Lecture 11. More Confidence Intervals

11.1 Differences of Means

Let $X_1, \dots, X_n$ be iid, each normal (μ_1, σ^2), and let $Y_1, \dots, Y_m$ be iid, each normal (μ_2, σ^2). Assume that $(X_1, \dots X_n)$ and $(Y_1, \dots, Y_m)$ are independent. We will construct a confidence interval for $\mu_1 - \mu_2$. In practice, the interval is often used in the following way. If the interval lies entirely to the left of 0, we have reason to believe that $\mu_1 < \mu_2$.

Since $\operatorname{Var}(\overline{X} - \overline{Y}) = \operatorname{Var}\overline{X} + \operatorname{Var}\overline{Y} = (\sigma^2/n) + (\sigma^2/m)$,

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \quad \text{is normal (0,1).}$$

Also, nS_1^2/σ^2 is $\chi^2(n-1)$ and mS_2^2/σ^2 is $\chi^2(m-1)$. But $\chi^2(r)$ is the sum of squares of r independent, normal (0,1) random variables, so

$$\frac{nS_1^2}{\sigma^2} + \frac{mS_2^2}{\sigma^2} \quad \text{is} \quad \chi^2(n+m-2).$$

Thus if

$$R = \sqrt{\left(\frac{nS_1^2 + mS_2^2}{n+m-2}\right)\left(\frac{1}{n} + \frac{1}{m}\right)}$$

then

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{R} \quad \text{is} \quad T(n+m-2).$$

Our assumption that both populations have the same variance is crucial, because the *unknown* variance can be cancelled.

If $P\{-b < T < b\} = .95$ we get a 95 percent confidence interval for $\mu_1 - \mu_2$:

$$-b < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{R} < b$$

or

$$(\overline{X} - \overline{Y}) - bR < \mu_1 - \mu_2 < (\overline{X} - \overline{Y}) + bR.$$

If the variances σ_1^2 and σ_2^2 are *known* but possibly *unequal*, then

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

is normal (0,1). If R_0 is the denominator of the above fraction, we can get a 95 percent confidence interval as before: $\Phi(b) - \Phi(-b) = 2\Phi(b) - 1 > .95$,

$$(\overline{X} - \overline{Y}) - bR_0 < \mu_1 - \mu_2 < (\overline{X} - \overline{Y}) + bR_0.$$

11.2 Example

Let Y_1 and Y_2 be binomial (n_1, p_1) and (n_2, p_2) respectively. Then

$$Y_1 = X_1 + \cdots + X_{n_1} \quad \text{and} \quad Y_2 = Z_1 + \cdots + Z_{n_2}$$

where the X_i and Z_j are indicators of success on trials i and j respectively. Assume that $X_1, \ldots X_{n_1}, Z_1, \ldots, Z_{n_2}$ are independent. Now $E(Y_1/n_1) = p_1$ and $\operatorname{Var}(Y_1/n_1) = n_1 p_1(1-p_1)/n_1^2 = p_1(1-p_1)/n_1$, with similar formulas for Y_2/n_2. Thus for large n,

$$\left(\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right) - (p_1 - p_2)$$

divided by

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

is approximately normal (0,1). But this expression cannot be used to construct confidence intervals for $p_1 - p_2$ because the denominator involves the *unknown* quantities p_1 and p_2. However, Y_1/n_1 converges in probability to p_1 and Y_2/n_2 converges in probability to p_2, and this justifies replacing p_1 by Y_1/n_1 and p_2 by Y_2/n_2 in the denominator.

11.3 The Variance

We will construct confidence intervals for the variance of a normal population. Let $X_1, \ldots, X_n$ be iid, each normal (μ, σ^2), so that nS^2/σ^2 is $\chi^2(n-1)$. If h_{n-1} is the $\chi^2(n-1)$ density and a and b are chosen so that $\int_a^b h_{n-1}(x)\,dx = 1 - \alpha$, then

$$P\{a < \frac{nS^2}{\sigma^2} < b\} = 1 - \alpha.$$

But $a < (nS^2)/\sigma^2 < b$ is equivalent to

$$\frac{nS^2}{b} < \sigma^2 < \frac{nS^2}{a}$$

so we have a confidence interval for σ^2 at confidence level $1 - \alpha$. In practice, a and b are chosen so that $\int_b^\infty h_{n-1}(x)\,dx = \int_{-\infty}^a h_{n-1}(x)\,dx$. For example, if H_{n-1} is the $\chi^2(n-1)$ distribution function and the confidence level is 95 percent, we take $H_{n-1}(a) = .025$ and $H_{n-1}(b) = 1 - .025 = .975$. This is optimal (the length of the confidence interval is minimized) when the density is symmetric about zero, and in the symmetric case we would have $a = -b$. In the nonsymmetric case (as we have here), the error is usually small.

In this example, μ is unknown. If the mean is known, we can make use of this knowledge to improve performance. Note that

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \quad \text{is} \quad \chi^2(n)$$

so if

$$W = \sum_{i=1}^{n} (X_i - \mu)^2$$

and we choose a and b so that $\int_a^b h_n(x)\,dx = 1 - \alpha$, then $P\{a < (W/\sigma^2) < b\} = 1 - \alpha$. The inequality defining the confidence interval can be written as

$$\frac{W}{b} < \sigma^2 < \frac{W}{a}.$$

11.4 Ratios of Variances

Here we see an application of the F distribution. Let $X_1, \dots, X_{n_1}$ be iid, each normal (μ_1, σ_1^2), and let $Y_1, \dots, Y_{n_2}$ be iid, each normal (μ_2, σ_2^2). Assume that $(X_1, \dots, X_{n_1})$ and $(Y_1, \dots, Y_{n_2})$ are independent. Then $n_i S_i^2/\sigma_i^2$ is $\chi^2(n_i - 1), i = 1, 2$. Thus

$$\frac{(n_2 S_2^2/\sigma_2^2)/(n_2 - 1)}{(n_1 S_1^2/\sigma_1^2)/(n_1 - 1)} \quad \text{is} \quad F(n_2 - 1, n_1 - 1).$$

Let V^2 be the unbiased version of the sample variance, i.e.,

$$V^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Then

$$\frac{V_2^2}{V_1^2} \frac{\sigma_1^2}{\sigma_2^2} \quad \text{is} \quad F(n_2 - 1, n_1 - 1)$$

and this allows construction of confidence intervals for σ_1^2/σ_2^2 in the usual way.

Problems

1. In (11.1), suppose the variances σ_1^2 and σ_2^2 are unknown and possibly unequal. Explain why the analysis of (11.1) breaks down.
2. In (11.1), again assume that the variances are unknown, but $\sigma_1^2 = c\sigma_2^2$ where c is a known positive constant. Show that confidence intervals for the difference of means can be constructed.

Lecture 12. Hypothesis Testing

12.1 Basic Terminology

In our general statistical model (Lecture 9), suppose that the set of possible values of θ is partitioned into two subsets A_0 and A_1, and the problem is to decide between the two possibilities $H_0 : \theta \in A_0$, the *null hypothesis*, and $H_1 : \theta \in A_1$, the *alternative*. Mathematically, it doesn't make any difference which possibility you call the null hypothesis, but in practice, H_0 is the "default setting". For example, $H_0 : \mu \le \mu_0$ might mean that a drug is no more effective than existing treatments, while $H_1 : \mu > \mu_0$ might mean that the drug is a significant improvement.

We observe x and make a decision via $\delta(x) = 0$ or 1. There are two types of errors. A *type 1 error* occurs if H_0 is true but $\delta(x) = 1$, in other words, we declare that H_1 is true. Thus in a type 1 error, we *reject* H_0 *when it is true.*

A *type 2 error* occurs if H_0 is false but $\delta(x) = 0$, i.e., we declare that H_0 is true. Thus in a type 2 error, we *accept* H_0 *when it is false.*

If H_0 [resp. H_1] means that a patient does not have [resp. does have] a particular disease, then a type 1 error is also called a *false positive*, and a type 2 error is also called a *false negative*.

If $\delta(x)$ is always 0, then a type 1 error can never occur, but a type 2 error will always occur. Symmetrically, if $\delta(x)$ is always 1, then there will always be a type 1 error, but never an error of type 2. Thus by ignoring the data altogether we can reduce one of the error probabilities to zero. To get *both* error probabilities to be small, in practice we must increase the sample size.

We say that H_0 [resp. H_1] is *simple* if A_0 [resp. A_1] contains only one element, *composite* if A_0 [resp. A_1] contains more than one element. So in the case of *simple hypothesis vs. simple alternative*, we are testing $\theta = \theta_0$ vs. $\theta = \theta_1$. The standard example is to test the hypothesis that X has density f_0 vs. the alternative that X has density f_1.

12.2 Likelihood Ratio Tests

In the case of simple hypothesis vs. simple alternative, if we require that the probability of a type 1 error be at most α and try to minimize the probability of a type 2 error, the optimal test turns out to be a *likelihood ratio test (LRT)*, defined as follows. Let $L(x)$, the *likelihood ratio*, be $f_1(x)/f_0(x)$, and let λ be a constant. If $L(x) > \lambda$, reject H_0; if $L(x) < \lambda$, accept H_0; if $L(x) = \lambda$, do anything.

Intuitively, if what we have observed seems significantly more likely under H_1, we will tend to reject H_0. If H_0 or H_1 is composite, there is no general optimality result as there is in the simple vs. simple case. In this situation, we resort to *basic statistical philosophy*: If, assuming that H_0 is true, we witness a rare event, we tend to reject H_0.

The statement that LRT's are optimal is the Neyman-Pearson lemma, to be proved at the end of the lecture. In many common examples (normal, Poisson, binomial, exponential), $L(x_1, \dots, x_n)$ can be expressed as a function of the sum of the observations, or equivalently as a function of the sample mean. This motivates consideration of tests based on $\sum_{i=1}^n X_i$ or on $\overline{X}$.

12.3 Example

Let $X_1, \dots, X_n$ be iid, each normal (θ, σ^2). We will test $H_0 : \theta \le \theta_0$ vs. $H_1 : \theta > \theta_0$. Under H_1, $\overline{X}$ will tend to be larger, so let's reject H_0 when $\overline{X} > c$. The *power function* of the test is defined by

$$K(\theta) = P_\theta\{\text{reject } H_0\},$$

the probability of rejecting the null hypothesis when the true parameter is θ. In this case,

$$P\{\overline{X} > c\} = P\left\{\frac{\overline{X} - \theta}{\sigma/\sqrt{n}} > \frac{c - \theta}{\sigma/\sqrt{n}}\right\} = 1 - \Phi\left(\frac{c - \theta}{\sigma/\sqrt{n}}\right)$$

(see Figure 12.1). Suppose that we specify the probability α of a type 1 error when $\theta = \theta_1$, and the probability β of a type 2 error when $\theta = \theta_2$. Then

$$K(\theta_1) = 1 - \Phi\left(\frac{c - \theta_1}{\sigma/\sqrt{n}}\right) = \alpha$$

and

$$K(\theta_2) = 1 - \Phi\left(\frac{c - \theta_2}{\sigma/\sqrt{n}}\right) = 1 - \beta.$$

If $\alpha, \beta, \sigma, \theta_1$ and θ_2 are known, we have two equations that can be solved for c and n.

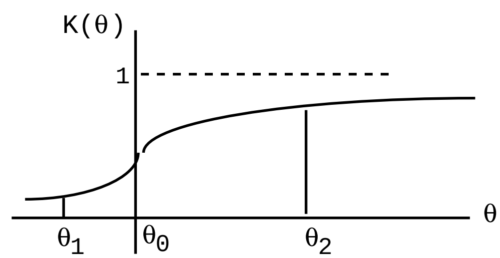

Figure 12.1

The *critical region* is the set of observations that lead to rejection. In this case, it is $\{(x_1, \dots, x_n) : n^{-1} \sum_{i=1}^n x_i > c\}$.

The *significance level* is the largest type 1 error probability. Here it is $K(\theta_0)$, since $K(\theta)$ increases with θ.

12.4 Example

Let H_0 : X is uniformly distributed on (0,1), so $f_0(x) = 1, 0 < x < 1$, and 0 elsewhere. Let $H_1 : f_1(x) = 3x^2, 0 < x < 1$, and 0 elsewhere. We take only one observation, and reject H_0 if $x > c$, where $0 < c < 1$. Then

$$K(0) = P_0\{X > c\} = 1 - c, \quad K(1) = P_1\{X > c\} = \int_c^1 3x^2\, dx = 1 - c^3.$$

If we specify the probability α of a type 1 error, then $\alpha = 1 - c$, which determines c. If β is the probability of a type 2 error, then $1 - \beta = 1 - c^3$, so $\beta = c^3$. Thus (see Figure 12.2)

$$\beta = (1 - \alpha)^3.$$

If $\alpha = .05$ then $\beta = (.95)^3 \approx .86$, which indicates that you usually can't do too well with only one observation.

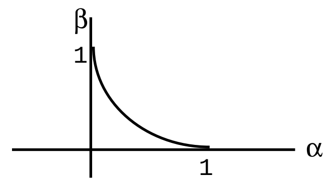

Figure 12.2

12.5 Tests Derived From Confidence Intervals

Let $X_1, \dots, X_n$ be iid, each normal (μ_0, σ^2). In Lecture 10, we found a confidence interval for μ_0, assuming σ^2 unknown, via

$$P\left\{-b < \frac{\overline{X} - \mu_0}{S/\sqrt{n-1}} < b\right\} = 2F_T(b) - 1 \quad \text{where} \quad T = \frac{\overline{X} - \mu_0}{S/\sqrt{n-1}}$$

has the T distribution with $n-1$ degrees of freedom.

Say $2F_T(b) - 1 = .95$, so that

$$P\left\{\left|\frac{\overline{X} - \mu_0}{S/\sqrt{n-1}}\right| \geq b\right\} = .05$$

If μ actually equals μ_0, we are witnessing an event of low probability. So it is natural to test $\mu = \mu_0$ vs. $\mu \neq \mu_0$ by rejecting if

$$\left|\frac{\overline{X} - \mu_0}{S/\sqrt{n-1}}\right| \geq b,$$

in other words, μ_0 does not belong to the confidence interval. As the true mean μ moves away from μ_0 in either direction, the probability of this event will increase, since $\overline{X} - \mu_0 = (\overline{X} - \mu) + (\mu - \mu_0)$.

Tests of $\theta = \theta_0$ vs. $\theta \neq \theta_0$ are called *two-sided*, as opposed to $\theta = \theta_0$ vs. $\theta > \theta_0$ (or $\theta = \theta_0$ vs. $\theta < \theta_0$), which are *one-sided*. In the present case, if we test $\mu = \mu_0$ vs. $\mu > \mu_0$, we reject if

$$\frac{\overline{X} - \mu_0}{S/\sqrt{n-1}} \geq b.$$

The power function $K(\mu)$ is difficult to compute for $\mu \neq \mu_0$, because $(\overline{X} - \mu_0)/(\sigma/\sqrt{n})$ no longer has mean zero. The "noncentral T distribution" becomes involved.

12.6 The Neyman-Pearson Lemma

Assume that we are testing the simple hypothesis that X has density f_0 vs. the simple alternative that X has density f_1. Let φ_λ be an LRT with parameter λ (a nonnegative constant), in other words, $\varphi_\lambda(x)$ is the probability of rejecting H_0 when x is observed, and

$$\varphi_\lambda(x) = \begin{cases} 1 & \text{if } L(x) > \lambda \\ 0 & \text{if } L(x) < \lambda \\ \text{anything} & \text{if } L(x) = \lambda. \end{cases}$$

Suppose that the probability of a type 1 error using φ_λ is α_λ, and the probability of a type 2 error is β_λ. Let φ be an arbitrary test with error probabilities α and β. If $\alpha \le \alpha_\lambda$ then $\beta \ge \beta_\lambda$. In other words, the LRT has maximum power among all tests at significance level α_λ.

Proof. We are going to assume that f_0 and f_1 are one-dimensional, but the argument works equally well when $X = (X_1, \dots, X_n)$ and the f_i are n-dimensional joint densities. We recall from basic probability theory the *theorem of total probability*, which says that if X has density f, then for any event A,

$$P(A) = \int_{-\infty}^{\infty} P(A|X = x) f(x)\, dx.$$

A companion theorem which we will also use later is the *theorem of total expectation*, which says that if X has density f, then for any random variable Y,

$$E(Y) = \int_{-\infty}^{\infty} E(Y|X = x) f(x)\, dx.$$

By the theorem of total probability,

$$\alpha = \int_{-\infty}^{\infty} \varphi(x) f_0(x)\, dx, \quad 1 - \beta = \int_{-\infty}^{\infty} \varphi(x) f_1(x)\, dx$$

and similarly

$$\alpha_\lambda = \int_{-\infty}^{\infty} \varphi_\lambda(x) f_0(x)\, dx, \quad 1 - \beta_\lambda = \int_{-\infty}^{\infty} \varphi_\lambda(x) f_1(x)\, dx.$$

We claim that for all x,

$$[\varphi_\lambda(x) - \varphi(x)][f_1(x) - \lambda f_0(x)] \ge 0.$$

For if $f_1(x) > \lambda f_0(x)$ then $L(x) > \lambda$, so $\varphi_\lambda(x) = 1 \ge \varphi(x)$, and if $f_1(x) < \lambda f_0(x)$ then $L(x) < \lambda$, so $\varphi_\lambda(x) = 0 \le \varphi(x)$, proving the assertion. Now if a function is always nonnegative, its integral must be nonnegative, so

$$\int_{-\infty}^{\infty} [\varphi_\lambda(x) - \varphi(x)][f_1(x) - \lambda f_0(x)]\, dx \ge 0.$$

The terms involving f_0 translate to statements about type 1 errors, and the terms involving f_1 translate to statements about type 2 errors. Thus

$$(1-\beta_\lambda) - (1-\beta) - \lambda\alpha_\lambda + \lambda\alpha \geq 0,$$

which says that $\beta - \beta_\lambda \geq \lambda(\alpha_\lambda - \alpha) \geq 0$, completing the proof. ♣

12.7 Randomization

If $L(x) = \lambda$, then "do anything" means that randomization is possible, e.g., we can flip a possibly biased coin to decide whether or not to accept H_0. (This may be significant in the discrete case, where $L(x) = \lambda$ may have positive probability.) Statisticians tend to frown on this practice because two statisticians can look at exactly the same data and come to different conclusions. It is possible to adjust the significance level (by replacing "do anything" by a definite choice of either H_0 or H_1) to avoid randomization.

Problems

1. Consider the problem of testing $\theta = \theta_0$ vs. $\theta > \theta_0$, where θ is the mean of a normal population with known variance. Assume that the sample size n is fixed. Show that the test given in Example 12.3 (reject H_0 if $\overline{X} > c$) is *uniformly most powerful*. In other words, if we test $\theta = \theta_0$ vs. $\theta = \theta_1$ for any given $\theta_1 > \theta_0$, and we specify the probability α of a type 1 error, then the probability β of a type 2 error is minimized.
2. It is desired to test the null hypothesis that a die is unbiased vs. the alternative that the die is loaded, with faces 1 and 2 having probability 1/4 and faces 3,4,5 and 6 having probability 1/8. The die is to be tossed once. Find a most powerful test at level $\alpha = .1$, and find the type 2 error probability β.
3. We wish to test a binomial random variable X with $n = 400$ and $H_0 : p = 1/2$ vs. $H_1 : p > 1/2$. The random variable $Y = (X - np)/\sqrt{np(1-p)} = (X - 200)/10$ is approximately normal (0,1), and we will reject H_0 if $Y > c$. If we specify $\alpha = .05$, then $c = 1.645$. Thus the critical region is $X > 216.45$. Suppose the actual result is $X = 220$, so that H_0 is rejected. Find the minimum value of α (sometimes called the *p-value*) for which the *given* data lead to the *opposite* conclusion (acceptance of H_0).

Lecture 13. Chi-Square Tests

13.1 Introduction

Let $X_1, \dots, X_k$ be multinomial, i.e., X_i is the number of occurrences of the event A_i in n generalized Bernoulli trials (Lecture 6). Then

$$P\{X_1 = n_1, \dots, X_k = n_k\} = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$$

where the n_i are nonnegative integers whose sum is n. Consider $k = 2$. Then X_1 is binomial (n, p_1) and $(X_1 - np_1)/\sqrt{np_1(1-p_1)} \approx$ normal(0,1). Consequently, the random variable $(X_1 - np_1)^2/np_1(1-p_1)$ is approximately $\chi^2(1)$. But

$$\frac{(X_1 - np_1)^2}{np_1(1-p_1)} = \frac{(X_1 - np_1)^2}{n}\left[\frac{1}{p_1} + \frac{1}{1-p_1}\right] = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}.$$

(Note that since $k = 2$ we have $p_2 = 1 - p_1$ and $X_1 - np_1 = n - X_2 - np_1 = np_2 - X_2 = -(X_2 - np_2)$, and the outer minus sign disappears when squaring.) Therefore $[(X_1 - np_1)^2/np_1] + [(X_2 - np_2)^2/np_2] \approx \chi^2(1)$. More generally, it can be shown that

$$Q = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \approx \chi^2(k-1).$$

where

$$\frac{(X_i - np_i)^2}{np_i} = \frac{(\text{observed frequency-expected frequency})^2}{\text{expected frequency}}.$$

We will consider three types of chi-square tests.

13.2 Goodness of Fit

We ask whether a random variable X has a specified distribution (normal, Poisson, etc.). The null hypothesis is that the multinomial probabilities are $\underline{p} = (p_1, \dots, p_k)$, and the alternative is that $\underline{p} \neq (p_1, \dots, p_k)$.

Suppose that $P\{\chi^2(k-1) > c\}$ is at the desired level of significance (for example, .05). If $Q > c$ we will reject H_0. The idea is that if H_0 is in fact true, we have witnessed a rare event, so rejection is reasonable. If H_0 is false, it is reasonable to expect that some of the X_i will be far from np_i, so Q will be large.

Some practical considerations: Take n large enough so that each $np_i \geq 5$. Each time a parameter is estimated from the sample, reduce the number of degrees of freedom by 1. (A typical case: The null hypothesis is that X is Poisson (λ), but the mean λ is unknown, and is estimated by the sample mean.)

13.3 Equality of Distributions

We ask whether two or more samples come from the same underlying distribution. The observed results are displayed in a *contingency table*. This is an $h \times k$ matrix whose rows are the samples and whose columns are the attributes to be observed. For example, row i might be $(7, 11, 15, 13, 4)$, with the interpretation that in a class of 50 students taught by method of instruction i, there were 7 grades of A, 11 of B, 15 of C, 13 of D and 4 of F. The null hypothesis H_0 is that there is no difference between the various methods of instruction, i.e., $P(A)$ is the same for each group, and similarly for the probabilities of the other grades. We estimate $P(A)$ from the sample by adding all entries in column A and dividing by the total number of observations in the entire experiment. We estimate $P(B), P(C), P(D)$ and $P(F)$ in a similar fashion. The expected frequencies in row i are found by multiplying the grade probabilities by the number of entries in row i.

If there are h groups (samples), each with k attributes, then each group generates a chi-square $(k-1)$, and $k-1$ probabilities are estimated from the sample (the last probability is determined). The number of degrees of freedom is $h(k-1)-(k-1)=(h-1)(k-1)$, call it r. If $P\{\chi^2(r) > c\}$ is the desired significance level, we reject H_0 if the chi-square statistic is greater than c.

13.4 Testing For Independence

Again we have a contingency table with h rows corresponding to the possible values x_i of a random variable X, and k columns corresponding to the possible values y_j of a random variable Y. We are testing the null hypothesis that X and Y are independent.

Let R_i be the sum of the entries in row i, and let C_j be the sum of the entries in column j. Then the sum of all observations is $T = \sum_i R_i = \sum_j C_j$. We estimate $P\{X = x_i\}$ by R_i/T, and $P\{Y = y_j\}$ by C_j/T. Under the independence hypothesis H_0, $P\{X = x_i, Y = y_j\} = P\{X = x_i\}P\{Y = y_j\} = R_iC_j/T^2$. Thus the expected frequency of (x_i, y_j) is R_iC_j/T. (This gives another way to calculate the expected frequencies in (13.3). In that case, we estimated the j-th column probability by C_j/T, and multiplied by the sum of the entries in row i, namely R_i.)

In an $h \times k$ contingency table, the number of degrees of freedom is $hk-1$ minus the number of estimated parameters:

$$hk - 1 - (h - 1 + k - 1) = hk - h - k + 1 = (h-1)(k-1).$$

The chi-square statistic is calculated as in (13.3). Similarly, if there are 3 attributes to be tested for independence and we form an $h \times k \times m$ contingency table, the number of degrees of freedom is

$$hkm - 1 - (h-1) + (k-1) + (m-1) = hkm - h - k - m + 2.$$

Problems

1. Use a chi-square procedure to test the null hypothesis that a random variable X has the following distribution:

$$P\{X = 1\} = .5, \quad P\{X = 2\} = .3, \quad P\{X = 3\} = .2$$

We take 100 independent observations of X, and it is observed that 1 occurs 40 times, 2 occurs 33 times, and 3 occurs 27 times. Determine whether or not we will reject the null hypothesis at significance level .05 .

2. Use a chi-square test to decide (at significance level .05) whether the two samples corresponding to the rows of the contingency table below came from the same underlying distribution.

	A	B	C
Sample 1	33	147	114
Sample 2	67	153	86

3. Suppose we are testing for independence in a 2×2 contingency table

$$\begin{array}{cc} a & b \\ c & d \end{array}$$

Show that the chi-square statistic is

$$\frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)}.$$

(The number of degrees of freedom is $1 \times 1 = 1$.)

Lecture 14. Sufficient Statistics

14.1 Definitions and Comments

Let $X_1, \dots, X_n$ be iid with $P\{X_i = 1\} = \theta$ and $P\{X_i = 0\} = 1 - \theta$, so $P\{X_i = x\} = \theta^x(1-\theta)^{1-x}, x = 0, 1$. Let Y be a *statistic* for θ, i.e., a function of the observables $X_1, \dots, X_n$. In this case we take $Y = X_1 + \cdots + X_n$, the total number of successes in n Bernoulli trials with probability θ of success on a given trial.

We claim that the conditional distribution of $X_1, \dots, X_n$ given Y is free of θ, in other words, does not depend on θ. We say that Y is *sufficient* for θ.

To prove this, note that

$$P_\theta\{X_1 = x_1, \dots, X_n = x_n | Y = y\} = \frac{P_\theta\{X_1 = x_1, \dots, X_n = x_n, Y = y\}}{P_\theta\{Y = y\}}.$$

This is 0 unless $y = x_1 + \cdots + x_n$, in which case we get

$$\frac{\theta^y(1-\theta)^{n-y}}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} = \frac{1}{\binom{n}{y}}.$$

For example, if we know that there were 3 heads in 5 tosses, the probability that the actual tosses were $HTHHT$ is $1/\binom{5}{3}$.

14.2 The Key Idea

For the purpose of making a statistical decision, we can ignore the individual random variables X_i and base the decision entirely on $X_1 + \cdots + X_n$.

Suppose that statistician A observes $X_1, \dots, X_n$ and makes a decision. Statistician B observes $X_1 + \cdots + X_n$ only, and constructs $X_1', \dots, X_n'$ according to the conditional distribution of $X_1, \dots, X_n$ given Y, i.e.,

$$P\{X_1' = x_1, \dots, X_n' = x_n | Y = y\} = \frac{1}{\binom{n}{y}}.$$

This construction is possible because the conditional distribution does not depend on the unknown parameter θ. We will show that under θ, $(X_1', \dots, X_n')$ and $(X_1, \dots, X_n)$ have exactly the same distribution, so anything A can do, B can do at least as well, even though B has less information.

Given $x_1, \dots, x_n$, let $y = x_1 + \cdots + x_n$. The only way we can have $X_1' = x_1, \dots, X_n' = x_n$ is if $Y = y$ and then B's experiment produces $X_1' = x_1, \dots, X_n' = x_n$ given y. Thus

$$P_\theta\{X_1' = x_1, \dots, X_n' = x_n\} = P_\theta\{Y = y\}P_\theta\{X_1' = x_1, \dots, X_n' = x_n | Y = y\}$$

$$= \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{1}{\binom{n}{y}} = \theta^y(1-\theta)^{n-y} = P_\theta\{X_1 = x_1, \dots, X_n = x_n\}.$$

14.3 The Factorization Theorem

Let $Y = u(X)$ be a statistic for θ; (X can be $(X_1, \dots, X_n)$, and usually is). Then Y is sufficient for θ if and only if the density $f_\theta(x)$ of X under θ can be factored as $f_\theta(x) = g(\theta, u(x))h(x)$.

[In the Bernoulli case, $f_\theta(x_1, \dots, x_n) = \theta^y(1-\theta)^{n-y}$ where $y = u(x) = \sum_{i=1}^n x_i$ and $h(x) = 1$.]

Proof. (Discrete case). If Y is sufficient, then

$$P_\theta\{X = x\} = P_\theta\{X = x, Y = u(x)\} = P_\theta\{Y = u(x)\}P\{X = x|Y = u(x)\}$$

$$= g(\theta, u(x))h(x).$$

Conversely, assume $f_\theta(x) = g(\theta, u(x))h(x)$. Then

$$P_\theta\{X = x|Y = y\} = \frac{P_\theta\{X = x, Y = y\}}{P_\theta\{Y = y\}}.$$

This is 0 unless $y = u(x)$, in which case it becomes

$$\frac{P_\theta\{X = x\}}{P_\theta\{Y = y\}} = \frac{g(\theta, u(x))h(x)}{\sum_{\{z:u(z)=y\}} g(\theta, u(z))h(z)}.$$

The g terms in both numerator and denominator are $g(\theta, y)$, which can be cancelled to obtain

$$P\{X = x|Y = y\} = \frac{h(x)}{\sum_{\{z:u(z)=y\}} h(z)}$$

which is free of θ. ♣

14.4 Example

Let $X_1, \dots, X_n$ be iid, each normal (μ, σ^2), so that

$$f_\theta(x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right].$$

Take $\theta = (\mu, \sigma^2)$ and let $\overline{x} = n^{-1}\sum_{i=1}^n x_i$, $s^2 = n^{-1}\sum_{i=1}^n (x_i - \overline{x})^2$. Then

$$x_i - \overline{x} = x_i - \mu - (\overline{x} - \mu)$$

and

$$s^2 = \frac{1}{n}\left[\sum_1^n (x_i - \mu)^2 - 2(\overline{x} - \mu)\sum_1^n (x_i - \mu) + n(\overline{x} - \mu)^2\right].$$

Thus

$$s^2 = \frac{1}{n}\sum_1^n (x_i - \mu)^2 - (\overline{x} - \mu)^2.$$

The joint density is given by

$$f_\theta(x_1, \dots, x_n) = (2\pi\sigma^2)^{-n/2} e^{-ns^2/2\sigma^2} e^{-n(\overline{x}-\mu)^2/2\sigma^2}.$$

If μ and σ^2 are both unknown then $(\overline{X}, S^2)$ is sufficient (take $h(x) = 1$). If σ^2 is known then we can take $h(x) = (2\pi\sigma^2)^{-n/2} e^{-ns^2/2\sigma^2}, \theta = \mu$, and $\overline{X}$ is sufficient. If μ is known then $(h(x) = 1)$ $\theta = \sigma^2$ and $\sum_{i=1}^n (X_i - \mu)^2$ is sufficient.

Problems

In Problems 1-6, show that the given statistic $u(X) = u(X_1, \dots, X_n)$ is sufficient for θ and find appropriate functions g and h for the factorization theorem to apply.

1. The X_i are Poisson (θ) and $u(X) = X_1 + \cdots + X_n$.
2. The X_i have density $A(\theta)B(x_i), 0 < x_i < \theta$ (and 0 elsewhere), where θ is a positive real number; $u(X) = \max X_i$. As a special case, the X_i are uniformly distributed between 0 and θ, and $A(\theta) = 1/\theta, B(x_i) = 1$ on $(0, \theta)$.
3. The X_i are geometric with parameter θ, i.e., if θ is the probability of success on a given Bernoulli trial, then $P_\theta\{X_i = x\} = (1-\theta)^x\theta$ is the probability that there will be x failures followed by the first success; $u(X) = \sum_{i=1}^n X_i$.
4. The X_i have the exponential density $(1/\theta)e^{-x/\theta}, x > 0$, and $u(X) = \sum_{i=1}^n X_i$.
5. The X_i have the beta density with parameters $a = \theta$ and $b = 2$, and $u(X) = \prod_{i=1}^n X_i$.
6. The X_i have the gamma density with parameters $\alpha = \theta$, β an arbitrary positive number, and $u(X) = \prod_{i=1}^n X_i$.
7. Show that the result in (14.2) that statistician B can do at least as well as statistician A, holds in the general case of arbitrary iid random variables X_i.

Lecture 15. Rao-Blackwell Theorem

15.1 Background From Basic Probability

To better understand the steps leading to the Rao-Blackwell theorem, consider a typical two stage experiment:

Step 1. Observe a random variable X with density $(1/2)x^2e^{-x}, x > 0$.

Step 2. If $X = x$, let Y be uniformly distributed on $(0, x)$.

Find $E(Y)$.

Method 1 via the joint density:

$$f(x,y) = f_X(x)f_Y(y|x) = \frac{1}{2}x^2e^{-x}(\frac{1}{x}) = \frac{1}{2}xe^{-x}, 0 < y < x.$$

In general, $E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,dx\,dy$. In this case, $g(x,y) = y$ and

$$E(Y) = \int_{x=0}^{\infty}\int_{y=0}^{x} y(1/2)xe^{-x}\,dy\,dx = \int_0^{\infty}(x^3/4)e^{-x}\,dx = \frac{3!}{4} = \frac{3}{2}.$$

Method 2 via the theorem of total expectation:

$$E(Y) = \int_{-\infty}^{\infty} f_X(x)E(Y|X = x)\,dx.$$

Method 2 works well when the conditional expectation is easy to compute. In this case it is $x/2$ by inspection. Thus

$$E(Y) = \int_0^{\infty}(1/2)x^2e^{-x}(x/2)\,dx = \frac{3}{2} \quad \text{as before.}$$

15.2 Comment On Notation

If, for example, it turns out that $E(Y|X = x) = x^2 + 3x + 4$, we can write $E(Y|X) = X^2 + 3X + 4$. Thus $E(Y|X)$ is a function $g(X)$ of the random variable X. When $X = x$ we have $g(x) = E(Y|X = x)$.

We now proceed to the Rao-Blackwell theorem via several preliminary lemmas.

15.3 Lemma

$E[E(X_2|X_1)] = E(X_2)$.

Proof. Let $g(X_1) = E(X_2|X_1)$. Then

$$E[g(X_1)] = \int_{-\infty}^{\infty} g(x)f_1(x)\,dx = \int_{-\infty}^{\infty} E(X_2|X_1 = x)f_1(x)\,dx = E(X_2)$$

by the theorem of total expectation. ♣

15.4 Lemma

If $\mu_i = E(X_i), i = 1, 2$, then

$$E[\{X_2 - E(X_2|X_1)\}\{E(X_2|X_1) - \mu_2\}] = 0.$$

Proof. The expectation is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x_2 - E(X_2|X_1 = x_1)][E(X_2|X_1 = x_1) - \mu_2]f_1(x_1)f_2(x_2|x_1)\,dx_1\,dx_2$$

$$= \int_{-\infty}^{\infty} f_1(x_1)[E(X_2|X_1 = x_1) - \mu_2]\int_{-\infty}^{\infty}[x_2 - E(X_2|X_1 = x_1)]f_2(x_2|x_1)\,dx_2\,dx_1.$$

The inner integral (with respect to x_2) is $E(X_2|X_1 = x_1) - E(X_2|X_1 = x_1) = 0$, and the result follows. ♣

15.5 Lemma

$\operatorname{Var} X_2 \geq \operatorname{Var}[E(X_2|X_1)]$.

Proof. We have

$$\operatorname{Var} X_2 = E[(X_2 - \mu_2)^2] = E\big([\{X_2 - E(X_2|X_1\} + \{E(X_2|X_1) - \mu_2\}]^2\big)$$

$$= E[\{X_2 - E(X_2|X_1)\}^2] + E[\{E(X_2|X_1) - \mu_2\}^2] \quad \text{by (15.4)}$$

$$\geq E[\{E(X_2|X_1) - \mu_2\}^2] \quad \text{since both terms are nonnegative.}$$

But by (15.3), $E[E(X_2|X_1)] = E(X_2) = \mu_2$, so the above term is the variance of $E(X_2|X_1)$. ♣

15.6 Lemma

Equality holds in (15.5) if and only if X_2 is a function of X_1.

Proof. The argument of (15.5) shows that equality holds iff $E[\{X_2 - E(X_2|X_1)\}^2] = 0$, in other words, $X_2 = E(X_2|X_1)$. This implies that X_2 is a function of X_1. Conversely, if $X_2 = h(X_1)$, then $E(X_2|X_1) = h(X_1) = X_2$, and therefore equality holds. ♣

15.7 Rao-Blackwell Theorem

Let $X_1, \dots, X_n$ be iid, each with density $f_\theta(x)$. Let $Y_1 = u_1(X_1, \dots, X_n)$ be a sufficient statistic for θ, and let $Y_2 = u_2(X_1, \dots, X_n)$ be an unbiased estimate of θ [or more generally, of a function of θ, say $r(\theta)$]. Then

(a) $\operatorname{Var}[E(Y_2|Y_1)] \leq \operatorname{Var} Y_2$, with strict inequality unless Y_2 is a function of Y_1 alone.

(b) $E[E(Y_2|Y_1)] = \theta$ [or more generally, $r(\theta)$].

Thus in searching for a minimum variance unbiased estimate of θ [or more generally, $r(\theta)$], we may restrict ourselves to functions of the sufficient statistic Y_1.

Proof. Part (a) follows from (15.5) and (15.6), and (b) follows from (15.3). ♣

15.8 Theorem

Let $Y_1 = u_1(X_1, \dots, X_n)$ be a sufficient statistic for θ. If the maximum likelihood estimate $\hat{\theta}$ of θ is unique, then $\hat{\theta}$ is a function of Y_1.

Proof. The joint density of the X_i can be factored as

$$f_\theta(x_1, \dots, x_n) = g(\theta, z)h(x_1, \dots, x_n)$$

where $z = u_1(x_1, \dots, x_n)$. Let θ_0 maximize $g(\theta, z)$. Given z, we find θ_0 by looking at all $g(\theta, z)$, so that θ_0 is a function of $u_1(X_1, \dots, X_n) = Y_1$. But θ_0 also maximizes $f_\theta(x_1, \dots, x_n)$, so by uniqueness, $\hat{\theta} = \theta_0$. ♣

In Lectures 15-17, we are developing methods for finding uniformly minimum variance unbiased estimates. Exercises will be deferred until Lecture 17.