Topic 10: Hypothesis Testing

Rohini Somanathan

Course 003, 2017

The Problem of Hypothesis Testing

A statistical hypothesis is an assertion or conjecture about the probability distribution of one or more random variables.

A test of a statistical hypothesis is a rule or procedure for deciding whether to reject that assertion.

Suppose we have a sample $x = (x_1, \ldots, x_n)$ from a density f. We have two hypotheses about f. On the basis of our sample, one of the hypotheses is accepted and the other is rejected.

The two hypotheses have different status:

- the null hypothesis, H_0, is the hypothesis under test. It is the conservative hypothesis, not to be rejected unless the evidence is clear
- the alternative hypothesis H_1 specifies the kind of departure from the null hypothesis that is of interest to us

A hypothesis is simple if it completely specifies the probability distribution, else it is composite.

Examples:

- Income $\sim$ log-normally with known variance but unknown mean. $H_0 : \mu \geq 8,000$ rupees per month, $H_1 : \mu < 8,000$
- We would like to know whether parents are more likely to have boys than girls. The probability of a boy child $\sim$ Bernoulli (p). $H_0 : p = \frac{1}{2}$ and $H_1 : p > \frac{1}{2}$

Statistical tests

Before deciding whether or not to accept H_0, we observe a random sample. Denote by S, the set of all possible sample outcomes.

A test procedure partitions S into two subsets, the acceptance region with values that lead us to accept H_0 and the critical region R , which has values which lead its rejection.

These sets are usually defined in terms of values taken by a test statistic (the same mean, the sample variance or functions of these). The critical values of a test statistic are the bounds of R.

When arriving at a decision based on a sample and a test, we may make two types of errors:

- H_0 may be rejected when it is true- a Type I error
- H_0 may be accepted when it is false- a Type II error

We use α and β to denote these errors.

The power function is useful in computing these errors and summarizes the properties of a test. We will define these functions.

We also identify the set of hypothesis testing problems for which there is an optimal test and characterize these tests.

The power function

The power function of a test is the probability of rejecting H_0 as a function of the parameter $\theta \in \Omega$. If we are using a test statistic T

$$\pi(\theta) = Pr(T \in R) \text{ for } \theta \in \Omega$$

Since the power function of a test specifies the probability of rejecting H_0 as a function of the real parameter value, we can evaluate our test by asking how often it leads to mistakes.

What is the power function of an ideal test ? Think of examples when such a test exists.

It is common to specify an upper bound α_0 on $\pi(\theta)$ for every value $\theta \in \Omega_0$. This bound α_0 is the level of significance of the test.

The size of a test, α is the maximum probability, among all values of $\theta \in \Omega_0$ of making an incorrect decision:

$$\alpha = \sup_{\theta_0 \in \Omega_0} \pi(\theta)$$

Given a level of significance α_0, only tests for which $\alpha \leq \alpha_0$ are admissible.

Example 1: Binomial distribution

The probability of a defective bolts in a shipment is given by a Bernoulli random variable.

We have the following null and alternative hypotheses: $H_0 : p \leq .02$, $H_1 : p > .02$.

Our sample consists of 200 bolts and our test statistic is number of defective items X in the sample. We want to find a test for which $\alpha_0 = .05$.

Let us now think of how our test statistic X behaves for different values of p. We want to find X such that $\alpha_0 \leq 0.05$

Since $X \sim \text{Bin}(n, p)$, the probability of X being greater than any given x is increasing in p, we can focus on $p = .02$. If $Pr(X > x) \leq .05$ for this p, it will be true for all smaller values of p.

It turns out that for $p = .02$, the probability that the number of defective items is greater than 7 is .049 (display 1-binomial(200,7,.02)).

$R = \{x : x > 7\}$ is therefore the test we choose and its size is 0.049.

In general, for discrete distributions, the size will typically be strictly smaller than α_0.

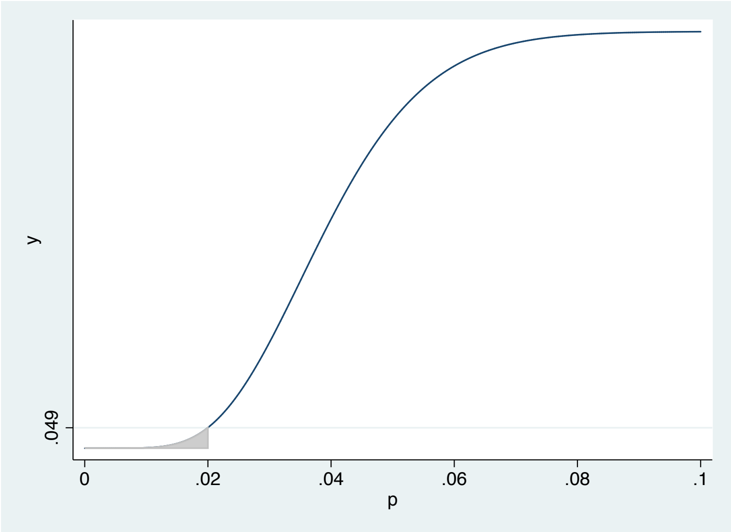
The size of tests which reject for more than 4, 5 and 6 defective pieces are .37. .21 and .11 respectively.

Example 1: power function

Let's graph the power function of this test:

twoway function y=1-binomial(200,7,x), range(0 .1) xtitle(p)||function y=1-binomial(200,7,x), range(0 .02) color(gs12) recast(area) legend(off) ylabel(.049)

(all on one line)



Can you mark off the two types of errors for different values of p?

What happens to this power function as we increase or decrease the critical region? (say $R = \{x : x > 6\}$ or $R = \{x : x > 8\}$.

Example 2 : Uniform distribution

A random sample is taken from a uniform distribution on $[0, \theta]$ and we would like to test

$$H_0 : 3 \leq \theta \leq 4 \qquad \text{against} \qquad H_1 : \theta < 3 \text{ or } \theta > 4$$

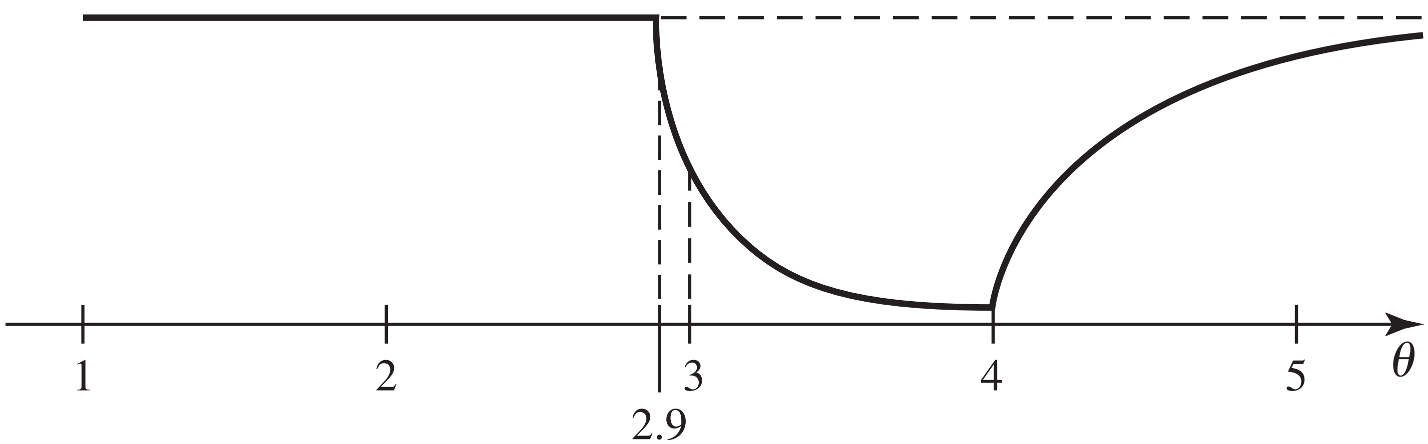
Our test procedure uses the M.L.E. of θ, $Y_n = \max(X_1, \dots, X_n)$ and rejects the null hypothesis whenever Y_n lies outside $[2.9, 4]$. What might be the rationale for this type of test?

The power function for this test is given by

$$\pi(\theta) = Pr(Y_n < 2.9|\theta) + Pr(Y_n > 4|\theta)$$

- What is the power of the test if $\theta < 2.9$?
- When θ takes values between 2.9 and 4, the probability that any sample value is less than 2.9 is given by $\frac{2.9}{\theta}$ and therefore $Pr(Y_n < 2.9|\theta) = (\frac{2.9}{\theta})^n$ and $Pr(Y_n > 4|\theta) = 0$. Therefore the power function $\pi(\theta) = (\frac{2.9}{\theta})^n$
- When $\theta > 4$, $\pi(\theta) = (\frac{2.9}{\theta})^n + [1 - (\frac{4}{\theta})^n]$

The power graph..example 2



Example 3: Normal distribution

$X \sim N(\mu, 100)$ and we are interested in testing $H_0 : \mu = 80$ against $H_1 : \mu > 80$.

Let $\bar{x}$ denote the mean of a sample $n = 25$ from this distribution and suppose we use the critical region $R = \{(x_1, x_2, \dots x_{25}) : \bar{x} > 83\}$.

The power function is

$$\pi(\mu) = P(\bar{X} > 83) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{83 - \mu}{2}\right) = 1 - \Phi(\frac{83 - \mu}{2})$$

The size of this test is the probability of Type 1 error: $\alpha = 1 - \Phi(\frac{3}{2}) = .067 = \pi(80)$

What are the values of $\pi(83)$, and $\pi(86)$?

$\pi(80)$ is given above, $\pi(83) = 0.5$, $\pi(86) = 1 - \Phi(-\frac{3}{2}) = \Phi(\frac{3}{2}) = .933$

stata: display normal(1.5)

We can sketch the graph of the power function using the command

stata: twoway function 1-normal((83-x)/2), range (70 90)

What is the p-value corresponding to $\bar{x} = 83.41$?

This is the smallest level of significance, α_0 at which a given hypothesis would be rejected based on the observed outcome of X?

In this case, $Pr(\bar{X} \geq 83.41) = 1 - \Phi(\frac{3.41}{2}) = .044.$

Can you find a test for which $\alpha_0 = .05$?

Testing simple hypotheses

We have so far focussed on understanding the properties of given tests. What does an optimal test look like and when does such a test exist?

Suppose that Ω_0 and Ω_1 contain only a single element each and our null and alternative hypotheses are given by

$$H_0 : \theta = \theta_0 \text{ and } H_1 : \theta = \theta_1$$

Denote by $f_i(x)$ the joint density function or p.f. of the observations in our sample under H_i:

$$f_i(x) = f(x_1|\theta_i)f(x_2|\theta_i)\dots f(x_n|\theta_i)$$

Denote the type I and type II errors by $\alpha(\delta)$ and $\beta(\delta)$ respectively:

$$\alpha(\delta) = Pr(\text{ Rejecting } H_0|\theta = \theta_0) \quad \text{and} \quad \beta(\delta) = Pr(\text{ Not Rejecting } H_0|\theta = \theta_1)$$

By always accepting H_0, we achieve $\alpha(\delta) = 0$ but then $\beta(\delta) = 1$. The converse is true if we always reject H_0.

It turns out that we can find an optimal test which minimizes any linear combination of $\alpha(\delta)$ and $\beta(\delta)$.

Optimal tests for simple hypotheses

Theorem (Minimizing the linear combination $a\alpha(\delta) + b\beta(\delta)$):

Let δ^* denote a test procedure such that the hypothesis H_0 is accepted if $af_0(x) > bf_1(x)$ and H_1 is accepted if $af_0(x) < bf_1(x)$. Either H_0 or H_1 may be accepted if $af_0(x) = bf_1(x)$. Then for any other test procedure δ,

$$a\alpha(\delta^*) + b\beta(\delta^*) \leq a\alpha(\delta) + b\beta(\delta)$$

So we reject whenever the likelihood ratio $\frac{f_1(x)}{f_0(x)} > \frac{a}{b}$. If we are minimizing the sum of errors, we would reject whenever $\frac{f_1(x)}{f_0(x)} > 1$.

Proof. (for discrete distributions)

$$a\alpha(\delta)+b\beta(\delta) = a\sum_{x\in R} f_0(x)+b\sum_{x\in R^c} f_1(x) = a\sum_{x\in R} f_0(x)+b\left[1-\sum_{x\in R} f_1(x)\right] = b+\sum_{x\in R}[af_0(x)-bf_1(x)]$$

The desired function $a\alpha(\delta) + b\beta(\delta)$ will be minimized if the critical region includes only those points for which $af_0(x) - bf_1(x) < 0$. We therefore reject when the likelihood ratio exceeds $\frac{a}{b}$.

□

Minimizing $\beta(\delta)$, given α_0

If we fix a level of significance α_0 we want a test procedure that minimizes $\beta(\delta)$, the type II error subject to $\alpha \leq \alpha_0$. We can obtain this by modifying the previous result.

The Neyman-Pearson Lemma : Let δ^* denote a test procedure such that, for some constant k, the hypothesis H_0 is accepted if $f_0(x) > kf_1(x)$ and H_1 is accepted if $f_0(x) < kf_1(x)$. Either H_0 or H_1 may be accepted if $f_0(x) = kf_1(x)$. If δ is any other test procedure such that $\alpha(\delta) \leq \alpha(\delta^*)$, then it follows that $\beta(\delta) \geq \beta(\delta^*)$. Furthermore if $\alpha(\delta) < \alpha(\delta^*)$ then $\beta(\delta) > \beta(\delta^*)$

This result implies that if we set a level of significance $\alpha_0 = .05$, we should try and find a value of k for which $\alpha(\delta^*) = .05$ This procedure will then have the minimum possible value of $\beta(\delta)$.

Proof. (for discrete distributions)

From the previous theorem we know that $\alpha(\delta^*) + k\beta(\delta^*) \leq \alpha(\delta) + k\beta(\delta)$. So if $\alpha(\delta) \leq \alpha(\delta^*)$, it follows that $\beta(\delta) \geq \beta(\delta^*)$

□

Neyman Pearson Lemma..example 1

Let $X_1 \dots X_n$ be a normal random sample, $X_i \sim N(\mu, 1)$.

$$H_0 : \theta = 0 \text{ and } H_1 : \theta = 1$$

We want a test procedure δ for which β is minimized given $\alpha_0 = .05$.

$$f_0(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum x_i^2} \text{ and } f_1(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum (x_i - 1)^2} \text{ so } \frac{f_1(x)}{f_0(x)} = e^{n(\bar{x}_n - \frac{1}{2})} > k$$

by the NP Lemma. This condition can be re-written in terms of our sample mean $\bar{x}_n$:

$$\bar{x}_n > k' = \frac{1}{2} + \frac{1}{n} \log k$$

How do we find k'? $Pr(\bar{X}_n > k' | \theta = 0) = Pr(Z > k'\sqrt{n})$. For $\alpha_0 = .05$, we have $k'\sqrt{n} = 1.645$ or $k' = \frac{1.645}{\sqrt{n}}$

Under this procedure, Type II error is

$$\beta(\delta^*) = Pr(\bar{X}_n < \frac{1.645}{\sqrt{n}} | \theta = 1) = Pr(Z < 1.645 - \sqrt{n})$$

For $n = 9$, $\beta(\delta^*) = 0.0877$ (display normal(1.645-3))

If instead, we want to minimize $2\alpha(\delta) + \beta(\delta)$, we choose $k' = \frac{1}{2} + \frac{1}{n}\log 2$, our optimal procedure rejects H_0 when $\bar{x}_n > 0.577$. In this case, $\alpha(\delta_0) = 0.0417$ (display 1-normal((.577)*3)) and $\beta(\delta_0) = 0.1022$ (display normal((.577-1)*3)) and the minimized value of $2\alpha(\delta) + \beta(\delta)$ is 0.186

Neyman Pearson Lemma..example 2

Let $X_1 \dots X_n$ be a sample from a Bernoulli distribution

$$H_0 : p = 0.2 \text{ and } H_1 : p = 0.4$$

How do we find a test procedure δ which limits us to an α_0 and minimizes β. let y denote values taken by $Y = \sum X_i$

$$f_0(x) = (0.2)^y (0.8)^{n-y} \text{ and } f_1(x) = (0.4)^y (0.6)^{n-y}$$

$$\frac{f_1(x)}{f_0(x)} = \left(\frac{3}{4}\right)^n \left(\frac{8}{3}\right)^y$$

The lemma tells us to use a procedure which rejects H_0 when the likelihood ratio is greater than a constant k. This condition can be re-written in terms of our sample mean $y > \frac{\log k + n \log \frac{4}{3}}{\log \frac{8}{3}} = k'$.

Now we would like to find k' such that $Pr(Y > k'|p = 0.2) = .05$.

We may not however be able to do since Y is discrete. If $n = 10$, we find that $Pr(Y > 3|p = 0.2) = .121$ and $Pr(Y > 4|p = 0.2) = .038$, (display 1-binomial(10,4,.2)) so we can decide to set one of these probabilities as the values of $\alpha(\delta)$ and use the corresponding values for k'.

Can you calculate $\beta(\delta)$ if δ rejects for $y > 4$?