Part IV

Hypothesis Testing

Topic 17

Simple Hypotheses

> *I can point to the particular moment when I understood how to formulate the undogmatic problem of the most powerful test of a simple statistical hypothesis against a fixed simple alternative. At the present time, the problem appears entirely trivial and within reach of a beginning undergraduate. But, with a degree of embarrassment, I must confess that it took something like half a decade of combined effort of E.S.P. and myself to put things straight.* - Jerzy Neymann in the Festschrift in honor of Herman Wold, 1970, E.S.P is Egon Sharpe Pearson

17.1 Overview and Terminology

Statistical hypothesis testing is designed to address the question: *Do the data provide sufficient evidence to conclude that we must depart from our original assumption concerning the state of nature?*

The logic of hypothesis testing is similar to the one a juror faces in a criminal trial: *Is the evidence provided by the prosecutor sufficient for the jury to depart from its original assumption that the defendant is not guilty of the charges brought before the court?*

Two of the jury's possible actions are

- **Find the defendant guilty**.
- **Find the defendant not guilty**.

The weight of evidence that is necessary to find the defendant guilty depends on the type of trial. In a criminal trial the stated standard is that the prosecution must prove that *the defendant is guilty beyond any reasonable doubt*. In civil trials, the burden of proof may be the intermediate level of *clear and convincing evidence* or the lower level of the *preponderance of evidence*.

Given the level of evidence needed, a prosecutors task is to present the evidence in the most powerful and convincing manner possible. We shall see these notions reflected in the nature of hypothesis testing.

The simplest set-up for understanding the issues of **statistical hypothesis**, is the case of two values θ_0, and θ_1 in the parameter space. We write the test, known as a **simple hypothesis** as

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

H_0 is called the **null hypothesis**. H_1 is called the **alternative hypothesis**.

We now frame the issue of hypothesis testing using the classical approach. In this approach, the possible actions are:

- **Reject the hypothesis**. Rejecting the hypothesis when it is true is called a **type I error** or a **false positive**. Its probability α is called the **size of the test** or the **significance level**. Sometimes, $1-\alpha$, the **true negatiive** is called the **specificity**. In symbols, we write

$$\alpha = P_{\theta_0}\{\text{reject } H_0\}.$$

- **Fail to reject the hypothesis**. Failing to reject the hypothesis when it is false is called a **type II error** or a **false negative**, has probability β. The **power of the test**, $1-\beta$, the probability of rejecting the test when it is indeed false, is also called the **true positive fraction** or the the **sensitivity**. In symbols, we write

$$\beta = P_{\theta_1}\{\text{fail to reject } H_0\} \quad \text{and} \quad 1-\beta = P_{\theta_1}\{\text{reject } H_0\}.$$

hypothesis tests			criminal trials		
				the defendant is	
	H_0 is true	H_1 is true		innocent	guilty
reject H_0	type I error	OK	convict		OK
fail to reject H_0	OK	type II error	do not convict	OK	

Thus, the *higher* level necessary to secure conviction in a criminal trial corresponds to having *lower* significance levels. This analogy should not be taken too far. The nature of the data and the decision making process is quite dissimilar. For example, the prosecutor and the defense attorney are not always out to find the most honest manner to present information. In statistical inference for hypothesis testing, the goal is something that all participants in this endeavor ought to share.

In addition, care should be taken not to be overly invested in a fixed value α for the significance level. As we continue to investigate the logic and methodology behind hypothesis testing, we will broaden and make more sophisticated our approach to evaluating hypotheses.

The decision for the test is often based on first determining a **critical region** C. Data $\mathbf{x}$ in this region is determined to be too unlikely to have occurred when the null hypothesis is true. Thus, the decision is

$$\text{reject } H_0 \quad \text{if and only if} \quad \mathbf{x} \in C.$$

Given a choice α for the size of the test, the choice of a critical region C is called **best** or **most powerful** if for any other choice of critical region C^* for a size α test, i.e., both critical region lead to the same type I error probability,

$$\alpha = P_{\theta_0}\{X \in C\} = P_{\theta_0}\{X \in C^*\},$$

but perhaps different type II error probabiities

$$\beta = P_{\theta_1}\{X \notin C\}, \quad \beta^* = P_{\theta_1}\{X \notin C^*\},$$

we have the lowest probability of a type II error, ($\beta \le \beta^*$) associated to the critical region C.

The two approaches to hypothesis testing, classical and Bayesian, begin with distinct starting points and end with different interpretations for implications of the data. Interestingly, both approaches result in a decision that is based on the values of a likelihood ratio. In the classical approach, we shall learn, based on the Neyman-Pearson lemma, that the decision is based on a level for this ratio based on setting the type I error probabilities. In the Bayesian approach, the decision on minimizing risk, a concept that we will soon define precisely.

17.2 The Neyman-Pearson Lemma

Many critical regions are either determined by the consequences of the **Neyman-Pearson lemma** or by using analogies of this fundamental lemma. Rather than presenting a proof of this lemma, we will provide some intuition for the choice of critical region through the following "game".

We will conduct a single observation X that can take values from -11 to 11 and based on that observation, decide whether or not to reject the null hypothesis. Basing a decision on a single observation, of course, is not the usual circumstance for hypothesis testing. We will first continue on this line of reasoning to articulate the logic behind the Neyman-Pearson lemma before examining more typical and reasonable data collection protocols.

To begin the game, corresponding to values for x running from -11 to 11, write a row of the number from 0 up to 10 and back down to 0 and add an additional 0 at each end. These numbers add to give 100. Now, scramble the numbers and write them under the first row. This can be created and displayed quickly in R using the commands:

```
> x<- -11:11
> L1<-c(0,0:10,9:0,0)
> L0<-sample(L0) #This provides a random perturbation of the values in L1.
> data.frame(x,L1,L0)
```

The top row, giving the values of L_1, represents the likelihood for one observation under the alternative hypothesis. The bottom row, giving the values of L_0, represents the likelihood under the null hypothesis. Note that the values for L_0 is a rearrangement of the values for L_1. Here is the output.

x	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
$L_1(x)$	0	0	1	2	3	4	5	6	7	8	9	10	9	8	7	6	5	4	3	2	1	0	0
$L_0(x)$	3	8	7	5	7	1	3	10	6	0	6	4	2	5	0	1	0	4	0	8	2	9	9

The goal is to pick values x so that the accumulated points (*the benefit*) increase as quickly as possible from the likelihood L_1 keeping points (*the cost*) from L_0 as low as possible. The natural start is to pick values of x so that $L_0(x) = 0$. Then, the benefit begins to add up without any cost. We find four such values for x and record their values along with running totals for L_1 and L_0.

x	-2	3	5	7	
L_1 total	8	15	20	23	
L_0 total	0	0	0	0	

Being ahead by a score of 23 to 0 can be translated into a best critical region C in the following way. If we take $C = \{-2, 3, 5, 7\}$, then, because the L_1-total is 23 points out of a possible 100, we find the power of the test

$$1 - \beta = P_1\{X \in C\} = 0.23$$

and the type II error $\beta = P_1\{X \notin C\} = 0.77$. Because the L_0-total is 0 points, the size of the test,

$$\alpha = P_0\{X \in C\} = 0$$

and there is *no* chance of type I error with this critical region.

Understanding the next choice is crucial. Candidates are

$$x = 4, \text{with } L_1(4) = 6 \text{ against } L_0(4) = 1 \quad \text{and} \quad x = 1, \text{with } L_1(1) = 9 \text{ against } L_0(1) = 2.$$

The choice 6 against 1 is better than 9 against 2. One way to see this is to note that choosing 6 against 1 twice will put us in a better place than the single choice of 9 against 2. Indeed, after choosing 6 against 1, a choice of 3 against 1 puts us in at least as good a position than the single choice of 9 against 2. The central point is that the best choice comes to picking the remaining value for x that has the *highest benefit-to-cost ratio* of $L_1(x)$ to $L_0(x)$

Now we can pick the next few candidates, keeping track of both the type I and type II error of the test with the choice of critical region being the chosen values of x.

x	-2	3	5	7	4	1	-6	0	-5	
$L_1(x)/L_0(x)$	∞	∞	∞	∞	6	9/2	4	5/2	5/3	
L_1 total	8	15	20	23	29	38	42	52	57	
L_0 total	0	0	0	0	1	3	4	8	11	
β	0.92	0.85	0.80	**0.77**	0.71	0.62	0.58	0.48	0.43	
α	0.00	0.00	0.00	**0.00**	0.01	0.03	0.04	0.08	0.11	

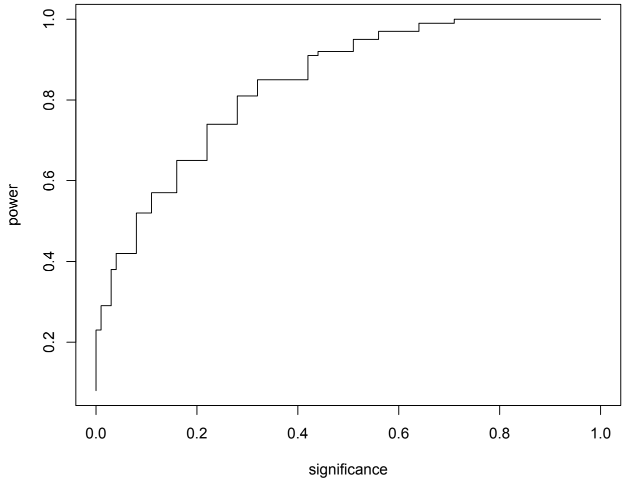

Figure 17.1: Receiver Operating Characteristic. The graph of $\alpha = P\{X \in C|H_0 \text{ is true}\}$ (significance) versus $1-\beta = P\{X \in C|H_1 \text{ is true}\}$ (power) in the example. The horizontal axis α is also called the **false positive fraction (FPF)**. The vertical axis $1-\beta$ is also called the **true positive fraction (TPF)**.

From this exercise we see how the likelihood ratio test is the choice for a most powerful test. For example, for these likelihoods, the last column states that for a $\alpha = 0.11$ level test, the best region consists of those values of x so that

$$\frac{L_1(x)}{L_0(x)} \geq \frac{5}{3}.$$

The type II error probability is $\beta = 0.43$ and thus the power is $1-\beta = 0.57$. In genuine examples, we will typically look for type II error probability much below 0.43 and we will make many observations. We now summarize carefully the insights from this game before examining more genuine examples. A proof of this theorem is provided in Section 17.4.

Theorem 17.1 (Neyman-Pearson Lemma). *Let $L(\theta|\mathbf{x})$ denote the likelihood function for the random variable X corresponding to the probability P_θ. If there exists a critical region C of size α and a nonnegative constant k_α such that*

$$\frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} \geq k_\alpha \quad \textit{for}\, \mathbf{x} \in C$$

and

$$\frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} < k_\alpha \quad \textit{for}\, \mathbf{x} \notin C, \tag{17.1}$$

then C is the most powerful critical region of size α.

We, thus, reject the null hypothesis if and only if the likelihood ratio exceeds a value k_α with

$$\alpha = P_{\theta_0}\left\{\frac{L(\theta_1|X)}{L(\theta_0|X)} \geq k_\alpha\right\}.$$

We shall learn that many of the standard tests use critical values for the t-statistic, the chi-square statistic, or the F-statistic. These critical values are related to the critical value k_α in extensions of the ideas of likelihood ratios. In a few pages, we will take a glance at the Bayesian approach to hypothesis testing.

17.2.1 The Receiver Operating Characteristic

Using R, we can complete the table for L_0 total and L_1 total.

```
> o<-order(L1/L0,decreasing=TRUE)
> sumL1<-cumsum(L1[o])
> sumL0<-cumsum(L0[o])
> significance<-sumL0/100
> power<-sumL1/100
> plot(significance,power,type="s")
> data.frame(x[o],L1[o],L0[o],sumL1,sumL0,power,significance)
```

Completing the curve, known as the **receiver operating characteristic (ROC)**, is shown in the figure above. The ROC shows the inevitable trade-offs between Type I and Type II errors. For example, by the mere fact that the graph is increasing, we can see that by setting a more rigorous test achieved by lowering α, the level of significance, (decreasing the value on the horizontal axis) necessarily reduces $1-\beta$, the power (decreasing the value on the vertical axis.). The unusual and slightly mystifying name is due to the fact that the ROC was first developed during World War II for detecting enemy objects in battlefields, Following the surprise military attack on Pearl Harbor in 1941, the United States saw the need to improve the prediction of the movement of aircraft from their radar signals.

Exercise 17.2. *Consider the following (ignorant) example. Flip a coin that gives heads with probability α. Ignore whatever data you have collected and reject if the coin turns up heads. This test has significance level α. Show that the receiver operating characteristic curve is the line through the origin having slope 1.*

This shows what a minimum acceptable ROC curve looks like - any hypothesis test ought be better than a coin toss that ignores the data. The ROC can be used as a test diagnostic. One commonly used is the area under the ROC, (AUC). For the example above, the AUC is 1/2. So any test should be improve on that value. The "nearly perfect test" would have have the power near to 1 for even very low significance level. In this case the AUC is very nearly equal to 1.

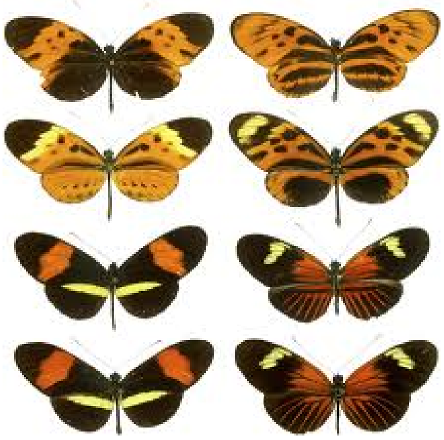

Figure 17.2: Heliconius butterflies

17.3 Examples

Example 17.3. *Mimicry is the similarity of one species to another in a manner that enhances the survivability of one or both species - the* **model** *and* **mimic** *. This similarity can be, for example, in appearance, behavior, sound, or scent. One method for producing a mimic species is* **hybridization**. *This results in the transferring of adaptations from the model species to the mimic. The genetic signature of this has recently been discovered in* Heliconius *butterflies. Padro-Diaz et al sequenced chromosomal regions both linked and unlinked to the red color locus and found a region that displays an almost perfect genotype by phenotype association across four species in the genus* Heliconius

Let's consider a model butterfly species with mean wingspan $\mu_0 = 10$ cm and a mimic species with mean wingspan $\mu_1 = 7$ cm. For both species, the wingspans have standard deviation $\sigma_0 = 3$ cm. Collect 16 specimen to decide if the mimic species has migrated into a given region. If we assume, for the null hypothesis, that the habitat under study is populated by the model species, then

- *a type I error is falsely concluding that the species is the mimic when indeed the model species is resident and*
- *a type II error is falsely concluding that the species is the model when indeed the mimic species has invaded.*

If our action is to begin an eradication program if the mimic has invaded, then a type I error would result in the eradication of the resident model species and a type II error would result in the letting the invasion by the mimic take its course.

To begin, we set a significance level. The choice of an $\alpha = 0.05$ test means that we are accepting a 5% chance of having this error. If the goal is to design a test that has the lowest type II error probability, then the Neyman-Pearson lemma tells us that the critical region is determined by a threshold level k_α for the likelihood ratio.

$$C = \left\{ \mathbf{x}; \frac{L(\mu_1|\mathbf{x})}{L(\mu_0|\mathbf{x})} \geq k_\alpha \right\}.$$

We next move to see how this critical region is determined.

Example 17.4. *Let $X = (X_1, \ldots, X_n)$ be independent normal observations with unknown mean and known variance σ_0^2. The hypothesis is*

$$H_0 : \mu = \mu_0 \quad \textit{versus} \quad H_1 : \mu = \mu_1. \tag{17.2}$$

For the moment consider the case in which $\mu_1 < \mu_0$. We look to determine the critical region.

$$\begin{aligned}
\frac{L(\mu_1|\mathbf{x})}{L(\mu_0|\mathbf{x})} &= \frac{\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp -\frac{(x_1-\mu_1)^2}{2\sigma_0^2} \cdots \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp -\frac{(x_n-\mu_1)^2}{2\sigma_0^2}}{\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp -\frac{(x_1-\mu_0)^2}{2\sigma_0^2} \cdots \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp -\frac{(x_n-\mu_1)^2}{2\sigma_0^2}} \\
&= \frac{\exp -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_1)^2}{\exp -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_0)^2} \\
&= \exp -\frac{1}{2\sigma_0^2} \sum_{i=1}^n \left((x_i - \mu_1)^2 - (x_i - \mu_0)^2 \right) \\
&= \exp -\frac{\mu_0 - \mu_1}{2\sigma_0^2} \sum_{i=1}^n (2x_i - \mu_1 - \mu_0)
\end{aligned}$$

Because the exponential function is increasing, the likelihood ratio test (17.1) is equivalent to

$$\frac{\mu_1 - \mu_0}{2\sigma_0^2} \sum_{i=1}^n (2x_i - \mu_1 - \mu_0), \tag{17.3}$$

exceeding some critical value. Continuing to simplify, this is equivalent to $\bar{x}$ bounded by some critical value,

$$\bar{x} \leq \tilde{k}_\alpha,$$

where $\tilde{k}_\alpha$ is chosen to satisfy

$$P_{\mu_0}\{\bar{X} \leq \tilde{k}_\alpha\} = \alpha.$$

(Note that division by the negative number $\mu_1 - \mu_0$ reverses the direction of the inequality.) Pay particular attention to the fact that the probability is computed under the null hypothesis specifying the mean to be μ_0. In this case, $\bar{X}$ is $N(\mu_0, \sigma_0/\sqrt{n})$ and consequently the standardized version of $\bar{X}$,

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}, \tag{17.4}$$

is a standard normal. Set z_α so that $P\{Z \leq -z_\alpha\} = \alpha$. (This can be determined in R *using the* `qnorm` *command.) Then, by rearranging (17.4), we can determine $\tilde{k}_\alpha$.*

$$\bar{X} \leq \mu_0 - z_\alpha \frac{\sigma_0}{\sqrt{n}} = \tilde{k}_\alpha.$$

.

Equivalently, we can use the standardized score Z as our test statistic and $-z_\alpha$ as the critical value. Note that the only role played by μ_1, the value of the mean under the alternative, is that is less than μ_0. However, it will play a role in determining the power of the test.

Exercise 17.5. *In the example above, give the value of* $\tilde{k}_\alpha$ *explicitly in terms of* $k_\alpha, \mu_0, \mu_1, \sigma_0^2$ *and* n.

Returning to the example of the model and mimic bird species, we now see, by the Neyman-Person lemma that the critical region can be defined as

$$C = \left\{\mathbf{x}; \bar{x} \le \tilde{k}_\alpha\right\} = \left\{\mathbf{x}; \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \le -z_\alpha\right\}.$$

Under the null hypothesis, $\bar{X}$ has a normal distribution with mean $\mu_0 = 10$ and standard deviation $\sigma/\sqrt{n} = 3/4$. This using the distribution function of the normal we can find either $\tilde{k}_\alpha$

```
> qnorm(0.05,10,3/4)
[1] 8.76636
```

or $-z_\alpha$,

```
> qnorm(0.05)
[1] -1.644854
```

Thus, the critical value is $\tilde{k}_\alpha = 8.767$ for the test statistic $\bar{x}$ and $-z_\alpha = -1.645$ for the test statistic z. Now let's look at data.

```
> x
 [1]  8.9  2.4 12.1 10.0  9.2  3.7 13.9  9.1  8.8  6.3 12.1 11.0 12.5  4.5  8.2 10.2
> mean(x)
[1] 8.93125
```

Then

$$\bar{x} = 8.931 \quad z = \frac{8.93124 - 10}{3/\sqrt{16}} = -1.425.$$

$\tilde{k}_\alpha = 8.766 < 8.931$ or $-z_\alpha = -1.645 < -1.425$ and we fail to reject the null hypothesis.

Exercise 17.6. *Modify the calculations in the example above to show that for the case* $\mu_0 < \mu_1$, *using the same value of* z_α *as above, the we reject the null hypothesis precisely when*

$$\bar{X} \ge \mu_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}}. \quad \textit{or} \quad Z \ge z_\alpha$$

Exercise 17.7. *Give an intuitive explanation why the power should*

- increase *as a function of* $|\mu_1 - \mu_0|$,
- decrease *as a function of* σ_0^2, *and*
- increase *as a function of* n.

Next we determine the type II error probability for the situation given by the previous exercise. We will be guided by the fact that

$$\frac{\bar{X} - \mu_1}{\sigma_0/\sqrt{n}}$$

is a *standard normal random variable* for the case that the *alternative hypothesis*, $H_1 : \mu = \mu_1$, is true.

For $\mu_1 > \mu_0$, we find that the type II error probability

$$\begin{aligned}\beta &= P_{\mu_1}\{X \notin C\} = P_{\mu_1}\{\bar{X} < \mu_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}}\} \\ &= P_{\mu_1}\left\{\frac{\bar{X}-\mu_1}{\sigma_0/\sqrt{n}} < z_\alpha - \frac{|\mu_1-\mu_0|}{\sigma_0/\sqrt{n}}\right\} = \Phi\left(z_\alpha - \frac{|\mu_1-\mu_0|}{\sigma_0/\sqrt{n}}\right)\end{aligned}$$

and the power

$$1 - \beta = 1 - \Phi\left(z_\alpha + \frac{|\mu_1-\mu_0|}{\sigma_0/\sqrt{n}}\right) \tag{17.5}$$

Exercise 17.8. *For* **sample size determination** *for the simple hypothesis (17.2) show that n^*, the number of observations to obtain type I error probability α and type II error probability β must satisfy*

$$n^* \geq \frac{\sigma_0^2}{(\mu_1 - \mu_0)^2}(z_\alpha + z_\beta)^2.$$

Notice that n^*

- *decreases* as a function of $|\mu_1 - \mu_0|$,
- *increases* as a function of σ_0^2, and
- *decreases* as a function of α and β. In other words, n^* *increases* as we decrease either type I or type II error.

Exercise 17.9. *Modify the calculations of power in (17.5) above to show that for the case $\mu_1 < \mu_0$ to show that*

$$1 - \beta = \Phi\left(-z_\alpha - \frac{\mu_1 - \mu_0}{\sigma_0/\sqrt{n}}\right). \qquad (17.6)$$

A type II error is falsely failing to conclude that the mimic species have inhabited the study area when indeed they have. To compute the probability of a type II error, note that for $\alpha = 0.05$, we substitute into (17.6),

$$-z_\alpha + \frac{\mu_0 - \mu_1}{\sigma_0/\sqrt{n}} = -1.645 + \frac{3}{3/\sqrt{16}} = 2.355$$

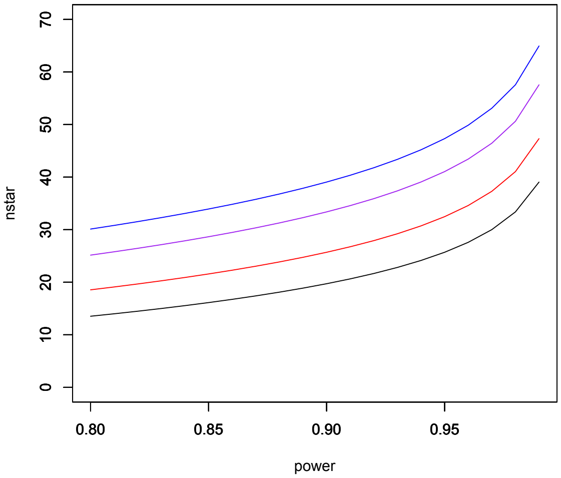

Figure 17.3: **Sample size determination** for the simple hypothesis (17.2) . Minimum sample sample size versus power for significance level $\alpha = 0.10$ (black), 0.05 (red), 0.02 (purple), and 0.01 (blue).

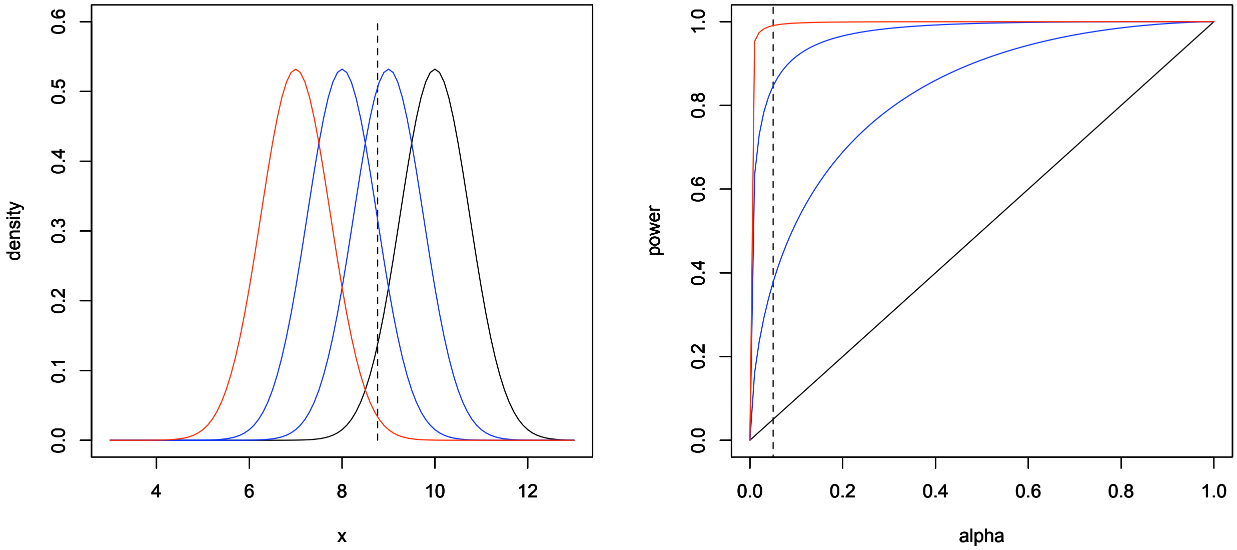

Figure 17.4: **Left**: (black) Density of $\bar{X}$ for normal data under the null hypothesis - $\mu_0 = 10$ and $\sigma_0/\sqrt{n} = 3/\sqrt{16} = 3/4$. With an $\alpha = 0.05$ level test, the critical value $\tilde{k}_\alpha = \mu_0 - z_\alpha\sigma_0/\sqrt{n} = 8.766$. Thus, the area to the left of the vertical dashed line and below the black density function is the significance level $\alpha = P_{\mu_0}\{\bar{X} \leq k_\alpha\}$. The alternatives shown are $\mu_1 = 9$ and 8 (in blue) and $\mu_1 = 7$ (in red). The areas below these curves and to the left of the dashed line is the power $1 - \beta = P_{\mu_1}\{\bar{X} \leq k_\alpha\}$. These values are 0.3777, 0.8466, and 0.9907 for respective alternatives $\mu_1 = 9, 8$ and 7. **Right**: The corresponding receiver operating characteristics curves of the power $1 - \beta$ versus the significance α using equation (17.6). The power for an $\alpha = 0.05$ test are indicated by the intersection of vertical dashed line and the receiver operating characteristics curves.

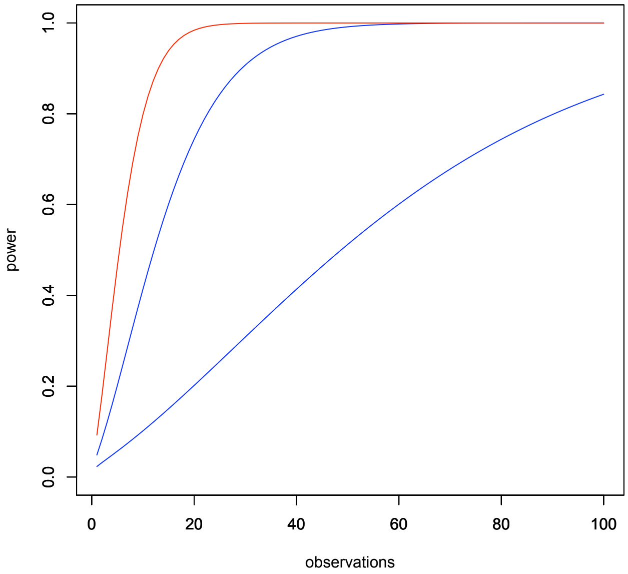

Figure 17.5: Power as a function of the number of observations for an $\alpha = 0.01$ level test. The null hypothesis - $\mu_0 = 10$. The alternatives shown are $\mu_1 = 9$ and 8 (in blue) and $\mu_1 = 7$ (in red). Here $\sigma_0 = 3$. The low level for α is chosen to reflect the desire to have a stringent criterion for rejecting the null hypothesis that the resident species is the model species.

```
>  pnorm(2.355)
[1] 0.9907386
```

and the type II error probability is $\beta = 1 - 0.9907 = 0.0093$, a bit under 1%.

Let's expand the examination of equation (17.6). As we move the alternative value μ_1 downward, the density of $\bar{X}$ moves leftward. The values for $\mu_1 = 9, 8,$ and 7 are displayed on the left in Figure 17.4. This shift in the values is a way of saying that the alternative is becoming more and more distinct as μ_1 decreases. The mimic species becomes easier and easier to detect. We express this by showing that the test is more and more powerful with decreasing values of μ_1. This is displayed by the increasing area under the density curve to the left of the dashed line from 0.377 for the alternative $\mu_1 = 9$ to 0.9907 for $\mu_1 = 7$. We can also see this relationship in the receiver operating characteristic graphed, the graph of the power $1 - \beta$ versus the significance α. This is displayed for the significance level $\alpha = 0.05$ by the dashed line.

Exercise 17.10. *Determine the power of the test for $\mu_0 = 10$ cm and $\mu_1 = 9, 8,$ and 7 cm with the significance level $\alpha = 0.01$. Does the power increase or decrease from its value when $\alpha = 0.01$? Explain your answer. How would the graphs in Figure 17.4 be altered to show this case?*

Often, we wish to know in advance the number of observations n needed to obtain a given power. In this case, we use (17.5) with a fixed value of α, the size of the test, and determine the power of the test as a function of n. We display this in Figure 17.5 with the value of $\alpha = 0.01$. Notice how the number of observations needed to achieve a desired power is high when the wingspan of the mimic species is close to that of the model species.

The example above is called the z-test. If n is sufficiently large, then even if the data are not normally distributed, $\bar{X}$ is well approximated by a normal distribution and, as long as the variance σ_0^2 is known, the z-test is used in this case. In addition, the z-test can be used when $g(\bar{X}_1, \ldots, \bar{X}_n)$ can be approximated by a normal distribution using the delta method.

Example 17.11 (Bernoulli trials)**.** *Here $X = (X_1, \ldots, X_n)$ is a sequence of Bernoulli trials with unknown success* *probability p, the likelihood*

$$\begin{aligned}L(p|\mathbf{x}) &= p^{x_1}(1-p)^{1-x_1}\cdots p^{x_n}(1-p)^{1-x_n} = p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)} \\ &= (1-p)^n\left(\frac{p}{1-p}\right)^{x_1+\cdots+x_n}\end{aligned}$$

For the test

$$H_0 : p = p_0 \quad \textit{versus} \quad H_1 : p = p_1$$

the likelihood ratio

$$\frac{L(p_1|\mathbf{x})}{L(p_0|\mathbf{x})} = \left(\frac{1-p_1}{1-p_0}\right)^n \left(\left(\frac{p_1}{1-p_1}\right)\Big/\left(\frac{p_0}{1-p_0}\right)\right)^{x_1+\cdots+x_n}. \qquad (17.7)$$

Exercise 17.12. *Show that the likelihood ratio (17.7) results in a test to reject H_0 whenever*

$$\sum_{i=1}^n x_i \ge \tilde{k}_\alpha \text{ when } p_0 < p_1 \quad \text{or} \quad \sum_{i=1}^n x_i \le \tilde{k}_\alpha \text{ when } p_0 > p_1. \qquad (17.8)$$

In words, if the alternative is a higher proportion than the null hypothesis, we reject H_0 when the data have too many successes. If the alternative is lower than the null, we eject H_0 when the data do not have enough successes .

In either situation, the number of successes $N = \sum_{i=1}^n X_i$ has a $Bin(n, p_0)$ distribution under the null hypothesis. Thus, in the case $p_0 < p_1$, we choose $\tilde{k}_\alpha$ so that

$$P_{p_0}\left\{\sum_{i=1}^n X_i \ge \tilde{k}_\alpha\right\} \le \alpha. \qquad (17.9)$$

In general, we cannot choose k_α to obtain exactly the value α. Thus, we take the minimum value of k_α to achieve the inequality in (17.9).

To give a concrete example take $p_0 = 0.6$ and $n = 20$ and look at a part of the cumulative distribution function.

x	$\cdots$	13	14	15	16	17	18	19	20
$F_N(x) = P\{N \le x\}$	$\cdots$	0.7500	0.8744	0.9491	0.9840	0.9964	0.9994	0.99996	1

If we take $\alpha = 0.05$, then

$$\begin{aligned}P\{N \ge 16\} &= 1 - P\{N \le 15\} = 1 - 0.9491 = 0.0509 > 0.05 \\ P\{N \ge 17\} &= 1 - P\{N \le 16\} = 1 - 0.9840 = 0.0160 < 0.05\end{aligned}$$

Consequently, we need to have at least 17 successes in order to reject H_0.

Exercise 17.13. *Find the critical region in the example above for $\alpha = 0.10$ and $\alpha = 0.01$. For what values of α is $C = \{16, 17, 18, 19, 20\}$ a critical region for the likelihood ratio test.*

Example 17.14. *If np_0 and $n(1-p_0)$ are sufficiently large, then, by the central limit theorem, $\sum_{i=1}^n X_i$ has approximately a normal distribution. If we write the sample proportion*

$$\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i,$$

then, under the null hypothesis, we can apply the central limit theorem to see that

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

is approximately a standard normal random variable and we perform the z-test as in the previous exercise.

For example, if we take $p_0 = 1/2$ and $p_1 = 3/5$ and $\alpha = 0.05$, then with 110 heads in 200 coin tosses

$$Z = \frac{0.55 - 0.50}{0.05/\sqrt{2}} = \sqrt{2}.$$

```
> qnorm(0.95)
[1] 1.644854
```

Thus, $\sqrt{2} < 1.645 = z_{0.05}$ and we fail to reject the null hypothesis.

Example 17.15. *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 $p_0 = 0.7$. We are checking to see if, for a particularly mild winter, this probability moved up to $p_1 = 0.8$. This leads us to consider the hypotheses*

$$H_0 : p = p_0 \quad \textit{versus} \quad H_1 : p = p_1.$$

for a test of the probability that a feral bee hive survives a winter. 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. For an α level test, the critical value is z_α where α is the probability that a standard normal is at least z_α. If the significance level is $\alpha = 0.05$, then we will reject H_0 for any value of $z > z_\alpha = 1.645$

For this study, 112 colonies have been chosen and 88 survive. Thus $\hat{p} = 0.7875$ and

$$z = \frac{0.7875 - 0.7}{\sqrt{0.7(1-0.7)/112}} = 1.979.$$

Consequently, reject H_0.

For both of these previous examples, the usual method is to compute the z-score with the continuity correction. We shall soon see this with the use of `prop.test` in R.

17.4 Summary

For a simple hypothesis

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

we have two possible action, **reject** H_0 and **fail to reject** H_0, this leads to two possible types of errors

error	probability	alternative names
type I	$\alpha = P_{\theta_0}\{\text{reject } H_0\}$	level significance false positive
type II	$\beta = P_{\theta_1}\{\text{fail to reject } H_0\}$	false negative

The probability $1 - \beta = P_{\theta_1}\{\text{reject } H_0\}$ is called the true positive probability or **power** or **sensitivity**. The probability $1 - \alpha = P_{\theta_0}\{\text{fail to reject } H_0\}$ is called the **specificity**.

The procedure is to set a **significance level** α and find a critical region C so that the type II error probability is as small as possible. The Neyman-Pearson lemma lets us know that in many cases the critical region is determined by setting a level k_α for the likelihood ratio.

$$C = \left\{ \mathbf{x}; \frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} \geq k_\alpha \right\}$$

We continue, showing the procedure in the examples above.

	normal observations $\mu_1 \geq \mu_0$	Bernoulli trials $p_1 > p_0$
Simplify likelihood ratio to obtain a test statistic $T(\mathbf{x})$	$\bar{x}$ $z = \frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}$	$\sum_{i=1}^n x_i$
Use the distribution of $T(\mathbf{x})$ under H_0 to set a critical value $\tilde{k}_\alpha$ so that $P_{\theta_0}\{T(X) \geq \tilde{k}_\alpha\} = \alpha$	$\bar{X} \sim N(\mu_0, \sigma_0/\sqrt{n})$ $Z \sim N(0,1)$	$\sum_{i=1}^n X_i \sim Bin(n, p_0)$
Determine type II error probability $\beta = P_{\theta_1}\{T(X) \geq \tilde{k}_\alpha\}$	$P_{\mu_1}\{\bar{X} \geq \tilde{k}_\alpha\}$	$P_{p_1}\{\sum_{i=1}^n X_i \geq \tilde{k}_\alpha\}$

17.5 Proof of the Neyman-Pearson Lemma

For completeness in exposition, we include a proof of the Neyman-Pearson lemma.

Let C be the α critical region determined by the likelihood ratio test. In addition, let C^* be a critical region for a second test of size α. In symbols,

$$P_{\theta_0}\{X \in C^*\} = P_{\theta_0}\{X \in C\} = \alpha \tag{17.10}$$

As before, we use the symbols β and β^* denote, respectively, the probability of type II error for the critical regions C and C^* respectively. The Neyman-Pearson lemma is the statement that $\beta^* \geq \beta$.

Divide both critical regions C and C^* into two disjoint subsets, the subset that the critical regions share $S = C \cap C^*$ and the subsets $E = C \backslash C^*$ and $E^* = C^* \backslash C$ that are exclusive to one region. In symbols, we write this as the disjoint unions

$$C = S \cup E, \quad \text{and} \quad C^* = S \cup E^*.$$

Thus under either parameter value $\theta_i, i = 1, 2$,

$$P_{\theta_i}\{X \in C\} = P_{\theta_i}\{X \in S\} + P_{\theta_i}\{X \in E\} \quad \text{and} \quad P_{\theta_i}\{X \in C^*\} = P_{\theta_i}\{X \in S\} + P_{\theta_i}\{X \in E^*\}.$$

(See Figure 17.5)

First, we will describe the proof in words.

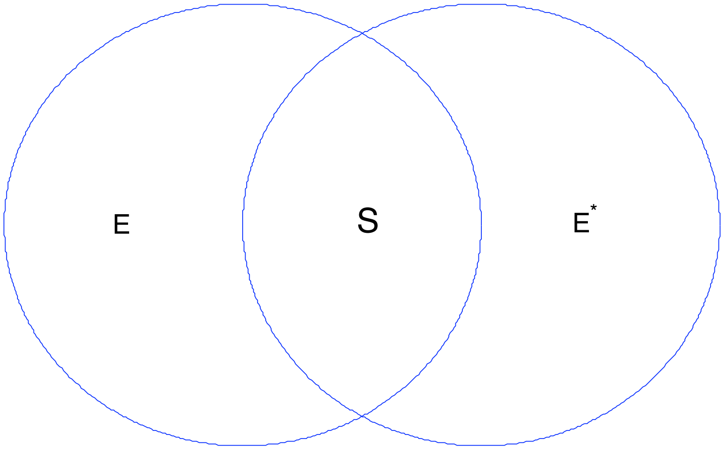

Figure 17.6: Critical region C as determined by the Neyman-Pearson lemma is indicated by the circle on the left. The circle on the right C^* is the critical region is for a second α level test. Thus, $C = S \cup E$ and $C^* = S \cup E^*$.

- The contribution to type I errors from data in S and for type II errors from data outside $E \cup E^*$ are the same for both tests. Consequently, we can focus on differences in types of error by examining the case in which the data land in either E and E^*.
- Because both test have level α, the probability that the data land in E or in E^* are the same under the null hypothesis.
- Under the likelihood ratio critical region, the null hypothesis is not rejected in E^*.
- Under the second test, the null hypothesis is not rejected in E.
- E^* is outside likelihood ratio critical region. So, under the alternative hypothesis, the probability that the data land in E^* is *at most* k_α times as large as it is under the null hypothesis. This contributes to the type II error for the likelihood ratio based test.
- E is in the likelihood ratio critical region. So, under the alternative hypothesis, the probability that the data land in E is *at least* k_α times as large as it is under the null hypothesis. This contributes a larger amount to the type II error for the second test than is added from E^* to the likelihood ratio based test.
- Thus, the type II error for the likelihood ratio based test is smaller than the type II error for the second test.

To carry out the proof, first consider the parameter value θ_0 and subtract from both sides in (17.10) the probability $P_{\theta_0}\{X \in S\}$ that the data land in the shared critical regions and thus would be rejected by both tests to obtain

$$P_{\theta_0}\{X \in E^*\} \geq P_{\theta_0}\{X \in E\}$$

or

$$P_{\theta_0}\{X \in E^*\} - P_{\theta_0}\{X \in E\} \geq 0. \tag{17.11}$$

Moving to the parameter value θ_1, the difference in the corresponding type II error probabilities is

$$\begin{aligned}\beta^* - \beta &= P_{\theta_1}\{X \notin C^*\} - P_{\theta_1}\{X \notin C\} \\ &= (1 - P_{\theta_1}\{X \in C^*\}) - (1 - P_{\theta_1}\{X \in C\}) = P_{\theta_1}\{X \in C\} - P_{\theta_1}\{X \in C^*\}.\end{aligned}$$

Now subtract from both of the integrals the quantity $P_{\theta_1}\{X \in S\}$, the probability that the hypothesis would be falsely rejected by both tests to obtain

$$\beta^* - \beta = P_{\theta_1}\{X \in E\} - P_{\theta_1}\{X \in E^*\} \tag{17.12}$$

We can use the likelihood ratio criterion on each of the two integrals above.

- For $\mathbf{x} \in E$, then $\mathbf{x}$ is in the critical region and consequently $L(\theta_1|\mathbf{x}) \geq k_\alpha L(\theta_0|\mathbf{x})$ and

$$P_{\theta_1}\{X \in E\} = \int_E L(\theta_1|\mathbf{x})\,d\mathbf{x} \geq k_\alpha \int_E L(\theta_0|\mathbf{x})\,d\mathbf{x} = k_\alpha P_{\theta_0}\{X \in E\}.$$

- For $\mathbf{x} \in E^*$, then $\mathbf{x}$ is not in the critical region and consequently $L(\theta_1|\mathbf{x}) \leq k_\alpha L(\theta_0|\mathbf{x})$ and

$$P_{\theta_1}\{X \in E^*\} = \int_{E*} L(\theta_1|\mathbf{x})\,d\mathbf{x} \leq k_\alpha \int_{E^*} L(\theta_0|\mathbf{x})\,d\mathbf{x} = k_\alpha P_{\theta_0}\{X \in E^*\}.$$

Apply these two inequalities to (17.12)

$$\beta^* - \beta \geq k_\alpha(P_{\theta_0}\{X \in E^*\} - P_{\theta_0}\{X \in E\}).$$

This difference is at least 0 by (17.11) and consequently $\beta^* \geq \beta$, i. e., the critical region C^* has at least as large type II error probability as that given by the likelihood ratio test.

NB. The integral will be placed by sums in the case of discrete random variables. For those who know some measure theory, we can maintain the inequalities above if the integral is taken with respect to some reference measure μ.

17.6 An Brief Introduction to the Bayesian Approach

As with other aspects of the Bayesian approach to statistics, hypothesis testing is closely aligned with Bayes theorem. For a simple hypothesis, we begin with a **prior probability** for each of the competing hypotheses.

$$\pi\{\theta_0\} = P\{H_0 \text{ is true}\} \quad \text{and} \quad \pi\{\theta_1\} = P\{H_1 \text{ is true}\}.$$

Naturally, $\pi\{\theta_0\} + \pi\{\theta_1\} = 1$. Although this is easy to state, the choice of a prior ought to be grounded in solid scientific reasoning.

As before, we collect data and with it compute the **posterior probabilities** of the two parameter values θ_0 and θ_1. This gives us the posterior probabilities that H_0 is true and H_1 is true.

We can see, in its formulation, the wide difference in perspective between the Bayesian and classical approaches.

- In the Bayesian approach, we begin with a prior probability that H_0 is true. In the classical approach, the assumption is that H_0 is true.
- In the Bayesian approach, we use the data and Bayes formula to compute the posterior probability that H_1 is true. In the classical approach, we use the data and a significance level to make a decision to reject H_0. The question: *What is the probability that H_1 is true?* has *no* meaning in the classical setting.
- The decision to reject H_0 in the Bayesian setting is based on minimizing risk using presumed losses for type I and type II errors. In classical statistics, the choice of type I error probability is used to construct a critical region. This choice is made with a view to making the type II error probability as small as possible. We reject H_0 whenever the data fall in the critical region.

Both approaches use as a basic concept, the likelihood function $L(\theta|\mathbf{x})$ for the data $\mathbf{x}$. Let $\tilde{\Theta}$ be a random variable taking on one of the two values θ_0, θ_1 and having a distribution equal to the prior probability π. Thus,

$$\pi\{\theta_i\} = P\{\tilde{\Theta} = \theta_i\}, \quad i = 0, 1.$$

Recall Bayes formula for events A and C,

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|C^c)P(C^c)}, \tag{17.13}$$

we set C to be the event that the alternative hypothesis is true and A to be the event that the data take on the value $\mathbf{x}$. In symbols,

$$C = \{\tilde{\Theta} = \theta_1\} = \{H_1 \text{ is true}\} \quad \text{and} \quad A = \{X = \mathbf{x}\}.$$

Focus for the moment on the case in which the data are discrete, we have the conditional probabilities for the alternative hypothesis.

$$P(A|C) = P_{\theta_1}\{X = \mathbf{x}\} = f_X(\mathbf{x}|\theta_1) = L(\theta_1|\mathbf{x}).$$

Similarly, for the null hypothesis,

$$P(A|C^c) = P_{\theta_0}\{X = \mathbf{x}\} = f_X(\mathbf{x}|\theta_0) = L(\theta_0|\mathbf{x}).$$

The posterior probability that H_1 is true can be written symbolically in several ways.

$$f_{\tilde{\Theta}|X}(\theta_1|\mathbf{x}) = P\{H_1 \text{ is true}|X = \mathbf{x}\} = P\{\tilde{\Theta} = \theta_1|X = \mathbf{x}\}$$

Returning to Bayes formula, we make the substitutions in (17.13),

$$f_{\tilde{\Theta}|X}(\theta_1|\mathbf{x}) = \frac{L(\theta_1|\mathbf{x})\pi\{\theta_1\}}{L(\theta_0|\mathbf{x})\pi\{\theta_0\} + L(\theta_1|\mathbf{x})\pi\{\theta_1\}}.$$

By making a similar argument involving limits, we can reach the same identity for the density of continuous random variables. The formula for the posterior probability can be more easily understood if we rewrite the expression above in terms of odds, i. e., as the ratio of probabilities.

$$\frac{f_{\tilde{\Theta}|X}(\theta_1|\mathbf{x})}{f_{\tilde{\Theta}|X}(\theta_0|\mathbf{x})} = \frac{P\{H_1 \text{ is true}|X=\mathbf{x}\}}{P\{H_0 \text{ is true}|X=\mathbf{x}\}} = \frac{P\{\tilde{\Theta}=\theta_1|X=\mathbf{x}\}}{P\{\tilde{\Theta}=\theta_0|X=\mathbf{x}\}} = \frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})}\cdot\frac{\pi\{\theta_1\}}{\pi\{\theta_0\}}. \tag{17.14}$$

With this expression we see that the posterior odds are equal to the likelihood ratio times the prior odds. In this case the likelihood ratio is called the **Bayes factor** of H_1 in favor of H_0.

$$B = \frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})}$$

(This is the reciprocal of the ratio used in the Neyman Pearson lemma. In general, pay particular attention to the choice of numerator and denominator in this ratio.)

The decision whether or not to reject H_0 depends on the values assigned for the loss obtained in making an incorrect conclusion. We begin by setting values for the loss. This can be a serious exercise in which a group of experts weighs the evidence for either adverse outcome. We will take a loss of 0 for making a correct decision, a loss of $\ell_{\rm I}$ for a type I error and $\ell_{\rm II}$ for a type II error. We summarize this in a table.

loss function table		
decision	H_0 is true	H_1 is true
H_0	0	$\ell_{\rm II}$
H_1	$\ell_{\rm I}$	0

The Bayes procedure is to make the decision that has the smaller posterior expected loss, also known as the **risk**. If the decision is H_0, the loss $\mathcal{L}_0(\mathbf{x})$ takes on two values

$$\mathcal{L}_0(\mathbf{x}) = \begin{cases} 0 & \text{with probability } P\{H_0 \text{ is true}|X=\mathbf{x}\}, \\ \ell_{\rm II} & \text{with probability } P\{H_1 \text{ is true}|X=\mathbf{x}\}. \end{cases}$$

The expected loss

$$E\mathcal{L}_0(\mathbf{x}) = \ell_{\rm II}P\{H_1 \text{ is true}|X=\mathbf{x}\} = \ell_{\rm II}(1-P\{H_0 \text{ is true}|X=\mathbf{x}\}) \tag{17.15}$$

is simply the product of the loss and the probability of incorrectly choosing H_1.

If the decision is H_1, the loss $\mathcal{L}_1(\mathbf{x})$ also takes on two values

$$\mathcal{L}_1(\mathbf{x}) = \begin{cases} \ell_{\rm I} & \text{with probability } P\{H_0 \text{ is true}|X=\mathbf{x}\}, \\ 0 & \text{with probability } P\{H_1 \text{ is true}|X=\mathbf{x}\}. \end{cases}$$

In this case, the expected loss

$$E\mathcal{L}_1(\mathbf{x}) = \ell_{\rm I}P\{H_0 \text{ is true}|X=\mathbf{x}\} \tag{17.16}$$

is a product of the loss and the probability of incorrectly choosing H_0.

We can now express the Bayesian procedure in symbols using the criterion of smaller posterior expected loss:

$$\text{decide on } H_1 \quad \text{if and only if} \quad E\mathcal{L}_1(\mathbf{x}) \le E\mathcal{L}_0(\mathbf{x}).$$

Now substituting for $E\mathcal{L}_0(\mathbf{x})$ and $E\mathcal{L}_1(\mathbf{x})$ in (17.15) and (17.16), we find that we make the decision on H_1 and reject H_0 if and only if

$$\begin{aligned} \ell_{\rm I}P\{H_0 \text{ is true}|X=\mathbf{x}\} &\le \ell_{\rm II}(1-P\{H_0 \text{ is true}|X=\mathbf{x}\}) \\ (\ell_{\rm I}+\ell_{\rm II})P\{H_0 \text{ is true}|X=\mathbf{x}\} &\le \ell_{\rm II} \\ P\{H_0 \text{ is true}|X=\mathbf{x}\} &\le \frac{\ell_{\rm II}}{\ell_{\rm I}+\ell_{\rm II}} \end{aligned}$$

or stated in terms of odds

$$\frac{P\{H_1 \text{ is true}|X = \mathbf{x}\}}{P\{H_0 \text{ is true}|X = \mathbf{x}\}} \ge \frac{\ell_{\mathrm{I}}}{\ell_{\mathrm{II}}}, \tag{17.17}$$

we reject H_0 whenever the posterior odds exceeds the ratio of the losses for each type of error.

As we saw in (17.14), this ratio of posterior odds is dependent on the ratio of prior odds. Taking this into account, we see that the criterion for rejecting H_0 is a level test for the likelihood ratio:

Reject H_0 if and only if the Bayes factor

$$B = \frac{L(\theta_1|\mathbf{x})}{L(\theta_0|\mathbf{x})} \ge \frac{\ell_{\mathrm{I}}/\pi\{\theta_1\}}{\ell_{\mathrm{II}}/\pi\{\theta_0\}}. \tag{17.18}$$

This is exactly the same type of criterion as that used in classical statistics. However, the rationale, thus the value for the ratio necessary to reject, is quite different. For example, the higher the value of the prior odds, the higher the likelihood ratio needed to reject H_0 under the Bayesian framework.

Example 17.16. *For normal observations with means μ_0 for the null hypothesis and μ_1 for the alternative hypothesis. If the variance has a known value, σ_0, we have from Example 17.4, the likelihood ratio*

$$\frac{L(\mu_1|\mathbf{x})}{L(\mu_0|\mathbf{x})} = \exp\frac{\mu_1 - \mu_0}{2\sigma_0^2}\sum_{i=1}^n (2x_i - \mu_1 - \mu_0) = \exp\left(\frac{\mu_1 - \mu_0}{2\sigma_0^2} n(2\bar{x} - \mu_1 - \mu_0)\right). \tag{17.19}$$

For Example 17.3 on the model and mime butterfly species, $\mu_0 = 10$, $\mu_1 = 7$, $\sigma_0 = 3$, and sample mean $\bar{x} = 8.931$ based on $n = 16$ observations, we find the likelihood ratio 0.1004. Thus,

$$\frac{P\{H_1 \text{ is true}|X = \mathbf{x}\}}{P\{H_0 \text{ is true}|X = \mathbf{x}\}} = \frac{P\{\tilde{M} = \mu_1|X = \mathbf{x}\}}{P\{\tilde{M} = \mu_0|X = \mathbf{x}\}} = 0.1004\frac{\pi\{\mu_1\}}{\pi\{\mu_0\}}.$$

where $\tilde{M}$ is a random variable having a distribution equal to the prior probability π for the model and mimic butterfly wingspan. Consequently, the posterior odds for the mimic vs. mime species is approximately ten times the prior odds.

Finally, the decision will depend on the ratio of $\ell_{\mathrm{II}}/\ell_{\mathrm{I}}$, i. e., the ratio of the loss due to eradication of the resident model species versus letting the invasion by the mimic take its course.

Exercise 17.17. *Substitute the likelihood ratio in 17.19 into 17.18 and solve in terms of $\bar{x}$. Use this to determine threshold values for $barx$ to reject H_0 for prior probabilities $\pi\{\mu_0\} = 0.05, 0.10, 0.20$ and lost ratios $\ell_{\mathrm{I}}/\ell_{\mathrm{II}} = 1/2, 1, 2$. What situations give the lowest and highest threshold values for $\bar{x}$? Explain your answer.*

Exercise 17.18. *Returning to a previous example, give the likelihood ratios for $n = 20$ Bernoulli trials with $p_0 = 0.6$ and $p_1 = 0.7$ for values $x = 0, \ldots, 20$ for the number of successes. Give the values for the number of successes in which the number of successes change the prior odds by a factor of 5 or more as given by the posterior odds.*

17.7 Answers to Selected Exercises

17.2 Flip a biased coin in which the probability of heads is α under both the null and alternative hypotheses and reject whenever heads turns up. Then

$$\alpha = P_{\theta_0}\{\text{heads}\} = P_{\theta_1}\{\text{heads}\} = 1 - \beta.$$

Thus, the receiver operating characteristic curve is the line through the origin having slope 1.

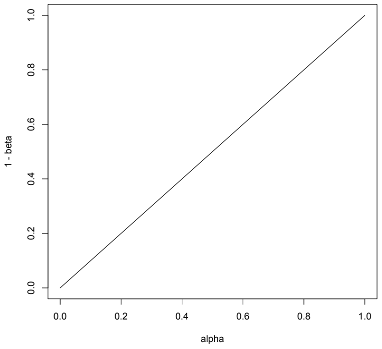

Figure 17.7: **Receiver operating Characteristic** based on a biased coin toss. Thus, any viable ROC should be above the line the the graph.

17.4. The likelihood ratio

$$\frac{L(\mu_1|\mathbf{x})}{L(\mu_0|\mathbf{x})} = \exp -\frac{\mu_0 - \mu_1}{2\sigma_0^2}\sum_{i=1}^n (2x_i - \mu_1 - \mu_0) \ge k_\alpha.$$

Thus,

$$\begin{aligned} \tfrac{\mu_1-\mu_0}{2\sigma_0^2} \textstyle\sum_{i=1}^n(2x_i-\mu_1-\mu_0) &\geq \ln k_\alpha \\ \textstyle\sum_{i=1}^n(2x_i-\mu_1-\mu_0) &\geq \tfrac{2\sigma_0^2}{\mu_1-\mu_0}\ln k_\alpha \\ 2\bar{x}-\mu_1-\mu_0 &\leq \tfrac{2\sigma_0^2}{n(\mu_1-\mu_0)}\ln k_\alpha \\ \bar{x} &\leq \tfrac{1}{2}\left(\tfrac{2\sigma_0^2}{n(\mu_1-\mu_0)}\ln k_\alpha+\mu_1+\mu_0\right)=\tilde{k}_\alpha \end{aligned}$$

Notice that since $\mu_1 < \mu_0$, division by $\mu_1 - \mu_0$ changes the direction of the inequality.

17.6. If c_α is the critical value in expression in (17.3) then

$$\frac{\mu_1-\mu_0}{2\sigma_0^2}\sum_{i=1}^n(2x_i-\mu_1-\mu_0) \geq c_\alpha$$

SInce $\mu_1 > \mu_0$, division by $\mu_1 - \mu_0$ does not change the direction of the inequality. The rest of the argument proceeds as before. we obtain that $\bar{x} \geq \tilde{k}_\alpha$.

17.7. If power means easier to distinguish using the data, then this is true when the means are farther apart, the measurements are less variable or the number of measurements increases. This can be seen explicitly is the power equation (17.5).

17.8. We shall do the case $\mu_1 > \mu_0$. the other case is similar.
From equation (17.5),

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

The goal is to choose n so that the argument argument $z_\alpha + \frac{|\mu_1-\mu_0|}{\sigma_0/\sqrt{n}}$ has probability β. However, we have that $-z_\beta$ has *lower* tail probability β. In other words, $\beta = \Phi(-z_\beta)$. Because Φ, the cumulative distribution function for the standard normal, is one-to-one,

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

Thus, n^*, any integer al least as large as n will have the desired type I and type II errors.

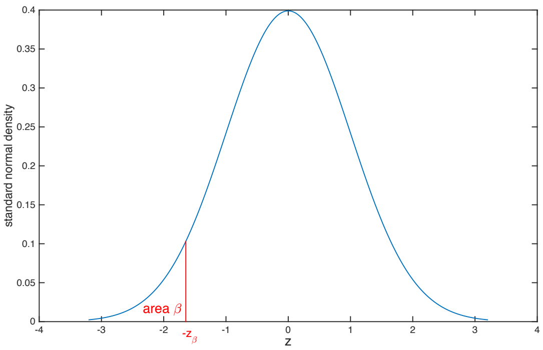

Figure 17.8: Plot of standard normal density function The value $-z_\beta$ has lower tail probabiility β. ($\beta = 0.05$ is shown.)

17.9. For $\mu_0 > \mu_1$,

$$\begin{aligned} \beta &= P_{\mu_1}\{X \notin C\} = P_{\mu_1}\{\bar{X} > \mu_0 - \frac{\sigma_0}{\sqrt{n}}z_\alpha\} \\ &= P_{\mu_1}\left\{\frac{\bar{X}-\mu_1}{\sigma_0/\sqrt{n}} > -z_\alpha - \frac{\mu_1-\mu_0}{\sigma_0/\sqrt{n}}\right\} = 1-\Phi\left(-z_\alpha - \frac{\mu_1-\mu_0}{\sigma_0/\sqrt{n}}\right) \end{aligned}$$

and the power

$$1-\beta = \Phi\left(-z_\alpha - \frac{\mu_1-\mu_0}{\sigma_0/\sqrt{n}}\right).$$

17.10. Interpreting equation (17.5) in R, we find that

```
> mu0<-10;sigma0<-3;n<-16
> zalpha<-qnorm(0.99)
> mu1<-c(9,8,7)
> power<-1-pnorm(zalpha-abs(mu1-mu0)/(sigma0/sqrt(n)))
> data.frame(mu1,power)
  mu1     power
1   9 0.1603514
2   8 0.6331918
3   7 0.9529005
```

Notice that the power has decreased from the case $\alpha = 0.05$. This could be anticipated. In reducing the significance level from $\alpha = 0.05$ to $\alpha = 0.01$, we make the criterion for rejecting more stringent by reducing he critical region C. The effect can be seen in FIgure 17.4. On the left side figure, the vertical dashed line is moved left to reduce the area under the black curve to the left of the dashed line. This, in turn, reduces the area under the other curves to the left of the dashed line. On the right figure, the vertical dashed line is moved left to the value $\alpha = 0.01$ and, because the ROC curve is increasing, the values for the power decreased.

17.12. For the likelihood ratio (17.7), take the logarithm to obtain

$$\ln\left(\frac{L(p_1|\mathbf{x})}{L(p_0|\mathbf{x})}\right) = n\ln\left(\frac{1-p_1}{1-p_0}\right) + (x_1+\cdots+x_n)\ln\left(\left(\frac{p_1}{1-p_1}\right)\Big/\left(\frac{p_0}{1-p_0}\right)\right) \ge \ln k_\alpha.$$

If $p_0 < p_1$ then the ratio in the expression for the logarithm in the second term is greater than 1 and consequently, the logarithm is positive. Thus, we isolate the sum $\sum_{i=1}^n x_i$ to give the test (17.8). For $p_0 > p_1$, the logarithm is negative and the direction of the inequality in (17.8) is reversed.

17.13. If we take $\alpha = 0.10$, then

$$P\{N \ge 15\} = 1 - P\{N \le 14\} = 1 - 0.8744 = 0.1256 > 0.10$$
$$P\{N \ge 16\} = 1 - P\{N \le 15\} = 1 - 0.9491 = 0.0509 < 0.10$$

Consequently, we need to have at least 16 successes in order to reject H_0. If we take $\alpha = 0.01$, then

$$P\{N \ge 17\} = 1 - P\{N \le 16\} = 1 - 0.9840 = 0.0160 > 0.01$$
$$P\{N \ge 18\} = 1 - P\{N \le 17\} = 1 - 0.9964 = 0.0036 < 0.01$$

Consequently, we need to have at least 18 successes in order to reject H_0. For $C = \{16, 17, 18, 19, 20\}$,

$$P\{N \in C\} = 1 - P\{N \le 15\} = 1 - 0.9491 = 0.0509.$$

Thus, α must be less that 0.0509 for C to be a critical region. In addition, $P\{N \ge 17\} = 0.0160$. Consequently, if we take any value for $\alpha < 0.0160$, then the critical region will be smaller than C.

17.17. Making the substitution of 17.19 into 17.18, we have

$$\begin{aligned}
\exp\left(\frac{\mu_0-\mu_1}{2\sigma_0^2}n(2\bar{x}-\mu_1-\mu_0)\right) &\le \frac{\ell_{\mathrm{II}}/\pi\{\theta_0\}}{\ell_{\mathrm{I}}/\pi\{\theta_1\}} \\
\frac{\mu_0-\mu_1}{2\sigma_0^2}n(2\bar{x}-\mu_1-\mu_0) &\le \ln\left(\frac{\ell_{\mathrm{II}}/\pi\{\theta_0\}}{\ell_{\mathrm{I}}/\pi\{\theta_1\}}\right) \\
2\bar{x}-\mu_1-\mu_0 &\le \frac{2\sigma_0^2}{n(\mu_0-\mu_1)}\ln\left(\frac{\ell_{\mathrm{II}}/\pi\{\theta_0\}}{\ell_{\mathrm{I}}/\pi\{\theta_1\}}\right) \\
\bar{x} &\le \frac{1}{2}\left(\frac{2\sigma_0^2}{n(\mu_0-\mu_1)}\ln\left(\frac{\ell_{\mathrm{II}}/\pi\{\theta_0\}}{\ell_{\mathrm{I}}/\pi\{\theta_1\}}\right)+\mu_1+\mu_0\right)
\end{aligned}$$

```
> mu0<-10;mu1<-7;sigma<-3;n<-16
> pi0<-c(0.05,0.10,0.20)
> lr<-1/2
> threshold<-(2*sigma^2/(n*(mu1-mu0))*log(lr*pi0/(1-pi0))+mu1+mu0)/2
> data.frame(pi0,threshold)
   pi0 threshold
1 0.05  9.182047
2 0.10  9.041945
3 0.20  8.889895
> threshold<-(2*sigma^2/(n*(mu1-mu0))*log(lr*pi0/(1-pi0))+mu1+mu0)/2
> data.frame(pi0,threshold)
   pi0 threshold
1 0.05  9.052082
2 0.10  8.911980
3 0.20  8.759930
> lr<-2
> threshold<-(2*sigma^2/(n*(mu1-mu0))*log(lr*pi0/(1-pi0))+mu1+mu0)/2
> data.frame(pi0,threshold)
   pi0 threshold
1 0.05  8.922117
2 0.10  8.782015
3 0.20  8.629965
> lr<-1
```

The lowest threshold value $\bar{x} = 8.62$ is for the case $\pi\{\theta_0\} = 0.20$ and $\ell_{\rm I}/\ell_{|\rm I} = 2$. This is the highest prior probability and the highest relative loss for a type I error. These both require stronger evidence to reject H_0 and thus need a more extreme and thus lower value for $\bar{x}$ to reject H_0.

The highest threshold value $\bar{x} = 9.18$ is for the case $\pi\{\theta_0\} = 0.05$ and $\ell_{\rm I}/\ell_{\rm II} = 1/2$. This is the lowest prior probability and the lowest relative loss for a type I error. These both require less evidence to reject H_0 and thus need a less extreme and thus higher value for $\bar{x}$ to reject H_0.

17.19. Using the reciprocal likelihood ratio formula in Example 17.9, we compute the Bayes factor B for

```
> x<-c(0:20)
> n<-20
> p0<-0.6
> p1<-0.7
> B<-((1-p1)/(1-p0))^n*((p1/(1-p1))/(p0/(1-p0)))^x
> data.frame(x[1:7],B[1:7],x[8:14],B[8:14],x[15:21],B[15:21])
  x.1.7.      B.1.7. x.8.14.    B.8.14. x.15.21.  B.15.21.
1      0 0.003171212       7 0.06989143       14  1.540361
2      1 0.004932996       8 0.10872001       15  2.396118
3      2 0.007673550       9 0.16912001       16  3.727294
4      3 0.011936633      10 0.26307558       17  5.798013
5      4 0.018568096      11 0.40922867       18  9.019132
6      5 0.028883705      12 0.63657794       19 14.029761
7      6 0.044930208      13 0.99023235       20 21.824072
```

Thus, values $x \leq 9$ increase the posterior odds in favor of H_0 by a factor greater than 5 ($B < 1/5$), values $x \geq 17$ increase the posterior odds in favor of H_1 by a factor greater than 5 ($B > 5$).

Topic 18

Composite Hypotheses

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 length, the reaction rate, the fraction displaying a particular behavior or having a particular opinion, the temperature, the kinetic energy, the Michaelis constant, the speed of light, mutation rate, the melting point, the probability that the dominant allele is expressed, the elasticity, the force, the mass, the parameter value θ_0 increase, decrease or change at all under under a different experimental condition?*

18.1 Partitioning the Parameter Space

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**.

For the three alternatives to the question posed above, let θ be one of the components in the parameter space, then

- 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**,

Rejection and failure to reject the null hypothesis, critical regions, C, and type I and type II errors have the same meaning for a composite hypotheses as it does with a simple hypothesis. Significance level and power will necessitate an extension of the ideas for simple hypotheses.

18.2 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

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

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

The ideal power function has

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

With this property for the power function, we would rarely reject the null hypothesis when it is true and rarely fail to reject the null hypothesis when it is false.

In reality, incorrect decisions are made. Thus, for $\theta \in \Theta_0$,

$$\pi(\theta) \text{ is the probability of making a type I error,}$$

i.e., rejecting the null hypothesis when it is indeed true. For $\theta \in \Theta_1$,

$$1 - \pi(\theta) \text{ is the probability of making a type II error,}$$

i.e., failing to reject the null hypothesis when it is false.

The goal is to make the chance for error small. The traditional method is analogous to that employed in the Neyman-Pearson lemma. 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. In other words, 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 $\theta \in \Theta_0$. Then, we look for a critical region that makes the power function as large as possible for values of the parameter $\theta \in \Theta_1$

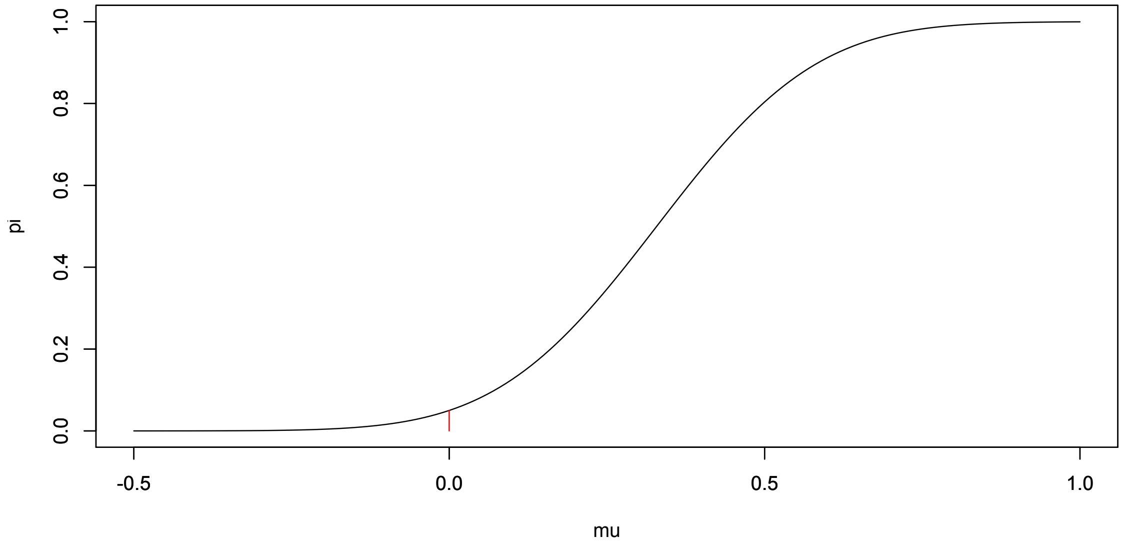

Figure 18.1: Power function for the one-sided test with alternative "greater". The size of the test α is given by the height of the red segment. Notice that $\pi(\mu) < \alpha$ for all $\mu < \mu_0$ and $\pi(\mu) > \alpha$ for all $\mu > \mu_0$

Example 18.1. *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*

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

As we shall see soon, the value of $k(\mu_0)$ depends on the level of the test.

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 the set $\Theta_0 = \{\mu; \mu \le \mu_0\}$ takes place for the value μ_0. Consequently, to obtain level α for the hypothesis test, set

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

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 $P\{Z \ge z_\alpha\} = \alpha$. Thus

$$P_{\mu_0}\{Z \ge z_\alpha\} = P_{\mu_0}\{\bar{X} \ge \mu_0 + \frac{\sigma_0}{\sqrt{n}} z_\alpha\}$$

and $k(\mu_0) = \mu_0 + (\sigma_0/\sqrt{n})z_\alpha$.

If μ is the true state of nature, then

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

is a standard normal random variable. We use this fact to determine the power function for this test.

$$\pi(\mu) = P_\mu\{\bar{X} \ge \frac{\sigma_0}{\sqrt{n}} z_\alpha + \mu_0\} = P_\mu\{\bar{X} - \mu \ge \frac{\sigma_0}{\sqrt{n}} z_\alpha - (\mu - \mu_0)\} \tag{18.1}$$

$$= P_\mu\left\{\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \ge z_\alpha - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right\} = 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right) \tag{18.2}$$

where Φ is the distribution function for a standard normal random variable.

We have seen the expression above 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 receiver operating characteristic as in Figure 17.2.
- If we fix μ_1 the alternative value and the significance level α, then we can determine the power as a function of the number of observations as in Figure 17.3.
- If we fix n and the significance level α, then we can determine the power function $\pi(\mu)$, the power as a function of the alternative value μ. An example of this function is shown in Figure 18.1.

Exercise 18.2. If the alternative is less than, show that

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

Returning to the example with a model species and its mimic. For the plot of the power function for $\mu_0 = 10$, $\sigma_0 = 3$, and $n = 16$ observations,

```
> zalpha<-qnorm(0.95)
> mu0<-10
> sigma0<-3
> mu<-(600:1100)/100
> n<-16
> z<--zalpha - (mu-mu0)/(sigma0/sqrt(n))
> pi<-pnorm(z)
> plot(mu,pi,type="l")
```

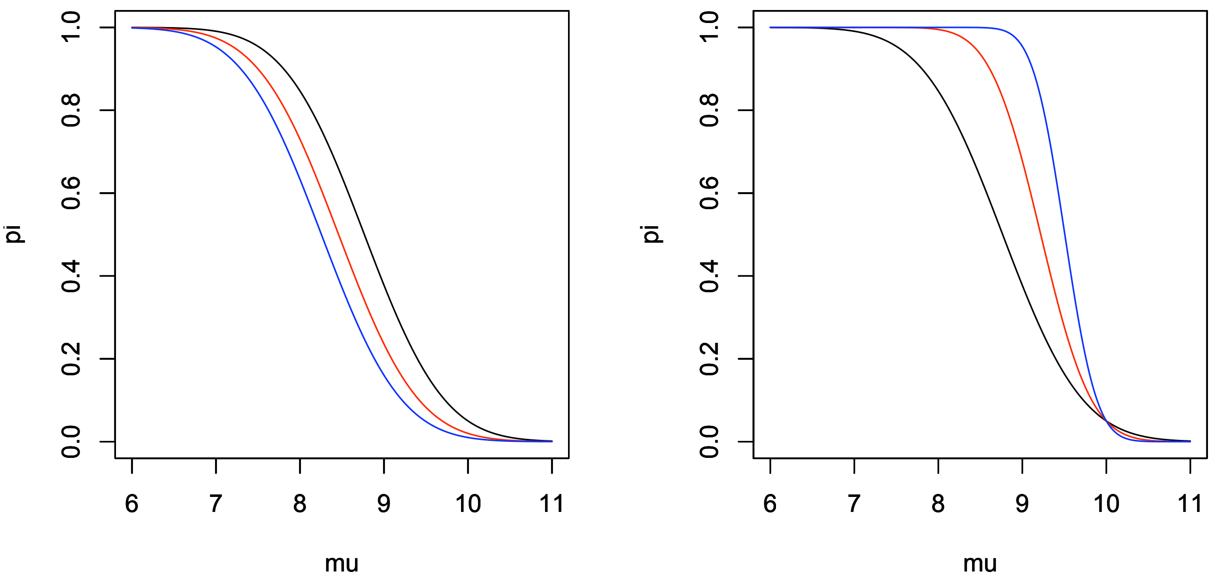

Figure 18.2: Power function for the one-sided test with alternative "less than". $\mu_0 = 10$, $\sigma_0 = 3$. Note, as argued in the text that π is a decreasing function. **(left)** $n = 16$, $\alpha = 0.05$ (black), 0.02 (red), and 0.01 (blue). Notice that lowering significance level α reduces power $\pi(\mu)$ for each value of μ. **(right)** $\alpha = 0.05$, $n = 15$ (black), 40 (red), and 100 (blue). Notice that increasing sample size n increases power $\pi(\mu)$ for each value of $\mu \leq \mu_0$ and decreases type I error probability for each value of $\mu > \mu_0$. For all 6 power curves, we have that $\pi(\mu_0) = \alpha$.

In Figure 18.2, we vary the values of the significance level α and the values of n, the number of observations in the graph of the power function π

Example 18.3 (mark and recapture)**.** *We may want to use mark and recapture as an experimental procedure to test whether or not a population has reached a dangerously low level. The variables in mark and recapture are*

- *t be the number captured and tagged,*
- *k be the number in the second capture,*
- *r the the number in the second capture that are tagged, and let*
- *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.

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

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

$$\pi(N) = P_N\{R \geq r_\alpha\} \leq \alpha \textit{ for all } N \geq N_0. \qquad (18.3)$$

As N increases, then recaptures become less likely and the probability in (18.3) decreases. Thus, we should set the value of r_α according to the parameter value N_0, the minimum value under the null hypothesis. Let's determine r_α

for several values of α using the example from the topic, Maximum Likelihood Estimation, and consider the case in which the critical population is $N_0 = 2000$.

```
> N0<-2000; t<-200; k<-400
> 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     49
2  0.02     51
3  0.01     53
```

For example, we must capture al least 49 that were tagged in order to reject H_0 at the $\alpha = 0.05$ level. In this case the estimate for N is $\hat{N} = [kt/r_\alpha] = 1632$. As anticipated, r_α increases and the critical regions shrinks as the value of α decreases.

Using the level r_α determined using the value N_0 for N, we see that the power function

$$\pi(N) = P_N\{R \geq r_\alpha\}.$$

R is a hypergeometric random variable with mass function

$$f_R(r) = P_N\{R = r\} = \frac{\binom{t}{r}\binom{N-t}{k-r}}{\binom{N}{k}}.$$

The plot for the case $\alpha = 0.05$ is given using the R *commands*

```
> N<-c(1300:2100)
> pi<-1-phyper(49,t,N-t,k)
> plot(N,pi,type="l",ylim=c(0,1))
```

We can increase power by increasing the size of k, the number the value in the second capture. This increases the value of r_α. For $\alpha = 0.05$, we have the table.

```
> k<-c(400,600,800)
> N0<-2000
> ralpha<-qhyper(0.95,t,N0-t,k)
> data.frame(k,ralpha)
    k ralpha
1 400     49
2 600     70
3 800     91
```

We show the impact on power $\pi(N)$ of both significance level α and the number in the recapture k in Figure 18.3.

Exercise 18.4. *Determine the type II error rate for $N = 1600$ with*

- *$k = 400$ and $\alpha = 0.05, 0.02,$ and 0.01, and*
- *$\alpha = 0.05$ and $k = 400, 600,$ and 800.*

Example 18.5. *For a* **two-sided test**

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

In this case, the parameter values for the null hypothesis Θ_0 consist of a single value, μ_0. We reject H_0 if $|\bar{X} - \mu_0|$ is too large. Under the null hypothesis,

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

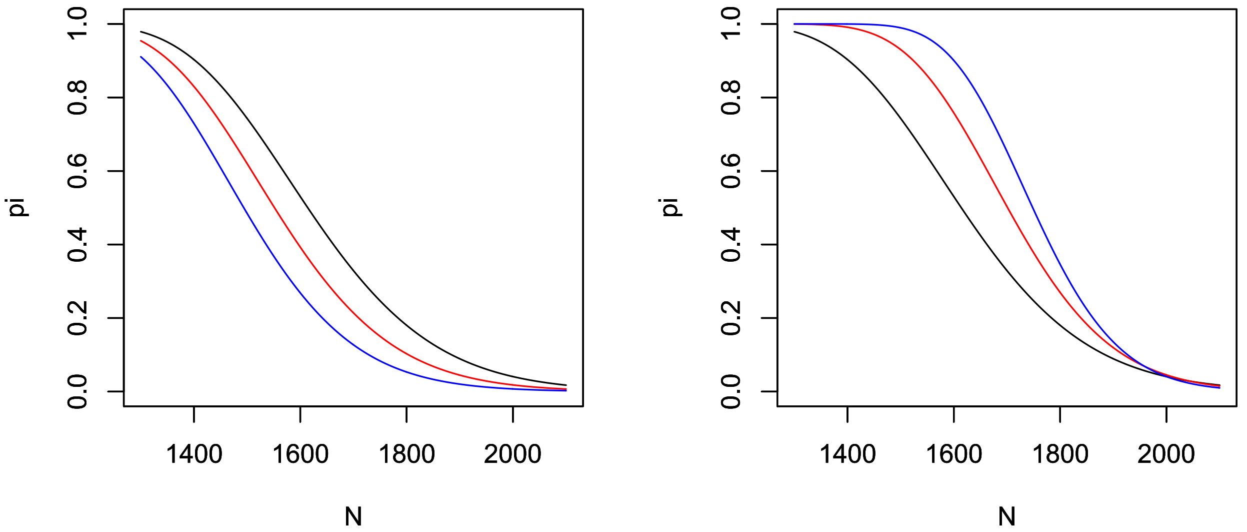

Figure 18.3: Power function for Lincoln-Peterson mark and recapture test for population $N_0 = 2000$ and $t = 200$ captured and tagged. **(left)** $k = 400$ recaptured $\alpha = 0.05$ (black), 0.02 (red), and 0.01 (blue). Notice that lower significance level α reduces power. **(right)** $\alpha = 0.05$, $k = 400$ (black), 600 (red), and 800 (blue). As expected, increased recapture size increases power.

is a standard normal random variable. For a significance level α*, choose* $z_{\alpha/2}$ *so that*

$$P\{Z \geq z_{\alpha/2}\} = P\{Z \leq -z_{\alpha/2}\} = \frac{\alpha}{2}.$$

Thus, $P\{|Z| \geq z_{\alpha/2}\} = \alpha$*. For data* $\mathbf{x} = (x_1, \ldots, x_n)$*, this leads to a critical region*

$$C = \left\{\mathbf{x}; \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| \geq z_{\alpha/2}\right\}.$$

If μ *is the actual mean, then*

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

is a standard normal random variable. We use this fact to determine the power function for this test

$$\begin{aligned}
\pi(\mu) &= P_\mu\{X \in C\} = 1 - P_\mu\{X \notin C\} = 1 - P_\mu\left\{\left|\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}\right| < z_{\alpha/2}\right\} \\
&= 1 - P_\mu\left\{-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} < z_{\alpha/2}\right\} = 1 - P_\mu\left\{-z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} < z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right\} \\
&= 1 - \Phi\left(z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right) + \Phi\left(-z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right)
\end{aligned}$$

If we do not know if the mimic is larger or smaller that the model, then we use a two-sided test. Below is the R *commands for the power function with* $\alpha = 0.05$ *and* $n = 16$ *observations.*

```
> zalpha = qnorm(.975)
> mu0<-10
> sigma0<-3
> mu<-(600:1400)/100
```

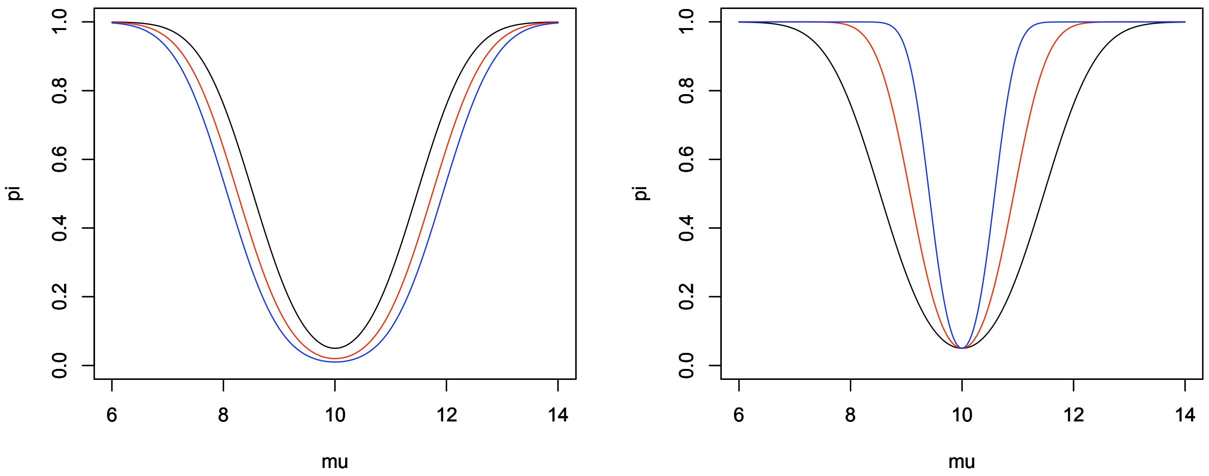

Figure 18.4: Power function for the two-sided test. $\mu_0 = 10$, $\sigma_0 = 3$. **(left)** $n = 16$, $\alpha = 0.05$ (black), 0.02 (red), and 0.01 (blue). Notice that lower significance level α reduces power. **(right)** $\alpha = 0.05$, $n = 15$ (black), 40 (red), and 100 (blue). As before, decreased significance level reduces power and increased sample size n increases power.

```
> n<-16
> pi<-1-pnorm(zalpha-(mu-mu0)/(sigma0/sqrt(n)))
  +pnorm(-zalpha-(mu-mu0)/(sigma0/sqrt(n)))
> plot(mu,pi,type="l")
```

We shall see in the the next topic how these tests follow from extensions of the likelihood ratio test for simple hypotheses.

The next example is unlikely to occur in any genuine scientific situation. It is included because it allows us to compute the power function explicitly from the distribution of the test statistic. We begin with an exercise.

Exercise 18.6. *For $X_1, X_2, \ldots, X_n$ independent $U(0, \theta)$ random variables, $\theta \in \Theta = (0, \infty)$. The density*

$$f_X(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_{(n)}$ denote the maximum of $X_1, X_2, \ldots, X_n$, then $X_{(n)}$ has distribution function

$$F_{X_{(n)}}(x) = P_\theta\{X_{(n)} \le x\} = \left(\frac{x}{\theta}\right)^n.$$

Example 18.7. *For $X_1, X_2, \ldots, X_n$ independent $U(0, \theta)$ random variables, take the null hypothesis that θ lands in some normal range of values $[\theta_L, \theta_R]$. The alternative is that θ lies outside the normal range.*

$$H_0 : \theta_L \le \theta \le \theta_R \quad \text{versus} \quad H_1 : \theta < \theta_L \text{ or } \theta > \theta_R.$$

Because θ is the highest possible value for an observation, if any of our observations X_i are greater than θ_R, then we are certain $\theta > \theta_R$ and we should reject H_0. On the other hand, all of the observations could be below θ_L and the maximum possible value θ might still land in the normal range.

Consequently, we will try to base a test based on the statistic $X_{(n)} = \max_{1 \le i \le n} X_i$ and reject H_0 if $X_{(n)} > \theta_R$ and too much smaller than θ_L, say $\tilde{\theta}$. We shall soon see that the choice of $\tilde{\theta}$ will depend on n the number of observations and on α, the size of the test.

The power function

$$\pi(\theta) = P_\theta\{X_{(n)} \le \tilde{\theta}\} + P_\theta\{X_{(n)} \ge \theta_R\}$$

We compute the power function in three cases - low, middle and high values for the parameter θ. The second case has the values of θ under the null hypothesis. The first and the third cases have the values for θ under the alternative hypothesis. An example of the power function is shown in Figure 18.5.

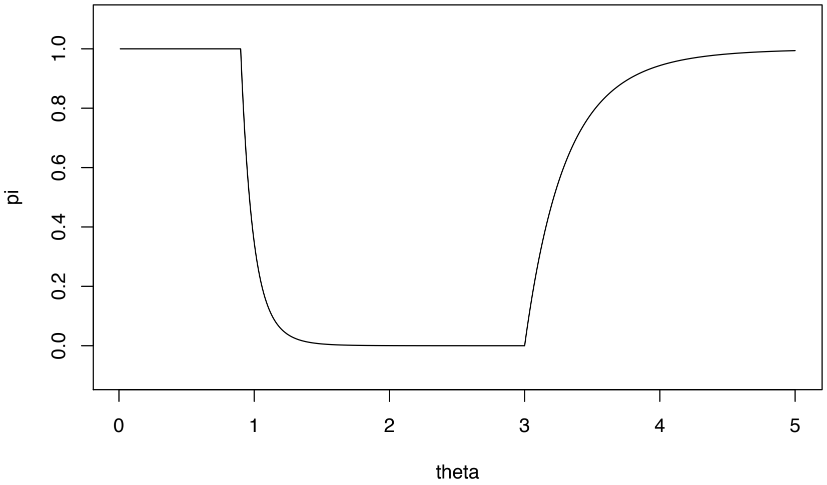

Figure 18.5: Power function for the test above with $\theta_L = 1, \theta_R = 3, \tilde{\theta} = 0.9$, and $n = 10$. The size of the test is $\pi(1) = 0.3487$.

Case 1. $\theta \le \tilde{\theta}$.

In this case all of the observations X_i must be less than θ which is in turn less than $\tilde{\theta}$. Thus, $X_{(n)}$ is certainly less than $\tilde{\theta}$ and

$$P_\theta\{X_{(n)} \le \tilde{\theta}\} = 1 \text{ and } P_\theta\{X_{(n)} \ge \theta_R\} = 0$$

and therefore $\pi(\theta) = 1$.

Case 2. $\tilde{\theta} < \theta \le \theta_R$.

Here $X_{(n)}$ can be less that $\tilde{\theta}$ but never greater than θ_R.

$$P_\theta\{X_{(n)} \le \tilde{\theta}\} = \left(\frac{\tilde{\theta}}{\theta}\right)^n \text{ and } P_\theta\{X_{(n)} \ge \theta_R\} = 0$$

and therefore $\pi(\theta) = (\tilde{\theta}/\theta)^n$.

Case 3. $\theta > \theta_R$.

Repeat the argument in Case 2 to conclude that

$$P_\theta\{X_{(n)} \le \tilde{\theta}\} = \left(\frac{\tilde{\theta}}{\theta}\right)^n$$

and that

$$P_\theta\{X_{(n)} \ge \theta_R\} = 1 - P_\theta\{X_{(n)} < \theta_R\} = 1 - \left(\frac{\theta_R}{\theta}\right)^n$$

and therefore $\pi(\theta) = (\tilde{\theta}/\theta)^n + 1 - (\theta_R/\theta)^n$.

The size of the test is the maximum value of the power function under the null hypothesis. This is case 2. Here, the power function

$$\pi(\theta) = \left(\frac{\tilde{\theta}}{\theta}\right)^n$$

decreases as a function of θ*. Thus, its maximum value takes place at* θ_L *and*

$$\alpha = \pi(\theta_L) = \left(\frac{\tilde{\theta}}{\theta_L}\right)^n$$

To achieve this level, we solve for $\tilde{\theta}$*, obtaining* $\tilde{\theta} = \theta_L \sqrt[n]{\alpha}$*. Note that* $\tilde{\theta}$ *increases with* α*. Consequently, we must expand the critical region in order to reduce the significance level. Also,* $\tilde{\theta}$ *increases with* n *and we can reduce the critical region while maintaining significance if we increase the sample size.*

The assessment of statistical power is an important aspect of experimental design. In practical terms, we can *increase* power by either *increasing effort* or *asking a less stringent question*. For example, we can increase effort

- **(mathematics)** by applying a more powerful test or a more rigorous design,
- **(engineering)** by designing a better measuring devise, reducing variance, or
- **(exersion)** by increasing sample size

We can ask a less stringent question

- by increasing the significance level and thus the ability to reject the null hypothesis or
- by increasing the difference between null value and the alternative value for detection of difference

These practical considerations will be useful in understanding the change in power resulting from a change in the experimental design and hypothesis testing.

18.3 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 not to choose, in advance, a significance level α of the test and then report "reject" or "fail to reject", but rather 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.

Example 18.8. *For the one-sided hypothesis test to see if the mimic had invaded,*

$$H_0 : \mu \ge \mu_0 \quad \text{versus} \quad H_1 : \mu < \mu_0.$$

with $\mu_0 = 10$ *cm,* $\sigma_0 = 3$ *cm and* $n = 16$ *observations. The test statistics is the sample mean* $\bar{x}$ *and the critical region is* $C = \{\mathbf{x}; \bar{x} \le k\}$

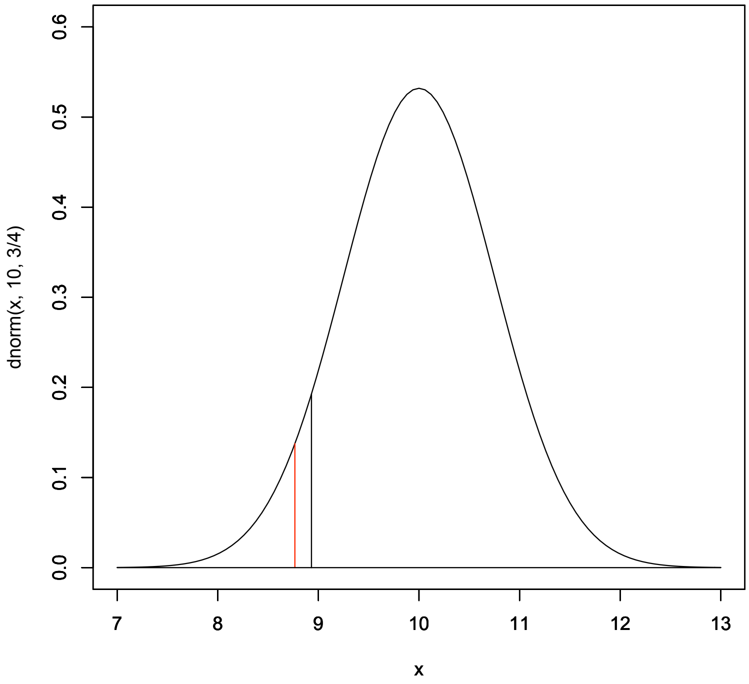

Figure 18.6: Under the null hypothesis, $\bar{X}$ has a normal distribution mean $\mu_0 = 10$ cm, standard deviation $3/\sqrt{16} = 3/4$ cm. The p-value, 0.077, is the area under the density curve to the left of the observed value of 8.931 for $\bar{x}$, The critical value, 8.767, for an $\alpha = 0.05$ level test is indicated by the red line. Because the p-vlaue is greater than the significance level, we cannot reject H_0.

Our data had sample mean $\bar{x} = 8.93125$ cm. The maximum value of the power function $\pi(\mu)$ for μ in the subset of the parameter space determined by the null hypothesis occurs for $\mu = \mu_0$. Consequently, the p-value is

$$P_{\mu_0}\{\bar{X} \leq 8.93125\}.$$

With the parameter value $\mu_0 = 10$ cm, $\bar{X}$ has mean 10 cm and standard deviation $3/\sqrt{16} = 3/4$. We can compute the p-value using R.

```
> pnorm(8.93125,10,3/4)
[1] 0.0770786
```

If the p-value is below a given significance level α, then we say that the result is **statistically significant** at the level α. For the previous example, we could not have rejected H_0 at the $\alpha = 0.05$ significance level. Indeed, we could not have rejected H_0 at any level below the p-value, 0.0770786. On the other hand, we would reject H_0 for any significance level above this value.

Many statistical software packages (including R, see the example below) do not need to have the significance level in order to perform a test procedure. This is especially important to note when setting up a hypothesis test for the purpose of deciding whether or not to reject H_0. In these circumstances, the significance level of a test is a value that should be decided *before* the data are viewed. After the test is performed, a report of the p-value adds information beyond simply saying that the results were or were not significant.

It is tempting to associate the p-value to a statement about the probability of the null or alternative hypothesis being true. Such a statement would have to be based on knowing which value of the parameter is the true state of nature. Assessing whether of not this parameter value is in Θ_0 is the reason for the testing procedure and the p-value was computed in knowledge of the data and our choice of Θ_0.

In the example above, the test is based on having a test statistic $S(\mathbf{x})$ (namely $\bar{x}$) fall below a level k_α, i.e., we have decision

$$\text{reject } H_0 \text{ if and only if } S(\mathbf{x}) \leq k_\alpha.$$

This choice of k_α is based on the choice of significance level α and the choice of $\theta_0 \in \Theta_0$ so that $\pi(\theta_0) = P_{\theta_0}\{S(X) \leq k_\alpha\} = \alpha$, the lowest value for the power function under the null hypothesis. If the *observed* data $\mathbf{x}$ takes the value $S(\mathbf{x}) = s$, then the p-value equals

$$P_{\theta_0}\{S(X) \leq s\}. \tag{18.4}$$

This is the lowest value for the significance level that would result in rejection of the null hypothesis *if we had chosen it in advance of seeing the data*.

Example 18.9. *Returning to the example on the proportion of hives that survive the winter, the appropriate composite hypothesis test to see if more that the usual normal of hives survive is*

$$H_0 : p \leq 0.7 \quad \textit{versus} \quad H_1 : p > 0.7.$$

The R *output shows a p-value of 3%.*

```
>  prop.test(88,112,0.7,alternative="greater")

1-sample proportions test with continuity correction

data:  88 out of 112, null probability 0.7
X-squared = 3.5208, df = 1, p-value = 0.0303
alternative hypothesis: true p is greater than 0.7
95 percent confidence interval:
 0.7107807 1.0000000
sample estimates:
        p
0.7857143
```

Exercise 18.10. *Is the hypothesis test above significant at the 5% level? the 1% level?*

In 2016, the American Statistical Association set for itself a task to make a statement on p-values. They note that it is all too easy to set a test, create a test statistic and compute a p-value. Proper statistical practice is much more than this and includes

- appropriately chosen techniques based on a thorough understanding of the phenomena under study,
- adequate visual and numerical summaries of the data,
- properly conducted analyses whose logic and quantitative approaches are clearly explained,
- correct interpretation of statistical results in context, and
- reproducibility of results via a thorough reporting.

Expressing a p-value is one of many approaches to summarize the results of a statistical investigation. The notion is that the smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis. This incompatibility is meant to cast doubt on the null hypothesis.

Under the logic of classical statistics, the p-value cannot be turned into a statement about the truth of the null hypothesis but rather is a statement about the data in relation to a specified statistical model stated as a hypothesis test. Moreover, the p-value is not meant to serve as a "bright line" between true and false. Part of this arises from the pedogogy of introducing of hypothesis testing in setting a significance level α as a part of the test.

These issues are compounded in most scientific considerations where multiple hypothesis testing makes interpretation of p-values difficult and calls on the authors for complete transparency of all statistical procedures including data collection and hypothesis testing. In addition, even strong statistical evidence of the incompatibility of the data with the null hypothesis may have very little practical or scientific meaning.

Many investigators engage in statistical analysis based on limited background and so often need to collaborate to find other appropriate statistical approached to decision making under uncertainty. Some appear in this book, e.g., Bayes factors, likelihood ratios, and false discovery rates, but there are many others.

18.4 Distribution of p-values and the Receiving Operating Characteristic

Let's return to the case of a simple hypotheses.

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

As before, for data $\mathbf{x}$, let $S(\mathbf{x})$ be a test statistic for this hypothesis, rejecting if the value of test statistic $S(\mathbf{x})$ is too low. If $S(\mathbf{x}) = s$, the p-value is $F_{S(X)}(s|\theta_0) = P_{\theta_0}\{S(X) \leq s\}$. Recalling out introductory example on model and mimic butterflies, the hypothesis on the mean wing span in centimeters, is

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

In this situation, the test statistic $S(X) = \bar{X}$ is $N(\mu_0, \sigma/\sqrt{n})$ under the null hypothesis. If the standard deviation is `sigma` with `n` observations, using `xbar` to denote the sample mean, we find the p-value with the command `pnorm(xbar,mu0,sigma/sqrt(n))`.

For a significance test at level α, there exists a critical value k_α so that

$$\alpha = P_{\theta_0}\{S(X) \leq k_\alpha\} = F_{S(X)}(k_\alpha|\theta_0)$$

and we reject the null hypothesis at level α if the value of the test statistic is below the critical value, i. e., $s < k_\alpha$. Thus, for significance level `alpha`, we determine k_α with the command `qnorm(alpha,mu0,sigma/sqrt(n))`.

For the parameter value θ_1, the power

$$1 - \beta(\alpha) = P_{\theta_1}\{S(X) \leq k_\alpha\}.$$

In other words, $1 - \beta(\alpha)$ is the probability, under the alternative parameter θ_1 that the p-value is less than α.

Define $F_R(\alpha) = 1 - \beta(\alpha)$, then F_R is a non-decreasing function on the interval $[0, 1]$ with $F_R(0) = 0$ and $F_R(1) = 1$. Thus, F_R is a *cumulative distribution function*. Recall that the **receiving operator characteristic** is the plot of the power as a function of significance. In other words, it is the plot $F_R(\alpha)$.

Exercise 18.11. *Show that the receiver operating characteristic gives the distribution function for the p-values for the alternative parameter value θ_1.*

The **area under the receiving operator characteristic, AUC**,

$$\int_0^1 F_R(\alpha)\, d\alpha.$$

is a general diagnostic for the overall power of a test. If the AUC is nearly 1, then the power has the very desirable property of increasing quickly for low significance levels.

Exercise 18.12. *Let S_i, $i = 0, 1$ be independent random variables that have the distributions of $S(X)$ under θ_i. The the area under the curve equals*

$$\int_{-\infty}^{\infty} F_1(s_0) f_0(s_0)\, ds_0 = P\{S_1 < S_0\}. \tag{18.5}$$

In words, for two independent samples of the test statistic, one under the null hypothesis and the other under the alternative, the area under the curve is the probability that the value under the alternative is smaller. We will see a similar expression in an alternative approach to t procedures. This will leads to the Wilcoxon ranked sum test and an interpretation associated to the area under the empirical receiving operator characteristic.

Exercise 18.13. *For $n = 16$ observations, standard deviation $\sigma = 3$ and $\mu_0 = 10$ centimeters, determine the values for the area under the receiver operator characteristics in Figure 17.3.*

μ_1	AUC
9	0.8271
8	0.9703
7	0.9977

Hint: *Use the* `integrate` *command for the integral in (18.5)*

Notice that, as expected, as the difference $\mu_0 - \mu_1$ increases, the mimic and the model butterfly are easier to distinguish and the AUC increases.

Exercise 18.14. *Simulate $P\{S_1 < S_0\}$ in the previous exercise and see how they match the values for the AUC.*

18.5 Multiple Hypothesis Testing

We now consider testing *multiple* hypotheses. This is common in the world of "big data" with thousands of hypothesis on many issues in subjects including genomics, internet searches, or financial transactions. For m hypotheses, let $p_1, \ldots, p_m$ be the p-values for m hypothesis tests.

18.5.1 Familywise Error Rate

The **familywise error rate** (FWER) is the probability of making even one type I error. If we set α_B for the significance level for a single test, then the simplest strategy is to employ the **Bonferroni correction**. This uses the Bonferroni inequality,

$$P(A_1 \cup \cdots \cup A_m) \leq P(A_1) + \cdots + P(A_m)$$

for events $A_1, \ldots, A_m$.

If A_i is the event of rejecting the null hypothesis when it is true, then $A_1 \cup \cdots \cup A_m$ is the event that at least one of the hypotheses is rejected when it is true. For each i, $P(A_i) = \alpha_B$ and so $\alpha = P(A_1 \cup \cdots \cup A_m) \leq m\alpha_B$. Thus, the Bonferroni correction is to reject if

$$p_i \leq \frac{\alpha}{m} \quad \text{for all } i.$$

Exercise 18.15. *For m independent, α_I level hypothesis tests, show that the familywise error $\alpha = 1 - (1-\alpha_I)^m$. Thus, $(1-\alpha)^{1/m} = 1 - \alpha_I$ and $\alpha_I = 1 - (1-\alpha)^{1/m}$ is the level necessary to obtain an α familywise error rate.*

This gives a cautionary take, if we take $\alpha = 0.05$ and $m = 20$, then the probability of one or more false positive tests, $1 - (1-0.05)^{20} \approx 0.64$, is well above 1/2. The Bonferroni correction, $\alpha_B = 0.05/20 = 0.0025$ and the independence correction, $\alpha_I = 1 - (1-0.05)^{1/20} = 0.0256$ will guarantee a familywise error rate $\alpha = 0.05$

Note that the second method allows for slightly higher values of α than the Bonferroni correction. However, it is far less general. For independent test statistics, **Fisher's method** for testing multiple works directly with the p-values. We begin with the following exercise.

Exercise 18.16. *Let θ_0 be the true state of nature. Assume that the distribution function for the text statistic $S(X)$ is continuous and strictly increasing for all possible values. Show that the p-value is uniformly distributed on the interval* $[0, 1]$.

In this circumstance, if the null hypothesis is true for all m hypotheses, then

$$p_1, \ldots, p_m \text{ are independent } U(0,1) \text{ random variables.}$$

Recall from the use of the probability transform that

$$-2\ln p_1, \ldots, -2\ln p_m \text{ are independent } Exp(1/2) \text{ random variables.}$$

So their sum

$$-2\ln p_1 - \cdots - 2\ln p_m \text{ is a } \Gamma(1/2, m) \text{ random variable.}$$

Thus this Γ random variable can serve as a test statistic for the multiple hypothesis that all the null hypotheses are true, rejecting if the sum above is sufficiency large. Traditionally, we use the fact that $\Gamma(1/2, m)$ is also a member of the chi-square family, namely, χ^2_{2m} and then use this as the distribution of $-2\ln p_1 - \cdots - 2\ln p_m$ under the multiple hypothesis that all m null hypotheses hold.

Example 18.17. *For 10 independent test consider the p-values*

```
> p
 [1] 0.0086 0.0164 0.6891 0.7671 0.2967 0.5465 0.0247 0.8235 0.9603 0.0041
```

The test statistic for Fisher's method

```
 > -2*sum(log(p))
[1] 41.5113
```

gives a p-value of 0.3% for the multiple test for all 10 hypotheses.

```
> 1-pchisq(-2*sum(log(p)),2*10)
[1] 0.003200711
```

18.5.2 False Discovery Rate

When the number of tests becomes very large, then having all hypotheses true is an extremely strict criterion. A more relaxed and often more valuable criterion is the **false discovery rate**..

Thus, we can model question *Is the null hypothesis hypothesis true?* as a sequence of Bernoulli trials. Let π_0 be the success parameter for the trials. Thus, with probability π_0, the null hypothesis is true and the p-values follow F_U, the uniform distribution on the interval $[0, 1]$. With probability $1 - \pi_0$, the null hypothesis is false and the p-values follow F_R, the distribution of the receiver operating characteristic. Taken together, we say that the p-values are distributed according to the mixture

$$F(x) = \pi F_U(x) + (1 - \pi)F_R(x) = \pi_0 x + (1 - \pi_0)F_R(x). \qquad (18.6)$$

Thus, if we reject whenever the p-value is below a chosen value α, then the type I error probability is α. From this we determine the false discovery rate, here defined as

$$q = P\{H_0 \text{ is true}|\text{reject } H_0\}.$$

Using Bayes formula

$$q = \frac{P\{\text{reject } H_0|H_0 \text{ is true}\}P\{H_0 \text{ is true}\}}{P\{\text{reject } H_0\}} = \frac{\alpha\pi_0}{F(\alpha)}.$$

An estimate of the false discovery rate can be determined from an estimate of π_0. This is determined by looking at the p-values and estimating the mixture in (18.6).

Example 18.18. *Consider a simple hypothesis*

$$H_0 : \mu = 0 \quad \textit{versus} \quad H_1 : \mu = 1.$$

for the mean μ based on 16 observations of normal random variable, variance 1. Thus, either the effect is not present ($\mu = 0$), or it is ($\mu = 1$). If we take the significance level $\alpha = 0.01$, then based on $n = 16$ observations, the test statistic $\bar{X}$ has standard deviation $1/\sqrt{16} = 1/4$,

```
> alpha<-0.01
> (kalpha<-qnorm(1-alpha,0,1/4))
[1] 0.581587
```

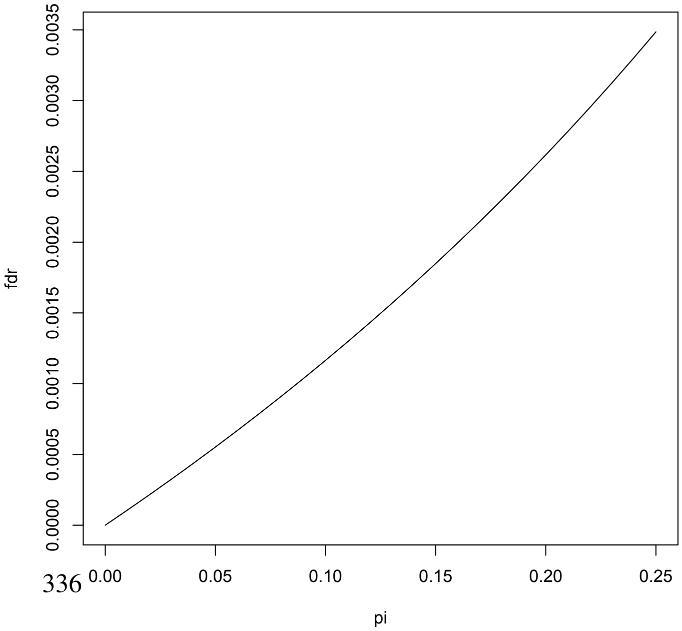

Figure 18.7: False discovery rate versus π. Here the significance level $\alpha = 0.01$, the power, $\beta = 0.953$.

and, thus, we reject H_0 if the sample mean $\bar{x} > k_\alpha = 0.581587$. The power, i.e.., the probability that we reject H_0 when H_1 is true,

```
> (p_1<-1-pnorm(xbar,1,1/4))
[1] 0.9529005
```

If we plot the false discovery rate versus π_0, the probability H_0 is true, then

```
> pi<-seq(0,0.25,0.01)
> fdr<-alpha*pi0/(alpha*pi0+p_1*(1-pi))
> plot(pi0,fdr,type="l")
```

In this case, for $\pi = 0.10$, we have a false discovery rate $q = 0.00116$, For 10,000 hypothesis, we have a mean of 11.6 false discoveries.

18.6 Answers to Selected Exercises

18.2. In this case the critical regions is $C = \{\mathbf{x}; \bar{x} \le k(\mu_0)\}$ for some value $k(\mu_0)$. To find this value, note that

$$P_{\mu_0}\{Z \le -z_\alpha\} = P_{\mu_0}\{\bar{X} \le -\frac{\sigma_0}{\sqrt{n}} z_\alpha + \mu_0\}$$

and $k(\mu_0) = -(\sigma_0/\sqrt{n})z_\alpha + \mu_0$. The power function

$$\begin{aligned}\pi(\mu) &= P_\mu\{\bar{X} \le -\frac{\sigma_0}{\sqrt{n}} z_\alpha + \mu_0\} = P_\mu\{\bar{X} - \mu \le -\frac{\sigma_0}{\sqrt{n}} z_\alpha - (\mu - \mu_0)\} \\ &= P_\mu\left\{\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \le -z_\alpha - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right\} = \Phi\left(-z_\alpha - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right).\end{aligned}$$

18.4. The type II error rate β is $1 - \pi(1600) = P_{1600}\{R < r_\alpha\}$. This is the distribution function of a hypergeometric random variable and thus these probabilities can be computed using the `phyper` command

- For varying significance, we have the R commands:

  ```
  > t<-200;N<-1600
  > k<-400
  > alpha<-c(0.05,0.02,0.01)
  > ralpha<-c(49,51,53)
  > beta<-1-phyper(ralpha-1,t,N-t,k)
  > data.frame(alpha,beta)
    alpha      beta
  1  0.05 0.5993010
  2  0.02 0.4609237
  3  0.01 0.3281095
  ```

  Notice that the type II error probability is high for $\alpha = 0.05$ and increases as α decreases.

- For varying recapture size, we continue with the R commands:

  ```
  > k<-c(400,600,800)
  > ralpha<-c(49,70,91)
  > beta<-1-phyper(ralpha-1,t,N-t,k)
  ```