Topic 6: Hypothesis Testing

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

Outline

1. Power of a Test

2. Relationship between Hypothesis Testing and Confidence Intervals

Power of a Test

Power

- Recall that α (the significance level) denotes the probability of committing a Type I error, and β denotes the probability of comitting a Type II error.

Power

- Recall that α (the significance level) denotes the probability of committing a Type I error, and β denotes the probability of comitting a Type II error.
- We can analogously define a quantity that represents the probability that a given test will lead to rejection of the null:

Power

Definition (Power)

Suppose that W is the test statistic and $\mathcal{R}$ is the rejection region for a test of a hypothesis involving the value of a parameter θ. Then the power of the test, denoted by power(θ), is the probability that the test will lead to rejection of H_0 when the actual parameter value is θ. That is,

$$\text{power}(\theta) = \mathbb{P}(W \in \mathcal{R} \text{ when the parameter value is } \theta)$$

Power

Theorem (Relationship between Power and β)

If θ_A is a value of θ in the alternative hypothesis H_A, then

$$\text{power}(\theta_A) = 1 - \beta(\theta_A)$$

where $\beta(\theta_A)$ denotes the probability of committing a Type II error when the true value of θ is θ_A.

Power

- As the notation suggests, we typically view power as a function of the true value of θ_A.

Power

- As the notation suggests, we typically view power as a function of the true value of θ_A.
- Plotting the power of a given test at a series of specified values in the alternative space yields a so-called **power curve**.

Power

- As the notation suggests, we typically view power as a function of the true value of θ_A.
- Plotting the power of a given test at a series of specified values in the alternative space yields a so-called **power curve**.
- Let's think through what the "ideal" power curve looks like.

Power

- As the notation suggests, we typically view power as a function of the true value of θ_A.
- Plotting the power of a given test at a series of specified values in the alternative space yields a so-called **power curve**.
- Let’s think through what the “ideal” power curve looks like.
- What would we like power(θ_0) to be?

Power

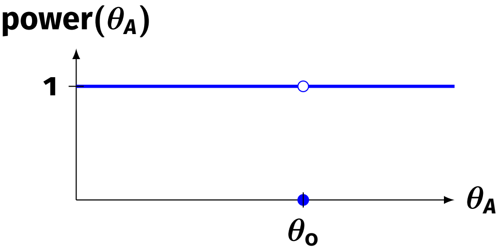
- As the notation suggests, we typically view power as a function of the true value of θ_A.
- Plotting the power of a given test at a series of specified values in the alternative space yields a so-called **power curve**.
- Let's think through what the "ideal" power curve looks like.
- What would we like power(θ_0) to be?
- Well, since power(θ_A) is, by definition and for any point θ_A, the probability of rejecting $H_0 : \theta = \theta_0$ when the true value of θ is θ_A, we'd like power$(\theta_0) = 0$.

Power

- Similarly, for any $\theta_A \neq \theta_0$, we'd like power$(\theta_A) = 1$.

Power

- Similarly, for any $\theta_A \neq \theta_0$, we'd like power$(\theta_A) = 1$.
- So, the ideal power curve for a test would look like



Power

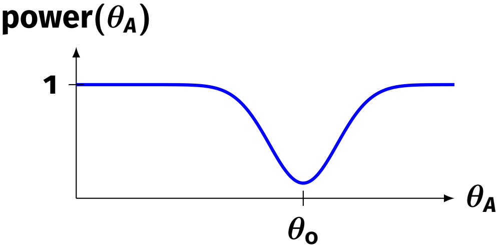
- Now, keep in mind that all tests are performed at a fixed α level of significance.

Power

- Now, keep in mind that all tests are performed at a fixed α level of significance.
- As we discussed before, it's impossible to simultaneously minimize α and β - hence, it's impossible to get a power of exactly zero.

Power

- Now, keep in mind that all tests are performed at a fixed α level of significance.
- As we discussed before, it's impossible to simultaneously minimize α and β - hence, it's impossible to get a power of exactly zero.
- A more realistic power curve for a test of $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$ might look like



Example

Example

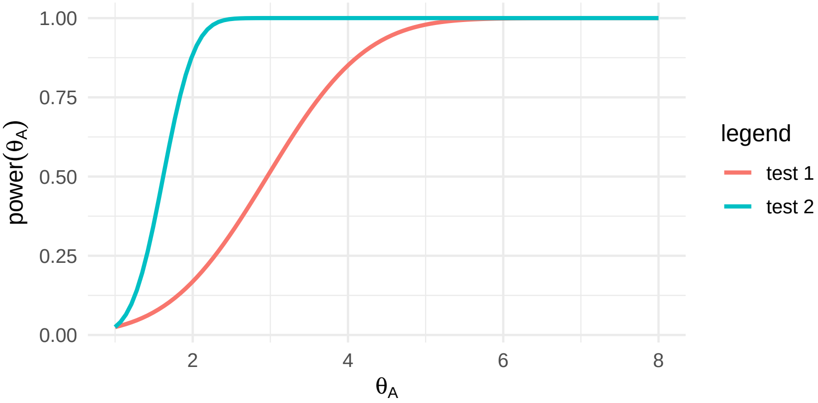
Let $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ for some unknown $\mu \in \mathbb{R}$, and suppose we wish to conduct a test of $H_0 : \mu = \mu_0$ vs $H_A : \mu > \mu_0$ at an $\alpha = 0.05$ level of significance. We propose two tests:

$$\textbf{Test 1: } \text{Reject } H_0 \text{ when } Y_1 - \mu_0 > \Phi^{-1}(0.975)$$

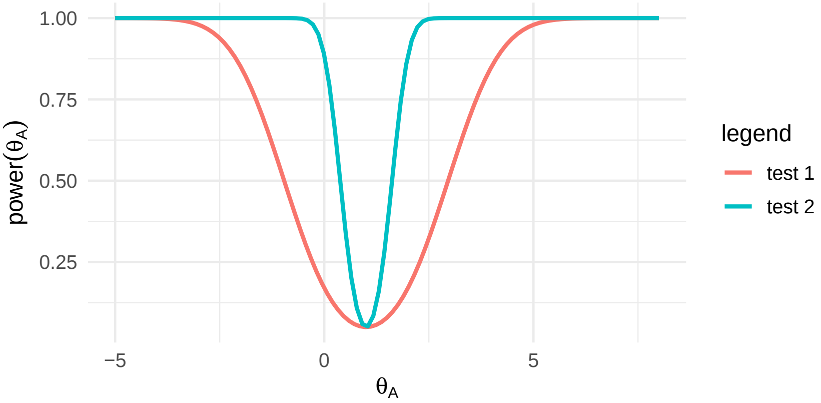
$$\textbf{Test 2: } \text{Reject } H_0 \text{ when } \frac{\overline{Y}_n - \mu_0}{1/\sqrt{n}} > \Phi^{-1}(0.975)$$

Derive expressions for the power functions for these two tests, and use this to determine if one test outperforms the other in terms of power for *all* values of θ in the alternative.

Power


1.00
0.75
0.50
0.25
0.00
power(θ_A)
2
4
6
8
θ_A
legend
test 1
test 2

Power



Power

- Since we want the power of our test to be 1 nearly everywhere, we often seek **uniformly most powerful tests**.

Power

- Since we want the power of our test to be 1 nearly everywhere, we often seek **uniformly most powerful tests**.
- In general, finding such tests is very challenging (and, indeed, such tests don't always exist).

Power

- Since we want the power of our test to be 1 nearly everywhere, we often seek **uniformly most powerful tests**.
- In general, finding such tests is very challenging (and, indeed, such tests don't always exist).
- However, if we restrict ourselves to a *simple-vs-simple* test, we actually *can* construct a most powerful test at a level α, using what is known as the **Neyman-Pearson Lemma**.

Neyman-Pearson Lemma

- Since we want the power of our test to be 1 nearly everywhere, we often seek **uniformly most powerful tests**.

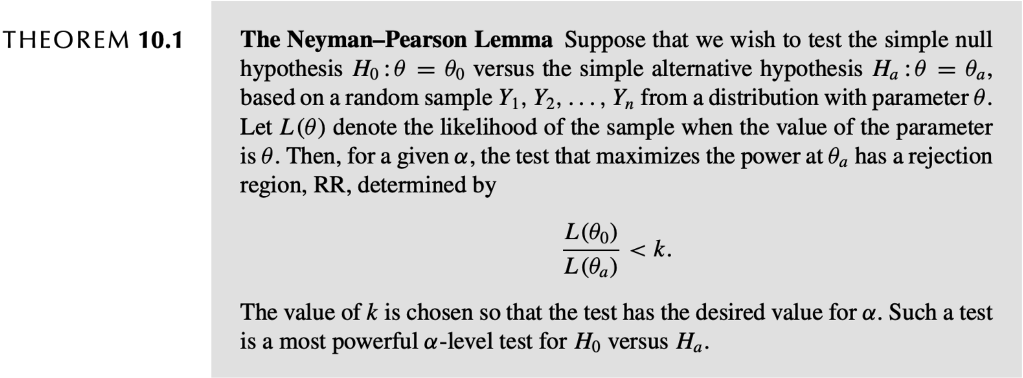
Neyman-Pearson Lemma

- Since we want the power of our test to be 1 nearly everywhere, we often seek **uniformly most powerful tests**.
- In general, finding such tests is very challenging (and, indeed, such tests don't always exist).

Neyman-Pearson Lemma

- Since we want the power of our test to be 1 nearly everywhere, we often seek **uniformly most powerful tests**.
- In general, finding such tests is very challenging (and, indeed, such tests don't always exist).
- However, if we restrict ourselves to a *simple-vs-simple* test, we actually *can* construct a most powerful test at a level α, using what is known as the **Neyman-Pearson Lemma**.

Neyman-Pearson Lemma



THEOREM **10.1**

The Neyman–Pearson Lemma Suppose that we wish to test the simple null hypothesis $H_0 : \theta = \theta_0$ versus the simple alternative hypothesis $H_a : \theta = \theta_a$, based on a random sample $Y_1, Y_2, \ldots, Y_n$ from a distribution with parameter θ. Let $L(\theta)$ denote the likelihood of the sample when the value of the parameter is θ. Then, for a given α, the test that maximizes the power at θ_a has a rejection region, RR, determined by

$$\frac{L(\theta_0)}{L(\theta_a)} < k.$$

The value of k is chosen so that the test has the desired value for α. Such a test is a most powerful α-level test for H_0 versus H_a.

Neyman-Pearson Lemma

- So, in the simple-vs-simple case (i.e. $H_0 : \theta = \theta_0$ vs $H_A : \theta = \theta_A$ for some $\theta_A \neq \theta_0$), we not only have the existence of a most powerful test, but we have its form!

Neyman-Pearson Lemma

- So, in the simple-vs-simple case (i.e. $H_0 : \theta = \theta_0$ vs $H_A : \theta = \theta_A$ for some $\theta_A \neq \theta_0$), we not only have the existence of a most powerful test, but we have its form!
- Indeed, the particular test described in the Neyman-Pearson Lemma is a special case of a broader class of tests, known as **Likelihood Ratio Tests** (LRTs).

Likelihood Ratio Test

Definition (Likelihood Ratio Test)

Consider hypotheses $H_0 : \theta \in \Omega_0$ and $H_A : \theta \in \Omega_A$. Define

$$\Lambda := \frac{\mathcal{L}(\hat{\Omega}_0)}{\mathcal{L}(\hat{\Omega})} = \frac{\max\limits_{\theta \in \Omega_0} \mathcal{L}_{\vec{\mathbf{Y}}}(\theta)}{\max\limits_{\theta \in \Omega_0 \cup \Omega_A} \mathcal{L}_{\vec{\mathbf{Y}}}(\theta)}$$

A **likelihood ratio test** (named as such because we call Λ a **likelihood ratio**) rejects H_0 whenever $\{\Lambda < k\}$.

Likelihood Ratio Test

- Note that the denominator is the maximum value of the likelihood, over the entire parameter space.

Likelihood Ratio Test

- Note that the denominator is the maximum value of the likelihood, over the entire parameter space.
- As such, in many cases we can rewrite the likelihood ratio itself as

$$\Lambda := \frac{\max\limits_{\theta \in \Omega_0} \mathcal{L}_{\vec{\mathbf{y}}}(\theta)}{\mathcal{L}_{\vec{\mathbf{y}}}(\widehat{\theta}_{\mathrm{MLE}})}$$

- Additionally, I've tried to match the definition of the LRT posited in the textbook - note that it applies to a *general* null hypothesis $H_0 : \theta \in \Omega_0$. Recall that in this class (PSTAT 120B), we almost always take $\Omega = \{\theta_0\}$ for some prespecified θ_0, which allows us to further simplify the likelihood ratio (as the next example demonstrates).

Example

Example

Let $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\theta)$. Construct the likelihood ratio test for $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$, using an α level of significance. You do not need to explicitly solve for constants; just derive the general form for the LRT.

Relationship between Hypothesis Testing and Confidence Intervals

$Z-$Test

- Let's, for the moment, return to a two-sided $Z-$Test.

$Z-$Test

- Let's, for the moment, return to a two-sided $Z-$Test.
- That is, take $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for known σ^2, and consider testing $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$.

$Z-$Test

- Let's, for the moment, return to a two-sided $Z-$Test.
- That is, take $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for known σ^2, and consider testing $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$.
- We previously saw that a test with significance level α rejects H_0 in favor of H_A whenever

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

Z—Test

- Equivalently, we fail to reject the null if

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

Z—Test

- Equivalently, we fail to reject the null if

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

- With a bit of algebra, we can see this is equivalent to failing to reject H_0 in favor of H_A when

$$\overline{Y}_n - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \leq \mu_0 \leq \overline{Y}_n + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Z—Test

- Equivalently, we fail to reject the null if

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

- With a bit of algebra, we can see this is equivalent to failing to reject H_0 in favor of H_A when

$$\overline{Y}_n - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \leq \mu_0 \leq \overline{Y}_n + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

- Do the endpoints of this interval look familiar?

Relationship between Hypothesis Testing and Confidence Intervals

Theorem (Hypothesis Testing and CIs)

Consider the setting of a two-sided $Z-$ or $T-$test. An equivalent formulation for the test at an α level of significance is to construct a $(1-\alpha) \times 100\%$ confidence interval for μ, and reject H_0 if μ_0 does not fall inside this CI.

Accepting vs. Failing to Reject

- As your textbook argues, this paradigm allows us to see why it pays to be careful with our language and say "fail to reject H_0" instead of "accept H_0."

Accepting vs. Failing to Reject

- As your textbook argues, this paradigm allows us to see why it pays to be careful with our language and say “fail to reject H_0” instead of “accept H_0.”
- Note that *any* value inside the confidence interval is an “acceptable” value for μ at a significance level α. There isn’t a *single* acceptable value, but an infinite number!

Accepting vs. Failing to Reject

- As your textbook argues, this paradigm allows us to see why it pays to be careful with our language and say “fail to reject H_0” instead of “accept H_0.”
- Note that *any* value inside the confidence interval is an “acceptable” value for μ at a significance level α. There isn’t a *single* acceptable value, but an infinite number!
- So, even if μ_0 falls within our CI, we cannot simply say that we “accept” the null - all we can say is that there isn’t enough evidence to reject it (i.e. we “fail to reject”).

Some Final Comments

- I **highly** encourage you to read Section 10.7 of the textbook, which is a two-page set of assorted comments on hypothesis testing.

Some Final Comments

- I **highly** encourage you to read Section 10.7 of the textbook, which is a two-page set of assorted comments on hypothesis testing.
- Hopefully I've convinced you that hypothesis testing is incredibly useful - indeed, you'll be using hypothesis tests a lot going forward!

Some Final Comments

- I **highly** encourage you to read Section 10.7 of the textbook, which is a two-page set of assorted comments on hypothesis testing.
- Hopefully I've convinced you that hypothesis testing is incredibly useful - indeed, you'll be using hypothesis tests a lot going forward!
- Section 10.7 contains some really nice thoughts and bits of guidance (e.g. what do we do if our null is of the form $H_0 : \theta \leq \theta_0$?)

Some Final Comments

- I'd also like to make a few comments of my own about hypothesis testing before closing out this lecture.

Some Final Comments

- I'd also like to make a few comments of my own about hypothesis testing before closing out this lecture.
- Firstly, there are still some questions we didn't fully answer.

Some Final Comments

- I'd also like to make a few comments of my own about hypothesis testing before closing out this lecture.
- Firstly, there are still some questions we didn't fully answer.
- For example, suppose I want to test the hypothesis that the average pollution levels in Seattle are the same as those in San Francisco.

Some Final Comments

- I’d also like to make a few comments of my own about hypothesis testing before closing out this lecture.
- Firstly, there are still some questions we didn’t fully answer.
- For example, suppose I want to test the hypothesis that the average pollution levels in Seattle are the same as those in San Francisco.
- This is a hypothesis test, but one that asks us to compare *two* different populations.

Some Final Comments

- I’d also like to make a few comments of my own about hypothesis testing before closing out this lecture.
- Firstly, there are still some questions we didn’t fully answer.
- For example, suppose I want to test the hypothesis that the average pollution levels in Seattle are the same as those in San Francisco.
- This is a hypothesis test, but one that asks us to compare *two* different populations.
- Indeed, there is a way to formulate tests for hypotheses like these - check out section 10.8 for a treatment of that.

Some Final Comments

- There also exists a very famous test for comparing two population variances (e.g. is the variance among all cat weights the same as the variance among all dog weights?)

Some Final Comments

- There also exists a very famous test for comparing two population variances (e.g. is the variance among all cat weights the same as the variance among all dog weights?)
- This is called an **$\boldsymbol{F}-$test**, which makes use of something called the $F-$distribution (you'll talk extensively about this in PSTAT 122).

Some Final Comments

- There also exists a very famous test for comparing two population variances (e.g. is the variance among all cat weights the same as the variance among all dog weights?)
- This is called an $\boldsymbol{F}$**—test**, which makes use of something called the F—distribution (you'll talk extensively about this in PSTAT 122).
- Check out section 10.9 of the textbook for a treatment of testing variances.

Some Final Comments

- There also exists a very famous test for comparing two population variances (e.g. is the variance among all cat weights the same as the variance among all dog weights?)
- This is called an **$F-$test**, which makes use of something called the $F-$distribution (you'll talk extensively about this in PSTAT 122).
- Check out section 10.9 of the textbook for a treatment of testing variances.

- There are also some very nice large-sample properties of the Likelihood Ratio Test, which is one of the reasons it remains a very popular method for constructing tests. Take a look at Section 10.11 for more information.