Lecture 19: Significance Tests, I

Relevant textbook passages:

Larsen–Marx [5]: Sections 7.2, 7.4, 7.5; 7.3, Appendices

19.1 A grossly oversimplified example

In the U.S., 6.9% of adults (16 million people) suffer from major depression, and the rate is 8.9% for your age group.\(^1\) Suppose you are employed by the FDA to decide whether some drug is effective in the treatment of major depression. There is a test whose outcome is subject to randomness due to variation in the physiology of patients. If the drug is ineffective, the test value \(X \sim N(0, 1)\), but if the drug is effective, the test values \(X \sim N(1, 1)\).

How do we decide on the basis of the test whether the drug is effective or not? Consider Figure 19.1, which shows the probability densities of the two normals. You can easily see that for values of \(x > 0.5\) the density from the \(\mu = 1\) Normal is larger, and for \(x < 0.5\) the density from the \(\mu = 0\) Normal is larger.

You might be tempted to say, let’s rule the drug to be effective if \(x > 1/2\). What if the drug is not effective? If \(\mu = 0\), then there is a 31% chance that

This means that there is a 31% chance that an ineffective drug will be ruled effective.

So what? Suppose I tell you that this drug will cost $275,000 over the life of a patient, and that taxpayers will foot the bill for the drug. Now what do you think about your decision rule? Suppose the cost were only $10 over the life of the patient?

I hope this suggests that the decision should be based on a loss function associated with the costs and benefits of the decision. That is, the exercise is one in statistical decision theory. Nevertheless, there is a principle sometimes advocated, called the Likelihood Principle that the decision be based solely on the picture of the densities, that is, on the likelihood functions. (My colleague and former professor Dave Grether refers to the advocates as the “likelihood brotherhood.”)

This is not entirely crazy. Most of “classical” statistical hypothesis testing is based on the likelihood principle, but there is still room in the classical approach for exercising judgment on how stringent to make the decision. This allows for different decision makers who have different cost/benefit considerations to use the same calculations, and adjust a single parameter of the test to suit their different loss functions.

Also, you should not stop here. As a researcher for the FDA you should not stop at running routine tests—your job is more than doing introductory statistics problem sets. You should be thinking about being smarter in making decisions. Perhaps there are characteristics of patients that can be used to identify those patients for whom the drug is effective. So you might want to collect data on each patient and “mine” it for information. Bear in mind that collecting data is costly, and in particular, certain medical tests have high price tags.\footnote{The economist in me compels me to point that there is a difference between price tags and real costs. The cost of an MRI machine is large, but once you have one, the cost of performing one more test is just the cost of the electric power and the technician’s time. The price tag is much higher.}

### 19.2 Hypothesis testing in the abstract

A typical data model or probability model is a density/likelihood function $f(x; \theta)$ depending on a datum $x$ from some set $\mathcal{X}$, and parameterized by $\theta$ belonging to some set $\Theta$. Having estimated parameters is usually not enough. We know that the estimates are random, and that if we are very very unlucky our estimates can be very misleading. Usually, we want to test some hypothesis about the probability model. There are two kinds of such hypotheses.

The first kind is a hypothesis about the parameter. A typical such hypothesis is of the form: $\theta \in A$, for some subset $A$ of the parameter space $\Theta$. For instance, we may be interested in whether the mean $\mu$ of a normal is greater than zero, $\mu \in [0, \infty)$. These tests are frequently referred to significance tests.
The second kind of hypothesis is whether our model is misspecified. Misspecification occurs when the data generating process is governed by a different functional form $g(x; \psi)$ for some possibly alternate parameter set. (Remember that we could take as our parameter space the set of all probability measures on $X$. But this space is so large we usually call this a nonparametric model.) An example of this might be whether the data come from any normal distribution, or whether an entirely different family, such as the Cauchy distribution, is a “better” model. These tests are frequently referred to specification tests.

We’ll start with the first kind of testing, parametric hypothesis testing in the context of a fixed data model $f$.

A bunch of definitions

- The data model: $f : X \times \Theta \to \mathbb{R}$. Typically, $X$ and $\Theta$ are subsets of some finite dimensional Euclidean spaces. $X$ is a random variable (or random vector) with either a density or a mass function given by $f(x; \theta)$. For convenience, I will typically refer to $f$ as a pdf. Recall that the likelihood function $L(\theta; x)$ is just $f(x; \theta)$ so I may also refer to $f$ as the likelihood function. Finally, when it is convenient, I may also write $f_\theta(x)$ for $f(x; \theta)$.

- The Null Hypothesis: Denoted $H_0$, it takes the form of statement $\theta \in \Theta_0 \subset \Theta$. It could be a simple as $\theta = 0$ (more properly $\theta \in \{0\}$). The next section elaborates on the role of the null hypothesis.

Often the null hypothesis is something the researcher hopes to prove false. For instance, if you want to show that a drug improves the cure rate for a disease, the null hypothesis is probably going to be that the difference in the cure rate (over the control group) is zero. If the drug is actually useful, then you should reject the null hypothesis.

- Alternative Hypothesis: $H_1$ is the hypothesis that $\theta \in \Theta_1 = \Theta \setminus \Theta_0$. The two sets $\Theta_0$ and $\Theta_1$ partition the parameter space $\Theta$. E.g., $\Theta = [0, \infty)$, $\Theta_0 = \{0\}$, and $\Theta_1 = (0, \infty)$. The null hypothesis is $\theta = 0$ and the alternative is $\theta > 0$.

- A simple hypothesis is that $\Theta_i$ has just one point, and a composite hypothesis is that $\Theta_i$ has more than one point.

Often a null hypothesis is simple and the alternative is composite, but that needn’t be.

The point of a test is to based on the datum (vector) $x$, either reject the null hypothesis in favor of the alternative, or to fail to reject the null hypothesis. (It is considered a faux pas to say that you accept the null hypothesis.)

How do you decide which hypothesis gets to be the null hypothesis, and which gets to be the alternative?
19.3 Choosing the null hypothesis

Brad Efron [1, pp. 556–557] gives a nice discussion of the role of the null and alternative hypotheses in scientific investigation. In the quotation below, the case he is referring to is known as **Bode’s Law**, namely that the the distance from the sun of the \( n \)th planet is of the form \( d_n = a + b2^n \). (Bode did not have a large sample of solar systems.) Efron is critiquing the analysis of Good [4] who was in turn critiquing Bode. In the quote, Model B is Bode’s Law, model \( B \) is Good’s alternative hypothesis, and C is Efron’s alternative. Note this is concerned with model specification rather than parametric hypotheses, but the comments on the role of the null hypothesis are still relevant.

The most interesting feature of this problem is the light it casts on the role of the null hypothesis in hypothesis testing. These terms are used here in Fisher’s sense in which the null hypothesis is by design a hypothesis of uninteresting structure compared to that which we are considering as an alternative, though it may contain its own interesting features. [...] However it is not necessary to believe in the null hypothesis in order to use it as a test against the alternative of interest. Very often, perhaps most of the time, we do not believe in the validity of the Fisherian null hypothesis, whether or not the test based on it accepts or rejects in the usual sense. [...] The null hypothesis in the context of this discussion plays the role of devil’s advocate, a competitor that an alternative of interest to us must soundly discredit in order to show its strength. [...] The conclusions of a significance test are bound to be less than completely satisfying given the indirect route of the argument. In the case at hand for instance, accepting C doesn’t mean we believe C is true (Figure B mildly discourages such a belief). All we can say is that a statistical model that is relatively uninteresting compared to Bode’s law would often yield data as “simple” as that actually observed, and this undermines the necessity of our belief in the law’s validity. Conversely even if we had decisively rejected C we still might fear that we had overlooked some other reasonable null hypothesis which would do better.

One should not be dismayed by the limitations of the Fisherian significance test since it is designed only to give us some direction toward the correct answer in situations like the present one where there is little data to work with. As more data accumulate in any given problem, significance testing becomes superfluous. [...] By definition “estimation” refers to situations where we believe we know all the possible relevant statistical models and we are simply trying to choose the correct one. Estimation is an inherently more
satisfying operation than significance testing, but demands more data or more theoretical knowledge from the statistician.

19.4 The abstract mechanics of a statistical test

A statistical test is made up of the following pieces:

• **Test statistic**: $T$, is a function of the data, and the null hypothesis. E.g., $T = (\bar{x} - \theta_0)/(s/\sqrt{n})$.

When we discussed estimators, I said that a statistic must be a function of the data, and not of the unknown parameters. If we have a simple null hypothesis, say $\theta = \theta_0$, the test statistic is allowed to depend on $\theta_0$ because in its role as the null hypothesis, $\theta_0$ is not unknown to us—we know quite well what its value is because we picked it.

Even if the null hypothesis is composite, but say of the form $\theta \leq \theta_0$, we can allow the test statistic depend on the boundary point $\theta_0$ (or any other fixed feature of the set $\Theta_0$).

• **Critical region**: $C$. If the value of $T$ belongs to $C$, the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise we fail to reject the null hypothesis.

The critical region is often either an interval (possibly infinite) or the complement of an interval. The the endpoint(s) define a critical values of the test. For example, if $C = [t^*, \infty)$, then $t^*$ is the critical value.

• **Significance level**: $\alpha$. This is, roughly speaking, the probability $P(T \in C)$ when $H_0$ is true. That is, it is the probability of rejecting the null when $T(x) = t$. If the test uses a critical region $C$, then $\phi$ is the indicator function $1_C$. When you would want to randomize? When the test statistic is discrete, and I largely ignore the issue, but read on to the next point.

• **Critical function**: $\phi$. This allows for randomizing when deciding whether or not to reject the null hypothesis. The value $\phi(t)$ is the probability of rejecting the null when $T(x) = t$. If the test uses a critical region $C$, then $\phi$ is the indicator function $1_C$. When you would want to randomize? When the test statistic is discrete, and I largely ignore the issue, but read on to the next point.

When the null hypothesis is simple, the meaning of $P(T \in C)$ is clear. The probability $P$ is actually $P_{\theta_0}$, the probability whose pdf is $f(\cdot; \theta_0)$. For a composite null hypothesis, we define the size $\alpha$ to be

$$\sup\{P_{\theta}(T \in C) : \theta \in \Theta_0\}.$$
That is, for a composite null hypothesis, the probability of rejecting the hypothesis for any parameter in the null hypothesis is no more than $\alpha$. The actual probability will depend on which $\theta \in \Theta_0$ is the “true” parameter.

Statisticians are inordinately fond of $\alpha = 0.05$ and $\alpha = 0.01$. This goes back to Ronald A. Fisher’s [2, 3] dicta in 1925. It also goes back to the pre-statistical software era in which you had to look up critical values in a table in a book, which listed only a few values of $\alpha$. When I was a freshman, one of the first books that everyone bought was the Chemical Rubber Company’s CRC Standard Math Tables [7]. You can see a typical page in Figure 19.2.

If the test statistic $T$ is a discrete random variable, for instance, a count of successes, it may be impossible to find a critical region of size exactly $\alpha$ due to the “lumpiness” of $T$. This is when you might want to randomize for a particular value $t$ of $T$, in order to get the probability of rejection exactly equal to $\alpha$. This is when to use a critical function instead of a critical region. When the null hypothesis is simple, $\theta = \theta_0$, then

$$E_{\theta_0} \phi(T) = \alpha.$$

- The significance level and critical region can only be computed if we know the distribution of $T$ given the parameters $\theta$. This is why so much effort has been devoted to figuring out the distribution of various test statistics.

19.5 Statistics means never having to say you’re certain

- A Type I error occurs when $H_0$ is rejected when in fact it is true. (False rejection.) This can only happen if we reject the hypothesis.

- A Type II error occurs when $H_0$ fails to be rejected when in fact it is false. (False acceptance.) This can only happen if we fail to reject the hypothesis.

- We cannot ever know if an error has occurred!

- The probability of committing a Type I error is $\alpha$, the significance level of the test. This is a choice made by the experimenter (we hope) before the experiment is conducted.

- Once we have chosen the critical region $C$ and the associated level of significance, with luck we can calculate the probability of a Type II error. This probability is frequently referred to as $\beta$. It depends on the actual parameter value, which of course we do not know.

- The complement of a type II error is to properly reject $H_0$ when it is indeed false. The probability of this is a function of the parameter $\theta$, namely $1 - \beta(\theta)$, and is called the power of the test. A graph of the power vs. $\theta$ is called the power curve of the test.
PERCENTAGE POINTS, STUDENT'S $t$-DISTRIBUTION

This table gives values of $t$ such that

$$F(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx$$

for $n$, the number of degrees of freedom, equal to $1, 2, \ldots, 30, 40, 60, 120, \infty$; and for $F(t) = 0.60, 0.75, 0.90, 0.95, 0.975, 0.99, 0.995,$ and $0.9995$. The $t$-distribution is symmetrical, so that $F(-t) = 1 - F(t)$

<table>
<thead>
<tr>
<th>$F$</th>
<th>.60</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
<th>.975</th>
<th>.99</th>
<th>.995</th>
<th>.9995</th>
</tr>
</thead>
<tbody>
<tr>
<td>.80</td>
<td>.280</td>
<td>.271</td>
<td>.262</td>
<td>.257</td>
<td>.256</td>
<td>.256</td>
<td>.256</td>
<td>.256</td>
</tr>
<tr>
<td>.00</td>
<td>.699</td>
<td>.686</td>
<td>.677</td>
<td>.675</td>
<td>.675</td>
<td>.675</td>
<td>.674</td>
<td>.673</td>
</tr>
<tr>
<td>1</td>
<td>.900</td>
<td>.890</td>
<td>.880</td>
<td>.870</td>
<td>.869</td>
<td>.869</td>
<td>.869</td>
<td>.869</td>
</tr>
<tr>
<td>2</td>
<td>.816</td>
<td>.810</td>
<td>.806</td>
<td>.802</td>
<td>.800</td>
<td>.799</td>
<td>.798</td>
<td>.797</td>
</tr>
<tr>
<td>3</td>
<td>.765</td>
<td>.759</td>
<td>.753</td>
<td>.749</td>
<td>.746</td>
<td>.745</td>
<td>.744</td>
<td>.744</td>
</tr>
<tr>
<td>4</td>
<td>.741</td>
<td>.734</td>
<td>.728</td>
<td>.724</td>
<td>.721</td>
<td>.720</td>
<td>.719</td>
<td>.718</td>
</tr>
<tr>
<td>5</td>
<td>.727</td>
<td>.720</td>
<td>.714</td>
<td>.710</td>
<td>.708</td>
<td>.706</td>
<td>.705</td>
<td>.704</td>
</tr>
<tr>
<td>6</td>
<td>.718</td>
<td>.711</td>
<td>.705</td>
<td>.701</td>
<td>.699</td>
<td>.697</td>
<td>.696</td>
<td>.695</td>
</tr>
<tr>
<td>7</td>
<td>.711</td>
<td>.706</td>
<td>.701</td>
<td>.697</td>
<td>.694</td>
<td>.691</td>
<td>.689</td>
<td>.688</td>
</tr>
<tr>
<td>8</td>
<td>.693</td>
<td>.688</td>
<td>.683</td>
<td>.679</td>
<td>.677</td>
<td>.674</td>
<td>.672</td>
<td>.671</td>
</tr>
<tr>
<td>9</td>
<td>.682</td>
<td>.677</td>
<td>.672</td>
<td>.669</td>
<td>.666</td>
<td>.664</td>
<td>.662</td>
<td>.661</td>
</tr>
<tr>
<td>10</td>
<td>.670</td>
<td>.666</td>
<td>.662</td>
<td>.659</td>
<td>.656</td>
<td>.654</td>
<td>.652</td>
<td>.651</td>
</tr>
<tr>
<td>12</td>
<td>.649</td>
<td>.645</td>
<td>.642</td>
<td>.639</td>
<td>.637</td>
<td>.635</td>
<td>.633</td>
<td>.632</td>
</tr>
<tr>
<td>13</td>
<td>.640</td>
<td>.637</td>
<td>.634</td>
<td>.631</td>
<td>.629</td>
<td>.627</td>
<td>.625</td>
<td>.624</td>
</tr>
<tr>
<td>14</td>
<td>.632</td>
<td>.629</td>
<td>.626</td>
<td>.623</td>
<td>.621</td>
<td>.619</td>
<td>.618</td>
<td>.616</td>
</tr>
<tr>
<td>15</td>
<td>.625</td>
<td>.622</td>
<td>.619</td>
<td>.617</td>
<td>.615</td>
<td>.613</td>
<td>.612</td>
<td>.610</td>
</tr>
<tr>
<td>16</td>
<td>.619</td>
<td>.616</td>
<td>.613</td>
<td>.611</td>
<td>.609</td>
<td>.607</td>
<td>.606</td>
<td>.605</td>
</tr>
<tr>
<td>17</td>
<td>.614</td>
<td>.611</td>
<td>.608</td>
<td>.606</td>
<td>.604</td>
<td>.603</td>
<td>.601</td>
<td>.600</td>
</tr>
<tr>
<td>18</td>
<td>.610</td>
<td>.608</td>
<td>.605</td>
<td>.603</td>
<td>.601</td>
<td>.599</td>
<td>.597</td>
<td>.596</td>
</tr>
<tr>
<td>19</td>
<td>.606</td>
<td>.604</td>
<td>.601</td>
<td>.599</td>
<td>.597</td>
<td>.595</td>
<td>.594</td>
<td>.593</td>
</tr>
<tr>
<td>20</td>
<td>.602</td>
<td>.599</td>
<td>.597</td>
<td>.595</td>
<td>.593</td>
<td>.592</td>
<td>.590</td>
<td>.589</td>
</tr>
</tbody>
</table>

*This table is abridged from the "Statistical Tables" of R. A. Fisher and Frank Yates published by Oliver & Boyd Ltd., Edinburgh and London, 1938. It is here published with the kind permission of the authors and their publishers.

Figure 19.2. A typical statistical table from the CRC Standard Math Tables [7].
• **Warning:** The classic bible of hypothesis testing by E. L. Lehmann [6, p. 61] uses $\beta$ to denote the power. Thus his $\beta$ is Larsen and Marx’s $1 - \beta$ and vice versa. I have not yet done a survey of my books to see which notation is more common.

• If we have two tests with the same significance level $\alpha$, if one is always more powerful than the other, then it is a (statistically) better test. (One way to increase the power is to get more data, which may be expensive.)

• The power of a test is influenced by the shape of the critical region. For instance, why do we usually take the critical region to be the regions where the density of $T$ is smallest? Because it makes the test more powerful. This is the essence of the **Neyman–Pearson Lemma**.

• Every test is characterized by its pair $(T, C)$ of test statistic and $T$ and critical region $C$. Let $\beta_{T,C}(\theta)$ denote the test’s probability of Type II error (false acceptance of the null) when the parameter is $\theta$,

$$\beta_{T,C}(\theta) = P_{\theta}(T \notin C).$$

If a test $(T^*, C^*)$ has size (significance level) $\alpha$, and if for every $\theta \in \Theta_1$,

$$1 - \beta_{T^*,C^*}(\theta) \geq \max\{1 - \beta_{T,C}(\theta) : \text{test } (T, C) \text{ has size } \alpha, \}$$

then we say that $(T^*, C^*)$ is **uniformly most powerful test** (UMP).

• UMPs are good to have, but they don’t always exist. We shall discuss situations where UMPs do exist, and what they look like, next time.

• When a UMP does not exist, many practitioners argue that a test should be selected to minimize a weighted average of $\alpha$ and $\beta$, where the weights reflect the tester’s concerns about the two types of errors.

**19.6 ★ Likelihood Ratio Tests for simple hypotheses**

Let us return to the simple example of Section 19.1 to make sense out of all this jargon and abstraction. We know that $X$ is either a Normal $N(0,1)$ random variable or a Normal $N(1,1)$ random variable. On the basis of $x$ we have to decide which.

The parameter set $\Theta$ for the mean $\mu$ has only two points, $\Theta = \{0, 1\}$. To avoid needles conclusion, let me say that the null hypothesis $H_0$, and the alternative $H_1$ are

$$H_0: \mu = 0, \quad H_1: \mu = 1.$$

Let me also write $f_0(x)$ for the pdf of $X$ under the null hypothesis, and $f_1(x)$ for the pdf under the alternative.

The same intuition that motivated our maximum likelihood estimation suggest that perhaps we ought to use a test like the following:
Let the test test statistic $T(x)$ just be $x$ itself. Define the **likelihood ratio**

$$\lambda(x) = \frac{f_1(x)}{f_0(x)}$$

**Warning:** I am writing the ratio the way that Lehmann [6, p. 64] writes it. Larsen–Marx [5, p. 380] invert the ratio. It’s merely a convention, but you have to know which one is being used or your inequalities will be reversed. I took an informal survey of my colleagues, to ask how they write likelihood ratios, and most of them agreed with Lehmann. However, Wikipedia agrees with Larsen and Marx.

For our normal case,

$$\lambda(x) = \frac{e^{-\frac{(x-1)^2}{2}}}{e^{-\frac{(x-0)^2}{2}}} = ce^x,$$

where $c = e^{-1/2}$.

A **likelihood ratio test** takes the form:

- Choose a cutoff $k > 0$,
- and reject $H_0$ if $\lambda(x) \geq k$,
- otherwise fail to reject $H_0$, or accept $H_0$ over $H_1$.

Note that if you invert the likelihood ratio the way that Larsen and Marx do, you want to reject $H_0$ if $\lambda \leq 1/k$.

So there is a one-parameter family of likelihood ratio tests parametrized by $k$. It is this freedom that lets you use the Likelihood Principle, but still accommodate cost/benefit considerations.

The likelihood ratio test is equivalent to

- Choose a cutoff $x^*$,
- and reject $H_0$ if $x \geq x^*$,
- otherwise fail to reject $H_0$, or accept $H_1$ over $H_0$.

The size $\alpha$ of the test is

$$P_0 (X \geq x^*) = 1 - \Phi(x^*).$$

The probability of False Acceptance

$$\beta(\theta_1) = P_1 (X < x^*) = \Phi(x^* - 1)$$

so the power is

$$1 - \Phi(x^* - 1).$$
Can we get more power?

19.6.1 Neyman–Pearson Fundamental Lemma  For testing a simple null versus a simple alternative, a likelihood ratio test maximizes the power, given the size.

Here is a reasonably convincing proof for the case of absolutely continuous test statistics, but if you want to dot all your is and cross all your ts, you need to use critical functions, not critical regions. See Lehmann [6, pp. 65–67] for a more complete proof.

Proof: Pick a point $c$ in the critical region, and a point $d$ in the non-critical region and imagine swapping tiny intervals about them:

\[
\Delta \alpha \approx -f_0(c) \delta + f_0(d) \varepsilon,
\]
\[
\Delta \beta \approx f_1(c) \delta - f_1(d) \varepsilon,
\]

where $\delta$ and $\varepsilon$ are the widths of the intervals around $c$ and $d$. Then

\[
\Delta \alpha = 0 \implies \varepsilon = \frac{f_0(c)}{f_0(d)} \delta.
\]

So

\[
\Delta \beta = \left[ f_1(c) - f_1(d) \frac{f_0(c)}{f_0(d)} \right] \delta
\]
\[
= \left[ \frac{f_1(c)}{f_0(c)} - \frac{f_1(d)}{f_0(d)} \right] f_0(c) \delta \geq 0.
\]

That is any small change to the critical region that keeps the significance (size) $\alpha$ constant (or reduces it) must increase $\beta$, the probability of Type II error, reducing the power.

So what can we do to get a more powerful test? Increasing the sample size reduces the standard deviation, so the power increases. The next chart shows the
effect of reducing the standard deviation by a factor of 4 (increasing the sample size by a factor of 16) for our toy problem. Of course, increasing the sample size can be expensive, especially if your experiment involves medical testing, colliding large hadrons, or mapping the climate on assorted planets.

![Graph showing α vs. β distributions](image)

### 19.7 The monotone likelihood ratio property

For composite hypotheses, likelihood ratio tests work best when the data model \( f(x; \theta) \) satisfies an additional property, known as the monotone likelihood ratio property. In this case, the Neyman–Pearson Fundamental Lemma generalizes, and likelihood ratio tests are UMP, they are characterized by critical values, and the notion of a \( p \)-value applies.

When \( \Theta \subset \mathbb{R} \), the probability model \( f(x; \theta) \) satisfies the **Monotone Likelihood Ratio Property (MLRP)** if there is real valued (one-dimensional) test statistic \( T(x) \), such that for every pair \( \theta, \theta' \), with

\[
\theta < \theta'
\]

the likelihood ratio

\[
\frac{f(x; \theta')}{f(x; \theta)}
\]

is nondecreasing in \( T(x) \).

We then say the likelihood ratio is monotone in \( T(x) \).

Another way to rewrite the MLRP is

\[
\left( \theta < \theta' \land T(x) < T(x') \right) \implies \frac{f(x; \theta')}{f(x; \theta)} \leq \frac{f(x'; \theta')}{f(x'; \theta)}.
\]

#### 19.7.1 Example
Consider the the Normal family \( N(\mu, 1) \). Here the one-dimensional parameter is \( \mu \) and the parameter space is \( \Theta = \mathbb{R} \). If we a sample \( n \) independent observations \( x_1, \ldots, x_n \), the sample mean \( \bar{x} \) is a sufficient statistic for \( \mu \) with a
Normal distribution with mean $\mu$ and variance $1/n$. Its density is

$$f(\bar{x}; \mu) = \frac{\sqrt{n}}{2\pi} e^{-(\bar{x} - \mu)^2/2}.$$ 

So for $\mu < \mu'$, the likelihood ratio is

$$\frac{e^{-(\bar{x} - \mu')^2/2}}{e^{-(\bar{x} - \mu)^2/2} e^{(\mu'^2 - \mu^2)/2}} = e^{(\mu' - \mu)\bar{x}},$$

which is clearly a strictly increasing function of $\bar{x}$. This remains true for $\sigma^2 \neq 1$. Thus the Normal family has the MLRP for $\mu$ with respect to $\bar{x}$ for fixed $\sigma^2$. □

**19.7.2 Example** The Poisson($\mu$) pmf is

$$f(k; \mu) = e^{-\mu} \frac{\mu^k}{k!} \quad (K = 0, 1, 2, \ldots).$$

The sample mean $\bar{x} = (k_1 + \cdots + k_n)/n$ of an independent sample of size $n$ has pmf

$$f(\bar{x}; \mu) = \frac{1}{k_1! \cdots k_n!} e^{-n\mu} \mu^{n\bar{x}}$$

For $\mu < \mu'$, the likelihood ratio is

$$\lambda(\bar{x}) = \frac{1}{k_1! \cdots k_n!} e^{-n\mu'} \frac{\mu'^{n\bar{x}}}{\mu^{n\bar{x}}} = e^{-n(\mu' - \mu)} \left(\frac{\mu'}{\mu}\right)^{n\bar{x}},$$

which is a strictly increasing function of $\bar{x}$. □

**19.8 UMP Tests for MLRP Densities**

When $\theta$ is a one-dimensional parameter, we say that the null hypothesis is one-sided if it is of the form

$$H_0: \theta \leq \bar{\theta} \quad \text{so} \quad H_1: \theta > \bar{\theta},$$

(or if we reverse the sense of the inequalities).

When the density has the MLRP for the statistic $T$, then a Uniformly Most Powerful Test exists. The next result may be found in Lehmann [6, Theorem 2, p. 68].

**19.8.1 Theorem** Let $\Theta$ be one-dimensional, and assume the probability model $f(x; \theta)$ has monotone likelihood ratio in $T(x)$. For testing the null hypothesis $H_0: \theta \leq \bar{\theta}$ against the alternative $H_1: \theta > \bar{\theta}$, there is a UMP test with a critical region of the form $[c^*, \infty)$. That is, there is a critical value $c^*$, so that the test rejects $H_0$ if $T(x) > c^*$. 


The size of the test, by definition

\[ \alpha = \sup \{ P_{\theta_0} (T(X) > c^*) : \theta \leq \theta_0 \} \]

is achieved for \( \theta = \theta_0 \), that is,

\[ P_{\theta_0} (T(X) > c^*) = \alpha. \]

In addition, the power function (Larsen and Marx’s \( 1 - \beta(\theta) \)) is strictly increasing in \( \theta \) (up to the point where it becomes 1, and then it is constant).

**Sketch of proof**: • If we test the simple null \( \theta = \theta_0 \) against a simple alternative \( \theta = \theta' \), where \( \theta' > \theta \) the Neyman–Pearson Lemma tells us the most powerful test is a likelihood ratio test.

• Because the likelihood ratio is monotone in \( T \), the test takes the form:

\[ \text{Reject } \theta = \theta_0 \text{ against the alternative } \theta = \theta' \text{ if } T > c^* \]

for some critical value \( c^* \).

• Find \( c^* \) to give you the desired level of significance \( \alpha \).

• Now observe that because of the MLRP the same test specified by \( c^* \) is also a likelihood ratio test of the null \( \theta = \theta_0 \) against the simple alternative \( \theta = \theta'' \) for any \( \theta'' > \theta_0 \), and it also has significance level \( \alpha \). So by the Neyman–Pearson Lemma, it is the most powerful such test.

• This means that the test with critical value \( c^* \) is Uniformly Most Powerful for testing the simple null \( \theta = \theta_0 \) against the *composite* hypothesis \( \theta > \theta^* \).

• The MLRP also implies the test is UMP for the composite null. For \( \theta < \theta_0 \), we have \( P_{\theta} (T > c^*) \leq P_{\theta_0} (T > c^*) \).

The details are spelled out in \[6, p. 69\], but you can probably work them out yourself.

In this setting, there is another quantity of interest.

• Given such a test we reject \( H_0 \) if \( T > c^* \), where \( c^* \) is chosen so that

\[ P_{\theta_0} (T > c^*) = \alpha. \]

Suppose the test statistic \( T \) has the value \( t \). The probability

\[ P_{\theta_0} (T > t) \]

is called the \textbf{p-value} of \( t \).

An equivalent description of the test is to reject \( H_0 \) whenever the p-value of the statistic is less than \( \alpha \).
19.9 Likelihood Ratio Tests for composite hypotheses without MLRP

Likelihood ratio tests can also be used with composite hypotheses even in the absence of the MLRP. For testing the Null Hypothesis $H_0: \theta \in \Theta_0$ versus the Alternative Hypothesis $H_1: \theta \in \Theta_1$, let $\hat{\theta}_0$ be the maximum likelihood estimator of $\theta$ over $\Theta_0$ and let $\hat{\theta}_1$ be the maximum likelihood estimator of $\theta$ over $\Theta_1$. That is,

$$L(\hat{\theta}_0(x); x) = \max\{L(\theta; x) : \theta \in \Theta_0\}$$

and

$$L(\hat{\theta}_1(x); x) = \max\{L(\theta; x) : \theta \in \Theta_1\}$$

**Warning:** Because it is easier to warn you than to change the following section, and make sure that I change it correctly everywhere, in this section I am going to write the likelihood ratio the way Larsen and Marx do, not the way Lehmann does. Then the ratio

$$\lambda(x) = \frac{L(\hat{\theta}_0(x); x)}{L(\hat{\theta}_1(x); x)} = \frac{\max\{L(\theta; x) : \theta \in \Theta_0\}}{\max\{L(\theta; x) : \theta \in \Theta_1\}}$$

may serve as a test of the null hypothesis $H_0: \theta \in \Theta_0$ versus the alternative $H_1: \theta \in \Theta_1$.

Now $\lambda(x)$ depends on the sample $x$, and so is a random variable, which L&M call $\Lambda$. According to Larsen–Marx [5, Definition 6.5.1, p. 381] you should choose a critical value $\lambda^*$ so that the null hypothesis is rejected if

$$0 \leq \lambda(x) \leq \lambda^*.$$ 

They assert the significance level of such a test is given by the $\alpha$ such that

$$P\left(\Lambda \leq \lambda^* \mid H_0 \text{ is true}\right) = \alpha. \quad \text{(*)}$$

**19.9.1 Remark** When $\Theta_0$ consists of just a single just a single parameter value $\theta_0$, I know how to make sense of L&M’s statement (*): Compute the probability using $\theta_0$. When $\Theta_0$ has more than one possible parameter, then (*) is ambiguous at best, and meaningless at worst. Here is what they mean. For each $\theta$, there is a density $f(x; \theta)$ of $x$. Then for each $\theta$,

$$P_\theta\left(\lambda(x) \leq \hat{\lambda}\right) = \int_{\lambda(x) \leq \hat{\lambda}} f(x; \theta) \, dx = \alpha. \quad \text{(**)}$$

defines a critical value $\hat{\lambda}^*(\theta)$ that makes (**) true. The critical value that should be used for the test is most stringent one,

$$\lambda^* = \min\{\hat{\lambda}^*(\theta) : \theta \in \Theta_0\}.$$
That way, for every $\theta \in \Theta_0$ the probability of a Type I error is no more than $\alpha$. Remember the role of the null hypothesis is the hypothesis that you want to thoroughly trounce before you are willing to give it up.

Now if Larsen and Marx were Bayesians, (*) would be less problematic. We could use the posterior density on $\theta$ to average over the $P_\theta$s.

Larsen and Marx called this a **generalized likelihood ratio test**, but others may drop the “generalized.” Again, the usefulness of such test is not because we attach magical properties to the likelihood function, but because test constructed in this way usually have desirable properties.

**Bibliography**


