JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section F

The Theoretical Sampling Distribution of the Sample Proportion and Its Estimate Based on a Single Sample

Sampling Distribution of the Sample Mean

- In the previous section we reviewed the results of simulations that resulted in estimates of what was formally called the sampling distribution of a sample proportion

- The sampling distribution of a sample proportion is a theoretical probability distribution
  - It describes the distribution of all sample proportions from all possible random samples of the same size taken from a population

Sampling Distribution of the Sample Mean

- In real research it is impossible to estimate the sampling distribution of a sample mean by actually taking multiple random samples from the same population (no research would ever happen if a study needed to be repeated multiple times) to understand this sampling behavior

- Simulations are useful to illustrate a concept, but not to highlight a practical approach!

- Luckily, there is some mathematical machinery that generalizes some of the patterns we saw in the simulation results

The Central Limit Theorem (CLT)

- The Central Limit Theorem (CLT) is a powerful mathematical tool that gives several useful results
  - The sampling distribution of sample proportions based on all samples of same size *n* is approximately normal
  - The mean of all sample proportions in the sampling distribution is the true mean of the population from which the samples were taken, *p*
  - The standard deviation in the sample proportions of size *n* is equal to $\sqrt{\frac{p \times (1-p)}{n}}$
  - This is often called the standard error of the sample proportion and sometimes written as $SE(\hat{p})$

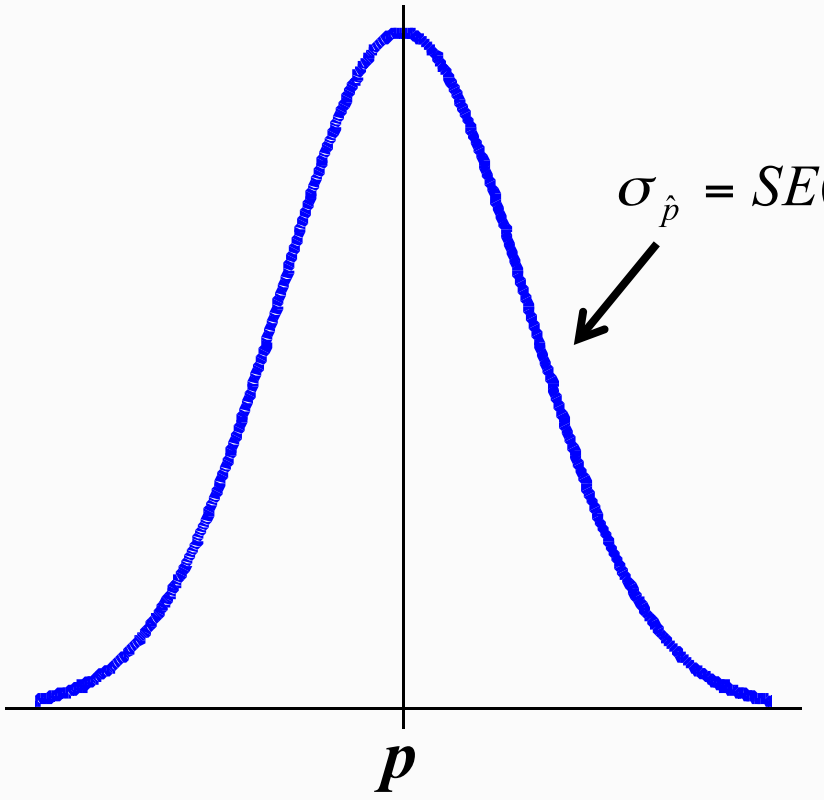
Example: Blood Pressure of Males

- Population distribution of individual insurance status
  - True proportion $p = 0.8$

Sample Sizes	Means of 500 Sample Proportions	Means of 5000 Sample Proportions	SD of 500 Sample Proportion	SD of 5000 Sample Proportions	SD of Sample Proportions (SE) by CLT
n = 20	0.805	0.799	0.094	0.090	0.089
n = 100	0.801	0.799	0.041	0.040	0.040
n = 1,000	0.799	0.80	0.012	0.012	0.012

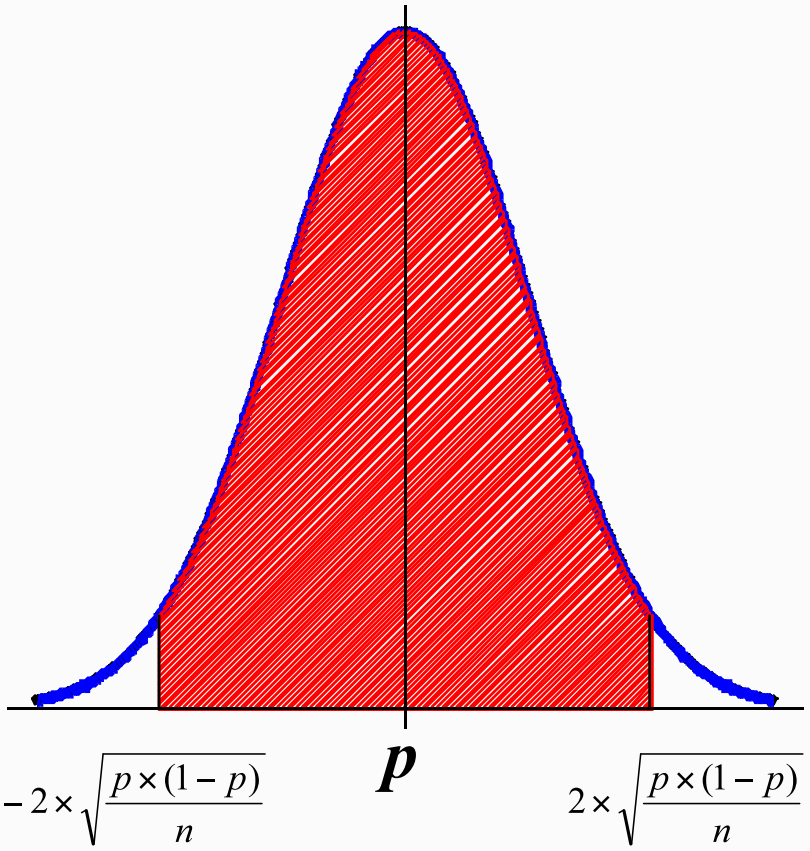
Recap: CLT

- So the CLT tells us the following:
  - When taking a random sample of binary measures of size *n* from a population with true proportion *p* the theoretical sampling distribution of sample proportions from all possible random samples of size *n* is:



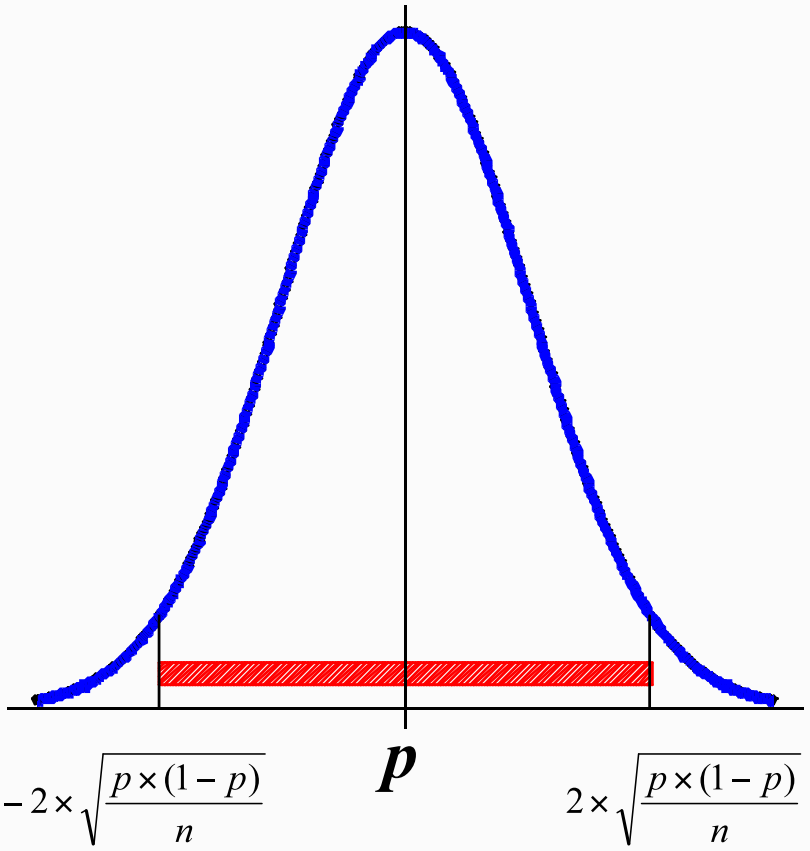
CLT: So What?

- So what good is this info?
  - Well using the properties of the normal curve, this shows that for most random samples we can take (95%), the sample proportion $\hat{p}$ will fall within 2 SEs of the true proportion *p*:



CLT: So What?

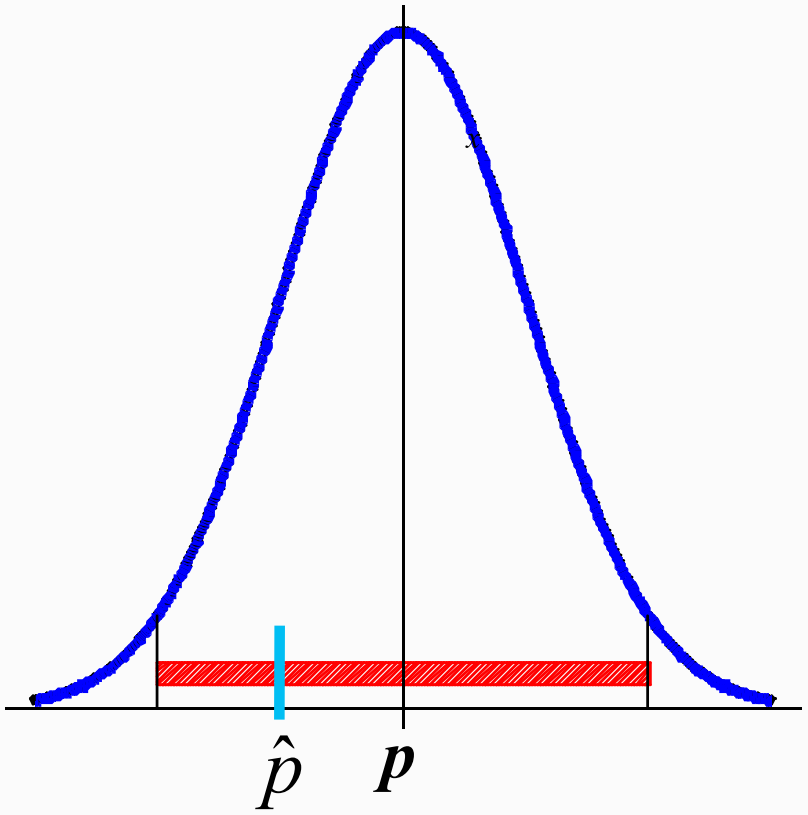
- So AGAIN what good is this info?
  - We are going to take a single sample of size *n* and get one $\hat{p}$
  - So we won't know *p* and if we did know *p* why would we care about the distribution of estimates of *p* from imperfect subsets of the population?



CLT: So What?

- We are going to take a single sample of size *n* and get one $\hat{p}$

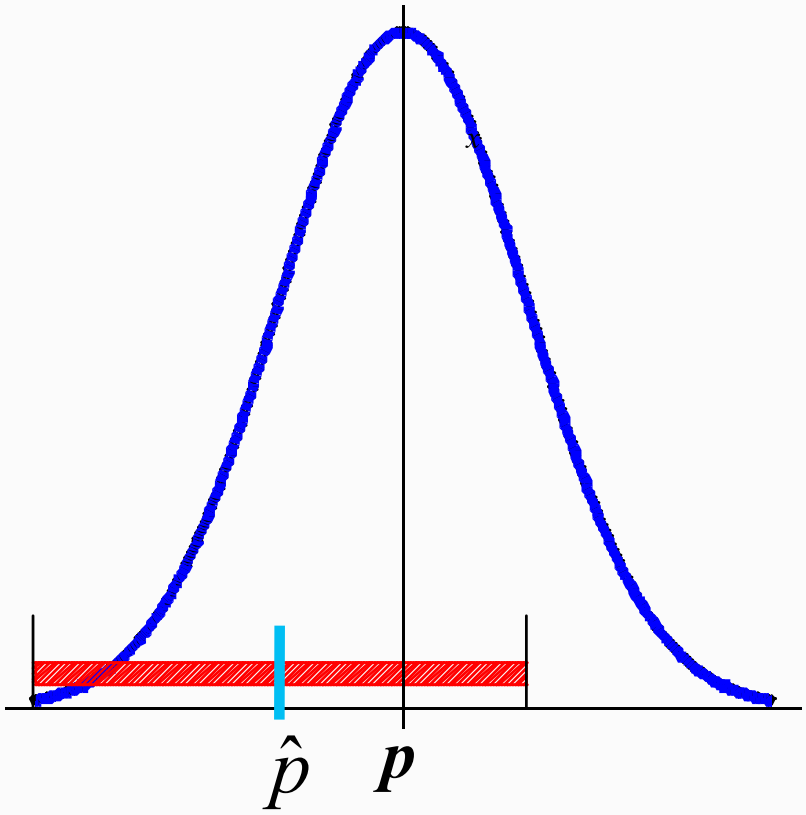
- But for most (95%) of the random samples we can get, our $\hat{p}$ will fall within +/- 2 SEs of *p*



CLT: So What?

- We are going to take a single sample of size *n* and get one $\hat{p}$

- So if we start at $\hat{p}$ and go 2 SEs in either direction, the interval created will contain *p* most (95 out of 100) of the time



Estimating a Confidence Interval

- Such and interval is a called a 95% confidence interval for the population proportion *p*
- Interval given by $\hat{p} \pm 2SE(\hat{p}) \rightarrow \bar{x} \pm 2 \times \sqrt{\frac{p \times (1 - p)}{n}}$
- Problem: we don't know *p*
  - Can estimate with $\hat{p}$, will detail this in next section
- What is interpretation of a confidence interval?

Interpretation of a 95%Confidence Interval (CI)

- Laypersons' range of "plausible" values for true proportion
  - Researcher never can observe true mean *p*
  - $\hat{p}$ is the best estimate based on a single sample
  - The 95% CI starts with this best estimate and additionally recognizes uncertainty in this quantity

- Technical
  - Were 100 random samples of size *n* taken from the same population, and 95% confidence intervals computed using each of these 100 samples, 95 of the 100 intervals would contain the values of true proportion *p* within the endpoints

Notes on Confidence Intervals

- Random sampling error
  - Confidence interval only accounts for random sampling error, not other systematic sources of error or bias

Notes on Confidence Intervals

- Are all CIs 95%?
  - No
  - It is the most commonly used
  - A 99% CI is wider
  - A 90% CI is narrower

- To change level of confidence adjust number of SE added and subtracted from $\hat{p}$
  - For a 99% CI, you need ± 2.6 SE
  - For a 95% CI, you need ± 2 SE
  - For a 90% CI, you need ± 1.65 SE

Summary

- What did we see with this set of examples

- A couple of trends:
  - Distribution of sample proportions tended to be approximately normal—even when original—and individual level data was not (binary outcome)
  - Variability in sample mean values decreased as the size of the sample each proportion was based upon increased

Clarification

- As with means for continuous data, variation in proportions values tied to the size of each sample selected in our exercise: NOT the number of samples