ECO220Y

Estimation:

Confidence Interval Estimator for Sample Proportions

Readings: Chapter 11 (skip 11.5)

Fall 2011

Lecture 10

Review: Sampling Distributions

- Sample proportion, $\hat{p}$
  1. $\hat{p} \sim N(p, p(1-p)/n)$
  2. Bell-shaped only if the rule of thumb holds: $p \pm 3\sqrt{p(1-p)/n}$
- Sample mean, $\bar{X}$
  1. $\bar{X} \sim N(\mu_x, \frac{\sigma_x^2}{n})$
  2. Use Central Limit Theorem to learn about the shape of the distribution
     1. if population is normal, then sampling distribution of $\bar{X}$ is normal for $n \geq 1$
     2. What would be the shape of the distribution of $\bar{X}$ if $n = 1$?
     3. If population is not normal, then $n \geq 30$ is sufficient
     4. For modest departures from normal, $n < 30$ is sufficient
- The idea is to "imagine" all possible values of a sample mean or sample proportion feasible with a given sample size, n
- Mean and standard deviation of the sampling distribution - parameters or statistics?

Warm-Up Example

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower.
For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?

1. The larger hospital
2. The smaller hospital
3. About the same

Warm-Up Example

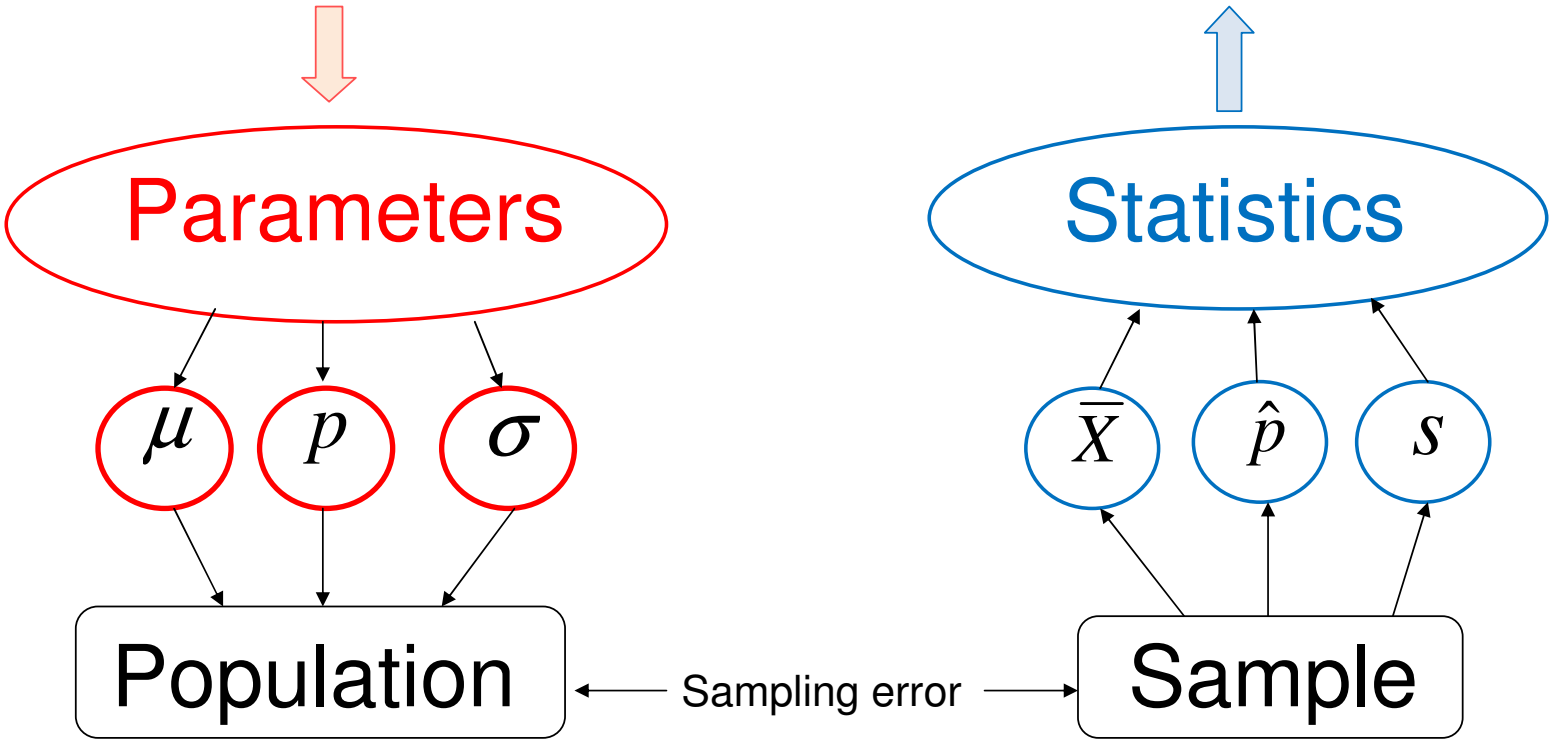
A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower.
For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?

1. The larger (21)
2. The smaller hospital (21)
3. About the same (53)

Statistical Inference

draws

conclusion from evidence



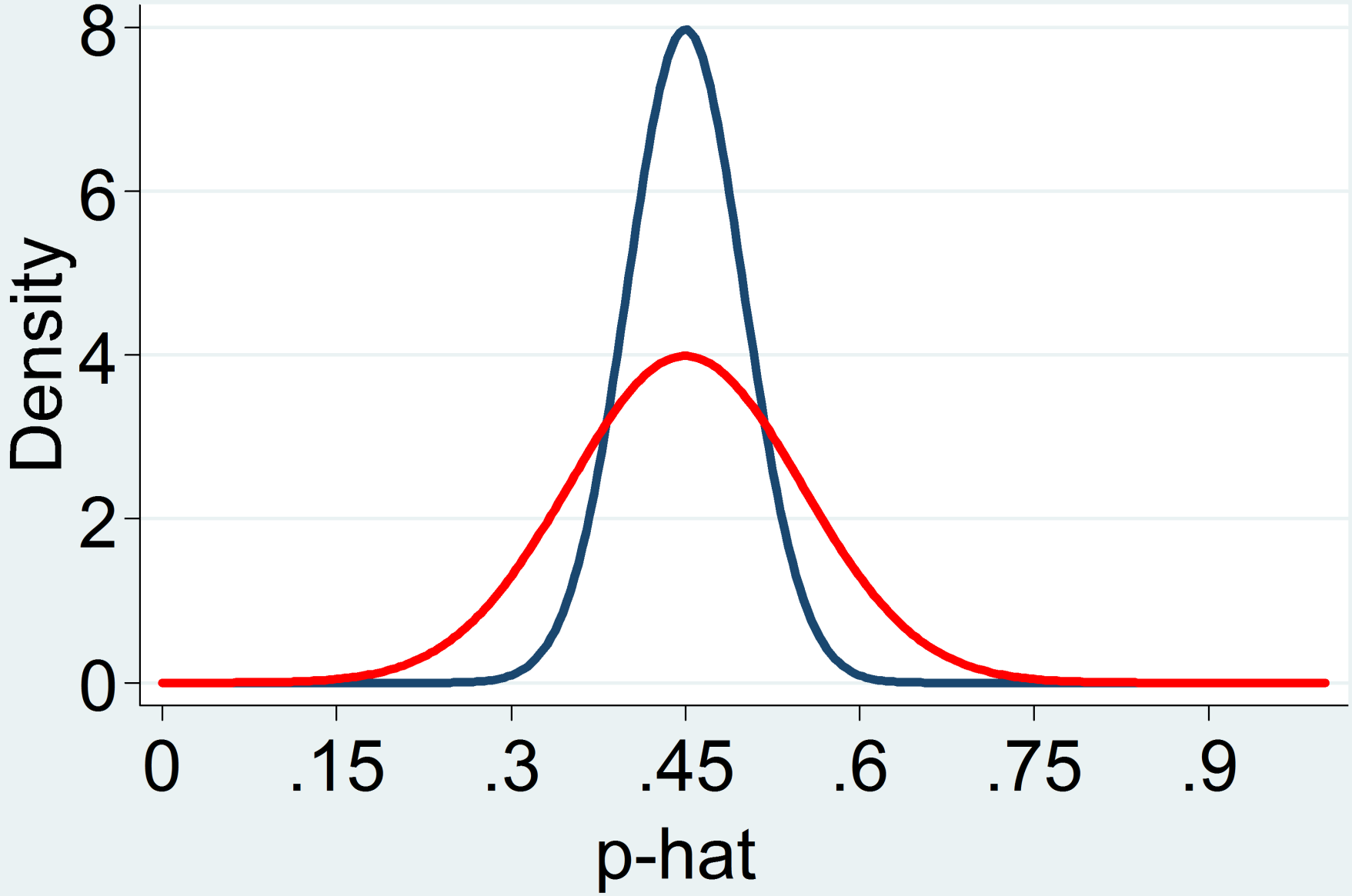
Statistical Inference

1. Descriptive Statistics ✓
2. Probability ✓
3. Inference (remaining time)
    1. Estimation
        - ★ Point estimator - uses a single value
        - ★ Confidence interval estimator - uses a range of values (Lectures 10 and 11)
    2. Hypothesis testing

Point Estimator

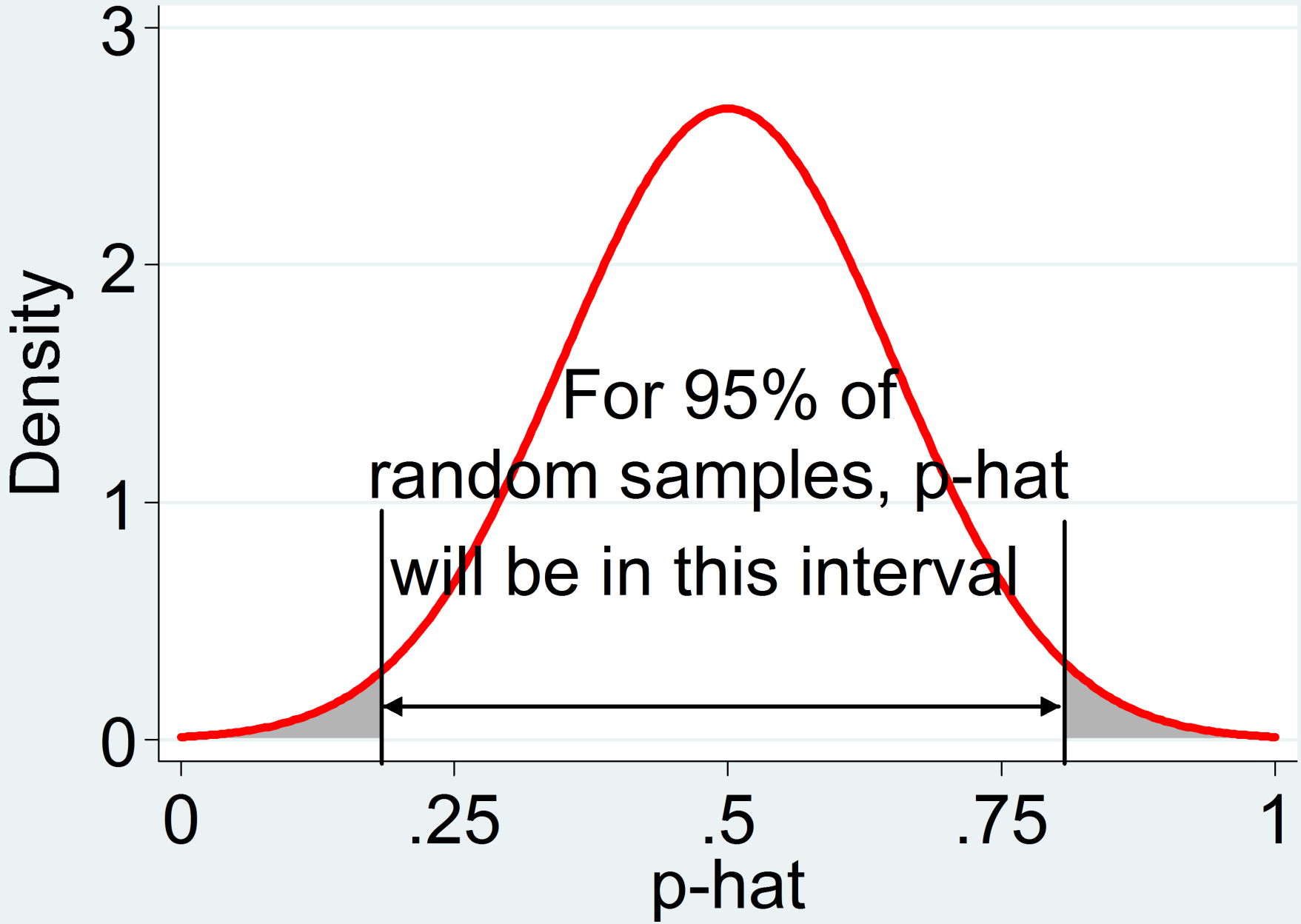
- A point estimator is a formula that produces a value - sample statistic
- Sample statistic ia an estimate of the population parameter
- A point estimator is a random variable - value differs from sample to sample
- Examples of estimators: sample mean and sample proportion
- Properties of good estimators:
  1. Unbiased (expected value of an estimator is equal to the value of the parameter it estimates)
  2. Consistent (variance $\longrightarrow 0$ as $n \longrightarrow \infty$
  3. Efficient (compares variances of two or more estimators)

Sampling Distribution of Sample Proportion



Confidence Interval Estimator

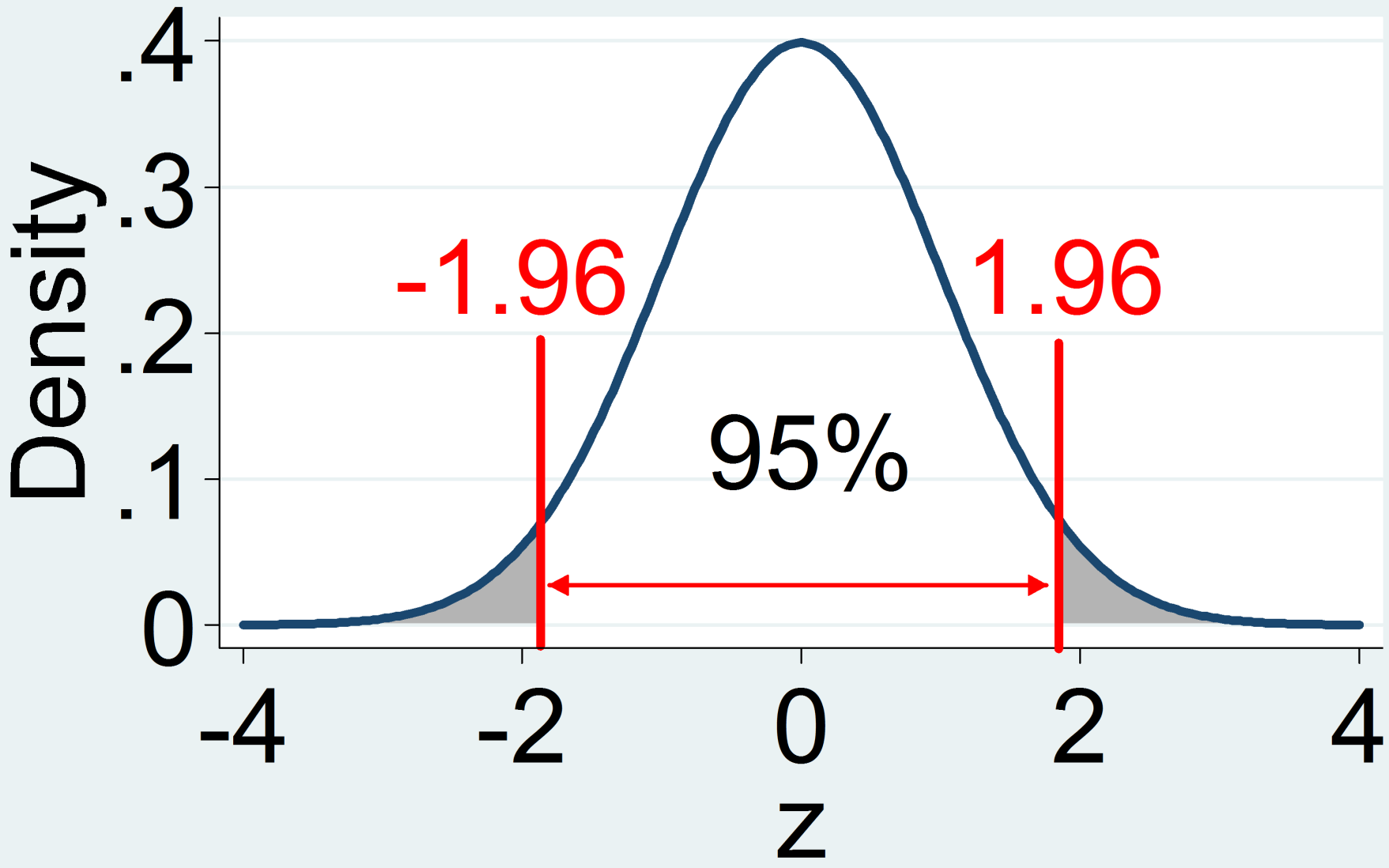
- As a rule, we would like to have more confidence about our estimate of the population parameter than the point estimate gives us.
- Instead of a point estimate, we turn to an interval estimator - a range of values around the point estimator.
- Our goal is to find a range of values around the point estimate in such a way that there is a large probability that this range would include the true population parameter.
- The probability in this case is called "a confidence level", and the range is called "a confidence interval".


Density
3
2
1
0
For 95% of
random samples, p-hat
will be in this interval
0
.25
.5
.75
1
p-hat

Confidence Interval

- We know that $\hat{p} \sim N(p, p(1-p)/n)$ under certain conditions
- If $\hat{p} \sim N$, then Empirical rule holds
- If Empirical rule holds, we know that about 95% random samples will produce $\hat{p}$ that is no more than 2 standard deviations away from the true population proportion, p
- In statistical notation, $P(p - 2SD(\hat{p}) < \hat{p} < p + 2SD(\hat{p})) \approx 0.95$
- To be more precise, $P(p - 1.96SD(\hat{p}) < \hat{p} < p + 1.96SD(\hat{p})) = 0.95$
- What is 1.96? How to get it?
- $P(-1.96 < Z < 1.96) = 0.95$

Standard Normal distribution



Confidence Interval Estimator for Sample Proportion

- 1.96 is a number of standard deviations away from the mean
- In other words, it's a value of z-score which corresponds to the desired level of confidence
- Z-score indicating the number of standard deviations in a confidence interval, is also called the critical value
- We denote confidence level as $1-\alpha$, where α is significance level
- In general, $P(p - z_{\alpha/2} * SD(\hat{p}) < \hat{p} < p + z_{\alpha/2} * SD(\hat{p})) = 1 - \alpha$
- Recall that standard deviation of $(\hat{p}) = \sqrt{p(1-p)/n}$ and re-write:
- $P(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \hat{p} < p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}) = 1 - \alpha$

CI Estimator For Population Proportion

- $\hat{p} \sim N(p, \sqrt{p(1-p)/n})$
- Can use simple math to derive confidence interval estimator for population proportion:

$$1-\alpha = P(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \hat{p} < p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}})$$

$$\downarrow \qquad \downarrow$$

$$1-\alpha = P(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}})$$

$$\downarrow \qquad \downarrow$$

$$\text{or} \quad \boxed{\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}}$$

Confidence Interval Estimator for p

Do we know p - population parameter?
What about $p(1-p)/n$?

CI Estimator for population proportion is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\sqrt{\hat{p}(1-\hat{p})/n}$ is the standard error of sample proportion, an estimate of the standard deviation.

Do not forget to check the rule of thumb:
$\hat{p} \pm 3\sqrt{\hat{p}(1-\hat{p})/n}$ within 0 and 1?

Alternative Significance Levels

- Significance level α is a chance that confidence interval excludes population parameter
- Confidence level $1 - \alpha$ is a chance that confidence interval includes population parameter

Confidence Level (1-α)	α	$\alpha/2$	$z_{\alpha/2}$
99%	1%	0.5%	2.58
98%	2%	1%	2.33
95%	5%	2.5%	1.96
90%	10%	5%	1.64

Mr Noxin - again!

- Assume, Mr Noxin have no idea about the fraction of voters who favour him
- He, however, is eager to estimate the proportion of his supporters
- Two of his friends volunteered to conduct a poll
- In one sample, 55 out of 100 said they would vote for Mr Noxin
- In another sample, 45 out of 100 said they would vote for him

Mr Noxin - again!

- $\hat{p_1} = 0.55, SE(\hat{p_1}) = \sqrt{0.55 * 0.45/100} = 0.0497$
- Lower bound of CI: $\hat{p} - 1.96 * SE(\hat{p}) = 0.55 - 1.96 * 0.0497 = 0.45$
- Upper bound of CI: $\hat{p} + 1.96 * SE(\hat{p}) = 0.55 - 1.96 * 0.0497 = 0.65$
- 95% chance that this interval (0.45, 0.65) includes population proportion

Mr Noxin - again!

- $\hat{p_2} = 0.45, SE(\hat{p_2}) = \sqrt{0.55 * 0.45/100} = 0.0497$
- Lower bound of CI: $\hat{p} - 1.96 * SE(\hat{p}) = 0.45 - 1.96 * 0.0497 = 0.35$
- Upper bound of CI: $\hat{p} + 1.96 * SE(\hat{p}) = 0.45 - 1.96 * 0.0497 = 0.55$
- 95% chance that this interval (0.35, 0.55) includes population proportion

Margin of Error

- What is the main reason for our uncertainty about the estimate of p?
- The spread in a confidence interval is called the margin of sampling error (ME)
- The margin of sampling error is $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- The more confident we want to be, the larger the margin of error must be. Why?
- Certainty vs Precision trade-off

Selecting Sample Size

For any desired margin of error (ME) we can choose a sample size:

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = ME$$

$$ME\sqrt{n} = z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}$$

$$\boxed{n = \left[\frac{z_{\alpha/2}}{ME}\right]^2 \hat{p}(1-\hat{p})}$$

Selecting Sample Size

Do we know $\hat{p}$ before we have collected information from the sample?

- Method 1: Use $\hat{p}$=0.5: This method gives sample size at least as big as you will need. Conservative approach.
- Method 2: Use $\hat{p}$ guess: This method gives you just the right sample size as long as your guess is correct. Efficient approach.

Mr Noxin - for the last time

- Mr Noxin would like to evaluate his chances of wining more precisely
- Specifically, he would like the margin of error in the estimate to be no more than 3%
- How many randomly selected voters do his friends need to ask about their favourite candidate?
- $n = \left[\frac{z_{\alpha/2}}{ME}\right]^2 \hat{p}(1-\hat{p}) = \frac{1.96^2}{0.03^2} * 0.5 * 0.5 = 1068$
- When computing the sample size, always round up!