8 Testing of Hypotheses and Confidence Regions

There are some problems we meet in statistical practice in which estimation of a parameter is not the primary goal; rather, we wish to use our data to decide whether we believe in a particular claim, theory, or conjecture. For example, a drug manufacturer may claim that a specific dose of its cholesterol lowering drug reduces the blood LDL level by 20 points after treatment for 2 months. We may carefully design an experiment and obtain some data on the level of LDL reduction in some subjects, and use that data to either accept the drug manufacturer's claim as believable, or reject the claim as an exaggeration. Properly formulated, such inference problems are called *testing of hypotheses* problems. Some think that testing of hypotheses is a more fundamental problem of inference than point estimation. Principal references for this chapter are Lehmann (1986), Bickel and Doksum (2001), Rao (1973), Ferguson (1967), Kendall and Stuart (1977), Welsh (1996), and Berger (1986); also see DasGupta (2008).

Testing problems, just like point estimation, require a model, a parameter space and an action space, a loss function, and then assessment of decision procedures by looking at risk functions. By tradition, decision procedures in testing problems are called *tests*.

Example 8.1. (Testing for a Normal Mean). Take the cholesterol lowering example. Suppose that in a clinical trial the observed LDL reductions of n subjects are $X_1, \cdots, X_n$. For the sake of working out an example, suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, 25)$. The drug manufacturer's claim is $\mu = 20$. If the manufacturer is not being truthful, we think that it is erring on the side of exaggeration, in which case the true value of μ is less than 20. These are our two *hypotheses*, the *null hypothesis* $H_0 : \mu = 20$, and the *alternative hypothesis*, $H_1 : \mu < 20$.

Notice that if H_0 is true, then that compels the underlying distribution to be $N(20, 25)$, which is a single fixed distribution with no unknown parameters in it; such a hypothesis is called a *simple hypothesis*. However, if the alternative hypothesis H_1 holds, the underlying distribution can be $N(\mu, 25)$ for any $\mu < 20$; thus, H_1 does not uniquely specify the underlying distribution. Such a hypothesis is called a *composite hypothesis*. So, in this example, we will test a simple null against a composite alternative. We will use our data to choose between these two hypotheses; in *two action testing problems*, the statistician's action space consists of only two actions, $\mathcal{A} = \{a_0, a_1\}$, where

$$a_0 \equiv \text{Accept } H_0;$$

$$a_1 \equiv \text{Reject } H_0.$$

Operationally, rejecting H_0 amounts to saying that between H_0 and H_1, we are picking H_1.

8.0.1 Testing as a Decision Problem

In general, for a general parameter θ of the distribution of the data, the null hypothesis is a statement of the form $H_0 : \theta \in \Theta_0$, and the alternative hypothesis is the statement $H_1 : \theta \in \Theta_1$, where Θ_0 is some suitable proper subset of the full parameter space Θ, and $\Theta_1 = \Theta - \Theta_0$. Thus, Θ_0 and Θ_1 are disjoint, and $\Theta_0 \cup \Theta_1 = \Theta$.
Here is a formal definition of what we call a test.

Definition 8.1. Let $X^{(n)} = (X_1, \cdots, X_n) \sim P = P_n$ be the vector of sample observations, and let $\theta = h(P) \in \Theta$ be a parameter of P. Given two hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, and the two point action space $\mathcal{A} = \{a_0, a_1\}$, a *pure test or test function* is any function $\phi(X_1, \cdots, X_n)$ taking values in the set $\mathcal{A}$.
Verbally, a pure test is a rule for you to pick exactly one of the null and the alternative hypothesis by using your data.

Example 8.2. (Testing for a Normal Mean). In the cholesterol example, our hypotheses are $H_0 : \mu = 20;\ H_1 : \mu < 20$. Each of the following is a pure test:

$$\text{Take action } a_1, \text{ i.e., reject } H_0 \text{ if } \bar{X} < 19.$$

$$\text{Take action } a_1, \text{ i.e., reject } H_0 \text{ if } M_n < 19,$$

where M_n denotes the median of the sample observations $X_1, \cdots, X_n$. For future reference, we denote these two tests as ϕ_1, ϕ_2 respectively. The first test ϕ_1 uses $\bar{X}$ as its *test statistic*, while the second test ϕ_2 uses M_n as its *test statistic*. There is choice in which test statistic you should use. Later in this chapter, we will have detailed discussions on the *optimal test statistic* to use in a given problem.

8.0.2 Evaluating a Test

Like estimation, testing is a decision problem. Hence, tests, which are the decision procedures in a testing problem, are evaluated on the basis of their risk functions. To compute a risk function, we first need a loss function. In testing problems, essentially the only loss function that has been studied seriously is the following *zero-one loss function*:

$$\begin{aligned} L(\theta, a_0) &= 0, \quad \text{if } \theta \in \Theta_0 \\ &= 1, \quad \text{if } \theta \in \Theta_1 \end{aligned}$$

$$\begin{aligned} L(\theta, a_1) &= 1, \quad \text{if } \theta \in \Theta_0 \\ &= 0, \quad \text{if } \theta \in \Theta_1 \end{aligned}$$

Thus, there is no loss in making the correct decision, and there is a loss of one unit in making an incorrect decision. The zero-one loss has been criticized because it penalizes

the constant one unit for incorrectly picking a_0 when the true $\theta \in \Theta_1$. The usual criticism is that if the true value of θ is very far from Θ_0, but we took action a_0, then such a bad mistake should be penalized more than taking action a_0 when the true value of θ is just outside Θ_0. These critics would be happier with a loss function such as

$$\begin{aligned} L(\theta, a_0) &= 0, && \text{if } \theta \in \Theta_0 \\ &= \text{dist}(\theta, \Theta_0)\ , && \text{if } \theta \in \Theta_1 \end{aligned}$$

where $\text{dist}(\theta, \Theta_0)$ denotes some suitable distance measure of θ from the null set Θ_0.

In spite of these criticisms, the use of the zero-one loss function for testing has not subsided since Jerzy Neyman and Egon Pearson wrote it down in their classic and seminal 1933 paper, Neyman and Pearson (1933).
Under this zero-one loss function, the risk function of a test is

$$\text{For } \theta \in \Theta_0, R(\theta, \phi) = 0 \times P_\theta(\phi(X_1, \cdots, X_n) = a_0) + 1 \times P_\theta(\phi(X_1, \cdots, X_n) = a_1)$$

$$= P_\theta(\phi(X_1, \cdots, X_n) = a_1) = P_\theta(\text{The corresponding test rejects } H_0);$$

and,

$$\text{For } \theta \in \Theta_1, R(\theta, \phi) = 0 \times P_\theta(\phi(X_1, \cdots, X_n) = a_1) + 1 \times P_\theta(\phi(X_1, \cdots, X_n) = a_0)$$

$$= P_\theta(\phi(X_1, \cdots, X_n) = a_0) = P_\theta(\text{The corresponding test accepts } H_0).$$

In words, the risk function, which is a function of θ, is the probability of rejecting the null if θ is a null value, and the risk function is the probability of accepting the null if θ is an alternative hypothesis value. These are such fundamental characteristics of a test, that statisticians have given them names for easy reference. They are called the *type I and type II error probabilities*, respectively:

$$\text{Type I Error Probability } = \alpha(\theta) = P_\theta(\text{Test rejects } H_0), \theta \in \Theta_0;$$

$$\text{Type II Error Probability } = \beta(\theta) = P_\theta(\text{Test accepts } H_0), \theta \in \Theta_1.$$

The probability of not making a type II error is called the *power of the test*; so, in general, power of a test is a function of θ for $\theta \in \Theta_1$:

$$\text{Power of the Test } = \gamma(\theta) = P_\theta(\text{Test rejects } H_0), \theta \in \Theta_1.$$

You could verbally interpret the two error probabilities and power as follows:

Type I Error Probability = The Probability that the Test Rejects a Null when the Null is True;

Type II Error Probability = The Probability that the Test Accepts a Null when the Null is False;

Power = The Probability that the Test Rejects a Null when the Null is False.

Hence, power reflects a test's ability to recognize that a proposed null hypothesis is false. So, we strive for a large power, and a small type I error probability. This is the same as saying that we are still striving for a low risk, when the loss function is zero-one.

8.0.3 Computing and Graphing the Error Probabilities

Let us return to our illustrative normal mean example and compute the power and the type I error for each of the two tests we are considering.

Example 8.3. (Testing for a Normal Mean). First we recall our hypotheses in this example, $H_0 : \mu = 20, H_1 : \mu < 20$. We will calculate the type I error probability and the power of each of our two tests, which we denote as ϕ_1, ϕ_2. First consider the test ϕ_1 which rejects H_0 if $\bar{X} < 19$. Its type I error probability is

$$\alpha(\phi_1) = P_{\mu=20}(\bar{X} < 19) = P(Z < \frac{19-20}{\frac{5}{\sqrt{n}}})$$

$$= P(Z < -\frac{\sqrt{n}}{5}) = 1 - \Phi(\frac{\sqrt{n}}{5}).$$

Because the null hypothesis is simple, the type I error probability is just one number; note that it depends on n, and as n increases, the type I error probability *decreases*. You would expect that.

Next, the power of this test is going to be a function, not a number, as the alternative hypothesis is composite. The power function is, for $\mu < 20$,

$$\gamma(\mu, \phi_1) = P_\mu(\bar{X} < 19) = P(Z < \frac{19-\mu}{\frac{5}{\sqrt{n}}})$$

$$= P(Z < \frac{\sqrt{n}(19-\mu)}{5}) = \Phi(\frac{\sqrt{n}(19-\mu)}{5}).$$

For fixed n, as a function of μ, it is decreasing in μ, as it should be, intuitively. As the true μ drifts away from the null value $\mu = 20$, any reasonable test should reject the null with a higher probability.

On the other hand, for fixed μ, as a function of n, the power is increasing in n, and this too makes sense intuitively.

Now we consider the second test ϕ_2 which rejects the null if $M_n < 19$, where M_n is the sample median. Now we have a little problem. The distribution of the sample median in the normal case does have an exact formula, but it is a very messy formula (see Section 1.25). Instead, we use the *large sample approximation*

$$M_n \approx N(\mu, \sigma^2 \frac{\pi}{2n});$$

(see Section 1.28).

Then, we can approximately calculate the type I error probability and the power function of the median based test ϕ_2. The type I error probability is

$$\alpha(\phi_2) = P_{\mu=20}(M_n < 19) \approx P(Z < \frac{19-20}{\sqrt{25\frac{\pi}{2n}}})$$

$$= P(Z < -\frac{\sqrt{2n}}{5\sqrt{\pi}}) = 1 - \Phi(\frac{\sqrt{2n}}{5\sqrt{\pi}}).$$

The power function is

$$\gamma(\mu, \phi_2) = P_{\mu}(M_n < 19) \approx P(Z < \frac{19-\mu}{\sqrt{25\frac{\pi}{2n}}})$$

$$= P(Z < \frac{\sqrt{2n}(19-\mu)}{5\sqrt{\pi}}) = \Phi(\frac{\sqrt{2n}(19-\mu)}{5\sqrt{\pi}}).$$

For purposes of illustration, we take $n = 50$ and compute the type I error probability of each test and also the power function; moreover, the two power functions will then be plotted.

First,

$$\alpha(\phi_1) = 1 - \Phi(\frac{\sqrt{n}}{5}) = 1 - \Phi(1.41) = .0793.$$

The power function of ϕ_1 is

$$\gamma(\mu, (\phi_1) = \Phi(\frac{\sqrt{n}(19-\mu)}{5}) = \Phi(1.41(19-\mu)).$$

On the other hand,

$$\alpha(\phi_2) = 1 - \Phi(\frac{\sqrt{2n}}{5\sqrt{\pi}}) = 1 - \Phi(1.13) = .1292.$$

The power function of ϕ_2 is

$$\gamma(\mu, (\phi_2) = \Phi(\frac{\sqrt{2n}(19-\mu)}{5\sqrt{\pi}}) = \Phi(1.13(19-\mu)).$$

The two power functions are plotted together for visual comparison.

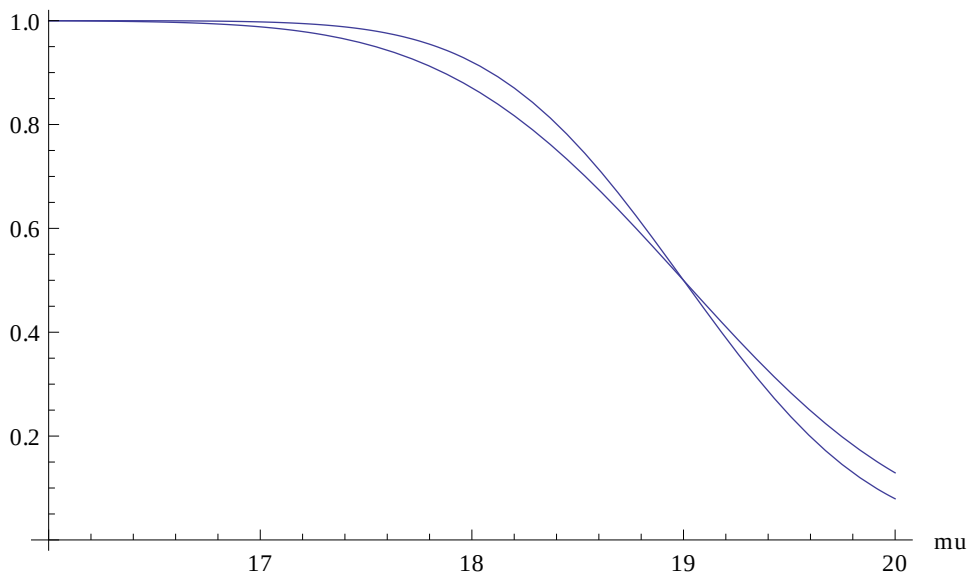
We see in the plot that

(a) The power function of both tests is a continuous monotone function.

(b), For each test, as $\mu \to 20$, the power function tends to the α value.

(c) For $19 < \mu < 20$, the median based test has larger power.

(d) At $\mu = 19$, both tests have power 0.5.



(e) For $\mu < 19$, the mean based test has larger power.

(f) The power function of both tests converges to one when $\mu << 20$.

These phenomena will be generally true. Even in simply stated and innocuous looking problems, you will often have to use large sample theory and limit theorems to compute a power function or a type I error probability. In most examples, power functions of common tests will be continuous functions of the parameters. At the null value, the power function coincides with the type I error probability. Between two tests, if one has a larger type I error probability, in exchange it will have a larger power for alternative values close to the null value. The power functions of two reasonable tests will often cross. And, as the alternative value drifts away from the null value, the power will monotonically converge to 1.

Example 8.4. (Testing for a Poisson Mean). Suppose the average number of defects per item in a particular manufacturing process was 0.2 before. This was considered too high, and a quality improvement plan was put in place for a period of time, say six weeks. Naturally, management wants to know if the quality improvement plan made a significant difference in the number of defects per item.

To formulate this problem, let X be the number of defects per item produced in the manufacturing process. We model X as a Poisson, $X \sim Poi(\lambda)$. Previously, λ was 0.2, and we want to know if it has lowered now. Thus, our hypotheses are $H_0 : \lambda = .2, H_1 : \lambda < .2$. To test the hypotheses, we sample n items from the process and count the number of defects in them, say $X_1, \cdots, X_n$. The assumption is that $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda)$. Consider

the test ϕ that rejects H_0 if $\bar{X} < .14$. This amounts to saying that management will conclude that the quality improvement plan had a significant effect if a sample from the process shows an average number of defects per item to be quite a bit smaller than the null value $\lambda = .2$; specifically, their threshold is $\bar{X} < .14$.

Our null hypothesis in this example is simple, while the alternative is composite. We will now compute the type I error probability and the power function of this test. To do explicit computation, we need a specific value of n; we use $n = 40$. For this computation, we will need to use the fact that if $X_i \overset{iid}{\sim} Poi(\lambda)$, then $\sum_{i=1}^{n} X_i \sim Poi(n\lambda)$.

The type I error probability is

$$\alpha = P_{\lambda=.2}(\bar{X} < .14) = P_{\lambda=.2}(\sum_{i=1}^{n} X_i < .14n)$$

$$= P(Poi(8) < 5.6) = P(Poi(8) \leq 5)$$

(since $40 \times .2 = 8$, and Poisson being integer valued, it is less than 5.6 only when it is less than or equal to 5)

$$= \sum_{x=0}^{5} \frac{e^{-8}8^x}{x!} = .1912.$$

The interpretation is that the test that the management has decided to use will wrongly conclude 19.12% of the times that an improvement in quality has occured although no improvement actually occured.

Now let us proceed to the power function. The power function is, for $\lambda < .2$,

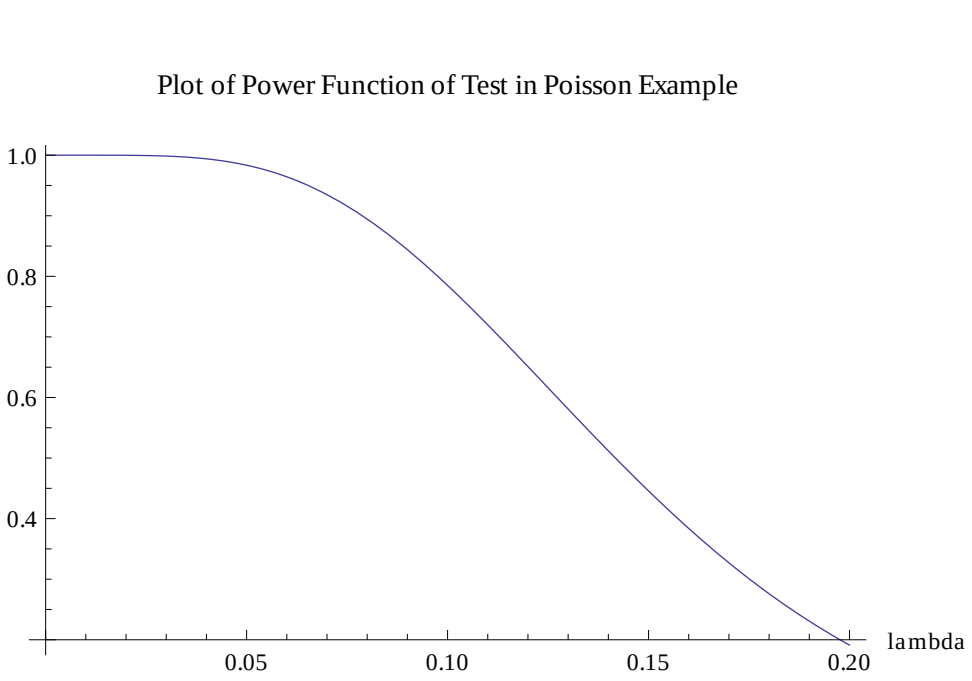
$$\gamma(\lambda) = P_\lambda(\bar{X} < .14) = P_\lambda(\sum_{i=1}^{40} X_i < 5.6)$$

$$= P(Poi(40\lambda) \leq 5) = \sum_{x=0}^{5} \frac{e^{-40\lambda}(40\lambda)^x}{x!}.$$

This can be computed easily for any given λ. We attach a plot of this power function.

For example, at $\lambda = .15$, the power is only about 0.45. That is, if quality improvement did occur, and the true value of λ actually lowered to .15 from what it was before, namely .2, our test will have only about a 45% probability of recognizing the quality improvement. Another way to say it is to say that with about a 55% probability, our test will tell us that no improvement has occured, although actually improvement did occur. On the other hand, if improvement was more drastic, and $\lambda = .08$, then the power is almost .90; i.e., if improvement is that drastic, then our test will have about a 90% probability of recognizing that quality improvement.

Notice also that exactly as in our preceding normal example, the power function increases monotonically to 1 as the value of λ drifts away from the null value $\lambda = .2$.



Example 8.5. (A Nonregular Example). We will see a power function with remarkable features in this example. In particular, the power function in this example has a point of nonsmoothness (a cusp), and the power is 1 for many values of the alternative!

Let $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta]$, and suppose we wish to test the simple null $H_0 : \theta = 1$ against the composite alternative $\theta \neq 1$. Such an alternative is called *a two sided alternative* because under the alternative, the parameter could take values on both sides (left or right) of the null value.

What test are we going to use? Arguing intuitively, if the null hypothesis is true, then the sample maximum should be below, but close to one. We may want to accept H_0 if $1 - \epsilon < X_{(n)} < 1$ for some suitable ϵ, and reject H_0 otherwise. How does such a test behave? In other words, what is its type I error probability and what is its power function? The type I error probability is

$$\alpha = P_{\theta=1}(X_{(n)} \leq 1 - \epsilon) + P_{\theta=1}(X_{(n)} \geq 1) = (1 - \epsilon)^n.$$

For the power function, we will have to consider three separate cases, $\theta > 1, \theta \leq 1 - \epsilon$, and $1 - \epsilon < \theta < 1$.

For $\theta > 1$, the power function is

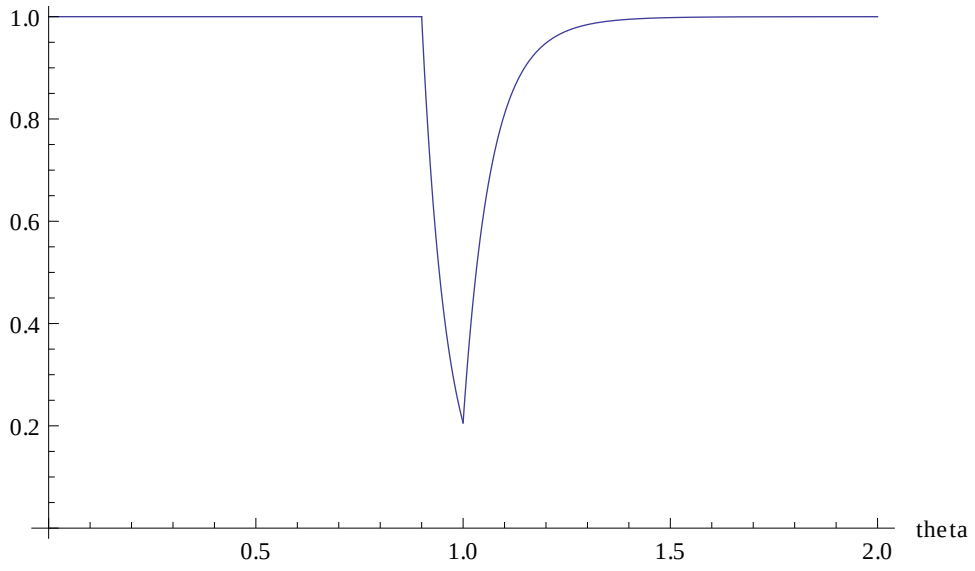
$$\gamma(\theta) = P_\theta(X_{(n)} \leq 1 - \epsilon) + P_\theta(X_{(n)} \geq 1) = (\frac{1 - \epsilon}{\theta})^n + 1 - (\frac{1}{\theta})^n.$$

For $\theta \leq 1 - \epsilon$,

$$\gamma(\theta) = P_\theta(X_{(n)} \leq 1 - \epsilon) + P_\theta(X_{(n)} \geq 1) = 1 + 0 = 1.$$

For $1 - \epsilon < \theta < 1$,

$$\gamma(\theta) = P_\theta(X_{(n)} \leq 1 - \epsilon) + P_\theta(X_{(n)} \geq 1) = (\frac{1 - \epsilon}{\theta})^n + 0 = (\frac{1 - \epsilon}{\theta})^n.$$



As an example, take $n = 15$ and $\epsilon = .1$. Then,

$$\alpha = .9^{15} = .2059.$$

The power at $\theta = 1.5$ is .9982. The power at $\theta = 0.5$ is 1, while the power at $\theta = 0.95$ is

$$(.9/.95)^{15} = .4444.$$

The plot of the power function shows the part with power one, the cusp at $\theta = 1$, and overall a very unusual power function.

Example 8.6. (Unconventional Problem). Suppose we have n iid observations from a density $f(x)$, where $f(x)$ is one of the following two densities, but we do not know which one:

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, -\infty < x < \infty;\ f_1(x) = \frac{1}{6}, -3 \le x \le 3.$$

In other words, the true density is either a standard normal, or $U[-3, 3]$. Thus, the two possible densities do not both belong to a single common parametric family. But, we could still artificially define a parameter θ taking only two values, $\theta = 0, 1$, and let our samples $X_1, \cdots, X_n \overset{iid}{\sim} f_\theta(x)$. We then test $H_0 : \theta = 0, H_1 : \theta = 1$.

Intuitively, what is a reasonable test for such a problem? On reflection, we realize that the $U[-3, 3]$ distribution is more spread out than the standard normal. For example, the variance of the $U[-3, 3]$ distribution is 3, while the standard normal has variance one. Based on this intuition, we consider the test that rejects H_0 for large values of s^2, the sample variance. For specificity, consider the test that rejects H_0 if $s^2 > 2$.

We will now calculate the type I error probability and the power of this test. Since the null and the alternative hypotheses are both simple, the type I error probability and the power are both a number, not functions. However, to calculate them we will need the distribution of s^2 under H_0 and under H_1. The distribution of s^2 under H_0 is not a problem, because when samples are from a standard normal distribution, $(n-1)s^2 \sim \chi^2_{n-1}$, i.e., we know the null distribution of s^2. But the distribution of s^2 under H_1 is not computable in closed form. So, for the power calculation, we must do an approximation, and this will be precisely explained. For specificity, we take $n = 20$.
The type I error probability of our test is

$$\alpha = P_{H_0}(s^2 > 2) = P((n-1)s^2 > 2(n-1)) = P(\chi^2(19) > 38) = .001.$$

For the power calculation, we use a normal approximation result (to be rigorously proved in Chapter 9) that for any distribution with four moments,

$$\sqrt{n}[s^2 - \sigma^2] \approx N(0, \mu_4 - \sigma^4),$$

where $\mu_4 = E(X_1 - \mu)^4$. In the $U[-3, 3]$ case,

$$\mu = 0, \sigma^2 = 3, \mu_4 = \frac{81}{5}.$$

Hence, by applying this normal approximation, the power of our test is

$$\gamma = P_{H_1}(s^2 > 2) = P_{H_1}(\frac{\sqrt{n}[s^2 - \sigma^2]}{\sqrt{\mu_4 - \sigma^4}} > \frac{\sqrt{n}[2 - \sigma^2]}{\sqrt{\mu_4 - \sigma^4}})$$

$$\approx P(Z > -1.67) = .9525.$$

8.1 Randomized and Nonrandomized Tests

In all of the examples in the previous section, our test was of the following type: choose a test statistic $T = T(X_1, \cdots, X_n)$ and reject the null when T belongs to a suitable interval of values (a, b). For instance, in the normal mean example, the test that rejects H_0 for $\bar{X} < 19$, chooses the test statistic $T = \bar{X}$ and rejects H_0 when T belongs to $(-\infty, 19)$. The point is, when our data values are available, we compute T and once we have T, we know immediately what our action should be, reject, or not reject.
Consider now the following simple example. Suppose we want to test whether the probability of heads p for some coin is .25 or .75. Suppose we toss the coin $n = 10$ times and obtain $X = 5$ heads. Based on this value of X, there does not seem to be any compelling rational reason to accept one of the hypotheses in preference to the other. Our data aren't telling us which hypothesis looks more credible. If we must choose, we may as well toss a fair coin and accept one of the two hypotheses at random. We have now effectively let

our final action be determined by the outcome of an auxiliary random experiment. Such tests which choose a final action on the basis of a new auxiliary random experiment are called *randomized tests*. The proper mathematical way to define them is to define a *test function*.

Definition 8.2. Given a general null H_0 and alternative H_1, any test of H_0 against H_1 is determined by a test function $\phi(X_1, \cdots, X_n)$ taking values in $[0, 1]$, where for any given $x_1, \cdots, x_n$,

$$\phi(x_1, \cdots, x_n) = P(\text{The test rejects } H_0 \text{ when } X_1 = x_1, \cdots, X_n = x_n).$$

An example will help understand what a test function is.

Example 8.7. (Illustrating a Test Function). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, 1)$ and we test $H_0 : \mu = 20$ against $H_1 : \mu < 20$. Consider the test that rejects H_0 if (and only if) $\bar{X} < 19$. For this test, the associated test function is

$$\begin{aligned} \phi(x_1, \cdots, x_n) &= 1, \quad \text{if} \quad \bar{x} < 19 \\ &= 0, \quad \text{if} \quad \bar{x} \geq 19. \end{aligned}$$

To put it in another way, if we define a set C as

$$C = \{(x_1, \cdots, x_n) : \bar{x} < 19\},$$

then in this case, $\phi(x_1, \cdots, x_n) = I_{(x_1, \cdots, x_n) \in C}$. Thus, this particular test function ϕ takes only the two boundary values $0, 1$. Such tests are *nonrandomized tests*, and the set C on which ϕ equals 1 is called the *critical or rejection region* of that nonrandomized test.
In contrast. consider the following example. Let $X \sim Bin(10, p)$, and we test $H_0 : p = .25$ against $H_1 : p = .75$. Consider the test which rejects H_0 if $X > 5$, accepts H_0 if $X < 5$, and if $X = 5$, rejects H_0 if an auxiliary fair coin toss results in heads and accepts H_0 if it results in tails. For this test, the associated test function is

$$\begin{aligned} \phi(x) &= 1, \quad \text{if} \quad x > 5 \\ &= 0, \quad \text{if} \quad x < 5 \\ &= 0.5, \quad \text{if} \quad x = 5. \end{aligned}$$

In this example, the test function takes not only the boundary values $0, 1$, but also a value in the interior of the interval $[0, 1]$; this is a genuinely randomized test. The nice thing about defining all tests through the concept of a test function is that we can handle nonrandomized and randomized tests within the same mathematical structure; nonrandomized tests are simply those which have as test functions an indicator function on their critical region C.

Within this all encompassing framework of test functions, type I error probabilities and the power function of a general test have the following representations:

$$\alpha(\theta,\phi) = P_\theta(H_0 \text{ will be rejected}) = E_\theta P_\theta(H_0 \text{ will be rejected}|X_1,\cdots,X_n)$$
$$= E_\theta[\phi(X_1,\cdots,X_n)], \theta \in \Theta_0;$$
$$\gamma(\theta,\phi) = P_\theta(H_0 \text{ will be rejected}) = E_\theta P_\theta(H_0 \text{ will be rejected}|X_1,\cdots,X_n)$$
$$= E_\theta[\phi(X_1,\cdots,X_n)], \theta \in \Theta_1.$$

Note the important point that this shows that type I error and power calculation is always the same mathematical calculation. For both, you calculate the expected value of the test function. For the type I error, calculate $E(\phi)$ under the null, while for the power, calculate $E(\phi)$ under the alternative. This is to be remembered.
Here is another example to illustrate this sort of a calculation.

Example 8.8. (Power Calculation for Randomized Tests). Let $X \sim Bin(10,p)$ and let $H_0 : p = .5, H_1 : p > .5$. Consider the randomized test that rejects with probability one if $X = 8, 9, 10$, and rejects with probability .5 if $X = 7$. We are going to compute its type I error probability and power function. The type I error probability is

$$\alpha = E_{p=.5}[\Phi(X)] = 1 \times P_{p=.5}(8 \le X \le 10) + .5 \times P_{p=.5}(X = 7)$$
$$= \sum_{x=8}^{10} \frac{\binom{10}{x}}{2^{10}} + .5\,\frac{\binom{10}{7}}{2^{10}} = .0547 + .5\,(.1172) = .1133.$$

The power function is

$$\gamma(p) = E_p[\Phi(X)] = 1 \times P_p(8 \le X \le 10) + .5 \times P_p(X = 7)$$
$$= \sum_{x=8}^{10} \binom{10}{x} p^x (1-p)^{10-x} + .5 \binom{10}{7} p^7 (1-p)^3, p > .5.$$

For example, at $p = .6$, the power is

$$\sum_{x=8}^{10} \binom{10}{x} .6^x .4^{10-x} + .5 \binom{10}{7} .6^7 .4^3 = .2748;$$

if we take an alternative value further away from the null value, say $p = .8$, the power is

$$\sum_{x=8}^{10} \binom{10}{x} .8^x .2^{10-x} + .5 \binom{10}{7} .8^7 .2^3 = .7785,$$

which is larger than the power at $p = .6$. If we take $p = .99$, the power is

$$\sum_{x=8}^{10} \binom{10}{x} .(.99)^x (.01)^{10-x} + .5 \binom{10}{7} (.99)^7 (.01)^3 = .9999.$$

The limit of the power as $p \to 1$ is one.

8.2 Most Powerful Tests and Neyman-Pearson Lemma

Associated with every test of a null hypothesis against an alternative, there is a type I error probability and a power. Suppose that we decide to use a particular test ϕ_1. Suppose that on calculation, we find that the type I error probability of ϕ_1 is 0.04 and its power against the particular alternative is .75. We find this power to be too low. Can we do something about the low power? Yes, we can find another test, say ϕ_2, with a power higher than .75, say .90; but typically, this larger power will come at the expense of a larger type I error probability. That is, ϕ_2 will have a larger type I error probability than ϕ_1. We cannot simultaneously achieve a very low type I error probability and a very large power, without the flexibility to increase the sample size n. For fixed n, looking for a test with a larger power (which is the same as saying a test with a smaller type II error Probability) will usually cause us to be ready for a larger type I error probability.

In view of this, in their classic 1933 paper, Jerzy Neyman and Egon Pearson suggested that we hold one of the two error probabilities at a prespecified level that we are willing to tolerate, and then minimize the other error probability. Specifically, Neyman and Pearson (1933) suggested that we look at those tests which have a type I error probability bounded above by some prechosen number $\alpha, 0 < \alpha < 1$, and look for a test that has the largest power among only these tests. For example, if the prechosen $\alpha = .05$, then we only look for a test with largest power among tests with type I error probability $\leq .05$. By choosing the bound on type I error probability to be .05, we are saying that we are willing to use a test that rejects the null once in twenty times if the null happened to be true; we can tolerate a type I error once in twenty surveys, but no more. With this threshold on the type I error rate, how large can we make the power?

Neyman and Pearson proved the remarkable result that if the null and the alternative are both simple hypotheses, then such a test with the largest possible power subject to a bound on the type I error probability always exists. Furthermore, they also gave a highly intutive explicit description of what this test with the largest power would look like. The test may be a randomized test in some problems, but the result of Neyman and Pearson is otherwise completely general. They solved the simple vs. simple problem in their formulation in complete generality by one unified result. This milestone is popularly known as *the Neyman-Pearson lemma*; the restriction to simple vs. simple problems is relaxable under additional structure in the problem, and will be treated in a later section. We first need a definition.

Definition 8.3. Let $X \sim P$, where P has density (pmf) f which equals either f_0 or f_1. Let $H_0 : f = f_0$ and $H_1 : f = f_1$ be a simple null and a simple alternative hypothesis. Fix $\alpha, 0 < \alpha < 1$. A (possibly randomized) test ϕ_0 is called most powerful of level α (MP level

α) if

$$(a) E_{f_0}[\phi_0(X)] \leq \alpha;$$

$$(b) E_{f_1}[\phi_0(X)] \geq E_{f_1}[\phi(X)] \text{ for any other test } \phi \text{ such that } E_{f_0}[\phi(X)] \leq \alpha.$$

Here is the part of the Neyman-Pearson lemma that is of the greatest use to us.

Theorem 8.1. (Neyman-Pearson Lemma)
Define the likelihood ratio

$$\Lambda(x) = \frac{f_1(x)}{f_0(x)}.$$

(a) Suppose a possibly randomized test $\phi_0(X)$ is of the following form:

$$\phi_0(x) = \begin{cases} 1, & \text{if} \quad \Lambda(x) > k \\ 0, & \text{if} \quad \Lambda(x) < k, \quad (*) \end{cases}$$

and that the constant k has been so chosen that

$$E_{f_0}[\phi_0(X)] = \alpha.$$

Then ϕ_0 is an MP level α test.
(b) Any MP level α test is of the form $(*)$ for some k.
Proof: Part (a) is a practical prescription for explicitly finding an MP level α test; we prove part (a) here.
Let ϕ be any other level α test; we will show that ϕ_0 has a power at least as large as that of ϕ.
Indeed,

$$E_{f_1}(\phi_1) - E_{f_1}(\phi_0) = \int [\phi_1 - \phi_0] f_1(x) dx$$

$$= \int [\phi_1 - \phi_0] \Lambda(x) f_0(x) dx = \int [\phi_1 - \phi_0][\Lambda(x) - k + k] f_0(x) dx$$

$$= k \int [\phi_1 - \phi_0] f_0(x) dx + \int [\phi_1 - \phi_0][\Lambda(x) - k] f_0(x) dx$$

Now,

$$\int [\phi_1 - \phi_0] f_0(x) dx = E_{f_0}(\phi_1) - E_{f_0}(\phi_0) \leq \alpha - \alpha = 0$$

$$\Rightarrow k \int [\phi_1 - \phi_0] f_0(x) dx \leq 0.$$

Next,

$$\int [\phi_1 - \phi_0][\Lambda(x) - k] f_0(x) dx$$

$$= \int_{\Lambda(x)>k} [\phi_1 - \phi_0][\Lambda(x) - k] f_0(x) dx + \int_{\Lambda(x)<k} [\phi_1 - \phi_0][\Lambda(x) - k] f_0(x) dx + \int_{\Lambda(x)=k} [\phi_1 - \phi_0][\Lambda(x) - k] f_0(x) dx.$$

On the set $\{\Lambda(x) > k\}$,

$$[\phi_1 - \phi_0][\Lambda(x) - k] \leq 0,$$

because by construction, $\phi_0 = 1$ and $\phi_1 \leq 1$ on $\{\Lambda(x) > k\}$.
On the set $\{\Lambda(x) < k\}$, also,

$$[\phi_1 - \phi_0][\Lambda(x) - k] \leq 0,$$

because by construction, $\phi_0 = 0$ and $\phi_1 \geq 0$ on $\{\Lambda(x) < k\}$.
And, on the set $\{\Lambda(x) = k\}$,

$$[\phi_1 - \phi_0][\Lambda(x) - k] = 0.$$

Putting these three cases together,

$$\int [\phi_1 - \phi_0][\Lambda(x) - k] f_0(x) dx \leq 0.$$

Therefore,

$$\int [\phi_1 - \phi_0] f_1(x) dx = k \int [\phi_1 - \phi_0] f_0(x) dx + \int [\phi_1 - \phi_0][\Lambda(x) - k] f_0(x) dx$$
$$\leq 0 + 0 = 0,$$

which proves part (a).
Let us now see a series of examples.

Example 8.9. (Discrete MP Test). This is an example due to Lehmann (1986). Let f_0, f_1, f_2 be three discrete distributions, and suppose we want to test $H_0 : f = f_0$ against respectively $H_1 : f = f_1$, $H_2 : f = f_2$. Our goal is to find the MP level α test for each alternative with $\alpha = .07$. Here are the pmfs f_0, f_1, f_2.

x	1	2	3	4	5	6
$f_0(x)$	.03	.02	.02	.01	0	.92
$f_1(x)$	.06	.05	.08	.02	.01	.78
$f_2(x)$	.09	.05	.12	0	.02	.72
$\frac{f_1(x)}{f_0(x)}$	2	2.5	4	2	∞	.85
$\frac{f_2(x)}{f_0(x)}$	3	2.5	6	0	∞	.78

According to the NP (Neyman-Pearson) lemma, we reject the null for those values of x for which the likelihood ratio is large. For the problem of testing H_0 against H_1, the likelihood ratio $\frac{f_1(x)}{f_0(x)}$ in decreasing order is $\infty, 4, 2.5, 2, 2, .85$, corresponding to $x = 5, 3, 2, 1, 4, 6$. Their probabilities under the null distribution f_0 are, respectively, $0, .02, .02, .03, .01, .92$. The first four values just suffice to give a total probability $0 + .02 + .02 + .03 = .07 = \alpha$.

Hence, the MP test at level $\alpha = .07$ rejects H_0 if $X = 5, 3, 2$, or 1 and accepts H_0 if $X = 4, 6$.

Next, for testing H_0 against H_2, the likelihood ratio $\frac{f_2(x)}{f_0(x)}$ in decreasing order is $\infty, 6, 3, 2.5, .78, 0$ corresponding to $x = 5, 3, 1, 2, 6, 4$. Their probabilities under the null distribution f_0 are, respectively, $0, .02, .03, .02, .92, .01$. Again, the first four values just suffice to give a total probability $0 + .02 + .03 + .02 = .07 = \alpha$. Hence, the MP test at level $\alpha = .07$ rejects H_0 if $X = 5, 3, 1$, or 2, and and accepts H_0 if $X = 4, 6$. Very interestingly, at the end, for each alternative, H_1 or H_2, the MP test at this level works out to be the same test; i.e., the rejection regions are exactly the same.

Example 8.10. (Most Powerful Tests for Normal Mean). Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ is considered known. Let μ_0, μ_1 be two specified numbers, $\mu_0 < \mu_1$, and suppose we want to test $H_0 : \mu = \mu_0$ against $\mu = \mu_1$, a simple null against a simple alternative.

Fix $\alpha, 0 < \alpha < 1$. By the NP (Neyman-Pearson) lemma, the most powerful level α test should reject for large values of the likelihood ratio

$$\Lambda(X_1, \cdots, X_n) = \frac{\frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum(X_i-\mu_1)^2}}{\frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum(X_i-\mu_0)^2}} = e^{\frac{\mu_1-\mu_0}{\sigma^2}\sum X_i} e^{-\frac{n}{2\sigma^2}(\mu_1^2-\mu_0^2)} > k$$

$$\Leftrightarrow e^{\frac{\mu_1-\mu_0}{\sigma^2}\sum X_i} > k^* \Leftrightarrow \frac{\mu_1-\mu_0}{\sigma^2}\sum X_i > \log k^*$$

$$\Leftrightarrow \bar{X} > c;$$

in the above k^*, c are appropriate constants, whose exact expressions are not needed by us. The only purpose of this calculation is to establish the fact that large values of the likelihood ratio correspond to large values of the sample mean in this problem. Hence, by the NP lemma, the MP test will reject H_0 for *sufficiently large values* of $\bar{X}$.

How large? In other words, what is the c that we ultimately want? This is also determined by the NP lemma; choose c to force the type I error of the test to be *exactly equal to* α.

$$\alpha = P_{\mu=\mu_0}(\bar{X} > c) = P_{\mu=\mu_0}(\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} > \frac{\sqrt{n}(c-\mu_0)}{\sigma})$$

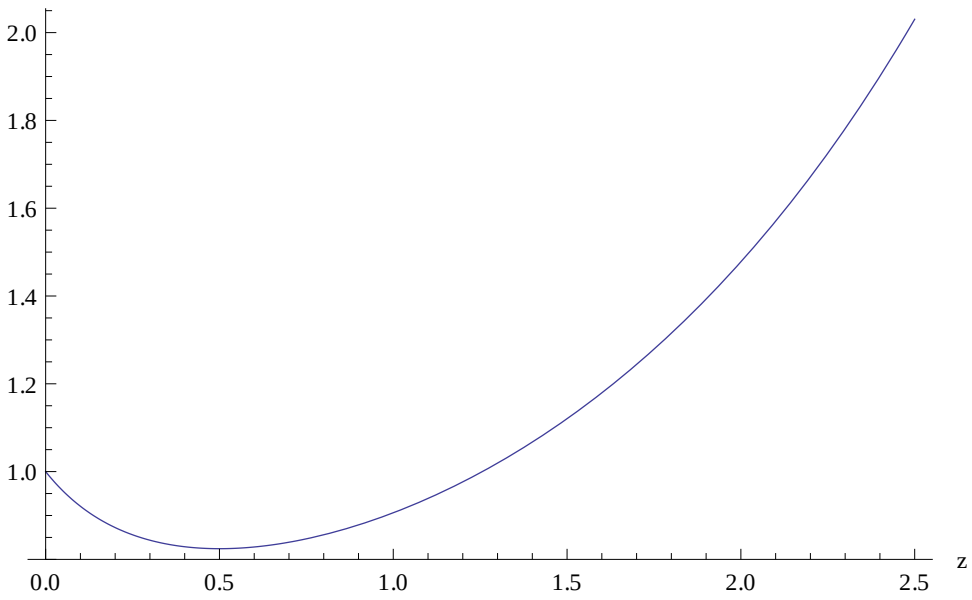
$$= P(Z > \frac{\sqrt{n}(c-\mu_0)}{\sigma}),$$

where $Z \sim N(0, 1)$. This forces

$$\frac{\sqrt{n}(c-\mu_0)}{\sigma} = z_\alpha \Rightarrow c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Therefore, the following test is of the Neyman-Pearson form and has type I error probability exactly equal to α, and hence is MP level α in this problem:

$$\text{Reject the null if } \bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}, \text{ and accept the null otherwise.}$$



Note that this is the z-test; it is a nonrandomized test.

Example 8.11. (MP Test of Normal vs. Cauchy). Suppose we have one observation $X \sim f$ where f is either a standard normal or a standard Cauchy density. We formulate the problem as one of testing

$$H_0 : f \text{ is } N(0,1);\ H_1 : f \text{ is } C(0,1).$$

Suppose we want to find an MP level α test, where $0 < \alpha < 1$. The final conclusion of this example is interesting. Let's proceed with deriving the MP level α test.
According to the NP lemma, any MP level α test will reject H_0 for large values of the likelihood ratio, which is

$$\Lambda(x) = \frac{\frac{1}{\pi}\frac{1}{1+x^2}}{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}} = c\frac{e^{x^2/2}}{1+x^2} = c\frac{e^z}{1+2z},$$

writing z for $\frac{x^2}{2}$.

We want to know for which values of z, $\frac{e^z}{1+2z}$ is larger than a threshold k. Here is where the problem gets unusual. A plot of the function $\frac{e^z}{1+2z}$ shows that depending on the value of k, the set of values of z for which $\frac{e^z}{1+2z} > k$ is either an one sided interval of the form $z > b$ or a union of two disjoint intervals $z < a, z > b$. It is a union of two disjoint intervals if the threshold $k < 1$. Hence, rejection regions of the MP test can take one of two different forms:

$$|X| > C, \text{ or } |X| < A \cup |X| > B.$$

In fact, if C is the unique positive root of the equation $\frac{e^z}{1+2z} = 1$, $(C = 1.2564)$, then

$$P_{H_0}(|X| > C) = .21.$$

For $\alpha \leq .21$, the MP level α test rejects H_0 when $|X| > z_{\alpha/2}$. But, for $\alpha > .21$, the MP level α test rejects H_0 when $|X| < A$ or $|X| > B$, where A, B depend on α, and have to be found numerically.

The intuition for the MP test to reject H_0 for both X near zero and X far away from zero when α is large is the following. Using a large α is tantamount to asking for a large power. That is, we are looking for a test which will know that the true f is not normal, but Cauchy, if it really was Cauchy. Now, the standard Cauchy has not only a heavier tail, but also a heavier peak than the standard normal. So, if X is either near zero or far away from zero, the test will see signature of the Cauchy, and reject H_0.

If $\alpha = .05$, the MP level α test rejects H_0 for $|X| > 1.96$. The power of this test is

$$\gamma = P_{C(0,1)}(|X| > 1.96) = .3003,$$

which is very small. The low power is caused partially by the sample size being only one.

Example 8.12. (Randomized MP Test for Binomial). Suppose $X \sim Bin(n, p)$ and we wish to test $H_0 : p = p_0$ against $H_1 : p = p_1$, where $p_1 > p_0$. The likelihood ratio is

$$\Lambda(x) = \frac{p_1^x(1-p_1)^{n-x}}{p_0^x(1-p_0)^{n-x}} = \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^x \left(\frac{1-p_1}{1-p_0}\right)^n > k$$

$$\Leftrightarrow \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^x > k^* \Leftrightarrow x > C,$$

since $p_1(1-p_0) > p_0(1-p_1)$ if $p_1 > p_0$.

Thus, according to the NP lemma, a test of the form

$$\phi(X) = \begin{cases} 1, & \text{if} \quad X > C \\ 0, & \text{if} \quad X < C \\ \gamma, & \text{if} \quad X = C \end{cases}$$

will be MP provided its type I error probability is exactly equal to the specified level α;

$$P_{p=p_0}(X > C) + \gamma P_{p=p_0}(X = C) = \alpha.$$

The cutoff C is the smallest integer such that $P_{p=p_0}(X > C) \leq \alpha$. Having found C, γ is found to solve the equation

$$P_{p=p_0}(X > C) + \gamma P_{p=p_0}(X = C) = \alpha \Leftrightarrow \gamma = \frac{\alpha - P_{p=p_0}(X > C)}{P_{p=p_0}(X = C)}.$$

For example, if $n = 20, p_0 = .5, \alpha = .05$, then by using a binomial table,

$$C = 14, \gamma = \frac{.05 - .0207}{.0370} = .7919.$$

So, finally, the MP test at level .05 is a randomized test; it rejects $H_0 : p = .5$ if $X \geq 15$ and rejects H_0 with probability .7919 if $X = 14$. Without the randomization, at this α level, we cannot have an MP test. If we change the α level to $\alpha = .0207$, the nonrandomized test which rejects H_0 for $X \geq 15$ and accepts H_0 otherwise will be MP. Thus, the necessity for randomization arises in the discrete cases for usual values of α, *if we insist on using an MP test*. In practice, users give up the insistence on using an MP test and use a nonrandomized test.

Example 8.13. (MP Test in Uniform). Let $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta]$, and suppose we wish to test $H_0 : \theta = 1$ against $H_1 : \theta = \theta_1 (> 1)$. In this case, the likelihood function equals

$$\Lambda(X_1, \cdots, X_n) = \frac{\frac{1}{\theta_1^n} I_{X_{(n)} \leq \theta_1}}{I_{X_{(n)} \leq 1}}.$$

This simplifies to

$$\Lambda(X_1, \cdots, X_n) = \begin{cases} \theta_1^{-n}, & \text{if} \quad X_{(n)} \leq 1 \\ \infty, & \text{if} \quad 1 < X_{(n)} \leq \theta_1 \end{cases}$$

Now choose $k = \theta_1^{-n}$, and consider the test

$$\phi(X_1, \cdots, X_n) = \begin{cases} 1, & \text{if} \quad \Lambda > k \\ 0, & \text{if} \quad \Lambda < k \\ \alpha, & \text{if} \quad \Lambda = k \end{cases}$$

Then, the type I error probability of this test is

$$1 \times P_{\theta=1}(1 < X_{(n)} \leq \theta_1) + \alpha \times P_{\theta=1}(X_{(n)} \leq 1) = 0 + \alpha = \alpha.$$

Therefore, by the NP lemma, this is an MP level α test for this problem. Interestingly, in this example, it can be shown that MP level α tests are not unique.

8.3 UMP Tests and Monotone Likelihood Ratio

The Neyman-Pearson lemma explicitly describes how to find an MP level α test for any simple null against any simple alternative at any specified level α. However, in practical problems, the alternative hypothesis is usually not simple; users have a class of potential alternatives in mind. But, as the alternative changes, the MP test could also change with the alternative. So, it would be satisfactory if for a certain class of alternatives, a single common test worked out to be MP at a given level simultaneously for all those alternatives.

If such a test exists, it is called *uniformly most powerful* (UMP) level α with respect to the class of alternatives provided.
It is not very common to have UMP tests to exist. You must have enough structure on the underlying distribution, as well as the type of the class of alternatives and the class of nulls for UMP tests to exist. The purpose of this section is to present essentially the only general class of hypothesis testing problems for which UMP tests exist. First we need a formal definition.

Definition 8.4. Let $X \sim P \in \mathcal{P}$. Let $\mathcal{P}_0, \mathcal{P}_1$ be two subclasses of the family of distributions $\mathcal{P}$. Consider testing $H_0 : P \in \mathcal{P}_0$ against $H_1 : P \in \mathcal{P}_1$. A test ϕ_0 is said to be UMP level α for testing H_0 against H_1 if

$$(a) E_P(\phi_0) \leq \alpha \, \forall P \in \mathcal{P}_0;$$

$$(b) E_P(\phi_0) \geq E_P(\phi) \, \forall P \in \mathcal{P}_1, \text{ and any } \phi \text{ such that } E_P(\phi) \leq \alpha \, \forall P \in \mathcal{P}_0.$$

To indicate the kinds of problems in which UMP tests do exist at any level α, we return to a previous important example.

Example 8.14. (UMP Test for Normal Mean). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. In Example 8.10, we derived the z-test as the MP level α test for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ (where $\mu_1 > \mu_0$). The z-test is nonrandomized and rejects the above H_0 if and only if $\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$.
Notice that the rejection region of this test depends on μ_0, α, σ and n, *but not on the specific alternative value* $\mu_1 > \mu_0$. Therefore, the argument in Example 8.10 could be repeated verbatim for $H_0 : \mu = \mu_0$ against $H_1^* : \mu = \mu_1^*$ for another $\mu_1^* > \mu_0$, *and we will end up with exactly the same MP test.* This immediately implies that for the simple null vs. composite alternative

$$H_0 : \mu = \mu_0; \; H_1 : \mu > \mu_0,$$

the z-test which rejects the above simple H_0 for $\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$, *is UMP level* α.
There is more to this. Suppose now that we expand even the null to the composite null $H_0^* : \mu \leq \mu_0$. Our claim is that the same z-test is UMP level α for the composite null $\mu \leq \mu_0$ against the composite alternative $\mu > \mu_0$. To make the proof clear, we need a little notation. Let

$$\mathcal{C}_\alpha = \{\phi : E_{\mu_0}(\phi) \leq \alpha\}; \; \mathcal{C}_\alpha^* = \{\phi : E_\mu(\phi) \leq \alpha \, \forall \mu \leq \mu_0\}.$$

What we so far know is that the z-test is UMP among the tests in the class $\mathcal{C}_\alpha$. Our claim is that it is UMP among the tests in the class $\mathcal{C}_\alpha^*$.
Notice that $\mathcal{C}_\alpha^* \subseteq \mathcal{C}_\alpha$. So, it might at first seem that it is trivially true that the z-test is UMP in the class $\mathcal{C}_\alpha^*$, because $\mathcal{C}_\alpha^*$ is a smaller class of competing tests than $\mathcal{C}_\alpha$. The reason

that it is not trivially true is that we only know so far that the z-test belongs to the larger class $\mathcal{C}_\alpha$; we do not yet know that it belongs to the smaller class $\mathcal{C}_\alpha^*$. If it didn't belong to $\mathcal{C}_\alpha^*$, it couldn't be the best test in $\mathcal{C}_\alpha^*$. So, this does require a proof!

This is easy to prove. Take $\mu < \mu_0$. Let ϕ_0 be the test function corresponding to the z-test. Then,

$$E_\mu(\phi_0) = P_\mu(\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}) = P_\mu(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > \frac{\sqrt{n}(\mu_0 - \mu + z_\alpha \frac{\sigma}{\sqrt{n}})}{\sigma})$$

$$= P_\mu(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} + z_\alpha) = \Phi(\frac{\sqrt{n}(\mu - \mu_0)}{\sigma} - z_\alpha),$$

since $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$. But, by choice $\mu - \mu_0 < 0$, and therefore,

$$\Phi(\frac{\sqrt{n}(\mu - \mu_0)}{\sigma} - z_\alpha) < \Phi(-z_\alpha) = \alpha.$$

This does prove that the z-test belongs to the smaller class $\mathcal{C}_\alpha^*$, and completes the proof of the following final result:

The z-test of Example 8.10 is UMP level α for $H_0 : \mu \le \mu_0,\ H_1 : \mu > \mu_0$.

There is a generalization of this normal case example to the entire one parameter regular Exponential family. In fact, the result generalizes to any family having a suitable monotonicity property called *monotone likelihood ratio* (MLR), and this monotone likelihood ratio property holds in particular for the one parameter Exponential family, and also for some distributions *not* within the Exponential family. Here is the definition of monotone likelihood ratio.

Definition 8.5. Let given $\theta \in \Theta \subseteq \mathcal{R}, X^{(n)} = (X_1, \cdots, X_n) \sim f(x_1, \cdots, x_n \,|\theta)$, a density or a pmf. The family $\{f(x_1, \cdots, x_n \,|\theta), \theta \in \Theta\}$ is said to be MLR in the statistic $T(X_1, \cdots, X_n)$ if for any pair $\theta_2, \theta_1 \in \Theta, \theta_2 > \theta_1$, the likelihood ratio

$$\Lambda(x_1, \cdots, x_n) = \frac{f(x_1, \cdots, x_n \,|\theta_2)}{f(x_1, \cdots, x_n \,|\theta_1)}$$

is monotone nondecreasing in $T(x_1, \cdots, x_n)$. That is, if two data vectors $(x_1, \cdots, x_n)$ and $(y_1, \cdots, y_n)$ are such that $T(y_1, \cdots, y_n) \ge T(x_1, \cdots, x_n)$, then $\Lambda(y_1, \cdots, y_n) \ge \Lambda(x_1, \cdots, x_n)$.

As remarked above, the one parameter Exponential family has the MLR property. Here is the exact result.

Theorem 8.2. Let $X_1, \cdots, X_n \overset{iid}{\sim} f(x \,|\eta) = e^{\eta T(x) - \psi(\eta)} h(x)$. Then the joint density (pmf) $f(x_1, \cdots, x_n \,|\eta)$ of $(X_1, \cdots, X_n)$ is MLR in the minimal sufficient statistic $\bar{T} = \frac{1}{n}\sum_{i=1}^n T(X_i)$.

Proof Take $\eta_1, \eta_2, \eta_2 > \eta_1$. The likelihood ratio is

$$\Lambda(x_1, \cdots, x_n) = \frac{e^{\eta_2 \sum_{i=1}^n T(X_i) - n\psi(\eta_2)}}{e^{\eta_1 \sum_{i=1}^n T(X_i) - n\psi(\eta_1)}}$$

$$e^{n[\psi(\eta_1)-\psi(\eta_2)]} \times e^{(\eta_2-\eta_1)\sum_{i=1}^n T(X_i)},$$

which is increasing in $\sum_{i=1}^n T(X_i)$, and hence increasing in $\bar{T}$ (it is increasing because $\eta_2 - \eta_1 > 0$).

Outside of the one parameter Exponential family, the MLR property is rare, although it does hold in some nonregular cases. Here is one such example.

Example 8.15. (MLR Property in Uniform). Let $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta]$. For $\theta_2 > \theta_1 > 0$, the likelihood ratio is

$$\Lambda(X_1, \cdots, X_n) = \begin{cases} \frac{\theta_1^n}{\theta_2^n}, & if \quad X_{(n)} \le \theta_1 \\ \infty \quad , if & \theta_1 < X_{(n)} \le \theta_2 \end{cases}$$

Thus, $\Lambda(X_1, \cdots, X_n)$ is monotone nondecreasing in $X_{(n)}$, and hence the $U[0, \theta]$ family is MLR in $X_{(n)}$. This example generalizes to more general nonregular families.

Some standard regular distributions which are not in the Exponential family are the location parameter Double exponential, Cauchy, and t, and none of these is MLR in any one dimensional statistic $T(X_1, \cdots, X_n)$.

Here is how the MLR property links to existence and explicit description of UMP tests for one sided composite nulls tested against one sided composite alternatives.

Theorem 8.3. Let $(X_1, \cdots, X_n) \sim f(x_1, \cdots, x_n \,|\theta)$. Suppose the family $f(x_1, \cdots, x_n \,|\theta)$ is MLR in some statistic $T(X_1, \cdots, X_n)$. Then, for testing $H_0 : g(\theta) \le c$ against $H_1 : g(\theta) > c$, where $g(\theta)$ is a nondecreasing function of θ, at any given level α,

(a) There is a UMP test.

(b) The MP level α test for the simple null $K_0 : g(\theta) = c$ against a simple alternative $K_1 : g(\theta) = d (d > c)$ is UMP level α for H_0 against H_1; here, the choice of the number $d > c$ does not matter.

Remark: In words, this theorem is saying that we can replace the composite null by a simple null at the boundary value $g(\theta) = c$ and replace the composite alternative by any simple alternative on the side of $g(\theta) > c$, and work out the MP test for this easy problem by using the Neyman-Pearson lemma. That test will also be UMP in our original composite null against the composite alternative problem. Thus, the MLR property enables us to do a marvelous mathematical reduction of the composite- composite testing problem to a simple-simple testing problem. A corollary of this theorem is the following.

Corollary 8.1. Consider an iid sample $X_1, \cdots, X_n$ from any density or pmf in the one parameter Exponential family $f(x \,|\eta) = e^{\eta T(x) - \psi(\eta)} h(x)$. Let $\mu = \mu(\eta) = E_\eta(T(X))$. Let $\bar{T} = \frac{1}{n}\sum_{i=1}^n T(X_i)$. For testing $H_0 : \mu \le \mu_0$ against $H_1 : \mu > \mu_0$, at any level α there is a

UMP test of the following form:

$$\phi_0(X_1, \cdots, X_n) = \begin{cases} 1, & if \quad \bar{T} > C \\ 0, & if \quad \bar{T} < C \\ \gamma, & if \quad \bar{T} = C \end{cases}$$

Here C, γ satisfy

$$P_{\mu_0}(\bar{T} > C) + \gamma P_{\mu_0}(\bar{T} = C) = \alpha.$$

Example 8.16. (UMP Test in Binomial). Let $X \sim Bin(n, p)$. Suppose $n = 20$ and we want to test $H_0 : p \leq .5$ against $H_1 : p > .5$ at the level $\alpha = .05$. In Example 8.12, we worked the MP test at level $\alpha = .05$ for testing the simple null $p = .5$ against the simple alternative $p = p_1(p_1 > .5)$. By virtue of Corollary 8.1, this same test is UMP for $H_0 : p \leq .5$ against $H_1 : p > .5$. That is, by using Example 8.12, the UMP test at level $\alpha = .05$ for this composite null against the composite alternative rejects H_0 for $X \geq 15$, rejects H_0 with probability .7919 if $X = 14$ and accepts H_0 if $X \leq 13$.

8.3.1 Nonexistence of UMP Tests

Unfortunately, throughout the Exponential family, UMP tests do not exist for testing $H_0 : \mu = \mu_0$ (a point null) against $H_1 : \mu \neq \mu_0$ (a two-sided alternative). Basically, for each one-sided alternative, $\mu < \mu_0$ and $\mu > \mu_0$, there is a unique UMP test, and those two tests are different. There cannot be a UMP test when the two one-sided alternatives are combined.

UMP tests also do not exist in Exponential family situations, *even for one sided alternatives*, if the family is multiparameter. Let us see an example. Suppose $X_1 \sim Poi(\lambda_1), X_2 \sim Poi(\lambda_2)$, and we want to test $H_0 : \lambda_1 \leq \lambda_2$ against $H_1 : \lambda_1 > \lambda_2$. Although the null and the alternative are both one sided, there are no UMP tests for this problem. Here is another example. Suppose $X \sim N(\mu, \sigma^2)$ and μ, σ are both unknown. Suppose we want to test $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$. The null and the alternative are both one sided; but there are no UMP tests for this problem.

In important multiparameter problems (problems that have nuisance parameters present) such as these, where no UMP tests in the class of all tests exist, one must suitably restrict consideration to smaller classes of tests. Restrictions are made in various ways, depending on the problem at hand. Standard reductions make use of restrictions to properties called invariance, or similarity, or unbiasedness, etc. In such restricted classes of tests, an UMP test often exists. The mathematical steps involved in deriving such UMP tests in restricted classes are quite sophisticated, and do not belong in the first course on inference. Standard references for these are Ferguson (1967) and Lehmann (1986).

Another approach to these problems is to give up the criterion of UMP altogether, and

use instead a general automatic recipe to produce a test, called the *likelihood ratio method*. Likelihood ratio tests (LRT) are extremely popular, and will be treated later in this chapter.

8.3.2 Sample Size Calculation

Power functions of UMP tests (or suitable other optimal tests) are often used to find the minimum sample size necessary to be able to detect a practically meaningful effect with a sufficiently high probability. For example, suppose we are testing that the mean μ of a normal distribution is zero against the alternative that it is greater than zero (a positive effect exists). Suppose that subject matter experts tell us that it'd be important to detect it if the effect was as large as $\mu = 1$. We can never be sure to detect an effect by using statistical sampling; but we can ask how large a sample would be necessary to detect such a practically important effect with a high probability, say 90% probability? Then, we go ahead and collect at least that many sample observations. Such a calculation is called a *sample size calculation*, and is quite common in some applied fields.

Example 8.17. (Sample Size Calculation). Let $X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2), \sigma$ being considered known. We wish to test $H_0 : \mu = 0$ against $H_1 : \mu > 0$ at level α; right now, we will keep α general. We wish to answer the following question: what is the smallest n for which the power of the UMP level α test at μ_1 is some specified number $1 - \epsilon$? Here, $\mu_1 > 0$ is an effect value that is considered practically important.
Recall that the UMP test is just the z-test which rejects H_0 when $\bar{X} > z_\alpha \frac{\sigma}{\sqrt{n}}$. Its power at any given μ_1 is

$$\gamma(\mu_1) = P_{\mu_1}(\bar{X} > z_\alpha \frac{\sigma}{\sqrt{n}}) = P_{\mu_1}(\frac{\sqrt{n}(\bar{X} - \mu_1)}{\sigma} > (\frac{\sqrt{n}(z_\alpha \frac{\sigma}{\sqrt{n}} - \mu_1)}{\sigma})$$

$$= P(\frac{\sqrt{n}(\bar{X} - \mu_1)}{\sigma} > z_\alpha - \frac{\sqrt{n}\mu_1}{\sigma}) = \Phi(\frac{\sqrt{n}\mu_1}{\sigma} - z_\alpha)$$

$$\geq 1 - \epsilon \Leftrightarrow \frac{\sqrt{n}\mu_1}{\sigma} - z_\alpha \geq z_\epsilon$$

$$\Leftrightarrow \sqrt{n} \geq \frac{\sigma(z_\alpha + z_\epsilon)}{\mu_1} \Leftrightarrow n \geq \left(\frac{\sigma(z_\alpha + z_\epsilon)}{\mu_1}\right)^2.$$

For example, if $\sigma = 1, \alpha = .05, \mu_1 = .5$ and $\epsilon = .1$, then,

$$n \geq \left(\frac{(1.645 + 1.28)}{.5}\right)^2 = 34.22.$$

So to detect an effect of magnituude $\mu_1 = .5$ with a 90% probability, at least 35 observations should be obtained. If we change ϵ to .01, this changes to $n \geq 64$. If we want to detect an effect as small as $\mu_1 = .1$ with a 99% probability (so that again, $\epsilon = .01$), the necessary n increases to 1581.

8.4 Likelihood Ratio Tests: A General Principle

As we have seen in the previous sections, it is quite rare for UMP tests at a given level to exist. UMP tests exist for special types of null and alternative hypotheses in special types of distributions. The question naturally arises if we can think of a general method to find tests that will apply to many problems, will be intuitively reasonable, and will have verifiable good properties. Fortunately, we do not have to search far for just such a general principle. The likelihood function takes us there. The tests we now discuss are called *likelihood ratio tests.*

The likelihood ratio test is a general omnibus test applicable, in principle, in most finite dimensional parametric problems. Thus, let $X^{(n)} = (X_1, \cdots, X_n)$ be the observed data with joint distribution $P_\theta = P_{\theta,n}$, $\theta \in \Theta$, and density (pmf) $f_\theta(x^{(n)})$. For testing

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta - \Theta_0,$$

the likelihood ratio test (LRT) rejects H_0 for small values of

$$\Lambda_n = \frac{\sup_{\theta \in \Theta_0} f_\theta(x^{(n)})}{\sup_{\theta \in \Theta} f_\theta(x^{(n)})}.$$

The motivation for using Λ_n comes from two sources:

(a) The case where H_0, H_1 are each simple, a most powerful (MP) test is found from Λ_n by the Neyman-Pearson lemma.

(b) The intuitive explanation that for small values of Λ_n we can better match the observed data with some value of θ outside of Θ_0.

Remark: LRTs are useful because they are omnibus tests and because UMP tests for a given sample size n generally do not exist outside of the Exponential family. Moreover, in many fundamental testing problems, the LRT works out to a test that makes common sense. These are the main reasons for the widespread acceptance of LRTs, much like the popularity of MLEs as point estimates. Lehmann (2006) is a recommended overview.

A point of caution is that the LRT is not a universally applicable test. There are important examples where the LRT simply cannot be used, because the null distribution of the LRT test statistic depends on nuisance parameters. Also, the exact distribution of the LRT statistic is very difficult or even impossible to find in many problems. That is the reason that in practical implementation of LRTs, large sample distributions of the LRT statistic become really important. We will see a major theorem on asymptotic behavior of the LRT statistic Λ_n later in Chapter 9.

We start with a series of examples, which illustrate various important aspects of the likelihood ratio method.

8.4.1 The t Test

Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Cconsider testing

$$H_0 : \mu = 0 \text{ vs.} H_1 : \mu \neq 0.$$

Let $\theta = (\mu, \sigma^2)$. Then,

$$\Lambda_n = \frac{\sup_{\theta \in \Theta_0}(1/\sigma^n) \exp\left(-\frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2\right)}{\sup_{\theta \in \Theta}(1/\sigma^n) \exp\left(-\frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2\right)} = \left(\frac{\sum_i (X_i - \bar{X})^2}{\sum_i X_i^2}\right)^{n/2},$$

by a straightforward calculation of mles of θ under H_0 and in the general parameter space. More precisely, under H_0, the MLE of the only parameter σ^2 is $\frac{1}{n}\sum_{i=1}^n X_i^2$, and in general, the MLEs of the two parameters are $\bar{X}$ and $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$; these are then substituted into the numerator and the denominator of Λ_n.
Define now

$$t_n = \frac{\sqrt{n}\bar{X}}{\sqrt{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2}}.$$

Then, Λ_n is a decreasing function of t_n^2 (a proof is provided below), and hence,

$$\Lambda_n < c \Leftrightarrow t_n^2 > b.$$

Since the rejection region of the LRT, by definition, is of the form $\Lambda_n < c$, this means that the LRT for this problem rejects H_0 for large values of t_n^2, or equivalently, for large values of $|t_n|$. That is, the LRT for testing for a normal mean when the variance is unknown is a test that rejects H_0 if $|t_n| > k$ for a suitable k. *This is the famous t-test.* It is a pleasant outcome that a famous test turns out to be the LRT in this important problem. In practice, the test would be used by comparing $|t_n|$ with a t-percentile, which can be read off from a t-table.
To pursue this example a little more, observe that

$$\begin{aligned} t_n^2 &= \frac{n\bar{X}^2}{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2} \\ &= \frac{\sum_i X_i^2 - \sum_i (X_i - \bar{X})^2}{\frac{1}{n-1}\sum_i (X_i - \bar{X})^2} \\ &= \frac{(n-1)\sum_i X_i^2}{\sum_i (X_i - \bar{X})^2} - (n-1) \\ &= (n-1)\Lambda_n^{-2/n} - (n-1). \end{aligned}$$

This gives,

$$\Lambda_n = \left(\frac{n-1}{t_n^2 + n - 1}\right)^{n/2}$$

$$\Rightarrow \quad \log \Lambda_n = \frac{n}{2} \log \frac{n-1}{t_n^2 + n - 1}$$

$$\Rightarrow -2 \log \Lambda_n = n \log \left(1 + \frac{t_n^2}{n-1}\right)$$

$$\approx n \frac{t_n^2}{n-1} \approx t_n^2.$$

When n is large $t_n \approx N(0,1)$ and hence $t_n^2 \approx \chi_1^2$. That $-2 \log \Lambda_n$ is approximately distributed as a chi square in large samples is a much more general phenomenon, and will be presented rigorously in Chapter 9.

The convergence of $-2 \log \Lambda_n$ to a chi-square with an appropriate degrees of freedom is a much more general and famous result of statistics. The result is widely used, because the exact distribution of Λ_n usually cannot be found.

8.4.2 More Examples

Example 8.18. (Equality of Poisson Means). Consider the problem of testing for equality of two Poisson means. Thus let $X_1, \cdots, X_m \overset{iid}{\sim} Poi(\mu_1)$ and $Y_1, \cdots, Y_n \overset{iid}{\sim} Poi(\mu_2)$, where we assume that all $m+n$ observations are independent. To derive the LRT, we need the MLEs of the parameters under H_0 and the MLEs in general. Under $H_0, \mu_1 = \mu_2 = \mu$ (say), and the MLE for μ is

$$\frac{\sum_i X_i + \sum_j Y_j}{m+n} = \frac{m\bar{X} + n\bar{Y}}{m+n}.$$

The unrestricted MLEs of μ_1, μ_2 are $\bar{X}, \bar{Y}$. Therefore, on an easy calculation,

$$\Lambda_{m,n} = \frac{\left((m\bar{X} + n\bar{Y})/(m+n)\right)^{m\bar{X}+n\bar{Y}}}{(\bar{X})^{m\bar{X}} (\bar{Y})^{n\bar{Y}}}$$

$$\Rightarrow -\log \Lambda_{m,n} = m\bar{X} \log \bar{X} + n\bar{Y} \log \bar{Y}$$

$$-(m\bar{X} + n\bar{Y}) \log \left(\frac{m}{m+n}\bar{X} + \frac{n}{m+n}\bar{Y}\right).$$

Significant additional simplification is not possible. The LRT statistic (and its logarithm) has a complicated formula, and the exact null distribution for given m, n has no closed form representation. However, once again, it may be proved that if $m, n \to \infty$ and $\frac{m}{m+n} \to \lambda (0 \le \lambda \le 1)$, then $-2 \log \Lambda_{m,n} \approx \chi_1^2$. Hence, in practice, the LRT test would be used by comparing $-2 \log \Lambda_{m,n}$ with a chi-square percentile, which can be read off from a chi-square table.

Example 8.19. (LRT in Multinomial Distribution). Consider a multinomial distribution with parameters $p_1, \cdots, p_r$ and n, where n is assumed to be known. Suppose we wish to test

$$H_0 : p_1 = p_2 = \cdots = p_r \ \text{ vs. } \ H_1 : H_0 \text{ is not true.}$$

A simple example would be one of testing that a given die is a fair die.
Let $n_1, \cdots, n_r$ denote the observed cell frequencies. Then, the unrestricted MLEs of the parameters are $\frac{n_i}{n}, i = 1, 2, \cdots, r$, and under H_0, the parameters are completely specified, namely each $p_i = \frac{1}{r}$. Hence, directly, from its definition, we get

$$\Lambda_n = \Pi_{i=1}^r \left(\frac{n}{rn_i}\right)^{n_i}$$

$$\Rightarrow -\log \Lambda_n = n(\log \frac{r}{n}) + \sum_{i=1}^r n_i \log n_i.$$

The LRT statistic has a messy form, and we cannot find its exact distribution. It can be shown that $-2 \log \Lambda_n \approx \chi^2_{r-1}$; the proof of it must await a general theorem that we will see in Chapter 9.

Example 8.20. (Example of Failure of LRT). Consider the so called Behrens-Fisher problem with

$$\begin{aligned} X_i, i = 1, \cdots, m, &\overset{iid}{\sim} \quad N(\mu_1, \sigma_1^2) \\ Y_j, j = 1, \cdots, n, &\overset{iid}{\sim} \quad N(\mu_2, \sigma_2^2). \end{aligned}$$

As usual, we assume that all $m + n$ observations are independent. We want to test

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2.$$

The unrestricted MLEs are easy:

$$\hat{\mu}_1 = \bar{X}, \ \hat{\mu}_2 = \bar{Y}, \ \hat{\sigma}_1^2 = \frac{1}{m}\sum_i (X_i - \bar{X})^2, \ \hat{\sigma}_2^2 = \frac{1}{n}\sum_j (Y_j - \bar{Y})^2.$$

Next, let $\hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_2^2$ denote the restricted mle of μ, σ_1^2 and σ_2^2 respectively, under H_0, where μ denotes the common value of μ_1 and μ_2 under H_0. The MLEs under H_0 satisfy the equations

$$\hat{\sigma}_1^2 = (\bar{X} - \hat{\mu})^2 + \frac{1}{m}\sum_i (X_i - \bar{X})^2;$$

$$\hat{\sigma}_2^2 = (\bar{Y} - \hat{\mu})^2 + \frac{1}{n}\sum_j (Y_j - \bar{Y})^2;$$

$$0 = \frac{m(\bar{X} - \hat{\mu})}{\hat{\sigma}_1^2} + \frac{n(\bar{Y} - \hat{\mu})}{\hat{\sigma}_2^2}.$$

Exercise 8.2. (Formulation). A manuscript submitted to a journal has 30 pages. As the journal's editor, you want to know if the number of grammatical errors in the manuscript is more than one per page on the average. Formulate this as a testing of hypothesis problem. Use notation and a specific model.

Exercise 8.3. (Plotting the Rejection Region). Suppose X_1, X_2 are two iid observations from $N(\mu, \sigma^2)$. We want to test $H_0 : \mu = 0$ against $\mu \neq 0$ by using the t-test.
(a) Plot the rejection region when $\alpha = .05$.
(b) Plot the rejection region when $\alpha = .01$.
(c) Is one of the two regions a superset of the other? Would you have expected that?

Exercise 8.4. (Error Probabilities). Suppose X_1, X_2, X_3 are iid $N(\mu, 1)$. You want to test $H_0 : \mu = 0$ against $H_1 : \mu > 0$. Consider the following two tests:
1) Reject H_0 if two or more of the observations are larger than 2.
2) Reject H_0 if the average oof the three observations are larger than 2.5.
(a) Find the type I error probability of each test.
(b) Find the power of each test at $\mu = 1, 2, 3$.

Exercise 8.5. (Slightly Unusual). Suppose X, Y are independent and that $X \sim Poi(\lambda), Y \sim Poi(2\lambda)$. We want to test $H_0 : \lambda = 1$ against $H_1 : \lambda = 2$. Find the type I error probability and the power of the test that rejects the null when $X + Y$ is larger than 4 but none of them is equal to zero.

Exercise 8.6. (Slightly Unusual). In coin tossing, a sequence of k or more consecutive heads is called a head-run of length k. A coin with probability p of heads has been tossed six times. We want to test $H_0 : p = .5$ against $H_1 : p > .5$.
Consider the following two tests:
1) Reject H_0 if the number of heads is 5 or more.
2) Reject H_0 if you observe a head-run oof length three.
(a) Find the type I error probability of each test.
(b) Find the power of each test at $p = .6, .9$.

Exercise 8.7. (Paradox of α-level Testing). (a) Suppose that based on one observation $X \sim N(\mu, 1)$, we want to test $H_0 : \mu = 0$ against $H_1 : \mu = 4$ at level $\alpha = .05$. Find the values of X for which the z test rejects $\mu = 0$.
(b) Now reformulate the problem as $H_0 : \mu = 4$ against $H_1 : \mu = 0$, still using level $\alpha = .05$. Find the values of X for which the z test rejects $\mu = 4$.
(c) Identify the values of X for which $\mu = 0$ is rejected if stated as a null and accepted if stated as an alternative.
(d) Generalize part (c) to n observations, general μ_0, μ_1, a general σ^2, and a general α.
(e) Do you consider this a paradox? Comment.

Exercise 8.8. (Differing Shapes of Rejection Regions). Let $X \sim C(\mu, 1)$; we want to test $H_0 : \mu = -1$ against $H_1 : \mu = 1$ based on the single X.
(a) Show that the form of the rejection region of the MP test depends on α.
(b) Explicitly find the MP test when $\alpha = \frac{1}{2}$.

Exercise 8.9. (Conceptual). Are MP tests at a given level unique? Prove it, or give a counterexample.

Exercise 8.10. (MP Tests for Beta).
(a) Let $X \sim Beta(\theta, \theta)$. Derive the MP level α test for $H_0 : \theta = 1$ against $H_1 : \theta = \theta_1 (< 1)$.
(b) Let $X_1, \cdots, X_n \overset{iid}{\sim} Beta(\theta, \theta)$. Derive the form of the MP level α test for $H_0 : \theta = 1$ against $H_1 : \theta = \theta_1 (< 1)$.
(c) Use the central limit theorem suitably to approximately find the MP level α test corresponding to part (b).

Exercise 8.11. (Uniform with Known Width). Let $X_1, \cdots, X_n \overset{iid}{\sim} U[\mu - 1, \mu + 1]$. Find an MP level α test for $H_0 : \mu = 0$ against $H_1 : \mu = 1$. Generalize to $H_1 : \mu = \mu_1 (> 0)$.

Exercise 8.12. (Normal vs. Double Exponential). Let $X \sim f$. We want to test $H_0 : f = N(0, \sigma^2)$ against $H_1 : f = \text{DoubleExp}(0, \tau)$, where σ, τ are specified. Derive the MP level α test.

Exercise 8.13. (Hardy-Weinberg Law). Consider a population with three kinds of individuals labeled as $1, 2, 3$, occuring according to the Hardy-Weinberg proportions, $\theta^2, 2\theta(1-\theta), (1-\theta)^2$. Suppose in a sample of n individuals, the frequencies of the three types are n_1, n_2, n_3.
Derive the form of the MP level α test for $H_0 : \theta = \theta_0$ against against $H_1 : \theta = \theta_1 (> \theta_0)$.

Exercise 8.14. (Conceptual). Suppose $X_1, \cdots, X_n$ are iid nonnegative integer valued random variables. Suggest a test for testing the null hypothesis that the common distribution of the X_i is some Poisson.

Exercise 8.15. (A General Inequality). Suppose an MP level α test for some simple null against a simple alternative is nonrandomized. Show that the constant k of the Neyman-Pearson lemma satisfies the inequalities

$$\frac{\beta}{1-\alpha} \leq k \leq \frac{1-\beta}{\alpha};$$

here β is the type II error probability of the MP test.

Exercise 8.16. (UMP Test). Consider the following three pmfs:

x	$f_0(x)$	$f_1(x)$	$f_2(x)$
0	1/3	0	1/3
1	1/3	1/3	0
2	1/3	2/3	2/3

(a) Let $\alpha = 1/3$. Show that there exists a UMP test of $H_0 : f = f_0$ against $H_1 : f = f_1$ or f_2.
(b) Let $\alpha = 2/3$. Show that there does not exist a UMP est of $H_0 : f = f_0$ against $H_1 : f = f_1$ or f_2.

Exercise 8.17. (UMP Test for Poisson). Let $X \sim Poi(\lambda)$.
(a) Find the MP test at level $\alpha = .05$ for $H_0 : \lambda = 1$ against $\lambda = \lambda_1$ and find its power at $\lambda_1 = 1.5, 2, 3, 3.01$.
(b) Is there a UMP test for $H_0 : \lambda = 1$ against $\lambda > 1$? Prove your claim.

Exercise 8.18. (UMP Test for Poisson). Let $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda)$.
(a) Find the form of the UMP test at level $\alpha = .05$ for $H_0 : \lambda = 1$ against $\lambda > 1$.
(b) Use the central limit theorem to approximately find the UMP test.
(c) Approximately find the power function of the UMP test.
(d) Plot it for $n = 20, 100$.
(e) Comment on what you learned from part (d).

Exercise 8.19. (UMP Test for Hypergeometric). Among $N = 40,000$ students in a university some unknown number D have registered to vote at the next election. Let $p = \frac{D}{N}$. Find the form of a UMP test at a general level for testing $H_0 : p \leq .5$ against $H_1 : p > .5$.

Exercise 8.20. (UMP Test for Normal Variance). Let $X_1, \cdots, X_n \overset{iid}{\sim} N(0, \sigma^2)$. Explicitly find the UMP test for $H_0 : \sigma \leq \sigma_0$ against $H_1 : \sigma > \sigma_0$ at a general level α.

Exercise 8.21. (UMP Test for Double Exponential). Let $X_1, \cdots, X_n \overset{iid}{\sim} \text{doubleExp}(0, \sigma)$. Find the form of the UMP test for $H_0 : \sigma \leq \sigma_0$ against $H_1 : \sigma > \sigma_0$ at a general level α.
(b) Use the central limit theorem suitably to approximately find the UMP test.

Exercise 8.22. (UMP Test for Variance Ratio). Let $X_1, \cdots, X_m \overset{iid}{\sim} N(0, \sigma_1^2)$. and $Y_1, \cdots, Y_n \overset{iid}{\sim} N(0, \sigma_2^2)$. Show that at any level α, there is a UMP test for $H_0 : \sigma_1^2 \leq \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$, and find it as explicitly as you can.

Exercise 8.23. (UMP Test for Gamma Shape Parameter). Let $X_1, \cdots, X_n \overset{iid}{\sim} G(\alpha, \lambda)$, where λ is known. Show that at any level α, there is a UMP test for $H_0 : \alpha \leq \alpha_0$ against $H_1 : \alpha > \alpha_0$, and find it as explicitly as you can.

Exercise 8.24. (Characterization of MLR Property). Let X be a single observation from $f(x\,|\theta) = f_0(x-\theta)$. Show that the family $f(x\,|\theta)$ is MLR in $T(X) = X$ if and only if $\log f_0$ is a concave function.

Exercise 8.25. Let X be a single observation from $C(\theta, 1)$. Show that this family is not MLR in $T(X) = X$.

Exercise 8.26. Generalize the previous exercise's result to a t distribution with a general degree of freedom, location parameter θ and known scale parameter $\sigma = 1$.

Exercise 8.27. (Power of the t-test). Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Show that the power of the t-test for testing $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ is increasing in $\frac{|\mu|}{\sigma}$.

Exercise 8.28. (Sample Size Calculation). Let $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda)$. Consider testing $H_0 : \lambda \leq \lambda_0$ against $H_1 : \lambda > \lambda_0$ and the nonrandomized test that rejects H_0 for large values of the sample mean.
(a) Approximately find the sample size n needed to attain a power of .90 at $\lambda = \lambda_1 (> \lambda_0)$.
(b) Make a table of the n values needed if $\lambda_0 = 1, \alpha = .05, \lambda_1 = 1.1, 2, 4$.

Exercise 8.29. (Smallest Detectable Effect). Let $X_1, \cdots, X_n \overset{iid}{\sim} Exp(\lambda)$. Consider testing $H_0 : \lambda \leq \lambda_0$ against $H_1 : \lambda > \lambda_0$ and the corresponding UMP level α test.
Suppose the sample size is some fixed number n_0. Call an effect value λ_1 detectable if the power of the UMP test at λ_1 is $\geq .9$. Approximately find the smallest detectable effect with a sample of size n_0 λ_1 if $\alpha = .05, .01$. You will need to use the central limit theorem.

Exercise 8.30. (LRT for Binomial).
(a) Derive the LRT statistic when $X \sim Bin(n, p)$ and $H_0 : p = \frac{1}{2}, H_1 : p \neq \frac{1}{2}$.
(b) Prove that the LRT statistic is a nondecreasing function of $|X - \frac{n}{2}|$.

Exercise 8.31. (Two Sample t-test). Suppose $X_1, \cdots, X_m \overset{iid}{\sim} N(\mu_1, \sigma^2), Y_1, \cdots, Y_n \overset{iid}{\sim} N(\mu_2, \sigma^2)$, where $\mu_1, (\mu_2, \sigma$ are unknown. We assume all $m + n$ observations are independent. Derive the LRT for $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_2 > \mu_1; H_2 : \mu_2 \neq \mu_1$.

Exercise 8.32. (Paired t-Test). Suppose $(X_i, Y_i), i = 1, 2, \cdots, n$ is an iid sample from a general bivariate normal distribution with five parameters. Derive the LRT of $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$.
Remark: You will notice that the test corresponds to taking the differences $D_i = X_i - Y_i$ and testing that they come from a normal distribution with mean zero.

Exercise 8.33. (Testing for Standard Normality). Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, both parameters unknown. Derive the LRT for $H_0 : \mu = 0, \sigma = 1$ against $H_1 : H_0$ is not true.

Exercise 8.34. (Testing for Sphericity). Suppose $X_1, X_2, \ldots, X_n$ are iid p-dimensional multivariate normals with mmean vector μ and coovariance matrix Σ. Derive the LRT for $H_0 : \Sigma = c^2 I$ against $H_1 : H_0$ is not true, where c is a specific constant.

Exercise 8.35. (ANOVA). Let $X_{ij} \overset{indep.}{\sim} N(\mu_i, \sigma^2), j = 1, 2, \cdots, n, i = 1, 2, \cdots, k$, all parameters unknown. Derive the LRT for $H_0 : \mu_1 = \cdots = \mu_k$ gainst $H_1 : H_0$ is not true.
Remark: The null distribution will be an F-distribution.

Exercise 8.36. (Unusual Problem). Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$. Derive the LRT for testing the null that μ is an integer against the alternative that it is not an integer.
Remark: The test statistic will involve the distance between the sample mean and the integer closest to the sample mean.

Exercise 8.37. (Interesting Nonregular LRT). Let $X_{ij} \overset{indep.}{\sim} U[0, \theta_i], j = 1, 2, \cdots, n, i = 1, 2$.
(a) Derive the LRT for testing $H_0 : \theta_1 = \theta_2$ against $H_1 : \theta_1 \neq \theta_2$.
(b) Show that you can work out the exact finite sample distribution of $-2 \log \Lambda$.
Remark: You will get a χ^2 distribution.

Exercise 8.38. (Computing P-values). Consider the following data on waiting times at a physician's office (in minutes):

$$20, 33, 25, 15, 40, 28, 22$$

We want to test that the mean waiting time is 25 minutes or less. Compute the P-value corresponding to these data if
(a) we assume normality and use the t-test;
(b) we assume an exponential waiting time and use the test that rejects the null for large values of the sample mean.

Exercise 8.39. (P-value Formula). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta]$. We want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ and the test that rejects the null for large values of the sample maximum.
(a) Derive an analytical formula for the P-value.
(b) Examine its asymptotic behavior under an alternative value θ_1; in particular, what is the expected P-value under an alternative θ_1?

Exercise 8.40. (P-values for Composite Nulls). Show that for a general null $H_0 : \theta \in \Theta_0$, the P-value is stochastically larger than a $U[0, 1]$ variable under any null value; i.e., $P_{\theta_0}(p \leq u) \leq u$, for all $u \in [0, 1]$ and any $\theta_0 \in \Theta_0$.

Exercise 8.41. (Log P-value Plot).
(a) Generate $n = 100$ iid Bernoullis with parameter $p = .6$ and compute the sequence of

P-values $p(X_1, \cdots, X_m), m = 1, 2, \cdots, n$ for $H_0 : p = .5$ and $H_1 : p > .5$. You may use the central limit theorem to find the P-values.
(b) Plot the points $(m, K_m), m = 1, 2, \cdots, n$, where $K_m = -\log p(X_1, \cdots, X_m)$.
(c) What kind of a general shape do you expect to see in this plot? Is that shape seen for your data?

Exercise 8.42. (Shortest Length Confidence Interval). Let $X \sim f(x\,|\mu) = e^{-(x-\mu)}, x \geq \mu$. Consider confidence intervals for μ of the form $X + a \leq \mu \leq X + b$.
(a) Show that the coverage probability iis zero if $a \geq 0$.
(b) If $a < 0$, show that the coverage probability of $X + a \leq \mu \leq X + b$ where $b > 0$ and $X + a \leq \mu \leq X$ are the same.
(c) For $a < b \leq 0$, show that the coverage probability of $X + a \leq \mu \leq X + b$ is $e^b - e^a$.
(d) Show that among these intervals, the shortest length interval corresponds to $a = \log \alpha, b = 0$, where $1 - \alpha$ is the desired coverage probability; i.e., $1 - \alpha = e^b - e^a$.

Exercise 8.43. (Likelihood Ratio Interval in Binomial).
(a) Derive the form of the LRT of $H_0 : p = p_0$ against $H_1 : p \neq p_0$ for $X \sim Bin(n, p)$.
(b) Use the method of inverting a test to hence derive a confiidence interval for p. Compute this interval explicitly for $n = 3$.

Exercise 8.44. (Coverage Probability of t-interval). Compute exactly the coverage probability of the t-confidence interval when the samples are iid $U[-1, 1]$ and the sample size is $n = 2$.

Exercise 8.45. (Coverage Probability of t-interval). Simulate the coverage probability of the t-confidence interval when the samples are iid from a standard lognormal density; use $\alpha = .05$ and $n = 20, 80$. For simulating standard lognormals, first simulate standard normals, and then exponentiate.

Exercise 8.46. (Length of the t-interval).
(a) For iid observations $X_1, \cdots, X_n$ from a general CDF F with variance σ^2, derive a formula for the expected length, say w_n, of the t confidence interval.
(b) Where does $\sqrt{n}w_n$ converge when $n \to \infty$?
(c) Notice that the expected length w_n is unbounded as a function of σ for any n; do you think there is some other interval for which the expected length is bounded as a function of σ?

Exercise 8.47. (A Beautiful Identity). Let $C = C(X_1, \cdots, X_n)$ be a confidence interval for a real parameter θ. Show that w_n, the expected length of C satisfies

$$E_\theta(C) = \int P_\theta(\theta' \in C) d\theta'.$$

That is, the larger the chance of covering false values of the parameter, larger would be the average length.

Exercise 8.48. (Posterior Probability of Null). The combined mass of three varieties of a subatomic particle is believed to be at most 0.28 (in some suitable unit). Experimental measurements that involve error have produced the following data:

$$.24, .27, .26, .28, .23, .25, .25, .26, .25, .23, .25, .27.$$

Suppose a model of $N(\mu, \sigma^2)$ is used.
(a) Take σ to be known as $\sigma = .02$. Calculate the posterior probability of $H_0 : \mu \leq .28$ for each of the following priors for μ:

$$N(.28, \sigma^2);\ N(.28, 2\sigma^2);\ U[.26, .30].$$

(b) Calculate the posterior probability of $H_0 : \mu \leq .28$ by using the joint improper prior (density) $\frac{1}{\sigma^2} d\mu d\sigma$ for (μ, σ).

Exercise 8.49. (Posterior Probability and Bayes Factor). Let $X \sim Bin(50, p)$ and suppose $X = 26$ is observed. We want to test $H_0 : p = .4$ against $H_1 : p > .4$. Let π_0 be the prior probability of H_0, and G the prior distribution on p under the alternative.
(a) Calculate the posterior probability of H_0 for each of the following cases:

$$\pi_0 = .25; G = U[.4, .6];\ \pi_0 = .5; G = U[.4, 1].$$

(b) Calculate the Bayes factor in favor of H_0 for each of these above two cases.
(c) Calculate the mid P-value.
(d) Compare all the results and comment.

Exercise 8.50. (Posterior Probability in Cauchy). Let X be a single observation from $C(\mu, 1)$. Suppose we want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Let the prior probability of H_0 be $\pi_0 > 0$ and the prior on μ under the alternative be G. Evaluate in closed form

$$\lim_{|x| \to \infty} P(H_0 \,|X = x).$$

Do you find the answer interesting? Can you explain the answer intuitively?

Exercise 8.51. (HPD Intervals). Let $X \sim Poi(\lambda)$ and suppose λ has a gamma prior density $ce^{-\lambda/\beta}\lambda^{\alpha-1}$. Suppose the observed value of X is $X = 5$.
(a) Find the 95% Bayes HPD interval for the following cases:

$$\lambda = 1, \alpha = 2;\ \lambda = 1, \alpha = 4.$$

(b) Find the 95% Bayes HPD interval if λ has the improper prior (density) $\frac{1}{\lambda}$.
(c) Compare all the results and comment.

Exercise 8.52. (Pivotal vs. HPD Interval). Suppose $X_1, \cdots, X_n$ are iid observations from a double exponential distribution with mean 0 and scale parameter σ.
(a) Show that $T(X_1, \cdots, X_n, \sigma) = \frac{\sum_{i=1}^{n} |X_i|}{\sigma}$ is a pivot.
(b) Hence find an exact $100(1-\alpha)\%$ confidence interval for σ.
(c) Take $\theta = \frac{1}{\sigma}$ to have a general gamma prior density. Derive an $100(1-\alpha)\%$ Bayes HPD interval for σ (be careful that HPD intervals for θ DO NOT automatically lead to HPD intervals for σ).

Exercise 8.53. HPD Interval in the German Tank Problem). Let $X_1, \cdots, X_n$ be iid discrete uniform on the set $\{1, 2, \cdots, N\}$. Suppose $n = 10$ and $X_{(n)} = 16$. Compute the 95% HPD interval for the following (improper) prior mass functions:

$$\frac{1}{N};\ \frac{1}{\sqrt{N}};\ 1.$$

Exercise 8.54. (HPD Intervals with Different Priors). Suppose $X \sim Bin(n, p)$. Compute the 95% HPD credible interval for p using the $Beta(\alpha, \alpha)$ priorr with $\alpha = \frac{1}{2}, 1, 2, 3, 4$; use $n = 25$ and $X = 3, 6, 9, 12$.

Exercise 8.55. (Conceptual). Consider a three cell multinomial distribution with cell probabilities p_1, p_2, p_3. How would you approach the problem of writing an interval estimate for $|p_1 - p_2|$?

Exercise 8.56. (Confidence Interval for Positive Normal Mean). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, 1), \mu > 0$.
(a) How will you derive a confidence interval for μ? Can you use the z-interval?
(b) How will you derive a Bayesian interval for μ? What priors will you use?

Exercise 8.57. (Benjamini-Hochberg Procedure). Suppose $m = 3, 5, 10$, the group samples are independent, and the group sample sizes are all 20. Simulate the FWER of the Benjamini-Hochberg procedure under the intersection null hypothesis that the means of m normal distributions are all zero; use $\alpha = .05$.

Exercise 8.58. (FWER Property of Holm Procedure). Give a proof that the Holm procedure in Theorem 8.7 dooes satisfy the FWER property in part (b) of the theorem.

Exercise 8.59. (False Discoveries Actually Made). Simulate the actual number of false discoveries made by the Benjamini-Hochberg procedure when $m = 10, 25, 50, 500$ and the nominal FDR control level is $\alpha = .10$. For convenience, you may take all the distributions to be $N(0, 1)$.

Exercise 8.60. (Donoho-Jin Procedure). Take $\beta = .75$, and $r = p_{HC}(\beta)$, to get a point on exactly the Donoho-Jin detection boundary. Simulate the higher criticism statistic by generating data from the mixture model in the Donoho-Jin setup and draw a histogram of the higher criticism values. Use $m = 100, 500, 1000$. Write a report.

Exercise 8.61. (Two Sample t-statistic). Consider independent samples $X_1, \cdots, X_m$ from a CDF F and another inde[endent set $Y_1, \cdots, Y_n$ from a CDF G. Assume that each of F, G has finite means and equal variance, μ_1, μ_2, σ^2. Let

$$s_p^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2},$$

where s_x^2, s_y^2 are the group sample variances. The two sample t-statistic is

$$t = t_{m,n} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{s_p^2(\frac{1}{m} + \frac{1}{n})}}.$$

Show that $t_{m,n} \sim t_{m+n-2}$.

Exercise 8.62. (Behrens-Fisher Problem). For each of the following cases, simulate the random degree of freedom of Welch's test :

$$m = n = 20, \sigma_1 = \sigma_2 = 1;\ m = 20, n = 50, \sigma_1 = 3, \sigma_2 = 1.$$

Exercise 8.63. (Behrens-Fisher Problem). Do a small simulation to find the power of Welch's test when the populations are normal with variances $1, 4$ and the sample sizes are $m = n = 25$; use selected values of the two group means.

8.10 References

Bahadur, R. (1960). Stochastic comparison of tests, Ann. Math. Statist., 31, 276-295.

Bahadur, R. (1967). Rates of convergence of estimates and test statistics, Ann. Math. Statist., 38,303-324.

Basu, D. (1975). Statistical information and likelihood, with discussions, Sankhya, Ser. A, 37, 1-71.

Basu, S. and DasGupta, A. (1995). Robustness of standard confidence intervals for location parameters, Ann. Statist., 23, 1433-1442.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, JRSSB, 57, 289-300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency, Ann. Stat., 29, 1165-1188.

Berger, J. (1986). Statistical Decision Theory and Bayesian Analysis, Springer, New York.

Berger, J. and Delampady, M. (1987). Testing precise hypotheses, with comments and a rejoinder, Statist. Sc., 2, 317-352.

Berger, J. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P-values evidence, JASA, 82, 112-122.

Best, D.J. and Rayner, J.C. (1987). Welch's approximate solution to the Behrens- Fisher

problem, Technometrics, 29, 205-210.
Bickel, P. and Doksum, K. (2001). Mathematical Statistics, Basic Ideas and Selected Topics, Vol.I, Prentice Hall, NJ.
Brown, L., Cai, T., and DasGupta, A. (2001). Interval estimation for a binomial proportion, Statist. Sc., 16, 101-133.
Brown, L., Cai, T., and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions, Ann. Statist., 30, 160-201.
Casella, G. and Berger, R. (1987). Reconciliing Bayesian and frequentist evidence in the one sided testing problem, with comments and a rejoinder, JASA, 82, 106-111.
DasGupta, A. (2008). Asymptotic Theory of Statistics and Probability, Springer, New York.
DasGupta, A. (2010). False vs. missed discoveries, Gaussian decision theory, and the Donsker-Varadhan principle, IMS Collections, 6, 1-21, IMS, Beachwood, Ohio.
Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogenous mixtures, Ann. Stat., 32, 962-994.
Dudoit, S. and van der Laan, M. (2008). Multiple Testing Procedures with Applications to Genomics, Springer, New York.
Dunnett, C.W. and Tamhane, A.C. (1991). Step-down multiple testing for comparing treatments with a control in unbalanced one-way layouts, Stat. in Med., 11, 1057-1063.
Efron, B. (2010). Large Scale Inference: Empirical Bayes Methods, Cambridge Univ. Press, Cambridge.
Ferguson, T. (1967). Mathematical Statistics: A Decision Theoretic Approach, Academic Press, New York.
Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlatted noise, Ann. Stat., 38, 1686-1732.
Hodges, J. and Lehmann, E. (1954). Testing the approximate validity of statistical hypotheses, JRSS, Ser. B, 26, 261-268.
Holm, S. (1979). A simple sequentially selective multiple procedure, Scand. Jour. Stat., 2, 65-70.
Hommel, G. (1988). A stage-wise rejective multiple test procedure based on a modified Bonferroni test, Biometrika, 75, 383-386.
Kendall, M. and Stuart, A. (1977). Advanced Theory of Statistics, Vol 2, Macmillan, New York.
Kim, S. and Cohen, A.S. (1998). On the Behrens-Fisher problem: A review, Jour. Ed. Behavior. Stat., 23, 356-377.
Lee, A. and Gurland, J. (1975). Size and power of tests for equality of means, JASA, 70, 933-941.
Lehmann, E. (1986). Testing Statistical Hypotheses, Wiley, New York.

Lehmann, E. (1993). The Fisher, Neyman-Pearson theories of testing hypothesis: One theory, or two?, JASA, 88, 1242-1249.

Lehmann, E. (2006). On likelihood ratio tests, Lecture Notes and Monographs, 49, 1-9, IMS, Beachwood, Ohio.

Lehmann, E. and Romano, J. (2005). Generalizations of the familywise error rate, Ann. Stat., 33, 1138-1151.

Lindley, D. (1957). A statistical paradox, Biometrika, 44, 187-192.

Linnik, J.V. (1963). On the Behrens-Fisher problem, Bull. Inst. Internat. Statist., 40, 833-841.

Miller, R. (1966). Simultaneous Statistical Inference, McGraw-Hill, New York.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability, Philos. Trans. Royal Soc. London, Ser. A, 236, 333-380.

Neyman, J. and Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses, Philos. Trans. Royal Soc. London, Ser. A, 231, 289-337.

Oh, Hyun-Sook, and DasGupta, A. (1999). Comparison of the P-value and posterior probability, Jour. Stat. Planning Inf., 76, 93-107.

Pfanzagl, J. (1974). On the Behrens-Fisher problem, Biometrika, 61, 39-47.

Rao, C.R. (1973). Linear Statistical Inference and its Applications, Wiley, New York.

Robinson, G. (1976). Properties of Student's-t and the Behrens-Fisher solutions to the two mens problem, Ann. Statist., 4, 963-971.

Sarkar, S. (2002). Some results on false discovery rates in stepwise multiple testing procedures, Ann. Stat., 30, 239-257.

Sellke, T., Bayarri, M., and Berger, J. (2001). Calibration of P-values for testing precise null hypotheses, Amer. Statist., 1, 62-71.

Serfling, R. (1980). Approximation Theorems of Mathematical Statistics, Wiley, New York.

Shaffer, J. (1995). Multiple hypothesis testing: A review, Annual Rev. Psyc., 46, 561-584.

Simes, R. (1986). An improved Bonferroni procedure for multiple tests of significance, Biometrika, 73, 751-754.

Sóric, B. (1989). Statistical 'discoveries' and effect-size estimation, JASA, 84, 608-610.

Storey, J. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value, Ann. Stat., 31, 2013-2035.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies, Proc. NAS, 100, 9440-9445.

Tukey, J.W. (1953). The Problem of Multiple Comparisons, Preprint, Princeton Univ., Dept. of Statistics.

van der Vaart, Aad (1998). Asymptotic Statistics, Cambridge University Press, Cambridge.