STAT 111, Week 6 - Hypothesis Tests

Leia Donaway — Swarthmore College

March 1, 2025

This week we are working from the Hypothesis Tests Week 6 guidelines from Professor Everson which are on Moodle. The second part is Likelihood Ratio Tests, which you can read more about in Rice 9.2, which these notes make heavy reference to.

A) State the Neyman-Pearson lemma and define the likelihood ratio test of two simple hypotheses.For the defect rate example in presentation 1, show that the likelihood ratio test rejects for large values of the sample proportion of defects.

From Rice page 332, "Suppose that H_0 and H_1 are simple hypotheses and that the test that rejects H_0 whenever the likelihood ratio is less than c and significance level α. Then any other test for which the significance level is less than or equal to α has power less than or equal to that of the likelihood ratio test."

For $X_1, \ldots, x_n \sim N(\mu, \sigma)$ with known variance σ, consider the simple hypotheses
$H_0 : \mu = \mu_0$ vs. $H_A : \mu = \mu_1$, at given significance level α.

The Likelihood Ratio is given as $\Lambda(x) = \frac{L(\mu_0|x)}{L(\mu_1|x)}$
Where we reject H_0 if $\Lambda(x) \leq c$ for some constant c chosen to achieve significance level α.

The Neyman-Pearson lemma states that among all tests at level α, the test that rejects for small values of the Likelihood Ratio is most powerful.

For the Defect Rate Example from presentation 1, we would test whether the defect rate θ is .1 or higher, using sample size $n = 400$. Our hypotheses are
$H_0 : \theta = \theta_0 = .1$ vs. $H_A : \theta = \theta_1 > .1$

Our Likelihood Ratio will be set up as $\Lambda = \frac{L(\theta_0|X_i)}{L(\theta_1|X_i)}$

For Binomial Random variables, $X_1, \ldots, X_n \sim^{iid} \text{Binom}(1, \theta)$, $L(\theta) = \theta^{\sum X_i}(1-\theta)^{n-\sum X_i}$ so $\Lambda = \frac{L(\theta_0)}{L(\theta_1)} = (\frac{\theta_0}{\theta_1})^{\sum X_i}(\frac{1-\theta_0}{1-\theta_1})^{n-\sum X_i} = [\frac{\theta_0}{\theta_1}\frac{(1-\theta_1)}{(1-\theta_0)}]^{\sum X_i}[\frac{1-\theta_0}{1-\theta_1}]^n = [(\frac{\theta_0}{1-\theta_0})(\frac{1-\theta_1}{\theta_1})]^{\sum X_i}[\frac{1-\theta_0}{1-\theta_1}]^n = [\frac{\theta_0/(1-\theta_1)}{\theta_1/(1-\theta_1)}]^{\sum X_i}[\frac{1-\theta_0}{1-\theta_1}]^n < 1$

So, when $\theta_1 > \theta_0$, Λ decreases as $\sum X_i$ increases.

The Neyman Pearson lemma says we should reject H_0 if Λ is small, which corresponds to when $\sum X_i$ is large. Another way to think about it is, when $\theta_a > .1$, which is what we're testing for, the ratio $\frac{.1}{\theta_a} < 1$. As $\sum X_i$ increases, the Likelihood Ratio decreases. Rejecting H_0 when $\sum X_i > c$ some constant means we're rejecting when the sample proportion $\frac{x}{n}$ of defects is high, so the test is working as we'd intuit.

B) Give a Bayesian justification for the likelihood ratio as a decision tool. Show that the conditional (posterior) probability or odds of H_0 decreases with the likelihood ratio, no matter what prior probabilities are assigned. Explain how the NP paradigm defines a decision rule without assigning prior probabilities (in order to remain 'objective'). Point out the limitations of this approach. For example, a test for Lyme disease is positive with probability 0.9 for a person who has Lyme (H_a), and with probability 0.05 for a person who does not have Lyme (H_0). Define a likelihood ratio test with significance level $\alpha = 0.05$.

In Bayesian statistics, the posterior probability of the tested hypothesis, given the data observed, is proportional to the product of the prior probability of the hypothesis and its Likelihood. So, if we want to show what happens to the conditional probability (odds) of H_0 and H_A, their posterior odds are
$\mathbb{P}(H_0|x) = \frac{\mathbb{P}(x|H_0)\mathbb{P}(H_0)}{\mathbb{P}(x)}$ $\mathbb{P}(H_A|x) = \frac{\mathbb{P}(x|H_A)\mathbb{P}(H_A)}{\mathbb{P}(x)}$ where x represents our data.
Then, the ratio of the posterior odds $\frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_A|x)} = \frac{\mathbb{P}(x|H_0)}{\mathbb{P}(x|H_A)} \times \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_A)}$. We can recognize each ratio as the posterior probability (odds) ratio = Likelihood Ratio $\times$ (prior) probability ratio. So, we see that no matter what prior probabilities are assigned, the conditional (posterior) probability decreases as the Likelihood Ratio does.

The Neyman-Pearson paradigm defines decision using only the value of the likelihood ratio, and does not assign prior probabilities. The rejection region is chosen to control the Type I error α. This is considered objective since it doesn't bring into account prior beliefs. However, there are limitations to this approach. Failure to consider prior beliefs can lead to decisions that don't make sense in real world contexts. An example of this might be false positives occurring for rare events, despite low α levels. Said a bit differently, when we exclude prior beliefs, we only have the ability to tailor α to our specific test, we want to ask when level of variation would be surprising enough to reject H_0? If we default to .05 due to it being standard, are we making the most sensical decision.

Lyme Disease Example. We are given that the test is correctly positive (shows positive for a patient with Lyme Disease) with probability 0.9, and falsely positive 5% of the time. $\mathbb{P}(+|\text{Lyme}) = .9$ and $\mathbb{P}(+|\text{No Lyme}) = 0.05$. We'll test H_0 : No Lyme disease, vs. H_A : Having Lyme, at significance level $\alpha = .05$

$\Lambda = \frac{\mathbb{P}(+|H_0)}{\mathbb{P}(+|H_A)} = \frac{0.05}{0.9} = 0.0\bar{5}$. For $\alpha = .05$, we'll choose our threshold constant c such that $\mathbb{P}(\Lambda \leq c|H_0) = 0.05$
In this simple example, we have a Likelihood Ratio greater than our threshold, leading us to always reject H_0 when a Lyme disease test comes back positive. However, this is a situation in which being able to include our prior beliefs would be almost as important as the test result. If someone is totally healthy or has no reason to believe the test should be positive, then a positive result is most likely false, but because of our setup, we reject H_0 for a positive test.

C) As an example where the LR test has higher power than an alternate test with the same significance level, consider a test of $H_0 : \mu = 0$ vs. $H_a : \mu = 3$ for the NFL home field advantage data with $n = 272$ and assuming the data are iid Normal with $\sigma = 14.0$ known. An alternative test to the usual z-test is based on the count Y of positive values, with $Y \sim \text{Binom}(n, 0.5)$ under H_0. Find an α close to 0.05 that you can achieve exactly with a Binomial test. Show that the test that rejects for large $\bar{x}$ has higher power than the test that rejects for large Y. Note, however, how the test based on Y (the "sign test") is valid even if the data are not Normal.

We want to look at an example where the power of the Likelihood Ratio test can be compared to that of an alternate hypothesis test of the same significance level. Using the NFL home team data as an example, we will do a sign test and a usual z-test for $\bar{x}$.

Alternative test (Sign Test):
Y = count of positive values (games won by the home team)
Y$\sim$ Binom$(272, 0.5)$ under H_0, as we assume there is no advantage. Find the smallest integer c, such that $\mathbb{P}(Y \geq c|H_0) \leq 0.05$
$\mathbb{P}(Y \geq a|H_0) = \sum_{y=a}^{272} \mathbb{P}(Y = y)$, checking each value until desired α, so $\mathbb{P}(Y \geq 151|H_0) = 0.04997$, thus at $\alpha = 0.04997$, reject H_0 for $Y \geq 151$.
In R: min(which(1-pbinom(0:(271),272,.5)))$\leq$0.05

Likelihood Ratio Test:
The sample mean $\bar{X} \sim N(0, \frac{14}{\sqrt{272}})$ under $H_0 : \mu = 0$, rejecting for large values of $\bar{X}$. We'll calculate c such that $\mathbb{P}(\bar{X} \geq c|H_0) = 0.04997$, $c = 1.396 \approx 1.4$. So, we reject H_0 when $\bar{X} \geq 1.4$.
In R: qnorm(1-.04997),0,14/sqrt(272))

Compute the power of each test under $H_A : \mu = 3$:
Under H_A, $\bar{X} \sim N(3, \frac{14}{\sqrt{272}})$ and the power is $\mathbb{P}(\bar{X} \geq 1.4|H_A) = \mathbb{P}(\frac{\bar{X}-3}{14/\sqrt{272}} \geq \frac{1.4-3}{14/\sqrt{272}}) \approx 0.970$ power
In R: 1-pnorm(1.4,3,14/sqrt(272))
Whereas for Y, each observation has $\mathbb{P}(x > 0) = \mathbb{P}(Z > \frac{0-3}{14}) = \mathbb{P}(Z > -0.214)$, p ≈ 0.58
$\mathbb{P}(Y \geq 151|\text{Binom}(272, 0.58)) \approx 0.854$ power
In R: 1-pbinom(150,272,(1-pnorm(0,3,14))

The Likelihood Ratio test for $\bar{X}$ is more powerful than the alternate sign test for hypotheses under the same α level, which is a demonstration of Neyman-Pearson. Then, why would we use the sign test? The sign test is nonparametric, so we need only to assume the data are independent to use it– it works when the data aren't distributed Normally.

D) Outline the formal proof of the Neyman Pearson lemma to show that no test of comparable significance level has higher power than the likelihood ratio test. See Rice 9.2.

See the following handwritten notes expanding on the Rice 9.2 proof of Neyman-Pearson.

Rice 9.2

D) Outline the formal proof of the Neyman Pearson Lemma

- Show that no test of comparable significance level has higher power than a LR Test.

Definitions and Background

Our hypotheses

$H_0: f(x) = f_0(x)$ vs. $H_1: f(x) = f_A(x)$

where $f(x)$ is the pdf of our data

will use the decision function $d(x) = \begin{cases} 0 & \text{if we fail to reject} \\ 1 & \text{if we reject } H_0 \end{cases}$

- $d(X)$ is a Bernoulli random variable, and so we know that $\mathbb{E}[d(X)] = \mathbb{P}(d(X)=1)$
- We know also that the significance level alpha is equal to $\alpha = \mathbb{P}_0(d(X)=1)$, so $\alpha = \mathbb{E}_0[d(X)]$ and power $= \mathbb{P}_A(d(X)=0) = \mathbb{E}_A[d(X)]$,

  with E_0 meaning the Expectation under the specification of H_0.
- These definitions of significance level and power make sense if we think about their definitions in terms of Type I / II Errors and how our decision function $d(x)$ is defined. Now, consider two tests
- Let $d(X)$ correspond to the outcome of the Likelihood Ratio Test, such that $d(x)=1$ IF $f_0(x) < c\, f_A(x)$ and $\mathbb{E}[d(X)] = \alpha$ AND

Let $d^*(x)$ be the decision function of a seperate, non-LRT for hypotheses. We know this test has a significance level comparable, so $\mathbb{E}_0[d^*(X)] \leq \mathbb{E}_0[d(X)] = \alpha$.

What is the Goal? We WTS $\mathbb{E}_A[d^*(X)] \leq \mathbb{E}_A[d(X)]$ because (remember) this means the power of any alternative test for hypotheses at our desired α level is less powerful than the Likelihood ratio test

Proof

- Rice proposes the inequality

$$d^*(x)[c f_A(x) - f_0(x)] \leq d(x)[c f_A(x) - f_0(x)]$$

which must be true because

- If $d(x) = 1$, or we reject H_0, we know $f_0(x) < c f_A(x)$, so $c f_A(x) - f_0(x) > 0$, mean the RHS is equal to $c f_A(x) - f_0(x)$, meaning if $d^*(x) = 0$, then $0 \leq c f_A(x) - f_0(x)$, if $d^*(x) = 1$ the LHS=RHS

- If $d(x) = 0$, then we accept H_0 so $c f_A(x) - f_0(x) \leq 0$ we have $d^*(x)[\text{negative}] \leq (0)(\text{negative})$ which also holds whether $d^* = 0$ or 1.

- From the key inequality, we can sum both sides with respect to x, the secret being that pdfs will become expectations definitionally.

$$c\mathbb{E}_A[d^*(X)] - \mathbb{E}_0[d^*(x)] \leq c\mathbb{E}_A[d(X)] - \mathbb{E}_0[d(X)]$$

rearranging: $\underbrace{\mathbb{E}_0[d(X)] - \mathbb{E}_0[d^*(X)]}_{\text{we have that this is nonnegative by assumption}} \leq c\mathbb{E}_A[d(X)] - \mathbb{E}_A[d^*(X)]$

from which we get $\mathbb{E}_A[d^*(X)] \leq \mathbb{E}_A[d(X)]$, as c is a + constant

□.