Chapter 7: Hypothesis Testing

DSCC 462
Computational Introduction to Statistics

Anson Kahng
Fall 2022

Hypothesis Testing

- We want to draw conclusions regarding population parameters using information contained in a sample of observations
- We made a confidence interval to determine a reasonable range of values that will occur with a given confidence level
- Can we now determine how probable it is to see the results in our sample given a hypothesized value of the population parameter?

Hypothesis Testing: Scenario

- A grande Starbucks drink is claimed to contain 16 oz of liquid on average
- Suppose we purchase a grande-sized coffee from Starbucks and measure 15.8 oz of coffee
- If we accept that our measuring error is normally distributed with mean 0 and standard deviation 0.4, is the observed coffee consistent with Starbucks' claims?
- What if we had measured 15.0 oz of coffee? Would we be more suspicious of Starbucks' claims?

Hypotheses

- Hypothesis testing is a procedure based on sample evidence and probability that is used to test events regarding population parameters
- We begin hypothesis testing by claiming that the population parameter of interest (usually the mean) is some given value, μ_0
- This statement is called the *null hypothesis*, H_0
- The null hypothesis is that of “no change” (or status quo)
- We believe the null hypothesis to be true unless *overwhelming evidence* exists to the contrary (“innocent until proven guilty”)

Hypotheses

- The *alternative hypothesis*, H_1 (sometimes denoted H_A), is a second statement that contradicts H_0
  - This is the hypothesis for which the investigator is trying to gather evidence in favor of
- Either H_0 or H_1 must be true (mutually exclusive, exhaustive)
- We need overwhelming evidence to conclude that H_1 is true

Hypotheses: Example

- In our Starbucks example, we are interested in determining if a grande coffee is in fact 16 oz on average
- Let μ be the average amount of coffee in ounces
- $H_0 : \mu = 16$ oz – the grande coffee is 16 oz on average
- $H_1 : \mu \neq 16$ oz – the grande coffee is not 16 oz on average

Types of Hypotheses

- There are three types of hypotheses we can consider:
- Lower-tailed (true mean is less than hypothesized mean):
  - $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$
- Upper-tailed (true mean is greater than hypothesized mean):
  - $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$
- Two-sided (true mean is not equal to the hypothesized mean):
  - $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$

Test Statistic

- Once we formulate our hypotheses, we need to draw a random sample of size n from the population of interest
- Calculate the sample statistic and compare to the population parameter
  - Compare $\bar{x}$ to the postulated μ_0
- Is $\bar{x}$ different enough from μ_0 to conclude that H_1 is true?
  - Calculate *test statistic*

Test Statistic

- Use test statistics to determine the probability of seeing a sample mean as extreme or more extreme than the one observed, given that the null hypothesis is true

- Relies on the sampling distribution of the test statistic

- Recall that if X has mean μ_0 and known variance σ^2, then by the CLT, $\bar{X} \sim N\left(\mu_0, \sigma/\sqrt{n}\right)$ for $n \geq 30$

- For a *z-test*, our test statistic is $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Significance

- If there is evidence that the sample could not have come from a population with the hypothesized parameter, we reject the null hypothesis
  - Given that H_0 is true, the probability of obtaining a sample statistic as extreme or more extreme than the observed statistic is sufficiently small
  - In this case, our data is more supportive of H_1
  - Conclude that the population parameter could not be μ_0
- Such a test is *statistically significant*
- Intuition: "If the null hypothesis were true, our observations are extremely unlikely. Therefore, this is evidence that the null hypothesis is not true"

Significance

- If it seems reasonable (i.e., not extremely unlikely) that the sample came from a population centered at the hypothesized mean, then we fail to reject the null hypothesis
- We do not *accept* H_0, we merely fail to reject it
- It is still possible that the population parameter does not equal μ_0, but the random sample we selected does not provide enough evidence to confirm this
    - This can be the case if the sample is too small
- Intuition: “Not enough evidence to disprove the null hypothesis”

Significance

- Calculate a probability to determine how unlikely it is to see your sample results if the null hypothesis is true
  - *p-value*
- If that probability is less than some *pre-specified* ***significance level***, α, then reject the null hypothesis ("sufficiently unlikely")
  - Typically, let $\alpha = 0.05$: Reject H_0 when the chance that the sample could have come from a population with mean μ_0 is less than or equal to 5%
- If $p \leq \alpha$, we reject H_0 (i.e., if the p-value is low, we reject the null hypothesis)
- If $p > \alpha$, we fail to reject H_0

Significance Level (α)

- Choosing the significance level α allows us to specify the “power” of the test
- If we want to be more conservative, we can choose $\alpha = 0.01$
- To be less conservative, choose $\alpha = 0.1$
- Note: We must specify α **before** the test is carried out
    - Otherwise, we may do science in reverse (fit hypotheses to results)

Illustration

p-values for z-tests

- We calculate our p-value as follows, for each of the three types of tests (*z-tests*):
- One-sided, lower-tailed hypothesis ($H_1 : \mu < \mu_0$):
  - `pnorm(z)`
- One-sided, upper-tailed hypothesis ($H_1 : \mu > \mu_0$):
  - `1-pnorm(z)`
- Two-sided hypothesis ($H_1 : \mu \neq \mu_0$):
  - If $z \leq 0$: `2*pnorm(z)`
  - If $z > 0$: `2*(1-pnorm(z))`

t-tests

- When σ^2 is also unknown, we substitute the sample variance s^2 and use the t distribution instead of the normal distribution
- The t-statistic is $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- This t-statistic has a t distribution with $n - 1$ degrees of freedom
- We determine the probability of seeing a test statistic t as extreme or more extreme as the one observed via a *t-test*

p-values for t-tests

- We calculate our p-value as follows, for each of the three types of tests (*t-tests*):
- One-sided, lower-tailed hypothesis ($H_1 : \mu < \mu_0$):
  - `pt(t,df)`
- One-sided, upper-tailed hypothesis ($H_1 : \mu > \mu_0$):
  - `1-pt(t,df)`
- Two-sided hypothesis ($H_1 : \mu \neq \mu_0$):
  - If $z \leq 0$: `2*pt(t,df)`
  - If $z > 0$: `2*(1-pt(t,df))`

Conducting a Hypothesis Test

1. Check the conditions required for the validity of the test
2. Define the parameter of interest in the context of the problem
3. State the desired significance level
4. State the null hypothesis
5. State the alternative hypothesis
6. Determine the proper test to use, and calculate the test statistic
7. Calculate the p-value or critical value
8. Make “reject/fail to reject” decision
9. State your conclusion in the context of the problem

Example: One-sided z-test

- Let's return to our Starbucks example where we want to test whether grande coffees actually come with less than 16 oz of coffee
- We buy 40 grande coffees from Starbucks and find that the average amount of coffee is 15.8 oz
- Recall that we know that $\sigma = 0.4$ oz
- Is there evidence that grande coffees are under-filled at the $\alpha = 0.05$ significance level?

Example: One-sided z-test

1. Check the conditions:
2. Parameter of interest:
3. Significance level:
4. Null hypothesis:
5. Alternative hypothesis:
6. Which test and test statistic:
7. p-value:
8. Accept or reject H_0?
9. Conclusion:

Example: One-sided z-test

1. Check the conditions: $n = 40 \geq 30$, so normality assumption holds (CLT)
2. Parameter of interest: μ = mean amount of coffee in a grande cup
3. Significance level: $\alpha = 0.05$
4. Null hypothesis: $H_0 : \mu \geq 16$ oz
5. Alternative hypothesis: $H_1 : \mu < 16$ oz
6. Which test and test statistic: σ is known, so z-test with $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{15.8 - 16}{0.4/\sqrt{40}} = -3.16$
7. p-value: $p = \Pr(Z < -3.16)$ = `pnorm(-3.16)` = 0.0008
8. Reject H_0
9. Conclusion: There is sufficient evidence to conclude that less than 16 oz is being poured into grande cups

Example: Two-sided z-test

- The average weight for men in 1960 was 166.3 lbs
- In 2002, 30 men were sampled and their average weight was 191 lbs
- Assume $\sigma = 50$ lbs is known
- Do the data suggest that the average weight of men is significantly different in 2002 as compared to 1960 at the $\alpha = 0.05$ significance level?

Example: Two-sided z-test

1. Check the conditions:
2. Parameter of interest:
3. Significance level:
4. Null hypothesis:
5. Alternative hypothesis:
6. Which test and test statistic:
7. p-value:
8. Accept or reject H_0?
9. Conclusion:

Example: Two-sided z-test

1. Check the conditions: $n = 30 \geq 30$, so the CLT applies
2. Parameter of interest: μ = average weight of men in 2002 (lbs)
3. Significance level: $\alpha = 0.05$
4. Null hypothesis: $H_0 : \mu = 166.3$ lbs
5. Alternative hypothesis: $H_1 : \mu \neq 166.3$ lbs
6. Which test and test statistic: σ is known, so z-test with $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{191 - 166.3}{50/\sqrt{30}} = 2.706$
7. p-value: $p = 2 \cdot \Pr(Z > 2.706)$ = `2*(1-pnorm(2.706))` = 2*(0.0034) = 0.007
8. Reject H_0
9. Conclusion: There is sufficient evidence to conclude that the average weight of men has significantly changed between 1960 and 2002

Example: Two-sided t-test

- The average US heart rate is 71.2 beats per minute (bpm)
- We sample 40 US Olympic athletes
- The average heart rate for this sample is $\bar{x} = 60.9$ bpm with a sample standard deviation of $s = 34.2$ bpm
- The underlying distribution of US Olympic athlete heart rates is approximately normal with an unknown mean μ and unknown standard deviation σ
- Does the average heart rate for US Olympic athletes differ from that of the general American population at the $\alpha = 0.05$ significance level?

Example: Two-sided t-test

1. Check the conditions:
2. Parameter of interest:
3. Significance level:
4. Null hypothesis:
5. Alternative hypothesis:
6. Which test and test statistic:
7. p-value:
8. Accept or reject H_0?
9. Conclusion:

Example: Two-sided t-test

1. Check the conditions: We have a normal population
2. Parameter of interest: μ = average heart rate for US Olympic athletes
3. Significance level: $\alpha = 0.05$
4. Null hypothesis: $H_0 : \mu = 71.2$ bpm
5. Alternative hypothesis: $H_1 : \mu \neq 71.2$ bpm
6. Which test and test statistic: σ is unknown, so t-test with $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{60.9 - 71.2}{34.2/\sqrt{40}} = -1.90$
7. p-value: $p = 2 \cdot \Pr(T < -1.90)$ = `2*pt(-1.90,df=39)` = 2*(0.0324) = 0.0658
8. Fail to reject H_0
9. Conclusion: There is insufficient evidence to conclude that the average heart rate of US Olympic athletes is significantly different from the average heart rate of all Americans

Example: One-sided t-test

- The national average MCAT score is 500 (range: 472-528)
- The University of Rochester believes its students score better, on average, than the rest of the nation
- A sample of 52 U of R medical students is taken
- The average scores for these students is 516, with a sample standard deviation of 18
- Do U of R medical students score higher, on average, than the rest of the nation? Evaluate at the $\alpha = 0.05$ significance level

Example: One-sided t-test

1. Check the conditions:
2. Parameter of interest:
3. Significance level:
4. Null hypothesis:
5. Alternative hypothesis:
6. Which test and test statistic:
7. p-value:
8. Accept or reject H_0?
9. Conclusion:

Example: One-sided t-test

1. Check the conditions: $n = 52 \geq 30$, so CLT applies
2. Parameter of interest: μ = average MCAT scores of U of R medical students
3. Significance level: $\alpha = 0.05$
4. Null hypothesis: $H_0 : \mu \leq 500$
5. Alternative hypothesis: $H_1 : \mu > 500$
6. Which test and test statistic: σ is unknown, so t-test with $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{516 - 500}{18/\sqrt{52}} = 6.41$
7. p-value: $p = \Pr(T > 6.41)$ = `1-pt(6.41,df=51)` = 2 x 10^{-8}
8. Reject H_0
9. Conclusion: There is sufficient evidence to conclude that U of R medical students score significantly higher on the MCAT than students nationwide

Hypothesis Tests vs. Confidence Intervals

Confidence Intervals and Hypothesis Tests

- Confidence intervals for sample means are mathematically equivalent to hypothesis tests
- For a two-sided z-test, any value of z that lies between -1.96 and 1.96 would result in a p-value greater than 0.05
  - In this case, the null hypothesis would not be rejected at $\alpha = 0.05$
- Any value of z that lies outside (-1.96, 1.96) would result in a p-value less than 0.05, and thus we would reject H_0
- We say that -1.96 and 1.96 are the *critical values* of the test statistic at the $\alpha = 0.05$ significance level (for 2-sided z-tests)
- Conversely, we fail to reject a null hypothesis at $\alpha = 0.05$ if μ_0 falls within the 95% confidence interval for μ
  - We reject a null hypothesis at $\alpha = 0.05$ if μ_0 lies outside the 95% confidence interval for μ

Confidence Intervals and Hypothesis Tests

- Let's revisit the men's weights example: $\mu_0 = 166.3$, $n = 30$, $\bar{x} = 191$, $\sigma = 50$
- A two-sided z-test at $\alpha = 0.05$ is equivalent to a two-sided 95% confidence interval
- Critical values for z are $z = \pm 1.96$
- Therefore, the two-sided 95% confidence interval is

$$\begin{aligned} CI &= 191 \pm 1.96 \frac{50}{\sqrt{30}} \\ &= 191 \pm 17.89 \\ &= (173.11, 208.89) \end{aligned}$$

- Since 166.3 falls outside this interval, we reject the null hypothesis and conclude that the average weight of men is not equal to 166.3 lbs

Confidence Intervals and Hypothesis Tests

- Confidence intervals and hypothesis tests can lead to the same conclusions
- However, the information provided by each is a bit different
- A confidence interval gives a reasonable range of values for μ based on a the uncertainty in our point estimate $\bar{x}$
- A hypothesis test helps us determine whether the postulated value of the mean is likely to be correct by providing a p-value
- Hypothesis tests are centered around a null hypothesis that we are interested in gathering evidence against in order to reject it in favor of our alternative supposition

Rejection Regions (Critical Values)

- Assume σ^2 is known (i.e., z-test)
- Let your test statistic be z
- For a two-sided z-test with $\alpha = 0.05$, our critical value is $z_{\alpha/2}$ = `qnorm(0.975)` = 1.96
  - If $|z| \geq z_{\alpha/2}$, reject H_0
  - If $|z| < z_{\alpha/2}$, fail to reject H_0
- For a one-sided (upper) z-test with $\alpha = 0.05$, our critical value is z_{α} = `qnorm(0.95)` = 1.645
  - If $z \geq z_{\alpha}$, reject H_0
  - If $z < z_{\alpha}$, fail to reject H_0
- For a one-sided (lower) z-test with $\alpha = 0.05$, our critical value is z_{α} = `qnorm(0.05)` = -1.645
  - If $z \leq z_{\alpha}$, reject H_0
  - If $z > z_{\alpha}$, fail to reject H_0

Rejection Region Example: One-sided z-test (lower)

- Back to Starbuck's example: $n = 40,\ \bar{x} = 15.8,\ \sigma = 40,\ \mu_0 = 16$
- $H_0 : \mu \geq 16$ oz, $H_1 : \mu < 16$ oz
- σ is known, so we use a z-test: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{15.8 - 16}{0.4/\sqrt{40}} = -3.16$
- Critical value: z_α = `qnorm(0.05)` = -1.645
- Since $-3.16 \leq -1.645$, reject H_0
- There is sufficient evidence to conclude that less than 16 oz of coffee is being poured into grande cups

Rejection Region Example: Two-sided z-test

- Back to men's weight example: $n = 30,\ \bar{x} = 191,\ \sigma = 50,\ \mu_0 = 166.3$
- $H_0 : \mu = 166.3$ lbs, $H_1 : \mu \neq 166.3$ lbs
- σ is known, so we use a z-test: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{191 - 166.3}{50/\sqrt{30}} = 2.706$
- Critical value: $z_{\alpha/2}$ = `qnorm(0.975)` = 1.96
- Since $2.706 \geq 1.96$, reject H_0
- There is sufficient evidence to conclude that the average weight of men has significantly changed between 1960 and 2002

Rejection Region Example: Two-sided t-test

- Back to US Olympic athletes' heart rates example:
  $n = 40,\ \bar{x} = 60.9,\ s = 34.2,\ \mu_0 = 71.2$

- $H_0 : \mu = 71.2$ bpm, $H_1 : \mu \neq 71.2$ bpm

- σ is unknown, so we use a t-test: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{60.9 - 71.2}{34.2/\sqrt{40}} = -1.90$

- Critical value: $t_{\alpha/2}$ = `qt(0.975,df=39)` = 2.023

- Since $|-1.90| < 2.023$, fail to reject H_0

- There is insufficient evidence to conclude that the average heart rate of US Olympic athletes is significantly different from the average heart rate of all Americans

Rejection Region Example: One-sided t-test (upper)

- Back to MCAT example: $n = 52,\ \bar{x} = 516,\ s = 18,\ \mu_0 = 500$
- $H_0 : \mu \leq 500, H_1 : \mu > 500$
- σ is unknown, so we use a t-test: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{516 - 500}{18/\sqrt{52}} = 6.41$
- Critical value: t_α = `qt(0.95,df=51)` = 1.675
- Since $6.41 \geq 1.675$, reject H_0
- There is sufficient evidence to conclude that U of R medical students score significantly higher on the MCAT than the nationwide average

Types of Error

- We can make two types of errors when performing a hypothesis test:

	$\mu = \mu_0$	$\mu \neq \mu_0$
Fail to reject	Correct	Incorrect (Type II)
Reject	Incorrect (Type I)	Correct

Example for two-sided test

Type I Error

- ***Type I*** error occurs if we reject a true null hypothesis ("false positive")
  - $H_0 : \mu = \mu_0$ is true, but we reject it
- Example: Send an innocent person to prison
- The chance that this happens is $\Pr($reject $H_0 \mid H_0$ is true$)$
- However, recall that $\alpha = \Pr($reject $H_0 \mid H_0$ is true$)$
- Thus, the significance level α is the probability of making a type I error
- We decide what α is for our test

Type II Error

- ***Type II*** error occurs if we fail to reject a false null hypothesis ("false negative")
    - $H_0 : \mu = \mu_0$ is false, but we fail to reject it
- Example: Let a guilty person go free
- The probability of making a type II error is denoted β
- $\beta = \Pr($do not reject $H_0 | H_0$ is false$)$
- Typically, we want β to be around 0.10 (or less)
- Holding all else constant, the smaller we make α, the larger β becomes
- Exact type II error depends on the *particular* alternative population mean μ_1

Type I and Type II Error, Illustrated

Types of Error: Example

- Setup: A pharmaceutical company has developed a cancer treatment and wants to know if it is effective
- H_0: The treatment is not effective
- H_1: The treatment is effective
- What are type I and type II errors in this context?

Types of Error: Example

- Setup: A pharmaceutical company has developed a cancer treatment and wants to know if it is effective
- H_0: The treatment is not effective
- H_1: The treatment is effective
- What are type I and type II errors in this context?
  - Type I: Reject H_0 when H_0 is true, which means concluding that the treatment is effective when it is not effective
  - Type II: Fail to reject H_0 when H_0 is false, which means deciding that the treatment is not effective when it really is effective

Power

- The *power* of a test is equal to $1 - \beta$
- The power is the probability of correctly rejecting the null hypothesis
    - Power = $\Pr(\text{reject } H_0 \,|\, H_0 \text{ is false})$
- Example: Convict the criminal who committed the crime
- Power must be computed for a particular alternative population mean μ_1

Power: Example

- Back to Starbucks (one-sided z-test): $n = 36,\ \alpha = 0.05,\ \sigma = 0.4$
- Suppose the true mean amount of coffee in a grande cup is actually 15.8 oz with a standard deviation of 0.4 oz
- $H_0 : \mu \geq 16$ oz, $H_1 : \mu < 16$ oz
- What is the value of β associated with a test of the null hypothesis $H_1 : \mu \geq 16$ oz?

Power: Example

- First, find the mean amount of coffee our sample must have in order for H_0 to be rejected (i.e., where is the cutoff?)

- One-sided z-test: z = `qnorm(0.05)` = -1.645

- Therefore, $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = -1.645 \rightarrow \bar{x} = 15.89$

- Interpretation: the null hypothesis will be rejected if our sample has a mean $\bar{x}$ that is less than or equal to 15.89 oz

    - If the sample mean is larger, we lack sufficient evidence to reject H_0

Power: Example

- Now, given this cutoff for rejecting the null hypothesis (15.89 oz), what is β?
- Answer: β is the probability of observing a sample mean greater than 15.89 given that the true population mean is 15.8 oz
- Let $\mu_1 = 15.8$ oz and determine what proportion of the distribution centered around this mean lies above 15.89 oz
- $z = \dfrac{15.89 - 15.8}{0.4/\sqrt{36}} = 1.35$
- Therefore, β = `1-pnorm(1.35)` = 0.0885
- Interpretation: The probability of failing to reject $H_0 = 16$ oz when the true population mean is $\mu_1 = 15.8$ oz is 0.0885
- Hence, the power of the test is $1 - \beta = 0.9115$

Power and μ_1

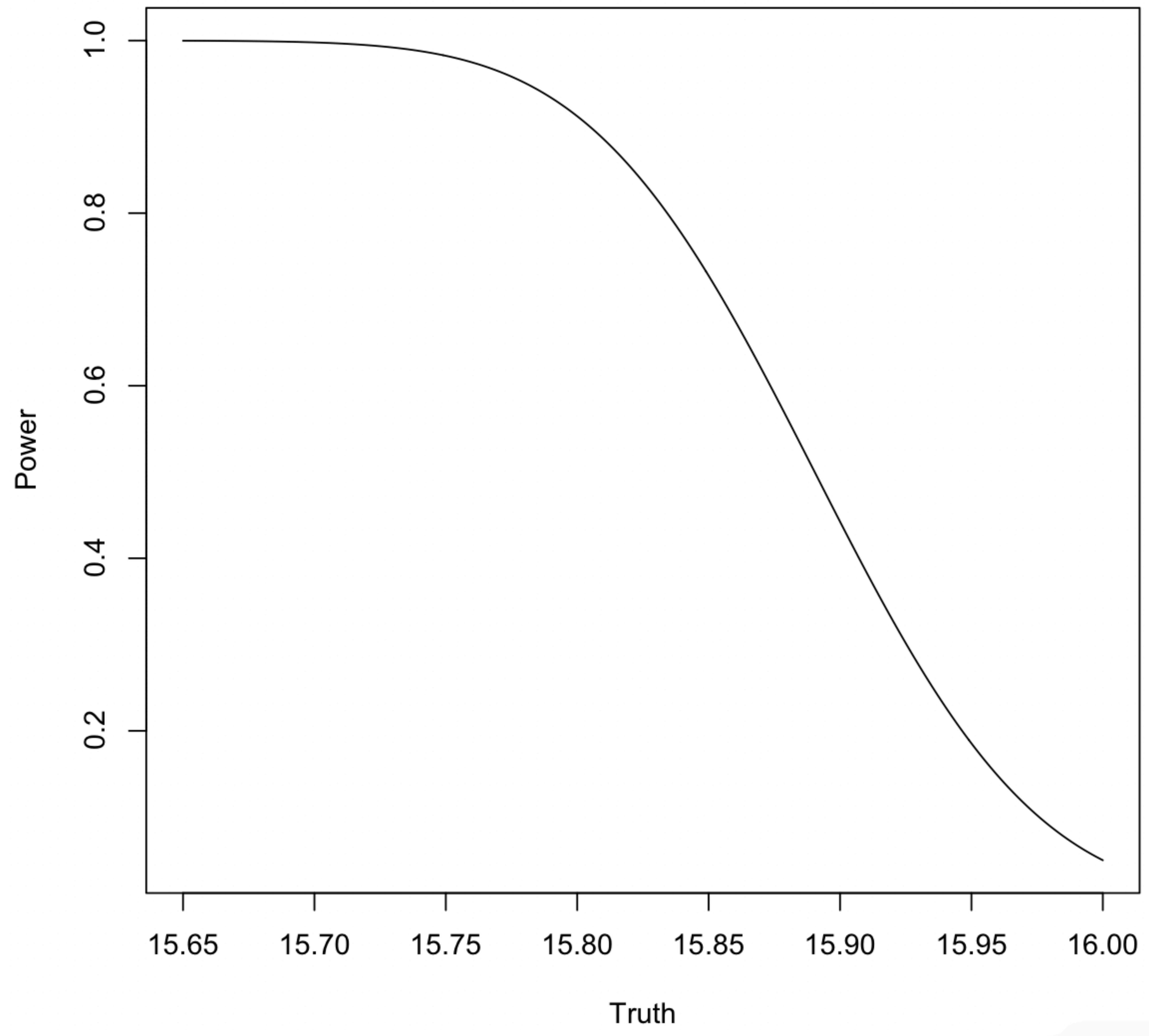
- To calculate β, note that we must have a specific value under the alternative to consider (i.e., a concrete value of μ_1)
- Under the null hypothesis, there are infinitely many options for what μ can be if the null hypothesis is rejected
- But, type II error is calculated for a single value of μ_1 that falls under the alternative hypothesis
- Different choices of the "truth" will lead to different values of β
- In general, the closer μ_1 is to μ_0, the harder it will be to reject H_0

Power Revisited

- In the Starbucks example ($n = 36,\ \alpha = 0.05,\ \sigma = 0.4,\ \mu_0 = 16,\ \mu_1 = 15.8$, one-sided lower z-test), we can more directly calculate power as follows:

$$\begin{aligned}
\text{Power} &= \Pr(\text{reject } H_0 \,|\, \mu = 15.8) \\
&= \Pr(\text{reject } \mu \geq 16 \,|\, \mu = 15.8) \\
&= \Pr(\overline{X} \leq 15.89 \,|\, \mu = 15.8) \\
&= \Pr(Z \leq 1.35) \\
&= 0.9115
\end{aligned}$$

Power Curve



Power
Truth
1.0
0.8
0.6
0.4
0.2
15.65
15.70
15.75
15.80
15.85
15.90
15.95
16.00

Hypothesis Testing: α and Power

- Ideally, we would like a test with small α and high power $(1 - \beta)$
- In practice, power less than 80% is considered insufficient and the test is deemed not worthwhile
- We can increase power (decrease β) by increasing α
- With larger sample sizes, we are less likely to commit either type of error (and thus, have higher power)
    - Intuition: Larger sample sizes = more sharply peaked distributions = less overlap in the two normal distributions = increased power

Sample Size Estimation

- Until now, we have taken our sample size to be fixed
- However, sample size (n) can impact the power of a test for a given significance level (α)
- Thus, if we want to achieve a certain power at a given significance level, we can calculate the sample size necessary to do so

Sample Size Estimation

- Back to Starbucks (one-sided z-test): $\sigma = 0.4,\ \alpha = \beta = 0.05,\ \mu_1 = 15.8$ oz, $H_0 : \mu \geq 16$ oz, $H_1 : \mu < 16$ oz
    - In other words, we want $\alpha = 0.05$ and power = 0.95 (so $\beta = 0.05$).
- What size sample (n) do we need?
    - First, find cutoff where we reject the null hypothesis:
    - Recall that we want a power of 0.95 ($\beta = 0.05$)
    - If the true mean were actually $\mu_1 = 15.8$, we want to reject the null hypothesis with probability 0.95
    - For $\beta = 0.05$, we have a z-score of

Sample Size Estimation

- Back to Starbucks (one-sided z-test): $\sigma = 0.4,\ \alpha = \beta = 0.05,\ \mu_1 = 15.8$ oz, $H_0 : \mu \geq 16$ oz, $H_1 : \mu < 16$ oz
  - In other words, we want $\alpha = 0.05$ and power = 0.95 (so $\beta = 0.05$).
- What size sample (n) do we need?
  - First, find cutoff where we reject the null hypothesis:
    z = `qnorm(0.05)` = -1.645, so we have $z = -1.645 = \dfrac{\bar{x} - 16}{0.4/\sqrt{n}} \rightarrow \bar{x} = 16 - 1.645\dfrac{0.4}{\sqrt{n}}$
  - Recall that we want a power of 0.95 ($\beta = 0.05$)
  - If the true mean were actually $\mu_1 = 15.8$, we want to reject the null hypothesis with probability 0.95
  - For $\beta = 0.05$, we have a z-score of z = `qnorm(0.95)` = 1.645

Sample Size Estimation

- We would like the cutoff for $\alpha = 0.05$, which is   , to correspond with the z-score for $\beta = 0.05$ of 1.645

- In other words:

- Thus, we need to sample $n =$ coffees

Sample Size Estimation

- We would like the cutoff for $\alpha = 0.05$, which is $\bar{x} = 16 - 1.645\dfrac{0.4}{\sqrt{n}}$, to correspond with the z-score for $\beta = 0.05$ of 1.645
- In other words:

$$
\begin{aligned}
z = 1.645 &= \frac{\bar{x} - \mu_1}{0.4/\sqrt{n}} \\
\implies \; 1.645 &= \frac{\left(16 - 1.645\frac{0.4}{\sqrt{n}}\right) - 15.8}{0.4/\sqrt{n}} \\
\implies \; 16 - 15.8 &= 1.645\frac{0.4}{\sqrt{n}} + 1.645\frac{0.4}{\sqrt{n}} \\
\implies \; n &= 43.3
\end{aligned}
$$

- Thus, we need to sample $n = 44$ coffees

Sample Size Estimation: One-sided z-test

- We can write this sample size calculation formula more generally for any one-sided hypothesis test
- Let z_α be the value that cuts off an area of α in the upper tail of the standard normal distribution
- Let z_β be the value of z that corresponds to a type II error probability of β
- Consider either set of one-sided hypotheses:
  - $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$
  - $H_0 : \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$
- If we want to achieve a power of $1 - \beta$ while keeping a significance level of α, our sample size formula is

$$n = \left\lceil \left(\frac{\sigma \cdot (z_\alpha + z_\beta)}{\mu_1 - \mu_0} \right)^2 \right\rceil$$

Sample Size Estimation: Two-sided z-test

- For a two-sided hypothesis test, instead of having α in the upper tail, we need $\alpha/2$ in the upper tail
- Let $z_{\alpha/2}$ be the value that cuts off an area of $\alpha/2$ in the upper tail of the standard normal distribution
- Let z_β be the value of z that corresponds to a type II error probability of β
- Consider the two-sided hypothesis:
  - $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$
- If we want to achieve a power of $1 - \beta$ while keeping a significance level of α, our sample size formula is

$$n = \left\lceil \left(\frac{\sigma \cdot (z_{\alpha/2} + z_\beta)}{\mu_1 - \mu_0} \right)^2 \right\rceil$$