Optimal Tests of Hypotheses
(Hogg Chapter Eight)

STAT 406-01: Mathematical Statistics II *

Spring Semester 2016

Contents

Tuesday 26 April 2016
– Read Section 8.1 of Hogg

1 Most Powerful Tests

We turn now to another application of likelihood methods: as tests for hypotheses. First, a reminder of some of the notation and definitions associated with classical (frequentist) hypothesis testing.

1.1 Review of Hypothesis Testing

In the classical framework, we define a test to reject or not reject a null hypothesis $\mathcal{H}_0$ in the context of an alternative hypothesis $\mathcal{H}_1$, based on a realization $\mathbf{x}$ of the n-dimensional random vector $\mathbf{X}$ (which is often but not always a sample drawn from a univariate distribution). We refer to the support space for the whole sample as $\mathcal{S}$, i.e., $\mathbf{X} \in \mathcal{S}$. We can define the test as partitioning the sample space into a critical region C and its compelement.

If $\mathbf{x} \in C \subset \mathcal{S}$ we reject $\mathcal{H}_0$, while if $\mathbf{x} \notin C$ (i.e., $\mathbf{x} \in C^c$) we do not reject $\mathcal{H}_0$. (Strictly speaking, we should not refer to the outcome of the test as accepting one hypothesis or the other.) We often think of this in terms of a parametrized distribution $f_{\mathbf{X}}(\mathbf{x};\theta)$ and define $\mathcal{H}_0$ to specify a value $\theta = \theta_0$ (Hogg calls this θ') for the parameter and $\mathcal{H}_1$ to specify some other value $\theta = \theta_1$ (which Hogg calls θ''). Since each hypothesis only specifies a single value for θ, we call these *point hypotheses*, and the parameter formalism is actually unneccessary; $f_{\mathbf{X}}(\mathbf{x};\theta_0) = f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)$ and $f_{\mathbf{X}}(\mathbf{x};\theta_1) = f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)$ can just be thought of as two different probsbility distributions relevant to $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively.

We define the size α of the critical region, also known as the significance or false-alarm probability of the test as the probability that we will reject $\mathcal{H}_0$ if it is true:

$$\alpha = P(\mathbf{X} \in C|\mathcal{H}_0) = \int_C f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x \tag{1.1}$$

The power of the test is the probability that we'll reject $\mathcal{H}_0$ if $\mathcal{H}_1$ is true:

$$\gamma = P(\mathbf{X} \in C|\mathcal{H}_1) = \int_C f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \tag{1.2}$$

We can also write this in terms of $\beta = 1 - \gamma$, which is known as the *false dismissal probability*.

Given different tests with the same false-alarm probability α, we'd naturally rather use the one with the lowest false-dismissal probability $\beta = 1 - \gamma$, i.e., the highest power γ. This is known as the *most powerful test*.

1.2 The Neyman-Pearson Lemma

There is a theorem, usually known as the Neyman-Pearson lemma, that shows how the most powerful test to of one point hypothesis $\mathcal{H}_0$ against another $\mathcal{H}_1$ can be constructed from the likelihood ratio

$$\Lambda(\mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} = \frac{L(\theta_0;\mathbf{x})}{L(\theta_1;\mathbf{x})} \tag{1.3}$$

We define C so that $\mathbf{x} \in C$ if and only if $\Lambda(\mathbf{x}) \leq k$ where k is defined by

$$P(\Lambda(\mathbf{X}) \leq k|\mathcal{H}_0) = \int_C f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x = \alpha \tag{1.4}$$

which ensures that the critical region C is of size α. I.e., we reject $\mathcal{H}_0$ if and only if $\Lambda(\mathbf{x}) \leq k$.

The Neyman-Pearson lemma states that the power of this test

$$\gamma = P(\Lambda(\mathbf{X}) \leq k|\mathcal{H}_1) = \int_C f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \tag{1.5}$$

is greater than or equal to the power of any other test with the same significance. I.e., if A is some other critical region with size α, so that

$$\int_A f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x = \alpha \tag{1.6}$$

the Neyman-Pearson lemma says that

$$\gamma(C) = \int_C f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \geq \int_A f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \tag{1.7}$$

The demonstration of the Neyman-Pearson lemma involves breaking up the regions C and A in terms of their overlap $C \cap A$. Evidentally, we can write

$$C = (C \cap A^c) \cup (C \cap A) \tag{1.8a}$$
$$A = (C^c \cap A) \cup (C \cap A) \tag{1.8b}$$

The contribution to both α and γ from $C \cap A$ cancel out of any comparison between C and A. So the Neyman-Pearson lemma is equivalent to the condition that

$$\gamma(C) - \int\limits_{C\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x = \int\limits_{C\cap A^c} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \geq \int\limits_{C^c\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \quad (1.9)$$

If we can prove that, we've proved the lemma

Now, by definition

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} \leq k \qquad \text{for } \mathbf{x} \in C \quad (1.10)$$

so

$$\int\limits_{C\cap A^c} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \geq \frac{1}{k} \int\limits_{C\cap A^c} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x \quad (1.11)$$

while

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} \geq k \qquad \text{for } \mathbf{x} \in C^c \quad (1.12)$$

so

$$\int\limits_{C^c\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \leq \frac{1}{k} \int\limits_{C^c\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x \quad (1.13)$$

But since the tests defined by C and A both have the same significance α, the integrals of $f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)$ over the non-overlapping regions must be the same:

$$\alpha - \int\limits_{C\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x = \int\limits_{C\cap A^c} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x = \int\limits_{C^c\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)\, d^n x \quad (1.14)$$

so the right-hand sides of (1.13) and (1.11) must be equal, which means

$$\int\limits_{C\cap A^c} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \geq \int\limits_{C^c\cap A} f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \quad (1.15)$$

which, as we've argued above, means that the power of the likelihood ratio test defined by C is greater than or equal to that defined by A.

1.3 Example: Gamma Distribution

$f(x;\theta) = \frac{x^{5-1}}{\theta^5} e^{-x/\theta}$, $0 < x < \infty$. Let $\mathcal{H}_0 : \theta = 1$ and $\mathcal{H}_1 : \theta = 2$. Likelihood is

$$L(\theta;\mathbf{x}) = \frac{(\prod_{i=1}^n x_i)^4}{\theta^{5n}} \exp\left(-\sum_{i=1}^n x_i/\theta\right) \quad (1.16)$$

so likelihood ratio is

$$\Lambda = \frac{L(1;\mathbf{x})}{L(2;\mathbf{x})} = 2^{5n} \exp\left(-\left[1-\frac{1}{2}\right]\sum_{i=1}^n x_i\right) = 2^{5n} e^{-y/2} \quad (1.17)$$

where $y = \sum_{i=1}^n x_i$. Note that the likelihood ratio depends only on the sufficient statistic y. A moment's thought will show that this will be true in general. If there are one or more sufficient statistics for the parameter(s) which distinguish $\mathcal{H}_0$ from $\mathcal{H}_1$, the likelihood ratio must be a function of those alone.

Note that in this case, Y is Gamma(n,1) of $\mathcal{H}_0$ is true and Gamma(n,2) of $\mathcal{H}_1$ is true, so we can figure out the significance and power of the test by using the percentiles of the Gamma (or equivalently chi-square) distribution.

Thursday 28 April 2016
– Read Section 8.2 of Hogg

1.4 Uniformly Most Powerful Tests

The Neyman-Pearson Lemma shows that when comparing point hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, a test based on the likelihood ratio $f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)/f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)$ gives the most powerful test (highest $\gamma = P(\mathbf{X} \in C|\mathcal{H}_1)$) at a given significance $\alpha = P(\mathbf{X} \in C|\mathcal{H}_0)$. We often wish to compare *composite hypotheses*, in which $\mathcal{H}_1$ and/or $\mathcal{H}_0$ can correspond to parametrized families of distributions. In particular, if $\mathbf{X}$ is a sample drawn from a distribution $f(x;\theta)$, we may have hypotheses $\mathcal{H}_0 : \theta \in \omega$ and $\mathcal{H}_1 : \theta \in \Omega$.

If we first consider the case where $\mathcal{H}_0$ is a point hypothesis $\theta = \theta_0$ and $\mathcal{H}_1$ is a composite hypothesis parametrized by θ, the power of a given test will depend on θ:

$$\gamma(\theta) = P(\mathbf{X} \in C;\theta) = \int_C f_{\mathbf{X}}(\mathbf{x};\theta) d^n x \tag{1.18}$$

If the same test is most powerful at a given α for every choice of θ, we say it is *uniformly most powerful* (UMP). For example, recall the example from the last class, where $\mathbf{X}$ is a sample of size n from a Gamma$(5,\theta)$ distribution and $\mathcal{H}_0$ is $\theta = \theta_0 = 1$. Now let $\mathcal{H}_1$ be the composite hypothesis $\theta > 1$. We know from the Neyman-Pearson lemma that at each θ the most powerful test will be the one given by the likelihood ratio:

$$C : \Lambda(\mathbf{x};\theta) = \frac{L(\theta_0;\mathbf{x})}{L(\theta;\mathbf{x})} \leq k(\theta) \tag{1.19}$$

where the constant $k(\theta)$ is defined by

$$P(\Lambda(\mathbf{X};\theta) \leq k(\theta)|\mathcal{H}_0) = \alpha \tag{1.20}$$

1.4.1 Example: One-Sided Hypothesis

We saw last time that

$$L(\theta;\mathbf{x}) = \frac{\prod_{i=1}^n x_i^4}{\theta^{5n}} e^{-y/\theta} \tag{1.21}$$

where y is the sufficient statistic $y = \sum_{i=1}^n x_i$, and Y is a Gamma$(5n,\theta)$ random variable. The likelihood ratio is

$$\begin{aligned}\Lambda(\theta;\mathbf{x}) = \frac{L(\theta_0;\mathbf{x})}{L(\theta;\mathbf{x})} &= \left(\frac{\theta}{\theta_0}\right)^{5n} \exp\left(-\left[\frac{1}{\theta_0} - \frac{1}{\theta}\right] y\right) \\ &= \theta^{5n} \exp\left(-\frac{\theta - 1}{\theta} y\right)\end{aligned} \tag{1.22}$$

As long as $\theta > 1$, this is a decreasing function of y, so a test which rejects $\mathcal{H}_0$ if $\Lambda \leq k(\theta)$ will be equivalent to one which does so if $Y \geq a$ for some corresponding a. This can be defined so that the critical region is of size α:

$$\alpha = P(Y \geq a|\mathcal{H}_0) = \gamma(\theta_0) = \int_a^\infty \frac{y^{5n-1}}{\Gamma(5n)\theta_0^{5n}} e^{-y/\theta_0}\, dy = \frac{\Gamma(5n,a)}{\Gamma(5)} \tag{1.23}$$

where

$$\Gamma(s,x) = \int_x^\infty t^{s-1}\, e^{-t}\, dt \tag{1.24}$$

is the upper incomplete Gamma function. In any event, the value of a doesn't depend on θ, as long as $\theta > 1$, so $Y \geq a$ is the UMP test of $\mathcal{H}_0$ in light of the composite hypothesis $\mathcal{H}_1$.

1.4.2 Example: Two-Sided Hypothesis

Note that there does not always exist a UMP test. Suppose instead the hypothesis $\mathcal{H}_1$ is $\theta \neq \theta_0 = 1$. For $\theta > 1$, the most

powerful test is to reject $\mathcal{H}_0$ if $Y \geq a$ as defined above. For $0 < \theta < 1$, the likelihood ratio

$$\Lambda(\theta; \mathbf{x}) = \theta^{5n} \exp\left(\frac{1-\theta}{\theta}\, y\right) \tag{1.25}$$

is a monotonically increasing function of y, so $\Lambda \leq k(\theta)$ is equivalent to $Y \leq b$, where b is defined by

$$\alpha = P(Y \leq b | \mathcal{H}_0) = \gamma(\theta_0) = \int_{-\infty}^{b} \frac{y^{5n-1}}{\Gamma(5n)\theta_0^{5n}} e^{-y/\theta_0}\, dy = \frac{\Gamma(5n) - \Gamma(5n, b)}{\Gamma(5n)} \tag{1.26}$$

where

$$\Gamma(s) - \Gamma(s, x) = \int_{-\infty}^{x} t^{s-1}\, e^{-t}\, dt \tag{1.27}$$

is the lower incomplete Gamma function. Thus the most powerful test depends on the value of θ: rejecting $\mathcal{H}_0$ when $Y \geq a$ is most powerful if $\theta > \theta_0$, while rejecting $\mathcal{H}_1$ when $Y \leq b$ is most powerful if $\theta < \theta_0$,

To illustrate, we'll plot the power function for the two tests assuming $\alpha = 0.10$ and $n = 4$. We'll get the percentiles and tail probabilities for the Gamma distribution using the SciPy stats package. For any distribution, the method `sf()` is the survival function (one minus the cdf), `cdf()` is the cdf, and `isf()` and `ppf()` are the inverses of these functions.

```
In [1]: from scipy.stats import gamma as gammadist

In [2]: alpha = 0.10

In [3]: n = 4

In [4]: a = gammadist(5*n).isf(alpha)

In [5]: b = gammadist(5*n).ppf(alpha)
```

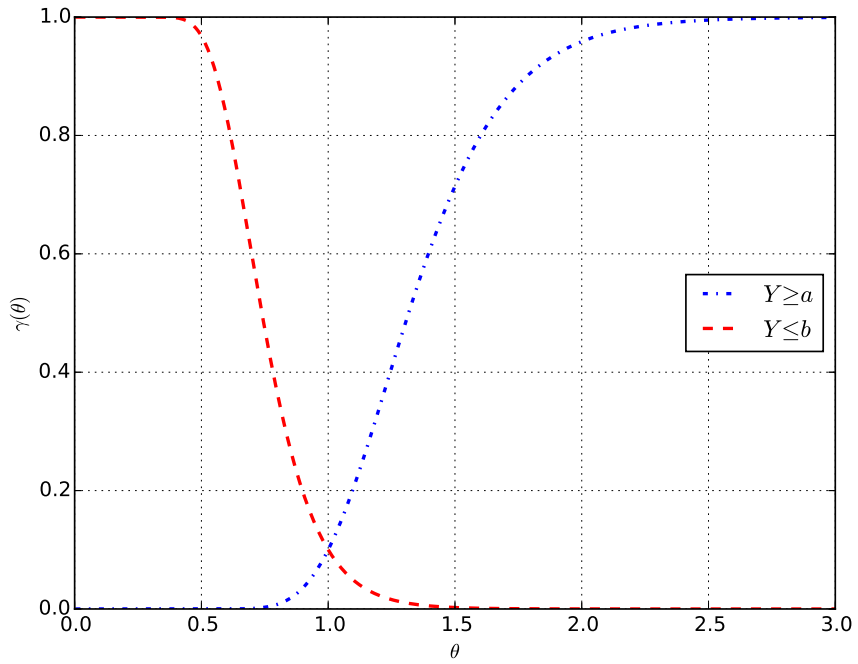
We check that the significance of the two tests is indeed 10%:

```
In [6]: print gammadist(5*n).sf(a)
0.1

In [7]: print gammadist(5*n).cdf(b)
0.1

In [8]: theta = linspace(0,3,1000)[1:]
```

The `[1:]` is to omit the first element in the array, so that we're not actually trying to use $\theta = 0$ as a value of the scale parameter.

```
In [9]: gamma_a = gammadist(5*n,scale=theta).sf(a)

In [10]: gamma_b = gammadist(5*n,scale=theta).cdf(b)

In [11]: figure();

In [12]: plot(theta,gamma_a,'b-',label=r'$Y\geq a$');

In [13]: plot(theta,gamma_b,'r--',lw=2,label=r'$Y\leq b$');

In [14]: legend(loc='center right');

In [15]: xlabel(r'$\theta$');

In [16]: ylabel(r'$\gamma(\theta)$');

In [17]: grid(True)

In [18]: savefig('notes08_power.eps',bbox_inches='tight')
```

Here are the two power curves:



Note that they intersect at $\gamma(1) = 0.1$, which makes sense since any test with significance α will satisfy $\gamma(\theta_0) = \alpha$. And as we can see, the power of the test $Y \geq a$ (solid blue) is higher for $\theta > 1$, while the power of the test $Y \leq b$ (dashed red) is higher for $\theta < 1$.

1.4.3 Example: Composite Null Hypothesis

Returning to the case where $\mathcal{H}_1$ is $\theta > 1$, we can extend the discussion to comparison between two composite hypotheses by letting $\mathcal{H}_0$ be $\theta \leq 1$. As a matter of convention, we define the significance α for a test with a composite null hypothesis as the worst-case false alarm rate:

$$\alpha = \max_{\theta \in \omega} P(\mathbf{X} \in C; \theta) \tag{1.28}$$

If we consider the test $Y \geq a$,

$$P(\mathbf{X} \in C; \theta) = \int_a^\infty \frac{y^{5n-1}}{\Gamma(5n)\theta^{5n}} e^{-y/\theta}\, dy = \int_{a/\theta}^\infty \frac{t^{5n-1}}{\Gamma(5n)} e^{-t}\, dt \tag{1.29}$$

Since the integrand is positive definite and independent of θ, we see that we can maximize the integral by making θ as large as possible subject to $\mathcal{H}_0$ (so that the lower limit a/θ is as small as possible and the integral is maximized), i.e., $\theta = \theta_0 = 1$. So the sigificance of the test for null hypothesis $\mathcal{H}_0 : \theta \leq \theta_0 = 1$ is the same as for $\mathcal{H}_0 : \theta = \theta_0 = 1$, and the rest of the problem proceeds as before, i.e., $Y \geq a$ defines the UMP for $\mathcal{H}_0 : \theta \leq 1$ vs $\mathcal{H}_1 : \theta > 1$.

1.4.4 The Karlin-Rubin Theorem

This is example is an illustration of a result known as the Karlin-Rubin theorem, which extends the Neyman-Pearson lemma to tests of one-sided composite hypotheses. If the likelihood function $L(\theta; \mathbf{x})$ can be written in terms of a single sufficient statistic $y = u(\mathbf{x})$ and has the property that, for any θ_0 and θ_1 in the parameter space and satisfying $\theta_0 < \theta_1$, the likelihood ratio

$$\Lambda(\mathbf{x}; \theta_0, \theta_1) = \frac{L(\theta_0; \boldsymbol{\theta})}{L(\theta_1; \boldsymbol{\theta})} \tag{1.30}$$

is a monotonically increasing or monotonically decreasing function of y, then there exists a uniformly most powerful test for $\mathcal{H}_0 : \theta \leq \theta'$ versus $\mathcal{H}_1 : \theta > \theta'$. We call this situation a *monotone likelihood ratio* (mlr).

1.4.5 Example: Exponential Family

The conditions for a monotone likelihood ratio seem a little restrictive, but there is a decent class of models which satisfy them.

If we consider a distribution which is a member of the regular exponential family, so

$$f(x;\theta) = \exp\left(\eta(\theta)K(x) + H(x) + q(\theta)\right) \tag{1.31}$$

then the likelihood for a sample of size n is

$$L(\theta;\mathbf{x}) = \exp\left(\eta(\theta)\sum_{i=1}^{n} K(x_i) + \sum_{i=1}^{n} H(x_i) + nq(\theta)\right) \tag{1.32}$$

and the likelihood ratio is

$$\Lambda(\mathbf{x};\theta_0,\theta_1) = \frac{L(\theta_0;\boldsymbol{\theta})}{L(\theta_1;\boldsymbol{\theta})} = \exp\left([\eta(\theta_0) - \eta(\theta_1)]y + n[q(\theta_0) - q(\theta_1)]\right) \tag{1.33}$$

where the dependence on the data is via the sufficient statistic $y = \sum_{i=1}^{n} K(x_i)$. We see that:

- If $\eta(\theta)$ is a monotonically increasing function of θ, then $\eta(\theta_0) - \eta(\theta_1) < 0$ for $\theta_0 < \theta_1$, and $\Lambda(\mathbf{x};\theta_0,\theta_1)$ is a monotonically decreasing function of y.
- If $\eta(\theta)$ is a monotonically decreasing function of θ, then $\eta(\theta_0) - \eta(\theta_1) > 0$ for $\theta_0 < \theta_1$, and $\Lambda(\mathbf{x};\theta_0,\theta_1)$ is a monotonically increasing function of y.

I.e., if $\eta(\theta)$ is a monotone function of θ, we have a monotone likelihood ratio, and the Karlin-Rubin theorem tells us there's a UMP test to distinguish between one-sided hypotheses.

Tuesday 3 May 2016
– Read Section 8.3 of Hogg

2 Likelihood Ratio Tests and UMP Unbiased Tests

2.1 Gamma Example Continued

Recall that last time, we saw that, for a sample of size n drawn from a Gamma$(5,\theta)$ distribution, a test which rejected $\mathcal{H}_0 : \theta = \theta_0 = 1$ if $\sum_{i=1}^{n} X_i \geq a$ was uniformly most powerful for comparing $\mathcal{H}_0$ to the one-sided alternative hypothesis $\theta > \theta_0 = 1$, while one with rejected $\mathcal{H}_0$ if $\sum_{i=1}^{n} X_i \leq b$ was UMP if the alternative hypothesis was $\theta < \theta_0 = 1$, but if the alternative hypothesis was $\mathcal{H}_1 = \theta \neq \theta_0 = 1$, there was no single test which was most powerful for every θ allowed under $\mathcal{H}_1$.

One method we've considered before for comparing composite hypotheses is a likelihood ratio test, where we maximize the likelihood subject to each hypothesis. In this case, where the null hypothesis is still the point hypothesis $\theta = \theta_0$, this becomes

$$\Lambda(\mathbf{x}) = \frac{L(\theta_0;\mathbf{x})}{L(\widehat{\theta}(\mathbf{x}),\mathbf{x})} \tag{2.1}$$

specifically in this example,

$$L(\theta;\mathbf{x}) = \frac{\prod_{i=1}^{n} x_i^4}{[\Gamma(5)]^n \theta^{5n}} e^{-y/\theta} \tag{2.2}$$

and the maximum likelihood solution $\widehat{\theta}$ is the solution to

$$0 = \frac{\partial}{\partial\theta}\left(-5n\ln\theta - \frac{y}{\theta}\right) = -\frac{5n}{\theta} + \frac{y}{\theta^2} \tag{2.3}$$

i.e., $\widehat{\theta} = \frac{y}{5n}$, which makes the likelihood ratio

$$\Lambda(\mathbf{x}) = \frac{L(1;\mathbf{x})}{L(\frac{y}{5n},\mathbf{x})} = \left(\frac{y}{5n}\right)^{5n} e^{-y+5n} \tag{2.4}$$

Since this goes to zero as y goes to zero or infinity, a test which rejects $\mathcal{H}_0$ if $\Lambda(\mathbf{x}) \leq k$ will be a two-sided test on Y, rejecting if

$$Y \leq c_1 \qquad \text{or} \qquad Y \geq c_2 \tag{2.5}$$

where $y = c_1$ and $y = c_2$ correspond to equal values of the likelihood, i.e.,

$$c_1^{5n} e^{-c_1} = c_2^{5n} e^{-c_2} \tag{2.6}$$

which, together with the specified significance α determines c_1 and c_2 via

$$1 - \alpha = \int_{c_1}^{c_2} f_Y(y;1)\,dy = \frac{1}{\Gamma(5n)} \int_{c_1}^{c_2} y^{5n-1}\, e^{-y}\, dy \tag{2.7}$$

Note that this likelihood ratio test is slightly different from the equal-tailed test defined by

$$\int_0^{c_1^{\text{sym}}} f_Y(y;1)\,dy = \frac{\alpha}{2} = \int_{c_2^{\text{sym}}}^{\infty} f_Y(y;1)\,dy \tag{2.8}$$

which is slightly easier to construct. (c_1^{sym} and c_2^{sym} can be written as the 100αth and $100(1-\alpha)$th percentiles of the Gamma$(5n, 1)$ distribution which applies to Y under the null hypothesis $\mathcal{H}_0$.) If the distribution function $f_Y(y;1)$ and the likelihood ratio $\Lambda(\mathbf{x})$ were symmetric functions in y, of course, the likelihood ratio test would be equal-tailed. This is the case in several of the examples in Hogg.

We can examine the power functions of the different options (the two one-sided tests, the equal-tailed test, and the likelihood ratio test) by continuing our plotting example from last time:

```
In [19]: c1_sym = gammadist(5*n).isf(0.5*alpha)

In [20]: c2_sym = gammadist(5*n).ppf(0.5*alpha)

In [21]: gamma_sym = (
   ....: gammadist(5*n,scale=theta).cdf(c2_sym)
   ....: + gammadist(5*n,scale=theta).sf(c1_sym) )

In [22]: plot(theta,gamma_sym,'k:',lw=2,
   ....: label=r'$Y\leq c_1^{\rm sym}$ '
   ....: + r'or $Y\geq c_2^{\rm sym}$');
```

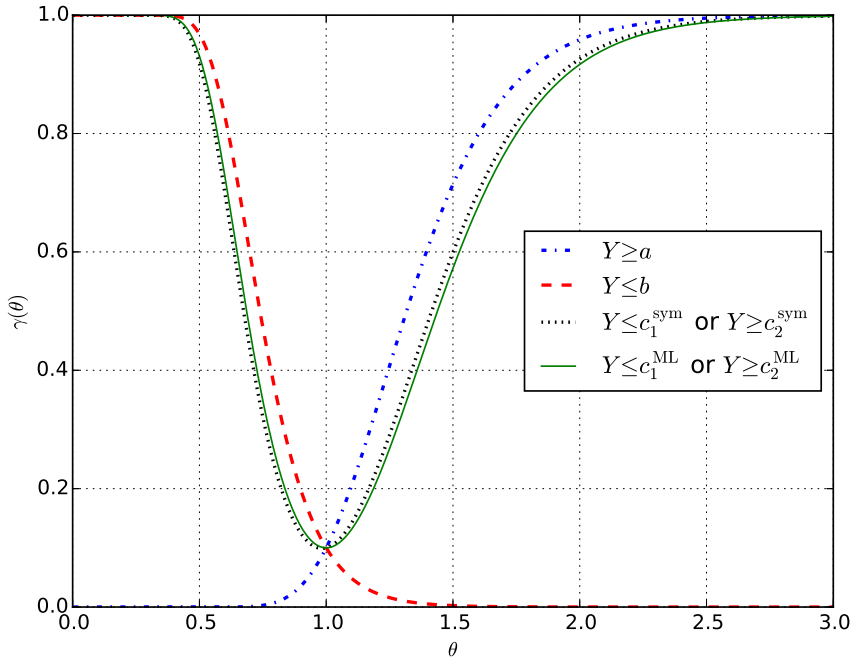
For the maximum likelihood test, we have to find the test which satisfies (2.6) from among the choices for which $P(Y \leq c_1|\mathcal{H}_0) + P(Y \geq c_2|\mathcal{H}_0) = \alpha$:

```
In [23]: alphaleft = linspace(0,alpha,1000)[1:-1]

In [24]: c1 = gammadist(5*n).ppf(alphaleft)

In [25]: c2 = gammadist(5*n).isf(alpha-alphaleft)
```

`c1` and `c2` are both 998-element arrays giving the lower and uppwe end of the range of y values for which we should not reject $\mathcal{H}_0$. Rather than doing some sophisticated root-finding, we just brute-force it and pick the pair which has the lowest value of $|c_1^{5n}e^{-c_1} - c_2^{5n}e^{-c_2}|$:

```
In [26]: ind_ML = argmin(abs(c1**(5*n)*exp(-c1)
   ....: -c2**(5*n)*exp(-c2)))

In [27]: c1_ML = c1[ind_ML]

In [28]: c2_ML = c2[ind_ML]
```

```
In [29]: gamma_ML = (
    ....: gammadist(5*n,scale=theta).cdf(c1_ML)
    ....: + gammadist(5*n,scale=theta).sf(c2_ML) )

In [30]: plot(theta,gamma_ML,'g-',
    ....: label=r'$Y\leq c_1^{\rm ML}$ '
    ....: + r'or $Y\geq c_2^{\rm ML}$');

In [31]: legend(loc='center right');

In [32]: savefig('notes08_ML.eps',bbox_inches='tight')
```



We see that the equal-tailed test is slightly different from the maximum likelihood test, thanks to the asymmetric distribution, but it's close. Both have the property of having a power function that goes to unity as you go away from $\theta = \theta_0 - 1$ in either direction, as opposed to the one-sided tests, whose power function goes to zero on the "wrong" side of $\theta = \theta_0$.

2.2 Unbiased Tests

One of the unsatisfactory properties of the one-sided hypothesis tests is that while they are the most powerful tests for some ranges of the parameter θ, for other ranges, the power is actually below the false alarm probability α. For instance, the test which rejects $\mathcal{H}_0$ when $Y \geq a$ has $\gamma(\theta) < \alpha$ when $\theta < \theta_0 = 1$. This means the test is more likely to reject $\mathcal{H}_0$ when it's true than when $\mathcal{H}_1$ is true, if it happens that $\theta < 1$. A test with this undesirable property is called biased. Conversely, an *unbiased test* is one for which $\gamma(\theta) \geq \alpha$ for all θ allowed by $\mathcal{H}_1$.

It is often the case that, while no uniformly most powerful test exists, there is an unbiased test which, at all values of θ, is more powerful than any other unbiased test. This is known as the *uniformly most powerful unbiased test*. It can be shown that in many cases (specifically the regular exponential family), the UMPU test is a two-sided test on the sufficient statistic Y, which rejects $\mathcal{H}_0$ if $Y \leq c_1$ or $Y \geq c_2$.

2.2.1 Illustration for Gamma example

Returning to the example of a sample from the Gamma$(5, \theta)$ distribution, consider a test which satisfies $\gamma(\theta) \geq \alpha$ when $\theta \neq \theta_0 = 1$. Since $\gamma(\theta_0) = \alpha$, that means the power function $\gamma(\theta)$ must have a minimum at $\theta = \theta_0$. Recalling that

$Y \sim \text{Gamma}(5n, \theta)$, we have

$$\begin{aligned}\gamma(\theta) &= 1 - \int_{c_1}^{c_2} f_Y(y;\theta)\,dy = 1 - \frac{1}{\Gamma(5n)} \int_{c_1}^{c_2} \frac{y^{5n-1}\,e^{-y/\theta}\,dy}{\theta^{5n}} \\ &= 1 - \frac{1}{\Gamma(5n)} \int_{c_1/\theta}^{c_2/\theta} t^{5n-1}\,e^{-t}\,dt \end{aligned} \tag{2.9}$$

where we have made the substitution $t = y/\theta$ in the last step. Requiring this to be a minimum at $\theta = \theta_0$ gives us

$$0 = \gamma'(\theta_0) = \frac{1}{\Gamma(5n)} \left(\frac{c_2}{\theta_0^2} c_2^{5n-1}\,e^{-c_2} - \frac{c_1}{\theta_0^2} c_1^{5n-1}\,e^{-c_1} \right) \tag{2.10}$$

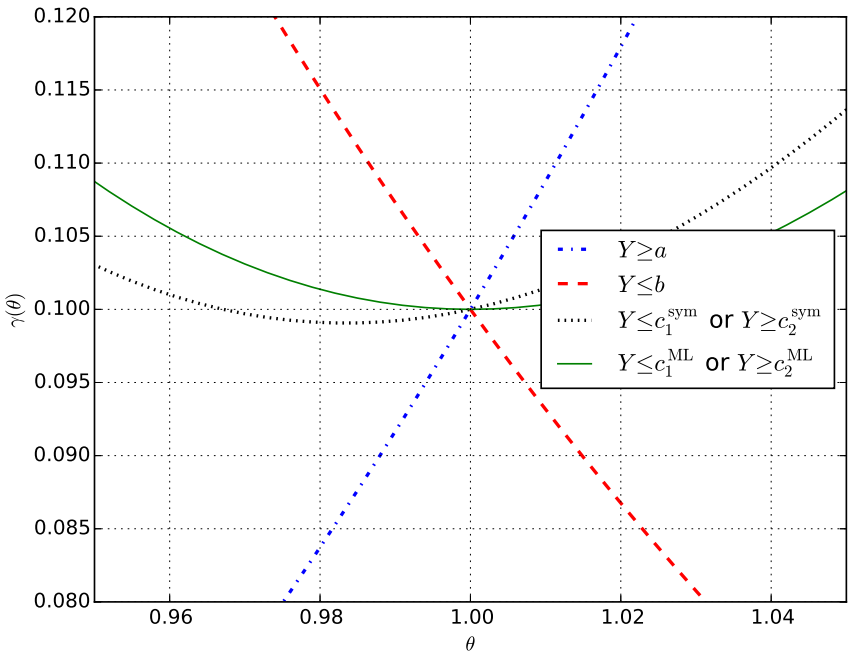
which gives us, using $\theta_0 = 1$,

$$c_2^{5n}\,e^{-c_2} = c_1^{5n}\,e^{-c_1} \tag{2.11}$$

If we look back, we see this is just the condition (2.6) associated with the likelihood ratio test, so that test is in fact the UMP unbiased test for this model.

We can verify this by zooming in on the power function plot above:

```
In [33]: xlim([.95,1.05]);

In [34]: ylim([0.08,0.12]);

In [35]: savefig('notes08_MLzoom.eps',bbox_inches='tight')
```



We see that the equal-tailed test has a power function which dips below $\alpha = 0.10$ for a small range of θ values, so in fact that test is not unbiased. We see that the likelihood ratio test is indeed unbiased, and it turns out to tbe the UMPU test.

Thursday 5 May 2016
– See Searle, http://arxiv.org/abs/0804.1161

3 Composite Hypothesis Testing with Priors

Consider now a slightly different situation. Suppose that the null hypothesis $\mathcal{H}_0$ is a point hypothesis, but the alternative hypothesis $\mathcal{H}_1$ is a composite hypothesis which allows for a range

of values of the model parameter(s) $\boldsymbol{\theta}$, but comes with a prior distribution $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)$ on those parameters. (We include the possibility of a p-dimensional parameter space, but it may also be that $p = 1$.) If we define a test which rejects $\mathcal{H}_0$ when

$$\frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} = \frac{L(\boldsymbol{\theta}_0;\mathbf{x})}{\int L(\boldsymbol{\theta};\mathbf{x})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)\, d^p\theta} \leq k \tag{3.1}$$

the Neyman-Pearson lemma tells us this is the most powerful test of $\mathcal{H}_0$ versus $\mathcal{H}_1$. This is "most powerful" in the sense of maximizing the power function

$$\begin{aligned}\gamma(\mathcal{H}_1) = P(\mathbf{X} \in C|\mathcal{H}_1) &= \int_C f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)\, d^n x \\ &= \int_C \int L(\boldsymbol{\theta};\mathbf{x})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)\, d^p\theta\, d^n x = \int \gamma(\boldsymbol{\theta})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)\, d^p\theta\end{aligned} \tag{3.2}$$

A few points to note:

- The test statistic in (3.1) is just the Bayes factor which we've already motivated using Bayes's theorem in the form

$$P(\mathcal{H}_i|\mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_i)\, P(\mathcal{H}_i)}{f_{\mathbf{X}}(\mathbf{x})} \tag{3.3}$$

  to write

$$\frac{P(\mathcal{H}_0|\mathbf{x})}{P(\mathcal{H}_1|\mathbf{x})} = \frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)} \tag{3.4}$$

- If there is a uniformly most powerful test which covers any $\boldsymbol{\theta}$ in the support of $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)$, this will also be the most powerful test of $\mathcal{H}_0$ against $\mathcal{H}_1$ for any prior distribution.
- A possible objection is that the frequentist hypothesis testing formalism only applies to the outcomes of repeated experiments, and isn't supposed to know about any prior distribution. The preprint by Searle referenced above actually refers to the outcome of a Monte Carlo experiment, where $\boldsymbol{\theta}$ is also randomly generated along with the realization of $\mathbf{X}$, so the relevant joint distribution is $f_{\mathbf{X}\boldsymbol{\Theta}}(\mathbf{x}, \boldsymbol{\theta}|\mathcal{H}_1) = f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)$. In that context, the Bayes factor constructed using the same prior as the Monte Carlo simulation gives the most powerful test, when attention is restricted to tests which define C using only $\mathbf{X}$ and not $\boldsymbol{\Theta}$.

3.1 Example: Mean of a Normal Distribution

Suppose that the data vector is $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ but rather than being a sample from a given distribution, they are independent random variables $X_1 \sim N(\theta_1, 1)$ and $X_2 \sim N(\theta_2, 1)$. Let the null hypothesis be $\mathcal{H}_0 : \theta_1 = 0 = \theta_2$ and the alternative hypothesis be $\mathcal{H}_1 : \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. The likelihood function is

$$L(\boldsymbol{\theta};\mathbf{x}) = \frac{1}{2\pi} \exp\left(-\frac{(x_1-\theta_1)^2}{2} - \frac{(x_2-\theta_2)^2}{2}\right) \tag{3.5}$$

The maximum likelihood solution is $\widehat{\theta}_1 = x_1$ and $\widehat{\theta}_2 = x_2$, so the likelihood ratio test would reject $\mathcal{H}_0$ if

$$\Lambda(\mathbf{x}) = \frac{L(\mathbf{0},\mathbf{x})}{L(\widehat{\boldsymbol{\theta}},\mathbf{x})} = \exp\left(-\frac{x_1^2 + x_2^2}{2}\right) \leq k \tag{3.6}$$

which means, for a test of significance α, rejecting $\mathcal{H}_0$ if

$$X_1{}^2 + X_2{}^2 \geq \chi^2_{2,\alpha} \tag{3.7}$$

On the other hand, if we have a prior $f_{\boldsymbol{\Theta}}(\theta_1, \theta_2|\mathcal{H}_1)$ the Neyman-Pearson lemma tells us the optimal test statistic is the Bayes factor

$$\frac{\exp\left(-\frac{x_1^2+x_2^2}{2}\right)}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\boldsymbol{\Theta}}(\theta_1,\theta_2|\mathcal{H}_1) \exp\left(-\frac{(x_1-\theta_1)^2}{2} - \frac{(x_2-\theta_2)^2}{2}\right) d\theta_1\, d\theta_2} \tag{3.8}$$

3.1.1 Case 1: Uniform Prior

Suppose that the prior $f_{\boldsymbol{\Theta}}(\theta_1, \theta_2|\mathcal{H}_1)$ is uniform in θ_1 and θ_2.[1] In that case the denominator is proportional to

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{(x_1-\theta_1)^2}{2} - \frac{(x_2-\theta_2)^2}{2}\right) d\theta_1\, d\theta_2 = \text{constant} \tag{3.9}$$

and the Bayes factor is proportional to the maximum-likelihood ratio

$$\exp\left(-\frac{x_1^2+x_2^2}{2}\right) \tag{3.10}$$

so using $x_1^2 + x_2^2$ as a test statistic does give the optimal test in light of the prior.

3.1.2 Case 2: Non-Uniform Prior

It could be that the prior on θ_1 and θ_2 is more complicated.[2] For instance, suppose $f_{\boldsymbol{\Theta}}(\theta_1, \theta_2) \propto (\theta_1\theta_2)^2$. Then the denominator of the Bayes factor is proportional to

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \theta_1^2\theta_2^2 \exp\left(-\frac{(x_1-\theta_1)^2}{2} - \frac{(x_2-\theta_2)^2}{2}\right) d\theta_1\, d\theta_2 \propto (1+x_1^2)(1+x_2^2) \tag{3.11}$$

which means the Bayes factor is proportional to

$$\frac{\exp\left(-\frac{x_1^2+x_2^2}{2}\right)}{(1+x_1^2)(1+x_2^2)} \tag{3.12}$$

and the optimal test rejects $\mathcal{H}_0$ if

$$X_1{}^2 + X_2{}^2 + 2\ln(1+X_1{}^2) + 2\ln(1+X_2{}^2) \geq c \tag{3.13}$$

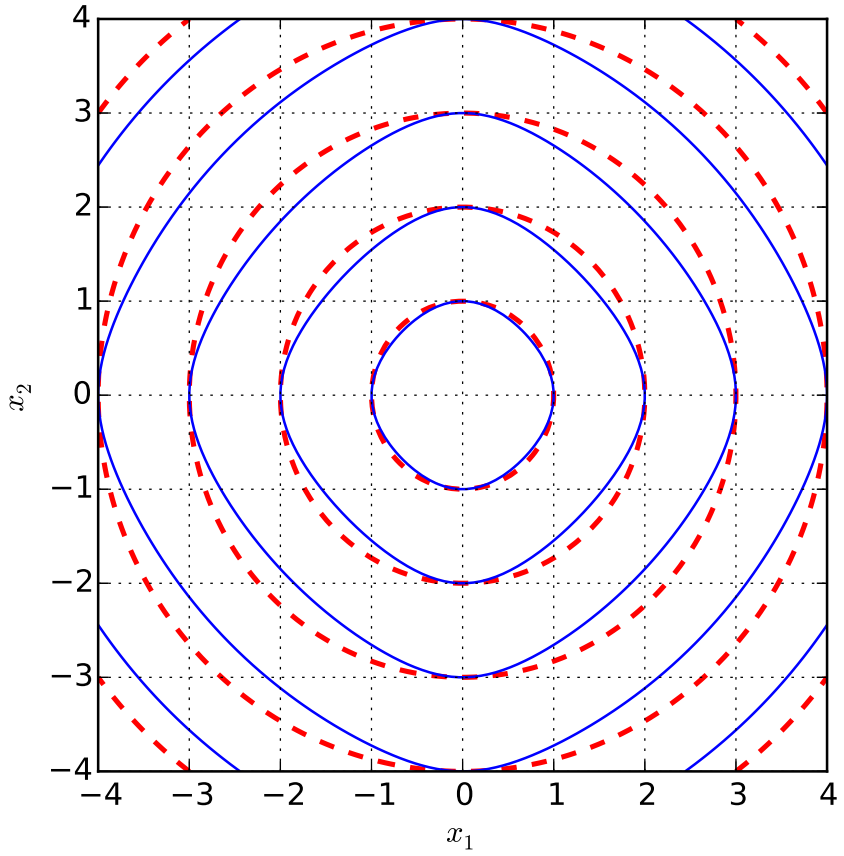
We can show the difference between the two families of critical regions on a contour plot:

```
In [1]: x = linspace(-4,4,100)

In [2]: x1grid, x2grid = meshgrid(x,x)

In [3]: Fgrid = x1grid**2 + x2grid**2

In [4]: Bgrid = ( Fgrid
   ...: + 2*log(1+x1grid**2) + 2*log(1+x2grid**2) )

In [5]: xlevels = arange(6)

In [6]: Blevels = xlevels**2 + 2*log(1+xlevels**2)

In [7]: Flevels = xlevels**2

In [8]: figure(figsize=(5,5));

In [9]: contour(x1grid,x2grid,Fgrid,Flevels,colors='r',
   ...: linewidths=2,linestyles='dashed');

In [10]: contour(x1grid,x2grid,Bgrid,Blevels,colors='b');

In [11]: grid(True)
```

[1] This is an improper prior, so technically the normalization factor will be zero and Bayes factor will be infinite, but we're only interested in the $\mathbf{x}$ dependence of the Bayes factor anyway. If we want to be careful, we can use a prior that is e.g., Gaussian with some large width, much greater than any of the other values in the problem.

[2] E.g., for the gravitational wave signal from a rotating or orbiting system, the amplitudes of the left- and right-circularly polarized parts of the signal are proportional to $\left(\frac{1\pm\cos\iota}{2}\right)^2$ where ι is an inclination angle, which is generally unknown with a uniform prior on $\cos\iota \in [-1,1]$, which produces a non-trivial prior distribution for the two amplitudes. See Whelan et al, *Classical and Quantum Gravity* **31**, 065002 (2014); http://arxiv.org/abs/1311.0065.

```
In [12]: xlabel(r'$x_1$');

In [13]: ylabel(r'$x_2$');

In [14]: savefig('notes08_contours.eps',
    ....: bbox_inches='tight')
```



Note that we drew the contours at arbitrary levels; to draw them at the same significance for the two tests, we'd probably need to estimate the significance of the Bayes-factor test numerically. (See figure 3 of Whelan et al 2014 for an example of this.)

Tuesday 10 May 2016 – Review for Final Exam

The exam is comprehensive, but with relatively more emphasis on chapter eight and the second half of chapter seven. Please come with questions and topics you'd like to go over.

Thursday 12 May 2016 – Review for Final Exam

The exam is comprehensive, but with relatively more emphasis on chapter eight and the second half of chapter seven. Please come with questions and topics you'd like to go over.