Hypothesis Testing

STA 032: Gateway to data science Lecture 21

Jingwei Xiong

May 19, 2023

Reminders/announcements

- Midterm 2 on May 24 at 9pm (in person, Wellman 6!)
    - Practice problems from homework
    - Read through the lecture notes
- Book problems: T/F and multiple choice questions on statistic
- R problems: T/F, multiple choice and code sort problems.

Recap

- More on confidence intervals:
  - Changing n and alpha
  - Simulation example
  - σ^2 unknown:
    - n large: $\left(\overline{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$
    - n small and X_i normal: $\left(\overline{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \overline{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$
  - Confidence interval for population proportion: $\left(\hat{p} - z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right)$

Recall: Sampling distributions

Knowing the sampling distribution can help us

- Estimate a population parameter as point estimate $\pm$ margin of error, where the margin of error is comprised of a measure of how confident we want to be and the sample statistic's variability.
- **Test whether a population parameter is equal to some value, by evaluating how likely it is that we have obtained the observed sample statistic, if the population parameter is indeed that value. (Coming soon: hypothesis testing)**

Today

- Introduction to hypothesis testing
  - Framework
  - Errors from hypothesis tests
- Hypothesis test for population mean

Example: Honor Court

Suppose a case of suspected cheating is brought to a university Honor Court. There are two opposing claims.

Student: I did not cheat on the exam.

Professor: The student did cheat on the exam.

This Honor Court assumes students are innocent until proven guilty. The professor must provide evidence to support their claim. They explain that they had two different versions of the exam and that the student on three separate problems used numbers from the *other* version of the exam.

The honor court members agree that this would be *extremely unlikely* if it were true that the student did not cheat. They agree that the professor's evidence is very strong and conclude that there is sufficient evidence to reject the student's claim that they did not cheat on the exam.

Hypothesis Testing Framework

Steps in hypothesis testing:

- Start with two claims about the population (often about the value of a population parameter or about some potential association between two variables in the population). Call them claim 1 and claim 2. (**They are competing**!)
- Choose a sampling strategy, draft an analysis plan, collect data, and summarize data
- Figure out how likely it is to see data like what we got, or more extreme results, if claim 1 is true.
- If our data would have been **unlikely** if claim 1 were true, then we reject claim 1 and deem claim 2 worthy of further study. Otherwise, we cannot reject claim 1.

Note: we never "accept" claim 1. The test only tells us if we have sufficient evidence to reject claim 1. The outcomes are (1) reject claim 1, (2) fail to reject claim 1.

> The power of the test is not symmetric on claim 1 and claim 2. We can only conclude whether to reject claim 1 (it's false) or fail to reject claim 1 (indecisive conclusion).

Hypothesis Testing Framework

- Claim 1 is called the **null hypothesis**, denoted H_0. Claim 1 states "nothing unusual is happening".
- Claim 2 is called the **alternative hypothesis**, denoted H_1 or H_A. Claim 2 challenges claim 1, and states that there is "something going on" or some relationship, or that our initial guess is incorrect.
    - If you want to make some new conclusion, **put it to claim 2**!
- In this example:
  H_0: Student did not cheat
  H_A: Student cheated
- Gather data (student have numbers from the other version of the exam in three separate problems.)
- Assess **how likely** we are to observe data, or more extreme results, **if** H_0 **were true** (**p-value**) (Student did not cheat)
- In Honor Court example, if the student did not cheat, it is very unlikely that the student would have numbers from the other version of the exam in three separate problems.

Example: Ultra Low Dose Contraceptives

- A certain ultra-low dose oral contraceptive pill is supposed to contain 0.02 mg of estrogen.
- If the dose is higher, the user may risk side effects better avoided, and if the dose is lower, the user may get pregnant.
- The manufacturer wishes to check whether the mean concentration in a large shipment is the needed 0.02 mg or not.
- A random sample of $n = 500$ pills is tested, and the sample mean concentration is 0.017 mg with a sample standard deviation of 0.008 mg.
- Is this sufficient evidence that the mean concentration is not 0.02 mg? What about if our sample mean was 0.019 mg? Hypothesis testing allows us to answer these questions.
- How do we set this up using our hypothesis testing framework?

Example: Ultra Low Dose Contraceptives

- State the claims
  - **Claim 1**: The shipment is consistent with a population mean of 0.02 mg estrogen. $H_0 : \mu = 0.02$
  - **Claim 2**: The shipment is not consistent with a population mean of 0.02 mg estrogen. $H_A : \mu \neq 0.02$
- Choose sampling strategy, design analysis plan, collect, and analyze data
  - Sample 500 pills at random and use a hypothesis test to evaluate whether they are consistent with a population with mean 0.02 mg estrogen
  - 500 pills were sampled with a sample mean $\bar{x} = 0.017$ and sample standard deviation $s = 0.008$
- Assess **how likely** we are to observe $\bar{x} = 0.017$, or more extreme results, if H_0 were true
  - We'll learn how to do this shortly, but for now assume the probability of getting a result like ours (or even more extreme) is just 0.01 if Claim 1 is true.

Example: Ultra Low Dose Contraceptives

- Conclusion: A probability of .01 is pretty unlikely. Reject claim 1.
- There is sufficient evidence to reject the null hypothesis that $\mu = .02$, that the population mean amount of estrogen is 0.02mg.
- Implications: the manufacturing procedure may not be consistent with one that produces pills at the required 0.02 mg dose. Further evaluation is needed to verify the dosage is sufficient to prevent breakthrough pregnancies.
- Suppose the probability of getting a result like ours, or more extreme, was relatively large, say 0.20 instead of 0.01.
- In this case, we would **fail to reject** Claim 1 and state that we do not have evidence to disprove Claim 1.

Two comments

1. We would not say that evidence leads us to accept Claim 1.
    - The concept is the same as that in the US judicial system
    - Defendants are "innocent until proven guilty"
    - We find someone "guilty" or "not guilty"
    - We do not proclaim someone "innocent"; we do not know if they are in fact innocent, but there is insufficient evidence to say they are guilty.
2. Hypothesis testing does not tell us the probability that Claim 1 is true.
    - We assumed claim 1 was true before we did our calculation, and thus we calculated a probability about data like ours or more extreme than ours under that assumption.
    - That's the conditional probability: **observing our data or more extreme conditioned on claim 1 is true**!

Hypothesis Testing Steps Again

1. State Claim 1 and Claim 2. Claim 1 states "nothing unusual is happening" and Claim 2 challenges it.

2. Finalize data collection and analysis plans, collect relevant data, and summarize it.

3. Assess how surprising it would be to see data like that, or even more extreme data, *if Claim 1 is really true*.

4. Draw conclusions.

Step 1: Examples

What are H_0 and H_A in each case?

1. Researchers would like to know whether a new intervention for informing children in developing countries of their HIV status is associated with different mental health quality of life.

2. Researchers would like to know if lead levels in the water from Flint exceed the EPA action level of 15 ppb.

3. The World Health Organization would like to know if the prevalence of the omicron variant this month is the same as last month.

Step 2

Step 2 is to make a plan for data collection and analysis, take a sample, and summarize the data.

- This involves defining a **test statistic** (T), which is a **random variable** that is computed from the data, e.g., a sample mean ($\overline{X}$). This choice depends on the question of interest, type of data (e.g., categorical or continuous) as well as the distribution of the data.
- We need to know the distribution of the test statistic (T) under the null hypothesis
- The type of test depends on this distribution, e.g., if our test statistic can be approximated by a normal distribution, we will use a **Z-test**

Step 3: Assess results

- Step 3 involves assessing the evidence in our data by calculating the probability of "getting data like ours, or more extreme than ours," if H_0 is actually true.
- From Step 2, we have the distribution of T
- In Step 3, we compute the value of the test statistic (t) based on the data collected, and calculate the probability of getting a test statistic that is equally or more extreme than the one that we got, based on the distribution of the test statistic (T)
- This is a **conditional probability** (conditional on H_0 being true), called a **p-value**.
- The **p-value** is the probability of getting a specific test statistic (t) based on the data, or one more extreme, if H_0 were true

Step 4: Draw conclusions

- Given the data and the p-value we calculated in Step 3, we make a decision about whether to reject claim 1
- Recall that the two possible outcomes are (1) Reject claim 1, and (2) fail to reject claim 1.
- Generally, we reject Claim 1 when the likelihood of seeing our data (or more extreme data) when Claim 1 is true would be relatively small.
- What qualifies as "relatively small" depends on the **significance level** of the test

Significance level

- We defined the significance level, α, when discussing confidence intervals:
  - Confidence level = $100(1-\alpha)$%, i.e., a 95% confidence interval will need $\alpha = .05$
  - P(CI contains true parameter) = $1-\alpha$.
- The significance level is also an important ingredient in a hypothesis test
  - It defines the tolerable **Type I error**: the probability of rejecting H_0 **when** H_0 **is actually true**.
  - The statistical property that is needed is P(reject H_0 | H_0 true) $= \alpha$
  - This means that when the null hypothesis is true, if we repeat the experiment a large number of times, we would expect to make the wrong decision only α (e.g., 5%) of the time

Decision rule

- The **decision rule**: reject H_0 if p-value $< \alpha$
    - We will demonstrate (in the next class) that this produces the required property that P(reject H_0 | H_0 true) $= \alpha$
- When the p-value $< \alpha$ you often see results described as "statistically significant."
- When the p-value is $\geq \alpha$ then we say we have insufficient evidence to reject H_0.
- There is a direct correspondence between hypothesis tests and confidence intervals; we will talk about this in the next lecture

Errors from hypothesis tests

A cat is on trial. Did it commit the crime? Evidence is presented as part of the trial by jury.

	Truly Innocent	**Truly Guilty**
Jury: Not Guilty	✓	☠
Jury: Guilty	☠	✓

Errors from hypothesis tests

During the trial, we assume the cat is innocent unless proven guilty. The right decisions are finding an innocent cat "not guilty" and convicting a guilty cat. We could make two mistakes: we could wrongly convict an innocent cat, or we could declare a guilty cat "not guilty."

	Truly Innocent	**Truly Guilty**
Jury: Not Guilty	✓	☠
Jury: Guilty	☠	✓

In a hypothesis testing framework: H_0 : Cat is innocent vs. H_A: Cat is guilty

	H_0 **true**	H_A **true**
Decision: Do Not Reject H_0	✓	☠
Decision: Reject H_0	☠	✓

Errors from hypothesis tests

Suppose we wish to test that the population mean equals some value, say μ_0.

Test of $H_0 : \mu = \mu_0$

	Truly $\mu = \mu_0$	**Truly** $\mu \neq \mu_0$
Decision: Do Not Reject H_0	✓	Type II ☠
Decision: Reject H_0	Type I ☠	✓

Type I error: rejecting H_0 when it is really true

- With significance level $\alpha = 0.05$, we have at most 5% to make a Type I error.

Type II error: not rejecting H_0 when it is false

Errors from hypothesis tests

- **Type I error**: rejecting H_0 when it is really true
- **Type II error**: not rejecting H_0 when it is false
- α is the maximum allowable Type I error rate.
- Type I errors involve incorrectly challenging the status quo, and are typically viewed as more severe than Type II errors. We specify α at the design stage of the study and use it in making decisions with hypothesis tests.
- The probability of making a Type II error is related to the **power** = **1 - P(type II error)** of the test, which is the probability of rejecting H_0 when H_0 is false (out of scope for this class).

Recall: Hypothesis Testing Framework

Steps in hypothesis testing:

- Start with two claims about the population, H_0 and H_A
- Choose a sampling strategy, collect data, and summarize data, i.e., define test statistic and compute statistic from the data
- Figure out how likely it is to see data like what we got, or more extreme results, if H_0 is true, i.e., compute p-value
- Draw conclusions, i.e., if our data would have been unlikely if H_0 were true, then reject H_0. Otherwise, do not reject H_0.

Hypothesis Testing for the Population Mean (population variance known)

Say X_i has mean μ and variance 4.

- **Step 1**: Start with two claims about the population

H_0: $\mu = 20$
H_A: $\mu \neq 20$

- **Step 2**: Choose a sampling strategy, collect data, and summarize data

Test statistic: By CLT, $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$ when n large. Collect a sample with $n = 100$. Under H_0, $Z = \frac{\overline{X} - 20}{2/\sqrt{100}} \approx N(0,1)$. From the sample, we get $\bar{x} = 21$

- **Step 3**: Figure out how likely it is to see data like what we got, or more extreme results, if H_0 is true.

To get the value of the test statistic based on our data (z), simply substitute $\bar{x} = 21$ to get $z = \frac{21-20}{2/\sqrt{100}} = 5$

Hypothesis Testing for the Population Mean (population variance known)

- **Step 3** (continued): Figure out how likely it is to see data like what we got, or more extreme results, if claim 1 is true.

Probability under H_0 of getting data like what we got, or more extreme, is $P(|Z| \geq |z|) = P(Z \geq 5 \text{ or } Z \leq -5)$.

```
2*pnorm(-5)
```

```
[1] 5.733031e-07
```

`2*pnorm(-5)` is very small (on the order of 10^{-7}).

- **Step 4**: If our data would have been unlikely if H_0 were true, then reject H_0. Otherwise, do not reject H_0

Using a significance level of $\alpha = .05$, $P(|Z| \geq 5) < \alpha$, so reject H_0. At a 5% level, there is sufficient evidence to reject the null hypothesis that $\mu = 20$.

Hypothesis Testing for the Population Mean

Say X_i has mean μ and standard deviation σ. The test statistic we will use is $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$. By CLT, $Z \approx N(0,1)$ when n large.

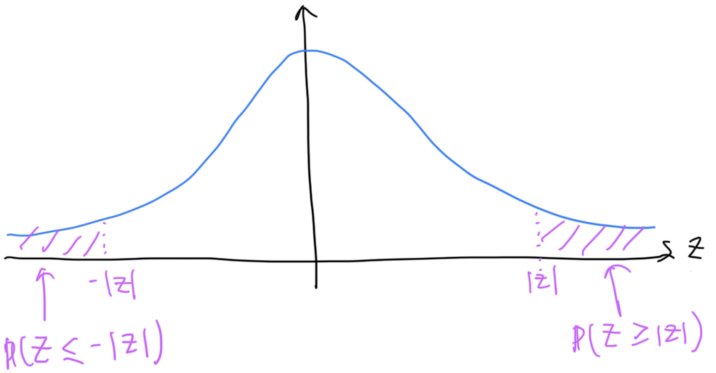
H_0: $\mu = \mu_0$
H_A: $\mu \neq \mu_0$

Under H_0, $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \approx N(0,1)$

Value of test statistic: $z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$

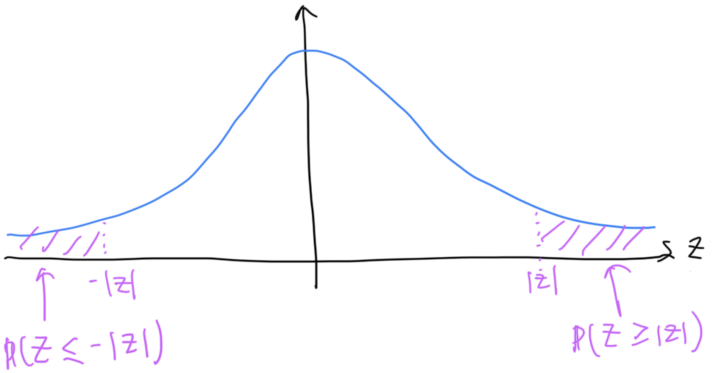
Decision rule: reject H_0 if $P(|Z| \geq |z|) = P(Z \geq |z| \text{ or } Z \leq -|z|) < \alpha$

More about p-values



- Note that all computations are done assuming that H_0 is true, i.e., to be precise, the decision rule is reject H_0 if the p-value $P(|Z| \geq |z| \mid H_0) = P(Z \geq |z| \text{ or } Z \leq -|z| \mid H_0) < \alpha$
- The blue distribution is the distribution under the null hypothesis
- $P(|Z| \geq |z|) = P(Z \geq |z|) + P(Z \leq -|z|)$ (shaded area)

More about p-values



- The value of the **test statistic** is z (value on horizontal axis)
- **p-value** = $P(|Z| \geq |z|)$ under H_0 (for a two-sided test; more details coming). It is the probability of getting a result as extreme as what we got, if H_0 were true.
- Recall the decision rule: reject H_0 if $P(|Z| \geq |z|) = P(Z \geq |z| \text{ or } Z \leq -|z|) < \alpha$. Alternatively, the p-value can be interpreted as the smallest significance level that we would reject H_0.

More about p-values

- The p-value is the probability of getting data like ours or more extreme data if H_0 were true
- Common misinterpretation: "p-value is the probability that H_0 is true". The p-value is calculated *assuming* that H_0 is true. It cannot be used to tell us how likely it is that assumption is correct.
- **Decision rule**: reject H_0 if p-value $< \alpha$
    - We will demonstrate that this produces the required property that P(reject H_0 | H_0 true) $= \alpha$

More about p-values: simulation example

Setup: $X_i \sim N(\mu, 4^2)$

H_0: $\mu = \mu_0 = 20$
H_A: $\mu \neq 20$

Under H_0, $Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

Value of test statistic: $z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$

Decision rule: reject H_0 if $P(|Z| \geq |z|) = P(Z \geq |z| \text{ or } Z \leq -|z|) < \alpha$

More about p-values: simulation example

1. Get 1000 draws from $N(20, 4^2)$. This is the distribution when H_0 is true.
2. Calculate the sample mean, $\bar{x}$, and the test statistic, $z = \frac{\bar{x}-20}{4/\sqrt{1000}}$
3. Calculate the p-value, $P(|Z| \geq |z|)$
4. Repeat steps (1)-(3) 10,000 times

```
set.seed(0)
myDraws <- t(replicate(10000, rnorm(1000, 20, 4)))
sampleMeans <- rowMeans(myDraws)
testStat <- (sampleMeans - 20)/(4/sqrt(1000))
pValues <- 2*pnorm(abs(testStat), lower.tail = FALSE)
```

More about p-values: simulation example

- The rejection rule is reject if p-value $< \alpha$. What proportion out of the 10,000 experiments should we expect to reject H_0?

This is given by the significance level α. When $\alpha = .05$, we should reject 5% of the time to produce the required property that P(reject H_0 | H_0 true) $= \alpha$ (more accurately, this is $\leq \alpha$, meaning that when H_0 is true we mistakenly reject 5% or less when $\alpha = .05$)

```
sum(pValues < .05)
```

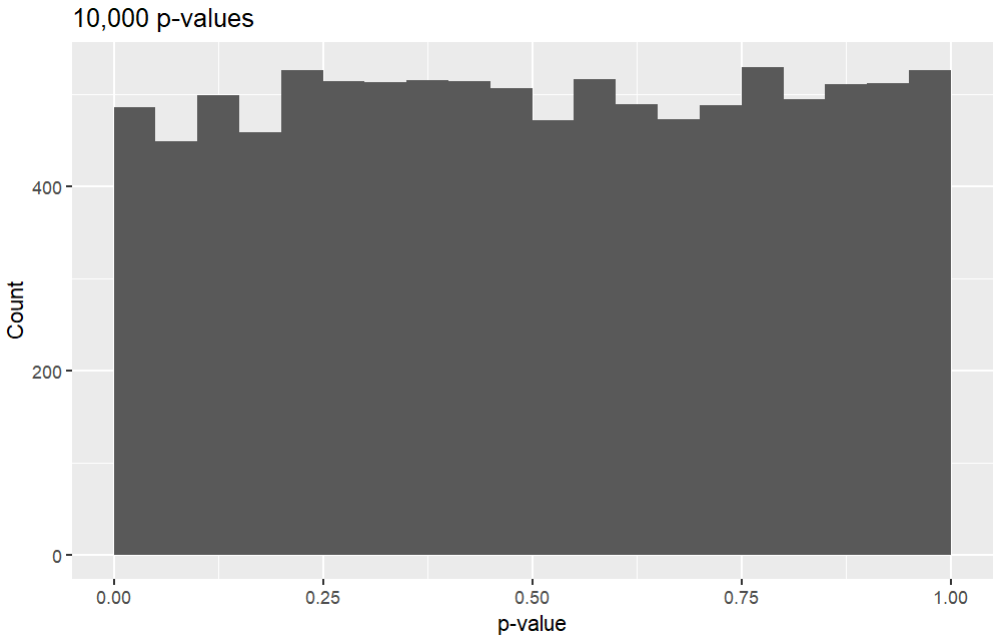
```
[1] 486
```

```
mean(pValues < .05)
```

```
[1] 0.0486
```

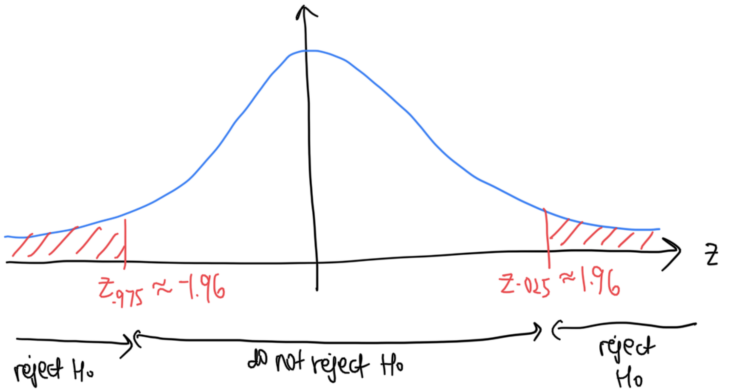
More about p-values: simulation example

```
ggplot(data.frame(pValues), aes(x = pValues)) +
  geom_histogram(binwidth = .05, boundary = 0) +
  labs(title = "10,000 p-values",
       x = "p-value",
       y = "Count")
```



- Note: the p-value is actually a random variable and it follows a uniform distribution (out of scope for this class)

Critical value approach



- The **rejection region** is $|z| > z_{\alpha/2}$ or $|z| > 1.96$ when $\alpha = .05$. This is a **portion of the x-axis**.
- The boundaries of the rejection region are called **critical values**.
- **Significance level** is the probability over the rejection region, the red area: $P(|Z| > z_{\frac{\alpha}{2}}) = \alpha$

Hypothesis Testing for the Population Mean (σ known)

Say X_i has mean μ and standard deviation σ. The test statistic we will use is $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$. By CLT, $Z \approx N(0,1)$ when n large.

H_0: $\mu = \mu_0$
H_A: $\mu \neq \mu_0$

Under H_0, $Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \approx N(0,1)$

Value of test statistic: $z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$

Decision rule:

- p-value approach: reject H_0 if $P(|Z| \geq z) = P(Z \geq |z| \text{ or } Z \leq -|z|) < \alpha$
- Critical value approach: reject if $|z| > z_{\alpha/2}$ or $|z| > 1.96$ when $\alpha = .05$

Hypothesis Testing for the Population Mean (σ unknown, n > 50)

Say X_i has mean μ and standard deviation σ. The test statistic we will use is $Z = \frac{\overline{X}-\mu}{S/\sqrt{n}}$. $Z \approx N(0,1)$ when n large. (Here notice that σ has been replaced by S)

H_0: $\mu = \mu_0$
H_A: $\mu \neq \mu_0$

Under H_0, $Z = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \approx N(0,1)$ (Here notice that σ has been replaced by S)

Value of test statistic: $z = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$ (replace σ by s)

Decision rule:

- p-value approach: reject H_0 if $P(|Z| \geq z) = P(Z \geq |z| \text{ or } Z \leq -|z|) < \alpha$
- Critical value approach: reject if $|z| > z_{\alpha/2}$ or $|z| > 1.96$ when $\alpha = .05$

Hypothesis Testing for the Population Mean (σ unknown, X_i i.i.d normal)

Say X_i are i.i.d normally distributed mean μ and standard deviation σ. The test statistic we will use is $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$. $T \sim t_{n-1}$. (Here notice that σ has been replaced by S)

H_0: $\mu = \mu_0$
H_A: $\mu \neq \mu_0$

Under H_0, $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$ (Here notice that σ has been replaced by S)

Value of test statistic: $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$ (replace σ by s)

Decision rule:

- p-value approach: reject H_0 if $P(|T| \geq t) = P(T \geq |t| \text{ or } T \leq -|t|) < \alpha$
- Critical value approach: reject if $|T| > t_{n-1,\alpha/2}$.

```
-qt(0.05/2, 20-1)
```

```
[1] 2.093024
```

Summary

- Hypothesis testing framework
  - Null and alternative hypotheses
  - Test statistics
  - p-values
  - Significance level
- Errors from hypothesis tests
  - Type I error
  - Type II error and power
- Hypothesis test for population mean