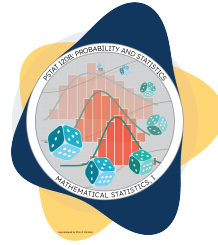
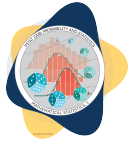


Topic 6: Hypothesis Testing

Ethan P. Marzban University of California, Santa Barbara PSTAT 120B

Outline

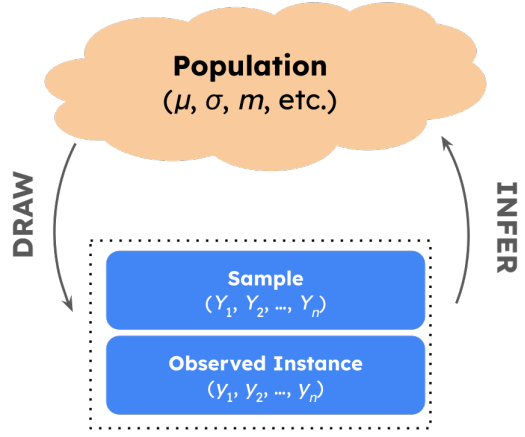


1. Basics of Hypothesis Tests

2. Test for a Mean; Normal Population

3. $p-$Values

Cycle of Inference



UC SANTA BARBARA
Department of Statistics
and Applied Probability

Leadup

- Up until now, we've treated "inference" as synonymous with "estimation" - given a population with unknown parameter θ, we sought to take samples $\vec{\boldsymbol{Y}}$ from this distribution and *estimate* the true value of θ.
  - We started off by constructing **point estimators**, which are rules used to compute estimates for a parameter.
  - More recently, we considered constructing **interval estimators** (aka **confidence intervals**), which are random intervals we believe, with some specified coverage probability, cover the true value of θ.

Leadup

- But, as I mentioned several lectures ago (when we first started talking about inferential statistics), estimation is not the only part of inferential statistics - we also have the notion of **hypothesis testing**.
- This is going to be the final topic for this course, and the topic we will explore for the next few lectures.
- As a disclaimer - hypothesis testing is an incredibly rich subfield of statistics, and I will by no means be able to to it justice in the few lectures we have together.
  - Rather, I hope to provide you with a *foundation* upon which you can build.

Game Plan

- Today we'll talk about the basics of hypothesis testing, and some general terminology.
- We'll then construct hypothesis tests in the normal setting (i.e. assuming a normally-distributed population), and introduce the notion of $p-$values in this context.
- Then (in the next set of slides) we'll talk about more general tests, and how to compare between two hypothesis tests.

Basics of Hypothesis Tests

UC SANTA BARBARA
Department of Statistics
and Applied Probability

Polydactyly

Source: https://www.treehugger.com/thing-didnt-know-polydactyl-cats-4864197

- **Polydactyly** refers to the situation in which a person/animal is born with extra fingers/toes.
- According to a Quora post, the average cat has about a 10% chance of being born with polydactyly.
- Quora would never lie to us... would it?

Leadup

- Say we collected a random sample of 100 cats.
- If we observed only 9 of these cats as being polydactyl, we probably wouldn't question the validity of the Quora statistic.
- But, say we observed 83 polydactyl cats in our sample of 100. Now we might start to question whether or not there's something wrong with the Quora statistic.
- The same thing goes for a situation in which we take a sample of 100 cats and observe only one polydactyl cat!

Leadup

- Now, of course, we wouldn't be able to definitively say that the Quora statistic was wrong. It's *possible* that the true percentage of polydactyl cats is 10%, and we just got *incredibly* lucky and observed 83 polydactyl cats in a sample of 100.
- But, it's pretty *unlikely*, meaning (assuming we did our sampling correctly), we might want to start questioning the Quora statistic.
- Believe it or not, this is the same idea we use to conduct a **hypothesis test**!

Components of a Test

- We start with a **null hypothesis**, which is a claim (which, for this class, will usually be about a population parameter).
  - For example, if p denotes the true proportion of all cats that are polydactyl, then our null hypothesis might be the statistic Quora gave - that p is 0.1.
- We use the notation H_0 to denote the null hypothesis. This is read "h - naught".
  - So, for example, in the polydactyl cat example, we would write $H_0 : p = 0.1$.
- We also need to provide an **alternative hypothesis**, notated H_A. As the name suggests, this is a claim that we think might be true if the null proves to be incorrect.

Components of a Test

- If our null takes the form $H_0 : \theta = \theta_0$ for some specified value of θ_0 (called the **null value**), there are four broad classes of alternatives we could consider:
  - **Lower-Tailed Alternative:** $H_A : \theta < \theta_0$
  - **Upper-Tailed Alternative:** $H_A : \theta > \theta_0$
  - **Two-Sided (aka Two-Tailed) Alternative:** $H_A : \theta \neq \theta_0$
  - **Simple-vs.-Simple** $H_A : \theta = \theta_A$ for some $\theta_A \neq \theta_0$.

- For example, a lower-tailed alternative to the Quora statistic would be $H_A : p < 0.1$.

Components of a Test

- For the purposes of this class, we will only consider null hypotheses of the form $H_0 : \theta = \theta_0$.
- Note: if we do this, a lower- or upper-tailed alternative should *not* include a sign of equality!
- For example, it is incorrect to phrase a lower-tailed alternative as $H_A : \theta \leq \theta_0$. This is because, what would we do if $\theta = \theta_0$? This case is covered by both the null and alternative, which is a contradiction!
- So, essentially, make sure that the null and alternative hypotheses cannot both be true at the same time.

Example

Example

A new restaurant claims that the average amount of time a customer spends waiting for a table is 5 minutes. Write down the null hypothesis, as well as a lower-tailed, upper-tailed, and two-sided alternative hypothesis.

Solution

- **ALWAYS** begin by defining parameters and notation!
  - To be clear, failure to do so may incur point penalties on quizzes/exams/homework.
- Let μ denote the true average wait time of customers at this restaurant.
- Our null value for μ is the value being provided by the restaurant - 5. Hence, our null hypothesis is

$$H_0 : \mu = 5$$

Solution

- A lower-tailed alternative is the claim that the true average wait time at this restaurant is *under* 5 minutes. Mathematically, this is given by $H_A : \mu < 5$.
- An upper-tailed alternative is the claim that the true average wait time at this restaurant is *over* 5 minutes. Mathematically, this is given by $H_A : \mu > 5$.
- A two-sided alternative is the claim that the true average wait time at this restaurant is *not* 5 minutes. Mathematically, this is given by $H_A : \mu \neq 5$.

- To stress: in reality, we would pick only *one* of these alternatives to test against the null.

Decision Rule

- So, those are what our null and alternative hypotheses are.
- A **hypothesis test** is essentially a rule that takes in data, and returns one of two conclusion: "reject the null in favor of the alternative", or "fail to reject the null in favor of the alternative." Such a rule we often call a **decision rule**, or **decision function**:

$$\texttt{decision(data)} = \begin{cases} \texttt{reject } H_0 \texttt{ in favor of } H_A \\ \texttt{fail to reject } H_0 \texttt{ in favor of } H_A \end{cases}$$

Leadup

- Now, before we begin discussing how to actually construct a decision rule with desirable properties, we should take a quick step back.
- Again, our decision will either be "reject H_0 in favor of H_A" or "fail to reject H_0 in favor of H_A."
- Behind the scenes, though, H_0 will either be true or false!
  - Sure, we'll never be able to definitively tell wehther it was true or not, but it will either be true or not - the true proportion of polydactyl cats is in fact either 10% or not 10%.
- This means there are essentially four states of the world:

States of the World

		State of H_0: True	State of H_0: False
Decision	**Reject H_0**	**BAD**	**GOOD**
	Fail to Reject H_0	**GOOD**	**BAD**

UC SANTA BARBARA
Department of Statistics and Applied Probability

Type I and Type II Errors

Decision	State of H_0: True	State of H_0: False
Reject H_0	**Type I Error**	
Fail to Reject H_0		**Type II Error**

UC SANTA BARBARA
Department of Statistics and Applied Probability

Type I and Type II Errors

Definition (Type I and Type II Errors)

A **Type I error** is made if H_0 is true but is rejected; a **Type II error** is made if H_0 is false but it is failed to be rejected.

Type I and Type II Errors

- Type I and Type II errors can be nicely phrased in terms of the US judicial system.
- The motto of the US judicial system is “innocent until proven guilty.” Hence, we can think of the null hypothesis as: a given person is innocent.
- A Type I error is akin to convicting an innocent person; i.e. they were innocent (the null was true) but we rejected their innocence.
- A Type II error is akin to letting a guilty person go free; i.e. they were not innocent (the null was false) but we failed to reject their innocence.

Type I and Type II Errors

- Now, as their names suggest, both Type I and Type II errors are errors - we'd like to commit them with as small probability as possible.
- But, as with many things in statistics (and life), there is a tradeoff.
- Going back to the judicial system, say I am a judge and I really want to reduce the probability that I commit a Type I error. That is, I want to *never* convict an innocent person.
- Well, to *never* convict an innocent person, I'd need to *never* convict *anyone*.
- But, in doing so, I would be letting every guilty person that passes through my courtroom free - i.e. I have drastically *increased* the probability of committing a Type II error.

Type I and Type II Errors

- So, we cannot simultaneously reduce the risk of committing both a Type I and Type II error.
- By the way, we can already see that the "risk of committing a Type __ error" has shown up a lot. As such, it will be useful to give these probabilities names:

Definition (α and β)

We define α to be the probability of committing a Type I error, and β to be the probability of committing a Type II error. α is usually also referred to as the **level of significance**.

Example

Example

The manager of a certain company claims that 20% of her employees work remotely. An external auditor wishes to test this manager's claims, and collects a representative sample of several employees from this company.

Phrase a Type I and Type II error in the context of this scenario, adopting a lower-tailed alternative.

Solutions

- Letting p denote the true proportion of employees at this company that work remotely, our null hypothesis is $H_0 : p = 0.2$.
- A Type I error is committed when the null is true but we reject it (in favor of the alternative). That is, a Type I error is committed if the true proportion of employees who work remotely *is* 20%, but the auditor determines that the true proportion is *lower* than 20%.
- A Type II error is committed when the null is false but we fail to reject it (in favor of the alternative). That is, a Type I error is committed if the true proportion of employees who work remotely is *lower* than 20%, but the auditor determines that the true proportion is *equal to* 20%.

Check your Understanding!

- There's a lot of terminology that's been thrown around already!
- We will, however, be using this terminology a *lot* going forward. So, please make sure you understand the following terms:
  - Null and Alternative Hypotheses
  - Decision Rules
  - Type I and Type II Errors, and significance level

Test for a Mean; Normal Population

Leadup

- Like we've done several times throughout this course, let's temporarily restrict ourselves to a specific case, and then eventually try to work our way back to generality.

Goal

Given $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with known σ^2, we wish to construct a hypothesis test of the following hypotheses:

$$\begin{bmatrix} H_0: & \mu = \mu_0 \\ H_A: & \mu \neq \mu_0 \end{bmatrix}$$

- By the way, what type of alternative hypothesis is this?

Leadup

- Recall that I said a hypothesis test is essentially a decision rule:

$$\texttt{decision(data)} = \begin{cases} \texttt{reject } H_0 \texttt{ in favor of } H_A \\ \texttt{fail to reject } H_0 \texttt{ in favor of } H_A \end{cases}$$

  - Going forward I'll use "f.t.r." as an abbreviation for "fail to reject." By the way, let's quickly discuss why I use the term "fail to reject" instead of "accept!"

Leadup

- Well, firstly note that, as the notation above suggests, our decision rule depends on *data*.
- Specifically, there are certain values of our data that, when fed into the decision rule, will result in a decision of "reject". The set of all such values is called the **rejection region** of the test, and is notated $\mathcal{R}$:

$$\texttt{decision(data)} = \begin{cases} \texttt{reject } H_0 & \text{if } \texttt{data} \in \mathcal{R} \\ \texttt{f.t.r. } H_0 & \text{otherwise} \end{cases}$$

Leadup

- It is customary to construct our decision rule to only depend on our data through a **statistic**.
- Since we're trying to test hypotheses pertaining to a population *mean*, it makes sense to have our decision rule depend on $\overline{Y}_n$, the sample mean:

$$\texttt{decision}(\overline{Y}_n) = \begin{cases} \texttt{reject} \;\; H_0 & \text{if } \overline{Y}_n \in \mathcal{R} \\ \texttt{f.t.r.} \;\; H_0 & \text{otherwise} \end{cases}$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Intuition

- Let's think heuristically about what our rejection region should look like.
- In other words, for what values of $\overline{Y}_n$ might we reject the null hypothesis that $H_0 : \mu = \mu_0$ in favor of the alternative that $H_A : \mu \neq \mu_0$?
- Maybe an example will help. Let μ denote the true average weight of a DSH cat; say $\mu_0 = 10$. That is, we're trying to see whether the true average weight of a DSH cat is 10 lbs or not 10 lbs.
- Clearly, if a particular sample of cats had an average weight of 30 lbs, we'd probably be inclined to reject H_0. Similarly, if a particular sample of cats had an average weight of 3 lbs, we'd also probably be inclined to reject H_0.

Intuition

- Another way to say this is: we'd be inclined to reject the null (that $\mu = 10$) if $\overline{Y}_n$ was either much smaller or much greater than 10.
- Mathematically, we can say that the *distance* between $\overline{Y}_n$ and 10 should be large in order for us to reject.
- Hence, a reasonable rejection region seems to be of the form

$$\{|\overline{Y}_n - 10| > c\}$$

for some constant c.

UC SANTA BARBARA
Department of Statistics and Applied Probability

Intuition

- So, going back to our general test for the mean, we now have the following form for the decision rule:

$$\texttt{decision}(\overline{Y}_n) = \begin{cases} \texttt{reject } H_0 & \text{if } \{|\overline{Y}_n - \mu_0| > c\} \\ \texttt{f.t.r.}H_0 & \text{otherwise} \end{cases}$$

  for some (as-of-yet undetermined) constant $c > 0$.

- So, really the only thing left to do is pick c, which we call the **critical value**!

UC SANTA BARBARA
Department of Statistics and Applied Probability

Intuition

- Okay, so how do we pick c?
- The answer will require us to go back to the notions of Type I and Type II errors.
- Recall that a Type I error is committed when we reject a true null. Furthermore, the level of significance α is the probability of committing a Type I error.
- Our null hypothesis is that $\mu = \mu_0$, and our decision rule is as stated on the previous slide. So, the event "a Type I error was committed" can be expressed as

$$\{|\overline{Y}_n - \mu_0| > c, \text{ when } \mu = \mu_0\}$$

Intuition

- Therefore, by the definition of α,

$$\mathbb{P}(|\overline{Y}_n - \mu_0| > c, \text{ when } \mu = \mu_0) = \alpha$$

- Statisticians always *begin* by specifying α. For example, we start off our test by saying - I want to use a $\alpha = 0.05$ level of significance (i.e. I want to commit a Type I error with probability 0.05). From this, we then solve the above probability equation to identify the appropriate value of c.
- Allow me to write $\mathbb{P}_{H_0}(A)$ to mean "the probability of an event A, when H_0 is true." This will help save us a lot of time.

Solving for the Critical Value

- The equation c must satisfy is

$$\mathbb{P}_{H_0}(|\overline{Y}_n - \mu_0| > c)$$

- Now, recall that our null hypothesis is just $H_0 : \mu = \mu_0$. So, the probability on the LHS is saying "the probability that $\overline{Y}_n$ is more than c units away from μ_0, assuming μ_0 is the true population mean."
- That is, under the null, we assume that μ_0 *is* the true value of μ, meaning

$$\frac{\sqrt{n}(\overline{Y}_n - \mu_0)}{\sigma} \overset{H_0}{\sim} \mathcal{N}(0, 1)$$

where the symbol "$\overset{H_0}{\sim}$" means "distributed as, under the null (i.e. assuming the null is true)."

UC SANTA BARBARA
Department of Statistics and Applied Probability

Solving for the Critical Value

- Hence:

$$\begin{aligned}
\mathbb{P}_{H_0}(|\overline{Y}_n - \mu_0| > c) &= \mathbb{P}_{H_0}\left(\overline{Y}_n - \mu_0 < -c\right) + \mathbb{P}_{H_0}\left(\overline{Y}_n - \mu_0 > -c\right) \\
&= \mathbb{P}_{H_0}\left(\frac{\overline{Y}_n - \mu_0}{\sigma/\sqrt{n}} < \frac{c}{\sigma/\sqrt{n}}\right) + \mathbb{P}_{H_0}\left(\frac{\overline{Y}_n - \mu_0}{\sigma/\sqrt{n}} > \frac{c}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-\frac{c}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) = 2 - 2\Phi\left(\frac{c}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

Solving for the Critical Value

- Therefore,

$$2 - 2\Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) = \alpha$$

meaning, solving for c,

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}$$

UC SANTA BARBARA
Department of Statistics
and Applied Probability

The Test

- So, to summarize: if $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and our null and alternative hypotheses are

$$\begin{bmatrix} H_0: & \mu = \mu_0 \\ H_A: & \mu \neq \mu_0 \end{bmatrix}$$

  a level-α test is:

$$\texttt{Reject } H_0 \quad \text{if} \quad |\overline{Y}_n - \mu_0| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}$$

The Test

- So, to summarize: if $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and our null and alternative hypotheses are

$$\begin{bmatrix} H_0: & \mu = \mu_0 \\ H_A: & \mu \neq \mu_0 \end{bmatrix}$$

  a level-α test is:

$$\texttt{Reject } H_0 \quad \text{if} \quad \left| \frac{\overline{Y}_n - \mu_0}{\sigma/\sqrt{n}} \right| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Two-Sided $Z-$Test

Theorem (Two-Sided $Z-$Test)

- **Assumptions:** $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for unknown $\mu \in \mathbb{R}$ but known $\sigma^2 > 0$; level α of significance.
- **Hypotheses:** $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$, for some specified $\mu_0 \in \mathbb{R}$
- **Test:** Reject H_0 in favor of H_A if:

$$\left| \frac{\overline{Y}_n - \mu_0}{\sigma/\sqrt{n}} \right| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Example

Example

A new restaurant claims that the average amount of time a customer spends waiting for a table is 5 minutes. To test this claim, a representative sample of 10 tables is taken, and their wait times (in minutes) are as follows:

$$\vec{\boldsymbol{y}} = (4,\ 5,\ 4,\ 4,\ 5,\ 3,\ 3,\ 1,\ 2,\ 1)$$

Assume that the standard deviation among all wait times is 1.5 minutes. Conduct a two-sided test of the restaurant's claims, using an $\alpha = 0.05$ level of significance. Phrase the conclusions of your test in the context of the problem.

Solution

- We really only need to plug into our previous theorem!
- First, we compute the value of the **test statistic**

$$\left|\frac{\bar{y}_n - \mu_0}{\sigma/\sqrt{n}}\right| = \left|\frac{3.2 - 5}{1.5/\sqrt{10}}\right| = |-3.795| = 3.795$$

- Next, we compute the critical value:

$$\Phi^{-1}\left(1 - \frac{0.05}{2}\right) \approx 1.96$$

- Finally, we compare: $3.795 > 1.96$ meaning the observed value of our test statistic has fallen into the rejection region, meaning we should reject the null.

Solution

- Now, I am a stickler about this - we shouldn't just say "reject!"
- A properly-stated conclusion to a hypothesis test should include mention of the level of significance and the alternative hypothesis, while being stated in the context of the problem.
- So, here's how I would phrase the conclusion of this particular test:

> *At a 5% level of significance, there was sufficient evidence to reject the null hypothesis that the true average wait time at this restaurant is 5 minutes in favor of the two-sided alternative that the true average weight time is not 5 minutes.*

Lower-Tailed $Z-$Test

Theorem (Lower-Tailed $Z-$Test)

- **Assumptions:** $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for unknown $\mu \in \mathbb{R}$ but known $\sigma^2 > 0$; level α of significance.
- **Hypotheses:** $H_0 : \mu = \mu_0$ vs $H_A : \mu < \mu_0$, for some specified $\mu_0 \in \mathbb{R}$
- **Test:** Reject H_0 in favor of H_A if:

$$\frac{\overline{Y}_n - \mu_0}{\sigma/\sqrt{n}} < \Phi^{-1}(\alpha)$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Upper-Tailed $Z-$Test

Theorem (Upper-Tailed $Z-$Test)

- **Assumptions:** $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for unknown $\mu \in \mathbb{R}$ but known $\sigma^2 > 0$; level α of significance.
- **Hypotheses:** $H_0 : \mu = \mu_0$ vs $H_A : \mu > \mu_0$, for some specified $\mu_0 \in \mathbb{R}$
- **Test:** Reject H_0 in favor of H_A if:

$$\frac{\overline{Y}_n - \mu_0}{\sigma/\sqrt{n}} > \Phi^{-1}(1-\alpha)$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Comments

- To derive these, you use a very similar argument to the one we used to construct the two-sided test.
- I ask you to work through one of them on your homework.
- Now, note that the $Z-$test requires a known population standard deviation σ. If we do *not* know σ, we can simply replace it with S_n (the sample standard deviation) and use what is known as a $T-$test:

$T-$Test

Theorem ($T-$Test)

- **Assumptions:** $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for unknown $\mu \in \mathbb{R}$ and unknown $\sigma^2 > 0$; level α of significance.
- **Test:** Reject H_0 in favor of H_A if:

$$\begin{cases} \left| \frac{\overline{Y}_n - \mu_0}{\sigma/\sqrt{n}} \right| > F^{-1}_{t_{n-1}}\left(1 - \frac{\alpha}{2}\right) & \text{(two-sided)} \\ \frac{\overline{Y}_n - \mu_0}{\sigma/\sqrt{n}} < F^{-1}_{t_{n-1}}(\alpha) & \text{(lower-tailed)} \\ \frac{\overline{Y}_n - \mu_0}{\sigma/\sqrt{n}} > F^{-1}_{t_{n-1}}(1-\alpha) & \text{(upper-tailed)} \end{cases}$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Example

Example

GauchoPop claims that their bottles contain, on average, 16 oz of soda. To test this claim, Ekaterina takes a random sample of 16 *GauchoPop*-brand soda bottles and finds that this sample contains an average of 15.5 oz of soda. She also computes the standard deviation of soda fills in her sample to be 2oz. Assume that the fill of a randomly-selected bottle is normally distributed.

Conduct a lower-tailed hypothesis test of *GauchoPop*'s claims, using a $\alpha = 0.05$ level of significance. Phrase your conclusion in the context of the problem.

Solution

- Let μ denote the true average fill of a randomly-selected *GauchoPop*-brand soda bottle. Our null and alternative hypotheses are therefore

$$\begin{bmatrix} H_0: & \mu = 16 \\ H_A: & \mu < 16 \end{bmatrix}$$

- Let $\vec{\boldsymbol{Y}} := \{Y_i\}_{i=1}^n$ denote the fills of a random sample of n *GauchoPop*-bottles; we're told to assume

$$Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

- Since we don't know σ^2, we need to use a $T-$test as opposed to a $Z-$test. Our test statistic takes the value

$$\frac{\bar{y}_n - \mu_0}{s_n/\sqrt{n}} = \frac{15.5 - 16}{2/\sqrt{16}} = -1$$

Solution

- Since we're using a $T-$test, the critical value of the test is given by

$$F_{t_{15}}^{-1}(0.05) = -1.75$$

- Note that our test statistic (-1) is *not* less than the critical value, meaning it does *not* fall into the rejection region. Hence, we fail to reject the null:

> *At a 5% level of significance, there was insufficient evidence to reject GauchoPop's claims that the average amount of fill in one of their soda bottles is 16 oz, in favor of the alternative that the true fill is less than 16 oz.*

p—Values

Leadup

- The framework discussed up until now is often referred to as the "critical value" framework, as our test ultimately took the form: compare the observed value of a test statistic with a *critical value* to determine whether or not the null should be rejected.
- The critical value clearly depends on the level of significance.
- There is an equivalent way of rephrasing our test, involving something known as a $p-$value.

$p-$Values

Definition ($p-$Value; from the Textbook)

If W is a test statistic, the ***p*−value**, or attained significance level, is the smallest level of significance α for which the observed data indicate that the null hypothesis should be rejected.

- Though this is a perfectly valid definition, I'm personally not too keen about it.
- Specifically, I find that it doesn't really tell us *how* to compute a $p-$value.

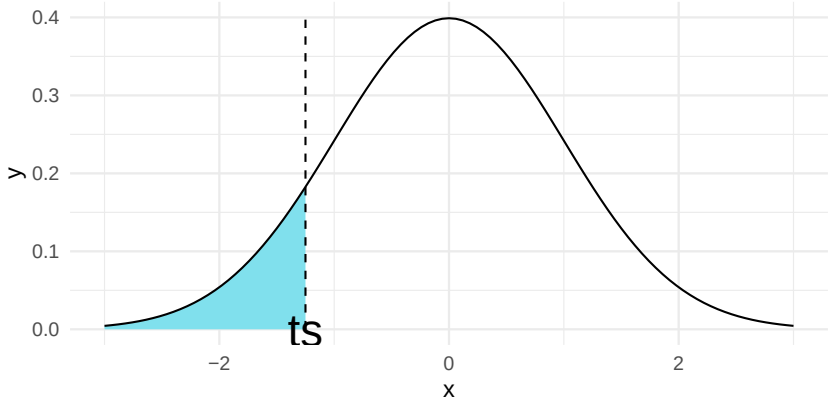
$p-$Values

Definition ($p-$Value; from the Textbook)

The $p-$value of a test statistic is the probability of observing something as or more extreme, in the direction of the null, as what was observed.

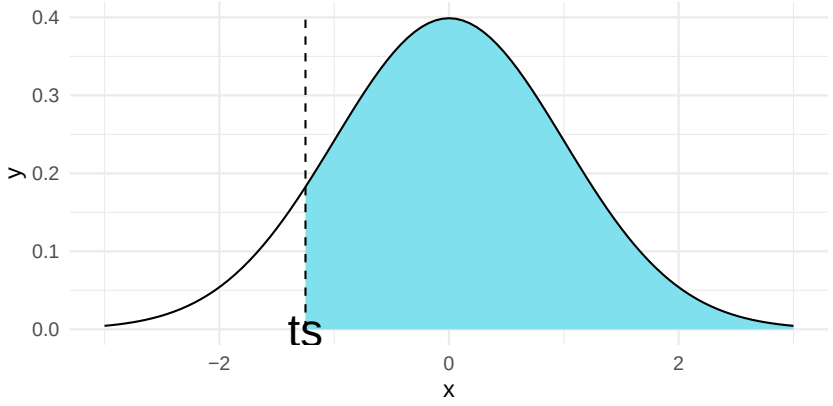
- In my opinion, this definition allows us (more directly) to draw a picture!

p—Values: Lower-Tailed Test



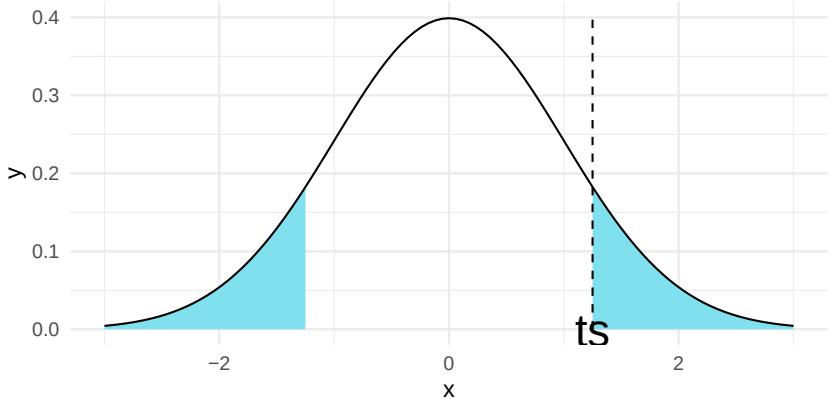
UC SANTA BARBARA
Department of Statistics
and Applied Probability

p—Values: Upper-Tailed Test



UC SANTA BARBARA
Department of Statistics
and Applied Probability

$p-$Values: Two-Sided Test



Example

Example

Suppose that a particular university's administration claims that the average commute time among all of their students is 10 minutes. A particular sample of 36 students randomly selected from this university had an average commute time of 9 minutes and a standard deviation of 3 minutes. Assume that commute times are independent and normally distributed.

Conduct a two-sided hypothesis test of the administration's claims, and report a $p-$value.

Solution

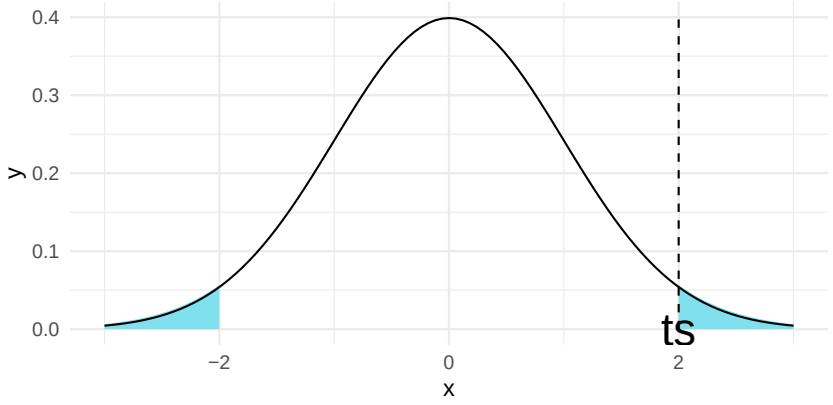
- Let μ denote the true average commute time; let $\vec{\boldsymbol{Y}} := \{Y_i\}_{i=1}^n$ denote the commute time of a randomly-selected set of n students.
- Our null hypothesis is $H_0 : \mu = 10$ and our alternative is $H_0 : \mu \neq 10$.
- The value of our test statistic is

$$\left| \frac{\overline{y}_n - \mu_0}{s_n/\sqrt{n}} = \frac{9 - 10}{3/\sqrt{36}} \right| = |-2|$$

- Since σ^2 is unknown, we use the $t_{35}-$distribution to compute our $p-$value:

UC SANTA BARBARA
Department of Statistics and Applied Probability

Solution



UC SANTA BARBARA
Department of Statistics
and Applied Probability

Solution

- Our $p-$value is therefore given by $2F_{t_{35}}(-2)$
- Note that the following are also all equivalent formulations for our $p-$value:
  - $2[1 - F_{t_{35}}(2)]$
  - $F_{t_{35}}(-2) + [1 - F_{t_{35}}(2)]$
- Using a computer software, we compute this to be 0.053. (On an exam, you would be asked to simply leave your answer in terms of the $t-$CDF.)

Using $p-$Values

- So, what do we do with $p-$values?
- We reject the null whenever our $p-$value is less than the level of significance.
- So, for example, if we used a 5% level of significance in our previous commute-time example, we would fail to reject the null since $0.053 > 0.05$.
- Note, crucially, that our decision is very dependent on the level of significance - for example, if we had instead used a 0.1 level of significance then we would have rejected the null in the previous example!

Critical Values vs. $p-$Values

- This leads fairly nicely into a discussion on which framework - critical values or $p-$values - is preferred.
- I'd like to stress that they are *equivalent* frameworks!
- Using $p-$values, in a sense, simply shifts the oenus of having to pick a level of signifcance to the reader. Using critical values allows us, as experimenters, to pick the level of significance and formulate our decisions.
- You will see both critical values and $p-$values "in the wild," in your future statistics/data science careers, so it pays to be able to understand them both.