12 Testing of Hypotheses.

The simplest kind of a testing of hypothesis is when we have two possible alternate models and based on the sample have to make a choice between them.

Suppose $f_0(x)$ and $f_1(x)$ are two possible densities on R and we have an observation x. We have to make a choice between the two. The **null hypothesis** says that f_0 is the true density. The **alternate hypothesis** says that f_1 is the true density. A decision procedure has to be of the following form. A region Ω called the **critical region**is picked. If $x \in \Omega$ we are critical of the null hypothesis and **reject** it in favor of the alternate. If $x \notin \Omega$ we **accept** the null hypothesis. The probability $\alpha = \int_\Omega f_)(x)dx$ of rejecting the null hypothesis when it is true is called the **type I** error or **size**. The probability $\beta = \int_{\Omega^c} f_1(x)dx$ of accepting the null hypothesis when it is false is called the **type II** error. $1 - \beta = \int_\Omega f_1(x)dx$ which is the probability of detecting that the null hypothesis when it is really false is called the **power** of the test. Clearly we would like to minimize $\int_\Omega f_0(x)dx$ while at the same time maximizing $\int_\Omega f_1(x)dx$. This procedure of accepting or rejecting the null hypothesis based on observations is called a **test** or a **test of hypothesis**. In our contest each choice of Ω provides a diffrent test for the same problem of deciding to either accept f_0 or reject it in favor of f_1.

We would like both α and β to be as small as possible. Of course there is a conflict. Minimizing α suggests making Ω small and making β small on the other hand needs Ω to be large. There has to be a trade off.

We cannot compare two tests that have different type I errors. If Ω_1 and Ω_2 are two different tests with the same size Ω_1 is called **more powerful** than Ω_2 if

$$\int_{\Omega_1} f_2(x)dx \geq \int_{\Omega_2} f_2(x)dx$$

while

$$\alpha = \int_{\Omega_1} f_1(x)dx = \int_{\Omega_2} f_1(x)dx$$

A test Ω of size α is said to be **most powerful** if it is more powerful than any other test of the same size.

The Neyman-Pearson lemma provides a way of making the right choice of Ω.

Theorem 12.1 (Neyman-Pearson Lemma). *Consider the family of sets*

$$A_\lambda = \{x : \frac{f_2(x)}{f_1(x)} \geq \lambda\}$$
$$B_\lambda = \{x : \frac{f_2(x)}{f_1(x)} > \lambda\}$$

Any Ω such that $A_\lambda \supset \Omega \supset B_\lambda$ for some $\lambda > 0$ is a most powerful test.

Proof. Let Ω be such that $A_\lambda \supset \Omega \supset B_\lambda$ for some $\lambda > 0$. Let the size of Ω be α. Let Ω_1 be any test of size α. We will prove that Ω is more powerful than Ω_1.

$$\begin{aligned}
\int_\Omega f_2(x)dx - \int_{\Omega_1} f_2(x)dx &= \int_{\Omega\cap\Omega_1^c} f_2(x)dx - \int_{\Omega^c\cap\Omega_1} f_2(x)dx \\
&\geq \lambda \int_{\Omega\cap\Omega_1^c} f_1(x)dx - \lambda \int_{\Omega^c\cap\Omega_1} f_1(x)dx \\
&= \lambda \int_\Omega f_1(x)dx - \lambda \int_{\Omega_1} f_1(x)dx \\
&= \lambda\alpha - \lambda\alpha = 0
\end{aligned}$$

□

We can have two probability distributions $p_0(x)$ and $p_1(x)$ instead of densities and nothin really changes except that the integrals are replaced by sums.

Examples.

1. Suppose we have n independent observations from a normal population with mean μ and variance 1. The null hypothesis is that $\mu = 0$ and the alternate $\mu = 1$.

   $$\log \frac{f_1(x_1, \dots, x_n)}{f_0(x_1, \dots, x_n)} = \frac{1}{2}\sum_i x_i^2 - \frac{1}{2}\sum_i (x_i - 1)^2 = \sum_i x_i - \frac{n}{2}$$

   Uniformly most powerful critical regions can be written in the form $\sqrt{n}\bar{x} \geq a$. Since the distribution of $\sqrt{n}\bar{x}$ under the null hypothesis is the normal distribution with mean 0 and variance 1, we pick a such that the size is α

   $$\frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-\frac{x^2}{2}} dx = \alpha$$

   medskip

2. Suppose we have n independent observations from a normal population with mean μ and variance 1. The null hypothesis is that $\mu = 0$ and the alternate $\mu = 2$.

$$\log \frac{f_1(x_1, \ldots, x_n)}{f_0(x_1, \ldots, x_n)} = \frac{1}{2}\sum_i x_i^2 - \frac{1}{2}\sum_i (x_i - 2)^2 = 2\sum_i x_i - 2n$$

Uniformly most powerful critical regions can again be written in the form $\sqrt{n}\bar{x} \geq a$. Since the distribution of $\sqrt{n}\bar{x}$ under the null hypothesis is the normal distribution with mean 0 and variance 1, we pick a such that the size is α

$$\frac{1}{\sqrt{2\pi}}\int_a^\infty e^{-\frac{x^2}{2}}\,dx = \alpha$$

In fact the test does not care if $\mu = 1$ or 2. So long as the alternative is any $\mu > 0$ we have the same family of most powerful tests.

3. Suppose we have n independent observations from a normal population with mean μ and variance 1. The null hypothesis is that $\mu = 0$ and the alternate $\mu = -1$.

$$\log \frac{f_1(x_1, \ldots, x_n)}{f_0(x_1, \ldots, x_n)} = \frac{1}{2}\sum_i x_i^2 - \frac{1}{2}\sum_i (x_i + 1)^2 = -\sum_i x_i - \frac{n}{2}$$

Uniformly most powerful critical regions can now be written in the form $\sqrt{n}\bar{x} \leq a$. Since the distribution of $\sqrt{n}\bar{x}$ under the null hypothesis is the normal distribution with mean 0 and variance 1, we pick a such that the size is α

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^a e^{-\frac{x^2}{2}}\,dx = \alpha$$

In fact the family of most powerful tests depend only on the sign of the alternative μ.

4. Suppose we have n independent observations from a normal population with mean 0 and variance θ. The null hypothesis is that $\theta = 1$ and the alternate $\theta = 2$.

$$\log \frac{f_1(x_1, \ldots, x_n)}{f_0(x_1, \ldots, x_n)} = -\frac{n}{2}\log\theta + \frac{1}{2}(1 - \frac{1}{\theta})\sum_i x_i^2$$

The most powerful critical regions look like $\sum_i x_i^2 \geq a$. Now we need the distribution of $S = \sum_i x_i^2$. It is the Gamma distribution with $\alpha = \frac{1}{2}$ and $p = \frac{n}{2}$. It has a special name. It is called the *chi-square* distribution with n degrees of freedom. The level a is detrmined from the size α by

$$\alpha = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \int_a^\infty e^{-\frac{x}{2}} x^{\frac{n}{2}-1} dx$$

5. Similarly if the alternate is some $\theta < 1$ the most powerful critical regions looks like $\sum_i x_i^2 \leq a$. We still need the chi-square distribution, but determine a so that

$$\alpha = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \int_0^a e^{-\frac{x}{2}} x^{\frac{n}{2}-1} dx$$

6. We look at some discrete distributions. We tossed a coin n times and obtained x heads. Is $p = 0.5$ or is it 0.7?

$$\log \frac{p_1(x)}{p_0(x)} = x \log \frac{0.7}{0.5} + (n-x) \log \frac{0.3}{0.5}$$

The critical regions of interest are of the form $x \geq k$ and k is determined so that

$$\sum_{x \geq k} \binom{n}{x} (0.5)^n = \alpha$$

This may not always be possible. Because we can change k only by integers and the probability may jump over α with out matching it. Usually this is not important. If we get to some α that is close enough that should be satisfactory. There is nothing special about the exact value of α. It should be small, because we do not want to change our beliefs on flimsy evidence. In practice one takes it to be 0.05 or 0.01. It is some times called **the level of significance**. If we really have to match it we need a fuzzy set. Maybe with $k-1$ we have $\alpha_1 > \alpha$ and with k we have $\alpha_2 < \alpha$. We do not need all of $k-1$ but only a fraction $\frac{\alpha-\alpha_2}{\alpha_1-\alpha_2}$ of $k-1$. But you cannot divide a point. If we get the observation $k-1$ our action is fuzzy or randomized. We perform a totally irrelevent random experiment (like generating random numbers) and based on its outcome, reject with probability $\frac{\alpha-\alpha_2}{\alpha_1-\alpha_2}$. Now the size is matched. One can prove a variant of the Neyman-Pearson lemma that among randomized tests this is the best. In other words we randomize only at the very edge to match the size.

More generally if we have a family $\{P_\theta : \theta \in \Theta\}$ of possible models the null hypothesis may be that $\theta \in \Theta_0$ and the alternative $\theta \in \Theta_1 = \Theta \backslash \Theta_0$. The null hypothesis is said to be **simple** if Θ_0 consists of a single point θ_0. The alternative is similarly simple if Θ_1 consists only of one point θ_1. So far we have cosidered testing a simple hypothesis agianst a simple alternative. Any hypothesis that is not simple is called **composite**.

While testing a simple hypothesis agianst a composite alternative, if it happens that the most powerful test of a given size α aginst the alternative θ_1 determined by the Neyman-Pearson lemma, is independent of $\theta_1 \in \Theta_1$, then we say we have a uniformly most powerful test against all the alternatives in Θ_1. It does not always exist. But it might. In the examples we considred if the alternative is on one side of the null hypothesis, then UMP tests exist. For two sided alternatives they usually do not.

For example if we are to test from a normal ppulation with mean μ and variance 1, the null hypothesis $\mathcal{H}_0 = \{\mu = 0\}$ against the alternative $\mathcal{H}_1 = \{\mu > 0\}$ critical regions of the form $\sqrt{n}\bar{x}_n > a$ will yield UMP tests. If on the other hand the alternative is $\mathcal{H}_1 = \{\mu \neq 0\}$ we do not have any UMP tests. In practice one settles for a critical region of the form $|\sqrt{n}\bar{x}_n| > a$.

For any test the function $P(\theta) = P_\theta[\Omega]$ is the power. Its value at θ_0 is the type I error or size. It is a measure of our ability to detect deviations from the null hypothesis. Since $P(\theta)$ is usually continuous in θ the power is close to the size α if $\theta \in \Theta_1$ is close to θ_0. On the other hand if $P_n(\theta)$ is the power of a good test based on n observations, it will happen that if we fix the size α, $P_n(\theta) \to 1$ as $n \to \infty$ for any $\theta \neq \theta_0$.

Notice that in the example of testing for the mean of the normal population $\mu = 0$ against the two sided alternative $\mu \neq 0$, if we use a region of the form $|\sqrt{n}\bar{x}_n| > a$, the level a is independent of n. The power is given by

$$\begin{aligned}
P_n(\mu) &= P_\mu[|\sqrt{n}\bar{x}_n| > a] \\
&= \frac{1}{\sqrt{2\pi}} \int_{|x|>a} e^{-\frac{(x-\sqrt{n}\mu)^2}{2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{|x+\sqrt{n}\mu|>a} e^{-\frac{x^2}{2}} dx \\
&= \Phi(-a-\sqrt{n}\mu) + 1 - \Phi(a-\sqrt{n}\mu) \\
&\to 1
\end{aligned}$$

if $n \to \infty$ as long as $\mu \neq 0$.

13 Composite Null Hypothesis.

The situation is more complex if the null hypothesis is composite and takes the form $\mathcal{H}_0 = \{\theta \in \Theta_0\}$. To find a critical region of size α we must find Ω such that $P_\theta(\Omega) \equiv \alpha$ for $\theta \in \Theta_0$. This may not be possible. It is more reasonable to insist only that $P_\theta(\Omega) \leq \alpha$ for $\theta \in \Theta_0$. So we defne the size as

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(\Omega)$$

But it is some times possible to find a test which is at the same level α. This often requires finding a statistic U whose distribution is the same for all $\theta \in \Theta_0$. Then a test based on this statistic would serve the purpose. In many problems there are such natuaral statistics. Let us look at some examples.

Examples.

1. Suppose $x_1, \ldots, x_n$ are n independent observations from a normal poulation with mean μ and variance θ. We want to test the null hypothesis that $\mu = 0$ against the alternative $\mu > 0$. We can not use just $\sqrt{n}\bar{x}_n$ because its distribution will involve θ an unknown parameter. If s^2 is the sample variance

$$s^2 = \frac{1}{n}\sum_i (x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_i x_i^2 - [\frac{1}{n}\sum_i x_i]^2 \tag{13.1}$$

then the quantity

$$t = \frac{\bar{x}_n}{s}\sqrt{n-1}$$

is called the Student's 't 'statistic with $n-1$ **degrees of freedom**. Its probability distribution is independent of θ and the density of the Student's t distribution with k degrees of freedom is given by

$$f_k(t) = \frac{c_k}{(1+\frac{t^2}{k})^{\frac{k+1}{2}}}$$

where the normalizing constant c_k is easily evaluated.

$$\begin{aligned} c_k^{-1} &= \int_{-\infty}^{\infty} \frac{dt}{(1+\frac{t^2}{k})^{\frac{k+1}{2}}} = \sqrt{k}\int_0^{\infty} \frac{dt}{\sqrt{t}(1+t)^{\frac{k+1}{2}}} \\ &= \sqrt{k}\int_0^1 u^{-\frac{1}{2}}(1-u)^{\frac{k}{2}-1}du = \sqrt{k}\,\beta(\frac{1}{2},\frac{k}{2}). \end{aligned}$$

For the two sided alternative one uses the same statistic, but a critical region of the form $|t| > a$.

2. Suppose we have n observations from a normal poulation with mean μ and variance θ and we are interested in testing the composite null hypothesis $\theta = 1$ against $\theta > 1$. We would use the statistic of sample variance defined in (13.1). The distribution of ns^2 is a chi-square with $n-1$ dgrees of freedom. We would use a critical region of the form $ns^2 > a$.

3. Suppose we have two sets of independent obsrvations from two normal poulations $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ from two normal poulations with means μ_1 and μ_2 and variances θ_1 and θ_2. The null hypothesis is $\theta_1 = \theta_2$ and the alternate may be $\theta_2 > \theta_1$. The 'F 'statistic is used here.

$$F = \frac{n_1 s_1^2}{n_2 s_2^2}\frac{(n_2-1)}{(n_1-1)}$$

has an F- distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. We will of course use a critical region of the form $F < a$.

14 Sampling Distributions.

It is clear that in designing tests based on various statistics, the distribution of these statistics become relevent. We will now collect some commonly known facts, mostly about statistics based on observations from normal populations.

1. If x is normally distributed with mean μ and variance $\theta = \sigma^2$, then the distribution of $y = \frac{x-\mu}{\sigma}$ is the normal with mean 0 and variance 1, which is often called the standard normal.

2. If $x_1, \ldots, x_n$ are independent normals with mean μ and variance σ^2 any linear combination $\sum_i a_i x_i$ is again normal with mean $(\sum_i a_i)\mu$ and variance $(\sum_i a_i^2)\sigma^2$.

3. Any a_i with $\sum_i a_i$ has a normal distribution with mean 0 and variance $(\sum_i a_i^2)\sigma^2$. In particular the distribution is independent of μ. Such linear functions are called contrasts.

4. The square of a standard normal is a Gamma$(\frac{1}{2}, \frac{1}{2})$.

$$P[x^2 \leq a] = 2\int_0^{\sqrt{a}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

   Differentiate with repect to a to obtain the correct Gamma density.

5. The sum of two independent random variables with distributions Gamma(α, k_1) and Gamma(α, k_2) is distributed according to Gamma$(\alpha, k_1 + k_2)$. The best way to see this is to use generating functions

$$\frac{\alpha^k}{\Gamma(k)} \int_0^\infty e^{-\lambda x} e^{-\alpha x} x^{k-1} dx = \frac{\alpha^k}{(\alpha + \lambda)^k}$$

   For an independent sum the generating function is the product of the individual generating functions.

6. Therefore the sum os squares of n standard normals is a Gamma $(\frac{1}{2}, \frac{n}{2})$ which is called a chi-square with n degrees of freedom. The degrees of freedom refers to the number of independent normals that have been squared and added. It is written as χ^2_n. Note that $E[\chi^2_n] = n$.

7. If we start from n independent standard normals and make an orthogonal linear transformation $y_i = \sum_{j=1}^n a_{i,j}x_j$ then the $\{y_i\}$ are again independent standard normals. To see this we use the fact

$$\frac{1}{(\sqrt{2\pi})^n}\exp[-\frac{1}{2}\sum x_i^2]\Pi dx_i = \frac{1}{(\sqrt{2\pi})^n}\exp[-\frac{1}{2}\sum y_i^2]\Pi dy_i$$

8. In particular we can take $a_{1,j} = \frac{1}{\sqrt{n}}$ for all j, and complete the rest of the matrix $\{a_{i,j}\}$ in any manner to be orthogonal. Then $y_1 = \sqrt{n}\bar{x}_n$ and $\sum_2^n y_i^2 = \sum_1^n y_i^2 - y_1^2 = \sum_1^n x_i^2 - n\bar{x}_n^2 = ns^2$, where s^2 is the sample variance (13.1). In particular $\bar{x}_n$ and ns^2 are independent and the distribution of ns^2 is a chi-square with $n-1$ degrees of freedom.

9. If we assume that $x_1, \ldots, x_n$ are normal with mean μ and variance 1, since $y_2, \ldots, y_n$ are contrasts, the distribution of $y_2, \ldots, y_n$ and ns^2 are independent of μ.

10. The distribution of $t_k = \frac{x}{\sqrt{\frac{\chi_k^2}{k}}} = \frac{x}{\sqrt{\chi_k^2}}\sqrt{k}$ can be calculated. We start with

$$f(x,S) = \frac{1}{\sqrt{2\pi}}\frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}e^{-\frac{x^2}{2}}e^{-\frac{S}{2}}S^{\frac{k}{2}-1}dxdS$$

which is the joint distribution of a standrad normal x and S which is a χ^2 with k degrees of freedom. We make a transformation from (x,S) to (t,S) where $x = t[\frac{S}{k}]^{\frac{1}{2}}$. The joint density of t and S is given by

$$f(t,S)dtdS = \frac{1}{\sqrt{k}\sqrt{2\pi}}\frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}e^{-\frac{t^2S}{2k}}e^{-\frac{S}{2}}S^{\frac{k-1}{2}}dtdS$$

If we integrate S out, we get the density of t as,

$$\begin{aligned} f(t)dt &= \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k}\sqrt{\pi}\Gamma(\frac{k}{2})}\frac{1}{(1+\frac{t^2}{k})^{\frac{k+1}{2}}}dt \\ &= \frac{1}{\beta(\frac{1}{2},\frac{k+1}{2})}\frac{1}{(1+\frac{t^2}{k})^{\frac{k+1}{2}}}\frac{dt}{\sqrt{k}} \end{aligned}$$

which the density of a 't' with k degrees of freedom.

11. We note that if $x_1, \dots, x_n$ are independent normals with mean 0 and variance σ, we can replace x_i by $y_i = \frac{x_i}{\sigma}$ and since t is scale invariant the distribution of t does not depend on σ.

12. Suppose we have two independent sets of observations from two normal populations with the the same variance, but perhaps with different means. Their sizes are n_1 and n_2 respectively and the two sample variances are s_1^2 and s_2^2. The 'F' is the ratio $F = \frac{\frac{n_1 s_1^2}{n_1-1}}{\frac{n_2 s_2^2}{n_2-1}}$ It is of the form $F = \frac{\frac{S_1}{k_1}}{\frac{S_2}{k_2}}$ where S_1 and S_2 are two independent chi-squares with k_1 and k_2 degrees of freedom. We can compute the density of F_{k_1,k_2}. We begin with

$$f(S_1, S_2)dS_1 dS_2 = \frac{1}{2^{\frac{k_1+k_2}{2}}\Gamma(\frac{k_1}{2})\Gamma(\frac{k_2}{2})} e^{-\frac{1}{2}(S_1+S_2)} S_1^{\frac{k_1}{2}-1} S_2^{\frac{k_2}{2}-1} dS_1 dS_2$$

We change variables from (S_1, S_2) to $(U = \frac{S_1}{S_2}\frac{k_2}{k_1}, S_2)$ to get the joint density

$$f(U, S_2)dU dS_2 = c_{k_1,k_2} e^{-\frac{1}{2}(U\frac{k_1}{k_2}S_2+S_2)} U^{\frac{k_1}{2}-1} S_2^{\frac{k_1+k_2}{2}-1} dS_1 dS_2$$

where c_{k_1,k_2} is the normalizing constant that we will not bother to keep track of. Integrating S_2 out produces the density of F distribution that depends on two parametrs k_1 and k_2.

$$f_{k_1,k_2}(U) = c'_{k_1,k_2} \frac{U^{\frac{k_1}{2}-1}}{(1+\frac{k_1}{k_2}U)^{\frac{k_1+k_2}{2}}}$$

15 Testing Composite Hypotheses.

In general if we have a composite null hypothesis of the form $\theta \in \Theta_0$ that we want to test a statistic that is often reasonable is the **likelihood ratio criterion** defined below.

$$\lambda = \frac{\sup_{\theta\in\Theta_0} L(\theta, x)}{\sup_{\theta\in\Theta} L(\theta, x)} \tag{15.1}$$

Clearly $0 < \lambda \leq 1$ and smaller the value of λ the less confident we are of our null hypothesis. Therefore the critical region is of the form $\lambda \leq c$.

Let us look at a few examples.

1. Suppose we have n observations from $N(\mu, \sigma^2)$ and we want to test $\mu = 0$ against the alternative $\mu \neq 0$.

$$L(\mu, \sigma^2, x_1, \dots, x_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp[-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2]$$

   Under the null hypothesis the MLE for σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_i x_i^2$. and the maximum of L is

$$L(0, \hat{\sigma}^2, x_1, \dots, x_n) = \frac{1}{(\sqrt{2\pi}S)^n} \exp[-\frac{n}{2}$$

   where $S = \sqrt{\frac{1}{n}\sum_i x_i^2}$. A similar calculation with out any assumptions on μ gives $\hat{\mu} = \bar{x}_n$ and $\hat{\sigma} = s = \sqrt{\frac{1}{n}\sum_i (x_i - \bar{x}_n)^2}$. This provides

$$L(\hat{\mu}, \hat{\sigma}^2, x_1, \dots, x_n) = \frac{1}{(\sqrt{2\pi}s)^n} \exp[-\frac{n}{2}]$$

   so that

$$\lambda = [\frac{s}{S}]^n$$

   We can use any monotonic functnction of λ and

$$|t| = |\frac{\bar{x}_n}{s}|\sqrt{n-1} = \sqrt{\frac{S^2 - s^2}{s^2}}\sqrt{n-1} = \sqrt{\lambda^{-\frac{2}{n}} - 1}\sqrt{n-1}$$

   is a monotonic decreasing function of λ.

2. Suppose we have n_1 observations from $N(\mu_1, \sigma_1^2)$ and n_2 observations from $N(\mu_2, \sigma_2^2)$. We wish to test $\sigma_1^2 = \sigma_2^2$. A similar calculation yields

$$\lambda = \frac{s_1^{n_1} s_2^{n_2}}{S^{n_1+n_2}} = \frac{R^{n_1}}{(\frac{n_2+n_1R^2}{n_2+n_1})^{\frac{n_1+n_2}{2}}}$$

   where s_1^2 and s_2^2 are the two sample variances, $S^2 = \frac{n_1s_1^2+n_2s_2^2}{n_1+n_2}$ and λ is written as function of the ratio $R = \frac{s_1}{s_2}$ which leads to the F test. Note that because λ is not monotone in R, the region $\lambda < c$ splits into two regions $R < r_0$ and $R > r_1$ giving us a two sided critial region for the F test.

16 Large Sample Tests.

When we have to test a simple hypothesis that $\theta = \theta_0$ against an alternative that may be one or two sided, it is natural to base the test on the maximum likelihood estimator $\hat{\theta}_n$ of θ based on n independent observations. We can look at the statistic

$$U_n = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{I(\hat{\theta}_n)}$$

The consistency of θ_n and the asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ imply that the distrbution of U_n converges to $N(0,1)$ as $n \to \infty$. We can base our tests on U_n.

Suppose we have a set of m parameters $\theta = \{\theta_1, \ldots, \theta_m\}$, and we want to test the composite hypothesis $\mathcal{H}_0 = \{\theta : \theta_1 = \theta_2 = \cdots = \theta_k = 0\}$ against the alternative that at least one of them is nonzero. The remaining $m-k$ parameters are some times called nuisance parameters. We can obtain maximum likelihood estimators $\hat{\theta}_1, \ldots, \hat{\theta}_m$ for all the parameters, based on n observations. With out loss of generality we can assume that the true value of all the parameters are 0. It is natural to base the test on $U_j = \sqrt{n}\hat{\theta}_j$ and take as critical region a set D in R^m and reject the null hypothesis if $U = (U_1, \ldots, U_m) \in D$. The joint distribution of the full set $U = (U_1, \ldots, U_m)$ is asymptotically the m variate normal $N(0, I^{-1}(0))$. If we take $J(0)$ as the square root of $I^{-1}(0)$, then $V = J(\hat{\theta})U$ has for its asymptotic distribution the standard multivariate normal $N(0, I)$. On R^m we have the $m-k$ dimensional subspace $S = \{U_1 = U_2 = \cdots = U_k = 0\}$. We need to measure how far U is from being in S and the D will consist of points that are too far away. The obvious step is to use orthogonal projections and define

$$Z = \inf_{u \in S} \|J(\hat{\theta})(U - u)\|^2 \simeq \inf_{v \in J_0 S} \|V - v\|^2$$

If X is distributed as a mutivariate normal $N(0, I)$ on R^m, then $\|X\|^2 = \sum_i x_i^2$ is distributed as a chi-square with m degrees of freedom. Suppose S is any subspace of dimension $m-k$, we can choose orthonormal coorinates $\{y_i\}$ such that $S = \{y : y_1 = y_2 \cdots = y_k = 0\}$ and

$$Z = \inf_{u \in S} \|x - u\|^2 = \sum_{i=1}^{k} y_i^2$$

is distributed as a chi-square with k degrees of freedom. We therefore conclude that asymptotically Z is a chi-square with k degrees of freedom. The critical region could then be of the form $Z > c$.

In fact far large samples the likelihood ratio criterion produces a test statistic of the form.

$$\begin{aligned}
-\log\lambda &= \sup_{\theta\in\Theta}\sum_i \log L(\theta, x_i) - \sup_{\theta\in\Theta_0}\sum_i \log L(\theta, x_i) \\
&= \sum_i \log L(\hat\theta_n, x_i) - \sup_{\theta\in\Theta_0}\sum_i \log L(\theta, x_i) \\
&\simeq \inf_{\theta\in\Theta_0} -\frac{1}{2}\sum_{1\le r,s\le m}[\frac{1}{n}\sum_i \frac{\partial^2 \log L(\theta, x_i)}{\partial\theta_r\partial\theta_s}]_{\theta=0}\sqrt{n}(\hat\theta_n^r - \theta_r)\sqrt{n}(\hat\theta_n^s - \theta_s) \\
&\simeq \inf_{u\in S}\frac{1}{2} < I(0)(U-u), (U-u) > \\
&= \inf_{v\in J(0)S}\frac{1}{2} < (V-v), (V-v) >
\end{aligned}$$

Therefore $-2\log\lambda$ is asymptotically a chi-square with k degrees of freedom.
Examples.

1. **Multinomial Ditributions.** Suppose we have categorical data of size N, divided into k categories with observed frequenies $\{f_i : 1 \le i \le k\}$ adding up to N. We want to test that the probabilties for the individual categories are given by $\{p_i : 1 \le i \le k\}$. The multinomial likelihood is

$$L(\pi_1, \pi_2, \dots, \pi_k) = \frac{N!}{n_1!n_2!\dots n_k!}\pi_1^{f_1}\cdots\pi_k^{f_k}$$

The MLE are $\hat\pi_i = \frac{f_i}{N}$.

$$\begin{aligned}
-2\log\lambda &= 2\sum f_i \log\hat\pi - 2\sum_i f_i \log p_i \\
&= N\sum_i \frac{(\hat\pi - p_i)^2}{p_i} \\
&= \sum_i \frac{f_i^2}{Np_i} - N
\end{aligned}$$

is a chi-square with $(k-1)$ degrees of freedom.

2. **Goodness of fit.** Here we want to test that $\pi_i = \pi_i(\theta)$ for some value of θ, where the functions $\pi_i(\cdot)$ are given. $\mathcal{H}_0$ is the range of $\{\pi_i(\theta)\}$ as θ varies over the admissible values. Calculate the MLE for θ from the likelihood function

$$L(\theta, f_1, \dots, f_k) = \frac{N!}{n_1!n_2!\dots n_k!}\pi_1(\theta)^{f_1}\cdots\pi_k(\theta)^{f_k}$$

Computation yields again

$$-2\log\lambda = \sum_i \frac{f_i^2}{N\pi(\hat{\theta})} - N$$

which is now a chi-square with $k-i-d$ degrees of freedom, where d is the dimension of the space pf parameters d.

2. **Testing for Associaton.** Suppose we have data in a two-way classification. Say for example a sample of size N classified into 9 categories according to two types of classification. One for smoking habits (no or light, moderate, heavy) and the other for respiratory problems (light or none, fair amount, serious). The frequencies $f_{i,j}$ are obtained from a study. The tobacco industry claims that if there is any association it is due to chance. How do you test? We use the multinomial model with $\pi_{i,j}$ for probabilities, which is 8 dimensional. Under null hypothesis $\pi_{i,j} = p_i q_j$, that claims that statistiaclly, smoking and respiratory problems have nothing to do with each other. The null hypothesis is described by 4 parameters. Our goodness of fit statistic will be a chi-square with 4 degrees of freedom.

3. **Two-by-two table.** Here we have just four frequencies arranged in a two-by-two table

f_{11}	f_{12}
f_{21}	f_{22}

A calculation shows that the statistic χ^2 with 1 degree of freedom is equal to

$$\frac{N(f_{11}f_{22} - f_{12}f_{21})^2}{(f_{11}+f_{12})(f_{11}+f_{21})(f_{12}+f_{22})(f_{21}+f_{22})}$$