STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 15

How do we compare two simple hypotheses

- This is the simplest situation for hypothesis testing.
- The two simple hypotheses can be formally treated as

$$\theta = \theta_0 \quad \text{versus} \quad \theta = \theta_1$$

  where each of θ_0 and θ_1 completely specifies $f(\mathbf{x}|\theta)$.
- We make a choice between these two possibilities based on the data $\mathbf{X}$ we observe.
- When $\mathbf{X}$ takes values in a set of possible values $\mathscr{R}$, we will *reject* θ_0 and choose θ_1.
- Otherwise, we accept, or do not reject, θ_0.

The likelihood ratio test

- The *likelihood ratio statistic*

$$\frac{f(\mathbf{X}|\theta_1)}{f(\mathbf{X}|\theta_0)}$$

  measures the relative evidence for the data under the two hypotheses.
- A rejection region based on this statistic is

$$\mathscr{R} = \left\{ \mathbf{x} : \quad \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} > K. \right\}$$

  for some constant K.
- It turns out that this is the "best" test to use with under certain sampling theory criteria.

What is a “good” test?

- Let us describe our testing problem as making a choice between

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

- We call H_0 the *null hypothesis* and H_1 the *alternative hypothesis*.
- Mathematically there is no difference between the null and the alternative, but in practice we choose one of them (H_0) to be the *default* or *baseline* case.

- Intuitively, a test is "good" if
    1. when H_0 is true, i.e. $\theta = \theta_0$, we are *more likely* to accept H_0.
    2. when H_0 is false, i.e. $\theta = \theta_1$ we are *more likely* to accept H_1 (i.e. reject H_0).
- Here the probabilistic term "likely" is with respect to the sampling distribution of the data $f(\mathbf{x}|\theta)$.
- It means that if we repeat the experiment many times, a small proportion of times we will make a mistake.
- In other words, we are make as few errors as possible as we repeat the experiment many times.

Two kinds of error

- Correspondingly, there are two types of errors that one can make:

  1. Decide H_1, i.e. $\theta = \theta_1$, when H_0 is true. This is called an error of the first type, or a *Type I error*.
  2. Decide H_0, i.e. $\theta = \theta_0$, when H_1 is true. This is called an error of the second type, or a *Type II error*.

- In addition, we can define two quantities

$$\alpha = \Pr(\text{decide } H_1 \mid H_0 \text{ is true}) - \text{Type I error rate.}$$
$$\beta = \Pr(\text{decide } H_0 \mid H_1 \text{ is true}) - \text{Type II error rate.}$$

- Using the characterization of tests by rejection regions

$$\alpha = \Pr(\mathbf{X} \text{ is in the rejection region} \mid \theta = \theta_0) = \Pr(\mathbf{X} \in \mathscr{R} | H_0)$$
$$\beta = \Pr(\mathbf{X} \text{ is not in the rejection region} \mid \theta = \theta_1) = \Pr(\mathbf{X} \notin \mathscr{R} | H_1)$$

- Question: How does the error rates change with the *size* of the rejection region $\mathscr{R}$?

A decision theoretic perspective

- Truth is either H_0 or H_1.
- We have two actions $\{0,1\}$—corresponding to choosing either hypothesis.
- Consider a test, i.e., a decision rule $\delta(\mathbf{X}) = \mathbf{1}_{\mathbf{X}\in\mathscr{R}}$.
- Also adopt the 0-1 loss $L(H_i, a) = \mathbf{1}(a \neq i)$ for $i = 0, 1$. That is, if we choose the wrong hypothesis, then the loss is 1; otherwise loss is 0.
- The risk function of the decision rule is given by

$$R_\delta(0) = \mathrm{E}L(H_0, \delta(\mathbf{X})) = \mathrm{E}\left(\mathbf{1}_{\mathbf{X}\in\mathscr{R}} \,|\, H_0\right) = \Pr(\mathbf{x} \in \mathscr{R} \,|\, H_0) = \alpha$$

and

$$R_\delta(1) = \mathrm{E}L(H_1, \delta(\mathbf{X})) = \mathrm{E}\left(\mathbf{1}_{\mathbf{X}\notin\mathscr{R}} \,|\, H_1\right) = \Pr(\mathbf{x} \notin \mathscr{R} \,|\, H_1) = \beta.$$

- Ideally we want both α and β to be small, even both to be zero.
  - For α to be small, we want to have a small rejection region.
  - For β to be small, we want to have a large rejection region.
- These two goals are contradicting. (Draw a figure.)
- Typically we cannot make both zero. So we must strike a balance. A formal theory is needed.

- A common strategy is to specify a permissible level for one of the two types of errors, most often for the Type I error α.
- This provides a criterion to control the *size* of the rejection region.
- Then given that level for α, find a test that attains the best level for the other, by custom β.
  - E.g. if we permit a 5% Type I error rate, what test (or rejection region) gives the lowest Type II error rate?
- It turns out that the *likelihood ratio test* is the best test for this purpose. That is, it gives the lowest Type II error β among all tests such that the type one error α is no larger than the permissible level. (We will prove this later.)

Example: Testing lot quality

- Consider a simple scenario where we stipulate that the failure rate of the lot λ is either

$$H_0 : \lambda = 1.0 \; (= \theta_0)$$

or

$$H_1 : \lambda = 2.0 \; (= \theta_1).$$

- Also suppose the data is a single measured lifetime X with density

$$f(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- The likelihood ratio test rejects when

$$\frac{f(X|\lambda_1)}{f(X|\lambda_0)} > K.$$

- Substituting the pdf of X into the test we get

$$\frac{2e^{-2X}}{e^{-X}} > K.$$

- That is, reject when

$$2e^{-X} > K,$$

  or equivalently, when

$$X < -\log(K/2) = C.$$

- So the corresponding rejection region is

$$\mathscr{R} = \{x : x < C\}.$$

- The test can be described as

  "Reject H_0 if $X < C$ and decide that the lot is bad; otherwise decide the lot is good."

- The theory that suggests the likelihood ratio test being the test of choice only guides us to the form of the test—in the current example "reject H_0 if $X < C$", but doesn't say anything about the value of C.
- To completely specify the test, however, we need to determine the value of C, or equivalently that of K.
- Note that C determines the "size" of rejection region $\mathscr{R}$.
- This is where we use the pre-specified α level of the test. This can in fact arise naturally in applications.

- Suppose the company can tolerate a 10% Type I error rate.
- In the current application a Type I error is when the lot is good $H_0 : \lambda = 1.0$ but we say its bad (i.e. reject H_0).
- In other words, the company allows a good lot to be deserted 10% of the times.
- Note that

$$\begin{aligned}\alpha &= \Pr(\text{Reject } H_0 \mid H_0 \text{ is true}) \\ &= \Pr(X < C|H_0 \text{ is true}) \\ &= \Pr(X < C|\lambda = 1.0) \\ &= 1 - e^{-C}\end{aligned}$$

- We set $\alpha = 0.1$ and solve for C, we get

$$C = -\log(1-\alpha) \approx 0.105.$$

- So the best level $\alpha = 0.1$ test, reject H_0 when $X < 0.105$.

- Once the test is completely specified by setting the Type I error rate, we can calculate what the Type II error rate is.
- The Type II error rate is

$$\begin{aligned}\beta &= \Pr(\text{accept } H_0 \mid H_1 \text{ is true}) = P(X \geq C|\lambda = 2.0) \\ &= e^{-2C} = e^{-2\times 0.105} \approx 0.81.\end{aligned}$$

- This is a very high error rate, but is the lowest possible for $\alpha = 0.1$.
- Why is the Type II error rate so high?
- The high Type II error rate simply reflects the fact that we only have one observation and the amount of information we get from our data is very limited.

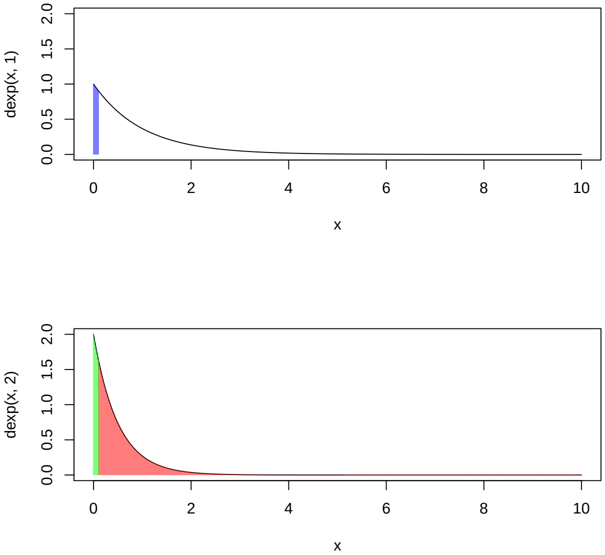
- Another characteristic of a test that is commonly used is the *power* of a test:

$$\pi = 1 - \beta = \Pr(\text{Reject } H_0 \mid H_1 \text{ is true}).$$

- That is, the power is simply one minus the Type II error rate.
- The power is the probability that one rejects the null hypothesis when it is indeed false.
- So minimizing β for a fixed level of α is the same as maximizing power π for a fixed level of α.
- For the ongoing example, the test of the likelihood ratio test with $\alpha = 0.1$ is

$$\pi = 1 - \beta = 1 - .81 = .19$$

  which is quite low.



Blue: Type I error rate; Red: Type II error rate; Green: Power.

Example: testing a normal mean, with known variance.

- Suppose the data are i.i.d. observation $X_1, X_2, \ldots, X_n$ from a $\mathrm{N}(\mu, \sigma^2)$.
- The parameter $\theta = (\mu, \sigma^2)$.
- A simple hypothesis test may be comparing

$$\theta_0 = (\mu_0, \sigma_0^2) \quad \text{vs. } \theta_1 = (\mu_1, \sigma_0^2)$$

where $\mu_0 < \mu_1$.

What is the likelihood ratio test in this case?

$$\frac{L(\theta_1)}{L(\theta_0)} = \frac{(2\pi)^{-n/2}\sigma_0^{-n} e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu_0)^2}}{(2\pi)^{-n/2}\sigma_0^{-n} e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu_1)^2}}$$

$$= e^{-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{n}(X_i-\mu_1)^2-\sum_{i=1}^{n}(X_i-\mu_0)^2\right]}$$

$$= e^{\frac{1}{\sigma_0^2}\left[(\mu_1-\mu_0)\sum_{i=1}^{n}X_i\right]} e^{-\frac{n}{2\sigma_0^2}(\mu_1^2-\mu_0^2)}$$

- *Because $\mu_1 > \mu_0$,* this ratio is $> K$ for some constant K if and only if $\sum_{i=1}^{n} X_i$ or $\bar{X}$ is greater than some constant C.
- So the likelihood ratio test rejects when $\bar{X} > C$, where C remains to be determined.
- Note that choosing C is equivalent to choosing K.

The choice of C

- Similar to the previous example, the theory (which says that the LR test is the "best" test to use) only gives the general form of the test—the rejection region is of the form $\bar{X} > C$.
- To determine the constant C, again we specify a threshold α for Type I error rate that we allow—the "size" of the test.

- Following from the definition of Type I error, we want to choose C such that
$$P(\bar{X} > C|H_0) \leq \alpha.$$
- Since we want to make $P(\bar{X} > C|H_1)$ as large as we can, we want to find the smallest C such that the above inequality holds.
- Now we know the sampling distribution of $\bar{X}$ under H_0 is $\mathrm{N}(\mu_0, \sigma_0^2/n)$, so

$$\begin{aligned}\alpha &= P(\bar{X} > C|H_0)\\ &= P\left(\frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} > \frac{C-\mu_0}{\sigma_0/\sqrt{n}}\middle| H_0\right)\\ &= P\left(Z > \frac{C-\mu_0}{\sigma_0/\sqrt{n}}\right)\end{aligned}$$

where Z is a standard normal random variable.

Therefore we set

$$1 - \Phi\left(\frac{C - \mu_0}{\sigma_0/\sqrt{n}}\right) = \alpha$$

and so

$$C = \mu_0 + \Phi^{-1}(1-\alpha)\frac{\sigma_0}{\sqrt{n}}.$$

- The *level* α *(or size* α*)* LR test thus rejects when

$$\bar{X} > \mu_0 + \Phi^{-1}(1-\alpha)\frac{\sigma_0}{\sqrt{n}}.$$

- For example, if $n = 10$, $\mu_0 = 15$, $\mu_1 = 17$, $\sigma_0 = 2$, and we choose $\alpha = 0.05$, then

$$C = 15 + 1.645 \times \frac{2}{\sqrt{10}} \approx 16.04.$$

  So the level 5% LR test rejects when $\bar{X} > 16.04$.

- Note that in the calculation for choosing C we have not used the alternative hypothesis $(\mu_1, \sigma_0) = (17, 2)$.
- Indeed, once we have the specific form of a test, e.g. $\bar{X} > C$, then the Type I error rate of this test is completely determined by the null hypothesis alone.
- So we used only the null hypothesis to pin down the constant C.
- However, the Type II error rate, and thus the power, of this test depend on the alternative hypothesis.

Type II error rate and power of the test

- The Type II error rate of our test reject when $\bar{X} > C$ is

$$\beta = P(\bar{X} \leq C | H_1).$$

- Note that under H_1, $\bar{X} \sim \mathrm{N}(\mu_1, \sigma_0^2/n)$ so

$$\beta = P\left(\frac{\bar{X} - \mu_1}{\sigma_0/\sqrt{n}} \leq \frac{C - \mu_1}{\sigma_0/\sqrt{n}} \,\Big|\, H_1\right) = \Phi\left(\frac{C - \mu_1}{\sigma_0/\sqrt{n}}\right).$$

- For $C = 16.04$, $\mu_1 = 17$, $\sigma_0 = 2$ and $n = 10$, we have

$$\beta = \Phi\left(\frac{16.04 - 17}{2/\sqrt{10}}\right) = P(Z \leq -1.52) = .0643.$$

- Correspondingly, the power of the test is

$$\pi = P(\bar{X} > C) = 1 - \beta = .9357.$$

The above examples illustrate a general procedure for carrying our simple hypothesis tests.

- First we choose a particular form of the test. In the previous examples we chose the likelihood ratio test.
- After deriving the specific form of the test, (e.g. reject when $\bar{X} > C$), we choose a particular Type I error rate threshold to determine the corresponding constant in the test.
- At this point the test is completely specified.
- Then we can compute the Type II error and the power of the test using the specific alternative.

Why LR test?

- This is because there is a simple elegant result due to Neyman and Pearson that states that the "best" level α test is the LR test.
- Here "best" means achieving the lowest Type II error rate, or equivalently the highest power, among all possible level α tests.
- Next we present the interesting result by Neyman and Pearson more formally.

The Neyman-Pearson (NP) Lemma

In the problem of testing simple hypotheses

$H_0 : X$ has distribution $f(x|\theta_0)$ vs. $H_1 : X$ has distribution $f(x|\theta_1)$

given a Type I error rate α, no test with the same or lower α has a lower Type II error rate β than the likelihood ratio test.

- Before we present a formal proof of the lemma, let us understand it intuitively.
- A gold miner analogy.
- (Draw a figure.)

Indicator representation of tests

- For any other test with a rejection region $\mathscr{R}$, the two error rates satisfy
$$\alpha_{\mathscr{R}} = P(X \in \mathscr{R} | \theta_0) = E(\mathbf{1}_{X \in \mathscr{R}} | \theta_0)$$
and
$$1 - \beta_{\mathscr{R}} = P(X \in \mathscr{R} | \theta_1) = E(\mathbf{1}_{X \in \mathscr{R}} | \theta_1).$$

Proof of NP Lemma

There are different ways to prove the lemma. One that uses the indicator function is particularly simple.

- The Type I error and Type II error rates of the LR test satisfy

$$\alpha_{LR} = P(X \in \mathscr{R}_{LR} | \theta_0) = E(\mathbf{1}_{X \in \mathscr{R}_{LR}} | \theta_0)$$

  and

$$1 - \beta_{LR} = P(X \in \mathscr{R}_{LR} | \theta_1) = E(\mathbf{1}_{X \in \mathscr{R}_{LR}} | \theta_1).$$

- Next, we claim that for all x,

$$\mathbf{1}_{x \in \mathscr{R}_{LR}} \big(f(x|\theta_1) - Kf(x|\theta_0)\big) \geq \mathbf{1}_{x \in \mathscr{R}} \big(f(x|\theta_1) - Kf(x|\theta_0)\big)$$

  for all possible level α rejection region $\mathscr{R}$.
- Why?
- Now that we know the inequality holds, after rearranging terms we get

$$\mathbf{1}_{x \in \mathscr{R}_{LR}} f(x|\theta_1) - \mathbf{1}_{x \in \mathscr{R}} f(x|\theta_1) \geq K \mathbf{1}_{x \in \mathscr{R}_{LR}} f(x|\theta_0) - K \mathbf{1}_{x \in \mathscr{R}} f(x|\theta_0).$$

- Now integrate both sides over x we get

$$E(\mathbf{1}_{X\in\mathscr{R}_{LR}}|\theta_1) - E(\mathbf{1}_{X\in\mathscr{R}}|\theta_1) \geq K\,E(\mathbf{1}_{X\in\mathscr{R}_{LR}}|\theta_0) - K\,E(\mathbf{1}_{X\in\mathscr{R}}|\theta_0).$$

- That is,

$$(1-\beta_{LR}) - (1-\beta_{\mathscr{R}}) \geq K\,(\alpha_{LR} - \alpha_{\mathscr{R}})$$

- Thus for $\alpha_{\mathscr{R}} \leq \alpha_{LR}$, we always have $\beta_{LR} \leq \beta_{\mathscr{R}}$. This completes the proof of the lemma.

Summary

- For simple hypotheses, the LR is “best”.
- We use the general form of the test

$$\frac{L(\theta_1)}{L(\theta_0)} > K$$

  to determine the specific form of the test in each problem.
- We then choose the corresponding constant using the pre-specified Type I error rate using the null hypothesis.
- We can then evaluate the Type II error rate and power of the test based on the alternative hypothesis.