Introduction to Hypothesis Testing

Keegan Korthauer
Department of Statistics
UW Madison

Hypothesis Test – Motivation

http://science.howstuffworks.com/transport/flight/modern/ejection-seat2.htm

- An engineer is designing a crew escape system that consists of an ejection seat and rocket motor that powers the seat

- The rocket motor contains a propellant, and for the ejection seat to function properly, the propellant should have a mean burning rate of 50 cm/sec

- If the burning rate is too low the seat may not function properly, or if too high, it may cause injury to the pilot

- **Question**: does the mean burning rate of the propellant equal 50 cm/sec, or something else?

Hypothesis Test – Motivation

A sample of **40 rocket motors** are tested. The sample mean and standard deviation of burning rate are **50.7 cm/s** and **2 cm/s** respectively

How certain are we that this sample with its mean of 50.7 cm/s could have have come from a population with **mean different than 50 cm/s**?

Hypothesis Testing

- We might think of obtaining a confidence interval for the mean burning rate μ
  - This doesn't tell us *directly* how confident we are that μ is different from 50

- The statement "μ is different from 50" is a **hypothesis** about the population mean μ

- To determine just how certain we are that this hypothesis is true, we will perform a **hypothesis test**

Null and Alternative Hypotheses

Hypothesis	Example	Conclusion
Null (H_0)	H_0: $\mu = 50$	The population mean burning rate is **equal to 50 cm/s**, and thus the ejection seat will function properly and safely
Alternative (H_1)	H_1: $\mu \neq 50$	The population mean emission rate is **different than 50 cm/s**, and thus the ejection seat will not function properly or cause injury

These correspond to a **two-sided** test – we care if the burning rate is too high **or** too low

Null and Alternative Hypotheses

Hypothesis	Example	Conclusion
Null (H_0)	H_0: $\mu \geq 50$	The population mean burning rate is **greater than or equal to 50 cm/s**, and thus the ejection seat will function properly
Alternative (H_1)	H_1: $\mu < 50$	The population mean emission rate is actually **less than 50 cm/s**, and thus the ejection seat will not function properly

These correspond to a **one-sided** test – here we are not concerned with a burning rate that is too high (for illustrative purposes only)

The Null Hypothesis is 'on trial'

- In a hypothesis test, we start out assuming the null hypothesis is **true** (i.e. that the null hypothesis **innocent until proven guilty**)

- The outcome of the test is a **p-value** which measures the strength of evidence *against* the null hypothesis provided by the sample

- Conclusion: two possible outcomes:
  - If the evidence against H_0 is strong, then we will **reject the null hypothesis**
  - If the evidence against H_0 is weak, then we will **fail to reject the null hypothesis**

Does NOT mean H_0 is true; only that we do not have enough evidence to reject

P-value: Definition and Properties

- **<u>p-value</u>: assuming H_0 is true, the probability that the test statistic would have a value *at least as extreme* as the one observed**

- Ranges between 0 and 1 (a probability) and measures the strength of the disagreement between the sample and H_0

- The smaller the p-value, the stronger the evidence against H_0

- If the p-value is sufficiently small, we may be willing to reject our assumption that H_0 is true in favor of H_1 (i.e. our decision is to reject the null hypothesis)

How Small is Sufficiently Small?

- We reject the null hypothesis when the p-value is small enough – how small depends on the situation
- Formally, we reject H_0 when $p < \alpha$ (this is called **statistical significance**)
- α is called the 'significance level'
- Rule of thumb is to use $\alpha = 0.05$
- In certain situations, we may desire a more conservative level (say $\alpha = 0.01$)

Steps to Perform a Hypothesis Test (HT)

1. Define H_0 and H_1
2. State the level of significance α
3. Construct the test statistic
4. Assume H_0 is true and evaluate the test statistic by finding the p-value
5. Make a conclusion based on the p-value

Use these steps in your HW and on the exam

Outline of HTs

We will learn how to perform hypothesis tests in various situations:

- Large-sample testing for a population mean
- Small-sample testing for a population mean
- One-sample test for a population proportion
- Large-sample testing for two population means
- Small-sample testing for two population means
- Two-sample test for population proportions
- Test for Paired Data
- and more!

LARGE-SAMPLE TEST FOR A POPULATION MEAN

Method

Step by step illustration

Large-Sample Testing for a Population Mean

Let $X_1,\ldots,X_n$ be a large ($n > 30$) sample from any population with mean μ and standard deviation σ. Approximate σ with *s* when unknown.

1. Set up the null H_0 and alternative H_1 hypotheses (see table below)
2. State the level of significance α you will use
3. Calculate the **z-score** (test statistic):

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

Use *s* when unknown!

4. Assume H_0 is true and calculate the P-value:

H_0	H_1	P-value
$\mu \le \mu_0$	$\mu > \mu_0$	Area to the right of *z*
$\mu \ge \mu_0$	$\mu < \mu_0$	Area to the left of *z*
$\mu = \mu_0$	$\mu \ne \mu_0$	Area to the left of -z plus area to the right of z

5. Make a conclusion based on the P-value

Ejection Seat Example

Statement of the problem:

In the ejection seat example previously described, we have a simple random sample of 40 rocket motors with mean burning rate of **50.7 cm/s** and sample standard deviation **2 cm/s**

Can we conclude that the mean burning rate is different from 50 cm/s at the 5% level?

Ejection Seat Example – Step 1

1. Set up the null and alternative hypotheses:

To *conclude that the mean burning rate is different from 50,* we must reject the null hypothesis that the mean burning rate is equal to 50 <- why not the other way around?

More formally, our null hypothesis is that $\mu = 50$ and we will reject the null if we have enough evidence to overturn it

Therefore, our μ_0 is 50 and our (two-sided) hypotheses are

$$\mathbf{H_0: \mu = 50}$$

$$\mathbf{H_1: \mu \neq 50}$$

Why can't we define the null and alternative like this?

$$H_0: \mu \neq 50$$

$$H_1: \mu = 50$$

In a HT, there are only two possible outcomes:

1. Reject H_0 (so conclude H_0 is false)
2. Fail to reject H_0 (so conclude H_0 is **plausible**)

If we had defined: H_0: $\mu \neq 50$, then these outcomes are:

1. Reject that $\mu \neq 50$ (so conclude that $\mu = 50$)
2. Fail to reject that $\mu \neq 50$ (so conclude it's plausible that $\mu \neq 50$)

Neither of these options result in **concluding $\mu \neq 50$ is true**

Ejection Seat Example – Step 2

2. State the level of significance α you will use

The question indicates that we will use the 5% level, so

$\alpha = 0.05$

Engine Seat Example – Step 3

3. Calculate the test statistic

We have a large sample (n > 30) so it is appropriate to use z-score for the test statistic

$$z = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} = \frac{50.7 - 50}{2 / \sqrt{40}} = 2.214$$

Ejection Seat Example – Step 4

4. Assume H_0 is true and calculate the P-value

- If H_0 is true, then the sample was drawn from a population with mean $\mu = 50$
- We estimate the population standard deviation σ with s=2
- Then under H_0 the CLT tells us that

$$\bar{X} \sim N(50, 2/\sqrt{40}) \Rightarrow z\text{-score} \sim N(0,1)$$

- P-value = P(test stat. at least as extreme as the one observed)

$$= P(Z > 2.214 \text{ or } Z < -2.214)$$

$$= P(Z > 2.214) + P(Z < -2.214)$$

$$\approx 2 * 0.0136 = 0.0272$$

Ejection Seat Example – Step 5

5. Make a conclusion based on the P-value

The p-value of 0.0272 indicates that we have evidence against H_0 because it is less than the significance level $\alpha = 0.05$ (p-value < α). Therefore we reject the null hypothesis that $\mu = 50$ and conclude that $\mu \neq 50$.

Based on the results of this hypothesis test, we can conclude that the mean burning rate is different from 50, so we should aim to improve the design of the rocket motor.

Interpreting a P-value

- The smaller the p-value, the more certain we are that H_0 is false

- But, the p-value does **NOT** represent the probability that the null hypothesis is false
  - Much like for CIs, we can only talk about probability in HTs in terms of repeated sampling out of a population – the test statistic is random in this case
  - The null hypothesis is either true or not true – there is no randomness involved there

Example – Problem 6.1.4

The pH of an acid solution used to etch aluminum varies somewhat from batch to batch. In a sample of 50 batches the mean pH was 2.6, with a standard deviation of 0.3. Let μ represent the mean pH for batches of this solution.

(a) Perform a hypothesis test at the 0.05 level with H_0: $\mu \leq 2.5$ and H_1: $\mu > 2.5$

(b) Either the mean pH is greater than 2.5, or the sample is in the most extreme __0.91__% of its distribution

Example – 6.1.9

A random sample of 126 international construction projects had an average profit margin (in %) of 8.24 with a standard deviation of 16.33.

Can we conclude that the mean profit margin μ for all international construction projects is less than 10% at the 0.05 level?

Statistical vs. Practical Significance

- When the p-value is less than the significance level, we the test has achieved *statistical significance* at that level

- Statistical significance doesn't guarantee *practical significance*
  - this will depend on the context of the problem
  - it is possible to have a highly statistically significant result of little to no practical value

Example - Statistical vs. Practical Significance

Say we know that a machine can produce fibers with mean breaking strength of 50N, but we have the option of purchasing a new (very expensive) machine that may be better (produce fibers with mean breaking strength greater than 50N)

To test if it is better, we sample 1000 random fibers from the new machine with an average of 50.1N and sample standard deviation 1N

- Can we conclude that the new machine is better (higher mean breaking strength)?

- Is this result of practical significance?

HTs and CIs

- For a population mean μ:
  - CI - collection of all values for μ that meet a certain level of plausibility
  - HT - specify a particular value of μ (null hypothesis) and determine how plausible it is
- They are related in the following ways:
  - the values contained in a 2-sided level $100(1-\alpha)\%$ CI for μ are all those whose p-value of a 2-sided HT will be greater than α
  - the values contained in a 1-sided level $100(1-\alpha)\%$ CI for μ are all those whose p-value of a 1-sided HT will be greater than α

Illustration: HTs and CIs

- Sample mean lifetime of 50 microdrills was 12.68 holes drilled, standard deviation was 6.83
- The 95% CI for the mean lifetime μ is (10.79, 14.57)
- A test of H_0: $\mu = 10.79$ vs H_1: $\mu \neq 10.79$ gives the test statistic z=1.96 which yields a p-value of $P(Z > 1.96 \text{ or } Z < -1.96) = 0.05$

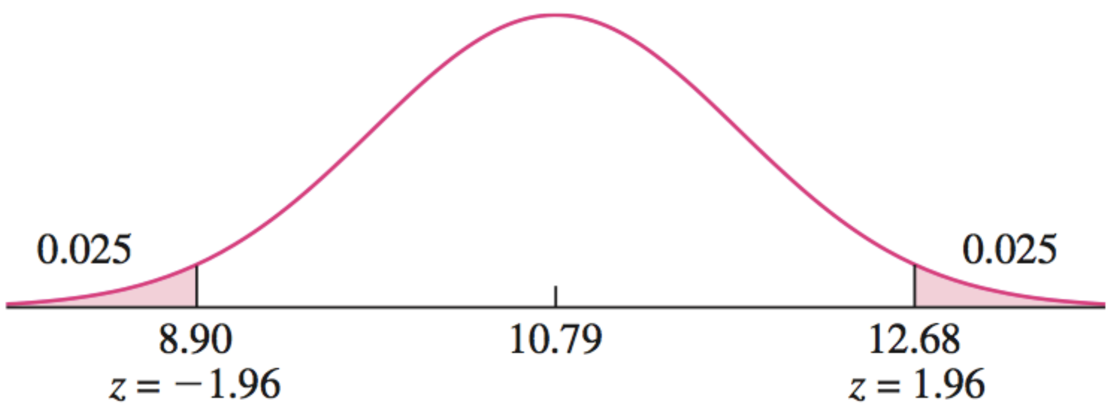

FIGURE 6.4 The sample mean $\overline{X}$ is equal to 12.68. Since 10.79 is an endpoint of a 95% confidence interval based on $\overline{X} = 12.68$, the P-value for testing $H_0: \mu = 10.79$ is equal to 0.05.

Example – HTs and CIs

In the previous example, the 95% CI for the mean lifetime μ is (10.79, 14.57)

Can we determine whether to reject the null at the 5% level for H_0: $\mu = 11.4$ vs H_1: $\mu \neq 11.4$ without additional calculations?

Yes; since 11.4 is contained in the 95% CI, we will fail to reject H_0 at the 5% level.

Can we determine whether to reject the null at the 1% level for H_0: $\mu = 15.2$ vs H_1: $\mu \neq 15.2$ without additional calculations?

No; we would need to calculate a 99% CI to determine whether we would reject at the 1% level.

Recap: Two **Key Concepts** in HT

1. We can reject H_0 or we can fail to reject H_0, but we can never accept that H_0 true

2. The p-value is **not** the probability that the null hypothesis is true

Next

- More Hypothesis Testing
  - Population Proportion
  - Small-sample mean

- HW7 will be posted today or tomorrow, due Friday 3/28

- Happy Spring Break!