Lecture 19

Rejecting Chance

Researchers are interested in answering direct questions often conduct *hypothesis test*.

We have already learned the basic question researchers ask when they conduct such a test:

Is the relationship observed in the sample large enough to be called statistically significant, or could it have been due to chance?

Thought Question 1:

In the courtroom, juries must make a **decision about the guilt or innocence of a defendant**. Suppose you are on the jury in a murder trial.

It is obviously a **mistake if the jury claims the suspect is guilty when in fact he or she is innocent**.

What is the **other type of mistake** the jury could make? Which is **more serious**?

Thought Question 2:

Suppose half (0.50) of a population would answer yes when asked if support the death penalty.

Random sample of 400 results in 220, or 0.55, who answer yes. Rule for Sample Proportions => potential sample proportions are approximately bell-shaped, with standard deviation of 0.025.

Find standardized score for observed value of 0.55. **How often** you would **expect to see** a standardized score **at least that large or larger**?

Thought Question 3:

Want to test a claim about the proportion of a population who have a certain trait.

Collect data and discover that *if* claim true, the **sample proportion you observed is so large that it falls at 99th percentile** of possible sample proportions for your sample size.

Would you believe claim and conclude that you just happened to get a weird sample, or would you reject the claim? What if result was at 70th percentile? At 99.99th percentile?

Thought Question 4:

Which is generally more serious when getting results of a medical diagnostic test:

a **false positive**, which tells you you have the disease when you don’t,

or a **false negative**, which tells you you do not have the disease when you do?

22.1 Using Data to Make Decisions

Examining Confidence Intervals:

Used CI to make decision about whether there was difference between two conditions by seeing if 0 was in interval or not.

Hypothesis Tests:

Is the relationship observed in sample large enough to be called statistically significant, or could it have been due to chance?

Example 1: Quarters or Semesters?

University currently on quarter system but may switch to semesters. Heard students may oppose semesters and want to test if a majority of students would oppose the switch.

Administrators must choose from **two hypotheses**:

1. There is **no clear preference** (or the switch is preferred), so there is no problem.
2. As rumored, a **majority of students oppose** the switch, so the administrators should reconsider their plan.

In **random sample of 400 students**, 220 or **55% oppose** switch. A clear majority of the sample are opposed.

If really no clear preference, how likely to observe sample results of this magnitude (55%) or larger, just by chance?

Example 1: Quarters or Semesters?

If no clear preference, rule for sample proportions …
If numerous samples of size 400 are taken, the frequency curve for the proportions from various samples will be **approximately bell-shaped**. **Mean** will be 0.50 and **standard deviation** will be:

$$\sqrt{\frac{(0.50)(1 - 0.50)}{400}} = 0.025.$$

Standardized score = z-score = (0.55 – 0.50)/0.025 = 2.00

Table 8.1: z-score of 2.00 falls between 1.96 and 2.05, the 97.5th and 98th percentiles. **If truly no preference, then we would observe a sample proportion as high as 55% (or higher) between 2% and 2.5% of the time.**

Example 1: Quarters or Semesters?

One of two things has happened:

1. **Really is no clear preference**, but by "luck" this sample resulted in an unusually high proportion opposed. So high that chance would lead to such a high value only about 2% of time.
2. **Really is a preference** against switching to the semester system. The proportion (of all students) against the switch is actually higher than 0.50.

Most researchers agree to rule out chance if "luck" would have produced such extreme results **less than 5% of the time**.

Conclude: proportion of students opposed to switching to semesters is ***statistically significantly higher than 50%***.

22.2 Basic Steps for Testing Hypotheses

1. Determine the **null** hypothesis and the **alternative** hypothesis.
2. Collect **data** and summarize with a single number called a **test statistic**.
3. Determine how **unlikely** test statistic would be ***if null hypothesis were true***.
4. Make a **decision**.

Step 1. Determine the hypotheses.

- **Null hypothesis**—hypothesis that says nothing is happening, status quo, no relationship, chance only.
- **Alternative (research) hypothesis** — hypothesis is reason data being collected; researcher suspects status quo belief is incorrect or that there is a relationship between two variables that has not been established before.

Common Practice about Hypothesis

- Null hypothesis is a status quo: no effects, no difference, previously accepted theory
- Alternative hypothesis is investigator’s suspect or belief
- The burden of proof is on the investigator to convince reviewers to abandon the null hypothesis.
- All statements are about population parameter, nothing about sample statistics shall appear in hypothesis.

Example 2: A Jury Trial

If on a jury, must presume defendant is innocent unless enough evidence to conclude is guilty.

Null hypothesis:	Defendant is ***innocent***.
Alternative hypothesis:	Defendant is ***guilty***.

- Trial held because prosecution believes status quo of innocence is incorrect.
- Prosecution collects evidence, like researchers collect data, in hope that jurors will be convinced that such evidence is extremely unlikely if the assumption of innocence were true.

Step 2. Collect data and summarize with a test statistic.

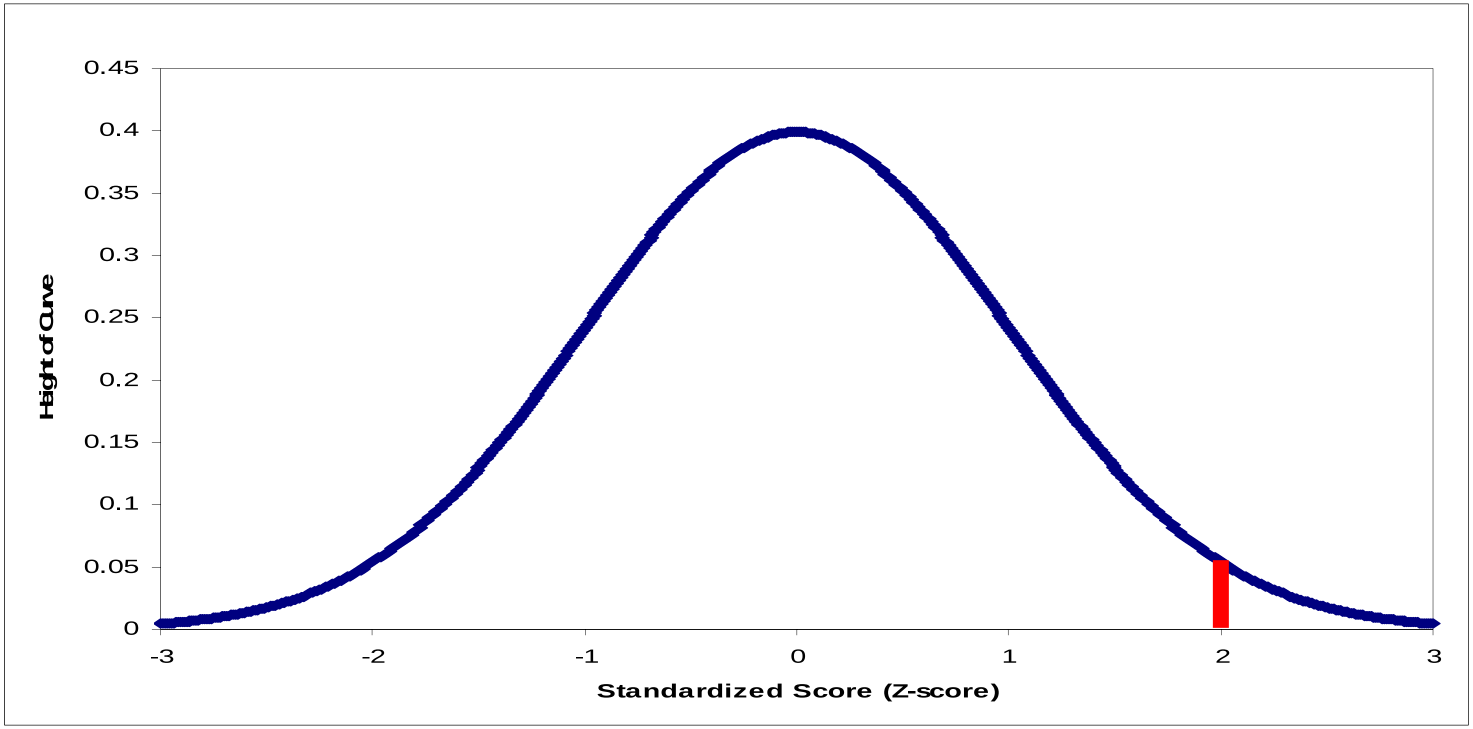
Decision in hypothesis test based on single summary of data – the **test statistic.**
e.g. ***chi-square*** test statistic and ***standard score***.

Step 3. Determine how unlikely test statistic would be if null hypothesis true.

If null hypothesis true, how likely to observe sample results of this magnitude or larger (in direction of the alternative) just by chance? **… called *p*-value.**

We found a z-score of +2.00 which falls between the 97.5th and 98th percentiles.

That is, if there is truly no preference, then we would expect to observe a sample proportion as high as this (or higher) between 2% and 2.5% of the time.



Test statistic

- A test is based on a statistic, which estimates the parameter that appears in the hypotheses
  - Point estimate
- Values of the estimate far from the parameter value in H_0 give evidence against H_0.
- H_a determines which direction will be counted as “far from the parameter value”.

Step 4. Make a Decision.

Choice 1: *p*-value not small enough to convincingly rule out chance. We **cannot reject the null** hypothesis as an explanation for the results. There is **no statistically significant difference** or relationship evidenced by the data.

Choice 2: *p*-value small enough to convincingly rule out chance. We **reject the null hypothesis** and *accept* the alternative hypothesis. There is a **statistically significant difference** or relationship evidenced by the data.

How small is small enough?

Standard is **5%,** also called **level of significance**.

One of two things has happened:

1. There really is no clear preference, but by “luck of the draw” this particular sample resulted in an unusually high proportion opposed to the switch. In fact, it is so high that chance would lead us to such a high value only slightly more than 2% of the time.

2. There really is a preference against switching to the semester system. The proportion (of all students) against the switch is actually higher than 0.50.

The p-value is computed by assuming that the null hypothesis is true, and then asking how likely we would be to observe such extreme results (or even more extreme results) under that assumption.

Describing *P-value*

- If the *P-value* is less than 1%, there is ***overwhelming evidence*** that supports the alternative hypothesis.
- If the *P-value* is between 1% and 5%, there is ***strong evidence*** that supports the alternative hypothesis.
- If the *P-value* is between 5% and 10% there is ***weak evidence*** that supports the alternative hypothesis.
- If the *P-value* exceeds 10%, there is ***no evidence*** that supports of the alternative hypothesis.

Significance Level α

- We need to make a conclusion after carrying out the hypothesis test. *What do we conclude?*
- We can compare the *P-value* with a fixed value that we regard as decisive.
- This amounts to announcing in advance how much evidence against H_0 we require in order to reject H_0.
- The decisive value is called the *significance level* of the test. It is denoted by α and the corresponding test is called a *level* α *test*.

Statistical Significance: If the *P-value* $\leq \alpha$, we say that the data are statistically significant at level α.

Most researchers agree, that by convention, we can rule out change if the “luck of the draw” would have produced such extreme results less than 5% of the time.

Therefore, in this case, the administrators should probably decide to rule out chance.

The proper conclusion is that, indeed, a majority is opposed to switching to the semester system.

When a relationship or value from a sample is so strong that we can effectively rule out chance this way, we say that the result is statistically significant.

In this case, we could say that the proportion of students who are opposed to switching to the semester system is statistically significantly higher than 50%.

Once we know how unlikely the results would have been if the null hypothesis were true, we must make one of two choices:

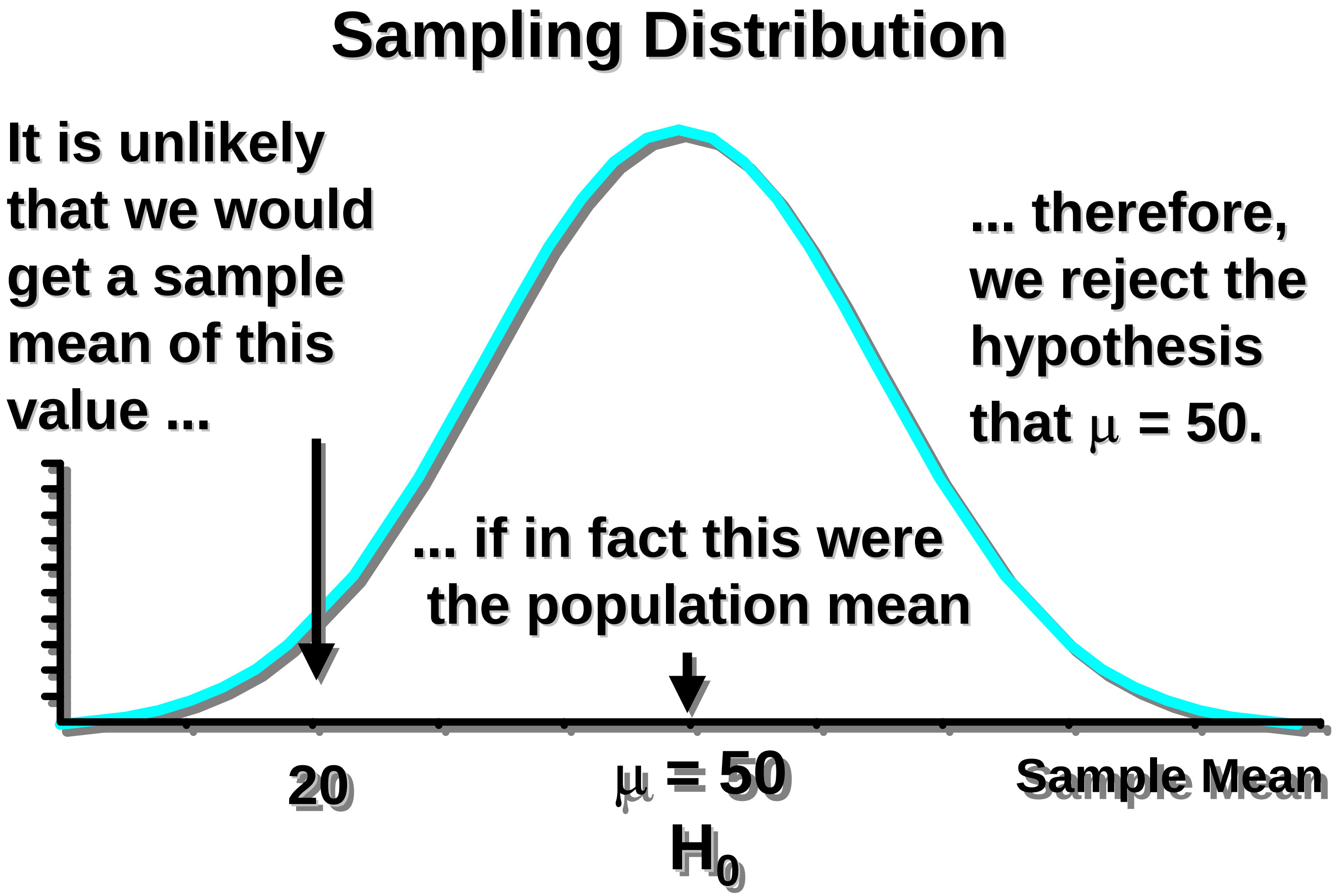
1. The p-value is not small enough to convincingly rule out change. Therefore we *cannot reject* the null hypothesis as an explanation for these results.

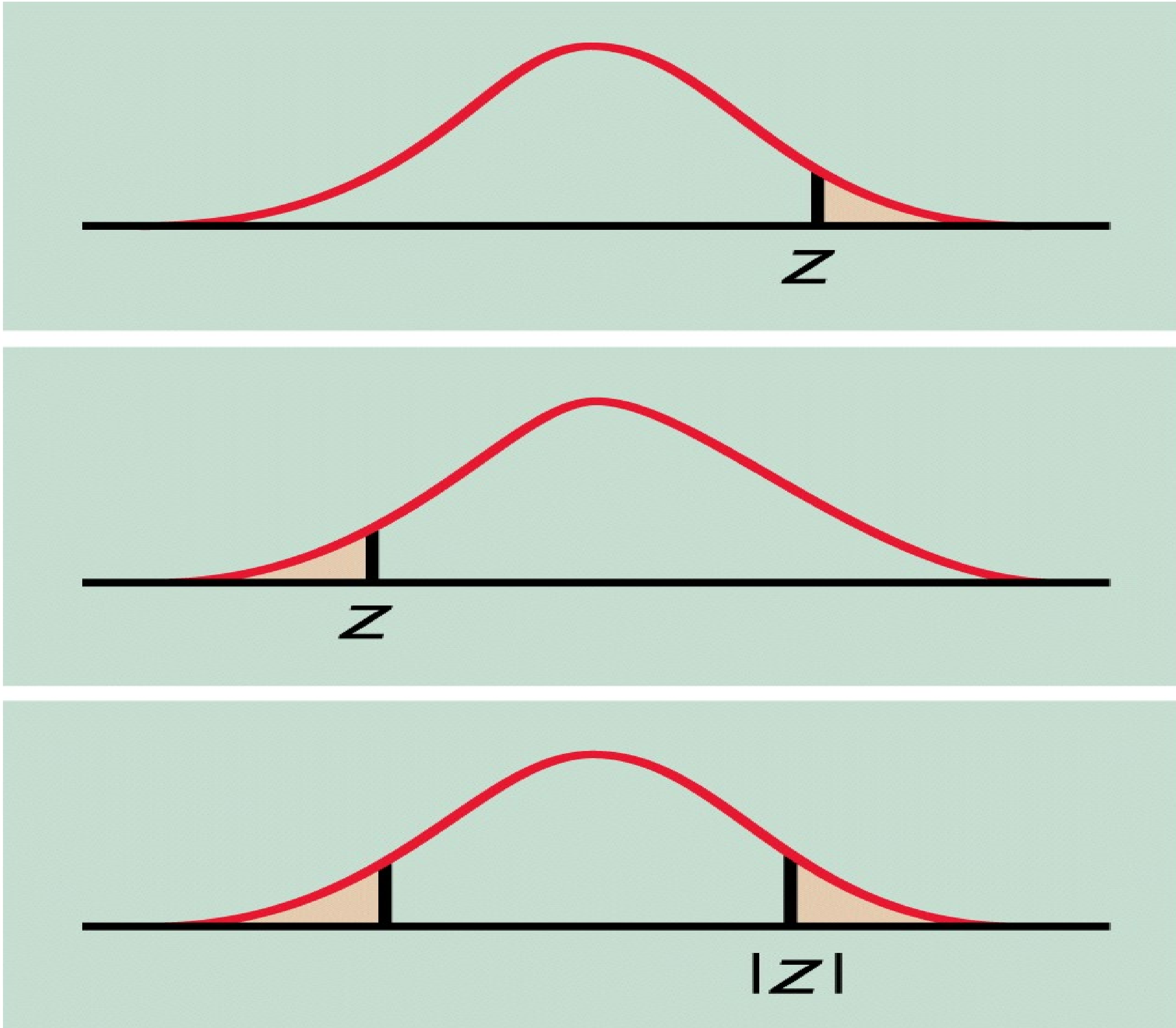
2. The p-value was small enough to convincingly rule out chance. We *reject* the null hypothesis and *accept* the alternative hypothesis.

You may be wondering how small the p-value must be in order to be small enough to rule out the null hypothesis. The standard used by most researchers is 5%.

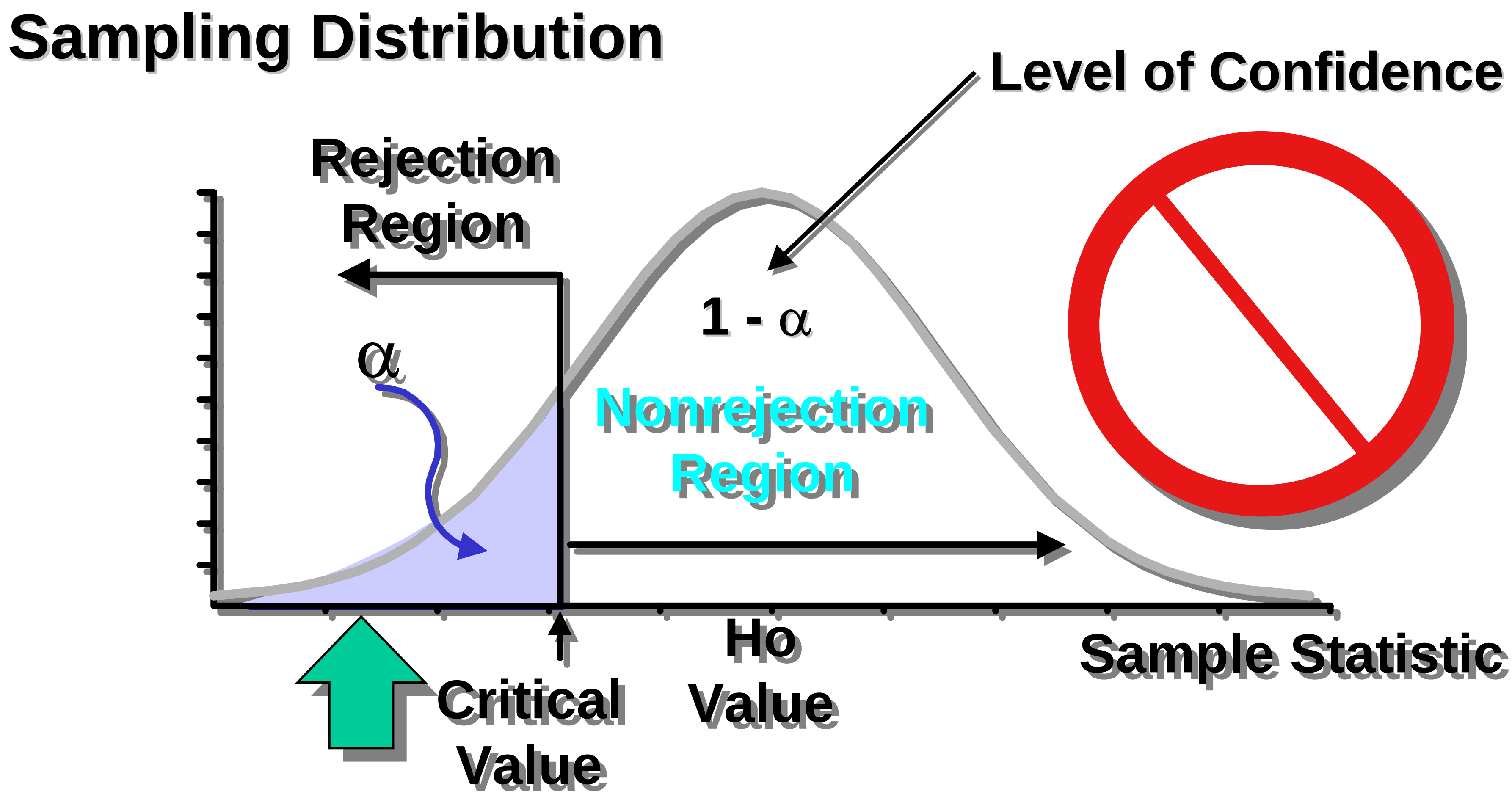
NOTE: We will state our result by saying that we reject the null hypothesis, or that we fail to reject the null hypothesis. So, even if we don't have convincing evidence to reject the null hypothesis, we **don't** say that we accept the null hypothesis. With a larger sample, we may be able to reject the null hypothesis.

One and Two Tailed Tests

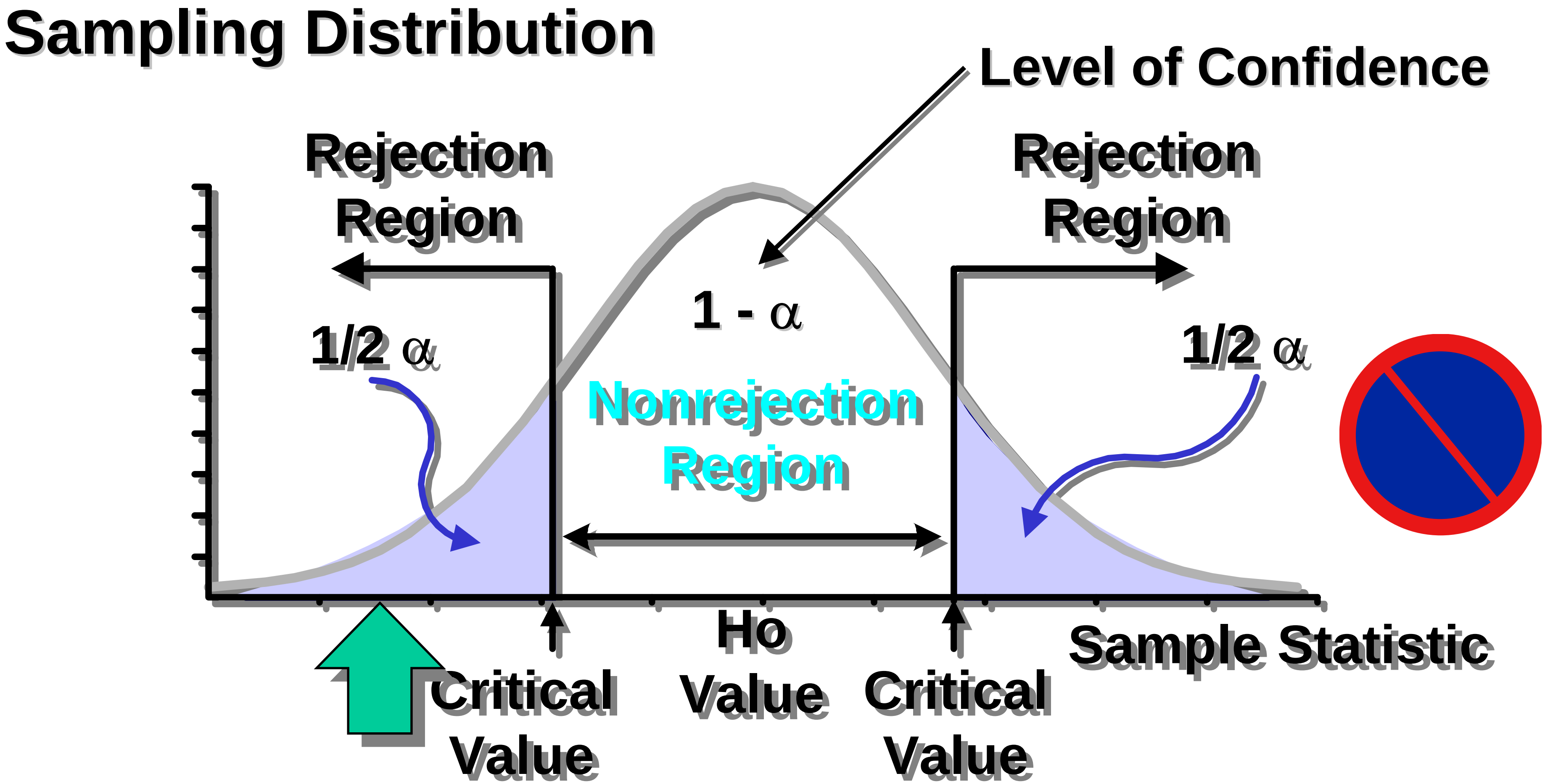



z
z
|z|

Rejection Region
(One-Tail Test)



Rejection Regions
(Two-Tailed Test)



22.3 Testing Hypotheses for Proportions

Step 1. Determine the null and alternative hypotheses.

Null hypothesis: The population proportion of interest **equals** the null value.

The alternative hypothesis is one of the following:

Alternative hypothesis: The population proportion of interest is **not equal to** the null value. [A *two-sided* hypothesis.]

Alternative hypothesis: The population proportion of interest is **greater than** the null value. [A *one-sided* hypothesis]

Alternative hypothesis: The population proportion of interest is **less than** the null value. [A *one-sided* hypothesis]

Step 2. Collect data and summarize with a test statistic.

The **test statistic** is a **standardized score.** It measures how far away the sample proportion is from the null value in standard deviation units.

Test statistic = *standardized score*

$$= z\text{-}score = \frac{\text{sample proportion} - \text{null value}}{\text{standard deviation}}$$

$$\text{where standard deviation} = \sqrt{\frac{(\text{null value}) \times (1 - \text{null value})}{\text{sample size}}}$$

Step 3. Determine how unlikely test statistic would be if null hypothesis true.

***p*-value** = probability of observing a standardized score as extreme or more extreme (in the direction specified in the alternative hypothesis) if the null hypothesis is true.

Alternative Hypothesis	*p*-value = proportion of bell-shaped curve:
Proportion is greater than null value	above the *z*-score test statistic value
Proportion is less than null value	below the *z*-score test statistic value
Proportion is not equal to null value	[above the absolute value of test statistic] × 2

Step 4. Make a decision.

Researcher **compares the *p*-value** to a (pre-)specified **level of significance**. Most common level of significance is 0.05.

If the ***p*-value is greater than the level of significance**:

- Do not reject the null hypothesis
- The true population proportion is not significantly different from the null value

If the ***p*-value is less than or equal to the level of significance**:

- Reject the null hypothesis
- Accept the alternative hypothesis
- The true population proportion is significantly different from the null value (*different* according to the alternative direction)

Example 1: Quarters or Semesters? *continued*

Step 1. Determine the null and alternative hypotheses.

Null hypothesis: The proportion of students at the university who oppose switching to semesters is 0.50.

Alternative hypothesis: The proportion of students at university who oppose switching to semesters is greater than 0.50.

Step 2. Collect data and summarize with test statistic.

In a **random sample of 400 students**, 220 or **55% oppose**.

So the *standard deviation* = $\sqrt{\dfrac{(0.50) \times (1 - 0.50)}{400}} = 0.025$.

Test statistic: $z = \dfrac{0.55 - 0.50}{0.025} = 2.00$

Example 1: Quarters or Semesters? *revisited*

Step 3. Determine the *p*-value.

Recall the alternative hypothesis: The proportion of students at university who oppose switching is **greater than** 0.50.

So *p*-value = proportion of bell-shaped curve *above* 2.00.

Table 8.1 => proportion is between 0.02 and 0.025.

Using computer/calculator: exact *p*-value = 0.0228.

Step 4. Make a decision.

The ***p*-value is less than or equal to 0.05**, so we conclude:

- Reject the null hypothesis
- Accept the alternative hypothesis
- The true population proportion opposing the switch to semesters is significantly *greater* than 0.50.

Example 3: Family Structure in Teen Survey

Government reports 67% of teens live with both parents but survey gave 84% => does survey population *differ*?

Step 1. Determine the null and alternative hypotheses.

Null hypothesis: For the population of teens represented by the survey, the proportion living with both parents is 0.67.

Alternative hypothesis: For population of teens represented by survey, proportion living with both parents is not equal to 0.67.

Step 2. Collect data and summarize with test statistic.

Survey of 1,987 teens => 84% living with both parents.

So the *standard deviation* = $\sqrt{\frac{(0.67) \times (1 - 0.67)}{1987}}$ = 0.0105.

Test statistic: $z = \frac{0.84 - 0.67}{0.0105} = 16$ *(extremely large!)*

Example 3: Family Structure in Teen Survey

Step 3. Determine the *p*-value.

Recall alternative hypothesis was two-sided.

So *p*-value = 2 × [proportion of bell-shaped curve above 16].
Table 8.1 => proportion is essentially 0.
Almost impossible to observe a sample of 1,987 teens with 84% living with both parents if only 67% of population do.

Step 4. Make a decision.

The ***p*-value is essentially 0**, so we conclude:

- Reject the null hypothesis
- Accept the alternative hypothesis
- The proportion of teens living with both parents in population represented by survey is significantly different from population of teens living with both parents in U.S.

22.4 What Can Go Wrong: The Two Types of Errors

Courtroom Analogy: ***Potential choices and errors***

Choice 1: We cannot rule out that defendant is innocent,
so he or she is set free without penalty.
Potential error: A criminal has been erroneously freed.

Choice 2: We believe enough evidence to conclude
the defendant is guilty.
Potential error: An innocent person falsely convicted
and guilty party remains free.

Choice 2 is usually seen as ***more serious***.

Decision Results

H_0: Innocent

Jury Trial			H_0 Test		
Verdict	Actual Situation		Decision	Actual Situation	
	Innocent	Guilty		H_0 True	H_0 False
Innocent	Correct	Error	Accept H_0	$1 - \alpha$	Type II Error (β)
Guilty	Error	Correct	Reject H_0	Type I Error (α)	Power ($1 - \beta$)

Medical Analogy: *False Positive vs False Negative*

Tested for a disease; most tests not 100% accurate.

Null hypothesis: You do not have the disease.
Alternative hypothesis: You have the disease.

Choice 1: Medical practitioner thinks you are healthy.
Test result weak enough to be "negative" for disease.
Potential error: You have disease but told you do not.
Your test was a **false negative**.

Choice 2: Medical practitioner thinks you have disease.
Test result strong enough to be "positive" for disease.
Potential error: You are healthy but told you're diseased.
Your test was a **false positive**.

Which is ***more serious***? Depends on disease and consequences.

The Two Types of Errors in Testing

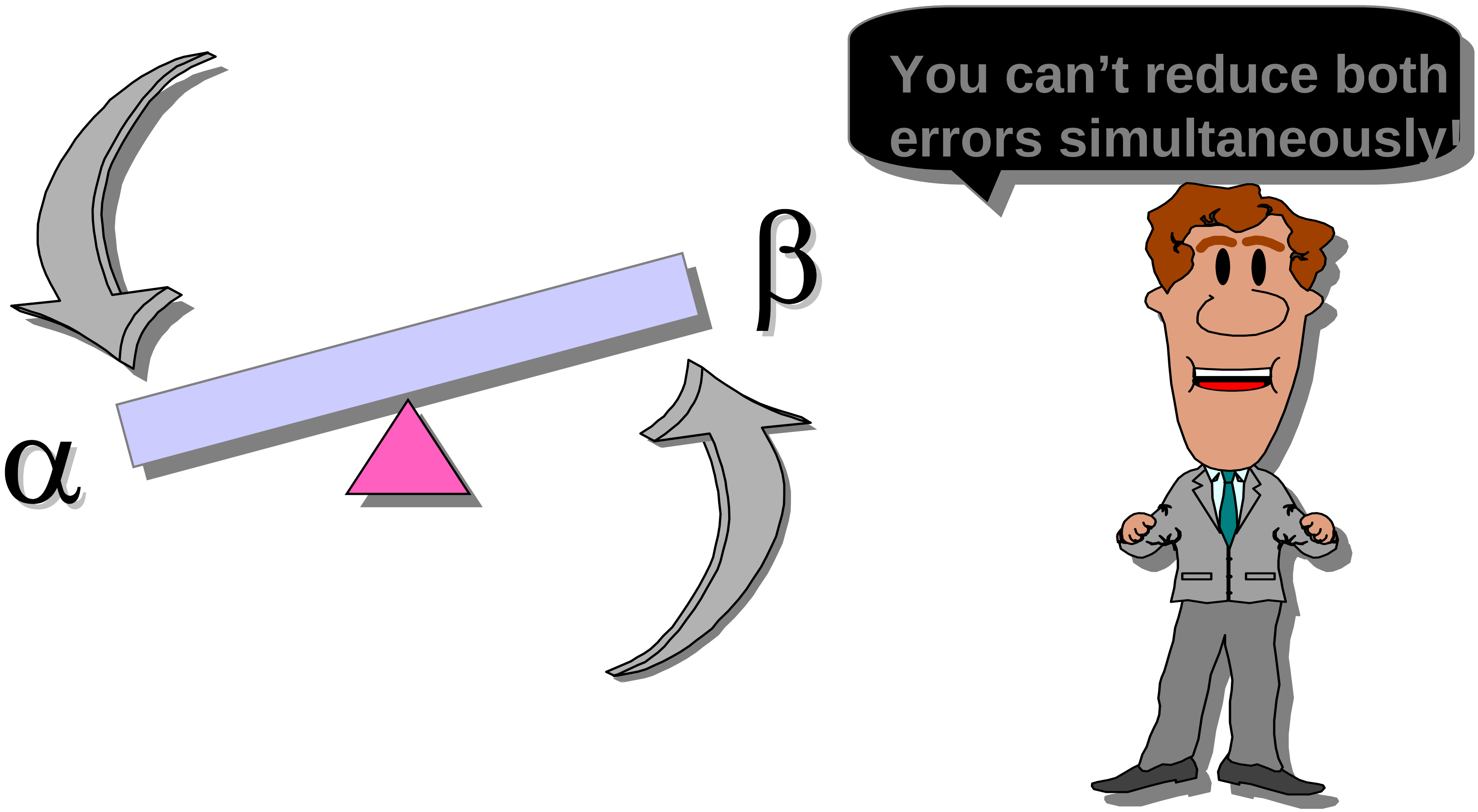
	True State of Nature	
Decision Made	Innocent, Healthy Null Hypothesis	Guilty, Diseased Alternative Hypothesis
Not guilty Healthy Don't reject null hypothesis	Correct ☺	Undeserved freedom False negative Type 2 error
Guilty Diseased Accept alternative hypothesis	Undeserved punishment False positive Type 1 error	Correct ☺

- **Type 1 error** can only be made if the null hypothesis is actually true.
- **Type 2 error** can only be made if the alternative hypothesis is actually true.

We are very concerned about wrongly reject H_0.

- Oliver Wendell Holmes (U.S. Supreme Court Justice: “Better to acquit 100 guilty men than convict one innocent one.”
- The justice procedure begins with the assumption that the defendant is innocent (null hypothesis is true)
- We review the evidence to find if there is enough evidence against H_0, i.e., support H_a.

α & β Have an Inverse Relationship



Probabilities Associated with Errors

We can only specify the conditional probability of making a type 1 error, given that the null hypothesis is true. That probability is called the **level of significance**, usually 0.05.

Level of Significance and Type I Errors

If null hypothesis is true, probability of making a type 1 error is equal to the level of significance, usually 0.05.
If null hypothesis is not true, a type 1 error cannot be made.

Type 2 Errors

A type 2 error is made if the alternative hypothesis is true, but you fail to choose it. The probability of doing that depends on which part of the alternative hypothesis is true, so computing the probability of making a type 2 error is not feasible.

Probabilities Associated with Errors

The Power of a Test

The **power** of a test is the probability of making the correct decision when the alternative hypothesis is true. If the population value falls close to the value specified in null hypothesis, then it is difficult to get enough evidence from the sample to conclusively choose the alternative hypothesis.

When to Reject the Null Hypothesis

In deciding whether to reject the null hypothesis consider the consequences of the two potential types of errors.

- If consequences of a type 1 error are very serious, then only reject null hypothesis if the *p*-value is very small.
- If type 2 error more serious, should be willing to reject null hypothesis with a moderately large *p*-value, 0.05 to 0.10.

Case Study 22.1: Testing for ESP

Description of the Experiments

Setup called *ganzfeld procedure*

- Two participants: one a sender, other a receiver.
- Two researchers: one the experimenter, other the assistant.
- Sender in one room focuses on either still picture (*static target*) or short video (*dynamic target*).
- Receiver in different room, white noise through headphones, looking at red light, with microphone to give continuous monologue about images/thoughts present in mind.
- Experimenter monitors procedure and listens to monologue.
- Assistant uses computer to randomly select the target.
- About 160 targets, half static and half dynamic.

Case Study 22.1: Testing for ESP

Quantifying the Results

To provide a comparison to chance …

- Three ***decoy*** targets are chosen from the set of the same type as the real target (static or dynamic).
- Any of the four targets (real and three decoys) could equally have been chosen to be the real target.
- Receiver shown the four potential targets and asked to decide which one the sender was watching.
- If receiver picks correct one = success.

Case Study 22.1: Testing for ESP

The Null and Alternative Hypotheses

Null hypothesis: Results due to chance guessing so probability of success is 0.25.

Alternative hypothesis: Results not due to chance guessing, so probability of success is higher than 0.25.

The Results

Sample proportion of successes: 122/355 = 0.344.

The *standard deviation* = $\sqrt{\dfrac{(0.25) \times (1 - 0.25)}{355}}$ = 0.023.

Test statistic: $z = (0.344 - 0.25)/0.023 = 4.09$

The *p*-value is about 0.00005. If chance alone were operating, we'd see results of this magnitude about 5 times in every 100,000 such experiments => a ***statistically significant result.***