Chapter 6. Testing Hypotheses.

In Chapter 5 we explored how in parametric statistical models we could address one particular inference problem, the problem of estimation, without the need for prior information about the parameter. In effect we adapted to the absence of an agreed quantitative expression of prior information by adopting a more limited inferential goal. Instead of seeking a full posterior description of the uncertainty about the value of the parameter, we accepted the goal of coming up with an estimate of its value, together with a description of the accuracy of the estimation procedure. In this chapter we take a different approach. As in Chapter 5 we address a limited question, again without assuming a quantitative description of prior information. But here we ask a different question. Rather than trying to pin down the value of the parameter from among a potentially infinite number of possibilities, we ask a simple dichotomous question: Given that the parameter is equal to one of two specified values (or in one of two specified sets of values), which of the two should be inferred from the data? Among two specific hypotheses about the parameter value, which should be accepted and which rejected? Before introducing a formal statement of the problem, let us consider three examples which illustrate different degrees of its complexity and will help serve as motivation.

Example 6.A (Pattern Recognition). It is becoming commonplace to encounter machine recognition of human handwriting. "Personal Organizers" such as the Palm Pilot must recognize hastily scrawled letters and digits; the U. S. Postal Service employs mechanical readers to turn handwritten zipcodes into barcodes. In each case a handwriting pattern is registered through an optical scanning device as a set of pixels, and then each such recorded pattern is assigned to a letter or digit. How is (or how should) this be accomplished? To focus the question, consider the following simplified version. A scheme is required to decide whether a handwritten digit is a "0" or a "6". That is, a pattern of dots (pixels) is recorded and we must decide if it would most reasonable have arisen from an attempt to write a "0" or a "6". Here the "parameter" is the intended numeral; θ = "0" or "6". The data X consists of the pattern of pixels. For example, if the grid were 20x30 pixels, there would be 2^{20x30} possible patterns. The model is a specification of the probability distribution $p(x|\theta)$, the respective probabilities of all possible pixel patterns for each of the two values of θ. This might be determined from a large training set, for example, a large database assembled (e.g. by the Postal Service) where it was known in each case which of the two digits was intended by the writer. That is, for a single observed pattern X that could have arisen from either "0" or "6", our problem would be to decide which digit should be assigned. It is not enough to assign a value according to whether $p(x|\theta) > .5$, since this might be true for neither or both values. Nor, as we shall see, is it enough to assign (for example) "6" if $p(x|\theta = \text{"6"}) > p(x|\theta = \text{"0"})$. A theory is needed.

(See Figure 6.1).

Example 6.B (Acceptance Sampling). Suppose we are designing a screening policy for a procurement office of a large tech corporation, to be used to judge acceptability of lots arriving from suppliers. To each arriving item there is (at least in principle) a time-to-failure X, the length of time it would operate continuously under stable conditions before it would fail. We might model this as being distributed

according to an exponential distribution with parameter $\theta > 0$, where $E(X) = 1/\theta$. We cannot, of course, test all items in each lot without destroying the lot, and so it is customary to test only a sample from each lot, producing as data a list of independently exponentially distributed failure times $X_1, X_2, \ldots, X_n$. Suppose a lot with $\theta \leq 1.0$ is considered "good" and one with $\theta > 1.0$ is considered "bad". How should we decide whether of not a particular lot is "good" or "bad" based upon such a list of data? Should we simply take the mean or the median of the X's and compare it with 1.0? In this example we are to decide between two sets of values for the parameter θ rather than between two specific values.

Example 6.C (Contingency Tables). In the 1880s Francis Galton sought to determine whether or not men and women select each other as marriage partners at least in part on the basis of height: Do tall men tend to marry tall women, do short women tend to marry short men, or do both select mates pretty much regardless of each others' height? A statistical reason underlay this question; he wanted to study the heritability of height in human populations and needed to know if it was necessary to incorporate the correlation of parental heights in the analysis. His analysis would be much simpler if he could assume the parental heights were independent. To address this issue he assembled the data in the following table.

Table 6.A. Galton's data on the relative heights of husbands and wives for 205 married couples (Galton, 1889, p. 206)

	Wife:	Tall	Medium	Short
	Tall	18	28	14
Husband:	Medium	20	51	28
	Short	12	25	9

Do these data support or refute the hypothesis of independence of spouses' heights? Let θ_{HT} stand for the probability that the husband of a randomly selected married couple is Tall, and θ_{WT} be the probability that the wife of a randomly selected married couple is Tall; these and the other similar probabilities are the marginal probabilities for the bivariate distribution of heights, i.e. θ_{HT} = Pr(Husband Tall and Wife Tall) + Pr(Husband Tall and Wife Medium) + Pr(Husband Tall and Wife Short). How then would we test if Pr(Husband Tall and Wife Tall) = $\theta_{HT} \cdot \theta_{WT}$, as well as for the other 8 combinations of heights? This problem is more complex that either of the previous two; it not only involves sets of possible values for the parameters of the bivariate distribution of heights, the sets are defined in terms of the marginal probabilities θ_{HT}, θ_{WT}, etc., which are not known and must be estimated from the data.

6.1 Testing Simple Hypotheses.

The simplest testing problem is that in which only two possible values of the parameter are being considered. While such a situation may seem so restrictive as to be useless in practice, that is not the case. Not only are a number of practical applications (such as the pattern recognition problem of Example 6.A) treatable in terms of this

framework, but because it yields a complete and general solution, it also provides the key to our treatment of more complex cases.

As in Chapter 5 we consider a setup where we observe data X (which may be a list $X_1, X_2, \ldots, X_n$), assumed to have a distribution $f(x|\theta)$ (or for a list, $f(x_1, x_2, \ldots, x_n|\theta)$), given as a probability function or density as appropriate. We shall define a <u>simple hypothesis</u> to be one that completely specifies the probability distribution of the data, and a <u>composite hypothesis</u> to be one that only specifies the distribution of the data as one of a set of probability distributions. Thus Example 6.A involves the comparison of two simple hypotheses (θ = "0" and θ = "6"), while Example 6.B involves the comparison of two composite hypotheses ($\theta \leq 1.0$ and $\theta > 1.0$). Formally we shall treat the problem of comparing two simple hypotheses as one of testing whether $\theta = \theta_0$ or $\theta = \theta_1$, where either value θ_0 or θ_1 would completely specify the distribution of the data. Our "test" will be described through a set of values for the data called a "Rejection Region." The procedure will be to decide upon a Rejection Region, and act accord to this plan:

If X is not in the Rejection Region decide $\theta = \theta_0$.

If X is in the Rejection Region decide $\theta = \theta_1$.

The problem then is to determine which values of the data correspond to each decision. Note that this is a strictly dichotomous decision problem: there are two and only two possible states of nature in the formulation, and two and only two actions are permitted – no indecision or indifference is permitted. If the data are a list of continuous measurements there will be a bewilderingly infinite number of possible Rejection Regions, many of them apparently sensible: do we act on the basis of the arithmetic mean or the median, for example? Even with a finite set of possibilities as in Example 6.A the number of choices is extremely large. And yet, surprisingly, this is one area in statistics where we are led to a simple and definitive best answer, or more correctly, a best class of answers. To test which of two simple hypotheses are indicated by the data one should always use a <u>Likelihood Ratio Test</u>: Take as the Rejection Region those X for which

$$\frac{f(X|\theta_1)}{f(X|\theta_0)} > K, \text{or equivalently, } f(X|\theta_1) > Kf(X|\theta_0).$$

This test is of course not complete without indicating how one might determine K. To complete it we need to introduce explicit performance criteria. To this end we introduce some further terminology.

Let us describe our testing problem as deciding between the hypotheses

H_0: $\theta = \theta_0$ (the "null hypothesis")

H_1: $\theta = \theta_1$ (the "alternative hypothesis").

There are clearly two kinds of possible errors: we might reject a true H_0 or we might accept a false H_0. Let

α = Pr(decide $H_1|H_0$ true), called the probability of a "Type I error"

β = Pr(decide $H_0|H_1$ true), called the probability of a "Type II error"

Equivalently one may write

α = Pr(X is in the Rejection Region$|\theta = \theta_0$)

β = Pr(X is not in the Rejection Region$|\theta = \theta_1$)

Clearly we want both α and β to be small; it would be ideal (but generally unattainable) to have both equal to zero. If, however, we are willing to specify a permissible level for one of them, say α, then best values for both β and K are determined, and an elegant theorem of mathematical statistics that dates from 1933 states the corresponding likelihood ratio test is the optimum test for that level of the probability of error. By custom we specify α and then determine β and K, although for simple hypotheses the two situations are symmetric. Before presenting the simple proof of this result (called the Neyman-Pearson Lemma), let us look as some examples of its use.

Example 6.B (for simple hypotheses). Consider the Acceptance Sampling problem, but with the artificial additional assumption that we stipulate that either

$\theta = 1.0$ (the lot is "good"), or

$\theta = 2.0$ (the lot is "bad"),

and that the data consists of a single measured lifetime X, assumed to have density

$$f(x|\theta) = \begin{cases} \theta e^{-\theta x} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

(Recall that the expected lifetime of a part is $E(X)=1/\theta$.)

We then are testing

H_0: $\theta = 1.0$ $(= \theta_0)$ (the "null hypothesis")

vs.

H_1: $\theta = 2.0$ $(= \theta_1)$ (the "alternative hypothesis"),

and the likelihood ratio test will reject H_0 when

$$\frac{f(x|\theta_1)}{f(x|\theta_0)} > K, \text{ or } \frac{2e^{-2x}}{1e^{-1x}} > K, \text{ or } 2e^{-2x+x} = 2e^{-x} > K.$$

Equivalently, reject H_0 when $e^{-x} > K/2$, or $X < -\log_e(K/2) = C$. Thus the test may simply be described as, "reject H_0 if $X < C$ and decide that the lot is bad; otherwise decide the lot is good."

The theory has only guided us to the form of the test ("reject if $X < C$"); the further step of determining C (which is equivalent to determining K) requires the use of the performance level we insist upon. To this end suppose that company policy will permit that the probability we accept a bad lot could be as high as 0.1 but no higher; that is, we take $\alpha = 0.1$. Then since α = Pr(X is in the Rejection Region$|\theta = \theta_0$) = P($X < C \mid \theta = 1.0$) = $1 - e^{-C}$ for this exponential distribution, taking $\alpha = 0.1$ yields $1 - e^{-C} = 0.1$, or $C = 0.105$. The test is then complete: Reject the lot if the sample item tested fails sooner than time 0.105.

(See Figure 6.2)

Once the test is specified we can consider the other probability of error, that of "Type II error" – that we will with this test accept a bad lot.

β = Pr(decide $H_0|H_1$ true) = P($X \geq C|\theta = 2.0$) = $e^{-2C} = e^{-.21} = 0.81$.

This is a rather large probability of error, reflecting the fact that we have relatively little information in a single observation. There is an evident tradeoff between α and β: if we want smaller α we must accept a larger β.

Another term we will use is the power of a test: The power γ is simply one minus the probability of a Type II error:

$\gamma = 1 - \beta = \Pr(\text{Reject } H_0 \mid H_0 \text{ false})$.

The term should be understood in the sense of measuring the discriminatory power of the test, the test's ability to discriminate correctly between the two hypotheses. In this example, $\gamma = 1 - .81 = .19$.

Example 6.D (Testing a normal mean, with known variance). Consider a situation in which the data are a list of continuous measurements $X_1, X_2, \ldots, X_n$ that we are content to model as independent, each distributed as a normal distribution, $N(\mu, \sigma^2)$. The parameter here is, as in Example 5.D, generally considered as the pair, $\theta = (\mu, \sigma^2)$, and the likelihood function is given by

$$L(\theta) = L(\mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}} \right]$$

$$= (2\pi)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2}.$$

A testing problem for this example might involve comparing two pairs $\theta_0 = (\mu_0, \sigma_0^2)$ and $\theta_1 = (\mu_1, \sigma_1^2)$. For illustrative purposes we shall consider here a more restrictive test, comparing the two pairs $\theta_0 = (\mu_0, \sigma_0^2)$ and $\theta_1 = (\mu_1, \sigma_0^2)$, where the variance is the same in both cases. We are, in effect, assuming that the variance is a known value σ_0^2 and that only the mean μ remains in doubt. Applications where this is a reasonable assumption are rare, but the simplification it allows helps to make a valuable point. More practical normal testing problems will be considered in a later chapter.

We consider then the testing problem,

H_0: $\theta = \theta_0 = (\mu_0, \sigma_0^2)$ vs.

H_1: $\theta = \theta_1 = (\mu_1, \sigma_0^2)$,

where we suppose for definiteness that $\mu_1 > \mu_0$. We are then presented with a list of data that come from one of two specific normal distributions, and asked to decide which of the two. Many possible classes of tests could be considered, for example tests based upon the arithmetic mean, on the sample median, on the largest (or smallest) sample values, or upon some strange combination of all of these. The Neyman-Pearson Lemma tells us to use a likelihood ratio test, which gives an unequivocal answer to the question. The likelihood ratio here is

$$\frac{L(\mu_1)}{L(\mu_0)} = \frac{(2\pi)^{-\frac{n}{2}}\sigma_0^{-n}e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu_1)^2}}{(2\pi)^{-\frac{n}{2}}\sigma_0^{-n}e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu_0)^2}}$$

$$= e^{-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{n}(X_i-\mu_1)^2-\sum_{i=1}^{n}(X_i-\mu_0)^2\right]}$$

$$= e^{-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{n}X_i^2-2\mu_1\sum_{i=1}^{n}X_i+n\mu_1^2-\sum_{i=1}^{n}X_i^2+2\mu_0\sum_{i=1}^{n}X_i-n\mu_0^2\right]}$$

$$= e^{-\frac{1}{2\sigma_0^2}\left[2(\mu_1-\mu_0)\sum_{i=1}^{n}X_i+n\mu_1^2-n\mu_0^2\right]}$$

$$= e^{\frac{1}{\sigma_0^2}\left[(\mu_1-\mu_0)\sum_{i=1}^{n}X_i\right]}e^{-\frac{n}{2\sigma_0^2}\left[\mu_1^2-\mu_0^2\right]}$$

Looked at as a function of the data, this will (since $\mu_1 - \mu_0 > 0$) be large when $\sum_{i=1}^{n} X_i$ is large, or equivalently when $\bar{X} > C$, where C is a constant to be determined.

Note that the theory is used here to derive the <u>form</u> of the test; it tells us that the test should be based upon only the arithmetic mean of the data, $\bar{X}$, and that large values of the $\bar{X}$ should be taken as evidence for H_1. In order to implement the test it remains to use the performance criteria to find <u>how</u> large, that is, to find C. We do not try to trace back through the calculations we have made, back to the "K" of the Neyman-Pearson Lemma, to determine the threshold or cutoff value C; rather we take the much simpler course of working directly with the statistic we have found, namely $\bar{X}$.

We determine C based upon the maximum level α we are willing to accept for the probability of rejecting H_0 when H_0 describes the true state of affairs. We know (Example 5.F, (5.50)) that for our model, $\bar{X}$ has a $N(\mu, \sigma^2/n)$ distribution, and our test rejects H_0 when $\bar{X} > C$. We might take as a limit for α a value such as 0.1 or 0.05 or 0.01, depending upon how strong we would insist the evidence be before rejecting H_0. Now

$$\begin{aligned}\alpha &= P(\bar{X} > C|H_0)\\ &= P\{(\bar{X}-\mu_0)/(\sigma_0/\sqrt{n}) > (C-\mu_0)/(\sigma_0/\sqrt{n})|H_0\}\\ &= P\{Z > (C-\mu_0)/(\sigma_0/\sqrt{n})\},\end{aligned}$$

where Z has a standard normal distribution, N(0,1). Then if $z_{1-\alpha}$ is the $(1-\alpha)$th percentage point of the standard normal (the value such that $P(Z \le z_{1-\alpha}) = 1 - \alpha$ and so $P(Z > z_{1-\alpha}) = \alpha$, to be found from a table), we have

$$C = \mu_0 + z_{1-\alpha}(\sigma_0/\sqrt{n}).$$

From this we can find $\beta = P(\bar{X} \le C|H_1)$ and the power of the test, $\pi = P(\bar{X} > C|H_1)$. For example, if n = 10, $\alpha = 0.05$, $\mu_0 = 15$, $\mu_1 = 17$, and $\sigma_0 = 2$, then since $z_{.95} = 1.645$, we have $C = 15 + 1.645(2/\sqrt{10}) = 15 + 1.04 = 16.04$. Then $\beta = P(\bar{X} \le 16.04|H_1) = P(Z \le (16.04 - 17)/(2/\sqrt{10})) = P(Z \le -1.52) = .0643$, and $\pi = 1 - .0643 = .9357$.

(Figure 6.3)

6.2 The Neyman-Pearson Lemma.

We consider the problem of testing

H_0: X has distribution $f(x|\theta_0)$ vs. H_1: X has distribution $f(x|\theta_1)$

The notation is not intended to be restrictive; f may be a discrete probability distribution or a density; θ may be a single parameter or a pair of parameters (or a more complicated description); and X may be a single variable or a list $X_1, X_2, ..., X_n$, in which case f is a joint distribution. The argument to be presented is perfectly general; all that is required is that the distribution of X be completely specified under each hypothesis. As before we let

α = P(reject $H_0|H_0$ true) = P(reject $H_0|\theta = \theta_0$) = probability of a type I error,

β = P(accept $H_0|H_1$ true) = P(accept $H_0|\theta = \theta_1$) = probability of a type II error.

The Neyman-Pearson Lemma: Given α, no test with the same or lower α has a lower β than the likelihood ratio test with the given α.

In other words, the best test is a likelihood ratio test (also called the Neyman-Pearson test, or NP test), and all that remains is to find the appropriate cutoff K for the given α. The likelihood ratio test rejects if X = x for any x satisfying

$$f(x|\theta_1) > Kf(x|\theta_0).$$

As we have seen in the examples of the previous section, usually the test is expressed in an equivalent form that is more convenient for implementation in the problem under consideration.

Proof: Since there are but two possible decisions, it will be mathematically convenient to describe the test by an "indicator function":

Let $I_{NP}(x) = 1$ if $f(x|\theta_1) > Kf(x|\theta_0)$

$= 0$ otherwise

This then gives a succinct description of the NP test, which rejects H_0 if and only if $I_{NP}(x) = 1$. This description is adopted because it allows us easily to link the test to its probabilities of error: Since $I_{NP}(X)$ is a random variable that is either 0 or 1, its expectation is $EI_{NP}(X) = 0 \cdot P\{I_{NP}(X)=0\}+1 \cdot P\{I_{NP}(X)=1\} = P\{I_{NP}(X)=1\}$.

In particular, let α_{NP} be the probability of a type I error for the NP test; we note then that $\alpha_{NP} = E(I_{NP}(X)|\theta_0)$, and also $1 - \beta_{NP} = E(I_{NP}(X)|\theta_1)$.

Let T be any other test with the same or smaller α. Let $I_T(x)$ be the "indicator function" of its rejection region; that is,

$I_T(x) = 1$ if test T rejects H_0 when the data X = x

$= 0$ otherwise.

Then just as with the NP test, we see that we can write T's probability of a type I error as $\alpha_T = E(I_T(X)|\theta_0)$, and $1 - \beta_T = E(I_T(X)|\theta_1)$.

Now we claim that for all x

$$I_{NP}(x)\{f(x|\theta_1) - Kf(x|\theta_0)\} \geq I_T(x)\{f(x|\theta_1) - Kf(x|\theta_0)\}$$

To see this, just consider all cases: If $I_{NP}(x) = 1$, the part in {} is ≥ 0 and since $I_{NP}(x) = 1 \geq I_T(x)$, the inequality holds. Similiarly, if $I_{NP}(x) = 0$, {} ≤ 0 and the inequality holds then too, since $I_T(x) \geq 0 = I_{NP}(x)$.

The proof is completed by showing that the Lemma can be deduced from this inequality. Multiplying the inequality out we get

$$I_{NP}(x)f(x|\theta_1) - KI_{NP}(x)f(x|\theta_0) \geq I_T(x)f(x|\theta_1) - KI_T(x)f(x|\theta_0).$$

If we then sum over all x (in the discrete case) or integrate over all x (in the continuous case – it will be a multiple dimensional integral if f is a joint distribution), we get

$$E[I_{NP}(X)|\theta_1] - KE[I_{NP}(X)|\theta_0] \geq E[I_T(X)|\theta_1] - KE[I_T(X)|\theta_0],$$

or, replacing these expectations by their expressions as probabilities,

$1 - \beta_{NP} - K\alpha_{NP} \geq 1 - \beta_T - K\alpha_T$, which leads to

$1 - \beta_{NP} \geq 1 - \beta_T + K(\alpha_{NP} - \alpha_T)$, and (since $\alpha_{NP} - \alpha_T \geq 0$ and $K \geq 0$), this implies

$1 - \beta_{NP} \geq 1 - \beta_T$, or

$\beta_T \geq \beta_{NP}$.

That is, the test T has at least as high a probability of a type II error as does the NP test.

QED

6.3 Uniformly Most Powerful Tests.

The Neyman-Pearson Lemma points to the solution of a restricted class of problems, the comparison of simple statistical hypotheses. It states that in these situations, for a given level α for the probability of a type I error, the likelihood ratio test will minimize the probability β of a type II error and hence maximize the discriminatory power π: The likelihood ratio test is the most powerful test at the given level α. But what if one of the hypotheses is composite, what if the situation is complicated (as in Examples 6.B and 6.C) by the need to compare hypotheses that allow for variety in the distribution of the data? In a limited class of such problems the likelihood ratio test answers the more complicated question, almost inadvertently.

Consider the problem of Example 6.D. Suppose that instead of testing

H_0: $\theta = \theta_0 = (\mu_0, \sigma_0^2)$ vs.

H_1: $\theta = \theta_1 = (\mu_1, \sigma_0^2)$ for a single specific $\mu_1 > \mu_0$,

we wish to test the same H_0 against the composite alternative H_1,

H_1: $\theta = \theta_1 = (\mu_1, \sigma_0^2)$ for any $\mu_1 > \mu_0$.

That is, we ask, "Is $\mu = \mu_0$, or is μ a larger value?", without specifying which larger value.

The problem that could potentially arise is that for every choice of an alternative μ_1 value, we could get a different test. We might get a different C for each μ_1, or a different form of rejection region. Fortunately, in this example (and some others) this problem does not arise. Inspection of the derivation of the form of the test in Example 6.D shows that we only used the fact that $\mu_1 > \mu_0$, the value of μ_1 did not otherwise come into play. The form of the test is $\overline{X} > C$ no matter which $\mu_1 > \mu_0$ is considered, and furthermore the calculation of C only involved μ_0, not μ_1. The values of α and β do depend upon μ_1, but the test does not. The conclusion is that the test is not only most powerful for a particular μ_1, it is most powerful against all $\mu_1 > \mu_0$. We would describe this by saying the test is uniformly most powerful against all alternatives $\mu_1 > \mu_0$. A graph of the power $\pi(\mu_1)$ vs. μ_1 is called the power function of the test, and it shows how the discriminatory power increases as the distance from the alternative μ_1 to the null hypothesis value μ_0 increases. (see Figure 6.3)

In some respects this is a happy mathematical accident. Even for this example there is no uniformly most powerful test if the class of alternatives is enlarged further, to

H_1'': $\theta = \theta_1 = (\mu_1, \sigma_0^2)$ for any $\mu_1 \neq \mu_0$.

To see this, just recall that for the set of alternatives $\mu_1 > \mu_0$ the most powerful test rejects H_0 when $\bar{X} > C$, while similar, symmetrical reasoning would show easily that the most powerful test for alternatives $\mu_1 < \mu_0$ would reject for $\bar{X} < C$[$= \mu_0 - z_{1-\alpha}(\sigma_0/\sqrt{n})$. Clearly the test that is most powerful for one set of alternative is essentially <u>least</u> powerful for the other, and therefore no uniformly most powerful test exists for H_0 vs. H_1''. Figure 6.4 shows the power functions for these two tests, together with that for a practical compromise test with the same level α that rejects when $|\bar{X} - \mu_0| > C$[$= z_{1-\alpha/2}(\sigma_0/\sqrt{n})$.

(Figure 6.4)

Example 6.E (Testing a binomial proportion.) Suppose we wish to test a hypothesis about a binomial parameter; for example, to test the hypothesis that a coin is fair vs. the set of alternatives that it is weighted in favor of heads. Our data then will be the count X of the number of successes in n independent trials, and a plausible model would be that X has a Binomial (n, π) distribution. We wish to test

H_0: $\pi = \pi_0$ vs.

H_1: $\pi = \pi_1$ for any $\pi_1 > \pi_0$.

Testing a coin for perfect balance would correspond to taking $\pi_0 = 0.5$.

An easy computation shows that the likelihood ratio is

$$\frac{p(X \mid \pi_1)}{p(X \mid \pi_0)} = \left[\frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}\right]^X \left[\frac{1-\pi_1}{1-\pi_0}\right]^n$$

Since both $\left[\frac{\pi_1}{\pi_0}\right] > 1$ and $\left[\frac{1-\pi_0}{1-\pi_1}\right] > 1$, it follows that $\left[\frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}\right] > 1$, and so the likelihood ratio for any $\pi_1 > \pi_0$ will be large when X is large: The most powerful test for each such π_1 (and hence the uniformly most powerful test) will be of the form, reject H_0 if $X > C$.

It remains to specify C, and here we run into a new wrinkle. The cutoff C is to be determined by α, but because X has a discrete distribution, not all α are feasible. For example, for n = 5 and $\pi_0 = 0.5$, the distribution of X is given by

x:	0	1	2	3	4	5
$p(x\mid\pi_0)$:	.03	.16	.31	.31	.16	.03

Since $\alpha = P(X > C \mid \pi = \pi_0)$, if we take C = 4 we will have $\alpha = .03$, while if we take C = 3, we will have $\alpha = .16 + .03 = .19$. Figure 6.5 shows the power function if we take $\alpha = .03$, $\beta(\pi_1) = P(X > 4 \mid \pi = \pi_1) = P(X = 5 \mid \pi = \pi_1) = \pi_1^5$. As in the previous example (and for essentially the same reason), there is no uniformly most powerful test for H_1'': $\pi = \pi_1$ for any $\pi_1 \neq \pi_0$.

(Figure 6.5)

In general, for testing composite hypotheses, no uniformly most powerful test is available. There are exceptions such as those in the previous two examples where "one-

sided" alternative hypotheses are considered, but in general, some sort of compromise must be made.

6.4 Testing Composite Hypotheses: The General Likelihood Ratio Test.

Consider now the general problem of testing composite hypotheses, that is, of testing

H_0: θ is one of a group of values Ω_0 vs.

H_1: θ is one of an alternative group of values Ω_1.

This would encompass all of the three Examples 6.A, 6.B, and 6.C, as well as an extraordinarily large number of other statistical applications. In the simple case where each group consists of but a single value (i.e. $\Omega_0 = \{\theta_0\}$ and $\Omega_1 = \{\theta_1\}$), this reduces to the case covered by the Neyman-Pearson Lemma, but in other cases, as we have seen, there is in general no single best test: even when $\Omega_0 = \{\theta_0\}$, different choices of Ω_1's may yield completely different best tests. In this section we discuss how an extension of the likelihood ratio idea can produce compromise solutions for the more general problem. The tests that result from this compromise may not be best in the Neyman-Pearson sense (they could not be, in general), but they will satisfy the assumed restrictions on α for all "null hypothesis" values Ω_0 and perform reasonably well against all alternatives Ω_1.

Intuitively, the idea is a simple one, akin to a tournament between two schools or nations. Each of the two combatants Ω_0 and Ω_1 are allowed to choose champions and the two champions then settle the matter according to agreed upon rules. For example, we might let Ω_0 and Ω_1 each choose restricted maximum likelihood estimates,

$$\hat{\theta}_0 = \max_{\theta \in \Omega_0} L(\theta) \text{ and } \hat{\theta}_1 = \max_{\theta \in \Omega_1} L(\theta),$$

and then settle the competition according to whether or not

$$\frac{L(\hat{\theta}_1)}{L(\hat{\theta}_0)} > K.$$

Indeed, that is essentially what we shall do, although for technical mathematical reasons we precede slightly differently. For the applications we have in mind, $\max_{\theta \in \Omega_1} L(\theta)$ may not exist, and so instead we shall consider a contest between $\hat{\theta}_0$ and the unrestricted maximum likelihood estimate, between
$\hat{\theta}_0$ which achieves $\max_{\theta \in \Omega_0} L(\theta)$, and $\hat{\theta}$ which achieves $\max_{\theta \in \Omega} L(\theta)$, where $\Omega = \Omega_0 \cup \Omega_1$. We would then evaluate the competition in terms of

$$\lambda = \frac{\max_{\theta \in \Omega_0} L(\theta)}{\max_{\theta \in \Omega} L(\theta)} = \frac{L\left(\hat{\theta}_0\right)}{L\left(\hat{\theta}_1\right)}.$$

There are two possibilities: supposing the unrestricted maximum is achieved at a unique $\hat{\theta}$, either it will occur for a θ in Ω_0 (in which case $\lambda = 1$), or it will occur for a θ in Ω_1 (in which case necessarily $0 < \lambda < 1$). For hypotheses Ω_0 that are genuinely restrictive the second of these will be the usual case regardless of which hypothesis is

true: generally an unrestricted maximum exceeds a restricted maximum. The question then would be whether or not λ is sufficiently smaller than 1 to reject the null hypothesis H_0. Accordingly, the general likelihood ratio test will reject H_0 when $\lambda < K$, where K is chosen so that

$$P\{\lambda < K|\theta\} \le \alpha \text{ for all } \theta \in \Theta_0.$$

In the special case where both hypotheses are simple, where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, we will have

$$\lambda = L(\theta_0)/\max\{L(\theta_0). L(\theta_1)\} = \min\{1, L(\theta_0)/L(\theta_1)\}.$$

Then $\lambda < K < 1$ if and only if $L(\theta_1)/L(\theta_0) > K^{-1} > 1$, and so as long as $K < 1$ (which is true as long as $\alpha < 1/2$, the usual situation) the general likelihood ratio test is equivalent to the Neyman-Pearson test. We shall refer to both by the term "likelihood ratio test," leaving the context to make clear whether we mean the generalization or the special case.

The general likelihood ratio test is intuitively appealing, but this appeal may be deceiving. The "selection of champions" mentioned above is done based upon the same data which are then used to perform the test, and this superiority may not translate into generally better performance. Indeed, there is no mathematical guarantee that the test's performance (as measured by its power function) exceeds that of other possible tests. In the case of Example 6.D (testing a normal mean with variance known) with the set of alternatives

$$H_1'': \ \theta = \theta_1 = (\theta_1, \sigma_0^2) \text{ for any } \theta_1 \neq \theta_0,$$

we have

$$\lambda = \frac{L(\theta_0)}{L(\bar{X})}$$

$$= e^{-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{n}(X_i-\theta_0)^2-\sum_{i=1}^{n}(X_i-\bar{X})^2\right]}$$

$$= e^{-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^{n}X_i^2-2\theta_0\sum_{i=1}^{n}X_i+n\theta_0^2-\sum_{i=1}^{n}X_i^2+2\bar{X}\sum_{i=1}^{n}X_i-n\bar{X}^2\right]}$$

$$= e^{-\frac{1}{2\sigma_0^2}\left[-2n\theta_0\bar{X}+n\theta_0^2+2n\bar{X}^2-n\bar{X}^2\right]}$$

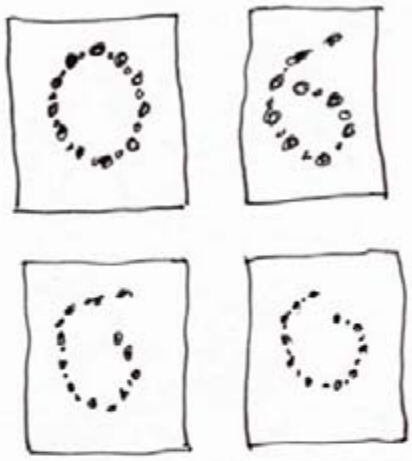
$$= e^{-\frac{n}{2\sigma_0^2}\left[\bar{X}^2-2\theta_0\bar{X}+\theta_0^2\right]}$$

$$= e^{-\frac{n}{2\sigma_0^2}(\bar{X}-\theta_0)^2}$$

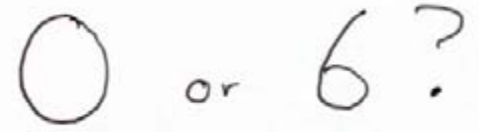
which will be < K when $|\bar{X} - \theta_0| > C''$, the same "compromise" test presented earlier. Thus here, where no uniformly most powerful test exists, the general likelihood ratio test produces a reasonable solution.

Not only do likelihood ratio tests tend to give "reasonable" answers, they give tractable solutions to some extremely complex statistical problems, and if the sample size is large they can be shown to have many good properties, at least approximately. Generally we will use the likelihood ratio formulation here as in the simple case to derive

simpler expressions for the form of the test; the application of the test is then carried out in the simpler setting. However with large samples the test may be carried out directly in terms of χ. We will present that result after exploring the use of these tests in a class of applications where they lead to “Chi-square tests”.



Four possible X's



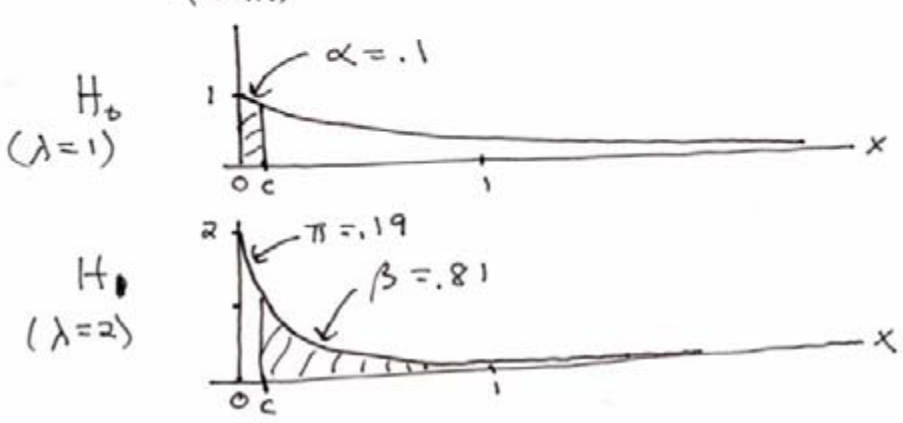
Two Θ's

Figure 6.1. Some patterns will be clear; others could arise from either Θ.

Figure 6.2

Densities of X under H_0 ("Good lot") and H_1 ("Bad lot").

(a) For $\alpha = .1$ $(C = .105)$



(b) For $\alpha = .2$ $(C = .223)$

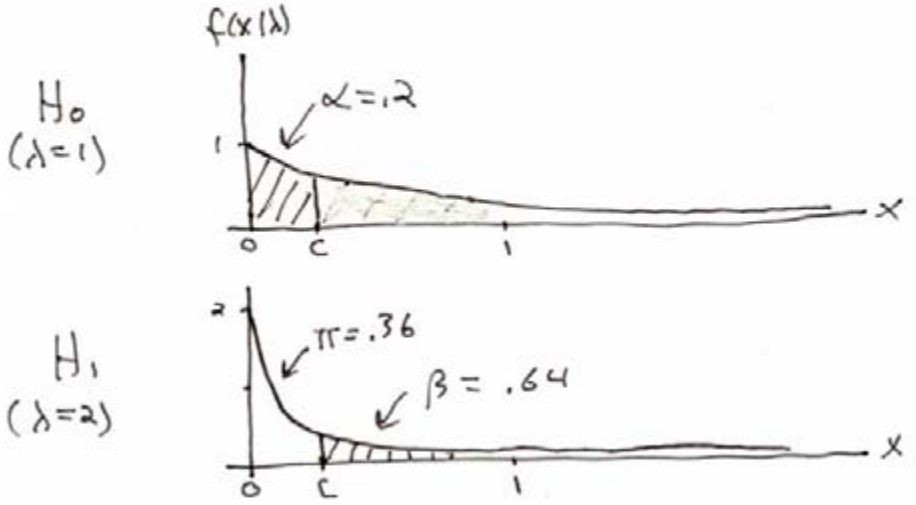
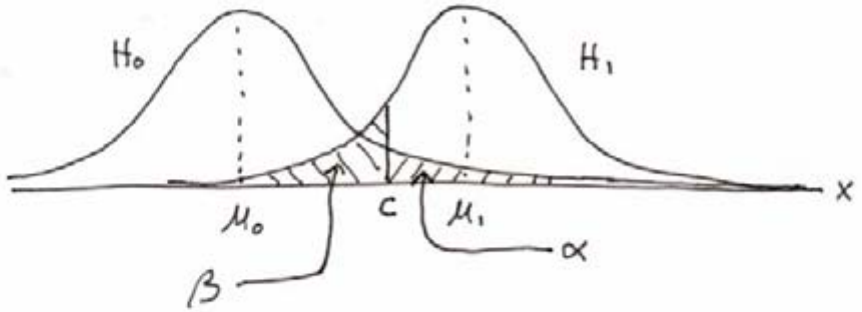
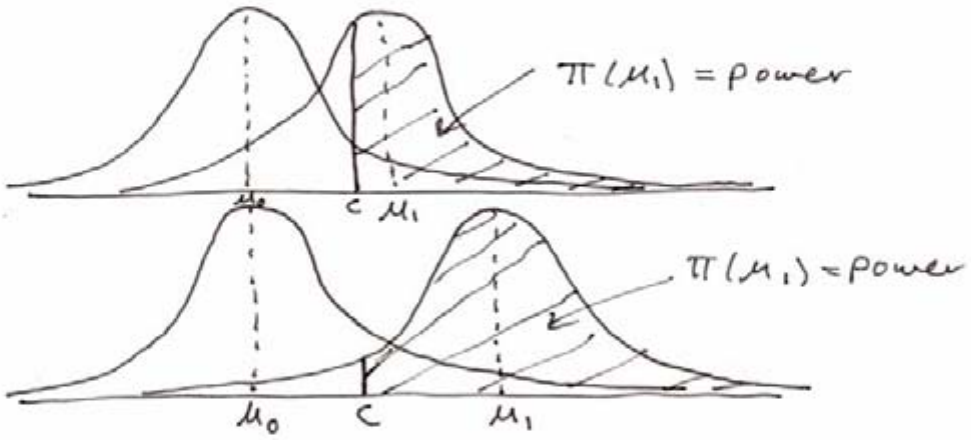
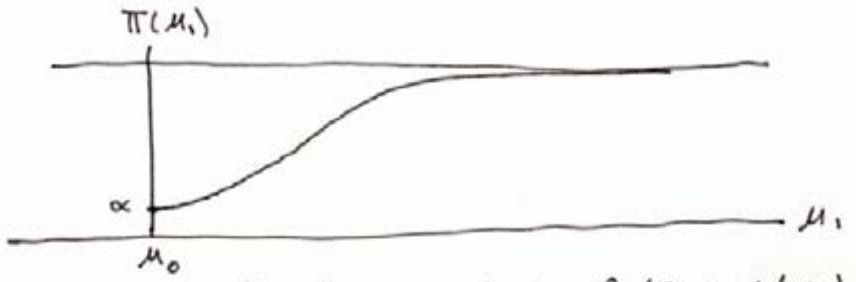


Figure 6.3



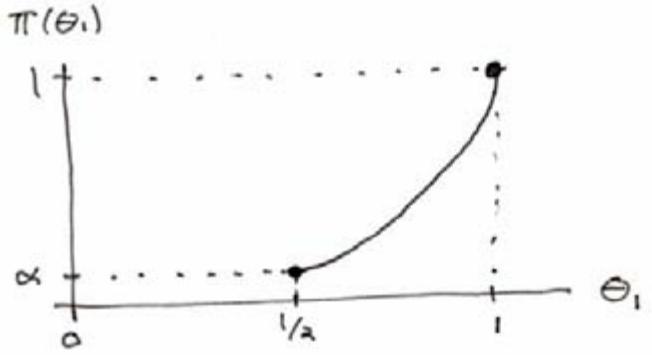
(a) Testing a simple hypothesis vs. a simple alternative.





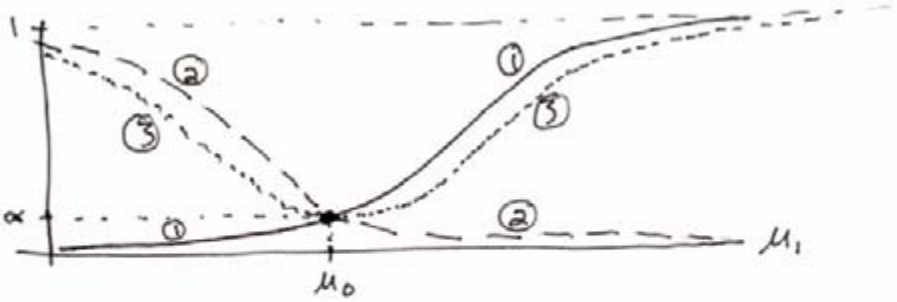
(b) The power function $\pi(\mu_1) = P_r(\text{Reject}(\mu_1))$, as a function of the alternative μ_1

Figure 6.4



The power function $\pi(\theta_1) = P(X > 4 \mid \theta_1)$ for the binomial test of $\theta = 1/2$ vs. $\theta > 1/2$, for $n = 5$ and $\alpha = .03$. $\pi(\theta_1) = \theta_1^5$ here.

Figure 6.5



The power functions $\pi(\mu_1) = P(\text{Reject} \mid \mu_1)$ for the level α tests:

① —— Reject if $\bar{X} > c$ (Best vs $\mu_1 > \mu_0$)

② - - - Reject if $\bar{X} < c'$ (Best vs. $\mu_1 < \mu_0$)

③ ..... Reject if $|\bar{X} - \mu_0| > c''$ ("Pretty good" overall)

There is <u>no</u> "uniformly most powerful" test.