Chapter 7. Hypothesis Testing

1. The General Problem

It is often necessary to make a decision, on the basis of available data from an experiment (carried out by yourself or by Nature), on whether a particular proposition H_o (theory, model, hypothesis) is true, or the converse H_1 is true. This decision problem is often encountered in scientific investigation. Economic examples of hypotheses are

(a) The commodities market is efficient (i.e., opportunities for arbitrage are absent).

(b) There is no discrimination on the basis of gender in the market for academic economists.

(c) Household energy consumption is a necessity, with an income elasticity not exceeding one.

(d) The survival curve for Japanese cars is less convex than that for Detroit cars.

Notice that none of these economically interesting hypotheses is framed directly as a precise statement about a probability law (e.g., a statement that the parameter in a family of probability densities for the observations from an experiment takes on a specific value). A challenging part of statistical analysis is to set out <u>maintained hypotheses</u> that will be accepted by the scientific community as true, and which in combination with the proposition under test give a probability law.

Deciding the truth or falsity of a <u>null</u> hypothesis H_o presents several general issues: the cost of mistakes, the selection and/or design of the experiment, and the choice of the test.

2. The Cost of Mistakes

Consider a two-by-two table that compares the truth with the result of the statistical decision. For now, think of each of the alternatives H_o and H_1 as determining a unique probability law for the observations; these are called simple alternatives. Later, we consider compound hypotheses or alternatives that are consistent with families of probability laws. There are two possible mistakes, *Type I* in which a true hypothesis is rejected, and *Type II* in which a false hypothesis is accepted. There are costs associated with these mistakes -- let C_I denote the cost associated with a Type I mistake, and C_{II} denote the cost associated with a Type II mistake. If the hypothesis is true, then there is a probability α that a particular decision procedure will result in rejection; this is also called the *Type I error probability* or the *significance level*. If the hypothesis is false, there is a probability β that it will be accepted; this is also called the *Type II error probability*. The probability $\pi \equiv 1-\beta$ is the probability that the hypothesis will be rejected when it is false, and is called the *power* of the decision procedure.

		Truth H_o	**Truth** H_1
Decision	**H_o accepted**	cost = 0 prob. = $1-\alpha$	cost = C_{II} prob. = β
Decision	**H_1 accepted**	cost = C_I prob. = α	cost = 0 prob. = $\pi \equiv 1-\beta$

This figure is in principle completely symmetric between the states H_o and H_1: You can call your favorite theory H_o and hope the evidence leads to it being accepted, or call it H_1 and hope the evidence leads to H_o being rejected. However, classical statistical analysis is oriented so that α is chosen by design, and β requires a sometimes complex calculation. Then, the Type I error is easier to control. Thus, in classical statistics, it is usually better to assign your theory between H_o and H_1 so that the more critical mistake becomes the Type I mistake. For example, suppose you set out to test your favorite theory. Your study will be convincing only if your theory passes a test which it would have a high (and known) probability of failing if it is in fact false. You can get such a stringent test by making your theory H_1 and selecting a null and a decision procedure for which α is known and small; then your theory will be rejected with large known probability $1-\alpha$ if it is in fact false. (This will not work if you pick a "straw horse" for the null that no one thinks is plausible.) Conversely, if you set out to do a convincing demolition of a theory that you think is false, then make it the null, so that there is a small known probability α of rejecting the hypothesis if it is in fact true.

A common case for hypothesis testing is that the null hypothesis H_o is simple, but the alternative hypothesis H_1 is compound, containing a family of possible probability laws. Then, the probability of a Type II error depends on which member of this family is true. Thus, the power of a test is a <u>function</u> of the specific probability law in a compound alternative. When both the null hypothesis and alternative are compound, the probability of a Type I error is a function of which member of the family of probability laws consistent with H_o is true. In classical statistics, the <u>significance</u> <u>level</u> is always defined to be the "worst case": the largest α for any probability law consistent with the null.

Given the experimental data available and the statistical procedure adopted, there will be a trade off between the probabilities of Type I and Type II errors.

When the cost C_I is much larger than the cost C_{II}, a good decision procedure will make α small relative to β. Conversely, when C_I is much smaller than C_{II}, the procedure will make α large relative to β. For example, suppose the null hypothesis is that a drug is sufficiently safe and effective to be released to the market. If the drug is critical for treatment of an otherwise fatal disease, then C_I is much larger than C_{II}, and the decision procedure should make α small. Conversely, a drug to reduce non-life-threatening wrinkles should be tested by a procedure that makes β small.

3. Design of the Experiment

One way to reduce the probability of Type I and Type II errors is to collect more observations by increasing sample size. One may also by clever design be able to get more information from a given sample size, or more relevant information from a given data collection budget. One has the widest scope for action when the data is being collected in a laboratory experiment that you can specify. For example, the Negative Income Experiments in the 1960's and 1970's were able to specify experimental treatments that presented subjects with different trade offs between wage and transfer income, so that labor supply responses could be observed. However, even in investigations where only natural experiments are available, important choices must be made on what events to study and what data to collect. For example, if a survey of 1000 households is to be made to determine the income elasticity of the demand for energy, one can get more precision by oversampling high income and low income households to get a greater spread of incomes.

There is an art to designing experiments or identifying natural experiments that allow tests of a null hypothesis without confounding by extraneous factors. For example, suppose one wishes to test the null hypothesis that Japanese cars have the same durability as Detroit cars. One might consider the following possible experiments:

(a) Determine the average age, by origin, of registered vehicles.

(b) Sample the age/make of scrapped cars as they arrive at junk yards.

(c) Draw a sample of individual new cars, and follow them longitudinally until they are scrapped.

(d) Draw a sample of individual new cars, and operate them on a test track under controlled conditions until they fail.

Experiment (a) is confounded by potential differences in historical purchase patterns; some of this could be removed by econometric methods that condition on the number of original purchases in earlier years. Experiments (a)-(c) are confounded by possible variations in usage patterns (urban/rural, young/old, winter roads/not). For example, if rural drivers who stress their cars less tend to buy Detroit cars, this factor rather than the intrinsic durability of the cars might make Detroit cars appear to last longer. One way to reduce this factor would be to assign drivers to car models randomly, as might be done for example for cars rented by Avis in the "compact" category. The ideal way to do this is a "double blind" experiment in which neither the subject nor the data recorder knows which "treatment" is being received, so there is no possibility that bias in selection or response could creep in. Most economic experimental treatments are obvious to aware subjects, so that "double blind" designs are impossible. This puts an additional burden on the researcher to carefully randomize assignment of treatments and to structure the treatments so that their form does not introduce factors that confound the experimental results.

4. Choice of the Decision Procedure

Suppose one thinks of hypothesis testing as a statistical decision problem, with a prior p_o that H_o is true and $p_1 = 1 - p_o$ that H_1 is true. Let $f(\mathbf{x}|H_o)$ denote the likelihood of $\mathbf{x}$ if H_o is true, and $f(\mathbf{x}|H_1)$ denote the likelihood if H_1 is true. Then, the posterior likelihood of H_o given $\mathbf{x}$ is, by application of Bayes Law, $q(H_o|\mathbf{x}) = f(\mathbf{x}|H_o)p_o/[f(\mathbf{x}|H_o)p_o + f(\mathbf{x}|H_1)p_1)]$. The expected cost of rejecting H_o given $\mathbf{x}$ is then $C_I q(H_o|\mathbf{x})$, and the expected cost of accepting H_o given $\mathbf{x}$ is $C_{II}q(H_1|\mathbf{x})$. The optimal decision rule is then to *reject* H_o for $\mathbf{x}$ in the *critical region* $\mathbf{C}$ where $C_I q(H_o|\mathbf{x}) < C_{II}q(H_1|\mathbf{x})$. This inequality simplifies to $C_I f(\mathbf{x}|H_o)p_o < C_{II}f(\mathbf{x}|H_1)p_1$, implying

$$\mathbf{x} \in \mathbf{C} \text{ (i.e., } \textit{reject } H_o\text{) if and only if } f(\mathbf{x}|H_1)/f(\mathbf{x}|H_o) > k \equiv C_I p_o/C_{II}p_1 \ .$$

The expression $f(\mathbf{x}|H_1)/f(\mathbf{x}|H_o)$ is termed the *likelihood ratio*. The optimal criterion is then to reject H_o if and only if the likelihood ratio exceeds a threshold k. The larger C_I or p_o, the larger this threshold.

A classical statistical treatment of this problem will also pick a *critical region* $\mathbf{C}$ of $\mathbf{x}$ for which H_o will be rejected, and will do so by maximizing power $\pi = \int_{\mathbf{C}} f(\mathbf{x}|H_1)d\mathbf{x}$ subject to the constraint $\alpha = \int_{\mathbf{C}} f(\mathbf{x}|H_o)d\mathbf{x}$. But this is accomplished by picking $\mathbf{C} = \{\mathbf{x}|f(\mathbf{x}|H_1)/f(\mathbf{x}|H_o) > k\}$, where k is a constant chosen so the constraint is satisfied. Hence, the optimal decision rule for various prior probabilities and costs and the classical statistical test procedure trace out the same families of procedures, and will coincide when the critical likelihood ratio k in the two approaches is the same.

In more general classical hypothesis testing situations where the alternative is compound, there is no longer an exact coincidence of the classical and statistical decision theory approaches to decisions. However, the likelihood ratio often remains a useful basis for constructing good test procedures, and in many cases a "best" test by some classical statistical criterion and a test utilizing the likelihood ratio criterion will be the same or nearly the same. In general, we will consider DGP which we maintain are members of a family $f(\mathbf{x},\theta)$ indexed by a parameter θ. The null hypothesis is that the true value θ_o of θ is contained in a set Θ_o, and the alternative is that it is contained in a set Θ_1, with Θ_o and Θ_1 partitioning the universe Θ of possible values of θ_o. A likelihood ratio critical region is usually defined in this general framework as a set of the form

$$\mathbf{C} = \{\mathbf{x} \mid \sup_{\theta\in\Theta_1} f(\mathbf{x},\theta) / \sup_{\theta\in\Theta_o} f(\mathbf{x},\theta) > k\} \ .$$

The likelihood ratio in this criterion is less than or equal to one when the maximum likelihood estimator of θ_o falls in Θ_o, and otherwise is greater than one. Then a critical region defined for some $k > 1$ will include the observed vectors $\mathbf{x}$ that are the least likely to have been generated by a DGP with a parameter in Θ_o. The significance level of the test is set by adjusting k. An equivalent way to characterize the likelihood ratio critical region $\mathbf{C}$ is in terms of the log likelihood function,

$$\mathbf{C} = \{\mathbf{x} \mid \sup_{\theta\in\Theta_1} \log f(\mathbf{x},\theta) - \sup_{\theta\in\Theta_o} \log f(\mathbf{x},\theta) > \kappa = \log k\} \ .$$

The analyst will often have available alternative testing procedures in a classical testing situation. For example, one procedure to test a hypothesis about a location parameter might be based on the sample mean, a second might be based on the sample median, and a third might be based on the likelihood ratio. Some of these procedures may be better than others in the sense of giving higher power for the same significance level. The ideal, as in the simple case, is to maximize the power given the significance level. When there is a compound alternative, so that power is a function of the alternative, one may be able to tailor the test to have high power against alternatives of particular importance. In a few cases, there will be a single procedure that will have uniformly best power against a whole range of alternatives. If so, this will be called a *uniformly most powerful* test.

The figure on the next page shows power functions for some alternative test procedures. The null hypothesis is that a parameter θ is zero. Power curves A and B equal 0.05 when H_o: $\theta = 0$ is true. Then, the significance level of these three procedures is $\alpha = 0.1$. The significance level of D is much higher, 0.5. Compare the curves A and B. Since A lies everywhere above B and has the same significance level, A is clearly the superior procedure. A comparison like A and B most commonly arises when A uses more data than B; that is, A corresponds to a larger sample. However, it is also possible to get a picture like this when A and B are using the same sample, but B makes poor use of the information in the sample.

Compare curves A and C. Curve C has significance level $\alpha = 0.1$, and has lower power than A against alternatives less than $\theta = 0$, but better power against alternatives greater than $\theta = 0$. Thus, A is a better test it we want to test against all alternatives, while C is a better test if we are mainly interested in alternatives to the right of $\theta = 0$ (i.e., we mostly want to test H_o: $\theta \le 0$). Compare curves A and D. Curve D has high power, but at the cost of a high probability of a Type I error. Thus, A and D represent a trade off between Type I and Type II errors.

Finally, suppose we are most interested in the alternative H_1: $\theta = 1.0$. The procedure giving curve A has power 0.52 against this alternative, and hence has a reasonable chance of discriminating between H_o and H_1. On the other hand, the procedure B has power 0.29, and much less chance of discriminating. We would conclude that the procedure A is a satisfactory statistical analysis, while B is of limited use.

5. Hypothesis Testing in Normal Populations

This section provides a summary of hypothesis test calculations for standard setups involving data drawn from a normal population, including power calculations. Assume that we start from a simple random sample of size n, giving i.i.d. observations $x_1,...,x_n$. Recall from Chapter 6.3 that the log likelihood of a normal random sample is

$$L(\mathbf{x},\mu,\sigma^2) = -\frac{n}{2}\cdot\text{Log}(2\pi) - \frac{n}{2}\cdot\text{Log }\sigma^2 - \frac{1}{2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2$$

$$= -\frac{n}{2}\cdot\text{Log}(2\pi) - \frac{n}{2}\cdot\text{Log }\sigma^2 - \frac{1}{2}\cdot\frac{(n-1)s^2}{\sigma^2} - \frac{n}{2}\cdot\frac{(\bar{x}-\mu)^2}{\sigma^2} \equiv n\cdot L(\bar{x},s^2,\mu,\sigma^2) .$$

where the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ are unbiased estimators of μ and σ^2, respectively. If Θ_o denotes the set of parameter values (μ,σ^2) consistent with a null hypothesis and Θ_1 denotes the set consistent with the alternative, then a likelihood ratio critical set for this hypothesis will take the form

$$\mathbf{C} = \{(\bar{x},s^2) \mid \sup_{\Theta_1} L(\bar{x},s^2,\mu,\sigma^2) - \sup_{\Theta_o} L(\bar{x},s^2,\mu,\sigma^2) > \kappa\} .$$

We consider a sequence of hypotheses and conditions. See Chapter 3.7 for the densities and other properties of the distributions used in this section, and Chapter 6.3 for the relation of these distributions to data from a normal population. The following table gives the statistical functions that are available in many econometrics software packages; the specific notation is that used in the Statistical Software Tools (SST) package. Tables at the end of most statistics texts can also be used to obtain values of the central versions of these distributions. The direct functions give the CDF probability for a specified argument, while the inverse functions give the argument that yields a specified probability.

Distribution	CDF	Inverse CDF
Normal	cumnorm(x)	invnorm(p)
Chi-Square	cumchi(x.df)	invchi(p,df)
F-Distribution	cumf(xdf1,df2)	invf(p,df1,df2)
T-Distribution	cumt(x,df)	invt(p,df)
Non-central Chi-Square	cumchi(x,df,λ)	NA
Non-central F-Distribution	cumf(x,df1,df2,λ)	NA
Non-Central T-Distribution	cumt(x,df,δ)	NA

In this table, df denotes degrees of freedom, and λ and δ are non-centrality parameters. Inverse CDF's are not available for non-central distributions in most packages, and are not needed. In most statistical packages, values of these functions can either be printed out or saved for further calculations. For example, in SST, the command calc cumnorm(1.7) will print out the probability that a standard normal random variable is less than 1.7, the command calc p = calc cumnorm(1.7) will store the result of this calculation in the variable p for further use, and a subsequent command calc p will also print out its value.

Problem 1: Testing the mean of a normal population that has known variance

Suppose a random sample of size n from a normal population with an unknown mean μ and a known variance σ^2. The null hypothesis is H_o: $\mu = \mu_o$, and the alternative is H_1: $\mu \neq \mu_o$. Verify that the likelihood ratio, $\text{Max}_\mu n(\mathbf{x}|\mu,1)/n(\mathbf{x}|\mu_o,1)$, is an increasing function of $(\bar{x} - \mu_o)^2$. Hence, a test equivalent to a likelihood ratio test can be based on $(\bar{x} - \mu_o)^2$. From Chapter 6.3(8), one has the result that under the null hypothesis, the statistic $(\bar{x} - \mu_o)^2/(\sigma^2/n) \equiv n(\bar{x} - \mu_o)^2/\sigma^2$ is distributed

χ_1^2. Alternately, from Chapter 6.3(5), the square root of this expression, $\dfrac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$, has a standard normal distribution.

Using the Chi-Square form of the statistic, the critical region will be values exceeding a critical level z_c, where z_c is chosen so that the selected significance level α satisfies $\chi_1^2(z_c) = 1 - \alpha$. For example, taking $\alpha = 0.05$ yields $z_c = 3.84146$. This comes from a statistical table, or from the SST command calc invchi(1-α,k), where α is the significance level and k is degrees of freedom. The test procedure <u>rejects</u> H_o whenever

(1) $$n(\bar{x} - \mu_o)^2/\sigma^2 > z_c = 3.84146 \;.$$

Consider the power of the Chi-square test against an alternative such as $\mu = \mu_1 \neq \mu_o$. The non-centrality parameter is

(2) $$\lambda = (\mu_1 - \mu_o)^2/(\sigma^2/n) \;.$$

For example, if $\mu_1 - \mu_o = 1.2$, $\sigma^2 = 25$, and $n = 100$, then $\lambda = 1.44 \cdot 100/25 = 5.76$. The power is calculated from the non-central Chi-square distribution (with 1 degree of freedom), and equals the probability that a random draw from this distribution exceeds z_c. This probability is readily calculated using the SST command π = calc 1 - cumchi(z_c,k,λ). In the example, π = calc 1 - cumchi(3.84146,1,5.76) = 0.67006. Then, a test with a five percent significance level has power of 67 percent against the alternative that the true mean is 1.2 units larger than hypothesized.

An equivalent test can be carried out using the standard normal distributed form $(\bar{x} - \mu_o)/(\sigma/\sqrt{n})$. The critical region will be values of this expression that in magnitude exceed a critical level w_c, where w_c is chosen for a specified significance level α so that a draw from a standard normal density has probability $\alpha/2$ of being below $-w_c$, and symmetrically a probability $\alpha/2$ of being above $+w_c$. One can find w_c from statistical tables, or by using a SST command w_c = calc invnorm(1-α/2). For example, if $\alpha = 0.05$, then w_c = calc invnorm(0.975) = 1.95996. The test <u>rejects</u> H_o whenever

(3) $$(\bar{x} - \mu_o)/(\sigma/\sqrt{n}) < -w_c \ \underline{\text{or}}\ (\bar{x} - \mu_o)/(\sigma/\sqrt{n}) > w_c \ .$$

For example, if n = 100, $\sigma = 5$, and $\mu_o = 0$, the critical region for a test with significance level $\alpha = 0.05$ is $10\bar{x}/5 < -1.95996$ or $10\bar{x}/5 > +1.95996$. Note that $w_c^2 = z_c$, so this test rejects exactly when the Chi-square test rejects. The power of the test above against the alternative $\mu = \mu_1 \neq \mu_o$ is the probability that the random variable $(\bar{x} - \mu_o)/(\sigma/\sqrt{n})$ lies in the critical region when $\bar{x} \sim N(\mu_1,\sigma^2)$. In this case, $(\bar{x} - \mu_1)/(\sigma/\sqrt{n}) \equiv Y$ is standard normal, and therefore $(\bar{x} - \mu_o)/(\sigma/\sqrt{n}) \equiv Y + \delta$, where

(4) $$\delta = (\mu_1 - \mu_o)/(\sigma/\sqrt{n}) \ .$$

Note that $\delta^2 = \lambda$, where λ is given in (2). The probability of rejection in the left tail is $\Pr((\bar{x} - \mu_o)/(\sigma/\sqrt{n}) < -w_c|\mu = \mu_1) = \Pr(Y < - w_c - \delta)$. For the right tail, $\Pr((\bar{x} - \mu_o)/(\sigma/\sqrt{n}) > w_c|\mu = \mu_1) = \Pr(Y > w_c - \delta)$. Using the fact that the standard normal is symmetric, we then have

(5) $$\pi = \Phi(-w_c - \delta) + 1 - \Phi(w_c - \delta) \equiv \Phi(-w_c - \delta) + \Phi(-w_c + \delta) \ .$$

This can be calculated using the SST command

(6) π = calc cumnorm($-w_c-\delta$) + cumnorm($-w_c+\delta$) .

For example, $\sigma = 5$, N = 100, $\mu_1 - \mu_o = 1.2$, $w_c = 1.95996$ give $\delta = 2.4$ and power π = calc cumnorm($-w_c$ - 2.4) + cumnorm($-w_c$ + 2.4) = 0.670. Note this is the same as the power of the Chi-square version of the test.

Suppose that instead of testing the null hypothesis H_o: $\mu = \mu_o$ against the alternative H_1: $\mu \neq \mu_o$, you want to test the <u>one-sided</u> hypothesis H_o: $\mu \le \mu_o$ against the alternative H_1: $\mu > \mu_o$. The likelihood ratio in this case is $\sup_{\mu>\mu_o} n(\mathbf{x}|\mu,\sigma^2)/\sup_{\mu\le\mu_o} n(\mathbf{x}|\mu,\sigma^2)$, which is constant for $\bar{x} \le \mu_o$ and is monotone increasing in $(\bar{x} - \mu_o)$ for $\bar{x} > \mu_o$. Hence, a test that rejects H_o for $\bar{x} - \mu_o$ large appears desirable. This suggests using a test based on the statistic $(\bar{x}-\mu_o)/(\sigma/\sqrt{n})$, which is normal with variance one, and has a non-positive mean under the null. Pick a critical level $w_c > 0$ such that

$$\sup_{\mu\le\mu_o} \text{Prob}((\bar{x}-\mu_o)/(\sigma/\sqrt{n}) > w_c) = \alpha \ .$$

Note that the sup is taken over all the possible true μ consistent with H_o, and that α is the selected significance level. The maximum probability of Type I error is achieved when $\mu = \mu_o$.[1] Then, w_c is determined to give probability α that a draw from a standard normal exceeds w_c. For example, if n = 100, $\alpha = 0.05$, $\sigma = 5$, and H_o is that $\mu \le 0$, then w_c = calc invnorm(0.95) = 1.64485. The power of the test of $\mu \le \mu_o = 0$ against the alternative $\mu = \mu_1 = 1.2$ is given by

$$\text{(6)} \qquad \begin{aligned} \pi = \ & \Pr((\bar{x}-\mu_o)/(\sigma/\sqrt{n}) > w_c|\mu = \mu_1) \equiv \Pr(Y \equiv (\bar{x}-\mu_1)/(\sigma/\sqrt{N}) > w_c - \delta) \\ & \equiv 1 - \Phi(w_c - \delta) \equiv \Phi(-w_c + \delta) \equiv \text{calc cumnorm}(-w_c + \delta) \ , \end{aligned}$$

where δ is given in (4). In the example, π = calc cumnorm(- 1.64485 + 2.4) = 0.775. Hence, a test which has a probability of at most $\alpha = 0.05$ of rejecting the null

---

[1] $\text{Prob}((\bar{x}-\mu_o)/(\sigma/\sqrt{n}) > w_c) \equiv \Pr(Y \equiv (\bar{x}-\mu)/(\sigma/\sqrt{N}) > w_c + (\mu_o - \mu)/(\sigma/\sqrt{N}))$, where μ is the true value. Since Y is standard normal, this probability is largest over $\mu \le \mu_o$ at $\mu = \mu_o$.

hypothesis when it is true has power 0.775 against the specific alternative $\mu_1 = 1.2$.

Problem 2. Testing the Mean of a Normal Population with Unknown Variance

This problem is identical to Problem 1, except that σ^2 must now be estimated. Use the estimator s^2 for σ^2 in the Problem 1 test statistics. From Chapter 6.3(8), the Chi-square test statistic with σ replaced by s, $F = (\bar{x} - \mu_o)^2/(s^2/n)$, has an F-distribution with degrees of freedom 1 and N-1. Hence, to test H_o: $\mu = \mu_o$ against the alternative H_1: $\mu \neq \mu_o$, find a critical level z_c such that a specified significance level α equals the probability that a draw from $F_{1,n-1}$ exceeds z_c. The SST function calc z_c = invf(1-α,1,n-1) gives this critical level; it can also be found in standard tables. For n = 100 and α = 0.05, the critical level is $z_c = 3.93694$.

The power of the test against an alternative μ_1 is the probability that the statistic F exceeds z_c. Under this alternative, F has a non-central F-distribution (from Chapter 3.9) with the non-centrality parameter $\lambda = (\mu_1 - \mu_o)^2/(\sigma^2/N)$ given in (2). Then, the power is given by

(7) π = calc 1 - cumf(z_c,1,N-1,λ) .

In the example with $\mu_1 - \mu_o = 1.2$ and $\sigma^2 = 25$, one has $\lambda = 144/25$, and the power is

(8) π = calc 1 - cumf(3.93694,1,99,144/25) = 0.662 .

The non-centrality parameter is defined using the <u>true</u> σ^2 rather than the estimate s^2. Calculating power at an <u>estimated</u> non-centrality parameter $\hat{\lambda} = (\mu_1 - \mu_o)^2/(s^2/n)$ introduces some error -- you will evaluate the power curve at a point somewhat different than you would like. For most practical purposes, you do not need an exact calculation of power; you are more interested in whether it is 0.1 or 0.9. Then, the error introduced by this approximation is harmless, and can be ignored. In particular, for sample sizes n > 100, this error is negligible.[2]

The test can be carried out equivalently using

(9) $$T = (\bar{x} - \mu_o)/(s/\sqrt{n}) ,$$

which by Chapter 6.3(7) has a t-distribution with n-1 degrees of freedom under H_o: $\mu = \mu_o$. For a significance level α, choose a critical level w_c, and reject the null hypothesis when $|T| > w_c$. The value of w_c satisfies $\alpha/2 = t_{n-1}(-w_c)$, and is given in standard tables, or in SST by w_c = invt(1-α/2,n-1). For the example $\alpha = 0.05$ and n = 100, this value is w_c = calc invt(.975,99) = 1.9842.

The power of the test is calculated as in Problem 1, replacing the normal distribution by the <u>non-central</u> <u>t-distribution</u>:

$$\pi = t_{n-1,\delta}(-w_c) + 1 - t_{n-1,\delta}(w_c) ,$$

where $\delta = (\mu_1 - \mu_o)/(\sigma/\sqrt{n})$ as in equation (4). Points of the non-central t are <u>not</u> in standard tables, but are provided by a SST function,

---

[2] Note that $\lambda/\hat{\lambda} = s^2/\sigma^2$, so $(N-1)\lambda/\hat{\lambda}$ is distributed χ^2_{n-1}. For the rare application where you really need to know how precise your power calculation is, you can form a <u>confidence</u> <u>interval</u> as follows: Given a "significance level" α, compute z_1 = calc invchi(α/2,n-1) and z_2 = calc invchi(1-α/2,n-1). Then, with probability α, $\lambda_1 \equiv z_1\hat{\lambda}/(n-1) < \lambda < z_1\hat{\lambda}/(n-1) \equiv \lambda_2$. The power π_1 calculated at λ_1 and the power π_2 calculated at λ_2 give a α-level confidence bound on the exact power. For example, $\alpha = 0.5$, N = 100, $\mu_1 - \mu_o = 1.2$, and $s^2 = 25$ imply $\hat{\lambda} = 144/25$, z_1 = calc invchi(.25,24) = 19.037, $\lambda_1 = 2.203$, and π_1 = calc 1-cumf(4.2597,1,24,2.203) = 0.297. Also, z_2 = calc invchi(.75,24) = 28.241, $\lambda_2 = 3.269$, and π_2 = calc 1-cumf(4.2597,1,24,3.269) = 0.411. Then, with probability 0.5, the exact power for the alternative $\mu_1 - \mu_o = 2$ is in the interval [0.297,0.411].

$$\pi = \text{cumt}(-w_c,n\text{-}1,\delta) + 1 - \text{cumt}(w_c,n\text{-}1,\delta) \ .$$

For the example $\alpha = 0.05$, N = 100, $\sigma = 5$, and $\mu_1\text{-}\mu_o = 1.2$ imply $\delta = 2.4$, and this formula gives π =

However, the approximation to the non-central T is needed to calculate the power of a test of the one-sided hypothesis H_o: $\mu \le \mu_o$. This approximation is $\pi \simeq \Phi((-w_c+\delta)[1+w_c^2/2(N\text{-}1)]^{-1/2})$, where w_c is the critical level given by w_c = calc invt(1-α,N-1), and H_o is rejected if $T > w_c$.

Problem 3. Testing the Variance of a Normal Population with Unknown Mean

Suppose H_o: $\sigma^2 = \sigma_o^2$ versus the alternative H_1: $\sigma^2 \ne \sigma_o^2$. Under the null, the statistic $X \equiv (N\text{-}1)s^2/\sigma_o^2$ is distributed χ^2_{n-1}. Then, a test with significance level α can be made by rejecting H_o if $X < z_{c1}$ or $X > z_{c2}$, where z_{c1} and z_{c2} are chosen so the probability is $\alpha/2$ that a draw from χ^2_{N-1} is less than z_{c1}, and $\alpha/2$ that it is greater than z_{c2}. These can be calculated using z_{c1} = calc invchi(α/2,N-1) and z_{c2} = calc invchi(1-α/2,N-1).

To calculate the power of the test against the alternative H_1: $\sigma^2 = \sigma_1^2$, note that in this case $(N\text{-}1)s^2/\sigma_1^2 = X\sigma_o^2/\sigma_1^2 \equiv Y$ is χ^2_{N-1}. Then,

$\pi = 1 - \Pr(z_{c1} \le X \le z_{c2}|\sigma^2=\sigma_1^2) = 1 - \Pr(z_{c1}\sigma_o^2/\sigma_1^2 \le Y \le z_{c2}\sigma_o^2/\sigma_1^2)$

= calc cumchi($z_{c1}\sigma_o^2/\sigma_1^2$,N-1) + 1 - cumchi($z_{c2}\sigma_o^2/\sigma_1^2$,N-1).

Problem 4. Testing the Equality of the Unknown Variances in Two Populations

Suppose independent random samples of sizes n_i are drawn from normal populations with means μ_i and variances σ_i^2, respectively, for i = 1,2. The null hypothesis is H_o: $\sigma_1^2 = \sigma_2^2$, and the alternative is $\sigma_1^2 \ne \sigma_2^2$. For each population, we know from 3.6 that $(n_i\text{-}1)s_i^2/\sigma_i^2$ has a Chi-square distribution with n_i-1 degrees of freedom.

Further, we know that the ratio of two independent Chi-square distributed random variables, each divided by its degrees of freedom, has an F-distribution with these respective degrees of freedom. Then, $\frac{s_1^2}{s_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}$ ~ $F(n_1\text{-}1,n_2\text{-}1)$. Under H_o, the statistic s_1^2/s_2^2 has this distribution, and one can form a critical region that has significance level α by choosing the lower and upper tails of the F-distribution so that each has probability $\alpha/2$. Under alternatives to the null, the ratio s_1^2/s_2^2, scaled by the ratio σ_2^2/σ_1^2, has a central F-distribution, and the power of the test is

$$\pi = F(c_L\cdot\sigma_1^2/\sigma_2^2,n_1\text{-}1,n_2\text{-}1) + 1 - F(c_U\cdot\sigma_1^2/\sigma_2^2,n_1\text{-}1,n_2\text{-}1),$$

where c_L and c_U are the lower and upper critical levels.

6. Hypothesis Testing in Large Samples

Consider data $\mathbf{x} = (x_1,...,x_n)$ obtained by simple random sampling from a population with density $f(x,\theta_o)$, where θ_o is an unknown finite-dimensional parameter contained in the interior of a set Θ. The sample DGP is $f(\mathbf{x},\theta_o) = \prod_{i=1}^{n} f(x_i,\theta_o)$, the log likelihood of an observation is $l(x,\theta) = \log f(x,\theta)$, and the log likelihood of the sample, normalized by 1/n, is $L_n(\mathbf{x},\theta) = \frac{1}{n}\sum_{i=1}^{n} l(x_n,\theta)$. Consider the maximum likelihood estimator $T_n(\mathbf{x})$, given by the value of θ that maximizes $L_n(\mathbf{x},\theta)$. Under general regularity conditions like those given in Chapter 6.4, the maximum likelihood estimator is consistent and asymptotically normal. This implies specifically that $\sqrt{n}(T_n(\mathbf{x})\text{-}\theta_o) \xrightarrow{d} Z_o$ with Z_o ~ $N(O,J^{-1})$ and J the Fisher information in an observation, $J = \mathbf{E}_{x|\theta_o}\nabla_\theta l(x,\theta_o)^2$. The Chapter 3.1.18 rule for limits of continuous transformations implies $\sqrt{n}\cdot J^{1/2}(T_n(\mathbf{x})\text{-}\theta_o) \xrightarrow{d} N(0,I)$, and hence that the quadratic form $W(\mathbf{x},\theta_o) \equiv n\cdot(T_n(\mathbf{x})-\theta_o)'J(T_n(\mathbf{x})-\theta_o) \equiv (T_n(\mathbf{x})-\theta_o)'\mathbf{V}(T_n(\mathbf{x}))^{-1}(T_n(\mathbf{x})-\theta_o) \xrightarrow{d} \chi^2_k$, the Chi-square distribution with k degrees of freedom, where k is the dimension of θ_o. When k = 1, this quadratic form equals the square of the difference between $T_n(\mathbf{x})$ and θ_o, divided by the variance $\mathbf{V}(T_n(\mathbf{x}))$ of $T_n(\mathbf{x})$. The square root of this expression, $(T_n(\mathbf{x})-\theta_o)/\sqrt{\mathbf{V}(T_n(\mathbf{x}))}$, converges in distribution to a standard normal.

Consider the null hypothesis H_o: $\theta = \theta_o$. When this null hypothesis is true, the quadratic form $W(\mathbf{x},\theta_o)$ has a limiting Chi-square distribution with k degrees of freedom. Then, a test of the hypothesis with a significance level α can be carried out by choosing a critical level *c* from the upper tail of the χ^2_k distribution so that the tail has probability α, and rejecting H_o when $W(\mathbf{x},\theta_o) > c$. We term $W(\mathbf{x},\theta_o)$ the *Wald statistic*.

Suppose an alternative H_1: $\theta = \theta_1$ to the null hypothesis is true. The power of the Wald test is the probability that the null hypothesis will be rejected when H_1 holds. But in this case, $\sqrt{n}\cdot J^{1/2}(T_n(\mathbf{x})-\theta_o) = \sqrt{n}\cdot J^{1/2}(T_n(\mathbf{x})-\theta_1) + \sqrt{n}\cdot J^{1/2}(\theta_1-\theta_o)$, with the first term converging in distribution to N(0,I). For fixed $\theta_1 \neq \theta_o$, the second term blows up. This implies that the probability that $\sqrt{n}\cdot J^{1/2}(T_n(\mathbf{x})-\theta_o)$ is small enough to accept the null hypothesis goes to zero, and the power of the test goes to one. A test with this property is called *consistent*, and consistency is usually taken to be a minimum requirement for a hypothesis testing procedure to be statistically satisfactory. A closer look at the power of a test in large samples is usually done by considering what is called *local power*. Suppose one takes a sequence of alternatives to the null hypothesis that get closer and closer to the null as sample size grows. Specifically, consider H_1: $\theta = \theta_o + \delta/\sqrt{n}$. For this sequence of alternatives, the term $\sqrt{n}\cdot J^{1/2}(\theta_1-\theta_o) = J^{1/2}\delta$ is a constant, and we have the result that $\sqrt{n}\cdot J^{1/2}(T_n(\mathbf{x})-\theta_o) \xrightarrow{d} N(J^{1/2}\delta,I)$. This implies that $n\cdot(T_n(\mathbf{x})-\theta_o)'J\cdot(T_n(\mathbf{x})-\theta_o)$, the Wald statistic, converges in distribution to a noncentral Chi-square distribution with k degrees of freedom and a noncentrality parameter $\delta'J\delta$. The local power of the test is the probability in the upper tail of this distribution above the critical level *c* for the Wald statistic. The local power will be a number between zero and one which provides useful information on the ability of the test to distinguish the null from nearby alternatives. In finite sample applications, the local power approximation can be used for a specific alternative θ_1 of interest by taking $\delta = \sqrt{n}\cdot(\theta_1-\theta_o)$ and using the noncentral Chi-square distribution as described above.

In practice, we do not know the Fisher Information J exactly, but must estimate it from the sample by

$$(10) \qquad \hat{J}_n = \mathbf{E}_n[\nabla_\theta l(x,T_n)][\nabla_\theta l(x_i,T_n)]' \equiv \frac{1}{n}\sum_{i=1}^{n}[\nabla_\theta l(x_i,T_n)][\nabla_\theta l(x_i,T_n)]'.$$

The expression in (10) is termed the *outer product of the score* $\nabla_\theta l(x_i,T_n)$ of an observation. When there is a single parameter, this reduces to the square of $\nabla_\theta l(x_i,T_n)$; otherwise, it is a k×k array of squares and cross-products of the components of $\nabla_\theta l(x_i,T_n)$. From the theorem in Chapter 6.4, $\hat{J}_n \xrightarrow{p} J$, and the rule 1.17 in Chapter 4 implies that replacing J by $\hat{J}_n$ in the Wald test statistic does not change its asymptotic distribution.

In the discussion of maximum likelihood estimation in Chapter 6.4 and the proof of its asymptotic normality, we established that when θ_o is the true parameter,

$$(11) \qquad \sqrt{n}\cdot(T_n(\mathbf{x})-\theta_o) = J^{-1}\cdot\nabla_\theta L(\mathbf{x},\theta_o) + o_p(1);$$

that is, the difference of the maximum likelihood estimator from the true parameter, normalized by $\sqrt{n}$, equals the score of the likelihood at θ_o, transformed by J^{-1}, plus asymptotically negligible terms. If we substitute (11) into the Wald statistic, we obtain $LM(\mathbf{x},\theta_o) = W(\mathbf{x},\theta_o) + o_p(1)$, where

(12) $\qquad LM(\mathbf{x},\theta_o) = [\nabla_\theta L(\mathbf{x},\theta_o)]'J^{-1}[\nabla_\theta L(\mathbf{x},\theta_o)].$

The statistic (12) is called the *Lagrange Multiplier* (LM) *statistic*, or the *score statistic*. The name Lagrange Multiplier comes from the fact that if we maximize $L_n(\mathbf{x},\theta)$ subject to the constraint $\theta_o - \theta = 0$ by setting up the Lagrangian $L_n(\mathbf{x},\theta) + \lambda(\theta_o-\theta)$, we obtain the first order condition $\lambda = \nabla_\theta L_n(\mathbf{x},\theta)$ and hence $LM(\mathbf{x},\theta_o) = \lambda^2 nJ^{-1}$. Because $LM(\mathbf{x},\theta_o)$ is *asymptotically equivalent* to the Wald statistic, it will have the same asymptotic distribution, so that the same rules apply for determining critical levels and calculating power. The Wald and LM statistics will have different numerical values in finite samples, and sometimes one will accept a null hypothesis when the other rejects. However, when sample sizes are large, their asymptotic equivalence implies that most of the time they will either both accept or both reject, and that they have the same power. In applications, J in (11) must be replaced by an estimate, either $\hat{J}_n$ from (10), or $\hat{J}_{on} = \mathbf{E}_n[\nabla_\theta l(x,\theta_o)][\nabla_\theta l(x_i,\theta_o)]'$ in which the score is evaluated at the hypothesized θ_o. Both converge in probability to J, and substitution of either in (12) leaves the asymptotic distribution of the LM statistic unchanged.

A major advantage of the LM form of the asymptotic test statistic is that it does not require that one compute the estimate $T_n(\mathbf{x})$. Computation of maximum likelihood estimates can sometimes be difficult. In these cases, the LM statistic avoids the difficulty.

The *generalized likelihood ratio criterion* was suggested in a number of simple tests of hypotheses as a good general procedure for obtaining test statistics. This method rejects H_O if

(13) $\qquad \kappa < \max_\theta L_n(\mathbf{x},\theta) - L_n(\mathbf{x},\theta_o),$

where κ is a constant that is adjusted to give the desired significance level for the test. A Taylor's expansion of $L_n(\mathbf{x},\theta_o)$ about $T_n(\mathbf{x})$ yields

(14) $$L_n(\mathbf{x},T_n(\mathbf{x})) - L_n(\mathbf{x},\theta_o) = \nabla_\theta L_n(\mathbf{x},T_n(\mathbf{x}))\cdot(T_n(\mathbf{x})-\theta_o) - \frac{1}{2}(T_n(\mathbf{x})-\theta_o)'\nabla_{\theta\theta}L_n(\mathbf{x},\tilde{\theta}_n)(T_n(\mathbf{x})-\theta_o),$$

where $\tilde{\theta}_n$ is between θ_O and θ_n. But $\nabla_\theta L_n(\mathbf{x},T_n(\mathbf{x})) = O$. Under the regularity conditions in Chapter 6.4, $\nabla_{\theta\theta}L_n(\mathbf{x},\tilde{\theta}_n) \xrightarrow{p} J$. (To make the last statement rigorous, one needs to either establish that the convergence in probability of $\nabla_{\theta\theta}L_n(\mathbf{x},\theta)$ to $J(\theta)$ is uniform in θ, or expand $\nabla_{\theta\theta}L_n(\mathbf{x},\tilde{\theta}_n)$ to first order about θ_o and argue that the first term goes in probability to -J and the second term goes in probability to zero.) Then, $LR(\mathbf{x},\theta_o) = 2\cdot n\cdot[L_n(\mathbf{x},T_n(\mathbf{x})) - L_n(\mathbf{x},\theta_o)]$, termed the *likelihood ratio statistic*, satisfies

(15) $$LR(\mathbf{x},\theta_o) = n\cdot(T_n(\mathbf{x})-\theta_o)'J(T_n(\mathbf{x})-\theta_o) + o_p(1) \equiv W(\mathbf{x},\theta_o) + o_p(1),$$

and the LR statistic is asymptotically equivalent to the Wald statistic. Therefore, the LR statistic will be asymptotically distributed Chi-square with k degrees of freedom, where k is the dimension of θ_o, and its local power is the same as that of the Wald statistic, and calculated in the same way.

The major advantage of the LR statistic is that its computation requires only the values of the log likelihood unrestricted and with the null imposed; it is unnecessary to obtain an estimate of J or perform any matrix calculations. We conclude that the trinity consisting of the Wald, LM, and LR statistics are all asymptotically equivalent, and provide completely substitutable ways of testing a hypothesis using a large sample approximation.