Exercises

20. Prove Theorem 9.

21. Prove Theorem 10.

22. Verify the identity

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

which we used in the proof of Theorem 11.

23. Use Theorem 11 to show that, for random samples of size n from a normal population with the variance σ^2, the sampling distribution of S^2 has the mean σ^2 and the variance $\frac{2\sigma^4}{n-1}$. (A general formula for the variance of S^2 for random samples from any population with finite second and fourth moments may be found in the book by H. Cramér listed among the references at the end of this chapter.)

24. Show that if $X_1, X_2, \ldots, X_n$ are independent random variables having the chi-square distribution with $\nu = 1$ and $Y_n = X_1 + X_2 + \cdots + X_n$, then the limiting distribution of

$$Z=\frac{\frac{Y_n}{n}-1}{\sqrt{2/n}}$$

as $n\to\infty$ is the standard normal distribution.

25. Based on the result of Exercise 24, show that if X is a random variable having a chi-square distribution with ν degrees of freedom and ν is large, the distribution of $\frac{X-\nu}{\sqrt{2\nu}}$ can be approximated with the standard normal distribution.

26. Use the method of Exercise 25 to find the approximate value of the probability that a random variable having a chi-square distribution with $\nu = 50$ will take on a value greater than 68.0.

27. If the range of X is the set of all positive real numbers, show that for $k > 0$ the probability that $\sqrt{2X}-\sqrt{2\nu}$ will take on a value less than k equals the probability that $\frac{X-\nu}{\sqrt{2\nu}}$ will take on a value less than $k+\frac{k^2}{2\sqrt{2\nu}}$.

28. Use the results of Exercises 25 and 27 to show that if X has a chi-square distribution with ν degrees of freedom, then for large ν the distribution of $\sqrt{2X}-\sqrt{2\nu}$ can be approximated with the standard normal distribution. Also, use this method of approximation to rework Exercise 26.

29. Find the percentage errors of the approximations of Exercises 26 and 28, given that the actual value of the probability (rounded to five decimals) is 0.04596.

30. (*Proof of the independence of* $\overline{X}$ *and* S^2 *for* $n = 2$) If X_1 and X_2 are independent random variables having the standard normal distribution, show that

(a) the joint density of X_1 and $\overline{X}$ is given by

$$f(x_1,\overline{x})=\frac{1}{\pi}\cdot e^{-\overline{x}^2}e^{-(x_1-\overline{x})^2}$$

for $-\infty<x_1<\infty$ and $-\infty<\overline{x}<\infty$;

(b) the joint density of $U = |X_1-\overline{X}|$ and $\overline{X}$ is given by

$$g(u,\overline{x})=\frac{2}{\pi}\cdot e^{-(\overline{x}^2+u^2)}$$

for $u > 0$ and $-\infty<\overline{x}<\infty$, since $f(x_1,\overline{x})$ is symmetrical about $\overline{x}$ for fixed $\overline{x}$;

(c) $S^2 = 2(X_1-\overline{X})^2 = 2U^2$;

(d) the joint density of $\overline{X}$ and S^2 is given by

$$h(s^2,\overline{x})=\frac{1}{\sqrt{\pi}}e^{-\overline{x}^2}\cdot\frac{1}{\sqrt{2\pi}}(s^2)^{-\frac{1}{2}}e^{-\frac{1}{2}s^2}$$

for $s^2 > 0$ and $-\infty<\overline{x}<\infty$, demonstrating that $\overline{X}$ and S^2 are independent.

31. (*Proof of the independence of* $\overline{X}$ *and* S^2) If $X_1, X_2, \ldots, X_n$ constitute a random sample from a normal population with the mean μ and the variance σ^2,

(a) find the conditional density of X_1 given $X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n$, and then set $X_1 = n\overline{X}-X_2-\cdots-X_n$ and use the transformation technique to find the conditional density of $\overline{X}$ given $X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n$;

(b) find the joint density of $\overline{X}, X_2, X_3, \ldots, X_n$ by multiplying the conditional density of $\overline{X}$ obtained in part (a) by the joint density of $X_2, X_3, \ldots, X_n$, and show that

$$g(x_2,x_3,\ldots,x_n|\overline{x})=\sqrt{n}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n-1}e^{-\frac{(n-1)s^2}{2\sigma^2}}$$

for $-\infty<x_i<\infty, i = 2, 3, \ldots, n$;

(c) show that the conditional moment-generating function of $\frac{(n-1)S^2}{\sigma^2}$ given $\overline{X}=\overline{x}$ is

$$E\left[e^{\frac{(n-1)S^2}{\sigma^2}\cdot t}\middle|\overline{x}\right]=(1-2t)^{-\frac{n-1}{2}}\qquad\text{for } t<\frac{1}{2}$$

Since this result is free of $\overline{x}$, it follows that $\overline{X}$ and S^2 are independent; it also shows that $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $n-1$ degrees of freedom.

This proof, due to J. Shuster, is listed among the references at the end of this chapter.

32. This question has been intentionally omitted for this edition.

33. Show that for $\nu > 2$ the variance of the t distribution with ν degrees of freedom is $\dfrac{\nu}{\nu-2}$. (*Hint*: Make the substitution $1+\dfrac{t^2}{\nu}=\dfrac{1}{u}$.)

34. Show that for the t distribution with $\nu > 4$ degrees of freedom

(a) $\mu_4 = \dfrac{3\nu^2}{(\nu-2)(\nu-4)}$;

(b) $\alpha_4 = 3 + \dfrac{6}{\nu-4}$.

(*Hint*: Make the substitution $1+\dfrac{t^2}{\nu}=\dfrac{1}{u}$.)

35. This question has been intentionally omitted for this edition.

36. By what name did we refer to the t distribution with $\nu = 1$ degree of freedom?

37. This question has been intentionally omitted for this edition.

38. Show that for $\nu_2 > 2$ the mean of the F distribution is $\dfrac{\nu_2}{\nu_2-2}$, making use of the definition of F in Theorem 14 and the fact that for a random variable V having the chi-square distribution with ν_2 degrees of freedom, $E\left(\dfrac{1}{V}\right) = \dfrac{1}{\nu_2-2}$.

39. Verify that if X has an F distribution with ν_1 and ν_2 degrees of freedom and $\nu_2 \to \infty$, the distribution of $Y = \nu_1 X$ approaches the chi-square distribution with ν_1 degrees of freedom.

40. Verify that if T has a t distribution with ν degrees of freedom, then $X = T^2$ has an F distribution with $\nu_1 = 1$ and $\nu_2 = \nu$ degrees of freedom.

41. If X has an F distribution with ν_1 and ν_2 degrees of freedom, show that $Y = \dfrac{1}{X}$ has an F distribution with ν_2 and ν_1 degrees of freedom.

42. Use the result of Exercise 41 to show that

$$f_{1-\alpha,\nu_1,\nu_2} = \frac{1}{f_{\alpha,\nu_2,\nu_1}}$$

43. Verify that if Y has a beta distribution with $\alpha = \dfrac{\nu_1}{2}$ and $\beta = \dfrac{\nu_2}{2}$, then

$$X = \frac{\nu_2 Y}{\nu_1(1-Y)}$$

has an F distribution with ν_1 and ν_2 degrees of freedom.

44. Show that the F distribution with 4 and 4 degrees of freedom is given by

$$g(f) = \begin{cases} 6f(1+f)^{-4} & \text{for } f > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and use this density to find the probability that for independent random samples of size $n = 5$ from normal populations with the same variance, S_1^2/S_2^2 will take on a value less than $\frac{1}{2}$ or greater than 2.

7 Order Statistics

The sampling distributions presented so far in this chapter depend on the assumption that the population from which the sample was taken has the normal distribution. This assumption often is satisfied, at least approximately for large samples, as illustrated by the central limit theorem. However, small samples sometimes must be used in practice, for example in statistical quality control or where taking and measuring a sample is very expensive. In an effort to deal with the problem of small samples in cases where it may be unreasonable to assume a normal population, statisticians have developed **nonparametric statistics**, whose sampling distributions do not depend upon any assumptions about the population from which the sample is taken. Statistical inferences based upon such statistics are called **nonparametric** inference. We will identify a class of nonparametric statistics called **order statistics** and discuss their statistical properties.

Consider a random sample of size n from an infinite population with a continuous density, and suppose that we arrange the values of $X_1, X_2, \ldots$, and X_n according to size. If we look upon the smallest of the x's as a value of the random variable Y_1, the next largest as a value of the random variable Y_2, the next largest after that as a value of the random variable $Y_3, \ldots$, and the largest as a value of the random variable Y_n, we refer to these Y's as **order statistics**. In particular, Y_1 is the first order statistic, Y_2 is the second order statistic, Y_3 is the third order statistic, and so on. (We are limiting this discussion to infinite populations with continuous densities so that there is zero probability that any two of the x's will be alike.)

To be more explicit, consider the case where $n = 2$ and the relationship between the values of the X's and the Y's is

$$y_1 = x_1 \quad \text{and} \quad y_2 = x_2 \quad \text{when} \quad x_1 < x_2$$
$$y_1 = x_2 \quad \text{and} \quad y_2 = x_1 \quad \text{when} \quad x_2 < x_1$$

Similarly, for $n = 3$ the relationship between the values of the respective random variables is

$$y_1 = x_1, \quad y_2 = x_2, \quad \text{and} \quad y_3 = x_3, \quad \text{when} \quad x_1 < x_2 < x_3$$
$$y_1 = x_1, \quad y_2 = x_3, \quad \text{and} \quad y_3 = x_2, \quad \text{when} \quad x_1 < x_3 < x_2$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$y_1 = x_3, \quad y_2 = x_2, \quad \text{and} \quad y_3 = x_1, \quad \text{when} \quad x_3 < x_2 < x_1$$

Let us now derive a formula for the probability density of the rth order statistic for $r = 1, 2, \ldots, n$.

THEOREM 16. For random samples of size n from an infinite population that has the value $f(x)$ at x, the probability density of the rth order statistic Y_r is given by

$$g_r(y_r) = \frac{n!}{(r-1)!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1} f(y_r)\left[\int_{y_r}^{\infty} f(x)\,dx\right]^{n-r}$$

for $-\infty < y_r < \infty$.

Proof Suppose that the real axis is divided into three intervals, one from $-\infty$ to y_r, a second from y_r to $y_r + h$ (where h is a positive constant), and the third from $y_r + h$ to ∞. Since the population we are sampling has the value $f(x)$ at x, the probability that $r - 1$ of the sample values fall into the first interval, 1 falls into the second interval, and $n - r$ fall into the third interval is

$$\frac{n!}{(r-1)!1!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1}\left[\int_{y_r}^{y_r+h} f(x)\,dx\right]\left[\int_{y_r+h}^{\infty} f(x)\,dx\right]^{n-r}$$

according to the formula for the multinomial distribution. Using the mean-value theorem for integrals from calculus, we have

$$\int_{y_r}^{y_r+h} f(x)\,dx = f(\xi)\cdot h \qquad \text{where } y_r \leqq \xi \leqq y_r + h$$

and if we let $h \to 0$, we finally get

$$g_r(y_r) = \frac{n!}{(r-1)!(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1} f(y_r)\left[\int_{y_r}^{\infty} f(x)\,dx\right]^{n-r}$$

for $-\infty < y_r < \infty$ for the probability density of the rth order statistic.

In particular, the sampling distribution of Y_1, the smallest value in a random sample of size n, is given by

$$g_1(y_1) = n \cdot f(y_1)\left[\int_{y_1}^{\infty} f(x)\,dx\right]^{n-1} \qquad \text{for } -\infty < y_1 < \infty$$

while the sampling distribution of Y_n, the largest value in a random sample of size n, is given by

$$g_n(y_n) = n \cdot f(y_n)\left[\int_{-\infty}^{y_n} f(x)\,dx\right]^{n-1} \qquad \text{for } -\infty < y_n < \infty$$

Also, in a random sample of size $n = 2m+1$ the **sample median** $\tilde{X}$ is Y_{m+1}, whose sampling distribution is given by

$$h(\tilde{x}) = \frac{(2m+1)!}{m!m!}\left[\int_{-\infty}^{\tilde{x}} f(x)\,dx\right]^{m} f(\tilde{x})\left[\int_{\tilde{x}}^{\infty} f(x)\,dx\right]^{m} \qquad \text{for } -\infty < \tilde{x} < \infty$$

[For random samples of size $n = 2m$, the median is defined as $\frac{1}{2}(Y_m + Y_{m+1})$.]

In some instances it is possible to perform the integrations required to obtain the densities of the various order statistics; for other populations there may be no choice but to approximate these integrals by using numerical methods.

EXAMPLE 4

Show that for random samples of size n from an exponential population with the parameter θ, the sampling distributions of Y_1 and Y_n are given by

$$g_1(y_1) = \begin{cases} \frac{n}{\theta} \cdot e^{-ny_1/\theta} & \text{for } y_1 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and

$$g_n(y_n) = \begin{cases} \frac{n}{\theta} \cdot e^{-y_n/\theta}[1 - e^{-y_n/\theta}]^{n-1} & \text{for } y_n > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and that, for random samples of size $n = 2m+1$ from this kind of population, the sampling distribution of the median is given by

$$h(\tilde{x}) = \begin{cases} \frac{(2m+1)!}{m!m!\theta} \cdot e^{-\tilde{x}(m+1)/\theta}[1 - e^{-\tilde{x}/\theta}]^{m} & \text{for } \tilde{x} > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Solution
The integrations required to obtain these results are straightforward, and they will be left to the reader in Exercise 45.

The following is an interesting result about the sampling distribution of the median, which holds when the population density is continuous and nonzero at the **population median** $\tilde{\mu}$, which is such that $\int_{-\infty}^{\tilde{\mu}} f(x)\,dx = \frac{1}{2}$.

THEOREM 17. For large n, the sampling distribution of the median for random samples of size $2n+1$ is approximately normal with the mean $\tilde{\mu}$ and the variance $\dfrac{1}{8[f(\tilde{\mu})]^2 n}$.

Note that for random samples of size $2n+1$ from a normal population we have $\mu = \tilde{\mu}$, so

$$f(\tilde{\mu}) = f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

and the variance of the median is approximately $\dfrac{\pi\sigma^2}{4n}$. If we compare this with the variance of the mean, which for random samples of size $2n+1$ from an infinite population is $\dfrac{\sigma^2}{2n+1}$, we find that for large samples from normal populations the mean is **more reliable** than the median; that is, the mean is subject to smaller chance fluctuations than the median.

Exercises

45. Verify the results of Example 4, that is, the sampling distributions of Y_1, Y_n, and $\tilde{X}$ shown there for random samples from an exponential population.

46. Find the sampling distributions of Y_1 and Y_n for random samples of size n from a continuous uniform population with $\alpha = 0$ and $\beta = 1$.

47. Find the sampling distribution of the median for random samples of size $2m+1$ from the population of Exercise 46.

48. Find the mean and the variance of the sampling distribution of Y_1 for random samples of size n from the population of Exercise 46.

49. Find the sampling distributions of Y_1 and Y_n for random samples of size n from a population having the beta distribution with $\alpha = 3$ and $\beta = 2$.

50. Find the sampling distribution of the median for random samples of size $2m+1$ from the population of Exercise 49.

51. Find the sampling distribution of Y_1 for random samples of size $n = 2$ taken
(a) without replacement from the finite population that consists of the first five positive integers;
(b) with replacement from the same population.
(*Hint*: Enumerate all possibilities.)

52. Duplicate the method used in the proof of Theorem 16 to show that the joint density of Y_1 and Y_n is given by

$$g(y_1, y_n) = n(n-1)f(y_1)f(y_n)\left[\int_{y_1}^{y_n} f(x)\,dx\right]^{n-2} \quad \text{for } -\infty < y_1 < y_n < \infty$$

and $g(y_1, y_n) = 0$ elsewhere.
(a) Use this result to find the joint density of Y_1 and Y_n for random samples of size n from an exponential population.
(b) Use this result to find the joint density of Y_1 and Y_n for the population of Exercise 46.

53. With reference to part (b) of Exercise 52, find the covariance of Y_1 and Y_n.

54. Use the formula for the joint density of Y_1 and Y_n shown in Exercise 52 and the transformation technique of several variables to find an expression for the joint density of Y_1 and the **sample range** $R = Y_n - Y_1$.

55. Use the result of Exercise 54 and that of part (a) of Exercise 52 to find the sampling distribution of R for random samples of size n from an exponential population.

56. Use the result of Exercise 54 to find the sampling distribution of R for random samples of size n from the continuous uniform population of Exercise 46.

57. Use the result of Exercise 56 to find the mean and the variance of the sampling distribution of R for random samples of size n from the continuous uniform population of Exercise 46.

58. There are many problems, particularly in industrial applications, in which we are interested in the proportion of a population that lies between certain limits. Such limits are called **tolerance limits**. The following steps lead to the sampling distribution of the statistic P, which is the proportion of a population (having a continuous density) that lies between the smallest and the largest values of a random sample of size n.

(a) Use the formula for the joint density of Y_1 and Y_n shown in Exercise 52 and the transformation technique of several variables to show that the joint density of Y_1 and P, whose values are given by

$$p = \int_{y_1}^{y_n} f(x)\,dx$$

is

$$h(y_1, p) = n(n-1)f(y_1)p^{n-2}$$

(b) Use the result of part (a) and the transformation technique of several variables to show that the joint density of P and W, whose values are given by

$$w = \int_{-\infty}^{y_1} f(x)\,dx$$

is

$$\varphi(w, p) = n(n-1)p^{n-2}$$

for $w > 0, p > 0, w + p < 1$, and $\varphi(w, p) = 0$ elsewhere.

(c) Use the result of part (b) to show that the marginal density of P is given by

$$g(p) = \begin{cases} n(n-1)p^{n-2}(1-p) & \text{for } 0 < p < 1 \\ 0 & \text{elsewhere} \end{cases}$$

This is the desired density of the proportion of the population that lies between the smallest and the largest values of a random sample of size n, and it is of interest to note that it does not depend on the form of the population distribution.

59. Use the result of Exercise 58 to show that, for the random variable P defined there,

$$E(P) = \frac{n-1}{n+1} \quad \text{and} \quad \text{var}(P) = \frac{2(n-1)}{(n+1)^2(n+2)}$$

What can we conclude from this about the distribution of P when n is large?

8 The Theory in Practice

More on Random Samples

While it is practically impossible to take a purely random sample, there are several methods that can be employed to assure that a sample is close enough to randomness to be useful in representing the distribution from which it came. In selecting a sample from a production line, *systematic sampling* can be used to select units at evenly spaced periods of time or having evenly spaced run numbers. In selecting a random sample from products in a warehouse, a *two-stage sampling process* can be used, numbering the containers and using a random device, such as a set of random numbers generated by a computer, to choose the containers. Then, a second set of random numbers can be used to select the unit or units in each container to be included in the sample. There are many other methods, employing mechanical devices or computer-generated random numbers, that can be used to aid in selecting a random sample.

Selection of a sample that reasonably can be regarded as random sometimes requires ingenuity, but it always requires care. Care should be taken to assure that only the specified distribution is represented. Thus, if a sample of product is meant to represent an entire production line, it should not be taken from the first shift only. Care should be taken to assure independence of the observations. Thus, the production-line sample should not be taken from a "chunk" of products produced at about the same time; they represent the same set of conditions and settings, and the resulting observations are closely related to each other. Human judgment in selecting samples usually includes personal bias, often unconscious, and such judgments should be avoided. Whenever possible, the use of mechanical devices or random numbers is preferable to methods involving personal choice.

The Assumption of Normality

It is not unusual to expect that errors are made in taking and recording observations. This phenomenon was described by early nineteenth-century astronomers who noted that different observers obtained somewhat different results when determining the location of a star.

Observational error can arise from one or both of two sources, **random error**, or statistical error, and **bias**. Random errors occur as the result of many imperfections of measurement; among these imperfections are imprecise markings on measurement scales, parallax (not viewing readings straight on) errors in setting up apparatus, slight differences in materials, expansion and contraction, minor changes in ambient conditions, and so forth. Bias occurs when there is a relatively consistent error, such as not obtaining a representative sample in a survey, using a measuring instrument that is not properly calibrated, and recording errors.

Errors involving bias can be corrected by discerning the source of the error and making appropriate "fixes" to eliminate the bias. Random error, however, is something we must live with, as no human endeavor can be made perfect in applications. Let us assume, however, that the many individual sources of random error, known or unknown, are additive. In fact this is usually the case, at least to a good approximation. Then we can write

$$X = \mu + E_1 + E_2 + \cdots + E_n$$

where the random variable X is an observed value, μ is the "true" value of the observation, and the E_i are the n random errors that affect the value of the observation. We shall assume that

$$E(X) = \mu + E(E_1) + E(E_2) + \cdots + E(E_n) = \mu$$

In other words, we are assuming that the random errors have a mean of zero, at least in the long run. We also can write

$$\text{var}(X) = (\mu + E_1 + E_2 + \cdots + E_n) = n\sigma^2$$

In other words, the variance of the *sum* of the random errors is $n\sigma^2$.

It follows that $\overline{X} = \mu + \overline{E}$, where $\overline{E}$ is the sample mean of the errors $E_1, E_2, \ldots, E_n$, and $\sigma^2{}_{\overline{X}} = \sigma^2/n$. The central limit theorem given by Theorem 3 allows us to conclude that

$$Z = \frac{\overline{X} - \mu}{\sigma\sqrt{n}}$$

is a random variable whose distribution as $n \to \infty$ is the standard normal distribution.

It is not difficult to see from this argument that most repeated measurements of the same phenomenon are, at least approximately, normally distributed. It is this conclusion that underscores the importance of the chi-square, t, and F distributions in applications that are based on the assumption of data from normally distributed populations. It also demonstrates why the normal distribution is of major importance in statistics.

Applied Exercises SECS. 1–3

In the following exercises it is assumed that all samples are drawn without replacement unless otherwise specified.

60. How many different samples of size $n = 3$ can be drawn from a finite population of size

(a) $N = 12$; **(b)** $N = 20$; **(c)** $N = 50$?

61. What is the probability of each possible sample if

(a) a random sample of size $n = 4$ is to be drawn from a finite population of size $N = 12$;

(b) a random sample of size $n = 5$ is to be drawn from a finite population of size $N = 22$?

62. If a random sample of size $n = 3$ is drawn from a finite population of size $N = 50$, what is the probability that a particular element of the population will be included in the sample?

63. For random samples from an infinite population, what happens to the standard error of the mean if the sample size is

(a) increased from 30 to 120;

(b) increased from 80 to 180;

(c) decreased from 450 to 50;

(d) decreased from 250 to 40?

64. Find the value of the finite population correction factor $\frac{N-n}{N-1}$ for

(a) $n = 5$ and $N = 200$;

(b) $n = 50$ and $N = 300$;

(c) $n = 200$ and $N = 800$.

65. A random sample of size $n = 100$ is taken from an infinite population with the mean $\mu = 75$ and the variance $\sigma^2 = 256$.

(a) Based on Chebyshev's theorem, with what probability can we assert that the value we obtain for $\overline{X}$ will fall between 67 and 83?

(b) Based on the central limit theorem, with what probability can we assert that the value we obtain for $\overline{X}$ will fall between 67 and 83?

66. A random sample of size $n = 81$ is taken from an infinite population with the mean $\mu = 128$ and the standard deviation $\sigma = 6.3$. With what probability can we assert that the value we obtain for $\overline{X}$ will not fall between 126.6 and 129.4 if we use

(a) Chebyshev's theorem;

(b) the central limit theorem?

67. Rework part (b) of Exercise 66, assuming that the population is not infinite but finite and of size $N = 400$.

68. A random sample of size $n = 225$ is to be taken from an exponential population with $\theta = 4$. Based on the central limit theorem, what is the probability that the mean of the sample will exceed 4.5?

69. A random sample of size $n = 200$ is to be taken from a uniform population with $\alpha = 24$ and $\beta = 48$. Based on the central limit theorem, what is the probability that the mean of the sample will be less than 35?

70. A random sample of size 64 is taken from a normal population with $\mu = 51.4$ and $\sigma = 6.8$. What is the probability that the mean of the sample will

(a) exceed 52.9;

(b) fall between 50.5 and 52.3;

(c) be less than 50.6?

71. A random sample of size 100 is taken from a normal population with $\sigma = 25$. What is the probability that the mean of the sample will differ from the mean of the population by 3 or more either way?

72. Independent random samples of sizes 400 are taken from each of two populations having equal means and the standard deviations $\sigma_1 = 20$ and $\sigma_2 = 30$. Using Chebyshev's theorem and the result of Exercise 2, what can we assert with a probability of at least 0.99 about the value we will get for $\overline{X}_1 - \overline{X}_2$? (By "independent" we mean that the samples satisfy the conditions of Exercise 2.)

73. Assume that the two populations of Exercise 72 are normal and use the result of Exercise 3 to find k such that

$$P(-k < \overline{X}_1 - \overline{X}_2 < k) = 0.99$$

74. Independent random samples of sizes $n_1 = 30$ and $n_2 = 50$ are taken from two normal populations having the means $\mu_1 = 78$ and $\mu_2 = 75$ and the variances $\sigma_1^2 = 150$ and $\sigma_2^2 = 200$. Use the results of Exercise 3 to find the probability that the mean of the first sample will exceed that of the second sample by at least 4.8.

75. The actual proportion of families in a certain city who own, rather than rent, their home is 0.70. If 84 families in this city are interviewed at random and their responses to the question of whether they own their home are looked upon as values of independent random variables having identical Bernoulli distributions with the parameter $\theta = 0.70$, with what probability can we assert that the value we obtain for the sample proportion $\hat{\Theta}$ will fall between 0.64 and 0.76, using the result of Exercise 4 and

(a) Chebyshev's theorem;

(b) the central limit theorem?

76. The actual proportion of men who favor a certain tax proposal is 0.40 and the corresponding proportion for women is 0.25; $n_1 = 500$ men and $n_2 = 400$

women are interviewed at random, and their individual responses are looked upon as the values of independent random variables having Bernoulli distributions with the respective parameters $\theta_1 = 0.40$ and $\theta_2 = 0.25$. What can we assert, according to Chebyshev's theorem, with a probability of at least 0.9375 about the value we will get for $\hat{\Theta}_1 - \hat{\Theta}_2$, the difference between the two sample proportions of favorable responses? Use the result of Exercise 5.

SECS. 4–6

(In Exercises 78 through 83, refer to Tables IV, V, and VI of "Statistical Tables.")

77. Integrate the appropriate chi-square density to find the probability that the variance of a random sample of size 5 from a normal population with $\sigma^2 = 25$ will fall between 20 and 30.

78. The claim that the variance of a normal population is $\sigma^2 = 25$ is to be rejected if the variance of a random sample of size 16 exceeds 54.668 or is less than 12.102. What is the probability that this claim will be rejected even though $\sigma^2 = 25$?

79. The claim that the variance of a normal population is $\sigma^2 = 4$ is to be rejected if the variance of a random sample of size 9 exceeds 7.7535. What is the probability that this claim will be rejected even though $\sigma^2 = 4$?

80. A random sample of size $n = 25$ from a normal population has the mean $\bar{x} = 47$ and the standard deviation $s = 7$. If we base our decision on the statistic of Theorem 13, can we say that the given information supports the conjecture that the mean of the population is $\mu = 42$?

81. A random sample of size $n = 12$ from a normal population has the mean $\bar{x} = 27.8$ and the variance $s^2 = 3.24$. If we base our decision on the statistic of Theorem 13, can we say that the given information supports the claim that the mean of the population is $\mu = 28.5$?

82. If S_1 and S_2 are the standard deviations of independent random samples of sizes $n_1 = 61$ and $n_2 = 31$ from normal populations with $\sigma_1^2 = 12$ and $\sigma_2^2 = 18$, find $P(S_1^2/S_2^2 > 1.16)$.

83. If S_1^2 and S_2^2 are the variances of independent random samples of sizes $n_1 = 10$ and $n_2 = 15$ from normal populations with equal variances, find $P(S_1^2/S_2^2 < 4.03)$.

84. Use a computer program to verify the five entries in Table IV of "Statistical Tables" corresponding to 11 degrees of freedom.

85. Use a computer program to verify the eight entries in Table V of "Statistical Tables" corresponding to 21 degrees of freedom.

86. Use a computer program to verify the five values of $f_{0.05}$ in Table VI of "Statistical Tables" corresponding to 7 and 6 to 10 degrees of freedom.

87. Use a computer program to verify the six values of $f_{0.01}$ in Table VI of "Statistical Tables" corresponding to $\nu_1 = 15$ and $\nu_2 = 12, 13, \ldots, 17$.

SEC. 7

88. Find the probability that in a random sample of size $n = 4$ from the continuous uniform population of Exercise 46, the smallest value will be at least 0.20.

89. Find the probability that in a random sample of size $n = 3$ from the beta population of Exercise 77, the largest value will be less than 0.90.

90. Use the result of Exercise 56 to find the probability that the range of a random sample of size $n = 5$ from the given uniform population will be at least 0.75.

91. Use the result of part (c) of Exercise 58 to find the probability that in a random sample of size $n = 10$ at least 80 percent of the population will lie between the smallest and largest values.

92. Use the result of part (c) of Exercise 58 to set up an equation in n whose solution will give the sample size that is required to be able to assert with probability $1-\alpha$ that the proportion of the population contained between the smallest and largest sample values is at least p. Show that for $p = 0.90$ and $\alpha = 0.05$ this equation can be written as

$$(0.90)^{n-1} = \frac{1}{2n+18}$$

This kind of equation is difficult to solve, but it can be shown that an approximate solution for n is given by

$$\frac{1}{2} + \frac{1}{4} \cdot \frac{1+p}{1-p} \cdot \chi^2_{\alpha,4}$$

where $\chi^2_{\alpha,4}$ must be looked up in Table V of "Statistical Tables". Use this method to find an approximate solution of the equation for $p = 0.90$ and $\alpha = 0.05$.

SEC. 8

93. Cans of food, stacked in a warehouse, are sampled to determine the proportion of damaged cans. Explain why a sample that includes only the top can in each stack would not be a random sample.

94. An inspector chooses a sample of parts coming from an automated lathe by visually inspecting all parts, and then including 10 percent of the "good" parts in the sample with the use of a table of random digits.

(a) Why does this method not produce a random sample of the production of the lathe?

(b) Of what population can this be considered to be a random sample?

95. Sections of aluminum sheet metal of various lengths, used for construction of airplane fuselages, are lined up on a conveyor belt that moves at a constant speed. A sample is selected by taking whatever section is passing in front of a station at five-minute intervals. Explain why this sample may not be random; that is, it is not an accurate representation of the population of all aluminum sections.

96. A process error may cause the oxide thicknesses on the surface of a silicon wafer to be "wavy," with a constant difference between the wave heights. What precautions are necessary in taking a random sample of oxide thicknesses at various positions on the wafer to assure that the observations are independent?

References

Necessary and sufficient conditions for the strongest form of the central limit theorem for independent random variables, the *Lindeberg–Feller* conditions, are given in

Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd ed. New York: John Wiley & Sons, Inc., 1968,

as well as in other advanced texts on probability theory.

Extensive tables of the normal, chi-square, F, and t distributions may be found in

Pearson, E. S., and Hartley, H. O., *Biometrika Tables for Statisticians*, Vol. I. New York: John Wiley & Sons, Inc., 1968.

A general formula for the variance of the sampling distribution of the second sample moment M_2 (which differs from S^2 only insofar as we divide by n instead of $n-1$) is derived in

Cramér, H., *Mathematical Methods of Statistics*. Princeton, N.J.: Princeton University Press, 1950,

and a proof of Theorem 17 is given in

Wilks, S. S., *Mathematical Statistics*. New York: John Wiley & Sons, Inc., 1962.

Proofs of the independence of $\overline{X}$ and S^2 for random samples from normal populations are given in many advanced texts on mathematical statistics. For instance, a proof based on moment-generating functions may be found in the above-mentioned book by S. S. Wilks, and a somewhat more elementary proof, illustrated for $n = 3$, may be found in

Keeping, E. S., *Introduction to Statistical Inference*. Princeton, N.J.: D. Van Nostrand Co., Inc., 1962.

The proof outlined in Exercise 48 is given in

Shuster, J., "A Simple Method of Teaching the Independence of $\overline{X}$ and S^2," *The American Statistician*, Vol. 27, No. 1, 1973.

Answers to Odd-Numbered Exercises

11 When we sample with replacement from a finite population, we satisfy the conditions for random sampling from an infinite population; that is, the random variables are independent and identically distributed.

17 $\mu = 13.0$; $\sigma^2 = 25.6$.

19 $s^2 = 4$.

29 21.9% and 5.53%.

47 $h(\tilde{x}) = \dfrac{(2m+1)!}{m!m!}\tilde{x}(1-\tilde{x})^m$ for $0 < x < 1$; $h(\tilde{x}) = 0$ elsewhere.

49 $g_1(y_1) = 12ny_1^2(1-y_1)(1-4y_1)^3$.

51 **(a)**

y_1	1	2	3	4
$g_1(y_1)$	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

(b)

y_1	1	2	3	4	5
$g_1(y_1)$	$\frac{9}{25}$	$\frac{7}{25}$	$\frac{5}{25}$	$\frac{3}{25}$	$\frac{1}{25}$

53 $\dfrac{1}{(n+1)^2(n+2)}$.

55 $f(R) = \dfrac{n-1}{\theta}e^{-R/\theta}[1-e^{-R/\theta}]^{n-2}$ for $R > 0$; $f(R) = 0$ elsewhere.

57 $E(R) = \dfrac{n-1}{n+1}$; $\sigma^2 = \dfrac{2(n-1)}{(n+1)^2(n+2)}$.

61 **(a)** $\frac{1}{495}$; **(b)** $\frac{1}{77}$.

63 **(a)** It is divided by 2. **(b)** It is divided by 1.5. **(c)** It is multiplied by 3. **(d)** It is multiplied by 2.5.

65 **(a)** 0.96; **(b)** 0.9999994.

67 0.0250.

69 0.0207.

71 0.2302.

73 4.63.

75 **(a)** 0.3056; **(b)** 0.7698.

77 0.216.

79 0.5.

81 $t = -1.347$; the data support the claim.

83 0.99.

89 0.851.

91 0.6242.

Decision Theory†

1 Introduction

In applied situations, mathematical expectations are often used as a guide in choosing among alternatives, that is, in making decisions, because it is generally considered rational to select alternatives with the "most promising" mathematical expectations—the ones that maximize expected profits, minimize expected losses, maximize expected sales, minimize expected costs, and so on.

Although this approach to decision making has great intuitive appeal, it is not without complications, for there are many problems in which it is difficult, if not impossible, to assign numerical values to the consequences of one's actions and to the probabilities of all eventualities.

EXAMPLE 1

A manufacturer of leather goods must decide whether to expand his plant capacity now or wait at least another year. His advisors tell him that if he expands now and economic conditions remain good, there will be a profit of \$164,000 during the next fiscal year; if he expands now and there is a recession, there will be a loss of \$40,000; if he waits at least another year and economic conditions remain good, there will be a profit of \$80,000; and if he waits at least another year and there is a recession, there will be a small profit of \$8,000. What should the manufacturer decide to do if he wants to minimize the expected loss during the next fiscal year and he feels that the odds are 2 to 1 that there will be a recession?

Solution

Schematically, all these "payoffs" can be represented as in the following table, where the entries are the losses that correspond to the various possibilities and, hence, gains are represented by negative numbers:

†Although the material in this chapter is basic to an understanding of the foundations of statistics, it is often omitted in a first course in mathematical statistics.