9 Properties of point estimators and finding them

9.1 Introduction

We consider several properties of estimators in this chapter, in particular efficiency, consistency and sufficient statistics. An estimator $\hat{\theta}_n$ is consistent if it converges to θ in a suitable sense as $n \to \infty$. An estimator $\hat{\theta}$ for θ is sufficient, if it contains all the information that we can extract from the random sample to estimate θ. If we have a sufficient statistic, then the Rao-Blackwell theorem gives a procedure for finding the unbiased estimator with the smallest variance. Until now we have been pulling our estimators out of the air. At the end of this chapter we consider two methods for finding estimators - the method of moments and maximum likelihood estimators.

9.2 Relative Efficiency

If we want to compare two unbiased estimator for θ, then we should compare their variances. The estimator with the smaller variance is better. Why? If we want to quantify this we can look at their ratio. The *relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$* is defined to be

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{var(\hat{\theta}_2)}{var(\hat{\theta}_1)} \tag{1}$$

If this is greater than 1, the estimator $\hat{\theta}_1$ is better than the estimator $\hat{\theta}_2$.

Example: Suppose that Y is uniformly distributed on $[0, \theta]$. (This is the population distribution.) The parameter θ is unknown. Let $Y_1, Y_2, \cdots, Y_n$ be a random sample. (So the Y_i are independent and each is uniform on $[0, \theta]$.) We compare two possible unbiased estimators for θ. The mean of Y is just $\theta/2$. So let

$$\hat{\theta}_1 = 2\overline{Y} \tag{2}$$

Then $\hat{\theta}_1$ is an unbiased estimator for θ. The variance of Y is $\sigma^2 = \theta^2/12$. So the variance of

$$\sigma_{\hat{\theta}_1} = \theta^2/(3n) \tag{3}$$

Now define

$$Y^{(n)} = \max\{Y_1, Y_2, \cdots, Y_n\} \tag{4}$$

and

$$\hat{\theta}_2 = \frac{n+1}{n} Y^{(n)} \tag{5}$$

In the homework we saw that $\hat{\theta}_2$ is an unbiased estimator. We computed the variance of $Y^{(n)}$ in the homework and found

$$var(Y^{(n)}) = \frac{n}{(n+2)(n+1)^2} \theta^2 \tag{6}$$

So

$$var(\hat{\theta}_2) = \frac{1}{(n+2)n} \theta^2 \tag{7}$$

So the relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{3}{n+2} \tag{8}$$

We see that θ_2 is a much better estimator than θ_1.

9.3 Consistency

We want to consider the behavior of estimators as the sample size goes to infinity. So we will denote the estimator by $\hat{\theta}_n$ to make the dependence on the sample size explicit.

We have seen before that for our estimators for the population mean, population proportion, of population variance, the variance of the estimator goes to zero as the sample size goes to infinity. Since these are unbiased estimators, the mean of $\hat{\theta}_n$ is θ. So as $n \to \infty$, the distribution of $\hat{\theta}_n$ is becoming more and more concentrated around θ. Does this mean that $\hat{\theta}_n$ is converging to θ in some sense? There are various senses in which a sequence of random variables can converge. In this section we will define one notion of convergence and show that if the variance of an unbiased estimator does to zero as $n \to \infty$, then $\hat{\theta}_n$ converges to θ in this sense. When this happens we say the estimator is *consistent.*

Definition 1. *A sequence of random variables X_n is said to converge to the random variable X in probability if for all $\epsilon > 0$, we have*

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0 \tag{9}$$

Definition 2. *A point estimator $\hat{\theta}_n$ for θ is consistent if $\hat{\theta}_n$ converges to θ in probability.*

Remark: Note that a constant is a RV. In the last definition we are thinking of the constant θ as a random variable.

The definition is a bit abstract, so we look at what it says in the context of an example. We flip a coin n times and let $\hat{p}_n$ be the number of heads we get divided by n. This is a sample proportion which estimates p, the probability that the coin comes up heads. Suppose $p = 1/2$. If $\hat{p}_n$ converges in probability to $1/2$, then $P(0.49 \leq \hat{p}_n \leq 0.51)$ converges to 1 as the number of flips goes to infinity, $P(0.499 \leq \hat{p}_n \leq 0.501)$ converges to 1 as the number of flips goes to infinity, and so on.

Theorem 1. *If $\hat{\theta}_n$ is an unbiased estimator for θ and $\sigma_{\hat{\theta}_n} \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is consistent.*

Proof: Follows from Chebyshev's inequality □

Corollary 1. *The following estimators are consistent*

- *The sample mean $\overline{Y}$ as an estimator for the population mean μ.*
- *The difference of two sample means $\overline{Y_1} - \overline{Y_2}$ drawn independently from two different populations as an estimator for the difference of the population means $\mu_1 - \mu_2$ if both sample sizes go to infinity.*
- *The sample proportion $\hat{p}$ as an estimator for the population proportion p.*
- *The difference of two sample proportions $\hat{p_1} - \hat{p_2}$ drawn independently from two different populations as an estimator for the difference of the population proportions $p_1 - p_2$ if both sample sizes go to infinity.*
- *The sample variance S^2 as an estimator for the population variance σ^2.*

The corollary says that S_n^2 converges in probability to σ^2. Does this imply that S_n converges in probability to σ? If so, S_n would be a consistent estimator for σ. The following theorem says it does.

Theorem 2. *Suppose that the estimator $\hat{\theta}_n$ converges in probability to the parameter θ and the estimator $\hat{\beta}_n$ converges in probability to the parameter β. Then*

1. *$\hat{\theta}_n + \hat{\beta}_n$ converges in probability to $\theta + \beta$.*

2. *$\hat{\theta}_n \times \hat{\beta}_n$ converges in probability to $\theta \times \beta$.*

3. *If $\beta \neq 0$, $\hat{\theta}_n / \hat{\beta}_n$ converges in probability to θ/β.*

4. *If $g(x)$ is a continuous function, then $g(\hat{\theta}_n)$ converges in probability to $g(\theta)$.*

Proof: We just prove the first statement. Fix an $\epsilon > 0$. We must show

$$\lim_{n\to\infty} P(|\hat{\theta}_n + \hat{\beta}_n - \theta - \beta| \leq \epsilon) = 1$$

Define events by

$$\begin{aligned} E_n &= \{|\hat{\theta}_n - \theta| \leq \epsilon/2\}, \\ F_n &= \{|\hat{\beta}_n - \beta| \leq \epsilon/2\}, \\ G_n &= \{|\hat{\theta}_n + \hat{\beta}_n - \theta - \beta| \leq \epsilon\} \end{aligned}$$

We know that as $n \to \infty$, $P(E_n)$ and $P(F_n)$ both converge to 1. The triangle inequality shows that $E_n \cap F_n \subset G_n$. So if we can show $P(E_n \cap F_n)$ converges to 1, then $P(G_n)$ converges to 1. We have

$$P(E_n \cap F_n) = 1 - P(E_n^c \cup F_n^c) \geq 1 - P(E_n^c) - P(F_n^c) \to 1$$

□

Note that since $g(x) = \sqrt{x}$ is a continuous function and S^2 is a consistent estimator for σ^2, the last statement in the theorem implies S is a consistent estimator for σ.

End of lecture on Tues, 2/13

Our first application of this theorem is to show that for unbiased estimators, if the variance goes to zero and the bias goes to zero then the estimator is consistent. Recall that the bias is defined to be $B(\hat{\theta}_n) = E[\hat{\theta}] - \theta$.

Theorem 3. *If $\hat{\theta}_n$ is an estimator for θ (possibly biased) such that $\sigma_{\hat{\theta}_n} \to 0$ as $n \to \infty$ and $B(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is consistent.*

Proof: We write the estimator as

$$\hat{\theta}_n = (\theta_n - B(\hat{\theta}_n)) + B(\hat{\theta}_n) \tag{10}$$

Note that the bias $B(\hat{\theta}_n)$ is a constant, i.e., not random. We can think of $\theta_n - B(\hat{\theta}_n)$ as another estimator for θ. It has the same variance as θ_n and so its variance goes to zero. Also, it is unbiased. So by our previous theorem $\theta_n - B(\hat{\theta}_n)$ converges to θ in probability. The sequence of constants $B(\hat{\theta}_n)$ converges to 0, so if we think of these as RV's they converges to 0 in probability. So by part (1) of the previous theorem $\hat{\theta}_n$ converges to θ in probability. □

Application: We have already seen that S_n^2 is a consistent estimator for σ^2. This followed from the fact that the variance of S_n^2 goes to zero. Note that we did not actually compute the variance of S_n^2. We illustrate the application of the previous proposition by giving another proof that S_n^2 is a consistent estimator. Recall

$$S_n^2 = \frac{n}{n+1} \left[\frac{1}{n} \sum_{i=1}^{n} Y_i^2 - (\overline{Y}_n)^2 \right] \tag{11}$$

The proof that this is a consistent estimator using the previous proposition was done in class, but is not included here.

The notion of convergence in probability is one of several senses in which a sequence of RV's can converge. Another important one is the following.

Definition 3. *Let X_n be sequence of RV's, X a RV. Let $F_{X_n}(t)$ and $F_X(t)$ be the CDF's. We say X_n converges to X in distribution if $F_{X_n}(t)$ conveges to $F_X(t)$ for all t such that $F_X(t)$ is continuous at t.*

Proposition 1. *If X_n converges to X in distribution and $F_X(t)$ is continuous at a and b, then $P(a \leq X_n \leq b)$ converges to $P(a \leq X \leq b)$ as $n \to \infty$.*

Partial proof:

$$P(a < X_n \leq b) = F_{X_n}(b) - F_{X_n}(a) \to F_X(b) - F_X(a) = P(a < X \leq b) \tag{12}$$

This is not quite the statement in the proposition since we have $a < \cdots$ rather than $a \leq \cdots$. For X this does not matter since the hypothesis that $F_X(t)$ is continuous at a implies $P(X = a) = 0$. But $P(X_n = a)$ can be nonzero. To finish the proof we would need to prove that this probability goes to 0 as $n \to \infty$. □

Note that the convergence in the central limit theorem is convergence in distribution.

Theorem 4. *If X_n converges in probability to X, then X_n converges to X in distribution.*

Proof: Let t be a point where $F_X(t)$ is continuous. We must show $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$. Let $\epsilon > 0$. Since F_X is continuous at t, there is a $\delta > 0$ such that

$$F_x(t+\delta) < F(t) + \epsilon/2, \quad F_X(t-\delta) > F(t) - \epsilon/2 \tag{13}$$

Since X_n converges in probability to X,

$$\lim_{n\to\infty} P(|X_n - X| \leq \delta) = 0 \tag{14}$$

So there is an N such that

$$n \geq N \Rightarrow P(|X_n - X| \leq \delta) \geq 1 - \epsilon/2 \tag{15}$$

Now let $n \geq N$. We will prove $F_{X_n}(t) \leq F_X(t) + \epsilon$ and $F_{X_n}(t) \geq F_X(t) - \epsilon$. (This will finish the proof.)

For the first inequality:

$$\begin{aligned} F_{X_n}(t) &= P(X_n \leq t) = P(X_n \leq t, |X_n - X| \leq \delta) + P(X_n \leq t, |X_n - X| > \delta) \\ &\leq P(X_n \leq t, |X_n - X| \leq \delta) + P(|X_n - X| > \delta) \end{aligned}$$

For the first term we note that if $X_n \leq t$ and $|X_n - X| \leq \delta$ then $X \leq t + \delta$. The second term above is bounded by $\epsilon/2$.

$$\begin{aligned} F_{X_n}(t) &\leq P(X \leq t+\delta) + \epsilon/2 = F_X(t+\delta) + \epsilon/2 \\ &\leq F_X(t) + \epsilon/2 + \epsilon/2 = F_X(t) + \epsilon \end{aligned}$$

For the second inequality:

$$\begin{aligned} P(X_n > t) &= P(X_n > t, |X_n - X| \leq \delta) + P(X_n > t, |X_n - X| > \delta) \\ &\leq P(X_n > t, |X_n - X| \leq \delta) + P(|X_n - X| > \delta) \end{aligned}$$

The first term is bounded by $P(X > t - \delta)$. The second term is bounded by $\epsilon/2$. So

$$P(X_n > t) \leq P(X > t - \delta) + \epsilon/2$$

So taking 1 minus the above,

$$F_{X_n}(t) \geq F_X(t - \delta) - \epsilon/2 \geq F_X(t - \delta) - \epsilon$$

□

An obvious question is whether the converse is true. It is not. In fact, the converse doesn't really make any sense. X_n can converge to X in distribution even if all the RV's are defines on different probability spacings. But the definition of convergence in probability requires that they all be defined on the same probability space. We do have the following partial converse

Theorem 5. *If X_n is a sequence of RV's that converges in distribution to a constant c, then X_n converges to c in probability.*

Proof: The CDF $F_c(t)$ of the "RV" c is 0 for $t < c$ and 1 for $t > c$. Let $\epsilon > 0$. Then

$$F_{X_n}(c + \delta) - F_{X_n}(c - \delta) = P(c - \delta < X_n \leq c + \delta) \tag{16}$$

As $n \to \infty$, the left side converges to $1 - 0 = 1$. So the probability on the right converges to 1. This shows X_n converges to c in probability. □

Theorem 6. (Slutsky's theorem) *Suppose that the sequence of random variables X_n converges to the random varable X in distribution, and the sequence of random variables Y_n converges in probability to the constant c. Then*

1. *$X_n + Y_n$ converges in distribution to $X + c$.*

2. *$X_n \times Y_n$ converges in distribution to cX.*

3. *If $c \neq 0$, X_n/Y_n converges in distribution to X/c.*

The proof of this theorem is much harder than the previous theorem and we do not give it.

We now turn to an important application of Slutsky's theorem. Recall that if the population is normal, then the statistic

$$T_n = \frac{\overline{Y} - \mu}{\sqrt{S_n^2/n}} \tag{17}$$

has a t-distribution with $n-1$ d.f. for any sample size. What if the population is not normal, but the sample size is large? Write T as a product:

$$T_n = \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}} \frac{\sqrt{\sigma^2}}{\sqrt{S^2}} = X_n Y_n, \tag{18}$$

$$X_n = \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}}, \quad Y_n = \frac{\sqrt{\sigma^2}}{\sqrt{S^2}} \tag{19}$$

The central limit theorem says that X_n converges in distribution to a standard normal. And we know that Y_n converges in probability to 1. So by Slutsky's theorem, T_n converges in distribution to a standard normal.

This implies that

$$P(-z_{\alpha/2} \le T_n \le z_{\alpha/2}) \to 1 - \alpha \tag{20}$$

So for large n,

$$P(-z_{\alpha/2} \le T_n \le z_{\alpha/2}) \approx 1 - \alpha \tag{21}$$

which we can rewrite as

$$P(\mu \in [\overline{Y}_n - z_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{Y}_n + z_{\alpha/2}\frac{S}{\sqrt{n}}]) \approx 1 - \alpha \tag{22}$$

This is the large sample confidence interval we derived before for μ.

We can do a similar thing for the confidence interval for a population proportion for large samples. Let

$$U_n = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \tag{23}$$

Rewrite this as

$$U_n = \frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{p(1-p)}} \tag{24}$$

The first term converges in distribution to standard normal. We know that $\hat{p}_n$ converges in probability to p. So $\frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{p(1-p)}$ converges in probability to 1. So by Slutsky's theorem, U_n converges in distribution to a standard normal. So for large n,

$$P(-z_{\alpha/2} \le U_n \le z_{\alpha/2}) \approx 1-\alpha \tag{25}$$

We can rewrite this as

$$P(p \in [\hat{p}_n - z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \overline{Y}_n + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}]) \approx 1-\alpha \tag{26}$$

This is the large sample confidence interval we derived before for p.

9.4 Sufficiency

Probability review - conditional probability

Recall that for two events A and B, the conditional probability of A given B is

$$P(A|B) = \frac{P(A\cap B)}{P(B)} \tag{27}$$

For discrete RV's this leads immediately to

Definition 4. *If X and Y are discrete RV's, then the conditional pmf of X given Y is*

$$f_{X|Y}(x|y) = P(X=x|Y=y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{28}$$

For each value of Y this given a conditional pmf for X. Note that if X and Y are independent, then $f_{X|Y}(x|y)$ is independent of y and is just $f_X(x)$. We can also define conditional joint pmf's :

Definition 5. *If $X_1, \cdots, X_n$ and Y are discrete RV's, then the conditional joint pmf of $X_1, \cdots, X_n$ given Y is*

$$\begin{aligned} f_{X_1,\cdots,X_n|Y}(x_1,\cdots,x_n|y) &= P(X_1=x_1,\cdots,X_n=x_n|Y=y) && (29) \\ &= \frac{f_{X_1,\cdots,X_n,Y}(x_1,\cdots,x_n,y)}{f_Y(y)} && (30) \end{aligned}$$

If X and Y are continuous RV's, then conditioning on $Y = y$ is problematic since this event has probability zero. You have to define it by a limiting process. We do not go through this argument here. The result is the following definitions.

Definition 6. *If X and Y are continuous RV's, then the conditional pdf of X given Y is*

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{31}$$

Definition 7. *If $X_1, \cdots, X_n$ and Y are continuous RV's, then the conditional joint pdf of $X_1, \cdots, X_n$ given Y is*

$$\begin{aligned} f_{X_1,\cdots,X_n|Y}(x_1, \cdots, x_n|y) &= P(X_1 = x_1, \cdots, X_n = x_n|Y = y) && (32) \\ &= \frac{f_{X_1,\cdots,X_n,Y}(x_1, \cdots, x_n, y)}{f_Y(y)} && (33) \end{aligned}$$

The definitions in the continuous case look exactly the same as in the discrete case, but the f's are now probability densitiies, so the definitions are really quite different.

We now end the probability review and introduce the idea of a sufficient statistic. Informally, a statistic U is sufficient for the population parameter θ, if the statistic contains all the information that the random sample has to tell us about the parameter. In other words, if statistician A is given all the values in the random sample and statistician B is only given the value of U the random sample, then statistician A cannot do a better job of estimating θ than statistician B. The formal definition is as follows.

Definition 8. *Let $Y_1, Y_2, \cdots, Y_n$ be a random sample from a population with an unknown parameter θ. Let U be a statistic, i.e., some function $U = g(Y_1, Y_2, \cdots, Y_n)$ of the random sample. We say that U is a sufficient statistic for θ if the conditional distribution of $Y_1, Y_2, \cdots, Y_n$ given U does not depend on θ.*

Example: We look at an example to see how this definition works. Suppose we are interested in a population proportion and the unknown parameter is the usual p. The estimator we have used before is $\hat{p}$, the sample proportion. Recall that this is just the number of successes in the sample divided by n.

If we let Y_i be 1 if the ith member of the sample is success, and $Y_1 = 0$ if it is failure, then $\hat{p}$ is equal to $\overline{Y}$. We will show that $\hat{p}$ is a sufficient statistic for p. It is a little easier to work with

$$U = \sum_{i=1}^{n} \tag{34}$$

We will show that U is a sufficient statistic. Since $\hat{p} = U/n$, this will show that $\hat{p}$ is a sufficient statistic. The distribution of U is just the binomial distribution:

$$f_U(u) = \binom{n}{u} p^u (1-p)^{n-u} \tag{35}$$

The distribution of $Y_1, Y_2, \cdots, Y_n$ is relatively simple. We get a factor of p for every Y_i that is 1 and a factor of $1-p$ for every $Y_i = 0$. So

$$f_{Y_1,\cdots,Y_n}(y_1, \cdots, y_n) = p^{\sum_i y_i} (1-p)^{\sum_i (1-y_i)} \tag{36}$$

where the sums on i run from 1 to n. Finally, we need the joint distribution of $Y_1, \cdots, Y_n, U$:

$$f_{Y_1,\cdots,Y_n,U}(y_1, \cdots, y_n, u) = \begin{cases} p^u (1-p)^{n-u} & \text{if } \sum_i y_i = u \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

So the conditional joint pmf of $Y_1, \cdots, Y_n$ given U is

$$f_{Y_1,\cdots,Y_n|U}(y_1, \cdots, y_n|u) = \begin{cases} \frac{1}{\binom{n}{u}} & \text{if } \sum_i y_i = u \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

This does not depend on p, so U is a sufficient statistic for p.

We continue the example by giving an example of a statistic that is not sufficient. We still consider a population proportion, but now let

$$U = \sum_{i=1}^{n-1} Y_i \tag{39}$$

Note that the sum only runs up to $n-1$. Intuitively we expect this is not a sufficient statistic since it does not incorporate Y_n. The distribution of U

is binomial, but with $n-1$ trials. Note that Y_n and U are independent. So the joint distribution of $Y_1, \cdots, Y_n, U$ is

$$f_{Y_1,\cdots,Y_n,U}(y_1,\cdots,y_n,u) = f_{Y_1,\cdots,Y_{n-1},U}(y_1,\cdots,y_{n-1},u) f_{Y_n}(y_n) \tag{40}$$

$$= \begin{cases} p^u(1-p)^{n-1-u} p^{Y_n}(1-p)^{1-Y_n} & \text{if } \sum_{i=1}^{n-1} y_i = u \\ 0 & \text{otherwise} \end{cases} \tag{41}$$

So the conditional joint pmf of $Y_1, \cdots, Y_n$ given U is

$$f_{Y_1,\cdots,Y_n|U}(y_1,\cdots,y_n|u) = \begin{cases} p^{Y_n}(1-p)^{1-Y_n} \frac{1}{\binom{n-1}{u}} & \text{if } \sum_{i=1}^{n-1} y_i = u \\ 0 & \text{otherwise} \end{cases} \tag{42}$$

This does depend on p, so this U is not a sufficient statistic for p.

The definition of sufficient statistic is not so easy to check. And it does not give us a method for finding sufficient statistics. Luckily there is a theorem which gives a much easier criterion for sufficiency and tells you how to find them. First we introduce some notation and terminology.

We are usually interested in a random variable Y which we think of as the population. It has a pmf or pdf $f(y)$. This pmf or pdf depends on one more unknown parameters. Let θ or $\theta_1, \cdots, \theta_k$ be these unknown parameters. We make the dependence of $f(y)$ on them explicit by writing it as $f(y|\theta)$ or $f(y|\theta_1, \cdots, \theta_k)$. To simplify the notation we will often just write $f(y|\theta)$ with the understanding that θ may be a single unknown parameter, or may be shorthand for several unknown parameters $\theta_1, \cdots, \theta_k$.

If we draw a random sample $Y_1, Y_2, \cdots, Y_n$ from our population, then the joint pmf or pdf of the random sample is $\prod_{i=1}^n f(y_i|\theta)$. We given this function a name.

Definition 9. *The likelihood function is*

$$L(y_1,\cdots,y_n|\theta) = \prod_{i=1}^n f(y_i|\theta) \tag{43}$$

Theorem 7. *(Factorization criterion) A statistic $U = U(Y_1, Y_2, \cdots, Y_n)$ is a sufficient statistic for the population parameter θ if the likehood function factors into a product of two non-negative functions:*

$$L(y_1,\cdots,y_n|\theta) = g(u,\theta) h(y_1, y_2, \cdots, y_n) \tag{44}$$

where g only depends on u and θ and h does not depend on θ.

Later we will give a proof of this theorem for the case that the random variables are discrete. This theorem does more than help you check that a given statistic is sufficient. It helps you find sufficient statistics. We illustrate this with some examples.

Example: We start with the binomial we have already looked at. So

$$\begin{aligned} L(y_1, \cdots, y_n|\theta) &= \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} && (45) \\ &= (1-p)^n \left[\frac{p}{1-p}\right]^{\sum_i y_i} && (46) \end{aligned}$$

where the sum on i is from 1 to n. So if we define $U = \sum_{i=1}^{n} Y_i$, then the above is of the form in the theorem with

$$\begin{aligned} g(u,p) &= (1-p)^n \left[\frac{p}{1-p}\right]^{u}, && (47) \\ h(y_1, \cdots, y_n) = 1 & && (48) \end{aligned}$$

This U is a sufficent statistic. This of course implies $U/n = \hat{p}$ is also a sufficient statistic.

End of lecture on Tues, 2/27

Remark: If U is a statistic and $V = f(U)$ is a function of U that is a sufficient statistic, then U is a sufficient statistic. This follows immediately from the factorization theorem. Since V is sufficient we can factor the likelihood function as $g(v,\theta)h(y_1, \cdots, y_n)$. But then we can express v as a function of u, so this becomes $g(v(u),\theta)h(y_1, \cdots, y_n)$, and we can think of g as a function of u and θ only. Caution: If U is sufficient and $V = f(U)$, then V need not be sufficient. (It will be if f is 1-1.)

Example: Suppose the population has an exponential distribution with mean β. So

$$f(y|\beta) = \frac{1}{\beta} e^{-y/\beta} 1(y \geq 0) \qquad (49)$$

Then

$$L(y_1, \cdots, y_n|\theta) = \beta^{-n} \exp(-\frac{1}{\beta}\sum_{i=1}^{n} y_i) \prod_{i=1}^{n} 1(y_i \geq 0) \qquad (50)$$

We can take $h(y_1, \cdots, y_n) = \prod_{i=1}^n 1(y_i \geq 0)$, and let $g(u, \beta)$ be the rest of the likelihood function where $U = \sum_{i=1}^n Y_i$. So U is a sufficient statistic. This implies that $\overline{Y}$ is also a sufficient statistic.

The last example might make you wonder if the sample mean $\overline{Y}$ is always a sufficient statistic for the population mean μ. The next example shows it is not, and in general it is not.

Example: Suppose the population has a uniform distribution on $[0, \theta]$ with θ unknown. Pay careful attention to the range here. The pdf of Y is not $1/\theta$. It is $1/\theta$ when $0 \leq y \leq \theta$ and 0 otherwise. We let $1(\cdots)$ denote the function that is 1 when $\cdots$ is true and it is 0 when $\cdots$ is false. So

$$f(y|\theta) = \frac{1}{\theta} 1(0 \leq Y \leq \theta) = \frac{1}{\theta} 1(0 \leq Y) 1(Y \leq \theta) \tag{51}$$

So the likelihood function is

$$L(y_1, \cdots, y_n|\theta) = \theta^{-n} \prod_i 1(0 \leq y_i) \prod_i 1(y_i \leq \theta) \tag{52}$$

We can put the factor θ^{-n} into $g(\theta, u)$ whatever U is. And we can put the factors of $1(0 \leq y_i)$ into $h(y_1, \cdots, y_n)$. The function $\prod_i 1(y_i \leq \theta)$ depends on θ, but there is no way to write this as a function of θ and $\overline{Y}$. So $\overline{Y}$ is not a sufficient statistic. If we only know $\sum_i y_i$, there is no way we can determine if all the y_i are $\leq \theta$. The function $\prod_i 1(y_i \leq \theta)$ is 1 if and only if all the $y_i \leq \theta$. So

$$\prod_i 1(y_i \leq \theta) = 1(\max\{y_1, \cdots, y_n\} \leq \theta) \tag{53}$$

But it is a function of $U = \max\{Y_1, \cdots, Y_n\}$ and θ, so we see that this max is a sufficient statistic.

Example: Suppose the population pdf is gamma with parameters α and β. So

$$f(y|\alpha, \beta) = \frac{y^{\alpha-1} e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)} \tag{54}$$

for $y \geq 0$. The likelihood function is

$$L(y_1, \cdots, y_n|\alpha, \beta) = [\prod_i y_i]^{\alpha-1} \exp(-\sum_i y_i/\beta)[\beta^\alpha \Gamma(\alpha)]^{-n} \tag{55}$$

If α is known, and β is unknown, then $\sum_i Y_i$ is a sufficient statistic for β. If β is known, and α is unknown, then $\prod_i Y_i$ is a sufficient statistic for α. But if both α and β are unknown, then there is no single sufficient statistic.

When the population has more than one unknown parameter, it is typically not possible to find a single sufficient statistic.

Definition 10. *Let $Y_1, Y_2, \cdots, Y_n$ be a random sample from a population with unknown parameter θ_1 and θ_2. Let U_1 and U_2 be two statistics, i.e., functions $U_i = g_i(Y_1, Y_2, \cdots, Y_n)$ of the random sample. We say that U_1, U_2 are sufficient statistics for θ_1, θ_2 if the conditional distribution of $Y_1, Y_2, \cdots, Y_n$ given U_1, U_2 does not depend on θ_1 or θ_2.*

The Factorization theorem for two parameters and two statistics is

Theorem 8. *(Factorization criterion) Statistics U_1 and U_2 are sufficient statistics for the population parameters θ_1, θ_2 if the likehood function factors into a product of two non-negative functions:*

$$L(y_1, \cdots, y_n|\theta_1, \theta_2) = g(u_1, u_2, \theta_1, \theta_2) h(y_1, y_2, \cdots, y_n) \tag{56}$$

where g only depends on u_1, u_2 and θ_1, θ_2 and h does not depend on θ_1 or θ_2.

Example: Suppose the population pdf is gamma with parameters α and β. So

$$f(y|\alpha, \beta) = \frac{y^{\alpha-1} e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)} \tag{57}$$

for $y \geq 0$. The likelihood function is

$$L(y_1, \cdots, y_n|\alpha, \beta) = [\prod_i y_i]^{\alpha-1} \exp(-\sum_i y_i/\beta)[\beta^\alpha \Gamma(\alpha)]^{-n} \tag{58}$$

So

$$U_1 = \prod_{i=1}^n Y_i, \quad U_2 = \sum_{i=1}^n Y_i \tag{59}$$

are sufficient statistics for α, β.

Remark: The number of sufficient statistics can be greater than the number of unknown parameters. For example, consider the examples above where the population had a gamma distribution. Suppose α is known and β is unknown. We saw that $\sum_i Y_i$ is a sufficient statistic. But the two statistics $\sum_i Y_i$ and $\prod_i Y_i$ are also sufficient statistics. As an extreme example, we can think of the entire random sample as n statistics. This n-tuple of statistics will be sufficient no matter what the population distribution is. Of course, we would like to have as few sufficient statistics as possible.

Recall that a sufficient statistic is a statistic that contains all the information that the random sample has to offer about θ. Intuitively, a minimal sufficient statistic is a sufficient statistic that does not contain any superfluous information.

Definition 11. *A sufficient statistic $U = U(Y_1, \cdots, Y_n)$ is a minimal sufficient statistic if for any sufficient statistic $T = T(Y_1, \cdots, Y_n)$, U is a function of T. (This means that $T(x_1, \cdots, x_n) = T(y_1, \cdots, y_n)$ implies $U(x_1, \cdots, x_n) = U(y_1, \cdots, y_n)$.)*

In all the examples we have considered we have been able to find a sufficient statistic. This is misleading. If we have one unknown parameter it is not always possible to find a single sufficient statistic. An example is the Weibull distribution:

$$f(y|\theta) = \theta y^{\theta-1} \exp(-y^\theta), y \geq 0 \tag{60}$$

The likelihood function is

$$\begin{aligned} L(y_1, \cdots, y_n, \theta) &= \theta^n \prod_{i=1}^n \left[y_i^{\theta-1} \exp(-y_i^\theta) \right] & (61) \\ &= \theta^n [\prod_{i=1}^n y_i]^{\theta-1} \exp(-\sum_i y_i^\theta) & (62) \end{aligned}$$

You cannot find a $u(y_1, \cdots, y_n)$ so that you write this in the form called for in the factorization theorem.

9.5 Mininum variance unbiased estimators (MVUE)

Suppose we restrict consideration to unbiased estimators for θ. In this case the mean square error (MSE) is equal to the variance of the estimator. So ideally we would like the find the unbiased estimator with the smallest variance. Note that the variance will usually depend on the population parameters. Ideally we would like to find an unbiased estimator of θ that has the smallest variance for all θ. Such an estimator is called a minimum variance unbiased estimator (MVUE). We start with a definition.

The next definition is not intuitive at all. Luckily, it holds for almost all the statistics and population distributions we typically consider.

Definition 12. *Let $U = U(Y_1, \cdots, Y_n)$ be a statistic for θ. U is complete if for every function g such that $E_\theta[g(U)] = 0$ for all θ, we have $g(U) = 0$.*

Note completeness is not just a property of the statistic. It also depends on the family of population distibutions that θ parameterizes.

Theorem 9. *Let U be a complete, sufficient statistic for θ. If $E[U] = \theta$, then U is the unique MVUE of θ. More generally, let $T = g(U)$ be a function of U. Let $\tau(\theta) = E_\theta[T] = E_\theta[g(U)]$. Then T is the unique MVUE of $\tau(\theta)$.*

Remark: The function g in the theorem can depend on n. For example, if $U = \sum_{i=1}^n$, then $\overline{Y} = U/n$ is a legitimate function of U.

The proof of this theorem is way beyond the scope of this course. It is a very useful theorem since it give a method for possibly finding the MVUBE for θ. First use the factorization theorem to find a sufficient statistic U. Then find a function of U whose expected value is θ. You may not be able to do all this, but if you succeed you will have the MVUE. We illustrate this approach with some examples.

Example: Consider an example we have looked at extensively - a population proportion. So the likelihood function

$$\begin{aligned} L(y_1, \cdots, y_n|\theta) &= \prod_{i=1}^n p^{y_i}(1-p)^{1-y_i} && (63) \\ &= (1-p)^n \left[\frac{p}{1-p}\right]^{\sum_i y_i} && (64) \end{aligned}$$

where the sum on i is from 1 to n. We saw before that $U = \sum_{i=1}^n Y_i$ is a sufficient statistic. We have $E[U] = np$. Let $\hat{p}$ denote U/n. So $\hat{p}$ is the usual sample proportion. It is an unbiased estimator of p and it is a function of the suffcient statistic U. Assuming it is complete, the theorem says it is a MVUE for p.

Example: Suppose the population has an exponential distribution with mean β. So

$$L(y_1, \cdots, y_n|\theta) = \beta^{-n} \exp(-\frac{1}{\beta} \sum_{i=1}^{n} y_i) \tag{65}$$

We saw before that $U = \sum_{i=1}^{n} Y_i$ is a sufficient statistic. Just as in the previous example, $\overline{Y} = U/n$ is an unbiased estimator which is a function of this sufficient statistic, so $\overline{Y}$ is the MVUE of of β. Recall that the variance of the exponential is $\sigma^2 = \beta^2$. Suppose we want a MVUE for σ^2. Note that $(\overline{Y})^2$ is a "reasonable" estimator for σ^2, but it is biased. In fact, we can compute its expected value. We know $var(\overline{Y}) = \beta^2/n$ and $E[\overline{Y}] = \beta$. So

$$E[(\overline{Y})^2] = var(\overline{Y}) + (E[\overline{Y}])^2 = \beta^2/n + \beta^2 = \frac{n+1}{n}\beta^2 = \frac{n+1}{n}\sigma^2 \tag{66}$$

So the MUVE for σ^2 is

$$\hat{\sigma^2} = \frac{n}{n+1}(\overline{Y})^2 \tag{67}$$

Example: Consider a population with a normal distribution with mean μ and variance σ^2. You will show in the homework that $\overline{Y}$ and S^2 are sufficient statistics for μ and σ^2. Since they are unbiased estimators for μ and σ^2, they are MVUE for them.

Example: Consider a population with distribution

$$f(y|\theta) = \frac{3y^2}{\theta^3} 1(0 \le y \le \theta) \tag{68}$$

Show that $U = \max\{Y_1, \cdots, Y_n\}$ is a sufficient statistic. So we need to find a function of U whose expected value is θ. We start by finding the distribution of U.

$$f_U(u|\theta) = \frac{3ny^{3n-1}}{\theta^{3n}} 1(0 \le u \le \theta) \tag{69}$$

From this we compute that

$$E[U] = \frac{3n}{3n+1}\theta \tag{70}$$

So if we let

$$\hat{\theta} = \frac{3n+1}{3n} U \tag{71}$$

then $\hat{\theta}$ is a function of the sufficient statistic U which is an unbiased estimator of θ. So $\hat{\theta}$ is the MVUE of θ.

We end this section with part of the theory behind the theorem we stated above which is of interest in its own right.

Theorem 10. *(Rao-Blackwell theorem) Let U be a sufficient statistic for θ. Suppose T is an unbiased estimator of $\tau(\theta)$. Let $\phi(U) = E[T|U]$. Then $\phi(U)$ is an unbiased estimator of θ and*

$$Var_\theta(\phi(U)) \leq Var_\theta(T) \tag{72}$$

for all θ.

If we have a sufficient statistic U and an unbiased estimator T, then we can get a better unbiased estimator than T by taking the conditional expectation $E[T|U]$. However, actually computing this conditional expectation may be very difficult. The Rao-Blackwell theorem involves the conditional expectation $[T|U]$. We need to review some probability to understand what this is.

GAP GAP GAP GAP

End of lecture on Tues, 3/13

9.6 Method of moments

Let Y be the RV associated with the population. Recall that the kth moment of Y is EY^k. We denote it by μ_k. It will be a function of the population parameters. For a random sample $Y_1, Y_2, \cdots, Y_n$ the sample moments are defined by

$$m_k = \frac{1}{n} \sum_{i=1}^{n} Y_i^k \tag{73}$$

Note that m_1 is the sample mean $\overline{Y}$. The sample moments are of course functions of the random sample.

Suppose we want to estimate k population parameters $\theta_1, \theta_2, \cdots, \theta_k$. The method of moments does this setting the first k sample momemts equal to the first k population moments, i.e., $\mu_i = m_i$ for $i = 1, 2, \cdots, k$. This gives a system of k equations in which we think of the random sample $Y_1, Y_2, \cdots, Y_n$ as known and $\theta_1, \cdots, \theta_k$ as the unknowns. We solve for $\theta_1, \cdots, \theta_k$ as functions of $Y_1, \cdots, Y_n$.

If we just have one parameter we want to estimate, then the method just becomes the follows. Find the population mean as a function of θ. Call it $\mu(\theta)$. We set $\mu(\theta) = \overline{Y}$ and then solve for θ as a function of $\overline{Y}$. Note that in the case of one unknown parameter, the method of moments will always give an estimator for θ that is a function of $\overline{Y}$.

The method of moments typically produces a consistent estimator(s). But they need not be unbiased and when there are unbiased they need not have minium variance.

Example: Suppose the population is uniformly distributed on $[0, \theta]$ with θ unknown. The population mean is $\mu = \theta/2$. So we set $\theta/2 = \overline{Y}$ and solve for θ. So our estimator is

$$\hat{\theta} = 2\overline{Y} \tag{74}$$

Example: Suppose the population is uniformly distributed on $[\theta_1, \theta_2]$. We want to use the method of moments to find estimators $\hat{\theta_1}$ and $\hat{\theta_2}$ for the two unknown parameters. The mean of the population RV is

$$\mu_1 = \frac{1}{2}(\theta_2 + \theta_1) \tag{75}$$

Its second moment is

$$\mu_2 = \int_{\theta_1}^{\theta_2} \frac{y^2}{\theta_2 - \theta_1}\, dy = \frac{1}{3}\frac{\theta_2^3 - \theta_1^3}{\theta_2 - \theta_1} = \frac{1}{3}(\theta_2^2 + \theta_1\theta_2 + \theta_1^2) \tag{76}$$

So we have the two equations

$$\begin{aligned} m_1 &= \frac{1}{2}(\hat{\theta_2} + \hat{\theta_1}), & (77) \\ m_2 &= \frac{1}{3}(\hat{\theta_2}^2 + \hat{\theta_1}\hat{\theta_2} + \hat{\theta_1}^2) & (78) \end{aligned}$$

We then solve them for $\hat{\theta_1}$ and $\hat{\theta_2}$ in terms of m_1 and m_2.

Example: Suppose the population pdf is gamma with parameters α and β. So

$$f(y|\alpha,\beta) = \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)} \tag{79}$$

for $y \geq 0$. The mean is $\mu = \alpha\beta$ and the variance is $\sigma^2 = \alpha\beta^2$. So

$$\mu_1 = \alpha\beta, \quad \mu_2 = \alpha\beta^2 + \alpha^2\beta^2 \tag{80}$$

So to find the method of moment estimators for α and β we must solve

$$m_1 = \hat{\alpha}\hat{\beta}, \quad m_2 = \hat{\alpha}\hat{\beta}^2 + \hat{\alpha}^2\hat{\beta}^2 \tag{81}$$

for $\hat{\alpha}$ and $\hat{\beta}$. Using the first equation we can rewrite the second eq. as

$$m_2 = m_1\hat{\beta} + m_1^2 \tag{82}$$

So

$$\hat{\beta} = \frac{m_2 - m_1^2}{m_1} \tag{83}$$

Then we get

$$\hat{\alpha} = \frac{m_1}{\hat{\beta}} = \frac{m_1^2}{m_2 - m_1^2} \tag{84}$$

We claim that these are consistent estimator of α and β

SHOW THIS

Consider the case where there is only one unknown parameter θ. The mean will be a function of θ, but we may not be able to compute this function (or its inverse) explicitly

Example Consider again the Weibull distribution:

$$f(y|\theta) = \theta y^{\theta-1}\exp(-y^\theta), y \geq 0 \tag{85}$$

The likelihood function is

$$\begin{aligned} L(y_1,\cdots,y_n,\theta) &= \theta^n \prod_{i=1}^n \left[y_i^{\theta-1}\exp(-y_i^\theta)\right] && (86) \\ &= \theta^n [\prod_{i=1}^n y_i]^{\theta-1}\exp(-\sum_i y_i^\theta) && (87) \end{aligned}$$

I put this in wolfram alpha, and it thought for a bit and then said "standard computation time exceeded." You can make a substitution $u = y^\theta$ and the new integral you get can be expressed in terms of the gamma function. But then you need to invert the gamma function which I don't know how to do explicit. But you could still find the estimator numerically.

9.7 Maximum likelihood estimation (MLE)

Recall that the likelihood function $L(y_1, \cdots, y_n|\theta)$ is the joint distribution of the random sample which depends on the population parameters θ. (θ here can be a single parameters or shorthand for $\theta_1, \cdots, \theta_k$.) We think of $y_1, \cdots, y_n$ as fixed and maximize $L(y_1, \cdots, y_n|\theta)$ as a function of θ. This defines (implicitly) θ as a function of $y_1, \cdots, y_n$.

Computational trick: Since the log function is monotonic, finding the θ that maximizes $L(y_1, \cdots, y_n|\theta)$ is equivalent to finding the θ that maximizes $\ln L(y_1, \cdots, y_n|\theta)$. We typically find where the max occurs by differentiating with respect to θ('s) and setting the derivative(s) to zero. Doing this for $\ln L$ instead of L is simpler since we have a sum to differentiate rather than a product.

Example: Suppose that the population has a Poisson distribution with unknown parameter λ. The likelihood function is

$$L(y_1, \cdots, y_n|\lambda) = \prod_{i=1}^{n} [e^{-\lambda} \frac{\lambda^{y_i}}{n!}] \tag{88}$$

So

$$\ln L(y_1, \cdots, y_n|\lambda) = \sum_{i=1}^{n} [-\lambda + y_i \ln \lambda - \ln(n!)] \tag{89}$$

Taking the derivative with respect to λ and setting it to 0 gives

$$-n + \sum_{i=1}^{n} \frac{y_i}{\lambda} = 0 \tag{90}$$

Solving for λ we get the MLE

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{Y} \tag{91}$$

Example: Suppose the population is uniform on $[0, \theta]$. The likelihood function is

$$L(y_1, \cdots, y_n) = \theta^{-n} \prod_{i=1}^{n} 1(0 \leq y_i \leq \theta) \tag{92}$$

We rewrite this as

$$L(y_1, \cdots, y_n) = \theta^{-n} 1(0 \leq \max\{y_1, \cdots, y_n\} \leq \theta) \tag{93}$$

The y_i are fixed, so the max is fixed. We want to maximize this as a function of θ. Note that the likelihood function is zero when $\theta < \max\{y_1, \cdots, y_n\}$. So the maximum will occur in the range $[\max\{y_1, \cdots, y_n\}, \infty)$. Note that θ^{-n} is a decreasing function of θ. So the maximum occurs at the endpoint $\theta = \max\{y_1, \cdots, y_n\}$. So the MLE is

$$\hat{\theta} = \max\{Y_1, \cdots, Y_n\} \tag{94}$$

End of lecture on Thurs, 3/15

Example Consider again the Weibull distribution:

$$f(y|\theta) = \theta y^{\theta-1} \exp(-y^\theta), y \geq 0 \tag{95}$$

The likelihood function is

$$\begin{aligned} L(y_1, \cdots, y_n, \theta) &= \theta^n \prod_{i=1}^{n} \left[y_i^{\theta-1} \exp(-y_i^\theta) \right] & (96) \\ &= \theta^n [\prod_{i=1}^{n} y_i]^{\theta-1} \exp(-\sum_i y_i^\theta) & (97) \end{aligned}$$

So

$$\ln L(y_1, \cdots, y_n, \theta) = n \ln(\theta) + (\theta - 1) \sum_{i=1}^{n} \ln(y_i) - \sum_i y_i^\theta \tag{98}$$

Taking derivative wrt θ and setting it to zero:

$$\frac{n}{\theta} + \sum_{i=1}^{n} \ln(y_i) - \sum_i y_i^\theta \ln(y_i) = 0 \tag{99}$$

Now all we need to do is solve for θ. Good luck. But we can do it numerically given specific values for the sample $y_1, \cdots, y_n$.

MLE and sufficient statistics Suppose we have a sufficient statistic for θ and we find the MLE estimator. By the factorization theorem

$$L(y_1, \cdots, y_n|\theta) = g(u, \theta)h(y_1, \cdots, y_n) \qquad (100)$$

where h does not depend on θ. So finding the θ that maximizes L is the same as finding the θ that maximizes $g(u, \theta)$. But this will obviously give θ as a function of u. So we conclude that if there is a sufficient statistic, then the MLE will be a function of this statistic.

Invariance principle Suppose that $\hat{\theta}$ is the MLE for the parameter θ. Let $\tau = t(\theta)$ be a function of θ. Suppose we want to find the MLE for τ. Then we have to think of the likelihood function as a function of τ: $L(y_1, \cdots, y_n|\tau)$. We can do this if $t(\theta)$ is invertible, i.e., if $t()$ is 1-1. Now we need to minimize $L(y_1, \cdots, y_n|\tau)$ as a function of τ. But we already know that as a function of θ the min occurs at $\hat{\theta}$. So as a function of τ it will occur at $t(\hat{\theta})$. This is called the invariance principle. It can be stated as

$$\widehat{t(\theta)} = t(\hat{\theta}) \qquad (101)$$

where both hat's mean the MLE.

Now suppose we have two unknown parameters for the population distribution. So the likehood function is $L(y_1, \cdots, y_n|\theta_1, \theta_2)$. We find the MLE estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ by maximizing L as a function of θ_1 and θ_2.

Example: Suppose that the population has a normal distribution with unknown mean μ and unknown variance σ^2. We seek maximum likelihood estimators for μ and σ^2. The likelihood function is

$$L(y_1, \cdots, y_n|\mu, \sigma) = \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(y_i - \mu)^2/\sigma^2) \right] \qquad (102)$$

So

$$\ln(L(y_1, \cdots, y_n|\mu, \sigma)) = -\frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i}^{n} (y_i - \mu)^2 \qquad (103)$$

Take derivative wrt μ and we find

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y} \tag{104}$$

Take derivative wrt σ^2 and we find

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 = 0 \tag{105}$$

Use the value of μ from above and we get

$$\widehat{\sigma^2} = \frac{1}{n} (y_i - \overline{y})^2 \tag{106}$$

9.8 MLE estimators for large samples

We end our discussion of MLE estimators with some material beyond the scope of the course. To keep things simple restrict our attention to one unknown population parameter.

Theorem 11. *Suppose the population has one unknown parameter θ and $\hat{\theta}_n$ is the MLE for θ. Under some regularity conditions on the family of distributions, $\hat{\theta}_n$ is a consistent estimator of θ, i.e., $\hat{\theta}_n$ converges to θ in probability.*

Of course this result does not tell us anything about how fast the MLE converges to θ. And it does not say anything about the distribution of the MLE estimator.

Definition 13. *An estimator $\hat{\theta}_n$ for θ is asymptotically normal if there is a constant σ_θ^2 such that $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to a normal distribution with mean* 0 *and variance σ_θ^2.*

Theorem 12. *Suppose the population has one unknown parameter θ and $\hat{\theta}_n$ is the MLE for θ. Under some regularity conditions on the family of distributions, $\hat{\theta}_n$ is asymptotically normal.*

There is a formula for σ_θ^2.

$$\sigma_{\theta_0}^2 = -\frac{1}{E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\ln(f(Y|\theta))|_{\theta=\theta_0}\right]} \tag{107}$$

Suppose that we want to estimate some function of θ, say $t(\theta)$, rather than θ itself. The invariance principle says the MLE is $t(\hat{\theta})$. Furthermore the estimator $t(\hat{\theta})$ is consistent and asymptotically normal for θ. The variance now is

$$\sigma_{t(\theta_0)}^2 = -\frac{(\frac{\partial t(\theta)}{\partial\theta})^2}{E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\ln(f(Y|\theta))|_{\theta=\theta_0}\right]} \tag{108}$$

When the estimator is asymptotically normal, we know (approximately) the distribution of the estimator and we can use this to construct confidence intervals. The confidence interval for $t(\theta)$ is

$$t(\hat{\theta}) \pm z_{\alpha/2}\frac{\sigma_{t(\hat{\theta})}}{\sqrt{n}} \tag{109}$$

Note that our formula for $\sigma_{t(\theta)}$ depends on the unknown parameter θ. So we have replace θ by $\hat{\theta}$ in the confidence interval.

Example: Consider a population proportion. So the unknown parameter is p and

$$f(y|p) = p^y(1-p)^{1-y} \tag{110}$$

and $y = 0, 1$.

$$\frac{\partial}{\partial p}\ln(f(Y|p) = \frac{Y}{p} - \frac{1-Y}{1-p} \tag{111}$$

$$\frac{\partial^2}{\partial^2 p}\ln(f(Y|p) = -\frac{Y}{p^2} + -\frac{1-Y}{(1-p)^2} \tag{112}$$

So

$$-E[\frac{\partial^2}{\partial^2 p}\ln(f(Y|p)] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)} \tag{113}$$

So $\sigma_\theta^2 = p(1-p)$. Thus .... which we knew already from CLT.

Now look at the population variance $p(1-p)$. By the invariance principle for MLE, the MLE estimator for this variance is $\hat{p}(1-\hat{p})$.

Confidence interval is

$$\hat{p}(1-\hat{p}) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})(1-2\hat{p})^2}{n}} \tag{114}$$