Chapter 8

Hypothesis Tests

Extensions on the Likelihood Ratio

One and Two Sided Tests

Outline

Introduction

For a composite hypothesis

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1,$$

we have seen critical regions defined by taking a statistic $T(\mathbf{x})$ and defining the critical region based on a critical value $\tilde{k}_\alpha$. For a one-sided test, we have seen critical regions

$$\{\mathbf{x}; T(\mathbf{x}) \geq \tilde{k}_\alpha\} \quad \text{or} \quad \{\mathbf{x}; T(\mathbf{x}) \leq \tilde{k}_\alpha\}.$$

For a two-sided test, we saw

$$\{\mathbf{x}; |T(\mathbf{x})| \geq \tilde{k}_\alpha\}.$$

$\tilde{k}_\alpha$ is determined by the level α. We thus use commands `qnorm`, `qbinom`, or `qhyper` when $T(\mathbf{x})$ has, respectively, a normal, binomial, or hypergeometric distribution under a appropriate choice of $\theta \in \Theta_0$. We now examine extensions of the likelihood ratio test for simple hypotheses that have desirable properties for a critical region.

One-Sided Tests

In testing for the invasion of a mimic butterfly by a model species, we collected a simple random sample modeled as independent normal observations with unknown mean and known variance σ_0^2.

We discovered, in the case of a simple hypothesis test,

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu = \mu_1$$

that the critical region as determined by the Neyman-Pearson lemma depends only on whether or not μ_1 was greater than μ_0. For example, if $\mu_1 > \mu_0$, then the critical region

$$C = \{\mathbf{x}; \bar{x} \geq \tilde{k}_\alpha\}$$

shows that we reject H_0 whenever the sample mean is higher than some threshold value $\tilde{k}_\alpha$ *irrespective* of the difference between μ_0 and μ_1.

One-Sided Tests

- If a test is most powerful against *each* possible alternative in a simple hypothesis test, when we can say that this test is in some sense *best overall* for a composite hypothesis?
- Does this test have the property that its power function π is greater for every value of $\theta \in \Theta_1$ than the power function of *any* other test? Such a test is called uniformly most powerful.
- We can hope for such a test if the procedures from simple hypotheses results in a common critical region for all values of the alternative.
- In the example above using independent normal data. In this case, the power function

$$\pi(\mu) = P_\mu\{\bar{X} \geq \tilde{k}_\alpha\}$$

increases as μ increases and so the test has the intuitive property of becoming more powerful with increasing μ.

Karlin-Rubin Theorem

In general, we look for a test statistic $T(\mathbf{x})$. Next, we check that the likelihood ratio,

$$\frac{L(\theta_2|\mathbf{x})}{L(\theta_1|\mathbf{x})}, \quad \theta_1 < \theta_2.$$

depends on the data $\mathbf{x}$ only through the value of statistic $T(\mathbf{x})$ and, in addition, this ratio is a monotone increasing function of $T(\mathbf{x})$.

Note that for any sufficient statistic, $T(\mathbf{x})$, we have by the Fisher-Neyman factorization theorem,

$$\frac{L(\theta_2|\mathbf{x})}{L(\theta_1|\mathbf{x})} = \frac{h(\mathbf{x})g(\theta_2, T(\mathbf{x}))}{h(\mathbf{x})g(\theta_1, T(\mathbf{x}))} = \frac{g(\theta_2, T(\mathbf{x}))}{g(\theta_1, T(\mathbf{x}))}.$$

and thus the likelihood ratio depends only on $T(\mathbf{x})$.

Karlin-Rubin Theorem

The Karlin-Rubin theorem states:

If these conditions hold, then for an appropriate value of $\tilde{k}_\alpha$,

$$C = \{\mathbf{x}; T(\mathbf{x}) \geq \tilde{k}_\alpha\}$$

is the critical region for a uniformly most powerful α level test for the one-sided alternative hypothesis

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

Proof. Let $\pi(\theta)$ be the power function for this test. We first show that $\pi(\theta)$ is an increasing function of θ.

Karlin-Rubin Theorem

We can write the monotone increasing function property in terms of the density function for T. For $\theta_1 < \theta_2,\ t_1 < t_2$.

$$
\begin{aligned}
\frac{f_T(t_2|\theta_2)}{f_T(t_2|\theta_1)} &\geq \frac{f_T(t_1|\theta_2)}{f_T(t_1|\theta_1)} \\
f_T(t_1|\theta_1)f_T(t_2|\theta_2) &\geq f_T(t_1|\theta_2)f_T(t_2|\theta_1)
\end{aligned}
$$

Now, integrate both sides with respect to t_1 on $(-\infty, t_2)$ to obtain

$$
\begin{aligned}
F_T(t_2|\theta_1)f_T(t_2|\theta_2) &\geq F_T(t_2|\theta_2)f_T(t_2|\theta_1) \\
\frac{f_T(t|\theta_2)}{f_T(t|\theta_1)} &\geq \frac{F_T(t|\theta_2)}{F_T(t|\theta_1)}
\end{aligned}
$$

for all t.

Karlin-Rubin Theorem

$$f_T(t_1|\theta_1)f_T(t_2|\theta_2) \geq f_T(t_1|\theta_2)f_T(t_2|\theta_1)$$

Now, integrate both sides with respect to t_2 on (t_1, ∞) to obtain

$$\begin{aligned} f_T(t_1|\theta_1)(1 - F_T(t_1|\theta_2)) &\geq f_T(t_1|\theta_2)(1 - F_T(t_2|\theta_1)) \\ \frac{1 - F_T(t|\theta_2)}{1 - F_T(t|\theta_1)} &\geq \frac{f_T(t|\theta_2)}{f_T(t|\theta_1)} \end{aligned}$$

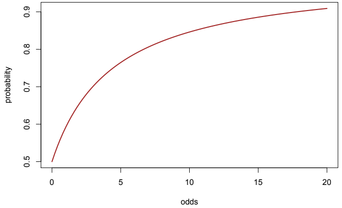
for all t. Thus,

$$\begin{aligned} \frac{1 - F_T(t|\theta_2)}{1 - F_T(t|\theta_1)} &\geq \frac{f_T(t_2|\theta_2)}{f_T(t_2|\theta_1)} \geq \frac{F_T(t|\theta_2)}{F_T(t|\theta_1)} \\ \frac{F_T(t|\theta_1)}{1 - F_T(t|\theta_1)} &\geq \frac{F_T(t|\theta_2)}{1 - F_T(t|\theta_2)} \end{aligned}$$

Karlin-Rubin Theorem

Now, the mapping from probability to odds $p \mapsto p/(1-p)$ is one-to-one and increasing. So is its inverse $o \mapsto o/(1+o)$

$$F_T(t|\theta_1) \geq F_T(t|\theta_2) \quad \text{for all } t$$



The power

$$\pi(\theta_1) = P_{\theta_1}\{T > \tilde{k}_\alpha\} = 1 - F_T(\tilde{k}_\alpha|\theta_1) \leq 1 - F_T(\tilde{k}_\alpha|\theta_2) = P_{\theta_2}\{T > \tilde{k}_\alpha\} = \pi(\theta_2)$$

and $\pi(\theta)$ is an increasing function of θ.

If we set $\tilde{k}_\alpha$ so that $\pi(\theta_0) = P_{\theta_0}\{T > \tilde{k}_\alpha\} = \alpha$, then $\pi(\theta) \leq \alpha$ for $\theta \leq \theta_0$ and so the test is an α-level test.

Karlin-Rubin Theorem

Next, set $\tilde{\theta} > \theta_0$ and consider the simple hypothesis

$$\tilde{H}_0 : \theta = \theta_0 \quad \text{versus} \quad \tilde{H}_1 : \theta = \tilde{\theta}_0.$$

Because the likelihood ratio is monotone in $T(\mathbf{X})$, the requirement that $T(\mathbf{X}) > \tilde{k}_\alpha$, is equivalent to the the likelihood ratio exceeding some value (say k_α). Thus, the critical region determined by a threshold level for $T(\mathbf{X})$ is also a threshold level for the likelihood ratio. Thus, by the Neyman-Pearson lemma, this critical region is most powerful.

Because this holds for every value of $\tilde{\theta}$, the test is simultaneously most powerful for every $\tilde{\theta} > \theta_0$, thus it uniformly most powerful. QED

Karlin-Rubin Theorem

A corresponding criterion holds for the one sided test a "less than" alternative.

Exercise. Verify that the likelihood ratio is an appropriate monotone function of the given test statistic, T.

1. For mark and recapture, use the hypothesis

$$H_0 : N \geq N_0 \quad \text{versus} \quad H_1 : N < N_0,$$

   use the test statistic $T(\mathbf{x}) = r(\mathbf{x})$, the number tagged in the second capture.

2. For $X = (X_1, \ldots, X_n)$ is a sequence of Bernoulli trials with unknown success probability p, and the one-sided test

$$H_0 : p \leq p_0 \quad \text{versus} \quad H_1 : p > p_0,$$

   use the test statistic $T(\mathbf{x}) = \hat{p}(\mathbf{x})$, the sample proportion of successes.

Binomial Test

If 20 out of 36 bee hives survive a severe winter, for an $\alpha = 0.05$ level test for

$$H_0 : p \geq 0.7 \quad \text{versus} \quad H_1 : p < 0.7,$$

we use the binomial distribution for the number of successes using `binom.test`.

```
> binom.test(20,36,p=0.7,alternative=c("less"))

Exact binomial test


data:  20 and 36
number of successes = 20, number of trials = 36, p-value = 0.04704
alternative hypothesis: true probability of success is less than 0.7
```

Exercise. Do we reject the hypothesis at the 5% level? the 1% level? Find the p-value using the `pbinom` command.

Proportion Test

If 250 out of 336 bee hives survive a mild winter, for an $\alpha = 0.05$ level test for

$$H_0 : p \leq 0.7 \quad \text{versus} \quad H_1 : p > 0.7,$$

we use the normal approximation for the number of successes using `prop.test`.

```
> prop.test(250,336,p=0.7,alternative=c("greater"))

1-sample proportions test with continuity correction

data:  250 out of 336, null probability 0.7
X-squared = 2.8981, df = 1, p-value = 0.04434
```

and we *reject* the null hypothesis.

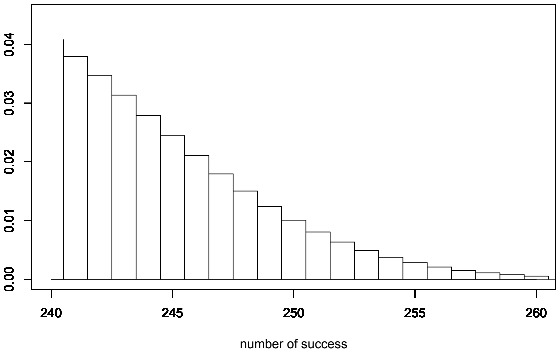
Continuity Correction

The p-value is $P\{X \geq 250\}$ where X is $Bin(336, 0.7)$. We compute this using R.

$$P\{X \geq 250\} = 1 - P\{X \leq 249\}$$

```
> 1-pbinom(249,336,0.7)
[1] 0.0428047
```

The command `prop.test` uses a normal approximation and a continuity correction to obtain a p-value 0.04434

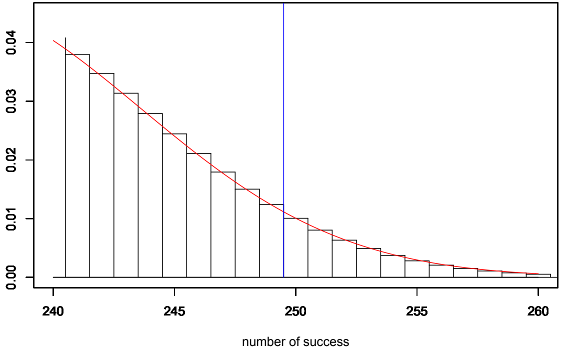


The p-value $P\{X \geq x\} = \sum_{y=x}^{n} P\{X = y\}$ can be realized as the area of rectangles, height $P\{X = y\}$ and width 1.

Continuity Correction

The rectangles look like a Riemann sum for the integral of the density of a $N(np_0, \sqrt{np_0(1-p_0)})$ random variable with lower limit $x - 1/2$.

```
> mu<-0.7*336
> sigma<-sqrt(336*0.7*0.3)
> x<-c(249,249.5,250)
> prob<-1-pnorm(x,mu,sigma)
> data.frame(x,prob)
      x       prob
1 249.0 0.05020625
2 249.5 0.04434199
3 250.0 0.03904269
```



The continuity correction replaces the binomial by finding the area under the normal density with lower limit $x - 1/2$.

Two-Sided Tests

- The likelihood ratio test is a popular choice for composite hypothesis when Θ_0 is a subspace Θ the parameter space.
- The rationale for this approach is that the null hypothesis is unlikely to be true if the maximum likelihood on Θ_0 is sufficiently smaller that the likelihood maximized over Θ. Let
  - $\hat{\theta}_0$ be the parameter value that maximizes the likelihood for $\theta \in \Theta_0$ and
  - $\hat{\theta}$ be the parameter value that maximizes the likelihood for $\theta \in \Theta$.
- The likelihood ratio

$$\Lambda(\mathbf{x}) = \frac{L(\hat{\theta}_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}.$$

Overview

We have two optimization problems - maximize $L(\theta|\mathbf{x})$ on the parameter space Θ and on the null hypothesis space Θ_0.

The critical region for an α-level likelihood ratio test is

$$\{\Lambda(\mathbf{x}) \leq \lambda_\alpha\}.$$

As with any α level test, λ_α is chosen so that

$$P_\theta\{\Lambda(X) \leq \lambda_\alpha\} \leq \alpha \text{ for all } \theta \in \Theta_0.$$

NB. This ratio is the reciprocal from the version given by the Neyman-Pearson lemma. Thus, the critical region consists of those values that are *below* a critical value.

Normal Observations

Consider the two-sided hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

Here the data are n independent $N(\mu, \sigma_0)$ random variables with known variance σ_0^2. The parameter space Θ is one dimensional giving the value μ for the mean. As we have seen before $\hat{\mu} = \bar{x}$. Θ_0 is the single point $\{\mu_0\}$ and so $\hat{\mu}_0 = \mu_0$.

$$L(\hat{\mu}_0|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp -\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i - \mu_0)^2, \quad L(\hat{\mu}|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp -\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i - \bar{x})^2$$

and

$$\Lambda(\mathbf{x}) = \exp -\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^n ((x_i - \mu_0)^2 - (x_i - \bar{x})^2)\right) = \exp -\frac{n}{2\sigma_0^2}(\bar{x} - \mu_0)^2.$$

Notice that

$$-2\ln\Lambda(\mathbf{x}) = \frac{n}{\sigma_0^2}(\bar{x} - \mu_0)^2 = \left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right)^2.$$

Normal Observations

Then, critical region,

$$\{\Lambda(\mathbf{x}) \leq \lambda_\alpha\} = \left\{ \left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \right)^2 \geq -2 \ln \lambda_\alpha \right\}.$$

Under the null hypothesis, $(\bar{X} - \mu_0)/(\sigma_0/\sqrt{n})$ is a standard normal random variable, and thus $-2 \ln \Lambda(X)$ is the square of a single standard normal. This is the defining property of a χ-square random variable with 1 degree of freedom.

Naturally we can use both

$$\left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \right)^2 \quad \text{and} \quad \left| \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \right|.$$

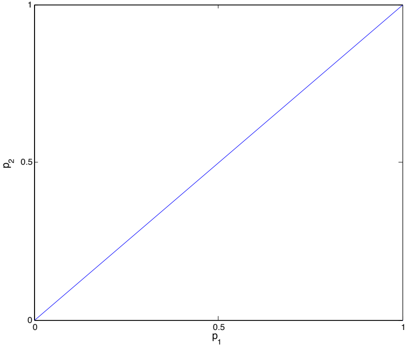
as a test statistic. We have seen the second choice in the example of a possible invasion of a model butterfly by a mimic.

Two-Sample Proportions

For the two-sided two-sample α-level likelihood ratio test for population proportions p_1 and p_2, based on the hypothesis

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 \neq p_2,$$

- we maximize the likelihood over the subspace $\Theta_0 = \{(p_1, p_2); p_1 = p_2\}$ (the blue line) and
- over the entire parameter space, $\Theta = [0, 1] \times [0, 1]$, shown as the square, and
- then take the ratio, simplify and make appropriate approximations.



The data are observations on n_1 Bernoulli trials, $x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}$ from the first population and, independently, n_2 Bernoulli trials, $x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}$ from the second.

Two-Sample Proportions

The likelihood ratio test is approximately equivalent to the critical region

$$|z| \geq z_{\alpha/2}$$

where

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

with $\hat{p}_i$, the sample proportion of successes from the observations from population i and $\hat{p}_0$, the pooled proportion

$$\hat{p}_0 = \frac{1}{n_1+n_2}\left((x_{1,1}+\cdots+x_{1,n_1})+(x_{2,1}+\cdots+x_{2,n_2})\right) = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}.$$

Two-Sample Proportions

The subsequent winter had 167 out of 250 hives surviving. To test if the two survival probabilities are significantly different:

```
> prop.test(c(250,167),c(332,250))

2-sample test for equality of proportions with continuity correction

data:  c(250, 167) out of c(332, 250)
X-squared = 4.664, df = 1, p-value = 0.0308
alternative hypothesis: two.sided
95 percent confidence interval:
 0.006942351 0.163081746
sample estimates:
  prop 1   prop 2
0.753012 0.668000
```

Two-Sample Proportions

Power analyses can be executed in R using the `power.prop.test` command. If we want to be able to detect a difference between two proportions $p_1 = 0.7$ and $p_2 = 0.6$ in a one-sided test with a significance level of $\alpha = 0.05$ and power $1 - \beta = 0.8$.

```
> power.prop.test(p1=0.70,p2=0.6,sig.level=0.05,power=0.8,
  alternative = c("one.sided"))
     Two-sample comparison of proportions power calculation
              n = 280.2581
             p1 = 0.7
             p2 = 0.6
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
 NOTE: n is number in *each* group
```

We will need a sample of $n = 281$ from each group.

Two-Sample Proportions

If we vary p_2 and determine the power.

```
> power.prop.test(n=250,p1=0.70,p2=c(0.6,0.65),sig.level=0.05,
   alternative = c("one.sided"))
             p2 = 0.60, 0.65
          power = 0.7589896, 0.3256442
```

Now, let's vary sample size.

```
> power.prop.test(n=c(250,350,450,550),p1=0.70,p2=0.60,sig.level=0.05,
   alternative = c("one.sided"))
              n = 250, 350, 450, 550
          power = 0.7589896, 0.8717915, 0.9342626, 0.9672670
```

Exercise. Determine the reduction in power when the significance level $\alpha = 0.02$ for the sample sizes above. Why is the power reduced?