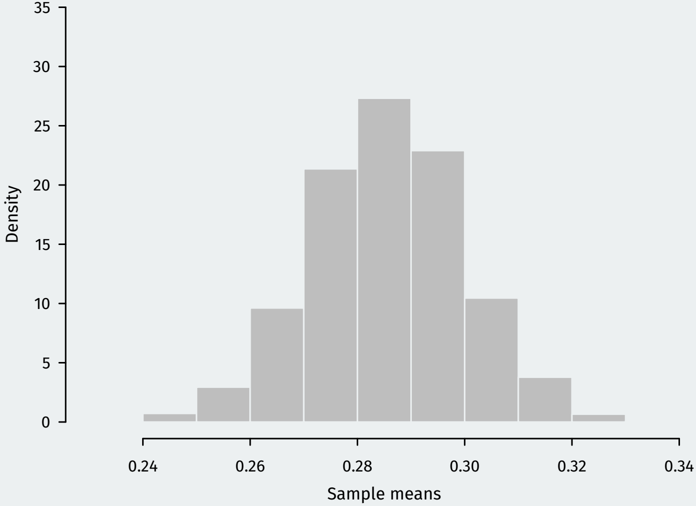
Let $X_1, \ldots, X_n$ be i.i.d. r.v.s from a distribution with mean μ and variance σ^2. Then, $\overline{X}_n$ will be approximately distributed $N(\mu, \sigma^2/n)$ in large samples.

- "Sample means tend to be normally distributed as samples get large."
- $\rightsquigarrow$ we know (an approx. of) the entire probability distribution of $\overline{X}_n$
  - Approximation is better as n goes up.
  - Does not depend on the distribution of X_i!

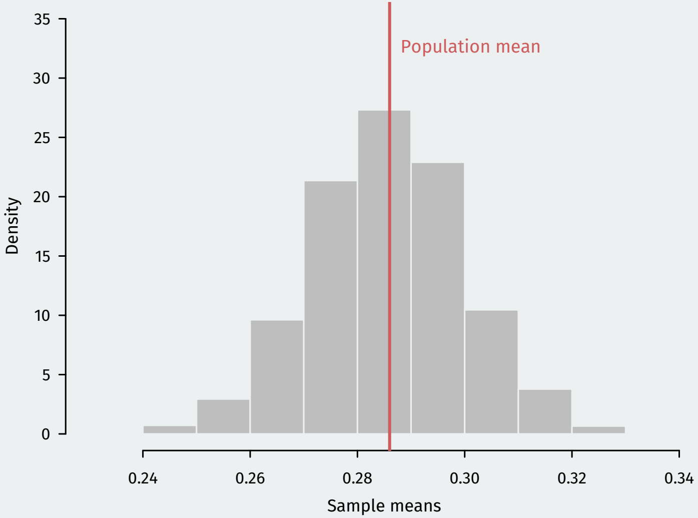
CLT simulation

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.

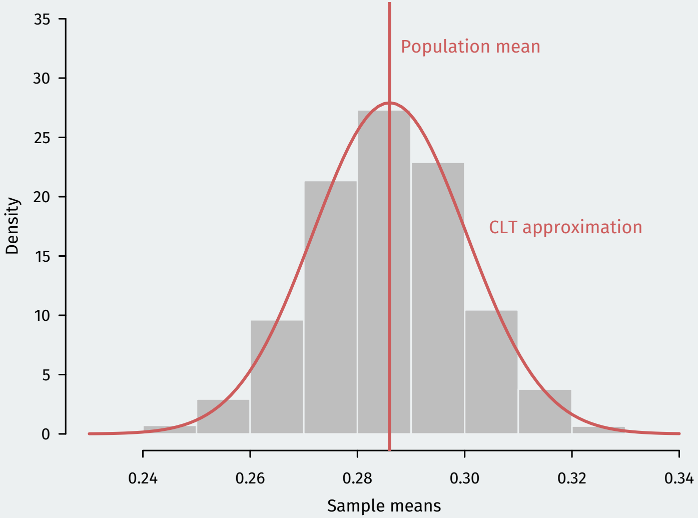
Histogram of sample means



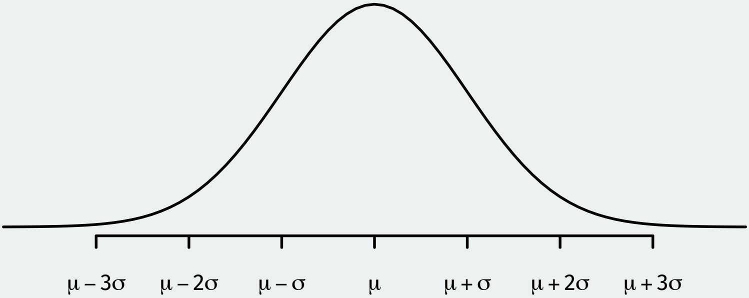
Histogram of sample means



Histogram of sample means

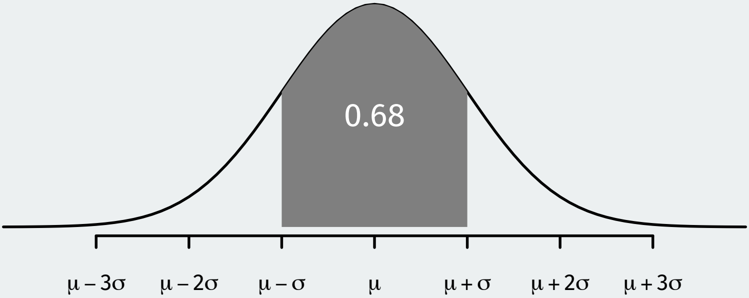


Empirical Rule for the Normal Distribution



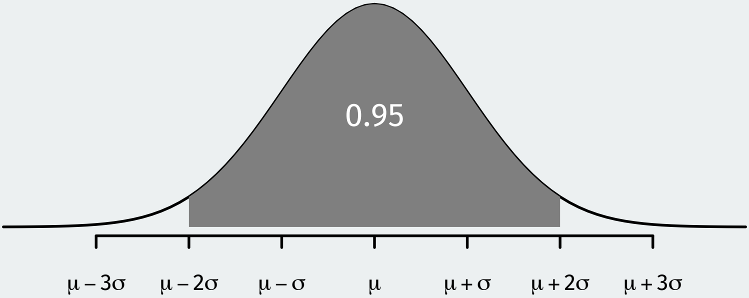
- If $X \sim N(\mu, \sigma^2)$, then:

Empirical Rule for the Normal Distribution



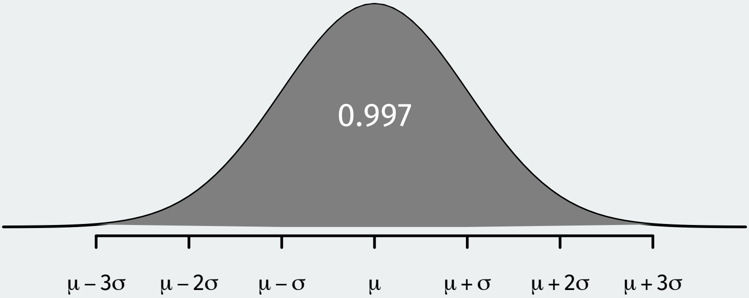
- If $X \sim N(\mu, \sigma^2)$, then:
  - $\approx$ 68% of the distribution of X is within 1 SD of the mean.

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?

- Why do we care about CLT?
  - We usually only 1 sample, so we'll only get 1 sample mean.
  - Implies our 1 sample mean will won't be too far from population mean.
- By CLT, sample mean $\approx$ normal with mean μ and SD $\frac{\sigma}{\sqrt{n}}$.
- By empirical rule, sample mean will be...
  - Between $\mu - 2 \times \frac{\sigma}{\sqrt{n}}$ and $\mu + 2 \times \frac{\sigma}{\sqrt{n}}$ 95% of the time.
- This will also help us create measure of uncertainty for our estimates.