

Lecture 6. Order Statistics

6.1 The Multinomial Formula

Suppose we pick a letter from $\{A, B, C\}$, with $P(A) = p_1 = .3, P(B) = p_2 = .5, P(C) = p_3 = .2$. If we do this independently 10 times, we will find the probability that the resulting sequence contains exactly 4 A's, 3 B's and 3 C's.

The probability of $AAAABBBCCC$, in that order, is $p_1^4 p_2^3 p_3^3$. To generate all favorable cases, select 4 positions out of 10 for the A's, then 3 positions out of the remaining 6 for the B's. The positions for the C's are then determined. One possibility is $BCAABACCAB$. The number of favorable cases is

$$\binom{10}{4} \binom{6}{3} = \frac{10!}{4!6!} \frac{6!}{3!3!} = \frac{10!}{4!3!3!}.$$

Therefore the probability of exactly 4 A's, 3 B's and 3 C's is

$$\frac{10!}{4!3!3!} (.3)^4 (.5)^3 (.2)^3$$

In general, consider n independent trials such that on each trial, the result is exactly one of the events A_1, \dots, A_r , with probabilities p_1, \dots, p_r respectively. Then the probability that A_1 occurs exactly n_1 times, \dots , A_r occurs exactly n_r times, is

$$p_1^{n_1} \cdots p_r^{n_r} \binom{n}{n_1} \binom{n-n_1}{n_2} \binom{n-n_1-n_2}{n_3} \cdots \binom{n-n_1-\cdots-n_{r-2}}{n_{r-1}} \binom{n_4}{n_r}$$

which reduces to the *multinomial formula*

$$\frac{n!}{n_1! \cdots n_r!} p_1^{n_1} \cdots p_r^{n_r}$$

where the p_i are nonnegative real numbers that sum to 1, and the n_i are nonnegative integers that sum to n .

Now let X_1, \dots, X_n be iid, each with density $f(x)$ and distribution function $F(x)$. Let $Y_1 < Y_2 < \cdots < Y_n$ be the X_i 's arranged in increasing order, so that Y_k is the k -th smallest. In particular, $Y_1 = \min X_i$ and $Y_n = \max X_i$. The Y_k 's are called the *order statistics* of the X_i 's

The distributions of Y_1 and Y_n can be computed without developing any new machinery. The probability that $Y_n \leq x$ is the probability that $X_i \leq x$ for all i , which is $\prod_{i=1}^n P\{X_i \leq x\}$ by independence. But $P\{X_i \leq x\}$ is $F(x)$ for all i , hence

$$F_{Y_n}(x) = [F(x)]^n \quad \text{and} \quad f_{Y_n}(x) = n[F(x)]^{n-1} f(x).$$

Similarly,

$$P\{Y_1 > x\} = \prod_{i=1}^n P\{X_i > x\} = [1 - F(x)]^n.$$

Therefore

$$F_{Y_1}(x) = 1 - [1 - F(x)]^n \quad \text{and} \quad f_{Y_1}(x) = n[1 - F(x)]^{n-1}f(x).$$

We compute $f_{Y_k}(x)$ by asking how it can happen that $x \leq Y_k \leq x + dx$ (see Figure 6.1). There must be $k - 1$ random variables less than x , one random variable between x and $x + dx$, and $n - k$ random variables greater than x . (We are taking dx so small that the probability that more one random variable falls in $[x, x + dx]$ is negligible, and $P\{X_i > x\}$ is essentially the same as $P\{X_i > x + dx\}$. Not everyone is comfortable with this reasoning, but the intuition is very strong and can be made precise.) By the multinomial formula,

$$f_{Y_k}(x) dx = \frac{n!}{(k-1)!1!(n-k)!} [F(x)]^{k-1} f(x) dx [1 - F(x)]^{n-k}$$

so

$$f_{Y_k}(x) = \frac{n!}{(k-1)!1!(n-k)!} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x).$$

Similar reasoning (see Figure 6.2) allows us to write down the joint density $f_{Y_j Y_k}(x, y)$ of Y_j and Y_k for $j < k$, namely

$$\frac{n!}{(j-1)!(k-j-1)!(n-k)!} [F(x)]^{j-1} [F(y) - F(x)]^{k-j-1} [1 - F(y)]^{n-k} f(x)f(y)$$

for $x < y$, and 0 elsewhere. [We drop the term $1!$ ($=1$), which we retained for emphasis in the formula for $f_{Y_k}(x)$.]

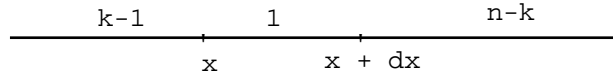


Figure 6.1

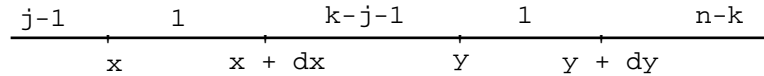


Figure 6.2

Problems

1. Let $Y_1 < Y_2 < Y_3$ be the order statistics of X_1, X_2 and X_3 , where the X_i are uniformly distributed between 0 and 1. Find the density of $Z = Y_3 - Y_1$.
2. The formulas derived in this lecture assume that we are in the continuous case (the distribution function F is continuous). The formulas do not apply if the X_i are discrete. Why not?

3. Consider order statistics where the $X_i, i = 1, \dots, n$, are uniformly distributed between 0 and 1. Show that Y_k has a beta distribution, and express the parameters α and β in terms of k and n .
4. In Problem 3, let $0 < p < 1$, and express $P\{Y_k > p\}$ as the probability of an event associated with a sequence of n Bernoulli trials with probability of success p on a given trial. Write $P\{Y_k > p\}$ as a finite sum involving n, p and k .

Lecture 7. The Weak Law of Large Numbers

7.1 Chebyshev's Inequality

- (a) If $X \geq 0$ and $a > 0$, then $P\{X \geq a\} \leq E(X)/a$.
- (b) If X is an arbitrary random variable, c any real number, and $\epsilon > 0, m > 0$, then $P\{|X - c| \geq \epsilon\} \leq E(|X - c|^m)/\epsilon^m$.
- (c) If X has finite mean μ and finite variance σ^2 , then $P\{|X - \mu| \geq k\sigma\} \leq 1/k^2$.

This is a universal bound, but it may be quite weak in a specific cases. For example, if X is normal (μ, σ^2) , abbreviated $N(\mu, \sigma^2)$, then

$$P\{|X - \mu| \geq 1.96\sigma\} = P\{|N(0, 1)| \geq 1.96\} = 2(1 - \Phi(1.96)) = .05$$

where Φ is the distribution function of a normal $(0, 1)$ random variable. But the Chebyshev bound is $1/(1.96)^2 = .26$.

Proof.

- (a) If X has density f , then

$$E(X) = \int_0^\infty xf(x) dx = \int_0^a xf(x) dx + \int_a^\infty xf(x) dx$$

so

$$E(X) \geq 0 + \int_a^\infty af(x) dx = aP\{X \geq a\}.$$

- (b) $P\{|X - c| \geq \epsilon\} = P\{|X - c|^m \geq \epsilon^m\} \leq E(|X - c|^m)/\epsilon^m$ by (a).
- (c) By (b) with $c = \mu, \epsilon = k\sigma, m = 2$, we have

$$P\{|X - \mu| \geq k\sigma\} \leq \frac{E[(X - \mu)^2]}{k^2\sigma^2} = \frac{1}{k^2}. \clubsuit$$

7.2 Weak Law of Large Numbers

Let X_1, \dots, X_n be iid with finite mean μ and finite variance σ^2 . For large n , the arithmetic average of the observations is very likely to be very close to the true mean μ . Formally, if $S_n = X_1 + \dots + X_n$, then for any $\epsilon > 0$,

$$P\left\{\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right\} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Proof.

$$P\left\{\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right\} = P\{|S_n - n\mu| \geq n\epsilon\} \leq \frac{E[(S_n - n\mu)^2]}{n^2\epsilon^2}$$

by Chebyshev (b). The term on the right is

$$\frac{\text{Var } S_n}{n^2\epsilon^2} = \frac{n\sigma^2}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \rightarrow 0. \clubsuit$$

7.3 Bernoulli Trials

Let $X_i = 1$ if there is a success on trial i , and $X_i = 0$ if there is a failure. Thus X_i is the indicator of a success on trial i , often written as $I[\text{Success on trial } i]$. Then S_n/n is the relative frequency of success, and for large n , this is very likely to be very close to the true probability p of success.

7.4 Definitions and Comments

The convergence illustrated by the weak law of large numbers is called *convergence in probability*. Explicitly, S_n/n converges in probability to μ . In general, $X_n \xrightarrow{P} X$ means that for every $\epsilon > 0$, $P\{|X_n - X| \geq \epsilon\} \rightarrow 0$ as $n \rightarrow \infty$. Thus for large n , X_n is very likely to be very close to X . If X_n converges in probability to X , then X_n converges to X *in distribution*: If F_n is the distribution function of X_n and F is the distribution function of X , then $F_n(x) \rightarrow F(x)$ at every x where F is continuous. To see that the continuity requirement is needed, look at Figure 7.1. In this example, X_n is uniformly distributed between 0 and $1/n$, and X is identically 0. We have $X_n \xrightarrow{P} 0$ because $P\{|X_n| \geq \epsilon\}$ is actually 0 for large n . However, $F_n(x) \rightarrow F(x)$ for $x \neq 0$, but not at $x = 0$.

To prove that convergence in probability implies convergence in distribution:

$$\begin{aligned} F_n(x) &= P\{X_n \leq x\} = P\{X_n \leq x, X > x + \epsilon\} + P\{X_n \leq x, X \leq x + \epsilon\} \\ &\leq P\{|X_n - X| \geq \epsilon\} + P\{X \leq x + \epsilon\} \\ &= P\{|X_n - X| \geq \epsilon\} + F(x + \epsilon) \\ F(x - \epsilon) &= P\{X \leq x - \epsilon\} = P\{X \leq x - \epsilon, X_n > x\} + P\{X \leq x - \epsilon, X_n \leq x\} \\ &\leq P\{|X_n - X| \geq \epsilon\} + P\{X_n \leq x\} \\ &= P\{|X_n - X| \geq \epsilon\} + F_n(x). \end{aligned}$$

Therefore

$$F(x - \epsilon) - P\{|X_n - X| \geq \epsilon\} \leq F_n(x) \leq P\{|X_n - X| \geq \epsilon\} + F(x + \epsilon).$$

Since X_n converges in probability to X , we have $P\{|X_n - X| \geq \epsilon\} \rightarrow 0$ as $n \rightarrow \infty$. If F is continuous at x , then $F(x - \epsilon)$ and $F(x + \epsilon)$ approach $F(x)$ as $\epsilon \rightarrow 0$. Thus $F_n(x)$ is boxed between two quantities that can be made arbitrarily close to $F(x)$, so $F_n(x) \rightarrow F(x)$. ♣

7.5 Some Sufficient Conditions

In practice, $P\{|X_n - X| \geq \epsilon\}$ may be difficult to compute, and it is useful to have sufficient conditions for convergence in probability that can often be easily checked.

(1) If $E[(X_n - X)^2] \rightarrow 0$ as $n \rightarrow \infty$, then $X_n \xrightarrow{P} X$.

(2) If $E(X_n) \rightarrow E(X)$ and $\text{Var}(X_n - X) \rightarrow 0$, then $X_n \xrightarrow{P} X$.

Proof. The first statement follows from Chebyshev (b):

$$P\{|X_n - X| \geq \epsilon\} \leq \frac{E[(X_n - X)^2]}{\epsilon^2} \rightarrow 0.$$

To prove (2), note that

$$E[(X_n - X)^2] = \text{Var}(X_n - X) + [E(X_n) - E(X)]^2 \rightarrow 0. \clubsuit$$

In this result, if X is identically equal to a constant c , then $\text{Var}(X_n - X)$ is simply $\text{Var} X_n$. Condition (2) then becomes $E(X_n) \rightarrow c$ and $\text{Var} X_n \rightarrow 0$, which implies that X_n converges in probability to c .

7.6 An Application

In normal sampling, let S_n^2 be the sample variance based on n observations. Let's show that S_n^2 is a *consistent estimate* of the true variance σ^2 , that is, $S_n^2 \xrightarrow{P} \sigma^2$. Since nS_n^2/σ^2 is $\chi^2(n-1)$, we have $E(nS_n^2/\sigma^2) = (n-1)$ and $\text{Var}(nS_n^2/\sigma^2) = 2(n-1)$. Thus $E(S_n^2) = (n-1)\sigma^2/n \rightarrow \sigma^2$ and $\text{Var}(S_n^2) = 2(n-1)\sigma^4/n^2 \rightarrow 0$, and the result follows.

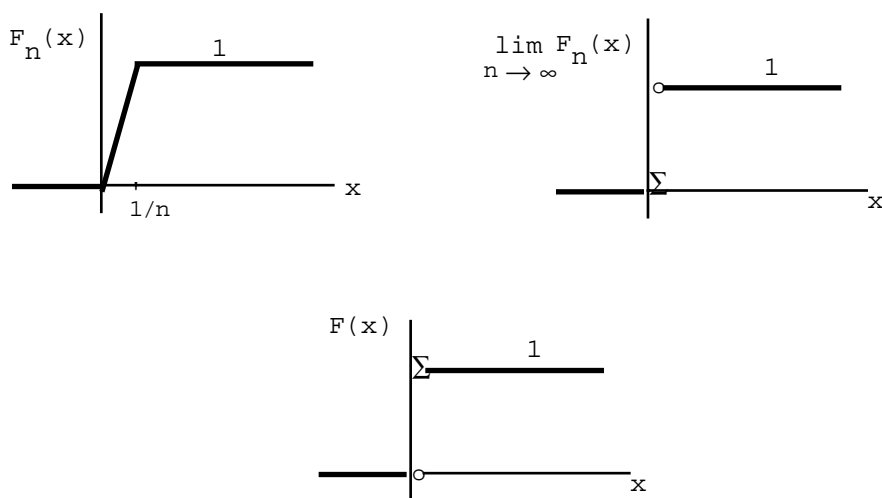


Figure 7.1

Problems

- Let X_1, \dots, X_n be independent, not necessarily identically distributed random variables. Assume that the X_i have finite means μ_i and finite variances σ_i^2 , and the variances are uniformly bounded, i.e., for some positive number M we have $\sigma_i^2 \leq M$ for all i . Show that $(S_n - E(S_n))/n$ converges in probability to 0. This is a generalization of the weak law of large numbers. For if $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$ for all i , then $E(S_n) = n\mu$, so $(S_n/n) - \mu \xrightarrow{P} 0$, i.e., $S_n/n \xrightarrow{P} \mu$.
- Toss an unbiased coin *once*. If heads, write down the sequence 10101010 \dots , and if tails, write down the sequence 01010101 \dots . If X_n is the n -th term of the sequence and $X = X_1$, show that X_n converges to X in distribution but not in probability.

3. Let X_1, \dots, X_n be iid with finite mean μ and finite variance σ^2 . Let \bar{X}_n be the sample mean $(X_1 + \dots + X_n)/n$. Find the *limiting distribution* of \bar{X}_n , i.e., find a random variable X such that $\bar{X}_n \xrightarrow{d} X$.
4. Let X_n be uniformly distributed between n and $n + 1$. Show that X_n does not have a limiting distribution. Intuitively, the probability has “run away” to infinity.

Lecture 8. The Central Limit Theorem

Intuitively, any random variable that can be regarded as the sum of a large number of small independent components is approximately normal. To formalize, we need the following result, stated without proof.

8.1 Theorem

If Y_n has moment-generating function M_n , Y has moment-generating function M , and $M_n(t) \rightarrow M(t)$ as $n \rightarrow \infty$ for all t in some open interval containing the origin, then $Y_n \xrightarrow{d} Y$.

8.2 Central Limit Theorem

Let X_1, X_2, \dots be iid, each with finite mean μ , finite variance σ^2 , and moment-generating function M . Then

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to a random variable that is normal $(0,1)$. Thus for large n , $\sum_{i=1}^n X_i$ is approximately normal.

We will give an informal sketch of the proof. The numerator of Y_n is $\sum_{i=1}^n (X_i - \mu)$, and the random variables $X_i - \mu$ are iid with mean 0 and variance σ^2 . Thus we may assume without loss of generality that $\mu = 0$. We have

$$M_{Y_n}(t) = E[e^{tY_n}] = E\left[\exp\left(\frac{t}{\sqrt{n}\sigma} \sum_{i=1}^n X_i\right)\right].$$

The moment-generating function of $\sum_{i=1}^n X_i$ is $[M(t)]^n$, so

$$M_{Y_n}(t) = \left[M\left(\frac{t}{\sqrt{n}\sigma}\right)\right]^n.$$

Now if the density of the X_i is $f(x)$, then

$$\begin{aligned} M\left(\frac{t}{\sqrt{n}\sigma}\right) &= \int_{-\infty}^{\infty} \exp\left(\frac{tx}{\sqrt{n}\sigma}\right) f(x) dx \\ &= \int_{-\infty}^{\infty} \left[1 + \frac{tx}{\sqrt{n}\sigma} + \frac{t^2 x^2}{2!n\sigma^2} + \frac{t^3 x^3}{3!n^{3/2}\sigma^3} + \dots\right] f(x) dx \\ &= 1 + 0 + \frac{t^2}{2n} + \frac{t^3 \mu_3}{6n^{3/2}\sigma^3} + \frac{t^4 \mu_4}{24n^2\sigma^4} + \dots \end{aligned}$$

where $\mu_k = E[(X_i)^k]$. If we neglect the terms after $t^2/2n$ we have, approximately,

$$M_{Y_n}(t) = \left(1 + \frac{t^2}{2n}\right)^n$$

which approaches the normal (0,1) moment-generating function $e^{t^2/2}$ as $n \rightarrow \infty$. This argument is very loose but it can be made precise by some estimates based on Taylor's formula with remainder.

We proved that if X_n converges in probability to X , then X_n convergence in distribution to X . There is a partial converse.

8.3 Theorem

If X_n converges in distribution to a *constant* c , then X_n converges in probability to X .

Proof. We estimate the probability that $|X_n - X| \geq \epsilon$, as follows.

$$\begin{aligned} P\{|X_n - X| \geq \epsilon\} &= P\{X_n \geq c + \epsilon\} + P\{X_n \leq c - \epsilon\} \\ &= 1 - P\{X_n < c + \epsilon\} + P\{X_n \leq c - \epsilon\} \end{aligned}$$

Now $P\{X_n \leq c + (\epsilon/2)\} \leq P\{X_n < c + \epsilon\}$, so

$$\begin{aligned} P\{|X_n - c| \geq \epsilon\} &\leq 1 - P\{X_n \leq c + (\epsilon/2)\} + P\{X_n \leq c - \epsilon\} \\ &= 1 - F_n(c + (\epsilon/2)) + F_n(c - \epsilon). \end{aligned}$$

where F_n is the distribution function of X_n . But as long as $x \neq c$, $F_n(x)$ converges to the distribution function of the constant c , so $F_n(x) \rightarrow 1$ if $x > c$, and $F_n(x) \rightarrow 0$ if $x < c$. Therefore $P\{|X_n - c| \geq \epsilon\} \rightarrow 1 - 1 + 0 = 0$ as $n \rightarrow \infty$. ♣

8.4 Remarks

If Y is binomial (n, p) , the *normal approximation to the binomial* allows us to regard Y as approximately normal with mean np and variance npq (with $q = 1 - p$). According to Box, Hunter and Hunter, "Statistics for Experimenters", page 130, the approximation works well in practice if $n > 5$ and

$$\frac{1}{\sqrt{n}} \left| \sqrt{\frac{q}{p}} - \sqrt{\frac{p}{q}} \right| < .3$$

If, for example, we wish to estimate the probability that $Y = 50$ or 51 or 52 , we may write this probability as $P\{49.5 < Y < 52.5\}$, and then evaluate as if Y were normal with mean np and variance $np(1 - p)$. This turns out to be slightly more accurate in practice than using $P\{50 \leq Y \leq 52\}$.

8.5 Simulation

Most computers can simulate a random variable that is uniformly distributed between 0 and 1. But what if we need a random variable with an arbitrary distribution function F ? For example, how would we simulate the random variable with the distribution function of Figure 8.1? The basic idea is illustrated in Figure 8.2. If $Y = F(X)$ where X has the

continuous distribution function F , then Y is uniformly distributed on $[0,1]$. (In Figure 8.2 we have, for $0 \leq y \leq 1$, $P\{Y \leq y\} = P\{X \leq x\} = F(x) = y$.)

Thus if X is uniformly distributed on $[0,1]$ and we want Y to have distribution function F , we set $X = F(Y)$, $Y = "F^{-1}(X)"$.

In Figure 8.1 we must be more precise:

Case 1. $0 \leq X \leq .3$. Let $X = (3/70)Y + (15/70)$, $Y = (70X - 15)/3$.

Case 2. $.3 \leq X \leq .8$. Let $Y = 4$, so $P\{Y = 4\} = .5$ as required.

Case 3. $.8 \leq X \leq 1$. Let $X = (1/10)Y + (4/10)$, $Y = 10X - 4$.

In Figure 8.1, replace the $F(y)$ -axis by an x -axis to visualize X versus Y . If $y = y_0$ corresponds to $x = x_0$ [i.e., $x_0 = F(y_0)$], then

$$P\{Y \leq y_0\} = P\{X \leq x_0\} = x_0 = F(y_0)$$

as desired.

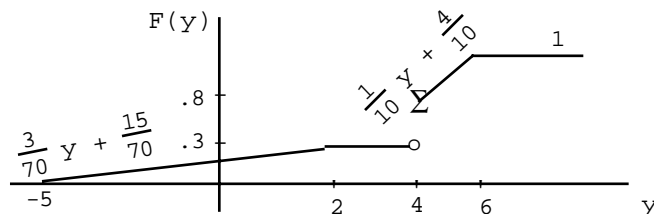


Figure 8.1

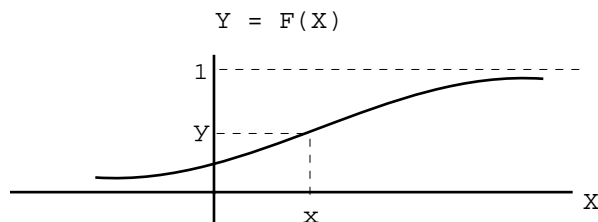


Figure 8.2

Problems

1. Let X_n be gamma (n, β) , i.e., X_n has the gamma distribution with parameters n and β . Show that X_n is a sum of n independent exponential random variables, and from this derive the limiting distribution of X_n/n .
2. Show that $\chi^2(n)$ is approximately normal for large n (with mean n and variance $2n$).

3. Let X_1, \dots, X_n be iid with density f . Let Y_n be the number of observations that fall into the interval (a, b) . Indicate how to use a normal approximation to calculate probabilities involving Y_n .
4. If we have 3 observations 6.45, 3.14, 4.93, and we round off to the nearest integer, we get 6, 3, 5. The sum of integers is 14, but the actual sum is 14.52. Let $X_i, i = 1, \dots, n$ be the round-off error of the i -th observation, and assume that the X_i are iid and uniformly distributed on $(-1/2, 1/2)$. Indicate how to use a normal approximation to calculate probabilities involving the total round off error $Y_n = \sum_{i=1}^n X_i$.
5. Let X_1, \dots, X_n be iid with continuous distribution function F , and let $Y_1 < \dots < Y_n$ be the order statistics of the X_i . Then $F(X_1), \dots, F(X_n)$ are iid and uniformly distributed on $[0, 1]$ (see the discussion of simulation), with order statistics $F(Y_1), \dots, F(Y_n)$. Show that $n(1 - F(Y_n))$ converges in distribution to an exponential random variable.

Lecture 9. Estimation

9.1 Introduction

In effect the statistician plays a game against nature, who first chooses the “state of nature” θ (a number or k -tuple of numbers in the usual case) and performs a random experiment. We do not know θ but we are allowed to observe the value of a random variable (or random vector) X , called the *observable*, with density $f_\theta(x)$.

After observing $X = x$ we estimate θ by $\delta(x)$, which is called a *point estimate* because it produces a single number which we hope is close to θ . The main alternative is an *interval estimate* or *confidence interval*, which will be discussed in Lectures 10 and 11.

For a point estimate $\delta(x)$ to make sense physically, it must depend *only on x , not on the unknown parameter θ* . There are many possible estimates, and there are no general rules for choosing a best estimate. Some practical considerations are:

- (a) How much does it cost to collect the data?
- (b) Is the performance of the estimate easy to measure, for example, can we compute $P\{|\delta(x) - \theta| < \epsilon\}$?
- (c) Are the advantages of the estimate appropriate for the problem at hand?

We will study several estimation methods:

1. Maximum likelihood estimates.

These estimates usually have highly desirable theoretical properties (consistency), and are frequently not difficult to compute.

2. Confidence intervals.

These estimates have a very useful practical feature. We construct an interval from the data, and we will know the probability that our (random) interval actually contains the unknown (but fixed) parameter.

3. Uniformly minimum variance unbiased estimates (UMVUE's).

Mathematical theory generates a large number of examples of these, but as we know, a biased estimate can sometimes be superior.

4. Bayes estimates.

These estimates are appropriate if it is reasonable to assume that the state of nature θ is a random variable with a known density.

In general, statistical theory produces many reasonable candidates, and practical experience will dictate the choice in a given physical situation.

9.2 Maximum Likelihood Estimates

We choose $\delta(x) = \hat{\theta}$, a value of θ that makes what we have observed as likely as possible. In other words, let $\hat{\theta}$ maximize the *likelihood function* $L(\theta) = f_\theta(x)$, with x fixed. This corresponds to basic statistical philosophy; if what we have observed is more likely under θ_2 than under θ_1 , we prefer θ_2 to θ_1 .

9.3 Example

Let X be binomial (n, θ) . Then the probability that $X = x$ when the true parameter is θ is

$$f_{\theta}(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, x = 0, 1, \dots, n.$$

Maximizing $f_{\theta}(x)$ is equivalent to maximizing $\ln f_{\theta}(x)$:

$$\frac{\partial}{\partial \theta} \ln f_{\theta}(x) = \frac{\partial}{\partial \theta} [x \ln \theta + (n - x) \ln(1 - \theta)] = \frac{x}{\theta} - \frac{n - x}{1 - \theta} = 0.$$

Thus $x - \theta x - \theta n + \theta x = 0$, so $\hat{\theta} = X/n$, the relative frequency of success.

Notation: $\hat{\theta}$ will be written in terms of random variables, in this case X/n rather than x/n . Thus $\hat{\theta}$ is itself a random variable.

We have $E(\hat{\theta}) = n\theta/n = \theta$, so $\hat{\theta}$ is *unbiased*. By the weak law of large numbers, $\hat{\theta} \xrightarrow{P} \theta$, i.e., $\hat{\theta}$ is *consistent*.

9.4 Example

Let X_1, \dots, X_n be iid, normal (μ, σ^2) , $\theta = (\mu, \sigma^2)$. Then, with $x = (x_1, \dots, x_n)$,

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

and

$$\ln f_{\theta}(x) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2;$$

$$\frac{\partial}{\partial \mu} \ln f_{\theta}(x) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0, \quad \sum_{i=1}^n x_i - n\mu = 0, \quad \mu = \bar{x};$$

$$\frac{\partial}{\partial \sigma} \ln f_{\theta}(x) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = \frac{n}{\sigma^3} \left[-\sigma^2 + \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \right] = 0$$

with $\mu = \bar{x}$. Thus

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2.$$

Case 1. μ and σ are both unknown. Then $\hat{\theta} = (\bar{X}, S^2)$.

Case 2. σ^2 is known. Then $\theta = \mu$ and $\hat{\theta} = \bar{X}$ as above. (Differentiation with respect to σ is omitted.)

Case 3. μ is known. Then $\theta = \sigma^2$ and the equation $(\partial/\partial\sigma) \ln f_\theta(x) = 0$ becomes

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

so

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

The sample mean \bar{X} is an unbiased and (by the weak law of large numbers) consistent estimate of μ . The sample variance S^2 is a biased but consistent estimate of σ^2 (see Lectures 4 and 7).

Notation: We will abbreviate maximum likelihood estimate by MLE.

9.5 The MLE of a Function of Theta

Suppose that for a fixed x , $f_\theta(x)$ is a maximum when $\theta = \theta_0$. Then the value of θ^2 when $f_\theta(x)$ is a maximum is θ_0^2 . Thus to get the MLE of θ^2 , we simply square the MLE of θ . In general, if h is any function, then

$$\widehat{h(\theta)} = h(\hat{\theta}).$$

If h is continuous, then consistency is preserved, in other words:

If h is continuous and $\hat{\theta} \xrightarrow{P} \theta$, then $h(\hat{\theta}) \xrightarrow{P} h(\theta)$.

Proof. Given $\epsilon > 0$, there exists $\delta > 0$ such that if $|\hat{\theta} - \theta| < \delta$, then $|h(\hat{\theta}) - h(\theta)| < \epsilon$. Consequently,

$$P\{|h(\hat{\theta}) - h(\theta)| \geq \epsilon\} \leq P\{|\hat{\theta} - \theta| \geq \delta\} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad \clubsuit$$

(To justify the above inequality, note that if the occurrence of an event A implies the occurrence of an event B , then $P(A) \leq P(B)$.)

9.6 The Method of Moments

This is sometimes a quick way to obtain reasonable estimates. We set the observed k -th moment $n^{-1} \sum_{i=1}^n x_i^k$ equal to the theoretical k -th moment $E(X_i^k)$ (which will depend on the unknown parameter θ). Or we set the observed k -th central moment $n^{-1} \sum_{i=1}^n (x_i - \mu)^k$ equal to the theoretical k -th central moment $E[(X_i - \mu)^k]$. For example, let X_1, \dots, X_n be iid, gamma with $\alpha = \theta_1, \beta = \theta_2$, with $\theta_1, \theta_2 > 0$. Then $E(X_i) = \alpha\beta = \theta_1\theta_2$ and $\text{Var } X_i = \alpha\beta^2 = \theta_1\theta_2^2$ (see Lecture 3). We set

$$\bar{X} = \theta_1\theta_2, \quad S^2 = \theta_1\theta_2^2$$

and solve to get estimates θ_i^* of $\theta_i, i = 1, 2$, namely

$$\theta_2^* = \frac{S^2}{\bar{X}}, \quad \theta_1^* = \frac{\bar{X}}{\theta_2^*} = \frac{\bar{X}^2}{S^2}$$

Problems

1. In this problem, X_1, \dots, X_n are iid with density $f_\theta(x)$ or probability function $p_\theta(x)$, and you are asked to find the MLE of θ .
 - (a) Poisson (θ), $\theta > 0$.
 - (b) $f_\theta(x) = \theta x^{\theta-1}$, $0 < x < 1$, where $\theta > 0$. The probability is concentrated near the origin when $\theta < 1$, and near 1 when $\theta > 1$.
 - (c) Exponential with parameter θ , i.e., $f_\theta(x) = (1/\theta)e^{-x/\theta}$, $x > 0$, where $\theta > 0$.
 - (d) $f_\theta(x) = (1/2)e^{-|x-\theta|}$, where θ and x are arbitrary real numbers.
 - (e) Translated exponential, i.e., $f_\theta(x) = e^{-(x-\theta)}$, where θ is an arbitrary real number and $x \geq \theta$.
2. let X_1, \dots, X_n be iid, each uniformly distributed between $\theta - (1/2)$ and $\theta + (1/2)$. Find more than one MLE of θ (so MLE's are not necessarily unique).
3. In each part of Problem 1, calculate $E(X_i)$ and derive an estimate based on the method of moments by setting the sample mean equal to the true mean. In each case, show that the estimate is consistent.
4. Let X be exponential with parameter θ , as in Problem 1(c). If $r > 0$, find the MLE of $P\{X \leq r\}$.
5. If X is binomial (n, θ) and a and b are integers with $0 \leq a \leq b \leq n$, find the MLE of $P\{a \leq X \leq b\}$.

Lecture 10. Confidence Intervals

10.1 Predicting an Election

There are two candidates A and B . If a voter is selected at random, the probability that the voter favors A is p , where p is fixed but *unknown*. We select n voters independently and ask their preference.

The number Y_n of A voters is binomial (n, p) , which (for sufficiently large n), is approximately normal with $\mu = np$ and $\sigma^2 = np(1-p)$. The relative frequency of A voters is Y_n/n . We wish to estimate the minimum value of n such that we can predict A 's percentage of the vote within 1 percent, with 95 percent confidence. Thus we want

$$P\left\{\left|\frac{Y_n}{n} - p\right| < .01\right\} > .95.$$

Note that $|(Y_n/n) - p| < .01$ means that p is within .01 of Y_n/n . So this inequality can be written as

$$\frac{Y_n}{n} - .01 < p < \frac{Y_n}{n} + .01.$$

Thus the probability that the *random* interval $I_n = ((Y_n/n) - .01, (Y_n/n) + .01)$ contains the true probability p is greater than .95. We say that I_n is a *95 percent confidence interval* for p .

In general, we find confidence intervals by calculating or estimating the probability of the event that is to occur with the desired level of confidence. In this case,

$$P\left\{\left|\frac{Y_n}{n} - p\right| < .01\right\} = P\{|Y_n - np| < .01n\} = P\left\{\left|\frac{Y_n - np}{\sqrt{np(1-p)}}\right| < \frac{.01\sqrt{n}}{\sqrt{p(1-p)}}\right\}$$

and this is approximately

$$\Phi\left(\frac{.01\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(\frac{-.01\sqrt{n}}{\sqrt{p(1-p)}}\right) = 2\Phi\left(\frac{.01\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 > .95$$

where Φ is the normal (0,1) distribution function. Since $1.95/2 = .975$ and $\Phi(1.96) = .975$, we have

$$\frac{.01\sqrt{n}}{\sqrt{p(1-p)}} > 1.96, \quad n > (196)^2 p(1-p).$$

But (by calculus) $p(1-p)$ is maximized when $1-2p=0$, $p=1/2$, $p(1-p)=1/4$. Thus $n > (196)^2/4 = (98)^2 = (100-2)^2 = 10000 - 400 + 4 = 9604$.

If we want to get within one tenth of one percent (.001) of p with 99 percent confidence, we repeat the above analysis with .01 replaced by .001, $1.99/2 = .995$ and $\Phi(2.6) = .995$. Thus

$$\frac{.001\sqrt{n}}{\sqrt{p(1-p)}} > 2.6, \quad n > (2600)^2/4 = (1300)^2 = 1,690,000.$$

To get within 3 percent with 95 percent confidence, we have

$$\frac{.03\sqrt{n}}{\sqrt{p(1-p)}} > 1.96, \quad n > \left(\frac{196}{3}\right)^2 \times \frac{1}{4} = 1067.$$

If the experiment is repeated independently a large number of times, it is very likely that our result will be within .03 of the true probability p at least 95 percent of the time. The usual statement “The margin of error of this poll is $\pm 3\%$ ” does not capture this idea.

Note that the accuracy of the prediction depends only on the *number of voters polled* and not in total number of votes in the population. But the model assumes sampling with replacement. (Theoretically, the same voter can be polled more than once since the voters are selected independently.) In practice, sampling is done without replacement, but if the number n of voters polled is small relative to the population size N , the error is very small.

The normal approximation to the binomial (based on the central limit theorem) is quite reliable, and is used in practice even for modest values of n ; see (8.4).

10.2 Estimating the Mean of a Normal Population

Let X_1, \dots, X_n be iid, each normal (μ, σ^2) . We will find a confidence interval for μ .

Case 1. The variance σ^2 is known. Then \bar{X} is normal $(\mu, \sigma^2/n)$, so

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ is normal } (0,1),$$

hence

$$P\{-b < \sqrt{n} \left(\frac{\bar{X} - \mu}{\sigma}\right) < b\} = \Phi(b) - \Phi(-b) = 2\Phi(b) - 1$$

and the inequality defining the confidence interval can be written as

$$\bar{X} - \frac{b\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{b\sigma}{\sqrt{n}}.$$

We choose a symmetrical interval to minimize the length, because the normal density with zero mean is symmetric about 0. The desired confidence level determines b , which then determines the confidence interval.

Case 2. The variance σ^2 is unknown. Recall from (5.1) that

$$\frac{\bar{X} - \mu}{S/\sqrt{n-1}} \text{ is } T(n-1)$$

hence

$$P\{-b < \frac{\bar{X} - \mu}{S/\sqrt{n-1}} < b\} = 2F_T(b) - 1$$

and the inequality defining the confidence interval can be written as

$$\bar{X} - \frac{bS}{\sqrt{n-1}} < \mu < \bar{X} + \frac{bS}{\sqrt{n-1}}.$$

10.3 A Correction Factor When Sampling Without Replacement

The following results will not be used and may be omitted, but it is interesting to measure quantitatively the effect of sampling without replacement. In the election prediction problem, let X_i be the indicator of success (i.e., selecting an A voter) on trial i . Then $P\{X_i = 1\} = p$ and $P\{X_i = 0\} = 1 - p$. If sampling is done with replacement, then the X_i are independent and the total number $X = X_1 + \cdots + X_n$ of A voters in the sample is binomial (n, p) . Thus the variance of X is $np(1-p)$. However, if sampling is done without replacement, then in effect we are drawing n balls from an urn containing N balls (where N is the size of the population), with Np balls labeled A and $N(1-p)$ labeled B . Recall from basic probability theory that

$$\text{Var } X = \sum_{i=1}^n \text{Var } X_i + 2 \sum_{i < j} \text{Cov}(X_i, X_j)$$

where Cov stands for covariance. (We will prove this in a later lecture.) If $i \neq j$, then

$$E(X_i X_j) = P\{X_i = X_j = 1\} = P\{X_1 X_2 = 1\} = \frac{Np}{N} \times \frac{Np-1}{N-1}$$

and

$$\text{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) = p \left(\frac{Np-1}{N-1} \right) - p^2$$

which reduces to $-p(1-p)/(N-1)$. Now $\text{Var } X_i = p(1-p)$, so

$$\text{Var } X = np(1-p) - 2 \binom{n}{2} \frac{p(1-p)}{N-1}$$

which reduces to

$$np(1-p) \left[1 - \frac{n-1}{N-1} \right] = np(1-p) \left[\frac{N-n}{N-1} \right].$$

Thus if SE is the *standard error* (the standard deviation of X), then SE (without replacement) = SE (with replacement) times a correction factor, where the correction factor is

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{1-(n/N)}{1-(1/N)}}.$$

The correction factor is less than 1, and approaches 1 as $N \rightarrow \infty$, as long as $n/N \rightarrow 0$.

Note also that in sampling without replacement, the probability of getting exactly k A 's in n trials is

$$\frac{\binom{Np}{k} \binom{N(1-p)}{n-k}}{\binom{N}{n}}$$

with the standard pattern $Np + N(1-p) = N$ and $k + (n-k) = n$.

Problems

1. In the normal case [see (10.2)], assume that σ^2 is known. Explain how to compute the length of the confidence interval for μ .
2. Continuing Problem 1, assume that σ^2 is unknown. Explain how to compute the length of the confidence interval for μ , in terms of the sample standard deviation S .
3. Continuing Problem 2, explain how to compute the expected length of the confidence interval for μ , in terms of the unknown standard deviation σ . (Note that when σ is unknown, we expect a larger interval since we have less information.)
4. Let X_1, \dots, X_n be iid, each gamma with parameters α and β . If α is known, explain how to compute a confidence interval for the mean $\mu = \alpha\beta$.
5. In the binomial case [see (10.1)], suppose we specify the level of confidence and the length of the confidence interval. Explain how to compute the minimum value of n .