Hypothesis Testing

Hypothesis testing is a statistical problem where you must choose, on the basis of data X, between two alternatives. We formalize this as the problem of choosing between two *hypotheses*: $H_o : \theta \in \Theta_0$ or $H_1 : \theta \in \Theta_1$ where Θ_0 and Θ_1 are a partition of the model $P_\theta; \theta \in \Theta$. That is $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

A rule for making the required choice can be described in two ways:

1. In terms of the set

$$C = \{X : \text{we choose } \Theta_1 \text{ if we observe } X\}$$

   called the *rejection* or *critical* region of the test.

2. In terms of a function $\phi(x)$ which is equal to 1 for those x for which we choose Θ_1 and 0 for those x for which we choose Θ_0.

For technical reasons which will come up soon I prefer to use the second description. However, each ϕ corresponds to a unique rejection region $R_\phi = \{x : \phi(x) = 1\}$.

The Neyman Pearson approach to hypothesis testing which we consider first treats the two hypotheses asymmetrically. The hypothesis H_o is referred to as the *null* hypothesis (because traditionally it has been the hypothesis that some treatment has no effect).

Definition: The power function of a test ϕ (or the corresponding critical region R_ϕ) is

$$\pi(\theta) = P_\theta(X \in R_\phi) = E_\theta(\phi(X))$$

We are interested here in **optimality** theory, that is, the problem of finding the best ϕ. A good ϕ will evidently have $\pi(\theta)$ small for $\theta \in \Theta_0$ and large for $\theta \in \Theta_1$. There is generally a trade off which can be made in many ways, however.

Simple versus Simple testing

Finding a best test is easiest when the hypotheses are very precise.

Definition: A hypothesis H_i is **simple** if Θ_i contains only a single value θ_i.

The simple versus simple testing problem arises when we test $\theta = \theta_0$ against $\theta = \theta_1$ so that Θ has only two points in it. This problem is of importance as a technical tool, not because it is a realistic situation.

Suppose that the model specifies that if $\theta = \theta_0$ then the density of X is $f_0(x)$ and if $\theta = \theta_1$ then the density of X is $f_1(x)$. How should we choose ϕ? To answer the question we begin by studying the problem of minimizing the total error probability.

We define a **Type I error** as the error made when $\theta = \theta_0$ but we choose H_1, that is, $X \in R_\phi$. The other kind of error, when $\theta = \theta_1$ but we choose H_0 is called a **Type II error**. We define the level of a simple versus simple test to be

$$\alpha = P_{\theta_0}(\text{We make a Type I error})$$

or

$$\alpha = P_{\theta_0}(X \in R_\phi) = E_{\theta_0}(\phi(X))$$

The other error probability is denoted β and defined as

$$\beta = P_{\theta_1}(X \notin R_\phi) = E_{\theta_1}(1 - \phi(X))$$

Suppose we want to minimize $\alpha + \beta$, the total error probability. We want to minimize

$$E_{\theta_0}(\phi(X)) + E_{\theta_1}(1 - \phi(X)) = \int [\phi(x) f_0(x) + (1 - \phi(x)) f_1(x)] dx$$

The problem is to choose, for each x, either the value 0 or the value 1, in such a way as to minimize the integral. But for each x the quantity

$$\phi(x) f_0(x) + (1 - \phi(x)) f_1(x)$$

can be chosen either to be $f_0(x)$ or $f_1(X)$. To make it small we take $\phi(x) = 1$ if $f_1(x) > f_0(x)$ and $\phi(x) = 0$ if $f_1(x) < f_0(x)$. It makes no difference what we do for those x for which $f_1(x) = f_0(x)$. Notice that we can divide both sides of these inequalities to rephrase the condition in terms of the **likelihood ratio** $f_1(x)/f_0(x)$.

Theorem 1 *For each fixed λ the quantity $\beta + \lambda\alpha$ is minimized by any ϕ which has*

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda \end{cases}$$

Neyman and Pearson suggested that in practise the two kinds of errors might well have unequal consequences. They suggested that rather than minimize any quantity of the form above you pick the more serious kind of error, label it **Type I** and require your rule to hold the probability α of a Type I error to be no more than some prespecified level α_0. (This value α_0 is typically 0.05 these days, chiefly for historical reasons.)

The Neyman and Pearson approach is then to minimize *beta* subject to the constraint $\alpha \leq \alpha_0$. Usually this is really equivalent to the constraint $\alpha = \alpha_0$ (because if you use $\alpha < \alpha_0$ you could make R larger and keep $\alpha \leq \alpha_0$ but make β smaller. For discrete models, however, this may not be possible. **Example**: Suppose X is Binomial(n, p) and either $p = p_0 = 1/2$ or $p = p_1 = 3/4$. If R is any critical region (so R is a subset of $\{0, 1, \ldots, n\}$) then

$$P_{1/2}(X \in R) = \frac{k}{2^n}$$

for some integer k. If we want $\alpha_0 = 0.05$ with say $n = 5$ for example we have to recognize that the possible values of α are $0, 1/32 = 0.03125, 2/32 = 0.0625$ and so on. For $\alpha_0 = 0.05$ we must use one of three rejection regions: R_1 which is the empty set, R_2 which is the set $x = 0$ or R_3 which is the set $x = 5$. These three regions have *alpha* equal to 0, 0.3125 and 0.3125 respectively and β equal to 1, $1-(1/4)^5$ and $1-(3/4)^5$ respectively so that R_3 minimizes β subject to $\alpha < 0.05$. If we raise α_0 slightly to 0.0625 then the possible rejection regions are R_1, R_2, R_3 and a fourth region $R_4 = R_2 \cup R_3$. The first three have the same α and β as before while R_4 has $\alpha = \alpha_0 = 0.0625$ an $\beta = 1 - (3/4)^5 - (1/4)^5$. Thus R_4 is optimal! The trouble is that this region says if all the trials are failures we should choose $p = 3/4$ rather than $p = 1/2$ even though the latter makes 5 failures much more likely than the former.

The problem in the example is one of discreteness. Here's how we get around the problem. First we expand the set of possible values of ϕ to include numbers between 0 and 1. Values of $\phi(x)$ between 0 and 1 represent the chance that we choose H_1 given that we observe x; the idea is that we actually toss a (biased) coin to decide! This tactic will show us the kinds of rejection regions which are sensible. In practise we then restrict our attention to levels α_0 for which the best ϕ is always either 0 or 1. In the binomial example we will insist that the value of α_0 be either 0 or $P_{\theta_0}(X \geq 5)$ or $P_{\theta_0}(X \geq 4)$ or ...

Here is a smaller example. There are 4 possible values of X and 2^4 possible rejection regions. Here is a table of the levels for each possible rejection region

R:

R	α
{}	0
{3}, {0}	1/8
{0,3}	2/8
{1}, {2}	3/8
{0,1}, {0,2}, {1,3}, {2,3}	4/8
{0,1,3}, {0,2,3}	5/8
{1,2}	6/8
{0,1,3}, {0,2,3}	7/8
{0,1,2,3}	

The best level 2/8 test has rejection region $\{0, 3\}$ and $\beta = 1 - [(3/4)^3 + (1/4)^3] = 36/64$. If, instead, we permit randomization then we will find that the best level test rejects when $X = 3$ and, when $X = 2$ tosses a coin which has chance 1/3 of landing heads, then rejects if you get heads. The level of this test is $1/8 + (1/3)(3/8) = 2/8$ and the probability of a Type II error is $\beta = 1 - [(3/4)^3 + (1/3)(3)(3/4)^2(1/4)] = 28/64$.

Definition: A hypothesis test is a function $\phi(x)$ whose values are always in $[0, 1]$. If we observe $X = x$ then we choose H_1 with conditional probability $\phi(X)$. In this case we have

$$\pi(\theta) = E_\theta(\phi(X))$$

$$\alpha = E_0(\phi(X))$$

and

$$\beta = E_1(\phi(X))$$

The Neyman Pearson Lemma

Theorem 2 *In testing f_0 against f_1 the probability β of a type II error is minimized, subject to $\alpha \leq \alpha_0$ by the test function:*

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda \\ \gamma & \frac{f_1(x)}{f_0(x)} = \lambda \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda \end{cases}$$

where λ is the largest constant such that

$$P_0(\frac{f_1(x)}{f_0(x)} \geq \lambda) \geq \alpha_0$$

and

$$P_0(\frac{f_1(x)}{f_0(x)} \leq \lambda) \geq 1 - \alpha_0$$

and where γ is any number chosen so that

$$E_0(\phi(X)) = P_0(\frac{f_1(x)}{f_0(x)} > \lambda) + \gamma P_0(\frac{f_1(x)}{f_0(x)} = \lambda) = \alpha_0$$

The value of γ is unique if $P_0(\frac{f_1(x)}{f_0(x)} = \lambda) > 0$.

Example: In the Binomial(n, p) with $p_0 = 1/2$ and $p_1 = 3/4$ the ratio f_1/f_0 is

$$3^x 2^{-n}$$

Now if $n = 5$ then this ratio must be one of the numbers 1, 3, 9, 27, 81, 243 divided by 32. Suppose we have $\alpha = 0.05$. The value of λ must be one of the possible values of f_1/f_0. If we try $\lambda = 343/32$ then

$$P_0(3^X 2^{-5} \geq 343/32) = P_0(X = 5) = 1/32 < 0.05$$

and

$$P_0(3^X 2^{-5} \geq 81/32) = P_0(X \geq 4) = 6/32 > 0.05$$

This means that $\lambda = 81/32$. Since

$$P_0(3^X 2^{-5} > 81/32) = P_0(X = 5) = 1/32$$

we must solve

$$P_0(X = 5) + \gamma P_0(X = 4) = 0.05$$

for γ and find

$$\gamma = \frac{0.05 - 1/32}{5/32} = 0.12$$

NOTE: No-one ever uses this procedure. Instead the value of α_0 used in discrete problems is chosen to be a possible value of the rejection probability

when $\gamma = 0$ (or $\gamma = 1$). When the sample size is large you can come very close to any desired α_0 with a non-randomized test.

If $\alpha_0 = 6/32$ then we can either take λ to be $343/32$ and $\gamma = 1$ or $\lambda = 81/32$ and $\gamma = 0$. However, our definition of λ in the theorem makes $\lambda = 81/32$ and $\gamma = 0$.

When the theorem is used for continuous distributions it can be the case that the CDF of $f_1(X)/f_0(X)$ has a flat spot where it is equal to $1 - \alpha_0$. This is the point of the word "largest" in the theorem.

Example: If $X_1, \ldots, X_n$ are iid $N(\mu, 1)$ and we have $\mu_0 = 0$ and $\mu_1 > 0$ then

$$\frac{f_1(X_1, \ldots, X_n)}{f_0(X_1, \ldots, X_n)} = \exp\{\mu_1 \sum X_i - n\mu_1^2/2 - \mu_0 \sum X_i + n\mu_2^2/2\}$$

which simplifies to

$$\exp\{\mu_1 \sum X_i - n\mu_1^2/2\}$$

We now have to choose λ so that

$$P_0(\exp\{\mu_1 \sum X_i - n\mu_1^2/2\} > \lambda) = \alpha_0$$

We can make it equal because in this case $f_1(X)/f_0(X)$ has a continuous distribution. Rewrite the probability as

$$P_0(\sum X_i > [\log(\lambda) + n\mu_1^2/2]/\mu_1) = 1 - \Phi([\log(\lambda) + n\mu_1^2/2]/[n^{1/2}\mu_1])$$

If z_α is notation for the usual upper α critical point of the normal distribution then we find

$$z_{\alpha_0} = [\log(\lambda) + n\mu_1^2/2]/[n^{1/2}\mu_1]$$

which you can solve to get a formula for λ in terms of z_{α_0}, n and μ_1.

The rejection region looks complicated: reject if a complicated statistic is larger than λ which has a complicated formula. But in calculating λ we re-expressed the rejection region in terms of

$$\frac{\sum X_i}{\sqrt{n}} > z_{\alpha_0}$$

The key feature is that this rejection region is the same for any $\mu_1 > 0$. [WARNING: in the algebra above I used $\mu_1 > 0$.] This is why the Neyman Pearson lemma is a lemma!

Definition: In the general problem of testing Θ_0 against Θ_1 the level of a test function ϕ is

$$\alpha = \sup_{\theta \in \Theta_0} E_\theta(\phi(X))$$

The power function is

$$\pi(\theta) = E_\theta(\phi(X))$$

A test ϕ^* is a Uniformly Most Powerful level α_0 test if

1. ϕ^* has level $\alpha \leq \alpha_o$
2. If ϕ has level $\alpha \leq \alpha_0$ then for every $\theta \in \Theta_1$ we have

$$E_\theta(\phi(X)) \leq E_\theta(\phi^*(X))$$

Application of the NP lemma: In the $N(\mu, 1)$ model consider $\Theta_1 = \{\mu > 0\}$ and $\Theta_0 = \{0\}$ or $\Theta_0 = \{\mu \leq 0\}$. The UMP level α_0 test of $H_0 : \mu \in \Theta_0$ against $H_1 : \mu \in \Theta_1$ is

$$\phi(X_1, \ldots, X_n) = 1(n^{1/2}\bar{X} > z_{\alpha_0})$$

Proof: For either choice of Θ_0 this test has level α_0 because for $\mu \leq 0$ we have

$$\begin{aligned} P_\mu(n^{1/2}\bar{X} > z_{\alpha_0}) &= P_\mu(n^{1/2}(\bar{X} - \mu) > z_{\alpha_0} - n^{1/2}\mu) \\ &= P(N(0,1) > z_{\alpha_0} - n^{1/2}\mu) \\ &\leq P(N(0,1) > z_{\alpha_0}) \\ &= \alpha_0 \end{aligned}$$

(Notice the use of $\mu \leq 0$. The central point is that the critical point is determined by the behaviour on the edge of the null hypothesis.)

Now if ϕ is any other level α_0 test then we have

$$E_0(\phi(X_1, \ldots, X_n)) \leq \alpha_0$$

Fix a $\mu > 0$. According to the NP lemma

$$E_\mu(\phi(X_1, \ldots, X_n)) \leq E_\mu(\phi_\mu(X_1, \ldots, X_n))$$

where ϕ_μ rejects if $f_\mu(X_1,\ldots,X_n)/f_0(X_1,\ldots,X_n) > \lambda$ for a suitable λ. But we just checked that this test had a rejection region of the form

$$n^{1/2}\bar{X} > z_{\alpha_0}$$

which is the rejection region of ϕ^*. The NP lemma produces the same test for every $\mu > 0$ chosen as an alternative. So we have shown that $\phi_\mu = \phi^*$ for any $\mu > 0$.

Proof of the Neyman Pearson lemma: Given a test ϕ with level strictly less than α_0 we can define the test

$$\phi^*(x) = \frac{1-\alpha_0}{1-\alpha}\phi(x) + \frac{\alpha_0 - \alpha}{1-\alpha}$$

has level α_0 and β smaller than that of ϕ. Hence we may assume without loss that $\alpha = \alpha_0$ and minimize β subject to $\alpha = \alpha_0$. However, the argument which follows doesn't actually need this.

Lagrange Multipliers

Suppose you want to minimize $f(x)$ subject to $g(x) = 0$. Consider first the function

$$h_\lambda(x) = f(x) + \lambda g(x)$$

If x_λ minimizes h_λ then for any other x

$$f(x_\lambda) \le f(x) + \lambda[g(x) - g(x_\lambda)]$$

Now suppose you can find a value of λ such that the solution x_λ has $g(x_\lambda) = 0$. Then for any x we have

$$f(x_\lambda) \le f(x) + \lambda g(x)$$

and for any x satisfying the constraint $g(x) = 0$ we have

$$f(x_\lambda) \le f(x)$$

This proves that for this special value of λ the quantity x_λ minimizes $f(x)$ subject to $g(x) = 0$.

Notice that to find x_λ you set the usual partial derivatives equal to 0; then to find the special x_λ you add in the condition $g(x\lambda) = 0$.

Proof of NP lemma

For each $\lambda > 0$ we have seen that ϕ_λ minimizes $\lambda\alpha + \beta$ where $\phi_\lambda = 1(f_1(x)/f_0(x) \geq \lambda)$.

As λ increases the level of ϕ_λ decreases from 1 when $\lambda = 0$ to 0 when $\lambda = \infty$. There is thus a value λ_0 where for $\lambda < \lambda_0$ the level is less than α_0 while for $\lambda > \lambda_0$ the level is at least α_0. Temporarily let $\delta = P_0(f_1(X)/f_0(X) = \lambda_0)$. If $\delta = 0$ define $\phi = \phi_\lambda$. If $\delta > 0$ define

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda_0 \\ \gamma & \frac{f_1(x)}{f_0(x)} = \lambda_0 \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda_0 \end{cases}$$

where $P_0(f_1(X)/f_0(X) < \lambda_0) + \gamma\delta = \alpha_0$. You can check that $\gamma \in [0, 1]$.

Now ϕ has level α_0 and according to the theorem above minimizes $\alpha + \lambda_0\beta$. Suppose ϕ^* is some other test with level $\alpha^* \leq \alpha_0$. Then

$$\lambda_0\alpha_\phi + \beta_\phi \leq \lambda_0\alpha_{\phi^*} + \beta_{\phi^*}$$

We can rearrange this as

$$\beta_{\phi^*} \geq \beta_\phi + (\alpha_\phi - \alpha_{\phi^*})\lambda_0$$

Since

$$\alpha_{\phi^*} \leq \alpha_0 = \alpha_\phi$$

the second term is non-negative and

$$\beta_{\phi^*} \geq \beta_\phi$$

which proves the Neyman Pearson Lemma.

Definition: In the general problem of testing Θ_0 against Θ_1 the level of a test function ϕ is

$$\alpha = \sup_{\theta \in \Theta_0} E_\theta(\phi(X))$$

The power function is

$$\pi(\theta) = E_\theta(\phi(X))$$

A test ϕ^* is a Uniformly Most Powerful level α_0 test if

1. ϕ^* has level $\alpha \leq \alpha_o$

2. If ϕ has level $\alpha \leq \alpha_0$ then for every $\theta \in \Theta_1$ we have

$$E_\theta(\phi(X)) \leq E_\theta(\phi^*(X))$$

Application of the NP lemma: In the $N(\mu, 1)$ model consider $\Theta_1 = \{\mu > 0\}$ and $\Theta_0 = \{0\}$ or $\Theta_0 = \{\mu \leq 0\}$. The UMP level α_0 test of $H_0 : \mu \in \Theta_0$ against $H_1 : \mu \in \Theta_1$ is

$$\phi(X_1, \ldots, X_n) = 1(n^{1/2}\bar{X} > z_{\alpha_0})$$

Proof: For either choice of Θ_0 this test has level α_0 because for $\mu \leq 0$ we have

$$\begin{aligned} P_\mu(n^{1/2}\bar{X} > z_{\alpha_0}) &= P_\mu(n^{1/2}(\bar{X} - \mu) > z_{\alpha_0} - n^{1/2}\mu) \\ &= P(N(0,1) > z_{\alpha_0} - n^{1/2}\mu) \\ &\leq P(N(0,1) > z_{\alpha_0}) \\ &= \alpha_0 \end{aligned}$$

(Notice the use of $\mu \leq 0$. The central point is that the critical point is determined by the behaviour on the edge of the null hypothesis.)

Now if ϕ is any other level α_0 test then we have

$$E_0(\phi(X_1, \ldots, X_n)) \leq \alpha_0$$

Fix a $\mu > 0$. According to the NP lemma

$$E_\mu(\phi(X_1, \ldots, X_n)) \leq E_\mu(\phi_\mu(X_1, \ldots, X_n))$$

where ϕ_μ rejects if $f_\mu(X_1, \ldots, X_n)/f_0(X_1, \ldots, X_n) > \lambda$ for a suitable λ. But we just checked that this test had a rejection region of the form

$$n^{1/2}\bar{X} > z_{\alpha_0}$$

which is the rejection region of ϕ^*. The NP lemma produces the same test for every $\mu > 0$ chosen as an alternative. So we have shown that $\phi_\mu = \phi^*$ for any $\mu > 0$.

This phenomenon is somewhat general. What happened was this. For any $\mu > \mu_0$ the likelihood ratio f_μ/f_0 is an increasing function of $\sum X_i$. The rejection region of the NP test is thus always a region of the form $\sum X_i > k$.

The value of the constant k is determined by the requirement that the test have level α_0 and this depends on μ_0 not on μ_1.
Definition: The family $f_\theta; \theta \in \Theta \subset R$ has monotone likelihood ratio with respect to a statistic $T(X)$ if for each $\theta_1 > \theta_0$ the likelihood ratio $f_{\theta_1}(X)/f_{\theta_0}(X)$ is a monotone increasing function of $T(X)$.

Theorem 3 *For a monotone likelihood ratio family the Uniformly Most Powerful level α test of $\theta \leq \theta_0$ (or of $\theta = \theta_0$) against the alternative $\theta > \theta_0$ is*

$$\phi(x) = \begin{cases} 1 & T(x) > t_\alpha \\ \gamma & T(X) = t_\alpha \\ 0 & T(x) < t_\alpha \end{cases}$$

where $P_0(T(X) > t_\alpha) + \gamma P_0(T(X) = t_\alpha) = \alpha_0$.

A typical family where this will work is a one parameter exponential family. In almost any other problem the method doesn't work and there is no uniformly most powerful test. For instance to test $\mu = \mu_0$ against the two sided alternative $\mu \neq \mu_0$ there is no UMP level α test. If there were its power at $\mu > \mu_0$ would have to be as high as that of the one sided level α test and so its rejection region would have to be the same as that test, rejecting for large positive values of $\bar{X} - \mu_0$. But it also has to have power as good as the one sided test for the alternative $\mu < \mu_0$ and so would have to reject for large negative values of $\bar{X} - \mu_0$. This would make its level too large.

The favourite test is the usual 2 sided test which rejects for large values of $|\bar{X} - \mu_0|$ with the critical value chosen appropriately. This test maximizes the power subject to two constraints: first, that the level be α and second that the test have power which is minimized at $\mu = \mu_0$. This second condition is really that the power on the alternative be larger than it is on the null.
Definition: A test ϕ of Θ_0 against Θ_1 is unbiased level α if it has level α and, for every $\theta \in \Theta_1$ we have

$$\pi(\theta) \geq \alpha\,.$$

When testing a point null hypothesis like $\mu = \mu_0$ this requires that the power function be minimized at μ_0 which will mean that if π is differentiable then

$$\pi'(\mu_0) = 0$$

We now apply that condition to the $N(\mu, 1)$ problem. If ϕ is any test function then

$$\pi'(\mu) = \frac{\partial}{\partial \mu} \int \phi(x) f(x, \mu) dx$$

We can differentiate this under the integral and use

$$\frac{\partial f(x, \mu)}{\partial \mu} = \sum (x_i - \mu) f(x, \mu)$$

to get the condition

$$\int \phi(x) \bar{x} f(x, \mu_0) dx = \mu_0 \alpha_0$$

Consider the problem of minimizing $\beta(\mu)$ subject to the two constraints $E_{\mu_0}(\phi(X)) = \alpha_0$ and $E_{\mu_0}(\bar{X}\phi(X)) = \mu_0 \alpha_0$. Now fix two values $\lambda_1 > 0$ and λ_2 and minimize

$$\lambda_1 \alpha + \lambda_2 E_{\mu_0}[(\bar{X} - \mu_0)\phi(X)] + \beta$$

The quantity in question is just

$$\int [\phi(x) f_0(x)(\lambda_1 + \lambda_2(\bar{X} - \mu_0)) + (1 - \phi(x)) f_1(x)] dx$$

As before this is minimized by

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda_1 + \lambda_2(\bar{X} - \mu_0) \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda_1 + \lambda_2(\bar{X} - \mu_0) \end{cases}$$

The likelihood ratio f_1/f_0 is simply

$$\exp\{n(\mu_1 - \mu_0)\bar{X} + n(\mu_0^2 - \mu_1^2)/2\}$$

and this exceeds the linear function

$$\lambda_1 + \lambda_2(\bar{X} - \mu_0)$$

for all $\bar{X}$ sufficiently large or small. That is, the quantity

$$\lambda_1 \alpha + \lambda_2 E_{\mu_0}[(\bar{X} - \mu_0)\phi(X)] + \beta$$

is minimized by a rejection region of the form

$$\{\bar{X} > K_U\} \cup \{\bar{X} < K_L\}$$

To satisfy the constraints we adjust K_U and K_L to get level α and $\pi'(\mu_0) = 0$. The second condition shows that the rejection region is symmetric about μ_0 and then we discover that the test rejects for

$$\sqrt{n}|\bar{X} - \mu_0| > z_{\alpha/2}$$

Now you have to mimic the Neyman Pearson lemma proof to check that if λ_1 and λ_2 are adjusted so that the unconstrained problem has the rejection region given then the resulting test minimizes β subject to the two constraints.

A test ϕ^* is a Uniformly Most Powerful Unbiased level α_0 test if

1. ϕ^* has level $\alpha \leq \alpha_0$.
2. ϕ^* is unbiased.
3. If ϕ has level $\alpha \leq \alpha_0$ then for every $\theta \in \Theta_1$ we have

$$E_\theta(\phi(X)) \leq E_\theta(\phi^*(X))$$

Conclusion: The two sided z test which rejects if

$$|Z| > z_{\alpha/2}$$

where

$$Z = n^{1/2}(\bar{X} - \mu_0)$$

is the uniformly most powerful unbiased test of $\mu = \mu_0$ against the two sided alternative $\mu \neq \mu_0$.

Nuisance Parameters

What good can be said about the t-test? It's UMPU.

Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ and that we want to test $\mu = \mu_0$ or $\mu \leq \mu_0$ against $\mu > \mu_0$. Notice that the parameter space is two dimensional and that the boundary between the null and alternatives is

$$\{(\mu, \sigma); \mu = \mu_0, \sigma > 0\}$$

If a test has $\pi(\mu, \sigma) \leq \alpha$ for all $\mu \leq \mu_0$ and $\pi(\mu, \sigma) \geq \alpha$ for all $\mu > \mu_0$ then we must have $\pi(\mu_0, \sigma) = \alpha$ for all σ because the power function of

any test must be continuous. (This actually uses the dominated convergence theorem; the power function is an integral.)

Now think of $\{(\mu, \sigma); \mu = \mu_0\}$ as a parameter space for a model. For this parameter space you can check that

$$S = \sum (X_i - \mu_0)^2$$

is a complete sufficient statistic. Remember that the definitions of both completeness and sufficiency depend on the parameter space. Suppose $\phi(\sum X_i, S)$ is an unbiased level α test. Then we have

$$E_{\mu_0,\sigma}(\phi(\sum X_i, S)) = \alpha$$

for all σ. Condition on S and get

$$E_{\mu_0,\sigma}[E(\phi(\sum X_i, S)|S)] = \alpha$$

for all σ. Sufficiency guarantees that

$$g(S) = E(\phi(\sum X_i, S)|S)$$

is a statistic and completeness that

$$g(S) \equiv \alpha$$

Now let us fix a single value of σ and a $\mu_1 > \mu_0$. To make our notation simpler I take $\mu_0 = 0$. Our observations above permit us to condition on $S = s$. Given $S = s$ we have a level α test which is a function of $\bar{X}$.

If we maximize the conditional power of this test for each s then we will maximize its power. What is the conditional model given $S = s$? That is, what is the conditional distribution of $\bar{X}$ given $S = s$? The answer is that the joint density of $\bar{X}, S$ is of the form

$$f_{\bar{X},S}(t, s) = h(s, t) \exp\{\theta_1 t + \theta_2 s + c(\theta_1, \theta_2)\}$$

where $\theta_1 = n\mu/\sigma^2$ and $\theta_2 = -1/\sigma^2$.

This makes the conditional density of $\bar{X}$ given $S = s$ of the form

$$f_{\bar{X}|s}(t|s) = h(s, t) \exp\{\theta_1 t + c^*(\theta_1, s)\}$$

Notice the disappearance of θ_2. Notice that the null hypothesis is actually $\theta_1 = 0$. This permits the application of the NP lemma to the conditional family to prove that the UMP unbiased test must have the form

$$\phi(\bar{X}, S) = 1(\bar{X} > K(S))$$

where $K(S)$ is chosen to make the conditional level α. The function $x \mapsto x/\sqrt{a - x^2}$ is increasing in x for each a so that we can rewrite ϕ in the form

$$\phi(\bar{X}, S) = 1(n^{1/2}\bar{X}/\sqrt{n[S/n - \bar{X}^2]/(n-1)} > K^*(S))$$

for some K^*. The quantity

$$T = \frac{n^{1/2}\bar{X}}{\sqrt{n[S/n - \bar{X}^2]/(n-1)}}$$

is the usual t statistic and is exactly independent of S (see Theorem 6.1.5 on page 262 in Casella and Berger). This guarantees that

$$K^*(S) = t_{n-1,\alpha}$$

and makes our UMPU test the usual t test.

Optimal tests

- A good test has $\pi(\theta)$ large on the alternative and small on the null.
- For one sided one parameter families with MLR a UMP test exists.
- For two sided or multiparameter families the best to be hoped for is UMP Unbiased or Invariant or Similar.
- Good tests are found as follows:

    1. Use the NP lemma to determine a good rejection region for a simple alternative.
    2. Try to express that region in terms of a statistic whose definition does not depend on the specific alternative.
    3. If this fails impose an additional criterion such as unbiasedness. Then mimic the NP lemma and again try to simplify the rejection region.

Likelihood Ratio tests

For general composite hypotheses optimality theory is not usually successful in producing an optimal test. instead we look for heuristics to guide our choices. The simplest approach is to consider the likelihood ratio

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}$$

and choose values of $\theta_1 \in \Theta_1$ and $\theta_0 \in \Theta_0$ which are reasonable estimates of θ assuming respectively the alternative or null hypothesis is true. The simplest method is to make each θ_i a maximum likelihood estimate, but maximized only over Θ_i.

Example 1: In the $N(\mu, 1)$ problem suppose we want to test $\mu \leq 0$ against $\mu > 0$. (Remember there is a UMP test.) The log likelihood function is

$$-n(\bar{X} - \mu)^2/2$$

If $\bar{X} > 0$ then this function has its global maximum in Θ_1 at $\bar{X}$. Thus $\hat{\mu}_1$ which maximizes $\ell(\mu)$ subject to $\mu > 0$ is $\bar{X}$ if $\bar{X} > 0$. When $\bar{X} \leq 0$ the maximum of $\ell(\mu)$ over $\mu > 0$ is on the boundary, at $\hat{\mu}_1 = 0$. (Technically this is in the null but in this case $\ell(0)$ is the supremum of the values $\ell(\mu)$ for $\mu > 0$. Similarly, the estimate $\hat{\mu}_0$ will be $\bar{X}$ if $\bar{X} \leq 0$ and 0 if $\bar{X} > 0$. It follows that

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} = \exp\{\ell(\hat{\mu}_1) - \ell(\hat{\mu}_0)\}$$

which simplifies to

$$\exp\{n\bar{X}|\bar{X}|/2\}$$

This is a monotone increasing function of $\bar{X}$ so the rejection region will be of the form $\bar{X} > K$. To get the level right the test will have to reject if $n^{1/2}\bar{X} > z_\alpha$. Notice that the *log likelihood ratio statistic*

$$\lambda \equiv 2\log(\frac{f_{\hat{\mu}_1}(X)}{f_{\hat{\mu}_0}(X)} = n\bar{X}|\bar{X}|$$

as a simpler statistic.

Example 2: In the $N(\mu, 1)$ problem suppose we make the null $\mu = 0$. Then the value of $\hat{\mu}_0$ is simply 0 while the maximum of the log-likelihood over the alternative $\mu \neq 0$ occurs at $\bar{X}$. This gives

$$\lambda = n\bar{X}^2$$

which has a χ^2_1 distribution. This test leads to the rejection region $\lambda > (z_{\alpha/2})^2$ which is the usual UMPU test.

Example 3: For the $N(\mu, \sigma^2)$ problem testing $\mu = 0$ against $\mu \neq 0$ we must find two estimates of μ, σ^2. The maximum of the likelihood over the alternative occurs at the global mle $\bar{X}, \hat{\sigma}^2$. We find

$$\ell(\hat{\mu}, \hat{\sigma}^2) = -n/2 - n\log(\hat{\sigma})$$

We also need to maximize ℓ over the null hypothesis. Recall

$$\ell(\mu, \sigma) = -\frac{1}{2\sigma^2}\sum(X_i - \mu)^2 - n\log(\sigma)$$

On the null hypothesis we have $\mu = 0$ and so we must find $\hat{\sigma}_0$ by maximizing

$$\ell(0, \sigma) = -\frac{1}{2\sigma^2}\sum X_i^2 - n\log(\sigma)$$

This leads to

$$\hat{\sigma}_0^2 = \sum X_i^2/n$$

and

$$\ell(0, \hat{\sigma}_0) = -n/2 - n\log(\hat{\sigma}_0)$$

This gives

$$\lambda = -n\log(\hat{\sigma}^2/\hat{\sigma}_0^2)$$

Since

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = \frac{\sum(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2 + n\bar{X}^2}$$

we can write

$$\lambda = n\log(1 + t^2/(n-1))$$

where

$$t = \frac{n^{1/2}\bar{X}}{s}$$

is the usual t statistic. The likelihood ratio test thus rejects for large values of $|t|$ which gives the usual test.

Notice that if n is large we have

$$\lambda \approx n[1 + t^2/(n-1) + O(n^{-2})] \approx t^2\,.$$

Since the t statistic is approximately standard normal if n is large we see that

$$\lambda = 2[\ell(\hat{\theta}_1) - \ell(\hat{\theta}_0)]$$

has nearly a χ^2_1 distribution.

This is a general phenomenon when the null hypothesis being tested is of the form $\phi = 0$. Here is the general theory. Suppose that the vector of $p + q$ parameters θ can be partitioned into $\theta = (\phi, \gamma)$ with ϕ a vector of p parameters and γ a vector of q parameters. To test $\phi = \phi_0$ we find two MLEs of θ. First the global mle $\hat{\theta} = (\hat{\phi}, \hat{\gamma})$ maximizes the likelihood over $\Theta_1 = \{\theta : \phi \neq \phi_0\}$ (because typically the probability that $\hat{\phi}$ is exactly ϕ_0 is 0).

Now we maximize the likelihood over the null hypothesis, that is we find $\hat{\theta}_0 = (\phi_0, \hat{\gamma}_0)$ to maximize

$$\ell(\phi_0, \gamma)$$

The log-likelihood ratio statistic is

$$2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

Now suppose that the true value of θ is ϕ_0, γ_0 (so that the null hypothesis is true). The score function is a vector of length $p+q$ and can be partitioned as $U = (U_\phi, U_\gamma)$. The Fisher information matrix can be partitioned as

$$\begin{bmatrix} I_{\phi\phi} & I_{\phi\gamma} \\ I_{\phi\gamma} & I_{\gamma\gamma} \end{bmatrix}.$$

According to our large sample theory for the mle we have

$$\hat{\theta} \approx \theta + I^{-1}U$$

and

$$\hat{\gamma}_0 \approx \gamma_0 + I^{-1}_{\gamma\gamma} U_\gamma$$

If you carry out a two term Taylor expansion of both $\ell(\hat{\theta})$ and $\ell(\hat{\theta}_0)$ around θ_0 you get

$$\ell(\hat{\theta}) \approx \ell(\theta_0) + U^t I^{-1} U + \frac{1}{2} U^t I^{-1} V(\theta) I^{-1} U$$

where V is the second derivative matrix of ℓ. Remember that $V \approx -I$ and you get

$$2[\ell(\hat{\theta}) - \ell(\theta_0)] \approx U^t I^{-1} U\,.$$

A similar expansion for $\hat{\theta}_0$ gives

$$2[\ell(\hat{\theta}_0) - \ell(\theta_0)] \approx U_\gamma^t I_{\gamma\gamma}^{-1} U_\gamma \, .$$

If you subtract these you find that

$$2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

can be written in the approximate form

$$U^t M U$$

for a suitable matrix M. It is now possible to use the general theory of the distribution of $X^t M X$ where X is $MVN(0, \Sigma)$ to demonstrate that

Theorem 4 *The log-likelihood ratio statistic*

$$\lambda = 2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

has, under the null hypothesis, approximately a χ_p^2 distribution.

Aside:

Theorem 5 *Suppose that $X \sim MVN(0, \Sigma)$ with Σ non-singular and M is a symmetric matrix. If $\Sigma M \Sigma M \Sigma = \Sigma M \Sigma$ then $X^t M X$ has a χ^2 distribution with degrees of freedom $\nu = trace(M\Sigma)$.*

Proof: We have $X = AZ$ where $AA^t = \Sigma$ and Z is standard multivariate normal. So $X^t M X = Z^t A^t M A Z$. Let $Q = A^t M A$. Since $AA^t = \Sigma$ the condition in the theorem is actually

$$AQQA^t = AQA^t$$

Since Σ is non-singular so is A. Multiply by A^{-1} on the left and $(A^t)^{-1}$ on the right to discover $QQ = Q$.

The matrix Q is symmetric and so can be written in the form $P\Lambda P^t$ where Λ is a diagonal matrix containing the eigenvalues of Q and P is an orthogonal matrix whose columns are the corresponding orthonormal eigenvectors. It follows that we can rewrite

$$Z^t Q Z = (P^t Z)^t \Lambda (PZ)$$

The variable $W = P^t Z$ is multivariate normal with mean 0 and variance covariance matrix $P^t P = I$; that is, W is standard multivariate normal. Now

$$W^t \Lambda W = \sum \lambda_i W_i^2$$

We have established that the general distribution of any quadratic form $X^t M X$ is a linear combination of χ^2 variables. Now go back to the condition $QQ = Q$. If λ is an eigenvalue of Q and $v \neq 0$ is a corresponding eigenvector then $QQv = Q(\lambda v) = \lambda Qv = \lambda^2 v$ but also $QQv = Qv = \lambda v$. Thus $\lambda(1 - \lambda)v = 0$. It follows that either $\lambda = 0$ or $\lambda = 1$. This means that the weights in the linear combination are all 1 or 0 and that $X^t M X$ has a χ^2 distribution with degrees of freedom, ν, equal to the number of λ_i which are equal to 1. This is the same as the sum of the λ_i so

$$\nu = trace(\Lambda)$$

But

$$\begin{aligned} trace(M\Sigma) &= trace(MAA^t) \\ &= trace(A^t MA) \\ &= trace(Q) \\ &= trace(P\Lambda P^t) \\ &= trace(\Lambda P^t P) \\ &= trace(\Lambda) \end{aligned}$$

In the application Σ is $\mathcal{I}$ the Fisher information and $M = \mathcal{I}^{-1} - J$ where

$$J = \begin{bmatrix} 0 & 0 \\ 0 & I_{\gamma\gamma}^{-1} \end{bmatrix}$$

It is easy to check that $M\Sigma$ becomes

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

where I is a $p \times p$ identity matrix. It follows that $M\Sigma M\Sigma = M\Sigma$ and that $trace(M\Sigma) = p$).

Confidence Sets

A level β confidence set for a parameter $\phi(\theta)$ is a random subset C, of the set of possible values of ϕ such that for each θ we have

$$P_\theta(\phi(\theta) \in C) \geq \beta$$

Confidence sets are very closely connected with hypothesis tests:

Suppose C is a level $\beta = 1 - \alpha$ confidence set for ϕ. To test $\phi = \phi_0$ we consider the test which rejects if $\phi \notin C$. This test has level α. Conversely, suppose that for each ϕ_0 we have available a level α test of $\phi = \phi_0$ who rejection region is say R_{ϕ_0}. Then if we define $C = \{\phi_0 : \phi = \phi_0 \text{ is not rejected}\}$ we get a level $1 - \alpha$ confidence for ϕ. The usual t test gives rise in this way to the usual t confidence intervals

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

which you know well.

Confidence sets from Pivots

Definition: A **pivot** (or pivotal quantity) is a function $g(\theta, X)$ whose distribution is the same for all θ. (As usual the θ in the pivot is the same θ as the one being used to calculate the distribution of $g(\theta, X)$.

Pivots can be used to generate confidence sets as follows. Pick a set A in the space of possible values for g. Let $\beta = P_\theta(g(\theta, X) \in A)$; since g is pivotal β is the same for all θ. Now given a data set X solve the relation

$$g(\theta, X) \in A$$

to get

$$\theta \in C(X, A)\,.$$

Example: The quantity

$$(n-1)s^2/\sigma^2$$

is a pivot in the $N(\mu, \sigma^2)$ model. It has a χ^2_{n-1} distribution. Given $\beta = 1 - \alpha$ consider the two points $\chi^2_{n-1,1-\alpha/2}$ and $\chi^2_{n-1,\alpha/2}$. Then

$$P(\chi^2_{n-1,1-\alpha/2} \leq (n-1)s^2/\sigma^2 \leq \chi^2_{n-1,\alpha/2}) = \beta$$

for all μ, σ. We can solve this relation to get

$$P(\frac{(n-1)^{1/2}s}{\chi_{n-1,\alpha/2}} \leq \sigma \leq \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha/2}}) = \beta$$

so that the interval from $(n-1)^{1/2}s/\chi_{n-1,\alpha/2}$ to $(n-1)^{1/2}s/\chi_{n-1,1-\alpha/2}$ is a level $1-\alpha$ confidence interval.

In the same model we also have

$$P(\chi^2_{n-1,1-\alpha} \le (n-1)s^2/\sigma^2) = \beta$$

which can be solved to get

$$P(\sigma \le \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha/2}}) = \beta$$

This gives a level $1-\alpha$ interval $(0, (n-1)^{1/2}s/\chi_{n-1,1-\alpha})$. The right hand end of this interval is usually called a confidence upper bound.

In general the interval from $(n-1)^{1/2}s/\chi_{n-1,\alpha_1}$ to $(n-1)^{1/2}s/\chi_{n-1,1-\alpha_2}$ has level $\beta = 1-\alpha_1-\alpha_2$. For a fixed value of β we can minimize the length of the resulting interval numerically. This sort of optimization is rarely used. See your homework for an example of the method.