Hypothesis Tests
Part 1: Formulation

DASC 512

Overview

- What is a hypothesis test?
- Defining the hypotheses
- Choosing a test statistic
- Setting the significance level
- Defining the rejection region
- *p*-values: what are they good for?

What is a Hypothesis Test?

A hypothesis test or significance test uses data to summarize the evidence about a hypothesis by comparing a point estimate of the parameter of interest to the value predicted by the hypothesis.

So we have the following elements:

- Hypothesis (null and alternative)
- Point estimate of parameter (test statistic)
- Predicted value of parameter (sampling distribution)
- Conclusion

We also must consider assumptions under which a test is valid.

Assumptions

Each type of hypothesis test requires varying assumptions pertaining to:

- Type of data: numerical or categorical
- Distribution: depends on the hypothesis test
- Sample size: must be sufficient to justify distributional assumption

And of course, they always assume independent and identically distributed (iid) data collection

To build a Hypothesis Test

1. Define the hypotheses (null and alternative)
   - Parameter of interest
   - Left-, right-, or two-tailed
2. Choose the significance level (α)
3. Choose a test statistic
4. Define the rejection region
   - This may be in terms of test statistic or *p*-value
5. Calculate the test statistic and/or *p*-value
6. Make a conclusion

Hypotheses – Null hypothesis

Hypothesis are formulated before analyzing the data.

The null hypothesis (H_0) is a statement that a parameter takes a particular value. This will be assumed to be true, but it cannot be proved to be true.

$$H_0: \mu = \mu_0$$

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_0: \pi = \pi_0$$

Hypotheses – Alternative hypothesis

The alternative hypothesis (H_a) states that a parameter falls in another range. This is usually a research hypothesis the investigator believes to be true. Data is collected to attempt to support this.

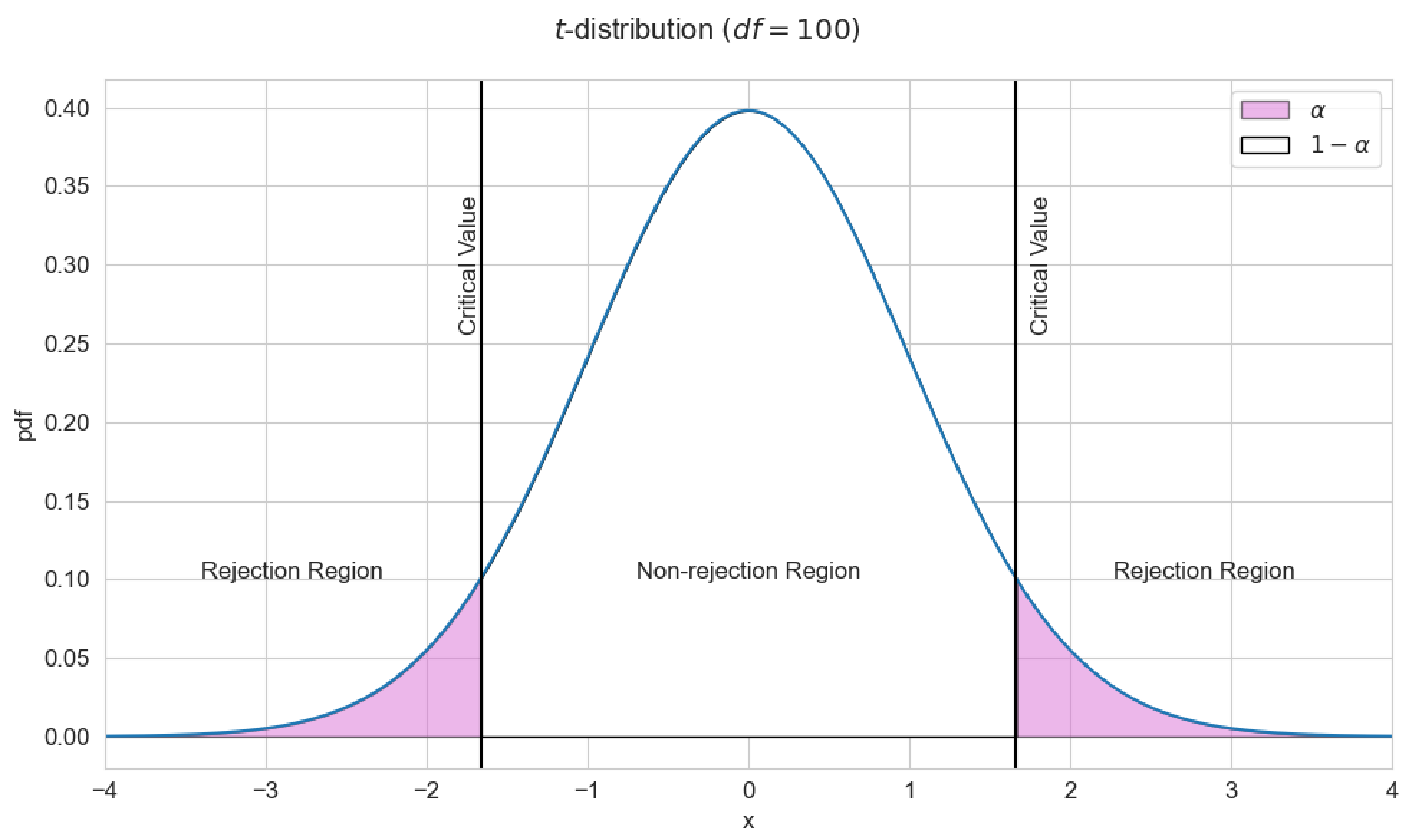
May be one-sided or two-sided.

$$H_a: \mu > \mu_0, \qquad H_a: \mu < \mu_0, \qquad H_a: \mu \neq \mu_0$$

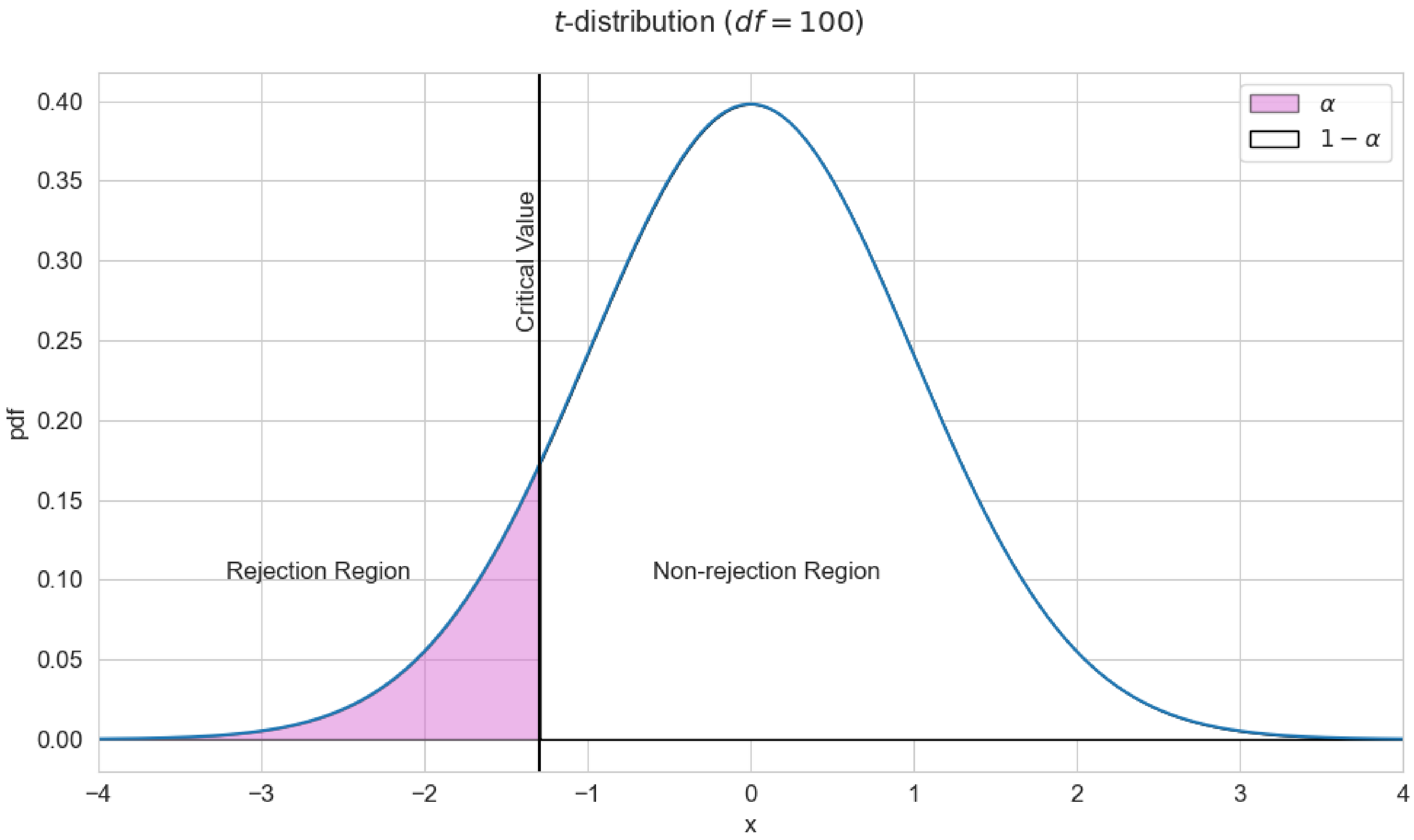
$$H_a: \sigma^2 > \sigma_0^2, \qquad H_a: \sigma^2 < \sigma_0^2, \qquad H_a: \sigma^2 \neq \sigma_0^2$$

$$H_a: \pi > \pi_0, \qquad H_a: \pi < \pi_0, \qquad H_a: \pi \neq \pi_0$$

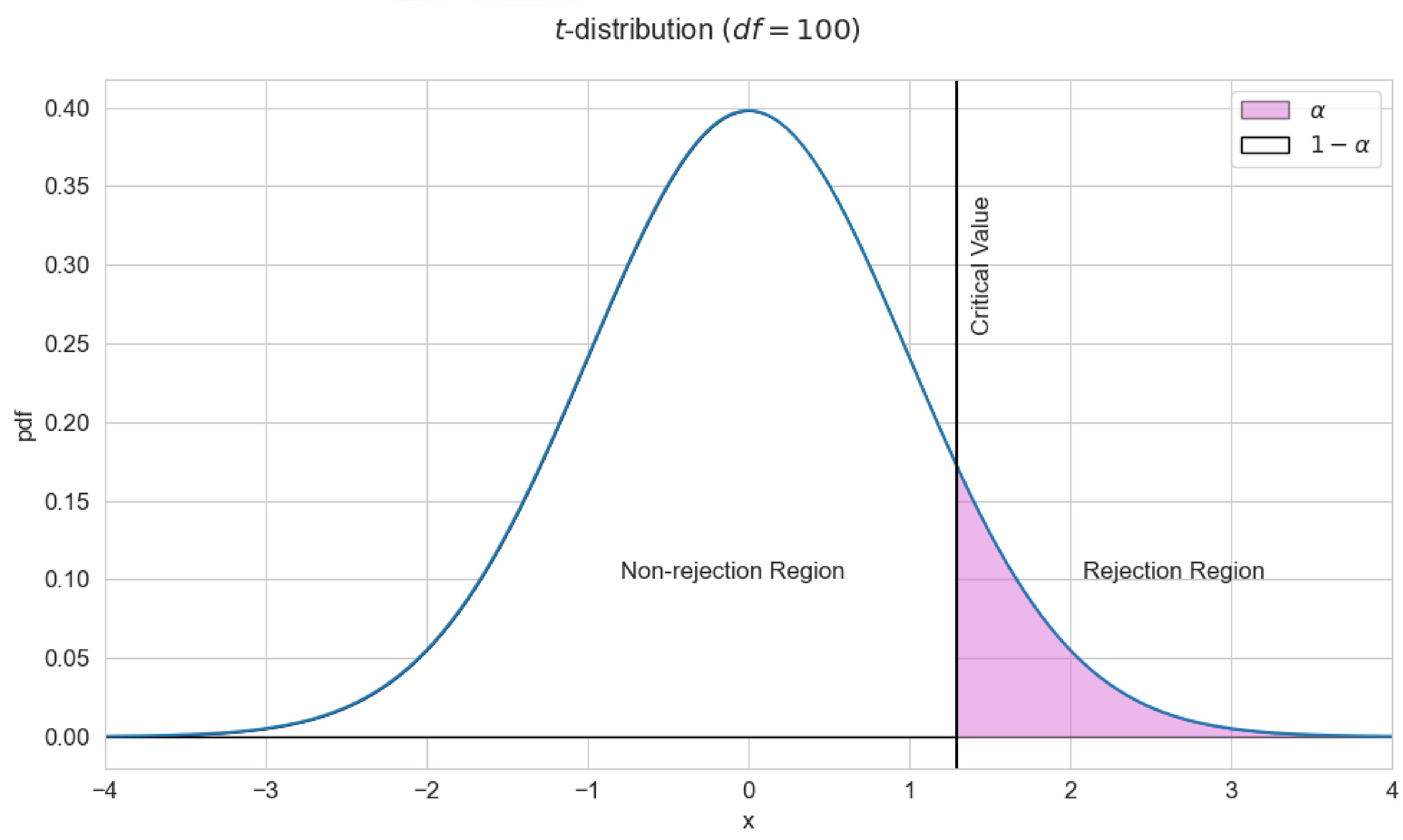
Significance: $H_a: \mu \neq \mu_0$



Significance: $H_a: \mu < \mu_0$



Significance: $H_a: \mu > \mu_0$



Test Statistic

We use data to calculate a test statistic, which summarizes how far the observed point estimate falls from the parameter value in H_0.

This value is examined in terms of the assumed sampling distribution.

Sampling distribution assumptions define the type of hypothesis test, based on factors such as

- Parameter of interest (mean, median, variance, proportion, distribution)
- Data collection method (paired observation, independence, sample size)

p-value

- The *p-value* is the probability, presuming that H_0 is true, that a sample would be collected with a test statistic that equals the observed value or a value even more extreme in the direction predicted by H_a

- Traditionally, p-values are used to choose between conclusions according to some pre-selected significance value α, which represents the probability of a Type I Error (confidence is $1 - \alpha$)

Statistical Errors

Testing can lead to two common errors:

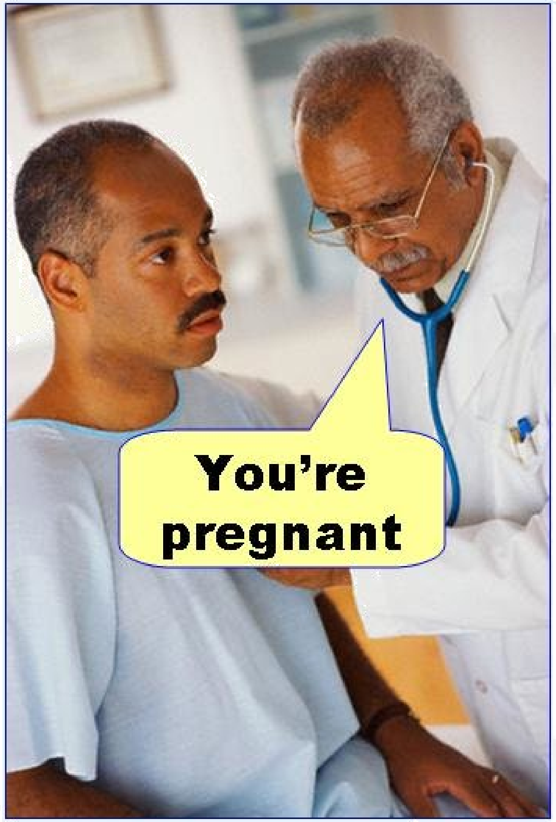
- Type I Error: H_0 is rejected although H_0 is true. (False Positive)
- Type II Error: H_0 is not rejected although H_a is true. (False Negative)

		Truth	
		H_0 True	H_a True
Researcher Concludes	H_0 True	Correct	Type II Error
	H_a True	Type I Error	Correct

Statistical Errors

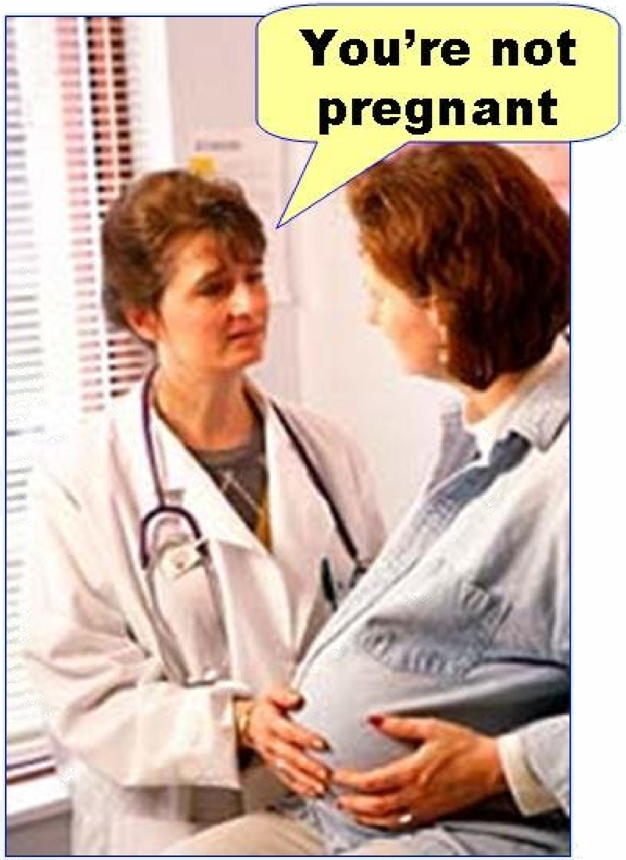
Type I error

(false positive)



Type II error

(false negative)



Conclusion

Evidence will lead us to one of two conclusions:

- Reject the null hypothesis in favor of the alternative hypothesis, or
- Fail to reject the null hypothesis for lack of contradictory evidence

We accept the null hypothesis.

We fail to reject the null hypothesis.

Example: Grocery Store Checkout

At a particular grocery store, the manager suspects that customers are taking longer to check out than they used to. In the past, the average checkout time of customers was 2 minutes.

The manager observes a random sample of 36 customers and finds that they took an average of 3.2 minutes to check out. Does this prove her claim?

To build a Hypothesis Test

1. Define the hypotheses (null and alternative)
   - Parameter of interest
   - Left-, right-, or two-tailed
2. Choose the significance level (α)
3. Choose a test statistic
4. Define the rejection region
   - This may be in terms of test statistic or *p*-value
5. Calculate the test statistic and/or *p*-value
6. Make a conclusion

Assumptions

Type of data: Quantitative (time)

Randomization: Iid random sampling

Population distribution: Exponential with mean of 2 minutes

$$X_i \sim Exp\left(\lambda = \frac{1}{2}\right)$$

Sample size: 36 – sufficient for CLT

Hypotheses

Null hypothesis: Checkout times have not changed

$$H_0: \mu = 2$$

Alternative hypothesis: Checkout times have increased (one-tailed)

$$H_a: \mu > 2$$

Significance

The manager does not want to highlight an imagined problem to the regional manager. She is willing to accept a one-in-ten chance of this happening.

$$\alpha = 0.1$$

Our rejection region is then when $z \geq z^*$, the critical value.

$$P(Z > z^*) = 0.1 \;\Rightarrow\; z^* = N_{ISF}(0.1) = 1.28$$

Test Statistic

Because the sample size is above 30, we assumed the sampling distribution to be a normal distribution.

$$\bar{x} = 3.2$$

$$z = \frac{(\bar{x} - \mu_0)}{\sqrt{\frac{\sigma_0^2}{n}}} = \frac{3.2 - 2}{\sqrt{\frac{4}{36}}} = \frac{1.2}{\frac{1}{3}} = 3.6 > 1.28$$

Conclusion

We reject the null hypothesis and conclude that the mean grocery store checkout time is now greater than 2 minutes.

We would typically also report the p-value.

P-value

The probability of observing this or a more extreme value is

$$p = P(Z \geq 3.6) = 0.00016$$

$$p = 0.00016 < 0.1 = \alpha$$

P-values: what are they (bad) for?

Publication bias: Journals historically publish results with $p < 0.05$.

Random chance: 5% of experiments have $p < 0.05$ randomly

Statistical vs. Practical Significance: With enough data, almost any hypothesis test will be significant. Effect size is more practical.

In 2016, the American Statistical Association released a statement about p-values with six principles underlying the proper use and interpretation of the p-value.

So what should I do instead?

A p-value is still useful, but it should not be treated as the final answer.

Ask yourself: if I was making this decision, what would I want to know?

Often a confidence interval is more useful than a p-value, but hypothesis tests are fundamental to statistical modeling techniques.

Back to the Grocery Store

$$\bar{x} = 3.2$$
$$s = \bar{x} = 3.2$$
$$SEM = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{3.2}{36}}$$
$$N_{PPF}\left(\nu = 35, \mu = 3.2, \sigma = \sqrt{\frac{3.2}{36}}, \alpha = 0.1\right) = 2.82$$

The manager can be 90% confident that checkout times have increased from 2 minutes by at least 49 seconds (0.82 minutes).

Recap

- What is a hypothesis test?
- Defining the hypotheses
- Choosing a test statistic
- Setting the significance level
- Defining the rejection region
- *p*-values: what are they good for?