Introduction to the Theory of Statistics Part 2

PM522b

Meredith Franklin

Division of Biostatistics, University of Southern California

Slides 5, 2019

Topics covered

- Hypothesis Testing
  - Methods for finding tests: Likelihood Ratio Test, Bayesian Tests
  - Methods for evaluating tests: Power, Size
  - Methods for evaluating tests: Uniformly Most Powerful Tests (UMP), Neyman-Pearson Lemma, Karlin-Rubin Theorem

Hypothesis Testing

- We now turn to hypothesis testing whereby we form a statement about a parameter θ and then perform a statistical test to determine the correctness of the statement.
- Hypothesis testing is similar to the scientific method: a scientist formulates a theory and then tests this theory against observation.
- In statistics we pose a theory concerning one or more population parameters (i.e. that they equal specified values), we then sample the population and compare our observations with our posed theory. If the observations disagree with the theory then we reject it.

Hypothesis Testing

Null and Alternative Hypotheses

- In parametric inference, a statistical hypothesis is a statement concerning an unknown parameter for the population distribution $f(x|\theta), x \in \mathbb{R}$ and $\theta \in \Theta$
- The statistical hypothesis is a statement about θ and the testing aims to prove its correctness.
- The hypothesis specifies that θ belongs to some subset of Θ. We define $\Theta_0 \subseteq \Theta$ and $\Theta_1 \subseteq \Theta$ with $\Theta_0 \cap \Theta_1 = 0$.
- The statement that $\theta \in \Theta_0$ is denoted by H_0 and is a statistical hypothesis that invalidates (nullifies) the statement under investigation. This is the null hypothesis.
- The statement that $\theta \in \Theta_1$ is denoted by H_1 and is the statistical hypothesis under investigation. This is the alternative hypothesis.
- In practice, after observing a sample it must be decided whether to accept (fail to reject) H_0 or to reject H_0 and decide H_1 is true.
- The alternative hypothesis H_1 is usually taken to be the negation of H_0.

Hypothesis Testing

Simple and Composite Hypotheses

- A hypothesis $H_i : \theta \in \Theta_i, i = 0, 1$ is called *simple* if the subset Θ_i contains only one element, $\Theta_i = \{\theta_i\}$
- Under a simple hypothesis, $H_i : \theta = \theta_i$, the population distribution $f(x|\theta_i), x \in \mathbb{R}$ is completely specified
- A hypothesis $H_i : \theta \in \Theta_i, i = 0, 1$ is called *composite* in Θ_i contains more than one element
- Under a composite hypothesis, $H_i : \theta = \theta_i$, the population distribution belongs to a family of distribution $f(x|\theta_i), x \in \mathbb{R}, \theta \in \Theta_i, i = 0, 1$
- There are two forms of composite hypotheses:
  - *one-sided* where they take the form $H_0 : \theta \leq \theta_0$ or $H_0 : \theta \geq \theta_0$ for the null hypothesis and subsequently $H_1 : \theta > \theta_1$ or $H_1 : \theta < \theta_1$ for the alternative hypothesis
  - *two-sided* where they take the form $H_0 : \theta = \theta_0$ for the null hypothesis and subsequently $H_1 : \theta \neq \theta_0$ for the alternative hypothesis

Hypothesis Testing

Critical and Acceptance Regions

- A statistical test of the null hypothesis $H_0 : \theta \in \Theta_0$ is a procedure by which, using the observed values $x_1, ..., x_n$ of a random sample $X_1, ..., X_n$, we come to the decision to reject H_0 and accepting H_1 OR accepting H_0 (not rejecting H_0).
- The subset of the sample space for which H_0 will be rejected is called the *rejection* or *critical region*, R.
- The compliment to the critical region is the *acceptance region*, A
- We need a test statistic $T(X_1, ..., X_n)$ that partitions (or maps) $x_1, ..., x_n$ into these two subset regions R and A.
- For example, a simple test could be that if $T(X_1, ..., X_n) = 1$ then we reject H_0, thus $R = \{(x_1, ..., x_n) : T(x_1, ..., x_n) = 1\}$ is our critical region.
- Another example, we could define a test where the critical region is $R = \{(x_1, ..., x_n) : \bar{x} > 0\}$.

Hypothesis Testing

In Class Example: Suppose a mayoral candidate claims she will get more than 50% of the votes in an election, and thereby be the winner. We do not believe this claim, so we would like to test the candidate's claim as a hypothesis test. Set up the elements of a statistical test (hypothesis test):

- null hypothesis
- alternative hypothesis
- test statistic
- rejection/critical region

Hypothesis Testing

Type I and Type II errors

- A statistical test of the null hypothesis $H_0 : \theta \in \Theta_0$ against an alternative hypothesis $H_1 : \theta \in \Theta_1$ leads to either a correct decision or one of the following errors:
  - Type I error = rejecting H_0 when it is true
  - Type II error = accepting H_0 when it is false

	H_0 is accepted	H_0 is rejected
H_0 is true	correct decision	type I error
H_0 is false	type II error	correct decision

Hypothesis Testing

Type I and Type II errors

The probability of rejecting the null hypothesis with the parameter θ restricted on the subsets Θ_0 and Θ_1 of the parameter space Θ can be expressed as the following:

Suppose the null hypothesis is true, $\theta \in \Theta_0$. The probability we will make an error in our decision occurs when our test statistic $T(X)$ falls in the rejection region: $P_\theta(X \in R) = \alpha$. This is the Type I error. Using this same logic, the probability that our test statistic $T(X)$ falls in the acceptance region is $P_\theta(X \in A) = 1 - P_\theta(X \in R)$

Suppose the alternative hypothesis is true, $\theta \in \Theta_1$. The probability we will make an error in our decision occurs when our test statistic $T(X)$ falls in the acceptance region: $P_\theta(X \in A)$. This is the Type II error. If we think about the probability that we reject the null hypothesis supposing the alternative hypothesis is true, we obtain the power. That is, for $\theta \in \Theta_1$, $P_\theta(X \in R) = \beta(\theta)$.

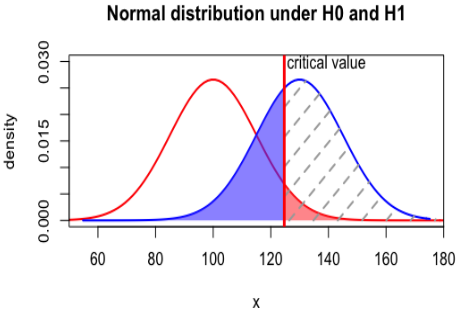
Hypothesis Testing

Power of a test (power function)

- $\beta(\theta) = P_\theta[(X_1, X_2, ..., X_n) \in R], \theta \in \Theta_1$ is the probability of rejecting H_0 given that H_1 is correct, and this is the correct decision of rejecting H_0
- The function $\beta(\theta)$ is called the **power** function of the the test and its value at a specific point $\theta = \theta_1 \in \Theta_1$ is the **power of the test** at θ_1
- The ideal power function is 0 for all $\theta \in \Theta_0$ and 1 for all $\theta \in \Theta_1$
- In a statistical test, it is desirable to keep the probabilities of Type I and Type II errors small. In searching for a good test, commonly the tests are restricted to control the Type I error probability at a specified level. Then within this class of tests, we search for those that have the smallest Type II error probability
- In controlling Type I error probabilities we have:
  - The **size** of a test, defined as $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$ for $0 \leq \alpha \leq 1$
  - The **level** of a test, defined as $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$ for $0 \leq \alpha \leq 1$

Hypothesis Testing

Example where we have expected value μ_0 under H_0 (100), expected value μ_1 under H_1 (130)



Hypothesis Testing

Significance Level and Most Powerful Test

- In testing a statistical hypothesis, a significance level α with $0 \leq \alpha \leq 1$ is taken among the tests satisfying
$$\alpha(\theta) \leq \alpha \text{ for all } \theta \in \Theta_0$$
- The test that minimizes the probability of Type II error for all $\theta \in \Theta_1$, or equivalently that maximizes the power for all $\theta \in \Theta_1$ is chosen.
- If the set Θ_1 contains one point, $\Theta_1 = \{\theta_1\}$, then the test is called the *most powerful test*
- If the set Θ_1 contains more than one point then the test is called the *uniformly most powerful test* (UMP)
- Usually significance level is chosen to be $\alpha \leq 0.05$, but a more conservative significance level is $\alpha \leq 0.01$.

Hypothesis Testing

p-Values

- It has become standard practice to report the **size** of the test, α used in the decision to reject or accept H_0 as it carries important information:
  - If α is small, the decision to reject H_0 is convincing (giving evidence that H_1 is true)
  - If α is large (usually larger than 0.05), the decision to reject H_0 is not very convincing because the hypothesis test has a large probability of incorrectly making that decision
- For a sample $x_1, ..., x_n$, a p-value is a test statistic satisfying $0 \leq p(x) \leq 1$, a p-value $p(X)$ is valid if

$$P(p(X) \leq \alpha) \leq \alpha$$

- The general approach for defining a valid p-value is to determine a test statistic $T(X)$ such that large values of T(x) (the observed value of the test statistic) give evidence that H_1 is true

$$p(x) = \sum_{\theta \in \Theta_0} P(T(X) \geq T(x))$$

Likelihood Ratio Test

The likelihood ratio test (LRT) is a very general method of deriving tests of hypothesis. The procedure works for both simple or composite hypotheses.

Likelihood Ratio Test

This method of hypothesis testing is related to maximum likelihood estimators.

- The generalized LRT is formulated as the ratio of the maximum probability of the observed sample being computed over the parameters in the null hypothesis, H_0 to the maximum probability of the observed sample over *all* possible parameters in Θ.

$$\lambda(x) = \frac{\sup_{\Theta_0} L(\theta|x)}{\sup_{\Theta} L(\theta|x)}$$

  $\lambda(x)$ is the likelihood ratio test statistic.

- Consider testing the null $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$. In terms of a simple hypothesis, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. The LRT statistic for this test is:

$$\lambda(x) = \frac{L(\theta_0|x)}{L(\theta_1|x)}$$

- The test for these statistics is defined by

$$R = \{x : \lambda(x) \leq c\} \text{ for } 0 \leq c \leq 1$$

Likelihood Ratio Test

Unrestricted and restricted maximization

- $\hat{\theta}$ is the MLE of θ, obtained by doing a maximization over all possible parameters Θ
- $\hat{\theta}_0$ can also be an MLE of θ, but obtained by doing a maximization over the restricted parameter space Θ_0
- That means $\hat{\theta}_0 = \hat{\theta}_0(x)$ is the value of the parameter $\theta \in \Theta_0$ that maximizes the likelihood $L(\theta|x)$. In this case, the LRT is

$$\begin{aligned}\lambda(x) &= \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)} \\ &= \frac{\sup_{\theta \in \Theta_0} L(\theta|x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} L(\theta|x_1, \ldots, x_n)}\end{aligned}$$

Likelihood Ratio Test

- The critical region for the LRT statistic is defined by:

$$R = \{x_1, ..., x_n : \lambda(x) \leq c\}$$

where c is a constant between 0 and 1 and will be for a given significance level α when it is chosen to satisfy:

$$\sup_{\theta \in \Theta_0} \{P[\lambda(x) \leq c_\alpha | \theta \in \Theta_0]\} = \alpha$$

Likelihood Ratio Test Simple Hypothesis

In class example for $X_1, ..., X_n \sim Exp(\theta)$. Recall the distributions of $f(X|\theta)$ known completely under H_0 and H_1 in a simple hyothesis.

Likelihood Ratio Test Composite Hypothesis

Example: LRT normal distribution

Testing the mean of $\mathrm{N}(\mu, \sigma^2)$ where σ^2 is known, namely $\mathrm{H}_0 : \mu = \mu_0$ against $\mathrm{H}_1 : \mu \neq \mu_0$. μ_0 is a number fixed by the experimenter before doing the experiment.

$$L(\mu|x_1, ..., x_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2]$$

$$\sup_{\mu \in \{\mu_0\}} L(\mu|x_1, ..., x_n) = L(\mu_0|x_1, ..., x_n)$$

$$\sup_{\mu \in \Theta} L(\mu|x_1, ..., x_n) = L(\hat{\mu}|x_1, ..., x_n)$$

where $\hat{\mu} = \bar{x}$ is the MLE of μ

$$\lambda(x) = \frac{L(\mu_0|x_1, ..., x_n)}{L(\hat{\mu}|x_1, ..., x_n)} = \frac{\exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_0)^2]}{\exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2]}$$

Likelihood Ratio Test

Example: LRT of normal distribution con't

Noting that

$$\sum_{i=1}^{n}(x_i - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \theta_0)^2$$

the LRT statistic becomes

$$\frac{n(\bar{x} - \theta_0)^2}{\sigma^2} = \left(\frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}}\right)^2$$

Because $\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}$ is a standard normal random variable, the above $\lambda(x)$ is the square of two standard normals which is a χ_1^2 random variable.

Likelihood Ratio Test

Example: LRT of normal distribution con't

$$\left(\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}\right)^2 \text{ and } \left|\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}\right|$$

Both are test statistics, with the first being the square of the critical value for the second. The critical region in terms of the second is:

$$R = \{x_1, ..., x_n : \left|\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}\right| \geq c\}$$

where c is a constant for a given significance level α, determined by

$$P(\left|\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}\right| \geq c|\mu = \mu_0) = \alpha$$

Since $\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}$ is a standard normal random variable we have a two-sided z-test. The test is to reject H_0 if $\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}} \geq z_{\alpha/2}$.

Likelihood Ratio Test

In class example with unknown mean and variance under a one-sided hypothesis.

Neyman-Pearson Lemma

This Lemma allows us to find the test of a given size α with the largest power (it is a most powerful test) and formulated around the simple hypothesis testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ where $\theta_0 \neq \theta_1$.

Basically the Neyman-Pearson lemma tells us that the best test for a simple hypothesis is a likelihood ratio test.

Neyman-Pearson Lemma

Neyman-Pearson Lemma

Let $X_1, .., X_n$ be a random sample from a distribution with parameter θ, where $\theta \in \Theta = \{\theta_0, \theta_1\}$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ where $\theta_0 \neq \theta_1$; the pdf or pmf corresponding to $\theta_i, i = 0, 1$ is $f(x|\theta_i)$ and corresponding likelihood function is $L(\theta_i|x)$.
If there exists a test at significance level α such that for some positive constant k,

$$\frac{L(\theta_1|x)}{L(\theta_0|x)} \geq k \text{ for each } x \in R \text{ (inside the critical region)}$$
$$\frac{L(\theta_1|x)}{L(\theta_0|x)} \leq k \text{ for each } x \notin R \text{ (outside the critical region)}$$

then R is the most powerful critical region of size α, and this is the most powerful test for testing the null against alternative hypothesis, $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$.

Neyman-Pearson Lemma

The Neyman-Pearson lemma gives us the most powerful test of size α for $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$.
Thus, the test based on the critical region $R = \{x : \lambda(x) \leq k\}$ has the largest power (smallest Type II error) of all tests with significance level α.
Among all tests with a given probability of Type I error, the likelihood ratio test minimizes the probability of a Type II error.

Proof Neyman-Pearson Lemma

For a continuous random variable, let R be the critical region of size α and A be another region of size α, both fitting the conditions of the Neyman-Pearson Lemma. Also, let $\int \cdots \int L(\theta|x_1, ..., x_n) dx_1, ..., dx_n$ be represented by $\int L(\theta)$ (integrated over a region).
Proof not required, shown on pp. 388-389 of CB.

Neyman-Pearson Lemma

Example: let $X_1, ..., X_n$ be iid $N(\mu, \sigma^2)$ where μ is unknown and σ^2 is known. For the hypothesis $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ where $\mu_1 > \mu_0$ find the most powerful test.

$$\frac{L(\theta_1|x)}{L(\theta_0|x)} = \frac{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_1)^2]}{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_0)^2]}$$

$$= \frac{\exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_1)^2}{\exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_0)^2}$$

$$= \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} ((x_i - \mu_1)^2 - (x_i - \mu_0)^2)$$

$$= \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i^2 - 2\mu_1 x_i + \mu_1^2) - \sum_{i=1}^{n} (x_i^2 - 2\mu_0 x_i + \mu_0^2)$$

Neyman-Pearson Lemma

Example con't

$$\exp -\frac{1}{2\sigma^2}(n(\mu_1^2 - \mu_0^2) - 2n\bar{x}(\mu_1 - \mu_0)) \geq k$$

Assume $\sigma^2 = 1$ and complete solving for k

Uniformly Most Powerful Tests

The Neyman-Pearson lemma gives us the most powerful test for a simple null hypothesis against a simple alternative hypothesis. It can be extended for composite alternative hypotheses by ensuring each simple alternative is accounted for.

Uniformly Most Powerful Test (UMP)

For a continuous random variable, let R be the critical region of size α. A test is the uniformly most powerful if it is a most powerful test against each simple alternative in the alternative (composite) hypothesis.
The critical region is called the most powerful critical region of size α.

LRT and Sufficient Statistics

If T(X) is a sufficient statistics for θ and $\lambda^*(t)$ and $\lambda(x)$ are the LRT statistics based on T(X) and X, then $\lambda^*(T(x)) = \lambda(X)$ for every X in the sample space.

Proof in class.

Monotone Likelihood Ratio

We defined the Uniformly Most Powerful test, but we must state when it exists.

For $X_1, ..., X_n$ with likelihood function $L(\theta|X) = \prod_{i=1}^{n} f(X_i|\theta)$ we define:

Monotone Likelihood Ratio Property

The family (or set) of distributions has Monotone Likelihood Ratio (MLR) if we can represent the likelihood ratio as

$$\frac{f(X|\theta_2)}{f(X|\theta_1)} = f(T(X), \theta_1, \theta_2) \text{ for } \theta_2 > \theta_1$$

Where the function $f(T(X), \theta_1, \theta_2)$ is non-decreasing in $T(X)$ (strictly increasing in T(X)).

Distributions having the MLR property include exponential, binomial, normal (unknown mean, known variance), and Poisson. Any regular exponential family with $g(t(x)|\theta) = h(t)c(\theta)\exp(w(\theta)t(x))$ has a MLR if $w(\theta)$ is a non-decreasing function.

Monotone Likelihood Ratio

MLR Example

Consider $X_1, .., X_n \sim N(\mu, 1)$. The pdf is

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2}}$$

and the likelihood is

$$f(X|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^2}$$

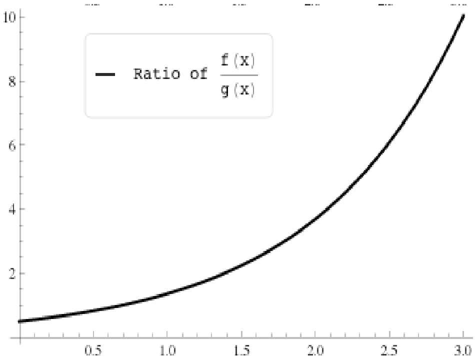
Then the likelihood ratio can be written as

$$\frac{f(X|\mu_2)}{f(X|\mu_1)} = e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu_2)^2} + e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu_1)^2} = e^{(\mu_2-\mu_1)\sum_{i=1}^{n}X_i - \frac{n}{2}(\mu_2^2-\mu_1^2)}$$

For $\mu_2 > \mu_1$ the likelihood ratio is increasing in $T(X) = \sum_{i=1}^{n} X_i$ and the MLR property holds.

Monotone Likelihood Ratio

The MLR property tells us that the likelihood ratio $f(X|\theta_2)/f(X|\theta_1)$ is a non-decreasing function of T(X) defined when $\theta_2 > \theta_1$



Parameter T(X) versus the ratio $f(X|\theta_2)/f(X|\theta_1)$, showing MLR property.
Source: Wikipedia

Monotone Likelihood Ratio

We link the MLR property with UMP tests:
If statistic T(X) has the MLR property then the UMP test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$ exists and can be expressed in terms of T(X) rather than in terms of the full likelihood ratio.

Monotone Likelihood Ratio

It holds true that if T(X) has MLR property then:

1. $T(X) > t_0$ is a UMP for $H_0 : \theta \leq \theta_0$ (or $\theta = \theta_0$) vs. $H_0 : \theta > \theta_0$
2. $T(X) < t_0$ is a UMP for $H_0 : \theta \geq \theta_0$ (or $\theta = \theta_0$) vs. $H_0 : \theta < \theta_0$

Monotone Likelihood Ratio

For 1), Consider $H_0 : \theta = \theta_0$ vs $H_0 : \theta > \theta_0$.
$T(X) > t_0$ is a UMP for $H_0 : \theta \leq \theta_0$ (or $\theta = \theta_0$) vs. $H_0 : \theta > \theta_0$

Karlin-Rubin Theorem

We link sufficient statistics and the MLR property with UMP tests through the Karlin-Rubin Theorem.

Suppose we are testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$. Let T(X) be a sufficient statistic, and the family of distributions for θ has MLR property in T(X). Then for any k, the test with rejection region $T(X) > k$ is the UMP test (given α) where $P(T(X) > k) = \alpha$.

By the same argument, we can apply this test to $H_0 : \theta = \theta_0$ vs $H_1 : \theta < \theta_0$ given $T(X) < k$ for the UMP α-level test where $P(T(X) < k) = \alpha$.

Non-Existence of UMP Tests

Two sided hypothesis do not have UMP tests. The reason is that the test that is UMP for $\theta < \theta_0$ is not the same as the test that is UMP for $\theta > \theta_0$. For the two sided case to exist, it would have to be most powerful across every value in the alternative hypothesis.