Chapter 5

CONTINUOUS PROBABILITY DISTRIBUTIONS

REVIEW OF KEY CONCEPTS

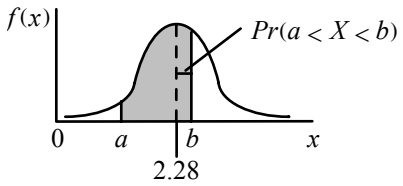
FEV (forced expiratory volume) is an important measure of lung function, representing the volume of air expelled in 1 second in a spirometry test. From the Tecumseh study, the distribution of FEV in children 10–14 years old has mean 2.28 liters, sd 0.56 liters. How can we define abnormal lung function in this age group? We could choose the lower 5% point or 10% point, or we could choose some absolute standard such as under 1.5 L and find the proportion of people who fall below this cutoff point. In any event, we need to make an assumption about the underlying distribution to calculate these probabilities. To accomplish this, we need to develop the concept of a *probability density function*.

SECTION 5.1 Probability Density Function

For distributions that take on a few discrete values, we could define probabilities of specific events; e.g., $Pr(0)$, $Pr(1)$, ... , $Pr(4)$ successes out of 4 color plates. For continuous distributions, there are an effectively infinite number of possibilities, which cannot be enumerated. How can we describe such distributions? By using a *probability density function*. A probability density function $f(x)$ has the following properties:

1. $f(x) \cong \dfrac{Pr(x < X < x + \Delta x)}{\Delta x}$ = "density" of probability at x as Δx approaches 0.
2. The total area under the density curve must equal 1.
3. The area under the density curve from a to $b = Pr(a < X < b)$.

Example: FEV in 10- 14-year-old children Probabilities of being in specific regions get smaller as one gets further from 2.28.



Density curve for FEV

The density curve does not have to be symmetric except in special cases such as for the normal distribution.

Example: Distribution of serum triglycerides.



Density curve for serum triglycerides

SECTION 5.2 Expected Value and Variance of a Continuous Random Variable

In Chapter 4, we learned that for a discrete random variable

$$E(X) = \sum x_i Pr(X = x_i) = \mu$$
$$Var(X) = \sum (x_i - \mu)^2 Pr(X = x_i) = \sigma^2$$

A similar definition holds for continuous random variables:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \mu$$
$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

In both cases, $\mu =$ average value for the random variable, $\sigma^2 =$ average squared distance of the random variable from the mean.

SECTION 5.3 Normal Probability Density Function

The normal distribution is the most commonly used continuous distribution. Its density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma} \right)^2 \right], -\infty < x < \infty$$

and is denoted as a $N(\mu, \sigma^2)$ distribution. Many random variables can only take on positive values but are well approximated by a normal distribution anyway, because the probability of obtaining a value < 0 is negligible. The normal density is a function of two parameters: μ, σ^2. These are the mean and variance for the distribution.

SECTION 5.4 Empirical and Symmetry Properties of the Normal Distribution

1. Distribution is symmetric about μ.
2. The points of inflection of the normal density are at $\mu + \sigma$ and $\mu - \sigma$, respectively.

3. The height of the normal density is inversely proportional to σ.
4. Probability $\cong 2/3$ that a normally distributed random variable will fall within $\mu \pm 1\sigma$
   $\cong 95\%$ that a normally distributed random variable will fall within $\mu \pm 2\sigma$ (actually $\mu \pm 1.96\sigma$)
   $\cong 99\%$ that a normally distributed random variable will fall within $\mu \pm 2.5\sigma$ (actually $\mu \pm 2.576\sigma$)

Example: Distribution of FEV in 10–14-year-old children is approximately normal with $\mu = 2.28$ L, $\sigma = 0.56$ L.

Range of FEV values	Approximate probability
$2.28 \pm 0.56 =$ (1.72, 2.84)	2/3
$2.28 \pm 1.12 =$ (1.16, 3.40)	95%
$2.28 \pm 1.40 =$ (0.88, 3.68)	99%

SECTION 5.5 Calculation of Probabilities for a Standard Normal Distribution

Suppose we want to compute $Pr(X < 1.5 \text{ L})$ for FEV in 10–14-year-old children. How can we do this? We first calculate probabilities for a standard normal distribution; i.e., a normal distribution with mean = 0 and variance = 1. In Table 3, column A (Appendix, text), we can obtain $\Phi(x) =$ cumulative probability distribution function $= \int_{-\infty}^{x} f(x)dx = Pr(X \le x)$ for positive values of x for a standard normal distribution. For negative values of x, we use the symmetry relationship $\Phi(x) = 1 - \Phi(-x)$. In general,

$$Pr(a \le X \le b) = \Phi(b) - \Phi(a).$$

Example: If $X \sim N(0, 1)$, what is

$$Pr(X \le 1.55),\ Pr(X \ge 1.55),\ Pr(1.55 \le X \le 2.75),\ Pr(-1.55 \le X \le 2.75)?$$

(The tilde, ~, stands for "is distributed as.")
Refer to Table 3 to obtain:

$$\begin{aligned}
Pr(X \le 1.55) &= \Phi(1.55) = .9394 \\
Pr(X \ge 1.55) &= 1 - \Phi(1.55) = .0606 \\
Pr(1.55 \le X \le 2.75) &= \Phi(2.75) - \Phi(1.55) = .9970 - .9394 = .0576 \\
Pr(-1.55 \le X \le 2.75) &= \Phi(2.75) - \Phi(-1.55) = \Phi(2.75) - [1 - \Phi(1.55)] \\
&= \Phi(2.75) + \Phi(1.55) - 1 \\
&= .9970 + .9394 - 1 = .9364
\end{aligned}$$

In general, we will not distinguish between $Pr(X < x)$ and $Pr(X \le x)$ for normal distributions, since the probabilities of specific values for continuous distributions (i.e., $Pr(X = x)$) is zero.

SECTION 5.6 Calculation of Probabilities for a General Normal Distribution

Suppose we want to compute $Pr(a < X < b)$ where $X \sim N(\mu, \sigma^2)$. We use the principle of standardization as follows:

$$Pr(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$Pr(X < a) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$Pr(X > b) = 1 - Pr(X < b) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$$

Thus, all computations for a general normal distribution can be performed based on percentiles of a standard normal distribution such as those given in Table 3 (Appendix, text).

Example: FEV in 10–14-year-old children

$$Pr(X < 1.5\text{ L}) = \Phi\left(\frac{1.5-2.28}{0.56}\right) = \Phi\left(\frac{-0.78}{0.56}\right) = \Phi(-1.393) = 1 - \Phi(1.393)$$
$$= 1 - .918 = .082.$$

Thus 8.2% of children will have FEV < 1.5 L.

SECTION 5.7 Inverse Normal Distribution

1. We define z_p to be the pth percentile of a standard normal distribution. We refer to the function z_p as the *inverse normal distribution function*. The probability is p that $Z \le z_p$ where Z follows a standard normal distribution (i.e., $\Phi(z_p) = p$).
2. The pth percentile of a general normal distribution with mean= μ and variance = σ^2 is $\mu + z_p\sigma$.

Example: What is the 5th percentile of the distribution of FEV in 10–14-year-old children? From Table 3 (Appendix, text), the 5th percentile of a standard normal distribution = -1.645. Therefore, the 5th percentile of the FEV distribution $= 2.28 - 1.645(0.56) = 1.36$ L. Thus, the interval (FEV < 1.36 L) is another way to specify an abnormal range of FEV values.

SECTION 5.8 Use of Electronic Tables for the Normal Distribution

Hemoglobin A1C (HgbAlc) is often used as a marker of compliance among insulin-dependent diabetic patients. Suppose that among diabetics who comply with taking their insulin shots that HgbAlc is normally distributed with mean = 8 and sd = 2.5. A diabetic patient comes to clinic with a HgbAlc of 12. Does this indicate that the patient is not complying with their insulin shots?

To answer this question we will compute $Pr\left[X > 12 \middle| X \sim N\left(8,\ 2.5^2\right)\right]$, where X = HgbAlc. We can use the NORMDIST function of Excel for this purpose. NORMDIST has 4 arguments (expressed as

NORMDIST(x, mean, sd, TYPE)), where TYPE = TRUE if the cumulative distribution function (cdf) is desired and TYPE = FALSE if the probability density function (pdf) is desired. Thus,

$$\text{NORMDIST}(x,\ \mu,\ \sigma,\ \text{TRUE}) = Pr\left[X \le x \middle| X \sim N(\mu,\ \sigma^2)\right]$$
$$\text{NORMDIST}(x,\ \mu,\ \sigma,\ \text{FALSE}) = \left[f(x) \middle| X \sim N(\mu,\ \sigma^2)\right] = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(x-\mu)/\sigma]^2}$$

Thus, to solve this problem, we compute $Pr\left[X \le 12 \middle| X \sim N(8,\ 2.5^2)\right]$ = NORMDIST(12, 8, 2.5, TRUE) and then subtract from 1 as shown on the spreadsheet. Thus,

$$Pr(X > 12) = 1 - .945 = .055.$$

Therefore, this is a slightly unusual event indicating that the patient may not be fully complying with their insulin shots.

Suppose we regard a subject who is in the upper 1% of the distribution of compliant patients as definitely not complying. What is this range of values?

We can use the Excel function NORMINV for this purpose. In general, NORMINV(p, μ, σ) provides the pth percentile of a normal distribution with mean = μ and sd = σ. Thus, we specify NORMINV(0.99, 8, 2.5) and obtain $x = 13.82$ (see spreadsheet below). Thus, a HgbAlc value ≥ 13.9 definitely indicates non-compliance. Notice that with the NORMDIST and NORMINV functions we can directly perform calculations for any normal distribution without referring to the standard normal distribution as we had to do when we were using a normal table (such as Table 3 (Appendix, text)).

If we have a specific reason to refer to the standard normal distribution, then we can use the Excel functions NORMSDIST and NORMSINV. For example, the National Center for Health Statistics uses a z-score scale to compare weights for a given child with age-sex specific norms. The weight z-score = (weight of child $-\mu)/\sigma$,where μ, σ are the mean and sd of weight for children of the same age-sex group as the child. Suppose a child has a weight z-score of 1.5. What is the weight percentile for that child? We can evaluate this by calculating NORMSDIST(1.5) $= \Phi(1.5) = Pr\left[X \le 1.5 \middle| X \sim N(0,\ 1)\right] = .933$ (see spreadsheet). Thus, the child is at the 93rd percentile for their age and sex group. Similarly, if a child is at the lower 10th percentile for their age-sex group, then their weight z-score $= z_{.10}$ and is given by NORMSINV(0.10) $= z_{0.10} = -1.28$ (see spreadsheet).

		Normal Distribution Functions of Excel	
NORMDIST	$Pr(X < 12)$	NORMDIST(12, 8, 2.5, TRUE)	0.945201
	$Pr(X \ge 12)$		0.054799
NORMINV		NORMINV(0.99, 8, 2.5)	13.81585
NORMSDIST		NORMSDIST(l.5)	0.933193
NORMSINV		NORMSINV(0.10)	–1.28155

SECTION 5.9 Normal Approximation to the Binomial Distribution

In children presenting with bilateral effusion (fluid in the middle ear of both ears) about 2/3 of children still have bilateral effusion after 4 weeks if left untreated. Suppose that if treated with antibiotics 118/198 or 59.6% have bilateral effusion after 4 weeks. Is treatment actually beneficial, or can these results be due to chance? Let X = number of treated children with bilateral effusion after 4 weeks. In general $Pr(X = k) = \binom{n}{k} p^k q^{n-k}$; in other words, X will follow a binomial distribution with parameters n and p. Suppose treatment was not beneficial at all. In this case, $n = 198$, $p = 2/3$. We want to know the

probability that 118 or fewer children will have bilateral effusion if the treatment is not beneficial at all. This is given by

$$Pr(X \le 118) = \sum_{k=0}^{118} \binom{198}{k} \left(\frac{2}{3}\right)^k \left(\frac{1}{3}\right)^{198-k}$$

This is very tedious to compute and instead we use the *normal approximation to the binomial distribution*.

We approximate the binomial random variable X by a normal random variable Y with the same mean (np) and variance (npq); i.e., $N(np, npq)$. This approximation should only be used if $npq \ge 5$. Under these conditions,

$$\begin{aligned} Pr(a \le X \le b) &\approx \Phi\left(\frac{b+\frac{1}{2}-\mu}{\sigma}\right) - \Phi\left(\frac{a-\frac{1}{2}-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b+\frac{1}{2}-np}{\sqrt{npq}}\right) - \Phi\left(\frac{a-\frac{1}{2}-np}{\sqrt{npq}}\right) \\ Pr(X \le a) &\approx \Phi\left(\frac{a+\frac{1}{2}-\mu}{\sigma}\right) = \Phi\left(\frac{a+\frac{1}{2}-np}{\sqrt{npq}}\right) \\ Pr(X \ge b) &= 1 - Pr(X \le b-1) \cong 1 - \Phi\left(\frac{b-\frac{1}{2}-\mu}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{b-\frac{1}{2}-np}{\sqrt{npq}}\right) \end{aligned}$$

Note: One can disregard the equal sign for Y but not for X, since Y is a continuous random variable, while X is a discrete random variable.

Example: Middle-ear effusion We wish to compute $Pr(X \le 118)$. We have $n = 198$, $p = 2/3$, $q = 1/3$, $a = 118$, $np = 132$, $npq = 44 \ge 5$. Therefore, we can use a normal approximation.

$$\begin{aligned} Pr(X \le 118) &\approx Pr(Y \le 118.5) = \Phi\left(\frac{118.5-132}{\sqrt{44}}\right) = \Phi\left(\frac{-13.5}{6.63}\right) = \Phi(-2.04) \\ &= 1 - \Phi(2.04) = 1 - .979 = .021 < .05 \end{aligned}$$

Since this probability is $< .05$, we conclude that treatment is beneficial.

Example: Color plates Suppose we identify a group of 100 people with a particular type of color blindness. We give each person the color plate test once and find that 32 identify the correct quadrant of the color plate. Are they doing significantly better than random?

If they are randomly selecting the quadrant, then $X =$ number of correct selections will be binomially distributed with parameters $n = 100$, $p = 1/4$. Since $npq = 18.75 \ge 5$, we can use a normal approximation. We approximate X by a normal distribution Y with mean $= np = 25$ and variance $= npq = 18.75$. We compute $Pr(X \ge 32)$ as follows:

$$\begin{aligned} Pr(X \ge 32) &\approx Pr(Y \ge 31.5) = 1 - \Phi\left(\frac{31.5-25}{\sqrt{18.75}}\right) = 1 - \Phi\left(\frac{6.5}{4.33}\right) \\ &= 1 - \Phi(1.50) = 1 - .933 = .067 > .05 \end{aligned}$$

This is not significantly different from random.

SECTION 5.10 Normal Approximation to the Poisson Distribution

Suppose we are studying a particular form of influenza. Usually 150 cases a year occur in a state. We observe 200 cases in 2004. Is this an unusual occurrence? We assume a Poisson distribution for $X =$ number of cases occurring in 1 year with parameters $\lambda = 150$, $T = 1$. We wish to compute $Pr(X \geq 200)$ based on the Poisson distribution. This will be tedious to do. Instead, we use a normal approximation. We approximate X by $Y \sim N(\mu, \mu)$, where $\mu = \lambda T$. Specifically

$$Pr(X = k) \cong Pr\left(k - \frac{1}{2} \leq Y \leq k + \frac{1}{2}\right) \text{ if } k > 0$$

$$Pr(X = 0) \cong Pr\left(Y \leq \frac{1}{2}\right)$$

$$Pr(a \leq X \leq b) \cong \Phi\left(\frac{b + \frac{1}{2} - \mu}{\sqrt{\mu}}\right) - \Phi\left(\frac{a - \frac{1}{2} - \mu}{\sqrt{u}}\right), \text{ where } a > 0$$

$$Pr(X \leq a) \approx \Phi\left(\frac{a + \frac{1}{2} - \mu}{\sqrt{u}}\right)$$

This approximation should only be used if $\mu \geq 10$.

Example: Influenza

$$Pr(X \geq 200) \approx Pr(Y \geq 199.5) = 1 - \Phi\left(\frac{199.5 - 150}{\sqrt{150}}\right) = 1 - \Phi\left(\frac{49.5}{12.25}\right)$$

$$= 1 - \Phi(4.04) < 10^{-4}$$

Thus, there is a significant increase in the number of cases in 2004.

PROBLEMS

5.1 What are the deciles of the standard normal distribution; that is, the 10, 20, 30, ... , 90 percentiles?

5.2 What are the quartiles of the standard normal distribution?

Hypertension
Blood pressure (BP) in childhood tends to increase with age, but differently for boys and girls. Suppose that for both boys and girls, mean systolic blood pressure is 95 mm Hg at 3 years of age and increases 1.5 mm Hg per year up to the age of 13. Furthermore, starting at age 13, the mean increases by 2 mm Hg per year for boys and 1 mm Hg per year for girls up to the age of 18. Finally, assume that blood pressure is normally distributed and that the standard deviation is 12 mm Hg for all age-sex groups.

5.3 What is the probability that an 11-year-old boy will have an SBP greater than 130 mm Hg?

5.4 What is the probability that a 15-year-old girl will have an SBP between 100 and 120 mm Hg?

5.5 What proportion of 17-year-old boys have an SBP between 120 and 140 mm Hg?

5.6 What is the probability that of 200 15-year-old boys, at least 10 will have an SBP of 130 mm Hg or greater?

5.7 What level of SBP is at the 80th percentile for 7-year-old boys?

5.8 What level of SBP is at the 70th percentile for 12-year-old girls?

5.9 Suppose that a task force of pediatricians decides that children over the 95th percentile, but not over the 99th percentile, for their age-sex group should be encouraged to take preventive nonpharmacologic measures to reduce their blood pressure, whereas those children

over the 99th percentile should receive antihypertensive drug therapy. Construct a table giving the appropriate BP levels to identify these groups for boys and girls for each year of age from 3 to 18.

Cancer
The incidence of breast cancer in 40–49-year-old women is approximately 1 new case per 1000 women per year.

5.10 What is the incidence of breast cancer over 10 years in women initially 40 years old?

5.11 Suppose we are planning a study based on an enrollment of 10,000 women. What is the probability of obtaining at least 120 new breast-cancer cases over a 10-year follow-up period?

Accident Epidemiology
Suppose the annual death rate from motor-vehicle accidents in 1980 was 10 per 100,000 in urban areas of the United States.

5.12 In an urban state with a population of 2 million, what is the probability of observing not more than 150 traffic fatalities in a given year?

In rural areas of the United States the annual death rate from motor-vehicle accidents is on the order of 100 per 100,000 population.

5.13 In a rural state with a population of 100,000, what is the probability of observing not more than 80 traffic fatalities in a given year?

5.14 How large should x be so that the probability of observing not more than x traffic fatalities in a given year in a rural state with population 100,000 is 5%?

5.15 How large should x be so that the probability of observing x or more traffic fatalities in a given year, in an urban state with population 500,000 is 10%?

Obstetrics
Assume that birthweights are normally distributed with a mean of 3400 g and a standard deviation of 700 g.

5.16 Find the probability of a low-birthweight child, where low birthweight is defined as ≤ 2500 g.

5.17 Find the probability of a very low birthweight child, where very low birthweight is defined as ≤ 2000 g.

5.18 Assuming that successive deliveries by the same woman have the same probability of being low birthweight, what is the probability that a woman with exactly 3 deliveries will have 2 or more low-birthweight deliveries?

Renal Disease
The presence of bacteria in a urine sample (bacteriuria) is sometimes associated with symptoms of kidney disease. Assume that a determination of bacteriuria has been made over a large population at one point in time and that 5% of those sampled are positive for bacteriuria.

5.19 Suppose that 500 people from this population are sampled. What is the probability that 50 or more people would be positive for bacteriuria?

Ophthalmology
A study was conducted among patients with retinitis pigmentosa, an ocular condition where pigment appears over the retina, resulting in substantial loss of vision in many cases [1]. The study was based on 94 patients who were seen annually at a baseline visit and at three annual follow-up visits. In this study, 90 patients provided visual-field measurements at each of the four examinations and are the subjects of the following data analyses. Visual field was transformed to the ln scale to better approximate normality and yielded the data given in Table 5.1.

Table 5.1 Visual-field measurements in retinitis-pigmentosa patients

Year of examination	Mean[a]	Standard deviation	n
Year 0 (baseline)	8.15	1.23	90
Year 3	8.01	1.33	90
Year 0–year 3	0.14	0.66	90

[a] ln (area of visual field) in degrees squared.

Source: Reprinted with permission of the *American Journal of Ophthalmology*, **99**, 240–251, 1985.

5.20 Assuming that change in visual field over 3 years is normally distributed when using the ln scale, what is the proportion of patients who showed a decline in visual field over 3 years?

5.21 What percentage of patients would be expected to show a decline of at least 20% in visual field over 3 years? (*Note*: In the ln scale this is equivalent to a decline of at least $\ln(1/0.8) = 0.223$).

5.22 Answer Problem 5.21 for a 50% decline over 3 years.

Cancer
A study of the Massachusetts Department of Health found 46 deaths due to cancer among women in the city of Bedford, MA over the period 1974–1978, where 30 deaths had been expected from statewide rates [2].

5.23 Write an expression for the probability of observing exactly k deaths due to cancer over this period if the statewide rates are correct.

5.24 Can the occurrence of 46 deaths due to cancer be attributed to chance? Specifically, what is the probability of observing at least 46 deaths due to cancer if the statewide rates are correct?

Hypertension
Suppose we want to recruit subjects for a hypertension treatment study and we feel that 10% of the population to be sampled is hypertensive.

5.25 If 100 subjects are required for the study and perfect cooperation is assumed, then how many people need to be sampled to be 80% sure of ascertaining at least 100 hypertensives?

5.26 How many people need to be sampled to be 90% sure of ascertaining at least 100 hypertensives?

Nutrition
The distribution of serum levels of alpha tocopherol (serum vitamin E) is approximately normal with mean 860 μg/dL and standard deviation 340 μg/dL.

5.27 What percentage of people have serum alpha tocopherol levels between 400 and 1000 μg/dL?

5.28 Suppose a person is identified as having toxic levels of alpha tocopherol if his or her serum level is > 2000 μg/dL. What percentage of people will be so identified?

5.29 A study is undertaken for evidence of toxicity among 2000 people who regularly take vitamin-E supplements. The investigators found that 4 people have serum alpha tocopherol levels > 2000 μg/dL. Is this an unusual number of people with toxic levels of serum alpha tocopherol?

Epidemiology
A major problem in performing longitudinal studies in medicine is that people initially entered into a study are lost to follow-up for various reasons.

5.30 Suppose we wish to evaluate our data after 2 years and anticipate that the probability a patient will be available for study after 2 years is 90%. How many patients should be entered into the study to be 80% sure of having at least 100 patients left at the end of this period?

5.31 How many patients should be entered to be 90% sure of having at least 150 patients after 4 years if the probability of remaining in the study after 4 years is 80%?

Pulmonary Disease

5.32 The usual annual death rate from asthma in England over the period 1862–1962 for people aged 5–34 was approximately 1 per 100,000. Suppose that in 1963 twenty deaths were observed in a group of 1 million people in this age group living in Britain. Is this number of deaths inconsistent with the preceding 100-year rate? In particular, what is the probability of observing 20 or more deaths in 1 year in a group of 1 million people? *Note*: This finding is both statistically and medically interesting, since it was found that the excess risk could be attributed to certain aerosols used by asthmatics in Britain during the period 1963–1967. The rate returned to normal during the period 1968–1972, when these types of aerosols were no longer used. (For further information see Speizer et al. [3].)

Hypertension
Blood-pressure measurements are known to be variable, and repeated measurements are essential to accurately characterize a person's BP status. Suppose a person is measured on n visits with k measurements per visit and the average of all nk diastolic blood pressure (DBP) measurements ($\bar{x}$) is used to classify a person as to BP status. Specifically, if $\bar{x} \geq 95$ mm Hg, then the person is classified as hypertensive; if $\bar{x} < 90$ mm Hg, then the person is classified as normotensive; and if $\bar{x} \geq 90$ mm Hg, and < 95 mm Hg, the person is classified as borderline. It is also assumed that a person's "true" blood pressure is μ, representing an average over a large number of visits with a large number of measurements per visit, and that $\bar{X}$ is normally distributed with mean μ and variance $= 27.7/n + 7.9/(nk)$.

5.33 If a person's true diastolic blood pressure is 100 mm Hg, then what is the probability that the person will be classified accurately (as hypertensive) if a single measurement is taken at 1 visit?

5.34 Is the probability in Problem 5.33 a measure of sensitivity, specificity, or predictive value?

5.35 If a person's true blood pressure is 85 mm Hg, then what is the probability that the person will be accurately classified (as normotensive) if 3 measurements are taken at each of 2 visits?

5.36 Is the probability in Problem 5.35 a measure of sensitivity, specificity, or predictive value?

5.37 Suppose we decide to take 2 measurements per visit. How many visits are needed so that the sensitivity and specificity in Problems 5.33 and 5.35 would each be at least 95%?

Hypertension
A study is planned to look at the effect of sodium restriction on lowering blood pressure. Nutritional counseling sessions are planned for the participants to encourage dietary sodium restriction. An important component in the study is validating the extent to which individuals comply with a sodium-restricted diet. This is usually accomplished by obtaining urine specimens and measuring sodium in the urine.

Assume that in free-living individuals (individuals with no sodium restriction) 24-hour urinary sodium excretion is normally distributed with mean 160.5 mEq/24 hrs and standard deviation $= 57.4$ mEq/24 hrs.

5.38 If 100 mEq/24 hr is the cutoff value chosen to measure compliance, then what percentage of noncompliant individuals—that is, individuals who do not restrict sodium—will have a 24-hour sodium level below this cutoff point?

Suppose that in an experimental study it is found that people who are on a sodium-restricted diet have 24-hour sodium excretion that is normally distributed with mean 57.5 mEq/24 hrs and standard deviation of 11.3 mEq/24 hrs.

5.39 What is the probability that a person who is on a sodium-restricted diet will have a 24-hour urinary sodium level above the cutoff point (100 mEq/24 hrs)?

5.40 Suppose the investigators in the study wish to change the cutoff value (from 100 mEq/24 hrs) to another value such that the misclassification probabilities in Problems 5.38 and 5.39 are the same. What should the new cutoff value be, and what are the misclassification probabilities corresponding to your answers to Problems 5.38 and 5.39?

5.41 The cutoff of 100 mEq/24 br is arbitrary. Suppose we regard a person eating a Na restricted diet as a true positive and a person eating a standard American diet as a true negative. Also, a person with a urinary Na < some cutoff c will be a test positive and a person with a urinary $\text{Na} \geq c$ will be a test negative. If we vary the cutoff c, then the sensitivity and specificity will change. Based on these data, plot the ROC curve for the 24 hr urinary Na test. Calculate the area under the ROC curve? What does it mean in this instance?

Hint: Using MINITAB, set up a column of successive integer values in increments of 5 from 0 to 300 as possible values for the cutoff (call this column C1). Use the normal cdf function of MINITAB to calculate the sensitivity and 1-specificity for each possible cutoff and store the values in separate columns (say C2 and C3). This only needs to be done once after C1 is created. You can then obtain the ROC curve by plotting C2 (y-axis) vs C3 (x-axis). Use the trapezoidal rule to calculate the area under the curve.

Table 5.2 shows data reported relating alcohol consumption and blood-pressure level among 18–39-year-old males [4].

TABLE 5.2 The relationship of blood pressure level to drinking behavior

	Systolic blood pressure (SBP)			Diastolic blood pressure (DBP)			Elevated blood pressure[b]
	Mean	sd	n	Mean	sd	n	
Nondrinker	120.2	10.7	96	74.8	10.2	96	11.5%
Heavy drinker[a]	123.5	12.8	124	78.5	9.2	124	17.7%

[a] >2.0 drinks per day.
[b] Either SBP ≥ 140 mm Hg or DBP ≥ 90 mm Hg.

5.42 If we assume that the distributions of SBP are normal, then what proportion of nondrinkers have

$$\text{SBP} \geq 140 \text{ mm Hg?}$$

5.43 What proportion of heavy drinkers have

$$\text{SBP} \geq 140 \text{ mm Hg?}$$

5.44 What proportion of heavy drinkers have "isolated systolic hypertension"; i.e., $\text{SBP} \geq 140$ mm Hg but $\text{DBP} < 90$ mm Hg?

5.45 What proportion of heavy drinkers have "isolated diastolic hypertension"; i.e., $\text{DBP} \geq 90$ mm Hg but $\text{SBP} < 140$ mm Hg?

Hypertension

The Pediatric Task Force Report on Blood Pressure Control in Children [5] reports blood-pressure norms for children by age and sex group. The mean ± standard deviation for 17-year-old boys for diastolic blood pressure (DBP) is 63.7 ± 11.4 mm Hg based on a large sample.

5.46 Suppose the 90th percentile of the distribution is the cutoff for elevated BP. If we assume that the distribution of blood pressure for 17-year-old boys is normally distributed, then what is the cutoff in mm Hg?

5.47 Another approach for defining elevated BP is to use 90 mm Hg as the cutoff (the standard for elevated adult DBP). What percentage of 17-year-old boys would have elevated BP using this approach?

5.48 Suppose there are 200 17-year-old boys in the 11th grade of whom 25 have elevated BP using the criteria in Problem 5.46. Are there an unusually large number of boys with elevated BP? Why or why not?

Environmental Health

5.49 A study was conducted relating particulate air pollution and daily mortality in Steubenville, Ohio [6]. On average over the last 10 years, there have been 3 deaths per day. Suppose that on 90 high-pollution days (where the total suspended particulates are in the highest quartile among all days) the death rate is 3.2 deaths per day or 288 deaths observed over the 90 high-pollution days. Are there an unusual number of deaths on high-pollution days?

Mental Health

A study was performed among three groups of male prisoners. Thirty six subjects between the ages of 18 and 45 were selected from the general inmate population at a Connecticut prison [7]. Three groups were identified from the inmate population: group A, a group of men with chronic aggressive

behavior, who were in prison for violent crimes such as aggravated assault or murder (the aggressive group); group B, a group of socially dominant men who were in prison for a variety of nonviolent crimes such as theft, check passing, or drug-related felonies and had asserted themselves into prestigious positions in prison hierarchy (the socially dominant group); group C, a nonaggressive group who were in prison for nonviolent crimes and were not socially dominant in prison hierarchy (the nonaggressive group). Group C was a group of "average" inmates and served as a control group. Twelve volunteers were obtained from each group who were willing to take blood tests. Blood samples were obtained on three consecutive days and a 3-day average value was computed for each subject. The mean ±1 sd for plasma testosterone and other variables are given in Table 5.3.

Table 5.3 Physical variables in three groups of male prisoners

	(A) Aggressive	(B) Socially dominant	(C) Non-aggressive
Testosterone (μg/mL ± *sd*)	10.10 ± 2.29[a]	8.36 ± 2.36[b]	5.99 ± 1.20
Weight (lb)	173.0 ± 13.8	182.2 ± 14.4	168.0 ± 25.2
Height (in)	71.2 ± 2.4	71.3 ± 2.2	69.5 ± 3.5
Age (yr)	29.08 ± 6.34	27.25 ± 3.59	28.4 ± 6.57

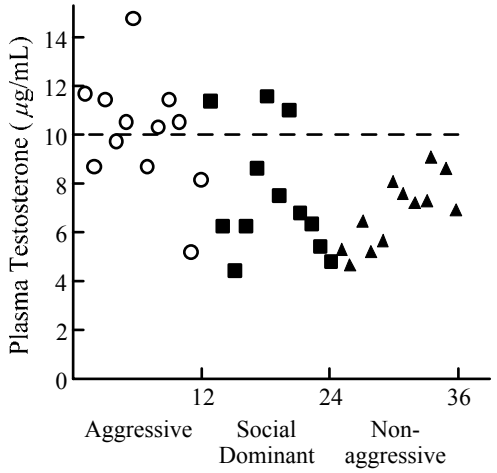
[a] Significant difference from nonaggressive group ($p < 0.01$).
[b] Significant difference from nonaggressive group ($p < 0.001$).

5.50 Suppose we assume that the distribution of mean plasma testosterone within each group is normally distributed. If values of mean plasma testosterone $> 10\,\mu$g/mL are considered high, then what percentage of subjects from each group would have high values?

5.51 The actual distribution of plasma-testosterone values within each group are plotted in Figure 5.1. Compare the observed and expected number of high values according to your answer to Problem 5.50. Do you think a normal distribution fits the data well? Why or why not?

5.52 The authors also remark that "there was a small variation in the plasma-testosterone values on successive days for each individual subject (±1.32 μg/mL standard deviation)." What is the coefficient of variation using the overall mean over all three groups ($n = 36$) as a typical mean level for a subject?

FIGURE 5.1 A plot of the mean values of plasma testosterone (μg/mL) for 36 subjects divided into the three groups: aggressive, social dominant, and nonaggressive



Ophthalmology

In ophthalmology, it is customary to obtain measurements on both the right eye (OD) and the left eye (OS). Suppose intraocular pressure (IOP) is measured on each eye and an average of the OD and OS values (denoted by the OU value) is computed. We wish to compute the variance of the IOP OU values for an individual person. Assume that the standard deviation is 4 mm Hg for the OD values and 4 mm Hg for the OS values (as estimated from the sample standard deviation of IOP from right and left eyes, respectively, from a sample of 100 people).

5.53 If the OD and OS values are assumed to be independent, then what is the variance of the IOP OU value for individual persons?

5.54 The correlation between OD and OS IOP values is typically about .8. What is the covariance between IOP OD values and IOP OS values for individual persons?

5.55 What is the variance of the IOP OU values for individual persons under the assumptions in Problem 5.54?

SOLUTIONS

5.1 We have

x	$\Phi(x)$	x	$\Phi(x)$
–1.28	.10	0.25	.60
–0.84	.20	0.52	.70
–0.52	.30	0.84	.80
–0.25	.40	1.28	.90
0.00	.50		

5.2 We have

x	–0.67	0.00	0.67
$\Phi(x)$	.25	.50	.75

5.3 The mean systolic blood pressure (SBP) for an 11-year-old boy $= 95 + 1.5 \times (11 - 3) = 107$. If X represents SBP

for an 11-year-old boy, then $X \sim N(107, 144)$. Therefore,

$$Pr(X > 130) = 1 - \Phi\left(\frac{130-107}{12}\right) = 1 - \Phi\left(\frac{23}{12}\right)$$
$$= 1 - \Phi(1.917) = 1 - .9724 = .028$$

5.4 The mean SBP for a 15-year-old girl

$$= 95 + 1.5 \times (13-3) + 1 \times (15-13) = 112,$$

while the standard deviation is 12. Therefore, if Y represents SBP for a 15-year-old girl, then

$$Pr(100 < Y < 120) = \Phi\left(\frac{120-112}{12}\right) - \Phi\left(\frac{100-112}{12}\right)$$
$$= \Phi(0.667) - \Phi(-1.00)$$
$$= .7475 - [1 - \Phi(1.00)]$$
$$= .7475 - .1587 = .589$$

5.5 The mean SBP for a 17-year-old boy

$$= 95 + 1.5 \times (13-3) + 2 \times (17-13) = 118,$$

while the standard deviation is 12. If Z represents SBP for a 17-year-old boy, then

$$Pr(120 < Z < 140) = \Phi\left(\frac{140-118}{12}\right) - \Phi\left(\frac{120-118}{12}\right)$$
$$= \Phi(1.833) - \Phi(0.167)$$
$$= .9666 - .5662 = .400$$

5.6 We first calculate the probability that one 15-year-old boy will have an SBP > 130. Let X_1 represent SBP for a 15-year-old boy. We have that the mean SBP $= 95 + 1.5 \times (13-3) + 2 \times (15-13) = 114$. Therefore,

$$Pr(X_1 > 130) = 1 - \Phi\left(\frac{130-114}{12}\right)$$
$$= 1 - \Phi\left(\frac{16}{12}\right) = 1 - \Phi(1.333) = .0912$$

Now let X represent the number of boys among 200 15-year-old boys with an SBP > 130. X is binomially distributed with parameters $n = 200$, $p = .0912$. We wish to compute $Pr(X \geq 10)$. We will approximate X by a normal distribution Y with

$$\text{mean} = np = 200 \times .0912 = 18.24$$

and

$$\text{variance} = npq = 200 \times .0912 \times .9088 = 16.58.$$

We have

$$Pr(X \geq 10) \cong Pr(Y > 9.5) = 1 - \Phi\left(\frac{9.5 - 18.24}{\sqrt{16.58}}\right)$$
$$= 1 - \Phi\left(\frac{-8.742}{4.072}\right)$$
$$= 1 - \Phi(-2.147) = \Phi(2.147) = .984$$

Therefore, there is a 98.4% probability that at least 10 of the 200 15-year-old boys will have an SBP of at least 130.

5.7 The mean SBP for 7-year-old boys is given by $95 + 1.5 \times (7-3) = 101$. From Table 3 (Appendix, text), the 80% point for a N(0, 1) distribution $= z_{0.80} = 0.84$. Therefore, since the distribution of SBP for 7-year-old boys is normal, with mean 101 and standard deviation 12, the 80% point $= 101 + 0.84 \times 12 = 111.1$ mm Hg.

5.8 From Table 3 (Appendix, text) the 70% point for a $N(0, 1)$ distribution $= z_{0.70} = 0.52$. Systolic blood pressure for 12-year-old girls is normally distributed with mean $95 + 1.5 \times (12-3) = 108.5$ and standard deviation $= 12$. Therefore, the 70% point for 12-year-old girls is given by $108.5 + 0.52 \times 12 = 114.7$ mm Hg.

5.9 Since $z_{0.95} = 1.645$, $z_{0.99} = 2.326$, it follows that in general, the 95th and 99th percentiles are given by $\mu + 1.645\sigma$, and $\mu + 2.326\sigma$ respectively, where $\mu =$ age-sex-specific mean SBP and $\sigma = 12$. For example, for 3-year-old boys,

the 95th percentile $= 95 + 1.645 \times 12 = 114.7$;
the 99th percentile $= 95 + 2.326 \times 12 = 122.9$.

These percentiles are presented in the table on page 56 for each year of age for 3-18-year-old boys and girls.

Therefore, for 3-year-old boys, the nonpharmacologic measure group is defined by SBP ≥ 115 and: SBP ≤ 122, while the antihypertensive-drug-therapy group is defined by SBP ≥ 123. A similar strategy is followed for the other age-sex subgroups.

5.10 .010

5.11 The number of breast-cancer cases (X) is binomally distributed with parameters $n = 10{,}000$, $p = .010$. This distribution is approximated by a normal distribution (Y) with mean $= np = 100$, variance $= npq = 99$. We have

$$Pr(X \geq 120) \approx Pr(Y \geq 119.5)$$
$$= 1 - \Phi\left(\frac{119.5 - 100}{\sqrt{99}}\right)$$
$$= 1 - \Phi(1.96) = .025$$

Boys				Girls			
		Percentiles				Percentiles	
Age	μ	95th	99th	Age	μ	95th	99th
3	95.0	114.7	122.9	3	95.0	114.7	122.9
4	96.5	116.2	124.4	4	96.5	116.2	124.4
5	98.0	117.7	125.9	5	98.0	117.7	125.9
6	99.5	119.2	127.4	6	99.5	119.2	127.4
7	101.0	120.7	128.9	7	101.0	120.7	128.9
8	102.5	122.2	130.4	8	102.5	122.2	130.4
9	104.0	123.7	131.9	9	104.0	123.7	131.9
10	105.5	125.2	133.4	10	105.5	125.2	133.4
11	107.0	126.7	134.9	11	107.0	126.7	134.9
12	108.5	128.2	136.4	12	108.5	128.2	136.4
13	110.0	129.7	137.9	13	110.0	129.7	137.9
14	112.0	131.7	139.9	14	111.0	130.7	138.9
15	114.0	133.7	141.9	15	112.0	131.7	139.9
16	116.0	135.7	143.9	16	113.0	132.7	140.9
17	118.0	137.7	145.9	17	114.0	133.7	141.9
18	120.0	139.7	147.9	18	115.0	134.7	142.9

5.12 Let X = number of traffic fatalities in the urban state in 1 year. We will assume that X is Poisson-distributed with parameter $\mu = 2 \times 10^6 \times (10/10^5) = 200$. We wish to compute $Pr(X \le 150)$. We will approximate X by a normal distribution Y with mean $= 200$, variance $= 200$. Compute

$$Pr(X \le 150) \cong Pr(Y \le 150.5) = \Phi\left(\frac{150.5 - 200}{\sqrt{200}}\right)$$
$$= \Phi(-3.50) = 1 - \Phi(3.50) = .0002$$

5.13 Let X = number of traffic fatalities in the rural state in 1 year. We will assume that X is Poisson-distributed with parameter $= \mu = 10^5 \times (100/10^5) = 100$. To compute $Pr(X \le 80)$, approximate X by a normal distribution Y with mean $= 100$, variance $= 100$. Therefore,

$$Pr(X \le 80) \cong Pr(Y \le 80.5) = \Phi\left(\frac{80.5 - 100}{\sqrt{100}}\right)$$
$$= \Phi(-1.95) = 1 - \Phi(1.95) = .026$$

5.14 We wish to compute x such that $Pr(X \le x) = .05$. We approximate X by a $N(100, 100)$ distribution (Y). Therefore, $Pr(X \le x)$ is approximated by

$$Pr(Y \le x + 0.5) = \Phi\left(\frac{x + 0.5 - 100}{10}\right) = .05 = \Phi(-1.645)$$

or

$$\frac{x + 0.5 - 100}{10} = -1.645$$

Solving for x, we obtain $x = 83.05$. Thus, x should be 83.

5.15 We wish to compute x such that $Pr(X \ge x) = .10$, where X is Poisson-distributed with parameter

$$\mu = 500{,}000 \times \frac{10}{10^5} = 50.$$

We approximate X by a $N(50, 50)$ distribution (Y). Therefore,

$$Pr(X \ge x) \cong Pr(Y \ge x - 0.5)$$
$$= 1 - \Phi\left(\frac{x - 0.5 - 50}{\sqrt{50}}\right) = .10$$

or

$$\Phi\left(\frac{50 - x + 0.5}{\sqrt{50}}\right) = 0.10 = \Phi(-1.28)$$

or

$$\frac{50 - x + 0.5}{\sqrt{50}} = -1.28$$

Solving for x, we obtain $x = 59.6$. Thus, x should be 60.

5.16 If X = birthweight, then $X \sim N(3400,\ 700^2)$. Thus,

$$Pr(X \le 2500) = \Phi\left(\frac{2500 - 3400}{700}\right) = \Phi(-1.286)$$
$$= 1 - \Phi(1.286)$$
$$= 1 - .9007 = .099$$

5.17 $\Pr(X \le 2000) = \Phi\left(\frac{2000 - 3400}{700}\right) = \Phi(-2.00)$
$$= 1 - \Phi(2.00) = 1 - .9772 = .023$$

5.18 Let X = number of low-birthweight deliveries. X is binomially distributed with parameters $n = 3$, $p = .099$. We wish to compute $Pr(X \geq 2)$.

$$Pr(X \geq 2) = \binom{3}{2}(.099)^2(.901) + \binom{3}{3}(.099)^3$$
$$= .0266 + .0010 = .028$$

5.19 Let X be the number of people positive for bacteriuria out of 500. X follows a binomial distribution with parameters $n = 500$ and $p = .05$. Compute $Pr(X \geq 50)$. X will be approximated by a normal random variable Y with mean $= 500 \times .05 = 25$ and variance $= 500 \times .05 \times .95 = 23.75$. Compute

$$Pr(X \geq 50) \approx Pr(Y \geq 49.5)$$
$$= 1 - \Phi\left(\frac{49.5 - 25}{\sqrt{23.75}}\right)$$
$$= 1 - \Phi(5.03) < .001$$

5.20 Let X = ln (visual-field year 0) –ln (visual-field year 3). We assume that $X \sim N(0.14,\ 0.66^2)$. A patient loses visual field if $X > 0$. Thus, we wish to compute

$$Pr(X > 0) = 1 - \Phi\left(\frac{0 - 0.14}{0.66}\right)$$
$$= 1 - \Phi(-0.212) = \Phi(0.212) = .584$$

Thus, about 58% of patients would be expected to decline in visual field over 3 years.

5.21 We wish to evaluate $Pr(X > 0.223)$. We have

$$Pr(X > 0.223) = 1 - \Phi\left(\frac{0.223 - 0.14}{0.66}\right)$$
$$= 1 - \Phi(0.126) = .450$$

Thus, about 45% of patients would be expected to decline at least 20% in visual field over 3 years.

5.22 The equivalent decline in the ln scale is

$$\ln(1/0.5) = 0.693.$$

We have

$$Pr(X > 0.693) = 1 - \Phi\left(\frac{0.693 - 0.14}{0.66}\right)$$
$$= 1 - \Phi(0.838) = .201$$

Thus, about 20% of patients would be expected to decline at least 50% in visual field over 3 years.

5.23 Suppose there are n women living in the city of Bedford, MA during 1974–1978 and that the probability of getting cancer for an individual woman over this period $= p$. If X = the number of women who actually get cancer, then X is binomially distributed with parameters n and p. Since n is large and p is small, we can approximate this distribution by a Poisson distribution with parameter $\mu = np = 30$. Thus,

$$Pr(k \text{ deaths}) = \frac{e^{-30}30^k}{k!},\ k = 0,\ 1,\ 2,\ \ldots$$

5.24 We wish to compute $Pr(X \geq 46)$. This is tedious to do with the Poisson distribution. Instead, we approximate the Poisson random variable X by a normal random variable Y with mean = variance $= 30$. Thus,

$$Pr(X \geq 46) \cong Pr(Y \geq 45.5)$$
$$= 1 - \Phi\left(\frac{45.5 - 30}{\sqrt{30}}\right)$$
$$= 1 - \Phi(2.83) = 1 - .9977 = .002$$

Thus, it is unlikely that the statewide cancer rate applies to the city of Bedford, MA.

5.25 Suppose n people are sampled for the study. The number of hypertensives X ascertained from this sample will be binomially distributed with parameters n and $p = .10$. For large n, from the normal approximation to the binomial, it follows that the distribution of X can be approximated by a normal distribution Y with mean $.1n$ and variance $n \times .1 \times .9 = .09n$. Also, we want n to be large enough so that $Pr(X \geq 100) = .8$. We have

$$Pr(X \geq 100) \approx Pr(Y \geq 99.5) = 1 - Pr(Y \leq 99.5)$$
$$= 1 - \Phi\left(\frac{99.5 - .1n}{\sqrt{.09n}}\right) = .8$$

or

$$\Phi\left(\frac{99.5 - .1n}{\sqrt{.09n}}\right) = .2$$

However, from the normal tables (Table 3, Appendix, text), $\Phi(0.84) = .8$ or $\Phi(-0.84) = .2$. Thus,

$$\frac{99.5 - 0.1n}{\sqrt{0.09n}} = -0.84$$

or

$$99.5 - 0.1n = -0.84\sqrt{0.09n}$$

or

$$.1n - 0.252\sqrt{n} - 99.5 = 0$$

If z^2 is substituted for n, this equation can be rewritten as $.1z^2 - 0.252z - 99.5 = 0$. The solution of this quadratic equation is given by

$$z = \sqrt{n} = \frac{0.252 \pm \sqrt{0.252^2 + 4(99.5)(0.1)}}{2(.1)}$$
$$= \frac{0.252 \pm \sqrt{39.864}}{.2} = \frac{0.252 \pm 6.314}{.2}$$
$$= -30.309 \text{ or } 32.829$$

or $n = z^2 = (-30.309)^2$ or $(32.829)^2 = 918.62$ or 1077.73. Since n must be larger than $100/0.1 = 1000$, we have $n = 1077.73$, or 1078. Thus, 1078 people need to be sampled to be 80% sure of recruiting 100 hypertensives.

5.26 The same approach is used as in Problem 5.25. Find n such that

$$Pr(Y \geq 99.5) = 1 - \Phi\left(\frac{99.5 - .1n}{\sqrt{.09n}}\right) = .9$$

or

$$\Phi\left(\frac{99.5 - .1n}{\sqrt{.09n}}\right) = .1$$

From the normal tables (Table 3, Appendix, text), $\Phi(1.28) = .9$ or $\Phi(-1.28) = .1$. Thus,

$$\frac{99.5 - .1n}{\sqrt{.09n}} = -1.28$$

or

$$99.5 - .1n = -0.384\sqrt{n}$$

or

$$.1n - 0.384\sqrt{n} - 99.5 = 0$$

Substitute z^2 for n and obtain the quadratic equation

$$.1z^2 - 0.384z - 99.5 = 0$$

The solution is given by

$$z = \frac{0.384 \pm \sqrt{0.384^2 + 4(.1)(99.5)}}{2(.1)}$$
$$= \frac{0.384 \pm \sqrt{39.947}}{.2} = \frac{0.384 \pm 6.320}{.2}$$
$$= -29.68 \text{ or } 33.52$$

Thus,

$$n = z^2 = (-29.68)^2 = 881.02 \text{ or } (33.52)^2 = 1123.72$$

Again use the root such that $n > 1000$, and thus $n = 1123.72$, or 1124 if rounded up to the nearest integer. Thus, 1124 people need to be sampled to be 90% sure of recruiting 100 hypertensives.

5.27 Let the random variable X represent serum alpha tocopherol. We have that $X \sim N(860, 340^2)$. We wish to compute

$$Pr(400 < X < 1000) = \Phi\left(\frac{1000 - 860}{340}\right) - \Phi\left(\frac{400 - 860}{340}\right)$$
$$= \Phi(0.412) - \Phi(-1.353)$$
$$= \Phi(0.412) - [1 - \Phi(1.353)]$$
$$= .6597 - .0880 = .572$$

Thus, 57% of people have serum alpha-tocopherol levels between 400 and 1000 μg/dL.

5.28 We wish to compute

$$Pr(X > 2000) = 1 - \Phi\left(\frac{2000 - 860}{340}\right)$$
$$= 1 - \Phi(3.353) = .0004$$

Thus, $< 1\%$ of people will have toxic levels of serum alpha tocopherol.

5.29 Let Y be the number of people with toxic levels of serum alpha tocopherol. From Problem 5.28, Y is binomially distributed with parameters $n = 2000$ and $p = .0004$. We wish to compute

$$Pr(Y \geq 4) = 1 - Pr(Y \leq 3).$$

We use the BINOMDIST function of Excel as follows:

$$Pr(Y \leq 3) = .991$$

Thus, $Pr(Y \geq 4) = 1 - .991 = .009$. Therefore, there are an unusual number of people with toxic levels in this group.

5.30 Suppose we enter n patients into the study. The number of patients, X, remaining in the study after two years will be binomially distributed with parameters n and $p = .90$. We want to choose n such that

$$Pr(X \geq 100) = .80.$$

We can approximate X by a normal random variable Y with mean $.9n$ and variance $n \times .9 \times .1 = .09n$. We have

$$Pr(X \geq 100) \cong Pr(Y \geq 99.5)$$
$$= 1 - \Phi\left(\frac{99.5 - .9n}{\sqrt{.09n}}\right) = .80$$

or

$$.20 = \Phi\left(\frac{99.5 - .9n}{\sqrt{.09n}}\right) = \Phi(-0.84)$$

Thus, we set $(99.5 - .9n)/\sqrt{.09n} = -0.84$, and obtain the equation

$$99.5 - .9n = -0.252\sqrt{n}$$

or

$$.9n - 0.252\sqrt{n} - 99.5 = 0$$

We substitute z^2 for n and rewrite the equation in the form

$$.9z^2 - 0.252z - 99.5 = 0$$

The solution to this quadratic equation is given by

$$z = \sqrt{n} = \frac{0.252 \pm \sqrt{0.252^2 + 4(.9)(99.5)}}{2(.9)}$$
$$= \frac{0.252 \pm \sqrt{358.26}}{1.8} = \frac{0.252 \pm 18.928}{1.8}$$
$$= -10.375 \text{ or } 10.655$$

Since $\sqrt{n}$ must be positive, we set $\sqrt{n} = 10.655$. Thus $n = (10.655)^2 = 113.5$ or 114 people. Thus, 114 people must be entered into the study in order to be 80% sure of having 100 patients left in the study after 2 years. Notice that this estimate is slightly larger than the intuitive answer of $100/.9 = 111.1$ or 112 which does not take into account the probabilistic nature of the binomial distribution.

5.31 Let $n =$ the number of patients entered into the study. The number of patients, *X*, remaining in the study after 4 years will be binomially distributed with parameters *n* and $p = .8$. We want to choose *n* such that

$$Pr(X \geq 150) = .90.$$

We can approximate *X* by a normal random variable *Y* with mean $.8n$ and variance $n \times .8 \times .2 = .16n$. We approximate $Pr(X \geq 150)$ by

$$Pr(Y \geq 149.5) = 1 - \Phi\left(\frac{149.5 - .8n}{\sqrt{.16n}}\right) = .90$$

Thus,

$$\Phi\left(\frac{149.5 - .8n}{\sqrt{.16n}}\right) = .10 = \Phi(-1.28)$$

and we set

$$\frac{149.5 - .8n}{\sqrt{.16n}} = -1.28$$

and obtain the equation

$$149.5 - .8n = -0.512\sqrt{n}$$

or

$$.8n - 0.512\sqrt{n} - 149.5 = 0$$

We substitute z^2 for n and rewrite the equation in the form

$$.8z^2 - 0.512z - 149.5 = 0$$

The solution to this quadratic equation is given by

$$z = \sqrt{n}$$
$$= \frac{0.512 \pm \sqrt{0.512^2 + 4(0.8)(149.5)}}{2(0.8)}$$
$$= \frac{0.512 \pm \sqrt{478.662}}{1.6}$$
$$= \frac{0.512 \pm 21.878}{1.6}$$
$$= -13.354 \text{ or } 13.994$$

Since $\sqrt{n}$ must be positive, we set $\sqrt{n} = 13.994$ or $n = 195.8$ or 196 people. Thus, 196 people must be entered into the study in order to be 90% sure of having 150 patients left in the study after 4 years. Once again, this is larger than the intuitive answer of $150/.8 = 187.5$ or 188 people. If we enrolled 188 people into the study, then we would have only a 50% probability that 150 subjects would remain in the study after 4 years, rather than the 90% probability that we require.

5.32 The number of deaths is a binomial random variable. However, since *n* is large (10^6), and *p* is small (10^{-5}), we can approximate this random variable by a Poisson distribution. Also, since $npq \cong 10$, we could use a normal approximation. We will use the normal approximation here since it is easier to use. We have $\mu = np = 10$, $\sigma^2 = npq \cong 10$. Thus, $X =$ number of deaths $\sim N(10, 10)$. We have

$$Pr(\geq 20 \text{ deaths}) \cong 1 - \Phi\left(\frac{19.5 - 10}{\sqrt{10}}\right)$$
$$= 1 - \Phi(3.004)$$
$$= .001$$

Thus, it is very unlikely that the same death rate prevailed in 1963, as in the previous 100 years.

5.33 If $n = 1$, $k = 1$, then

$$\overline{X} \sim N\left(100, \frac{27.7}{1} + \frac{7.9}{1}\right) = N(100, 35.6).$$

We wish to compute

$$Pr\left(\overline{X} \geq 95\right) = 1 - \Phi\left(\frac{95 - 100}{\sqrt{35.6}}\right)$$
$$= 1 - \Phi(-0.838)$$
$$= \Phi(0.838) = .799$$

Thus, there is an 80% chance of classifying such a person as hypertensive.

5.34 The probability in Problem 5.33 is

$$Pr\left(\overline{X} \geq 95 | \mu = 100\right) = Pr(\text{test } + | \text{true } +)$$
$$= \text{sensitivity}.$$

5.35 If $n = 2$, $k = 3$, then

$$\overline{X} \sim N\left(85, \frac{27.7}{2} + \frac{7.9}{6}\right) = N(85, 15.2).$$

We wish to compute

$$Pr\left(\overline{X} < 90\right) = \Phi\left(\frac{90 - 85}{\sqrt{15.2}}\right)$$
$$= \Phi(1.284) = .900$$

Thus, there is a 90% chance of classifying such a person as normotensive.

5.36 The probability in Problem 5.35 is

$$Pr\left(\overline{X} < 90 | \mu = 85\right) = Pr(\text{test } - | \text{true } -)$$
$$= \text{specificity}.$$

5.37 If we take 2 measurements per visit, then

$$\overline{X} \sim N\left(\mu, \frac{27.7}{n} + \frac{7.9}{2n}\right) = N\left(\mu, \frac{31.65}{n}\right).$$

We have that

$$\text{Sensitivity} = \Pr\left[\overline{X} \geq 95 \middle| \overline{X} \sim N\left(100, \frac{31.65}{n}\right)\right]$$
$$= 1 - \Phi\left(\frac{95 - 100}{\sqrt{\frac{31.65}{n}}}\right) = .95$$

or

$$\Phi\left(\frac{-5\sqrt{n}}{\sqrt{31.65}}\right) = .05$$

or

$$\Phi\left(\frac{5\sqrt{n}}{\sqrt{31.65}}\right) = .95$$

Since $\Phi(1.645) = .95$, it follows that

$$\frac{5\sqrt{n}}{\sqrt{31.65}} = 1.645$$

or

$$\sqrt{n} = \frac{1.645 \times \sqrt{31.65}}{5} = 1.851$$

or

$$n = 1.851^2 = 3.43$$

Thus, we would need at least 4 visits to ensure that the sensitivity is at least 95%. Similarly, we have that

$$\text{Specificity} = Pr\left[\overline{X} < 90 \middle| \overline{X} \sim N\left(85, \frac{31.65}{n}\right)\right]$$
$$= \Phi\left(\frac{90 - 85}{\sqrt{\frac{31.65}{n}}}\right)$$
$$= \Phi\left(\frac{5\sqrt{n}}{\sqrt{\frac{31.65}{n}}}\right) = .95$$

This is exactly the same criterion as was used earlier for calculating sensitivity. Therefore, we need at least 4 visits so that the sensitivity and specificity are each at least 95%.

5.38 If a person does not restrict sodium, then his or her urinary sodium excretion will follow a normal distribution with mean $= 160.5$ and sd $= 57.4$. Therefore, the probability that his or her urinary sodium will fall below 100 is given by

$$\Phi\left(\frac{100-160.5}{57.4}\right) = \Phi(-1.054)$$
$$= 1 - \Phi(1.054)$$
$$= 1 - .854 = .146.$$

Thus, there is a 14.6% chance that a noncompliant person will have a urinary sodium < 100 mEq/24 hours.

5.39 If a person is on a sodium-restricted diet, then his or her urinary sodium excretion will follow a normal distribution with mean $= 57.5$ and sd $= 11.3$. In this case, the probability that his or her urinary sodium will be $> 100 =$

$$1 - \Phi\left(\frac{100-57.5}{11.3}\right) = 1 - \Phi(3.76) = .0001.$$

Thus, the probability that a compliant person has a urinary sodium > 100 is negligible (.0001).

5.40 Suppose the new cutoff value $= x$. The misclassification probabilities in Problems 5.38 and 5.39 can be written as $\Phi[(x - 160.5)/57.4]$ and

$$1 - \Phi\left(\frac{x-57.5}{11.3}\right) = \Phi\left(\frac{57.5-x}{11.3}\right),$$

respectively. We want these probabilities to be the same. Thus we set

$$\frac{x-160.5}{57.4} = \frac{57.5-x}{11.3}$$

Solving for x, we obtain $x = 74.4$. Thus, the cutoff value should be 74.4 mEq/24 hours. The misclassification probabilities corresponding to this cutoff are

$$\Phi\left(\frac{74.4-160.5}{57.4}\right) = \Phi(-1.499)$$
$$= 1 - \Phi(1.499)$$
$$= 1 - .9331 = .067$$
$$= \Phi\left(\frac{57.5-74.4}{11.3}\right).$$

Thus, there is a 6.7% probability that people not on a sodium-restricted diet will be designated as compliers, and also a 6.7% probability that people on a sodium-restricted diet will be designated as noncompliers.

5.41 We refer to the MINITAB spreadsheet below. Column C1 gives the range of possible cutoff values. For a cutoff value of c, we have

$$\text{Sensitivity} = Pr[X \le c | X \sim N(57.5,\ 11.3^2)]$$
$$= \Phi\left(\frac{c-57.5}{11.3}\right)$$
$$\text{Specificity} = Pr[X > c | X \sim N(160.5,\ 57.4^2)]$$
$$= 1 - \Phi\left(\frac{c-160.5}{57.4}\right)$$
$$1 - \text{Specificity} = \Phi\left(\frac{c-160.5}{57.4}\right)$$

Using the normal cdf function of MINITAB we compute the sensitivity and (1 – specificity) in columns C2 and C3, respectively. The ROC curve is obtained by plotting the sensitivity (y-axis) vs (1 – specificity) (x-axis) (see Figure below). The area under the ROC curve is calculated from the trapexoidal rule by

$$\text{Area} = \sum_{i=1}^{n}\left(\frac{1}{2}\right)(y_i + y_{i-1})(x_i - x_{i-1})$$

where $(x_0,\ y_0) = (0,\ 0)$ and $(x_n,\ y_n) = (1,\ 1)$.

Note: We have created two extra columns C4 and C5 for this purpose using the Lag function of MINITAB (under the Stat menu, Time Series submenu) which creates a new variable that is the value in a column one row up from the current row. Thus, C4 representing sensitivity at point i-1 was created as the Lag of C2 and C5 representing 1 – specificity at point i-1 was created as the Lag of C3.

Column C6 has the individual areas under the curve between x_i and x_{i-1} while Column C7 has the sum of the individual areas. The total area = .96 and can be interpreted as the probability that for 2 randomly selected subjects, one with a Na restricted diet and one with a standard American diet, the subject with the Na restricted diet has the lower urinary Na.

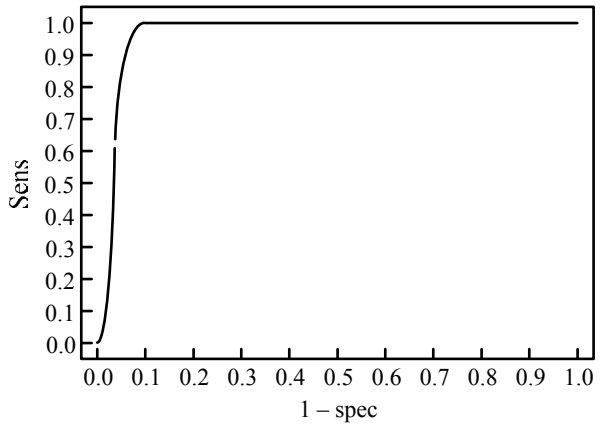
```
Worksheet size: 3500 cells

MTB > Retrieve 'C:\MTBSEW\PR541T.DOC'.
Retrieving worksheet from file: C:\MTBSEW\PR541T.DOC
'Worksheet was saved on 12/24/1999
MTB > Set c1
DATA> 1( 0 : 350 / 5 )1
DATA> End.
MTB > CDF c1 c2;
SUBC> Normal 57.5 11.3.
MTB > CDF c1 c3;
SUBC> Normal 160.5 57.4.
MTB > Lag 1 c2 c4.
MTB > Lag 1 c3 c5.
MTB > Let c6 = (c3-c5)*(c2+c4)/2
MTB > Let c7 = sum(c6)
MTB > Save 'C:\MTBSEW\PR541T.DOC';
SUBC> Replace.
Saving worksheet in file: C:\MTBSEW\PR541T.DOC
* NOTE * Existing file replaced.
MTB > Print c1-c7.
```

ROW	c	sens	1-spec	sens_i-1	1-spe_i1	area_i	area
1	0	0.00000	0.002586	*	*	*	0.960187
2	5	0.00000	0.003374	0.00000	0.002586	0.0000000	
3	10	0.00001	0.004371	0.00000	0.003374	0.0000000	
4	15	0.00008	0.005625	0.00001	0.004371	0.0000001	
5	20	0.00045	0.007188	0.00008	0.005625	0.0000004	
6	25	0.00201	0.009122	0.00045	0.007188	0.0000024	
7	30	0.00747	0.011497	0.00201	0.009122	0.0000113	
8	35	0.02323	0.014393	0.00747	0.011497	0.0000445	
9	40	0.06073	0.017895	0.02323	0.014393	0.0001470	
10	45	0.13432	0.022100	0.06073	0.017895	0.0004101	
11	50	0.25344	0.027109	0.13432	0.022100	0.0009712	
12	55	0.41245	0.033033	0.25344	0.027109	0.0019722	
13	60	0.58755	0.039984	0.41245	0.033033	0.0034757	
14	65	0.74656	0.048080	0.58755	0.039984	0.0054003	
15	70	0.86568	0.057437	0.74656	0.048080	0.0075434	
16	75	0.93927	0.068172	0.86568	0.057437	0.0096874	
17	80	0.97677	0.080392	0.93927	0.068172	0.0117074	
18	85	0.99253	0.094199	0.97677	0.080392	0.0135953	
19	90	0.99799	0.109682	0.99253	0.094199	0.0154089	
20	95	0.99955	0.126911	0.99799	0.109682	0.0172081	
21	100	0.99992	0.145940	0.99955	0.126911	0.0190237	
22	105	0.99999	0.166797	0.99992	0.145940	0.0208563	
23	110	1.00000	0.189486	0.99999	0.166797	0.0226888	
24	115	1.00000	0.213981	1.00000	0.189486	0.0244949	
25	120	1.00000	0.240226	1.00000	0.213981	0.0262450	
26	125	1.00000	0.268134	1.00000	0.240226	0.0279076	
27	130	1.00000	0.297585	1.00000	0.268134	0.0294513	
28	135	1.00000	0.328431	1.00000	0.297585	0.0308456	
29	140	1.00000	0.360492	1.00000	0.328431	0.0320618	
30	145	1.00000	0.393567	1.00000	0.360492	0.0330743	
31	150	1.00000	0.427428	1.00000	0.393567	0.0338610	
32	155	1.00000	0.461832	1.00000	0.427428	0.0344045	
33	160	1.00000	0.496525	1.00000	0.461832	0.0346927	
34	165	1.00000	0.531244	1.00000	0.496525	0.0347190	
35	170	1.00000	0.565727	1.00000	0.531244	0.0344828	
36	175	1.00000	0.599716	1.00000	0.565727	0.0339895	
37	180	1.00000	0.632967	1.00000	0.599716	0.0332503	
38	185	1.00000	0.665248	1.00000	0.632967	0.0322814	
39	190	1.00000	0.696352	1.00000	0.665248	0.0311040	
40	195	1.00000	0.726095	1.00000	0.696352	0.0297430	
41	200	1.00000	0.754322	1.00000	0.726095	0.0282269	
42	205	1.00000	0.780907	1.00000	0.754322	0.0265855	
43	210	1.00000	0.805758	1.00000	0.780907	0.0248505	
44	215	1.00000	0.828811	1.00000	0.805758	0.0230532	
45	220	1.00000	0.850035	1.00000	0.828811	0.0212243	
46	225	1.00000	0.869428	1.00000	0.850035	0.0193930	
47	230	1.00000	0.887014	1.00000	0.869428	0.0175858	
48	235	1.00000	0.902841	1.00000	0.887014	0.0158265	
49	240	1.00000	0.916976	1.00000	0.902841	0.0141357	
50	245	1.00000	0.929506	1.00000	0.916976	0.0125300	
51	250	1.00000	0.940529	1.00000	0.929506	0.0110230	
52	255	1.00000	0.950153	1.00000	0.940529	0.0096239	
53	260	1.00000	0.958492	1.00000	0.950153	0.0083389	
54	265	1.00000	0.965663	1.00000	0.958492	0.0071708	
55	270	1.00000	0.971783	1.00000	0.965663	0.0061199	
56	275	1.00000	0.976966	1.00000	0.971783	0.0051834	
57	280	1.00000	0. 981323	1.00000	0.976966	0.0043572	
58	285	1.00000	0.984958	1.00000	0.981323	0.0036349	
59	290	1.00000	0.987968	1.00000	0.984958	0.0030095	
60	295	1.00000	0.990441	1.00000	0.987968	0.0024728	
61	300	1.00000	0.992457	1.00000	0.990441	0.0020165	
62	305	1.00000	0.994089	1.00000	0.992457	0.0016320	
63	310	1.00000	0.995400	1.00000	0.994089	0.0013108	
64	315	1.00000	0.996445	1.00000	0.995400	0.0010449	
65	320	1.00000	0.997271	1.00000	0.996445	0.0008267	
66	325	1.00000	0.997921	1.00000	0.997271	0.0006490	
67	330	1.00000	0.998426	1.00000	0.997921	0.0005057	
68	335	1.00000	0.998817	1.00000	0.998426	0.0003911	
69	340	1.00000	0.999117	1.00000	0.998817	0.0003002	
70	345	1.00000	0.999346	1.00000	0.999117	0.0002286	
71	350	1.00000	0.999519	1.00000	0.999346	0.0001729	

MTB >

ROC curve



5.42 We have that

$$\begin{aligned} Pr(\text{SBP} \geq 140) &= 1 - Pr(\text{SBP} < 140) \\ &= 1 - \Phi\left(\frac{140 - 120.2}{10.7}\right) \\ &= 1 - \Phi(1.85) \\ &= 1 - .968 = .032 \end{aligned}$$

5.43 $$\begin{aligned} Pr(\text{SBP} \geq 140) &= 1 - Pr(\text{SBP} < 140) \\ &= 1 - \Phi\left(\frac{140 - 123.5}{12.8}\right) \\ &= 1 - \Phi(1.289) \\ &= 1 - .901 = .099 \end{aligned}$$

5.44 Let

$$A = \{\text{SBP} \geq 140 \text{ mm Hg}\},$$
$$B = \{\text{DBP} \geq 90 \text{ mm Hg}\}.$$

We wish to compute $Pr(A \cap \overline{B})$. We are given that $Pr(A \cup B) = .177$. Furthermore, from the addition law of probability we have

$$\begin{aligned} Pr(A \cup B) &= Pr(A) + Pr(B) - Pr(A \cap B) \\ &= Pr(B) + Pr(A \cap \overline{B}) \end{aligned}$$

Therefore, by subtraction we have

$$Pr(A \cap \overline{B}) = Pr(A \cup B) - Pr(B)$$

To compute $Pr(B)$, we have

$$\begin{aligned} Pr(B) &= Pr(\text{DBP} \geq 90) = 1 - Pr(\text{DBP} < 90) \\ &= 1 - \Phi\left(\frac{90 - 78.5}{9.2}\right) \\ &= 1 - \Phi(1.25) = 1 - .8944 \\ &= .106 \end{aligned}$$

Thus, $Pr(A \cap \overline{B}) = .177 - .106 = .071$.

5.45 We wish to compute $Pr(\overline{A} \cap B)$. If we reverse the roles of A and B, we have

$$Pr(\overline{A} \cap B) = Pr(A \cup B) - Pr(A)$$

From Problem 5.43 we have $Pr(A) = .099$. Therefore,

$$Pr(\overline{A} \cap B) = .177 - .099 = .078$$

5.46 The 90th percentile of a standard normal distribution $= 1.28$. Thus, the 90th percentile for diastolic blood pressure (DBP) for 17-year-old boys

$$= 63.7 + 11.4(1.28) = 78.3 \text{ mm Hg.}$$

5.47 Let $X = \text{DBP}$ for 17-year-old boys. We wish to compute $Pr(X \geq 90)$. We have

$$\begin{aligned} Pr(X \geq 90) &= 1 - \Phi\left(\frac{90 - 63.7}{11.4}\right) \\ &= 1 - \Phi(2.307) \\ &= 1 - .9895 = .0105. \end{aligned}$$

Thus, approximately 1% of 17-year-old boys have DBP ≥ 90 mm Hg.

5.48 We wish to compute $Pr(X \geq 25)$ where X is a random variable depicting the number of boys with elevated DBP. X is binomially distributed with parameters $n = 200$ and $p = .10$. We will approximate X by a normal random variable Y with

$$\text{mean} = np = 200(.1) = 20$$

and

$$\text{variance} = npq = 200(.1)(.9) = 18.$$

$$\begin{aligned} Pr(X \geq 25) &\cong Pr(Y \geq 24.5) \\ &= 1 - \Phi\left(\frac{24.5 - 20}{\sqrt{18}}\right) \\ &= 1 - \Phi(1.061) \\ &= 1 - .856 = .144 \end{aligned}$$

Thus, since this probability is $> .05$, there are not an unusually large number of boys with elevated DBP.

5.49 We observe 288 deaths over 90 high-pollution days, whereas we expect 270 deaths if pollution has no effect on mortality. If pollution has no effect on mortality, then the observed number of deaths over 90 days (X) should be Poisson-distributed with parameters, $\lambda = 3.0$ and $T = 90$. Thus, we wish to compute

$$Pr(X \geq 288 | \lambda = 3.0,\ T = 90) = \sum_{k=288}^{\infty} e^{-\mu} \frac{\mu^k}{k!}$$
$$= 1 - \sum_{k=0}^{287} e^{-\mu} \frac{\mu^k}{k!},$$

where

$$\mu = \lambda T = 3(90) = 270 .$$

This is very tedious to compute, and instead we use the normal approximation (Y) to the Poisson distribution, which has mean = variance $= \mu = 270$. We wish to compute

$$Pr(Y \geq 287.5) = 1 - Pr(Y \leq 287.5) = 1 - \Phi\left(\frac{287.5 - 270}{\sqrt{270}}\right)$$
$$= 1 - \Phi\left(\frac{17.5}{16.43}\right) = 1 - \Phi(1.065)$$
$$= 1 - .857 = .143.$$

Since this probability is $> .05$, there are *not* an unusual number of deaths on high-pollution days.

5.50 If the distribution of $X =$ mean plasma testosterone within a group is normally distributed with mean $= \mu$ and variance $= \sigma^2$, then the probability of a high value is given by

$$Pr(X > 10) = 1 - Pr(X < 10)$$
$$= 1 - \Phi\left(\frac{10 - \mu}{\sigma}\right)$$

Thus, referring to Table 5.3, we have Group A:

$$Pr(X > 10) = 1 - \Phi\left(\frac{10 - 10.10}{2.29}\right) = 1 - \Phi\left(\frac{-0.10}{2.29}\right)$$
$$= 1 - \Phi(-0.044) = \Phi(0.044) = .517$$

Group B:

$$Pr(X > 10) = 1 - \Phi\left(\frac{10 - 8.36}{2.36}\right) = 1 - \Phi\left(\frac{1.64}{2.36}\right)$$
$$= 1 - \Phi(0.695) = 1 - .756 = .244$$

Group C:

$$Pr(X > 10) = 1 - \Phi\left(\frac{10 - 5.99}{1.20}\right) = 1 - \Phi\left(\frac{4.01}{1.20}\right)$$
$$= 1 - \Phi(3.342) = 1 - .9996 = .0004$$

5.51 From Figure 5.1, we see that there are 7 high values in Group A, 3 high values in group B, and 0 high values in group C. The expected number of high values if the distribution of mean plasma testosterone within a group is normal are as follows:

group A: $12 \times .517 = 6.2$;
group B: $12 \times .244 = 2.9$;
group C: $12 \times .0004 = .005$.

Thus, the observed and expected number of high values agree very closely, which implies that a normal distribution is fitting the data well within specific groups.

5.52 The coefficient of variation (CV) is defined as $100\% \times s/\bar{x}$. In this case, s refers to the within-person standard deviation over the 3 values for a given person. Also $\bar{x} =$ overall mean in the study

$$\frac{5.99 + 8.36 + 10.10}{3} = 8.15 .$$

Thus, $CV = 100\% \times 1.32/8.15 = 16.2\%$.

5.53 We use Equation 5.10. We have $L = (X_1 + X_2)/2$, where $X_1 =$ IOP, OD and $X_2 =$ IOP, OS. Therefore,

$$\text{Var}(L) = \left(\frac{1}{4}\right)\text{Var}(X_1) + \left(\frac{1}{4}\right)\text{Var}(X_2)$$
$$= \left(\frac{1}{4}\right)(16) + \left(\frac{1}{4}\right)(16) = 8.$$

5.54 From Definition 5.13 we have:

$$Cov(X_1,\ X_2) = \rho\ \text{sd}(X_1)\text{sd}(X_2)$$
$$= .8(4)(4) = 12.8.$$

5.55 From Equation 5.12 (or more simply Equation 5.11), we have

$$\text{Var}(0.5X_1 + 0.5X_2) = 0.25\text{Var}(X_1) + 0.25\text{Var}(X_2)$$
$$+2(0.5)(0.5)Cov(X_1,\ X_2)$$
$$= 0.25(4)^2 + 0.25(4)^2 + 0.5(12.8)$$
$$= 14.4$$

Notice how the variance of IOP OU under the assumption of correlated IOP values (Variance $= 14.4$) is much larger than the variance under the assumption of independence (From Problem 5.53, Variance $= 8$) and is almost as large as the variance of an IOP value from a single eye (Variance $= 16$).

REFERENCES

[1] Berson, E. L., Sandberg, M. A., Rosner, B., Birch, D. G., & Hanson, A. H. (1985). Natural course of retinitis pigmentosa over a three-year interval. *American Journal of Ophthalmology*, **99**, 240–251.

[2] Boston Globe, April 25, 1980.

[3] Speizer, F. E., Doll, R., Heaf, P., & Strang, L. B. (1968, February 8). Investigation into use of drugs preceding death from asthma. *British Medical Journal*, **1**, 339–343.

[4] Weissfeld, J. L., Johnson, E. H., Brock, B. M., Hawthorne, V. M. (1988, September). Sex and age interactions in the association between alcohol and blood pressure. *American Journal of Epidemiology*, **128**(3), 559–569.

[5] Report of the Second Task Force on Blood Pressure Control in Children—1987. (1987, January). National Heart, Lung and Blood Institute, Bethesda, Maryland. *Pediatrics*, **79**(l), 1–25.

[6] Schwartz, J., & Dockery, D. W. (1992). Particulate air pollution and daily mortality in Steubenville, Ohio. *American Journal of Epidemiology*, **135**(l), 12-19.

[7] Ehrenkranz, J., Bliss, E., & Sheard, M. H. (1974). Plasma testosterone: Correlation with aggressive behavior and dominance in men. *Psychosomatic Medicine*, **36**(6), 469–475.