Statistical Testing

Jacob M. Montgomery

2017

Statistical Testing

Overview

- So far we have talked about a number of ways to make *estimates* about population parameters.

Overview

- So far we have talked about a number of ways to make *estimates* about population parameters.
- The other major focus of statistics is in conducting tests.

Overview

- So far we have talked about a number of ways to make *estimates* about population parameters.
- The other major focus of statistics is in conducting tests.
- What is a test?

Overview

- So far we have talked about a number of ways to make *estimates* about population parameters.
- The other major focus of statistics is in conducting tests.
- What is a test?
  - Definition
  - LR Test
  - Wald Test
- What criteria can we use to decide if our test is a good one?
  - Errors
  - Power/Power functions
  - Size/level

What are we testing anyways?

A ***hypothesis*** *is a statement about a population parameter.*

What are we testing anyways?

*A **hypothesis** is a statement about a population parameter.*

*The two complementary hypotheses in a hypothesis testing problem are called the **null hypothesis** and the **alternative hypotheses**. They are denoted by H_0 and H_1, respectively.*

- First, we choose some range/set of values for $\theta \in \Theta$ that represents our null hypothesis.

$$H_0 : \theta \in \Theta_0$$

- First, we choose some range/set of values for $\theta \in \Theta$ that represents our null hypothesis.

$$H_0 : \theta \in \Theta_0$$

- Second, we set up the alternative as the complement of this set.

$$H_1 : \theta \in \Theta_0^c$$

So what is a test?

A **hypothesis testing procedure** or **hypothesis test** is a rule that specifies:

1. For which sample values the decision is made to accept H_0 as true.
2. For which sample values H_0 is rejected and H_1 is accepted as true.

- The subset of the sample space for which H_0 will be rejected is called the rejection region.
- This is often written as

$$R = \{t(\mathbf{x}) > c\}$$

  where c is some "critical value."
- The complement of the rejection region is called the acceptance region.

$$R^c = \{t(\mathbf{x}) < c\}$$

Example 1: Likelihood ratio test

Let $X_1, \ldots, X_n$ be a random sample of a population with a pdf or pmf $f(x|\theta)$. The likelihood function is then

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Example 1: Likelihood ratio test

Let $X_1, \ldots, X_n$ be a random sample of a population with a pdf or pmf $f(x|\theta)$. The likelihood function is then

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

The **likelihood ratio statistic** for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta_0} L(\theta)}{\sup_{\theta} L(\theta)}$$

Example 1: Likelihood ratio test

Let $X_1, \ldots, X_n$ be a random sample of a population with a pdf or pmf $f(x|\theta)$. The likelihood function is then

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

The **likelihood ratio statistic** for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta_0} L(\theta)}{\sup_{\theta} L(\theta)}$$

The **likelihood ratio test** is any test that has a rejection region of the form $\{\lambda(\mathbf{x}) \leq c\}$, where c is any number satisfying $0 \leq c \leq 1$.

Understanding the LRT

- The numerator is the value of within H_0 that maximizes the likelihood.
- The denominator is the value that maximizes $L(\theta)$ over all possible values. This is just the MLE.

Understanding the LRT

- The numerator is the value of within H_0 that maximizes the likelihood.
- The denominator is the value that maximizes $L(\theta)$ over all possible values. This is just the MLE.
- The maximum and minimum values are therefore 1 and 0. Why?

Understanding the LRT

- The numerator is the value of within H_0 that maximizes the likelihood.
- The denominator is the value that maximizes $L(\theta)$ over all possible values. This is just the MLE.
- The maximum and minimum values are therefore 1 and 0. Why?
- After the midterm we will discuss why this statistic is useful, and later in this session we'll discuss how to choose c. For the moment let's practice calculating $\lambda(\mathbf{x})$.

LRT for normal data

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, 1)$ population. We are testing the null hypothesis that $H_0 : \theta = \theta_0$. We have already established that the MLE is $\bar{x}$. Thus λ will be

$$\lambda(\mathbf{x}) = \frac{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n}(x_i - \theta_0)^2/2]}{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n}(x_i - \bar{x})^2/2]}$$

LRT for normal data

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, 1)$ population. We are testing the null hypothesis that $H_0 : \theta = \theta_0$. We have already established that the MLE is $\bar{x}$. Thus λ will be

$$\lambda(\mathbf{x}) = \frac{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n}(x_i - \theta_0)^2/2]}{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n}(x_i - \bar{x})^2/2]}$$

$$= \exp\left[\left(-\sum_{i=1}^{n}(x_i - \theta_0)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\right)/2\right]$$

LRT for normal data

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, 1)$ population. We are testing the null hypothesis that $H_0 : \theta = \theta_0$. We have already established that the MLE is $\bar{x}$. Thus λ will be

$$\lambda(\mathbf{x}) = \frac{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n}(x_i - \theta_0)^2/2]}{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n}(x_i - \bar{x})^2/2]}$$

$$= \exp\left[\left(-\sum_{i=1}^{n}(x_i - \theta_0)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\right)/2\right]$$

$$= \exp\left[\left(-\sum_{i=1}^{n}(x_i - \bar{x})^2 - n(\bar{x} - \theta_0)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\right)/2\right]$$

LRT for normal data

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, 1)$ population. We are testing the null hypothesis that $H_0 : \theta = \theta_0$. We have already established that the MLE is $\bar{x}$. Thus λ will be

$$\lambda(\mathbf{x}) = \frac{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n}(x_i - \theta_0)^2/2]}{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n}(x_i - \bar{x})^2/2]}$$

$$= \exp\left[\left(-\sum_{i=1}^{n}(x_i - \theta_0)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\right)/2\right]$$

$$= \exp\left[\left(-\sum_{i=1}^{n}(x_i - \bar{x})^2 - n(\bar{x} - \theta_0)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2\right)/2\right]$$

$$= \exp[-n(\bar{x} - \theta_0)^2/2]$$

- We have now found an expression of $\lambda(\mathbf{x})$, but it is pretty complicated. The test would be something like

$$\left\{ \exp[-n(\bar{x} - \theta_0)^2/2] < c \right\}$$

- So let's try and recognize and simplify so we can use a simpler function of x.

$$|\bar{x} - \theta_0| \geq \sqrt{-2(\ln(c))/n}$$

- Or even

$$|\bar{x} - \theta_0| < c^*$$

  which should look pretty familliar.

Example 2: Wald test

- Let's do something simpler (but related).
- Assume we are testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$
- Assume further that $\hat{\theta}$ is asymptotically normal.

Example 2: Wald test

- Let's do something simpler (but related).
- Assume we are testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$
- Assume further that $\hat{\theta}$ is asymptotically normal.
- $$\frac{(\hat{\theta} - \theta_0)}{\hat{se}} \approx N(0, 1)$$

Example 2: Wald test

- Let's do something simpler (but related).
- Assume we are testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$
- Assume further that $\hat{\theta}$ is asymptotically normal.
- $$\frac{(\hat{\theta} - \theta_0)}{\hat{se}} \approx N(0, 1)$$
- $W = \frac{(\hat{\theta} - \theta_0)}{\hat{se}}$ is our statistic. What is the test?

Example 2: Wald test

- Let's do something simpler (but related).
- Assume we are testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$
- Assume further that $\hat{\theta}$ is asymptotically normal.
- $$\frac{(\hat{\theta} - \theta_0)}{\hat{se}} \approx N(0, 1)$$
- $W = \frac{(\hat{\theta} - \theta_0)}{\hat{se}}$ is our statistic. What is the test?
- The size α **Wald test** is: reject H_0 when $|W| > z_{\alpha/2}$.

Two class exercises

1. Go back to our normal data example and put back in σ but still assume it is known. Show how we could use the Wald test.
2. Go back to our normal example, but now assume that σ is unknown. Find the LR test. It is useful to know that:

$$\frac{\bar{X} - \theta}{S/\sqrt{n}} \sim t(n-1)$$

Evaluating tests

- There are two criteria we use to evaluate a test.
- The **power** of a test is "power to reject." The idea is that:
  - When the null hypothesis is false, we want a test that will reject it with a high probability (ideally one).

Evaluating tests

- There are two criteria we use to evaluate a test.
- The **power** of a test is "power to reject." The idea is that:
  - When the null hypothesis is false, we want a test that will reject it with a high probability (ideally one).
  - When the null hypothesis is true, we want a test that will reject it with a low probability (ideally zero).

Evaluating tests

- There are two criteria we use to evaluate a test.
- The **power** of a test is "power to reject." The idea is that:
  - When the null hypothesis is false, we want a test that will reject it with a high probability (ideally one).
  - When the null hypothesis is true, we want a test that will reject it with a low probability (ideally zero).
- The **size** of the test is the the maximum probability that we will reject the null hypothesis assuming that the null hypothesis is true.
- The important thing to remember here is that we use these terms to evaluate a test. The test is not usually derived based on these criteria.

Errors and power

- Let R denote the "rejection region" for our statistic such that we reject whenever $\mathbf{x} \in R$.

Errors and power

- Let R denote the "rejection region" for our statistic such that we reject whenever $\mathbf{x} \in R$.
- A **Type I** error occurs if $\theta \in \Theta_0$, which can be written as:

$$P(\mathbf{X} \in R)$$

Errors and power

- Let R denote the "rejection region" for our statistic such that we reject whenever $\mathbf{x} \in R$.
- A **Type I** error occurs if $\theta \in \Theta_0$, which can be written as:

$$P(\mathbf{X} \in R)$$

- A **Type II** error occurs if $\theta \in \Theta_0^c$ and $\mathbf{x} \in R^c$. This can be written as

$$P(\mathbf{X} \in R^c) = 1 - p(\mathbf{X} \in R)$$

The power function

- From this we can see that $P(\mathbf{X} \in R)$ contains all of the information we need to evaluate the erorrs.
  - If $\theta \in \Theta_0$ then $P(\mathbf{X} \in R)$ is the probability of a Type I error.
  - If $\theta \in \Theta_0^c$ then $P(\mathbf{X} \in R)$ is one minus the probability of a Type II error.

The power function

- From this we can see that $P(\mathbf{X} \in R)$ contains all of the information we need to evaluate the erorrs.
  - If $\theta \in \Theta_0$ then $P(\mathbf{X} \in R)$ is the probability of a Type I error.
  - If $\theta \in \Theta_0^c$ then $P(\mathbf{X} \in R)$ is one minus the probability of a Type II error.

*The **power function** of a test with rejection region R is the function of θ defined by*

$$\beta(\theta) = P(\mathbf{X} \in R).$$

The power function

- From this we can see that $P(\mathbf{X} \in R)$ contains all of the information we need to evaluate the erorrs.
  - If $\theta \in \Theta_0$ then $P(\mathbf{X} \in R)$ is the probability of a Type I error.
  - If $\theta \in \Theta_0^c$ then $P(\mathbf{X} \in R)$ is one minus the probability of a Type II error.

*The **power function** of a test with rejection region R is the function of θ defined by*

$$\beta(\theta) = P(\mathbf{X} \in R).$$

- A good test is one with a power function near 1 when H_0 is false and near 0 when H_0 is true. - You can think of this as the "power to reject."

Example: Binomial power function

Let $X \sim \textit{binomial}(5, \theta)$. Consider testing $H_0 : \theta \leq 1/2$ versus $H_1 : \theta > 1/2$. I propose a test that we should reject when we observe all successes.

$$\beta(\theta) = P(X \in R) = P(X = 5) = \theta^5$$

- Let's plot the power function.
- Evaluate it in terms of Type I and Type II errors.

Let $X \sim \textit{binomial}(5, \theta)$. Consider testing $H_0 : \theta \leq 1/2$ versus $H_1 : \theta > 1/2$. I now propose a test that we should reject when we observe 3 *or more* successes.

- Find the power function
- Plot it.
- Evaluate it in terms of Type I and Type II errors.
- Which is better?

Example: Normal power function

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population, where σ^2 is known. An LRT of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ is a test that rejects if

$$\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c.$$

Example: Normal power function

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population, where σ^2 is known. An LRT of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ is a test that rejects if

$$\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c.$$

The power function for this test (where c is some constant) is:

$$\beta(\theta) = P\left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c\right)$$

Example: Normal power function

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population, where σ^2 is known. An LRT of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ is a test that rejects if

$$\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c.$$

The power function for this test (where c is some constant) is:

$$\beta(\theta) = P\left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c\right)$$

$$= P\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} - \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} > c\right)$$

Example: Normal power function

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ population, where σ^2 is known. An LRT of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ is a test that rejects if

$$\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c.$$

The power function for this test (where c is some constant) is:

$$\beta(\theta) = P\left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c\right)$$

$$= P\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} - \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} > c\right)$$

$$= P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)$$

$$\beta(\theta) = P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)$$

- Let $\theta_0 = 0$, $n = 50$, $\sigma = 1$, and $c = 1.25$.
- Plot the power function and discuss.
- What happens when you increase c?
- What happens when you increase n?

Thinking about size and level

- For any fixed sample size, it is usually impossible to make both types of error arbitrarily small.
- Moreoever, as we have seen, there is usually a tradeoff.
- A common approach is to choose a maximum value of Type I error we are willing to tolerate and then search for a test that has the smallest probabilty of Type II error that match that criteria.

Defining size and level

For $0 \leq \alpha \leq 1$*, a test with power function* $\beta(\theta)$ *is a* ***size*** α *test if* $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.

Defining size and level

For $0 \leq \alpha \leq 1$*, a test with power function* $\beta(\theta)$ *is a* ***size*** α *test if* $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$*.*

For $0 \leq \alpha \leq 1$*, a test with power function* $\beta(\theta)$ *is a* ***level*** α *test if* $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$

Defining size and level

For $0 \le \alpha \le 1$, *a test with power function* $\beta(\theta)$ *is a* ***size*** α *test if* $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.

For $0 \le \alpha \le 1$, *a test with power function* $\beta(\theta)$ *is a* ***level*** α *test if* $\sup_{\theta \in \Theta_0} \beta(\theta) \le \alpha$

- Note that many authors are very loose in using size and level and these definitions are not cannonical.

Example: Size of a test for normal data

Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ where σ is kown. We want to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Let $\bar{X}$ be our statistic of interest. Therefore we want to set up a test such that we reject when $\bar{X} > c$ or

$$\beta(\mu) = P(\bar{X} > c)$$

Example: Size of a test for normal data

Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ where σ is kown. We want to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Let $\bar{X}$ be our statistic of interest. Therefore we want to set up a test such that we reject when $\bar{X} > c$ or

$$\beta(\mu) = P(\bar{X} > c)$$

1. Subtract μ from both sides and divide by the standard error.
2. Note that the left side is a standard normal and replace with Z.

Example: Size of a test for normal data

Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ where σ is kown. We want to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Let $\bar{X}$ be our statistic of interest. Therefore we want to set up a test such that we reject when $\bar{X} > c$ or

$$\beta(\mu) = P(\bar{X} > c)$$

1. Subtract μ from both sides and divide by the standard error.
2. Note that the left side is a standard normal and replace with Z.
3. $$\beta(\mu) = 1 - \Phi\left(\frac{c - \mu}{\sigma/\sqrt{n}}\right)$$

$$\beta(\mu) = 1 - \Phi\left(\frac{c - \mu}{\sigma/\sqrt{n}}\right)$$

4. This function is increasing in μ. Really?

$$\beta(\mu) = 1 - \Phi\left(\frac{c - \mu}{\sigma/\sqrt{n}}\right)$$

4. This function is increasing in μ. Really?
5. Thus what is the largest value β could take on and still remain in $\theta \in \Theta_0$? That is, what vale of μ will we have?

$$\beta(\mu) = 1 - \Phi\left(\frac{c - \mu}{\sigma/\sqrt{n}}\right)$$

4. This function is increasing in μ. Really?
5. Thus what is the largest value β could take on and still remain in $\theta \in \Theta_0$? That is, what vale of μ will we have?
6. Substitute in that value for μ and replace $\beta(\cdot)$ with α.

7. Now we just need to solve.

$$\alpha = 1 - \Phi(\sqrt{n}c/\sigma)$$

7. Now we just need to solve.

$$\alpha = 1 - \Phi(\sqrt{n}c/\sigma)$$

$$1 - \alpha = \Phi(\sqrt{n}c/\sigma)$$
$$\Phi^{-1}(1 - \alpha) = \sqrt{n}c/\sigma$$

7. Now we just need to solve.

$$\alpha = 1 - \Phi(\sqrt{n}c/\sigma)$$

$$1 - \alpha = \Phi(\sqrt{n}c/\sigma)$$
$$\Phi^{-1}(1 - \alpha) = \sqrt{n}c/\sigma$$

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

7. Now we just need to solve.

$$\alpha = 1 - \Phi(\sqrt{n}c/\sigma)$$

$$1 - \alpha = \Phi(\sqrt{n}c/\sigma)$$

$$\Phi^{-1}(1 - \alpha) = \sqrt{n}c/\sigma$$

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

8. Now go back. This is saying that we reject if

$$\bar{X} > \frac{\sigma \Phi^{-1}(1-\alpha)}{\sqrt{n}}$$

$$\bar{X} > \frac{\sigma z_\alpha}{\sqrt{n}}$$

9. This can be re-written as we reject when:

$$\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha$$

More advanced topics (some to be covered later)

- For some class of tests of level α we can sometimes identify the uniformly most powerful test.
- The perils of using p-values as tests.
- Multiple tests
- Goodness of fit tests