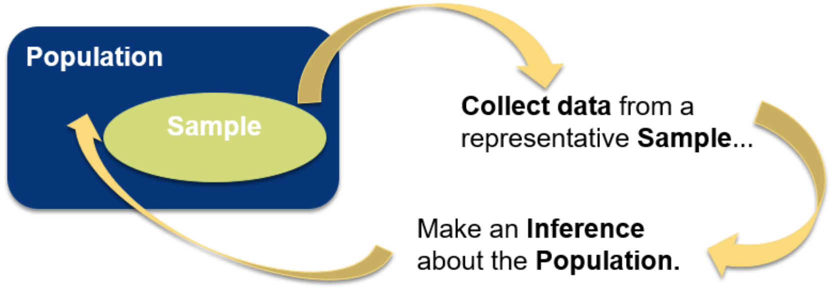
CHAPTER 5

Foundation for Inference

5.1 Point Estimates and sampling variability

Point Estimates and Error

- We are often interested in *population parameters*.



- Since complete populations are difficult (or impossible) to collect data on, we use *sample statistics* as *point estimates* for the unknown population parameters of interest.
  - Example: A poll suggested the US Presidents approval rating is 45%.

Point Estimate and Error

- When the parameter is a proportion, it is often denotated by p, and we refer to the sample proportion as $\hat{p}$.
- The population proportion (parameter) is a fixed number, but in practice, we do not know its value.

< Ex > Heart Transplant Operations at St. George’s Hospital

Data from St. George’s Hospital: 80% (8 out of 10 patients) died within 30 days of their heart transplant. What is the true mortality rate at the hospital?

Parameter: true mortality rate at the hospital $=p$

Statistic: mortality rate from the collected data at the hospital $=p=8/10=0.80$

Examples:

For each of the following situations , state whether the parameter of interest is a mean or a proportion.

- A poll shows that 64% of Americans personally worry a great deal about federal spending and the budget deficit.
- In a survey, smart phone users are asked whether or not they use a web-based taxi service.
- In a survey, smart phone users are asked how many times they used a web-based taxi service over the last year.

Point Estimates and Error

The difference we observe from the sample versus the parameter is called the **error** in the estimate. Generally the error consists of two aspects:

1. Sampling error (sampling uncertainty):
   - Describes how much an estimate will tend to vary from one sample to the next.

2. Bias:
   - Describe a systematic tendency to over- or under-estimate the true population value.

Understanding the Variability of a Point Estimate

Futures Company provides clients with research about maintaining and improving their business. They use a web interface to collect data from between 1000 and 2500 potential customers using 30- to 40- minute surveys. Assume 1650 out of 2500 potential customers in a sample show strong interest in a product.

The proportion of the sample ($\hat{p}$)who are interested is

$$\hat{p} = {}^{1650}\!/_{2500} = 0.66 = 66\%$$

The number $\hat{p} = 0.66$ is a statistic.

The corresponding parameter is the proportion (p) of all potential customers who would have expressed interest in this product had they been asked.

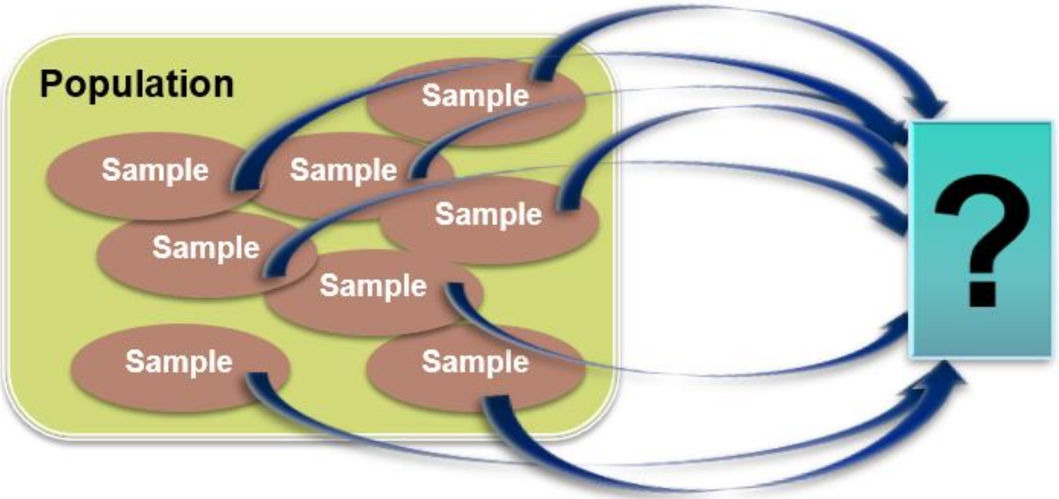
Understanding the Variability of a Point Estimate

Now suppose that if Futures Company took a second random sample of 2500 customers

The new sample would have different people in it.

It is almost certain that there would **not** be exactly 1650 positive responses. Hence, the value of the statistic $\hat{p}$ will vary from sample to sample.

We ask:“What would happen if we took many samples?”



Understanding the Variability of a Point Estimate

Statistical inference involves using information from a sample to draw conclusions about a wider population.

Statistical inference is based on one idea: “What would happen if we took many samples?”

Take a large number of samples from the same population.

Calculate the sample proportion $\hat{p}$ for each sample.

Make a histogram of the values of $\hat{p}$

Examine the distribution displayed in the histogram for shape, center, and spread, as well as outliers or other deviations.

However, in practice it is often too expensive to take too many samples from a large population – so we can imitate taking many samples with simulation.

One set of 16 tosses: 9 heads, 7 tails; p = 9/16 = 0.5625 = 56.25% success rate

Repeat the previous step many times and keep track of the results. The following is a dot plot of 10,000 different p values. This is an idea of a “Sampling Distribution” of a sample proportion.

The distribution is unimodal and roughly symmetric and it is centered around 0.50

Example

For the Future Company: We can simulate drawing SRSs of size 100 from the population of potential customers. Suppose that we know 60% of the population have interest in the product. Then the true value of the parameter we want to estimate is $p = 0.6$.

Note: Of course, we would not sample in practice if we already knew that $p = 0.6$. We are sampling to understand how sampling behaves.

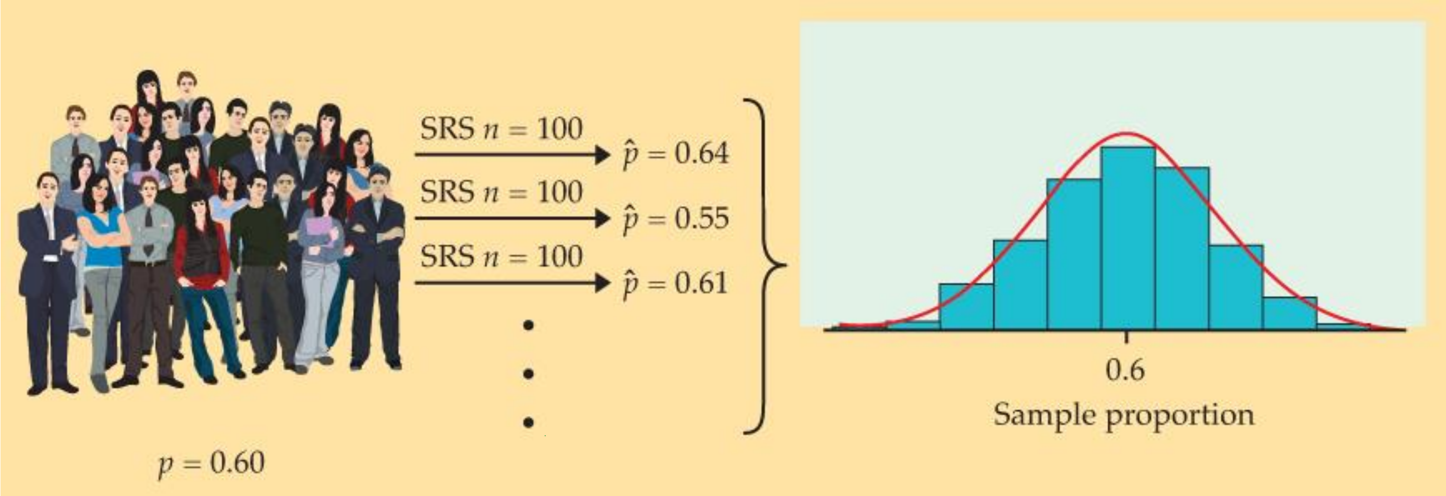
Example Cont. (For the Future Company)

Because all digits are equally likely, this assignment produces a population proportion p = 0.6.

Then simulate an SRS of 100 people from the population by taking 100 consecutive digits. The statistic $\hat{p}$ is the proportion of 0s to 5s in the sample.

we can use a computer to generate random numbers.

The figure below illustrates the process of choosing many samples and finding the sample proportion $\hat{p}$ for each one.

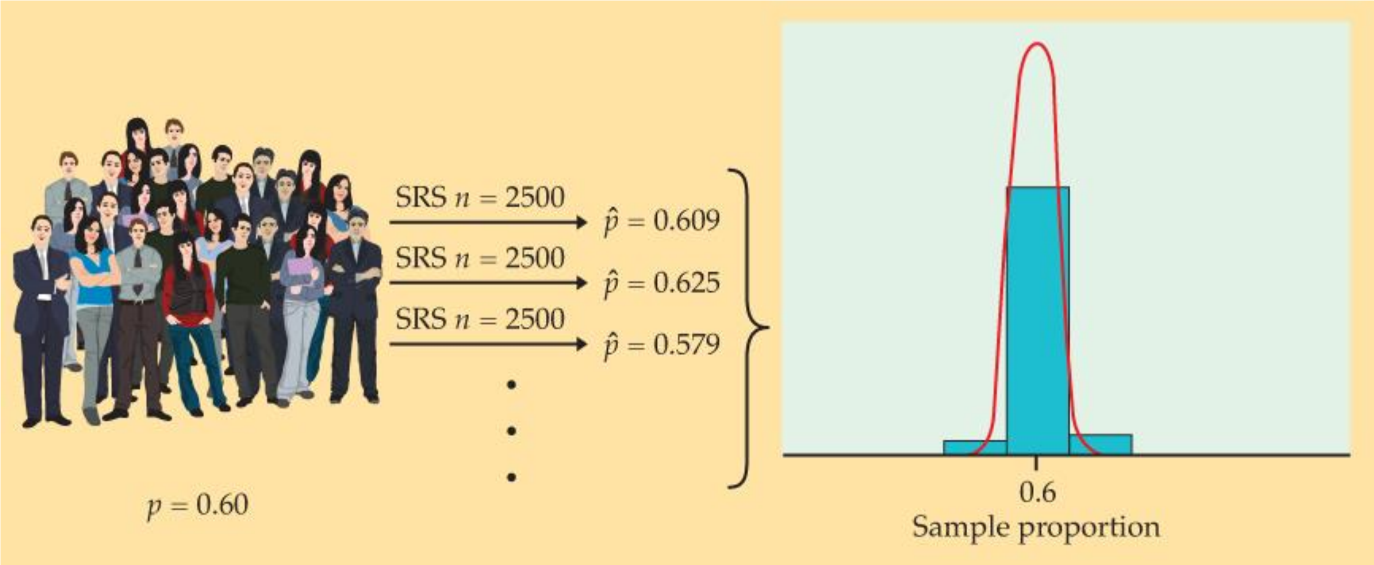


Example Cont.

But we know that Futures Company samples 2500 people, not just 100.

The figure below shows the process of choosing 1000 SRSs, each of size 2500, from a population in which the true proportion p = 0.6.

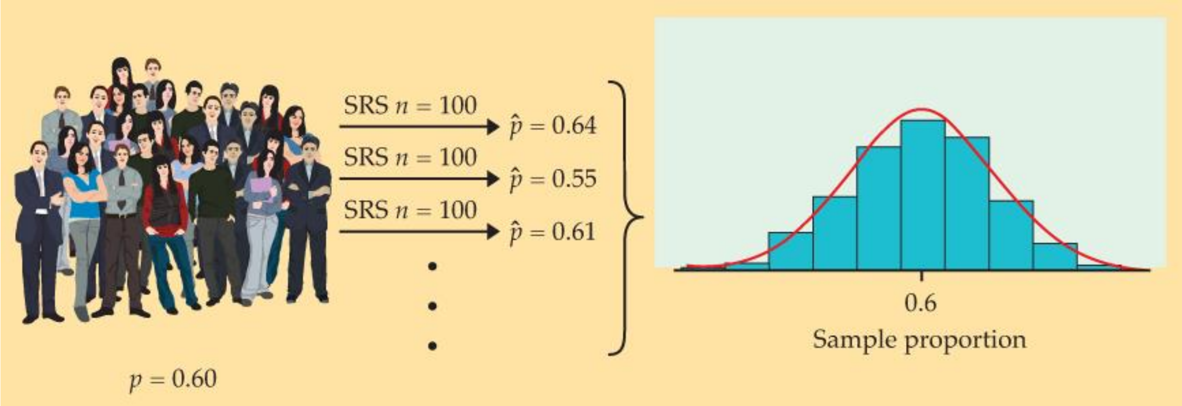
The 1000 values of $\hat{p}$ from these samples are given in the histogram below.

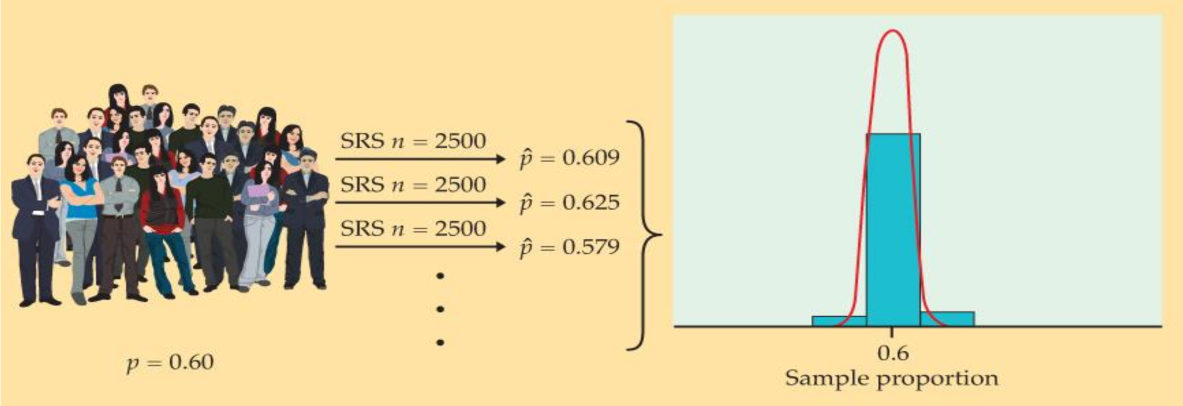


Example Cont.

We note that the previous figures are drawn on the same scale, so that we can see what happens when we increase the size of our sample from 100 to 2500.

These histograms display the sampling distribution of the statistic $\hat{p}$ for two sample sizes.





The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Central Limit Theorem

When observations are independent and the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with the following mean and standard error:

$$\mu_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

In order for the central limit theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$, which is called success- failure condition.

When p is Unknown

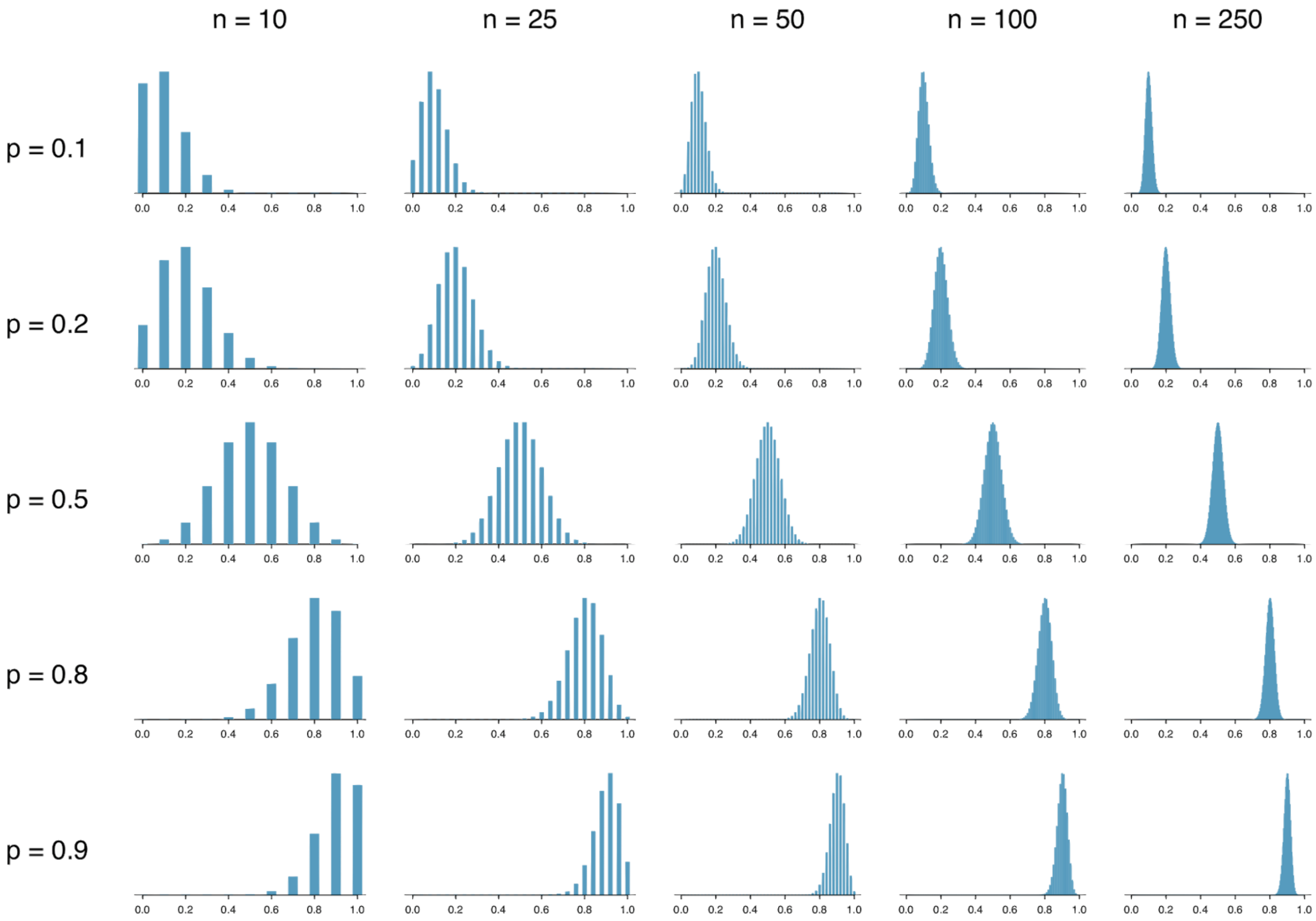
The CLT states

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

with the condition that np and $n(1 - p)$ are at least 10.

However, we often don't know the value of p, the population proportion. In these cases we substitute $\hat{p}$ for p.

What Happens when np and/or n(1-p) <10



Practice Problem

In a random sample 765 adult in the united states, 322 say they could not cover a $400 unexpected expense without borrowing money or going in debt.

a) What population is under consideration in the data set?

b) What parameter is being estimated?

c) What is the point estimate for the parameter?

d) What is the name of the statistic can we use to measure the uncertainty of the point estimate?

e) Compute the value from part (d) for this context.

f) A cable news pundit thinks the value is actually 50%. Should she be surprised by the data?

Suppose that 15% of all adults have hearing trouble.
We'll randomly sample 120 adults from the population. Let p = proportion of adults with hearing trouble in a sample of 120 adults

(a) Will the sampling distribution of p be approximate normal? Check the conditions.

(b) Determine the mean and standard error of p .

(c) Using the 68-95-99.7 rule, about 95% of p values will be between __________ and __________

(d) What is the probability that less than 12% of adults with hearing trouble in a sample? Use the normal distribution.

5.2 Confidence intervals for a proportion

Statistical Inference

Statistical inference is the process of making conclusion about the data. It also provides a statement of how much confidence we can place in our conclusion.

The two common types of statistical inference:

1) Confidence interval
2) Tests of significance

Confidence intervals

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic to estimate a parameter is like fishing in a lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish, but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Constructing a Confidence Interval

Confidence Intervals: estimate unknown population parameters by providing a set of plausible values and attaching a level of certainty.

The Big Idea: All confidence intervals we construct will have a form similar to this:

$$\boldsymbol{point\ estimate \pm z^* \times SE}$$

Confidence Interval for a Population Proportion

If a point estimate closely follow a normal model with standard error SE, then a confidence interval for the population parameter is

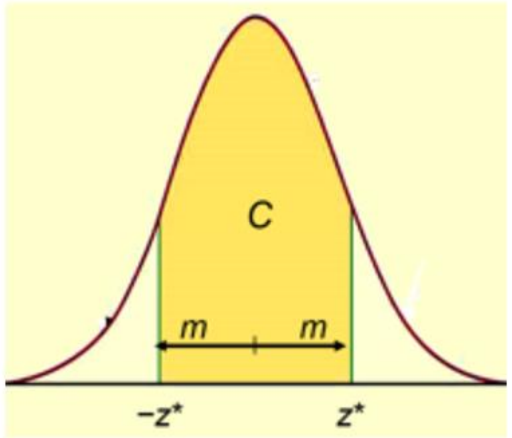
$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

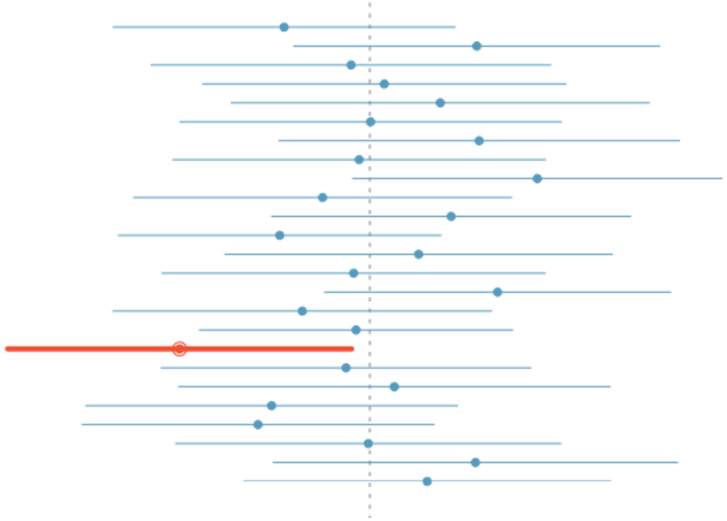
where z^* correspond to the confidence level selected.

Interpreting a Confident level

- The confidence level, denoted as C, is the overall capture rate if the method is used many times.

- The sample proportion will vary from sample to sample, but when we use the method "*estimate* $\pm$ *margin of error*" to get each interval, $C \times 100\%$ of all intervals capture the unknown population mean μ.





Changing the Confidence Level

$$\boldsymbol{point\ estimate \pm z^{*} \times SE}$$

- In a confidence interval, $z^{*} \times SE$ is called the *margin of error*, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust z^* in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, and 99%.
- For a 95% confidence interval, $z^* = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate z^* for any confidence level.

How to Find the Value of z^*

Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%. Find the z^* associated with each confidence level.

Confidence Interval for a Population Proportion

There are for steps to constructing the confidence interval:

1. Identify $\hat{p}$ and n, and determine what confidence level you wish to use.
2. Verify the conditions to ensure $\hat{p}$ is nearly normal. For one proportion confidence intervals, use $\hat{p}$ in place of p to check the success-failure condition.
3. If the conditions hold, compute SE using $\hat{p}$, find z^*, and construct the interval.
4. Interpret the confidence interval in the context of the problem.

Example

Pew Research asked a representative sample of 850 American Facebook users how accurately they feel the list of categories Facebook has listed for them on the page of their supposed interests actually represents them and their interests. 67% of the respondents said that the listed categories were accurate. Estimate the true proportion of American Facebook users who think Facebook categorizes their interests accurately. Compute a 95% confidence interval for the population proportion.

5.3 Hypothesis Testing for a Proportion

Statistical Inference

Two most common types of statistical inference:

***1. Confidence intervals*:**

- **Question we ask:** What is the value of the unknown parameter?
- **Goal:** To estimate a population parameter by giving a set of plausible values.

2. Tests of significance:

- **Question we ask:** Is the value of parameter different from some specific number?

  i.e. We are testing claims about a parameter.
- **Goal:** To assess evidence in the data about some claim concerning a population.

Tests of Significance

- A procedure for comparing *observed (actual) data* with a *claim* (also called a hypothesis) whose truth we want to assess.
- The **claim is a statement about a parameter**, like the population proportion p or the population mean μ.
- We express the results of a significance test in terms of a probability, called the P-value, that measures how well the data and the claim agree.

Test of Significance

We assume the simple conditions from Section 5.2 hold:

1. We have a perfect SRS (satisfies independence).

2. The expected numbers of success and failures are both at least 10 (Satisfies Normal approximation).

Four Steps of Tests of Significance

Tests of Significance: Four Steps

1. State the null and alternative **hypotheses.**
2. Calculate the value of the **test statistic.**
3. Find the ***P*-value** for the observed data**.**
4. State a **conclusion.**

Step 1: Stating Hypotheses

Significance test starts with a statement of the claims we want to compare.

Null hypothesis (H_0): often represent a claim to be. The null hypothesis is a statement of "no effect" or "no difference in the true parameter."

Alternative hypothesis (H_a): The claim about the population for which we're trying to find evidence.

- **One-sided alternative test:** If the parameter is either -
  1. *larger* than the null hypothesis value, or
  2. *smaller than* the null hypothesis value.

- **Two-sided alternative test:** If the parameter is *different* (not equal) *from* the null value.

Example

1. A college is trying a new student registration system and would like to know if there is sufficient evidence to conclude that more than 60% of the students favor the new system.

2. Ships arriving in the US ports are inspected by customs officials for contaminated cargo. Assume, for certain port, that 20% of the ships arriving in the previous year contained cargo that was contaminated. A random selection of 50 ships in the current year included five that had contaminated cargo. Does the data suggest that the proportion of ships arriving in the port with contaminated cargos decreased in the current year?

3. A production process will normally produce defective part 0.2% of the time. In a random sample of 1400 parts, three defectives are observed. Is there evidence to indicate that the defective rate of the process has changed?

Step 2: Test Statistic

Test statistic:

- Calculated from sample data
- Measures how far the data diverge from what we would expect if the null hypothesis H_0 were true.

$$z = \frac{estimate\ - hypothesisized\ value}{standard\ deviation\ of\ the\ estimate} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Large values of the test statistic show that the data are not consistent with H_0.

Example 1 Cont.

A college is trying a new student registration system and would like to know if there is sufficient evidence to conclude that more than 60% of the students favor the new system. In a random sample of 520 students, 352 said that the new registration system is superior. calculate the test statistic.

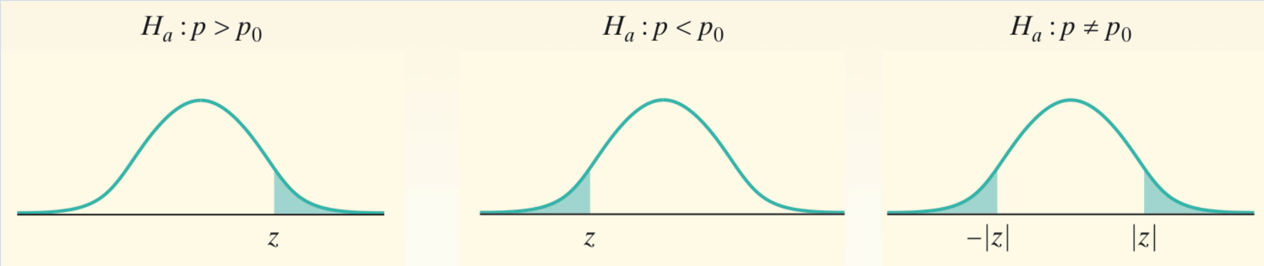
Step 3: P-Value

***P*-value:**

We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.

- Small *P*-values, the stronger the evidence against H_0.
- Large *P*-values fail to give convincing evidence against H_0.

Find the *P*-value by calculating the probability of getting a *z* statistic this large or larger in the direction specified by the alternative hypothesis H_a:



Example 1 Cont.

A college is trying a new student registration system and would like to know if there is sufficient evidence to conclude that more than 60% of the students favor the new system. In a random sample of 520 students, 352 said that the new registration system is superior. Calculate the P-value.

Step 4: Conclusion

There are only two decisions based on the strength of the evidence against the null hypothesis (and in favor of the alternative hypothesis).

Compare the P-value with a fixed value that we regard as decisive, called the **significance level.** We write it as α, the Greek letter alpha.

$$P\text{-value} < \alpha \rightarrow \text{reject } H_0 \rightarrow \text{conclude } H_a \text{ (in context)}$$

$$P\text{-value} \geq \alpha \rightarrow \text{fail to reject } H_0 \rightarrow \text{cannot conclude } H_a \text{ (in context)}$$

If the P-value is smaller than α, we say that the data are **statistically significant at level α.**

Note: A fail-to-reject H_0 decision in a significance test does not mean that H_0 is true. For that reason, you should never "accept H_0" or use language implying that you believe H_0 is true.

Example 1 Cont.

A college is trying a new student registration system and would like to know if there is sufficient evidence to conclude that more than 60% of the students favor the new system. State your conclusion based on your test statistic and p-value.

Decision Errors

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Truth about the population	
		H_0 true	H_0 false (H_a true)
Conclusion based on sample	Reject H_0	**Type I error (α)**	*Correct conclusion*
	Fail to reject H_0	*Correct conclusion*	**Type II error (β)**

If we reject H_0 when H_0 is true, we have committed a **Type I error.**

If we fail to reject H_0 when H_0 is false, we have committed a **Type II error.**

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial, then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty.
- Declaring the defendant guilty when they are actually innocent.

Which error do you think is the worse error to make?

"better that ten guilty persons escape than that one innocent suffer"
- William Blackstone

Practice Problem 1

A study suggests that 60% of college student spend 10 or more hours per week communicating with others online. You believe that this is incorrect and decide to collect your own sample for a hypothesis test. You randomly sample 160 students from your dorm and find that 70% spent 10 or more hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$$H_0 = \hat{p} < 0.6$$
$$H_A = \hat{p} > 0.7$$

Practice Problem 2

A potato-chip producer has just received a truckload of potatoes from its main supplier. If the producer determines that more than 8% of the potatoes in the shipment have blemishes, the truck will be sent away to get another load from the supplier. A supervisor selects a random sample of 500 potatoes from the truck. An inspection reveals that 47 of the potatoes have blemishes. Carry out a significance test at the $\alpha = 0.10$ significance level. What should the producer conclude?

Practice Problem 3

A food safety inspector is called upon to investigate a restaurant with a few customer reports of poor sanitation practices. The food safety inspector use a hypothesis testing framework to evaluate whether regulations are not being met. If he decide the restaurant is in gross violation, its license to serve food will be revoked.

a) Write the hypotheses in words.

b) What is Type I Error in this context?

c) What is Type II Error in this context?

Practice Problem 4

For Americans using library services, the American Library Association (ALA) claims that 67% borrow books. A library director feels that this is not true so she randomly selected 100 borrowers and finds that 82 borrowed books. Can she show that ALA claim is incorrect?

Practice Problem 5

Ships arriving in the US ports are inspected by customs officials for contaminated cargo. Assume, for certain port, that 20% of the ships arriving in the previous year contained cargo that was contaminated. A random selection of 50 ships in the current year included five that had contaminated cargo. Does the data suggest that the proportion of ships arriving in the port with contaminated cargos decreased in the current year at $\alpha = 0.01$?