Chapter 8

Hypothesis Testing

Multiple Testing

Outline

Partitioning the Parameter Space

Simple hypotheses limit us to a decision between one of two possible states of nature. This limitation does not allow us, under the procedures of hypothesis testing to address the basic question:

> *Does the parameter value θ_0 increase, decrease or change at all under under a different experimental condition?*

This leads us to consider composite hypotheses. In this case, the parameter space Θ is divided into two disjoint regions, Θ_0 and Θ_1. The hypothesis test is now written

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

Again, H_0 is called the null hypothesis and H_1 the alternative hypothesis.

Partitioning the Parameter Space

For the three alternatives to the question posed above, we have

- increase would lead to the choices $\Theta_0 = \{\theta; \theta \leq \theta_0\}$ and $\Theta_1 = \{\theta; \theta > \theta_0\}$,
- decrease would lead to the choices $\Theta_0 = \{\theta; \theta \geq \theta_0\}$ and $\Theta_1 = \{\theta; \theta < \theta_0\}$, and
- change would lead to the choices $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta; \theta \neq \theta_0\}$

for some choice of parameter value θ_0. The effect that we are meant to show, here the nature of the change, is contained in Θ_1. The first two options given above are called one-sided tests. The third is called a two-sided test.

Rejecting the null hypothesis, critical regions, and type I and type II errors have the same meaning for a composite hypotheses. Significance level and power will necessitate an extension of the ideas for simple hypotheses.

The Power Function

Power is now a function of the parameter value θ. If our test is to reject H_0 whenever the data fall in a critical region C, then the power function is defined as

$$\pi(\theta) = P_\theta\{X \in C\},$$

the probability of rejecting the null hypothesis for a given parameter value.

- For $\theta \in \Theta_0$, $\quad \pi(\theta)$ is the probability of making a type I error, i.e., rejecting the null hypothesis when it is indeed true.
- For $\theta \in \Theta_1$, $\quad 1 - \pi(\theta)$ is the probability of making a type II error, i.e., failing to reject the null hypothesis when it is false.

The ideal power function has

$$\pi(\theta) \approx 0 \text{ for all } \theta \in \Theta_0 \text{ and } \pi(\theta) \approx 1 \text{ for all } \theta \in \Theta_1.$$

The Power Function

- The goal is to make the chance for error small.
- One strategy is to consider a method analogous to that employed in the Neyman-Pearson lemma. Thus, we must *simultaneously*,
  - fix a (significance) level α, now defined to be the largest value of $\pi(\theta)$ in the region Θ_0 defined by the null hypothesis,

    By focusing on the value of the parameter in Θ_0 that is most likely to result in an error, we insure that the probability of a type I error is no more that α *irrespective* of the value for $\theta \in \Theta_0$.
  - and look for a critical region that makes the power function as large as possible for values of the parameter $\theta \in \Theta_1$.

The Power Function

Example. Let $X_1, X_2, \ldots, X_n$ be independent $N(\mu, \sigma_0)$ random variables with σ_0 known and μ unknown. For the composite hypothesis for the one-sided test

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0,$$

we use the test statistic from the likelihood ratio test and reject H_0 if the statistic $\bar{x}$ is too large. Thus, the critical region

$$C = \{\mathbf{x}; \bar{x} \geq k(\mu_0)\}.$$

If μ is the true mean, then the power function

$$\pi(\mu) = P_\mu\{X \in C\} = P_\mu\{\bar{X} \geq k(\mu_0)\}.$$

The value of $k(\mu_0)$ depends on the level α of the test.

The Power Function

- As the actual mean μ increases, then the probability that the sample mean $\bar{X}$ exceeds a particular value $k(\mu_0)$ also increases.
- In other words, π is an increasing function.
- Thus, the maximum value of π on $\Theta_0 = \{\mu; \mu \leq \mu_0\}$ takes place for $\mu = \mu_0$.
- Consequently, to obtain level α for the hypothesis test, set

$$\alpha = \pi(\mu_0) = P_{\mu_0}\{\bar{X} \geq k(\mu_0)\}.$$

We now use this to find the value $k(\mu_0)$. When μ_0 is the value of the mean, we standardize to give a standard normal random variable

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}.$$

Choose z_α so that $P\{Z \geq z_\alpha\} = \alpha$. Thus, $P_{\mu_0}\{Z \geq z_\alpha\} = P_{\mu_0}\{\bar{X} \geq \mu_0 + \frac{\sigma_0}{\sqrt{n}} z_\alpha\}$ and $k(\mu_0) = \mu_0 + (\sigma_0/\sqrt{n}) z_\alpha$.

The Power Function

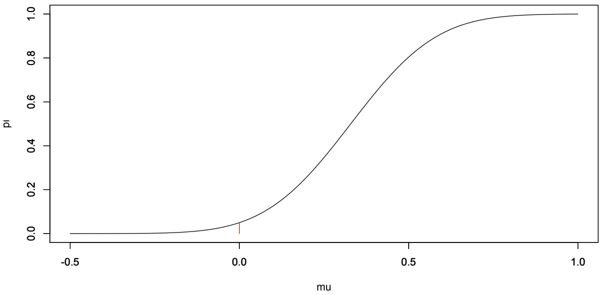
Exercise. If μ is the true state of nature, then

$$Z = \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}}$$

is a standard normal random variable. Use this to show that the power function

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

where Φ is the distribution function for the standard normal.



Power function for the one-sided test with alternative greater. The size of the test α is given by the height of the red segment. Notice that $\pi(\mu) < \alpha$ for all $\mu < \mu_0$ and $\pi(\mu) > \alpha$ for all $\mu > \mu_0$.

The Power Function

We have seen the expression

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

in several contexts.

- If we fix n, the number of observations and the alternative value $\mu = \mu_1 > \mu_0$ and determine the power $1 - \beta$ as a function of the significance level α, then we have the receiving operator characteristic.
- If we fix μ_1 the alternative value and the significance level α, then we can determine the power as a function of n the number of observations.
- If we fix n and the significance level α, then we can determine the power function, $\pi(\mu)$, as a function of the alternative value μ.

Exercise. Give the appropriate expression for π for a less than alternative and use this to plot the power function for the example with a model species and its mimic. Take $\alpha = 0.05$, $\mu_0 = 10$, $\sigma_0 = 3$, and $n = 16$ observations,

The Power Function

To compute sample size for chosen type I and type II errors, let $\mu_1 > \mu_0$.

$$\beta = \Phi\left(z_\alpha - \frac{|\mu_1 - \mu_0|}{\sigma_0/\sqrt{n}}\right)$$

Choose n so that $z_\alpha + \frac{|\mu_1-\mu_0|}{\sigma_0/\sqrt{n}}$ has probability β. However, $\beta = \Phi(-z_\beta)$.

$$\begin{aligned}
-z_\beta &= z_\alpha - \frac{|\mu_1 - \mu_0|}{\sigma_0/\sqrt{n}} \\
\sqrt{n}\frac{|\mu_1 - \mu_0|}{\sigma_0} &= z_\alpha + z_\beta \\
\sqrt{n} &= \frac{\sigma_0}{|\mu_1 - \mu_0|}(z_\alpha + z_\beta) \\
n &= \frac{\sigma_0^2}{(\mu_1 - \mu_0)^2}(z_\alpha + z_\beta)^2
\end{aligned}$$

Choose n^*, any integer al least as large as n.

Mark and Recapture

Mark and recapture can be used as experimental procedure to test whether or not a population has reached a dangerously low level. The variables are

- t be the number captured and tagged,
- k be the number in the second capture,
- r be the number in the second capture that are tagged, and
- N be the total population.

If N_0 is the level that a wildlife biologist say is dangerously low, then the natural hypothesis is one-sided.

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

Mark and Recapture

The data are used to compute r, the number in the second capture that are tagged. The likelihood function for N is the hypergeometric distribution,

$$L(N|r) = \frac{\binom{t}{r}\binom{N-t}{k-r}}{\binom{N}{k}}$$

The maximum likelihood estimate is $\hat{N} = [tk/r]$. Thus, higher values for r lead us to lower estimates for N. Let R be the (random) number in the second capture that are tagged, then, for an α level test, we look for the minimum value r_α so that

$$\pi(N) = P_N\{R \geq r_\alpha\} \leq \alpha \textit{ for all } N \geq N_0.$$

As N increases, then recaptures become less likely and the probability above decreases. Thus, we set the value of r_α according to the parameter value N_0, the minimum value under the null hypothesis.

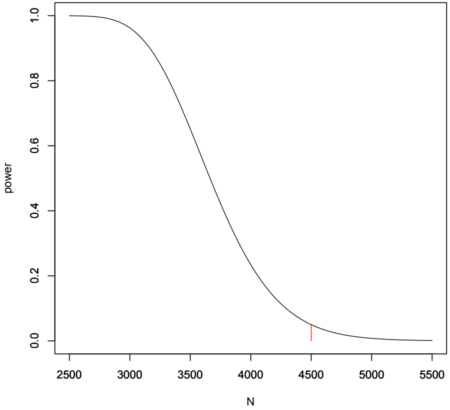
Mark and Recapture

To determine r_α for $\alpha = 0.05, 0.02, 0.01$,

```
> N0<-4500;t<-400;k<-500
> alpha<-c(0.05,0.02,0.01)
> ralpha<-qhyper(1-alpha,t,N0-t,k)
> data.frame(alpha,ralpha)
  alpha ralpha
1  0.05     54
2  0.02     57
3  0.01     59
```

The power curve $\pi(N) = P_N\{R \geq r_{0.05}\}$ is given using the R commands

```
> N<-c(2500:5500)
> power<-1-phyper(54,t,N-t,k)
> plot(N,power,type="l",ylim=c(0,1))
```



Power curve. Vertical red segment at $N = N_0 = 4000$ has height $\alpha = 0.05$.

Mark and Recapture

Note that we must capture al least $r_\alpha = 54$ that were tagged in order to reject H_0 at the $\alpha = 0.05$ level. In this case the estimate for N is

$$\hat{N} = \left[\frac{kt}{r_\alpha}\right] = 3703$$

is well below $N_0 = 4500$.

Exercise. Determine the type II error rate for $N = 4500$ with

- $k = 800$ and $\alpha = 0.05$, 0.02, 0.01, and
- $\alpha = 0.05$ and $k = 600$, 800, and 1000.

Sample Proportion

Honey bees store honey for the winter. This honey serves both as nourishment and insulation from the cold. Typically for a given region, the probability of survival of a feral bee hive over the winter is $p_0 = 0.7$. To check whether this winter has a different survival probability, we consider the hypotheses

$$H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0.$$

If we use the central limit theorem, then, under the null hypothesis,

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

has a distribution approximately that of a standard normal random variable. We reject if $|z|$ is too big.

Sample Proportion

For an α level test, the critical value is $z_{\alpha/2}$. The critical region

$$C = \left\{ \hat{p}; \left| \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right| > z_{\alpha/2} \right\}.$$

For this study, we examine 336 colonies and find that 250 survive.

Exercise. For $\alpha = 0.05$, determine whether or not we reject H_0.

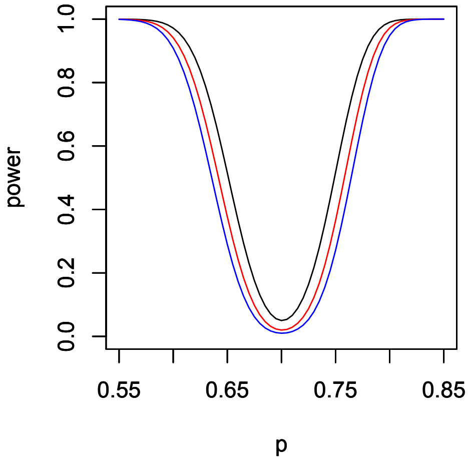
Exercise. Show that

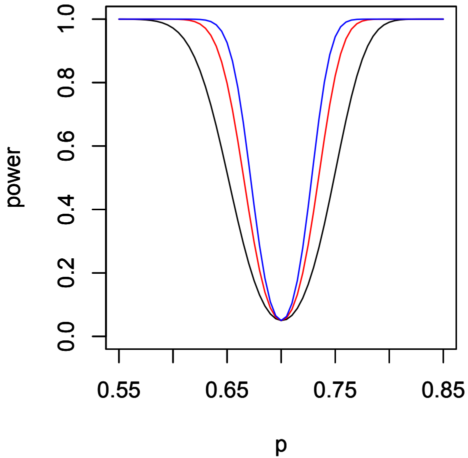
$$-z_{\alpha/2} < \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < z_{\alpha/2}$$

if and only if

$$\frac{p_0 - p}{\sqrt{p(1-p)/n}} - z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{p(1-p)}} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < \frac{p_0 - p}{\sqrt{p(1-p)/n}} + z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{p(1-p)}}.$$

Sample Proportion





Power curve. (left) $n = 336$, $\alpha = 0.05, 0.02, 0.01$. (right) $\alpha = 0.05$, $n = 336, 672, 1008$.

The *p*-value

- The report of *reject* the null hypothesis does not describe the strength of the evidence because it fails to give us the sense of whether or not a small change in the values in the data could have resulted in a different decision.
- Consequently, one common method is to report the value of the test statistic and to give all the values for α that would lead to the rejection of H_0.
- The *p*-value is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. In this way, we provide an assessment of the strength of evidence against H_0.
- Consequently, a very low *p*-value indicates strong evidence against the null hypothesis.

The p-value

We can see how this works with the example on winter survival of beehives using the R command `prop.test`.

```
> prop.test(250,336,p=0.7)

1-sample proportions test with continuity correction

data:  250 out of 336, null probability 0.7
X-squared = 2.8981, df = 1, p-value = 0.08868
alternative hypothesis: true p is not equal to 0.7
95 percent confidence interval:
 0.6932518 0.7891515
sample estimates:
        p
0.7440476
```

The p-value states that we could reject H_0 for any significance level above this value.

The *p*-value

- Under the null hypothesis, $\hat{p}$ has approximately normal, mean $p_0 = 0.7$.
- The *p*-value, 0.089, is the area under the density curve outside the test statistic values

$$|z| = \left| \frac{\hat{p} - p_0}{p_0(1 - p_0)/n} \right| = 1.702$$

(indicated in red),

- The critical value, 1.96, for an $\alpha = 0.05$ level test. (indicated in blue).
- Because the *p*-value is greater than the significance level, we cannot reject H_0 at the 5% level.