Although the Laplace distribution is a mathematical rarity, it nicely illustrates a very practical point: If a population has thick tails, so that outlying observations are likely to occur, then the sample mean has larger variance—because it takes into account all the observations, even the distant outliers that the sample median ignores. In Chapter 16 we will pursue this issue further.

PROBLEMS

7-1 Assuming as usual that samples are random, answer True or False; if False, correct it.

a. Samples are used for making inferences about the population from which they are drawn.

b. μ is a random variable (varying from sample to sample), and is an unbiased estimator of the parameter $\overline{X}$.

c. If we double the sample size, we halve the standard error of $\overline{X}$, and consequently double its accuracy in estimating the population mean.

d. The sample proportion P is an unbiased estimator of the population proportion π.

7-2 Based on a random sample of 2 observations, consider two competing estimators of the population mean μ:

$$\overline{X} \equiv \frac{1}{2}X_1 + \frac{1}{2}X_2$$

$$\text{and } U \equiv \frac{1}{3}X_1 + \frac{2}{3}X_2$$

a. Are they unbiased?

b. Which estimator is more efficient? How much more efficient?

7-3 An economist gathers a random sample of 500 observations, and loses the records of the last 180. This leaves only 320 observations from which to calculate the sample mean. What is the efficiency of this, relative to what could have been obtained from the whole sample?

7-4 What is the efficiency of the sample median relative to the sample mean in estimating the center of a normal population? [Hint: Recall from (7-6) that the efficiency of the mean relative to the median was 157%.]

7-5 **a.** Answer True or False; if False, correct it.

In both Problems 7-3 and 7-4 we have examples of estimates that are only 64% efficient. In Problem 7-3, this inefficiency was

obvious, because 36% of the observations were lost in calculating $\overline{X}$. In Problem 7-4, the inefficiency was more subtle, because it was caused merely by using the sample median instead of the sample mean. However, in terms of results—producing an estimate with more variance than necessary—both inefficiencies are equally damaging.

b. In view of part **a**, what advice would you give to a researcher who spends \$100,000 collecting data, and \$100 analyzing it?

7-6 Suppose that a surveyor is trying to determine the area of a rectangular field, in which the measured length X and the measured width Y are independent random variables that fluctuate widely about the true values, according to the following probability distributions:

x	$p(x)$
8	1/4
10	1/4
11	1/2

y	$p(y)$
4	1/2
6	1/2

The calculated area $A = XY$ of course is a random variable, and is used to estimate the true area. If the true length and width are 10 and 5, respectively,

a. Is X an unbiased estimator of the true length?

b. Is Y an unbiased estimator of the true width?

c. Is A an unbiased estimator of the true area? (Hint: see Problem 5-38)

7-7 **a.** To guide long-term planning, an automobile executive commissioned two independent sample surveys to estimate the proportion π of car owners who intend to buy a smaller car next time. The first survey showed a proportion $P_1 = 60/200 = 30\%$. The second and larger survey showed a proportion $P_2 = 240/1000 = 24\%$. To get an overall estimate, the simple average $P^* = 27\%$ was taken. What is the variance of this estimate? [Hint: You may assume simple random sampling, so that var $P \simeq P(1 - P)/n$]

b. The first poll is clearly less reliable than the second. So it was proposed to just throw the first away, and use the estimate $P_2 = 24\%$. What is the variance of this estimate? What then is its efficiency relative to P^*?

c. The best estimate of all, of course, would count each *observation* equally (not each *sample* equally). That is, take the overall proportion in favor, $P = (60 + 240)/(200 + 1000) = 25\%$. What is the variance of this estimate? Then what is its efficiency relative to P^*?

d. True or False? If False, correct it:
It is important to know the reliability of your sources. For exam-

7-9 A large chain of shops specializing in tuneups has to choose one of four gauges to measure the gap in a spark plug. When tested, each gauge showed a slight error (in hundredths of mm.):

Gauge	*A*	*B*	*C*	*D*
bias	none	−10	5	2
standard dev.	10	none	5	8

Which gauge has the smallest MSE (greatest accuracy)?

7-10 A market survey of young business executives was undertaken to determine what sort of computer would suit a combination of their professional and personal needs. Since those with more children were thought to be more likely to buy a home computer, one of the questions each executive was asked was, "How many children do you have?"

Unfortunately, those with more children tend to have less time and inclination to reply to the survey, as the following table shows:

	Total Population (Target)		*Subpopulation Who Would Respond*	
x = Number of Children Over 5 Years Old	*Frequency f*	*Rel. Frequency f/N*	*Frequency f*	*Rel. Frequency f/N*
0	20,000	.40	6,200	.62
1	12,000	.24	2,100	.21
2	10,000	.20	1,200	.12
3	6,000	.12	400	.04
4	2,000	.04	100	.01
	$N = 50{,}000$	1.00	$N = 10{,}000$	1.00

Two types of sample survey were proposed:

i. *High volume*, with 1000 executives sampled, and with no follow up. Their overall response rate would be 10,000/50,000 = 20% as given by the table, yielding 200 replies.

ii. *High quality*, with 25 executives sampled, and enough follow-up to get a 100% response rate.

a. Calculate the mean number of children in the population μ.

b. In estimating μ, does either survey have a sample mean $\overline{X}$ that is unbiased?

c. Which survey has the smallest MSE (greatest accuracy)?

***7-11** In Problem 7-10, note how the response rate of executives drops as the number of children X increases. For example, when $X = 0$, the response rate is 6200/20,000 = 31%, while for $X = 4$, the response

b. Similarly, we may write:

$$s_*^2 = \left(\frac{n-1}{n+1}\right) s^2 = \left(1 - \frac{2}{n+1}\right) s^2$$

And since $2/(n+1)$ tends to zero, this is also asymptotically unbiased.

REMARKS

We have shown that both MSD and s_*^2 are asymptotically unbiased. And s^2 itself is unbiased for *any* sample size n. It could further be shown that all three estimators have variance that approaches zero, so that they are all consistent.

Which of the three estimators should we use? Since all three are consistent, we need a stronger criterion to make a final choice, such as efficiency. For many populations, including the normal, it turns out that s_*^2 is most efficient.

C—CONCLUSIONS

Although consistency has an abstract definition, it often provides a useful preliminary criterion for sorting out estimators.

Nevertheless, to finally sort out the best estimator, a stronger criterion such as efficiency is required—as we saw in Example 7-5. Another familiar example will illustrate: In estimating the center of a normal population, both the sample mean and median satisfy the consistency criterion. To choose between them, efficiency is the criterion that will finally select the winner (the sample mean).

*PROBLEMS

7-12 The population of American personal incomes is skewed to the right (as we saw in Figure 2-5, for men in 1975, for example). Which of the following will be consistent estimators of the population mean μ?

a. From a random sample of incomes, the sample mean? The sample median? The sample mode?

b. Repeat part **a**, for a sample of incomes drawn at random from the cities over one million.

7-13 When S successes occur in n trials, the sample proportion $P = S/n$ customarily is used as an estimator of the probability of success π. However, sometimes there are good reasons to use the estimator

$P^* \equiv (S + 1)/(n + 2)$. Alternatively, P^* can be written as a linear combination of the familiar estimator P:

$$P^* = \frac{nP + 1}{n + 2} = \left(\frac{n}{n + 2}\right) P + \left(\frac{1}{n + 2}\right)$$

a. What is the MSE of P? Is it consistent?

b. What is the MSE of P^*? Is it consistent? (Hint: Calculate the mean and variance of P^*, in terms of the familiar mean and variance of P.)

c. To decide which estimator is better, P or P^*, does consistency help? What criterion *would* help?

d. Tabulate the efficiency of P^* relative to P, for example when $n = 10$ and $\pi = 0, .1, .2, \ldots, .9, 1.0$.

e. State some possible circumstances when you might prefer to use P^* instead of P to estimate π.

CHAPTER 7 SUMMARY

7-1 Statistics such as $\overline{X}$ from random samples (colored blue) are used to estimate parameters such as μ from populations (gray).

7-2 An estimator is called unbiased if, on average, it is exactly on target. An unbiased estimator is called efficient if it has the smallest variance.

7-3 For estimators with bias as well as variance, minimum MSE (mean squared error) is the appropriate measure of efficiency. MSE remains disappointingly high for estimators with persistent bias, such as nonresponse bias.

***7-4** A consistent estimator is one that eventually is on target. (Not only on target on average, but the whole sampling distribution gets squeezed onto the target, as the sample size n increases infinitely.)

REVIEW PROBLEMS

7-14 An estimator that has small variance (but may be biased) is called *precise*. An estimator that has small MSE is called *accurate*. To illustrate: A standard 100-gm mass was weighed many many times on a scale A, and the distribution of measurements is graphed below. A similar distribution was obtained for scale B, and finally for scale C.

ii. $W_2 = \frac{3}{4}U + \frac{1}{4}V$ (weighted average)

iii. $W_3 = 1U + 0V$ (drop the less accurate estimate)

a. Which are unbiased?

b. Intuitively, which would you guess is the best estimator? The worst?

c. Check out your guess in part **b** by making the appropriate calculations.

***d.** Intuitively, W_2 works well because it gives only ⅓ as much weight to the component (V) that has 3 times the standard deviation.

Is it possible to do even better than W_2? Suggest some possibilities, and then check them out.

7-17 A processor of sheet metal produces a large number of square plates, whose size must be cut within a specified tolerance. To measure the final product, a slightly worn gauge is used: Its measurement error is normally distributed with a mean $\mu = 0$ and standard deviation $\sigma = .10$ inch. To improve the accuracy, and to protect against blunders, two independent measurements of a plate's length are taken with this gauge, say X_1 and X_2. To find the area of a plate, the quality control manager is in a dilemma:

i. Should he square first, and then average:

$$\frac{X_1^2 + X_2^2}{2}$$

ii. Should he average first, and then square:

$$\left(\frac{X_1 + X_2}{2}\right)^2$$

a. Are methods **i** and **ii** really different, or are they just two different ways of saying the same thing? (Hint: Try a simulation. Suppose, for example, the two measured lengths are $X_1 = 5.9$ and $X_2 = 6.1$.)

b. Which has less bias? [Hint: See equation (4-36).]

c. As an alternative estimator of the area, what is the bias of X_1X_2? (Hint: See Problem 5-38.)

***7-18** A free-trade agreement has opened up a new market of 50 million potential customers for personal computers, and a market survey of these customers is being planned. People with higher incomes are more likely to buy a computer within the next 6 months, and also more likely to respond to a phone survey, as the following table shows:

Income Level	Proportion Who Will Buy	Total Population (Target) Frequency f (millions)	Subpopulation Who Would Respond Frequency f (millions)
$0–20,000	2%	40	7
20–40,000	4%	5	1
40–80,000	10%	3	1
over 80,000	20%	2	1
		N = 50	N = 10

a. In the 50 million population, how many will buy a computer? Answer as a total figure, and then as a percentage.

b. A market survey of 1000 random phone calls would bring about how many replies? Among these replies, the percentage P who will buy is a natural estimator of the population percentage in **a**. What is the bias and MSE of P?

c. A smaller survey was also considered, with just 100 calls but enough follow-up to get a 100% response. What is the bias and MSE of the resulting estimator P^*?

7-19 To interpret MSE concretely, we could take its square root to get the "typical" error (more precisely, the Root-Mean-Square or RMS error—just like we took the square root of the variance to get the standard deviation).

In Problem 7-18, calculate this RMS error:

a. For P and P^*, the two competing estimators of the percentage of the population who will buy.

b. For the two corresponding estimators of the *total* number in the population who will buy (the "market size").

7-20 ***A Final Challenge: How Much Follow-Up Should a Survey Use?***

A market survey was being planned to estimate the number of drug circulars physicians have read in the past seven days. Physicians who read more were also more likely to respond to the survey, as the following table shows:

X = Number of Circulars Read	Whole Target Population Frequency f	Subpopulation Responding to First Contact Frequency f	Subpopulation Responding to First or Second Contact Frequency f
0	40,000	2,000	14,000
1	5,000	1,000	2,000
2	5,000	2,000	4,000
	N = 50,000	N = 5,000	N = 20,000