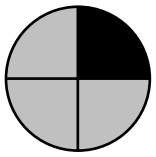
Chapter 3

PROBABILITY

REVIEW OF KEY CONCEPTS

SECTION 3.1 Frequency Definition of Probability

What is probability? Suppose we want to set up a test for color blindness. We use a plate divided into four quadrants. One of four color quadrants has a particular color. The other three quadrants are of the same color, but different from the first quadrant.



Thus, the probability that a color-blind person will pick the correct quadrant at random = 1/4. What does "at random" mean? Suppose the test is performed by many color-blind people and the following results are obtained:

Number of color-blind people	% correct
20	6/20 correct = .30
100	24/ 100 correct = .24
1000	255/1000 correct = .255

As the number of color-blind people taking the test is increased, the proportion of correct trials will approach a number p (in this case, .25), which we call the *probability*. This is the *frequency definition of probability*.

All probabilities must be between 0 and 1. Probabilities are defined over events. Two events are mutually exclusive if they cannot occur at the same time. Probabilities of mutually exclusive events must add; e.g., suppose we repeat the test 4 times for a single color-blind person. Let the event $E_1 =$ exactly 1 out of 4 correct, $E_2 =$ exactly 2 out of 4 correct:

$$Pr(E_1) + Pr(E_2) = Pr(E_1 \text{ or } E_2) = Pr(E_1 \cup E_2)$$

E_1 and E_2 are mutually exclusive events because they cannot occur at the same time.

SECTION 3.2 Multiplication Law of Probability

When can we multiply probabilities? Let the events A, B be defined by

$$A = \text{1st selection is correct}$$
$$B = \text{2nd selection is correct}$$

Then $Pr(A \cap B) =$ probability both selections are correct $= Pr(A) \times Pr(B) = 1/16$. Two probabilities can be multiplied if the events are *independent*.

Consider another example: suppose we have a group of 6-month-old children with two normal ears at their routine 6-month checkup. Suppose there is a 10% chance that a child will have fluid in the middle ear at an exam 1 month later in a specific ear, while the probability that both ears are affected (called "bilateral middle-ear effusion") is .07. Are the ears independent? No, because

$$Pr \text{ (bilateral middle-ear effusion)} = .07 > .1 \times .1 = .01$$

This is an example of *dependent events*. The middle-ear status of both ears of the same child are dependent events, because there is often a common reason why both ears get infected at the same time (e.g., exposure of the child to other affected children in a day-care center).

SECTION 3.3 Addition Law of Probability

Let A = right ear affected, B = left ear affected. What is $Pr(A \cup B) = Pr(\text{either ear affected})$?

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

This is known as the addition law of probability. For the *ear* example,

$$Pr \text{ (either ear affected)} = .1 + .1 - .07 = .13:$$

13% have at least one ear affected
7% have bilateral middle-ear effusion (both ears affected)
6% have unilateral middle-ear effusion (only one ear affected)

For the *color plate* example, A = (1st selection correct), B = (2nd selection correct)

$$\begin{aligned} Pr(A \cup B) &= Pr(\text{at least 1 of 2 selections are correct}) \\ &= Pr(A) + Pr(B) - Pr(A \cap B) \\ &= Pr(A) + Pr(B) - Pr(A) \times Pr(B) \\ &= \frac{1}{4} + \frac{1}{4} - \left(\frac{1}{4}\right)^2 = \frac{7}{16} \end{aligned}$$

SECTION 3.4 Conditional Probability

The conditional probability of B given A is defined as $Pr(B \cap A)/Pr(A)$ and is denoted by $Pr(B|A)$. It corresponds to the proportion of times that B occurs among the subset of occasions when A occurs. For the

ear example, let A = right ear affected, B = left ear affected, and $\overline{A}$ = the event that the right ear is *not* affected.

$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} = \frac{.07}{.10} = 70\% = \text{conditional probability of } B \text{ given } A$$

$$Pr(B|\overline{A}) = \frac{Pr(B \cap \overline{A})}{Pr(\overline{A})} = \frac{Pr(B) - Pr(A \cap B)}{.90} = \frac{.10 - .07}{.90} = \frac{.03}{.90} = \frac{1}{30} \approx .03$$

In words, $Pr(B|A)$ = probability that the left ear is affected given that the right ear is affected = 70%; $Pr(B|\overline{A})$ = probability that the left ear is affected given that the right ear is *not* affected = 3%. Stated another way, among children whose right ear is affected, 70% also have an affected left ear. Similarly, among children whose right ear is not affected, only 3% have an affected left ear.

3.4.1 Relative Risk

The Relative Risk of B given A is defined as $Pr(B|A)/Pr(B|\overline{A})$. For the ear example, if A = right ear affected and B = left ear affected, then

$$\text{Relative risk} = RR = \frac{Pr(B|A)}{Pr(B|\overline{A})} = \frac{\frac{7}{10}}{\frac{1}{30}} = 21$$

The left ear is 21 times as likely to be affected if the right ear is affected than if the right ear is unaffected.

There was an outbreak of Legionnaire's disease in Austin, Minnesota in 1957. Subsequent investigation focused on employment at a meat-packing plant as a possible cause. The illness rate per 1000 subjects among all adults in the town is given in the following table:

Employment status	%	Total	Number ill	Illness rate per 1000	*RR*
Employed at meat-packing plant	19	4,718	46	9.7	6.1
Not employed at meat-packing plant	81	19,897	32	1.6	
	100	24,615	78	3.2	

The relative risk $(RR) = 9.7/1.6 = 6.1$, indicating that employees at the meat packing plant were six times more likely to get Legionnaire's disease than persons not employed at the plant.

If two events are independent, then $Pr(B|A) = Pr(B|\overline{A}) = Pr(B)$ and $RR = 1$. For the color plate example, let A = 1st selection correct, B = 2nd selection correct; $Pr(B|\overline{A}) = Pr(B) = Pr(B|A) = 1/4$ and $RR = 1$. Thus, the probability that the 2nd selection is correct $= 1/4$ regardless of whether the 1st selection is correct or not.

SECTION 3.5 Total Probability Rule

The total probability rule specifies the relationship between conditional and unconditional probabilities:

$$Pr(B) = Pr(B|A)Pr(A) + Pr(B|\overline{A})Pr(\overline{A})$$

In words, the unconditional probability of B is a weighted average of the conditional probabilities of B when A occurs and when A does not occur, where the weights are $Pr(A)$ and $Pr(\overline{A})$, respectively. For

example, in the case of Legionnaire's disease: A = work at meat-packing plant, B = Legionnaire's disease. Suppose $Pr(A) = .19$.

$$Pr(B) = \frac{3.2}{1000} = .19 \times \frac{9.7}{1000} + .81 \times \frac{1.6}{1000}$$

SECTION 3.6 Sensitivity, Specificity, Predictive Values of Screening Tests

The angiogram is the standard test used to diagnose the occurrence of stroke. However, some patients experience side effects from this test, and some investigators have attempted to use a noninvasive test as an alternative. Sixty-four patients with transient monocular blindness, or TMB (where a person temporarily loses vision in one eye), were given both tests. The sample was selected to have about equal numbers of angiogram-positive and -negative patients. The results were as follows:

Angiogram	Noninvasive test	*n*
–	–	21
–	+	8
+	–	3
+	+	32
		64

How can we compare the two tests? Sensitivity, specificity and predictive value (positive and negative) are commonly used measures for describing the accuracy of screening tests. If we assume that the angiogram is the gold standard, then

$$\textit{Sensitivity} \text{ is defined as } Pr(\text{test } + | \text{true } +) = \frac{32}{35} = .914$$

$$\textit{Specificity} \text{ is defined as } Pr(\text{test } - | \text{true } -) = \frac{21}{29} = .724$$

We would like to convert sensitivity and specificity into predictive values:

$$\text{Predictive value positive } (PV+) \text{ is defined as } Pr(\text{true } + | \text{test } +)$$

It can be shown that

$$PV+ = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

where prevalence is the proportion of true positives. Assume the prevalence of strokes is 20% among TMB patients:

$$PV+ = \frac{.914 \times .20}{0.914(.20) + 0.276(.80)} = \frac{.1829}{.1829 + .2207} = \frac{.1829}{.4035} = .453$$

$$\text{Predictive value negative } (PV-) \text{ is defined as } Pr(\text{true} - | \text{test} -)$$

It can be shown that

$$PV- = \frac{\text{specificity} \times (1 - \text{prevalence})}{\text{specificity} \times (1 - \text{prevalence}) + (1 - \text{sensitivity})\text{prevalence}}$$
$$= \frac{.724(.80)}{.724(.80) + .086(.20)} = \frac{.5793}{.5793 + .0171} = \frac{.5793}{.5965} = .971$$

3.6.1 ROC Curves

Sometimes the criteria for designating a subject as positive on a screening test is arbitrary. To summarize the accuracy of the screening test we vary the cutpoint denoting positivity and calculate sensitivity and specificity for different cutpoints. The accuracy of the test can be displayed graphically by plotting the sensitivity vs 1-specificity for each possible cutpoint. The resulting curve is called an ROC curve (or receiver operating characteristic curve). The area under the curve can be shown to be a good measure of the overall accuracy of the test. To interpret the area under the ROC curve, if low values correspond to poorer outcome, and we pick a random affected subject and a random normal subject, then the area under the ROC curve = the probability that the affected subject will have a lower score than the normal subject.

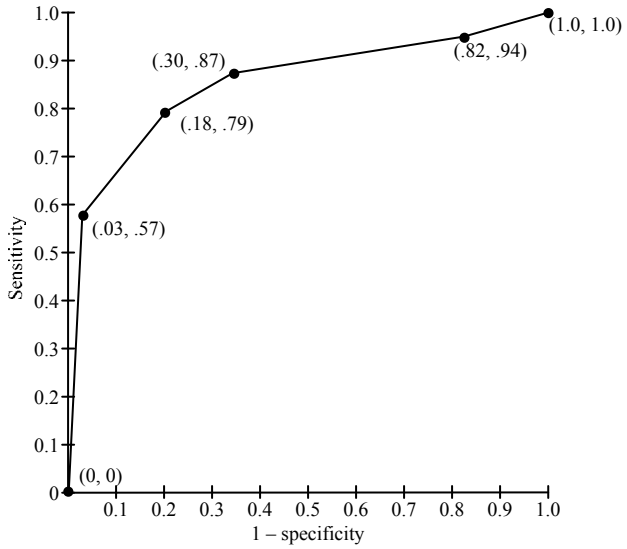
For example, in one study, 4 participating readers used two different types of film, PACS film and plain film to evaluate abnormality based on radiographic images. There was a 5 point rating scale with a lower score indicating abnormality. The issue is what cutpoint to use to designate abnormality. The results are given below for reader 1 for PACS film [1]:

	Score					
True status	1	2	3	4	5	Total
Negative	1	5	4	17	6	33
Positive	38	15	5	5	4	67
Total	39	20	9	22	10	100

To obtain an ROC curve, we consider different cutpoints for determining abnormality. Suppose we use a criterion of ≤ 1 to designate abnormal and ≥ 2 to designate normal. The sensitivity will be 38/67 = .57 and the specificity will be 32/33 = .97. For each of the possible cutpoints we have

Cutpoint for Abnormality	Sensitivity	Specificity
≤ 0	.00	1.00
≤ 1	.57	.97
≤ 2	.79	.82
≤ 3	.87	.70
≤ 4	.94	.18
≤ 5	1.00	.00

The resulting ROC curve is shown below.



The area under the ROC curve = .855. Thus, a randomly selected affected person will have a lower score than a randomly affected normal person about 86% of the time. *Note*: if the affected and normal person have the same score, then this outcome is counted as 1/2 of a success in calculating the proportion of affected-normal pairs where the affected person has a lower score.

SECTION 3.7 Bayes' Theorem

The determination of predictive value positive and negative is a particular application of a more general principle (*Bayes' theorem*). If A = symptom(s) and B = disease, then

$$Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A|B)Pr(B) + Pr\left(A\middle|\overline{B}\right)Pr\left(\overline{B}\right)}$$

More generally, if there are k disease states such that each person has one and only one disease state (which could include being normal),

$$Pr(B_i|A) = \frac{Pr(A|B_i)Pr(B_i)}{\sum_{j=1}^{k} Pr(A|B_j)Pr(B_j)}, \; i = 1, \ldots, k$$

where $B_i = i$th disease state and A = symptom(s). Bayes' theorem is used to compute the probability of different disease states given the occurrence of one or more symptoms. To use Bayes' theorem, we need the prevalence of each of the disease states $(Pr(B_i))$ as well as how frequently different symptoms occur among patients in a given disease state $(Pr(A|B_i))$.

PROBLEMS

Let A = {serum cholesterol = 250–299},
B = {serum cholesterol ≥ 300},
C = {serum cholesterol ≤ 280}.

3.1 Are the events A and B mutually exclusive?

3.2 Are the events A and C mutually exclusive?

3.3 Suppose $Pr(A) = .2$, $Pr(B) = .1$. What is

Pr(serum cholesterol ≥ 250)?

3.4 What does $A \cup C$ mean?

3.5 What does $A \cap C$ mean?

3.6 What does $B \cup C$ mean?

3.7 What does $B \cap C$ mean?

3.8 Are the events B and C mutually exclusive?

3.9 What does the event $\overline{B}$ mean? What is its probability?

Suppose that the gender of successive offspring in the same family are independent events and that the probability of a male or female offspring is .5.

3.10 What is the probability of two successive female offspring?

3.11 What is the probability that exactly one of two successive children will be female?

3.12 Suppose that three successive offspring are male. What is the probability that a fourth child will be male?

Cardiovascular Disease
A survey was performed among people 65 years of age and older who underwent open-heart surgery. It was found that 30% of patients died within 90 days of the operation, whereas an additional 25% of those who survived 90 days died within 5 years after the operation.

3.13 What is the probability that a patient undergoing open-heart surgery will die within 5 years?

3.14 What is the mortality incidence (per patient month) in patients receiving this operation in the first 90 days after the operation? (Assume that 90 days = 3 months.)

3.15 Answer the same question as in Problem 3.14 for the period from 90 days to 5 years after the operation.

3.16 Can you tell if the operation prolongs life from the data presented? If not, then what additional data do you need?

A study relating smoking history to several measures of cardiopulmonary disability was recently reported [2]. The data in Table 3.1 were presented relating the number of people with different disabilities according to cigarette-smoking status.

Table 3.1 Number of people with selected cardiopulmonary disabilities versus cigarette-smoking status

	Cigarette-smoking status			
Disability	None (n = 656)	Ex (n = 826)	Current < 15 g/day (n = 955)	Current ≥ 15 g/day (n = 654)
Shortness of breath	7	15	18	13
Angina	15	19	19	16
Possible infarction	3	7	8	6

3.17 What is the prevalence of angina among light current smokers (< 15g/day)?

3.18 What is the relative risk of ex-smokers, light current smokers, and heavy current smokers, respectively, for shortness of breath as compared with nonsmokers?

3.19 Answer Problem 3.18 for angina.

3.20 Answer Problem 3.18 for possible infarction.

Pulmonary Disease
Pulmonary embolism is a relatively common condition that necessitates hospitalization and also often occurs in patients hospitalized for other reasons. An oxygen tension (arterial Po_2) < 90 mm Hg is one of the important criteria used in diagnosing this condition. Suppose that the sensitivity of this test is 95%, the specificity is 75%, and the estimated prevalence is 20% (i.e., a doctor estimates that a patient has a 20% chance of pulmonary embolism before performing the test).

3.21 What is the predictive value positive of this test? What does it mean in words?

3.22 What is the predictive value negative of this test? What does it mean in words?

3.23 Answer Problem 3.21 if the estimated prevalence is 80%.

3.24 Answer Problem 3.22 if the estimated prevalence is 80%.

Environmental Health, Pediatrics

3.25 Suppose that a company plans to build a lead smelter in a community and that the city council wishes to assess the health effects of the smelter. In particular, there is concern from previous literature that children living very close to the smelter will experience unusually high rates of lead poisoning in the first 3 years of life. The projected rates of lead poisoning over this time period are 50 per 100,000 for those children living within 2 km of the smelter, 20 per 100,000 for children living > 2 km but ≤ 5 km from the smelter, and 5 per 100,000 for children living > 5 km from the smelter. If 80% of the children live more than 5 km from the smelter, 15% live > 2 km but ≤ 5 km from the smelter, and the remainder live ≤ 2 km from the smelter, then what is the overall probability that a child from this community will get lead poisoning?

Diabetes

The prevalence of diabetes in adults at least 20 years old has been studied in Tecumseh, Michigan [3]. The age-sex-specific prevalence (per 1000) is given in Table 3.2.

Table 3.2 Age-sex-specific prevalence of diabetes in Tecumseh, MI (per 1000)

	Sex	
Age group (years)	Male	Female
20–39	5	7
40–54	23	31
55+	57	89

Source: Reprinted with permission from the *American Journal of Epidemiology*, 116(6), 971–980.

3.26 Suppose we plan a new study in a town that consists of 48% males and 52% females. Of the males, 40% are ages 20–39, 32% are 40–54, and 28% are 55+. Of the females, 44% are ages 20–39, 37% are 40–54, and 19% are 55+. Assuming that the Tecumseh prevalence rates hold, what is the expected prevalence of diabetes in the new study?

3.27 What proportion of diabetics in the new study would be expected in each of the six age-sex groups?

Cancer

Table 3.3 shows the annual incidence rates for colon cancer, lung cancer, and stomach cancer in males ages 50 years and older from the Connecticut Tumor Registry, 1963–1965 [4].

Table 3.3 Average annual incidence per 100,000 males for colon, lung, and stomach cancer from the Connecticut Tumor Registry 1963–1965

	Ages		
Type of cancer	50–54	55–59	60–64
Colon	35.7	60.3	98.9
Lung	76.1	137.5	231.7
Stomach	20.8	39.1	46.0

Source: Reprinted from *Cancer Incidence in Five Continents* II, 1970, with permission of Springer-Verlag, Berlin.

3.28 What is the probability that a 57-year-old, disease-free male will develop lung cancer over the next year?

3.29 What is the probability that a 55-year-old, disease-free male will develop colon cancer over the next 5 years?

3.30 Suppose there is a cohort of 1000 50-year-old men who have never had cancer. How many colon cancers would be expected to develop in this cohort over a 15-year period?

3.31 Answer Problem 3.30 for lung cancer.

3.32 Answer Problem 3.30 for stomach cancer.

Cardiovascular Disease

The relationship between physical fitness and cardiovascular-disease mortality was recently studied in a group of railroad working men, ages 22–79 [5]. Data were presented relating baseline exercise-test heart rate and coronary heart-disease mortality (Table 3.4).

Table 3.4 Relationship between baseline exercise-test heart rate and coronary heart-disease mortality

Exercise-test heart rate (beats/min)	Coronary heart-disease mortality (20 years) (per 100)
≤ 105	9.1
106–115	8.7
116–127	11.6
> 127	13.2

Suppose that 20, 30, 30, and 20% of the population, respectively, have exercise-test heart rates of ≤ 105, 106–115, 116–127, > 127 beats/minute. Suppose a test is positive if the exercise-test heart rate is > 127 beats/min and negative otherwise.

3.33 What is the probability of a positive test among men who have died over the 20-year period? Is there a name for this quantity?

3.34 What is the probability of a positive test among men who survived the 20-year period? Is there a name for this quantity?

3.35 What is the probability of death among men with a negative test? Is there a name for this quantity?

Cardiovascular Disease
Exercise testing has sometimes been used to diagnose patients with coronary-artery disease. One test criterion that has been used to identify those with disease is the abnormal ejection-fraction criterion; that is, an absolute rise of less than 0.05 with exercise. The validity of this noninvasive test was assessed in 196 patients versus coronary angiography, the gold standard, a procedure that can unequivocally diagnose the disease but the administration of which carries some risk for the patient. A sensitivity of 79% and a specificity of 68% were found for the exercise test in this group.

3.36 What does the sensitivity mean in words in this setting?

3.37 What does the specificity mean in words in this setting?

Suppose a new patient undergoes exercise testing and a physician feels before seeing the exercise-test results that the patient has a 20% chance of having coronary-artery disease.

3.38 If the exercise test is positive, then what is the probability that such a patient has disease?

3.39 Answer Problem 3.38 if the exercise test is negative.

Nutrition
The food-frequency questionnaire (FFQ) is a commonly used method for assessing dietary intake, where individuals are asked to record the number of times per week they usually eat for each of about 100 food items over the previous year. It has the advantage of being easy to administer to large groups of people, but has the disadvantage of being subject to recall error. The gold standard instrument for assessing dietary intake is the diet record (DR), where people are asked to record each food item eaten on a daily basis over a 1-week period. To investigate the accuracy of the FFQ, both the FFQ and the DR were administered to 173 participants in the United States. The reporting of alcohol consumption with each instrument is given in Table 3.5, where alcohol is coded as alcohol = some drinking, versus no alcohol = no drinking.

Table 3.5 Actual drinking habits as determined from the diet record cross-classified by self-reported drinking status from the food-frequency questionnaire

	Food-frequency questionnaire		
Diet record	Alcohol	No alcohol	Total
Alcohol	138	18	156
No alcohol	1	16	17
Total	139	34	173

3.40 What is the sensitivity of the FFQ?

3.41 What is the specificity of the FFQ?

Let us treat the 173 participants in this study as representative of people who would use the FFQ.

3.42 What is the predictive value positive of the FFQ?

3.43 What is the predictive value negative of the FFQ?

3.44 Suppose the questionnaire is to be administered in a different country where the true proportion of drinkers is 80%. What would be the predictive value positive if administered in this setting?

Genetics
Two healthy parents have a child with a severe autosomal recessive condition that cannot be identified by prenatal diagnosis. They realize that the risk of this condition for subsequent offspring is 1/4, but wish to embark on a second pregnancy. During the early stages of the pregnancy, an ultrasound test determines that there are twins.

3.45 Suppose that there are monozygotic, or MZ (identical) twins. What is the probability that both twins are affected? one twin affected? neither twin affected? Are the outcomes for the two MZ twins independent or dependent events?

3.46 Suppose that there are dizygotic, or DZ (fraternal) twins. What is the probability that both twins are affected? one twin affected? neither twin affected? Are the outcomes for the two DZ twins independent or dependent events?

3.47 Suppose there is a 1/3 probability of MZ twins and a 2/3 probability of DZ twins. What is the overall probability that both twins are affected? One twin affected? Neither affected?

3.48 Suppose we learn that both twins are affected but don't know whether they are MZ or DZ twins. What is the probability that they are MZ twins given this additional information?

Cerebrovascular Disease
Atrial fibrillation (AF) is a common cardiac condition in the elderly (e.g., former President George H.W. Bush has this condition) characterized by an abnormal heart rhythm that greatly increases the risk of stroke. The following estimates of the prevalence rate of AF and the incidence rate of stroke for people with and without AF by age from the Framingham Heart Study are given in Table 3.6 [6].

3.49 What does an incidence rate of 48.9 strokes per 1000 person-years among 70–79-year-olds with AF mean in Table 3.6?

3.50 What is the relative risk of stroke for people with AF compared with people without AF in each age group? Does the relative risk seem to be the same for different age groups?

Table 3.6 Relationship between atrial fibrillation and stroke

		Incidence rate of stroke per 1000 person-years	
Age group	Prevalence of AF (%)	No AF	AF
60-69	1.8	4.5	21.2
70-79	4.7	9.0	48.9
80-89	10.2	14.3	71.4

Suppose we screen 500 subjects from the general population of age 60–89, of whom 200 are 60–69, 200 are 70–79, and 100 are 80–89.

3.51 What is the incidence rate of stroke in the screened population over a 1-year period?

3.52 Suppose the study of 500 subjects is a "pilot study" for a larger study. How many 60-89 year old subjects need to be screened if we wish to observe an average of 50 strokes in the larger study over 1 year?

Radiology

It is well-known that there is variation among readers in evaluating radiologic images. For this purpose, we present evaluations from a 2nd reader of the same 100 images for the study described in Section 3.6.1 of the Study Guide. The results are given in Table 3.7.

Table 3.7 Ratings from Reader 2 using PACS film for the study described in Section 3.6.1

	Score					
True Status	1	2	3	4	5	Total
Negative	0	3	7	14	9	33
Positive	0	49	11	4	3	67
Total	0	52	18	18	12	100

Suppose different cutpoints are considered for positivity, viz, ≤ 0, ≤ 1, ≤ 2, ≤ 3, ≤ 4 and ≤ 5.

3.53 Compute the sensitivity of the test for each cutpoint for Reader 2.

3.54 Compute the specificity of the test for each cutpoint for Reader 2.

3.55 Draw the ROC curve for Reader 2.

3.56 Compare the accuracy of the test for Reader 2 vs Reader 1.

SOLUTIONS

3.1 Yes

3.2 No

3.3 .3

3.4 $A \cup C =$ {serum cholesterol ≤ 299}

3.5 $A \cap C =$ {$250 \le$ serum cholesterol ≤ 280}

3.6 $B \cup C =$ {serum cholesterol ≤ 280 or ≥ 300}

3.7 $B \cap C$ is the empty set; that is, it can never occur.

3.8 Yes

3.9 $\overline{B} =$ {serum cholesterol < 300}. $Pr(\overline{B}) = .9$

3.10 Let $A_1 =$ {1st offspring is a male}, $A_2 =$ {2nd offspring is a male}. $Pr(\overline{A}_1 \cap \overline{A}_2) = .5 \times .5 = .25$

3.11 $Pr(\overline{A}_1 \cap A_2) + Pr(A_1 \cap \overline{A}_2) = .5 \times .5 + .5 \times .5 = .50$

3.12 The probability = .5, because the sex of successive offspring are independent events.

3.13 Probability $= .30 + (1 - .30) \times .25 = .475$

3.14 10% per patient-month.

3.15 The mortality incidence per month = .25/57 months = 0.44% per patient-month.

3.16 No. For comparison, mortality data on a control group of patients with the same clinical condition as the patients who underwent open-heart surgery, but who did not have the operation are needed.

3.17 The prevalence of angina among light smokers

$$19/955 = .020.$$

3.18 Relative risk of shortness of breath for

$$\text{Ex-smokers vs. nonsmokers} = \frac{\frac{15}{826}}{\frac{7}{656}} = \frac{.0182}{.0107} = 1.7$$

$$\text{Light smokers vs. nonsmokers} = \frac{\frac{18}{955}}{\frac{7}{656}} = \frac{.0188}{.0107} = 1.8$$

$$\text{Heavy smokers vs. nonsmokers} = \frac{\frac{13}{654}}{\frac{7}{656}} = \frac{.0199}{.0107} = 1.9$$

3.19 Relative risk of angina for

$$\text{Ex-smokers vs. nonsmokers} = \frac{\frac{19}{826}}{\frac{15}{656}} = \frac{.0230}{.0229} = 1.0$$

$$\text{Light smokers vs. nonsmokers} = \frac{\frac{19}{955}}{\frac{15}{656}} = \frac{.0199}{.0229} = .9$$

$$\text{Heavy smokers vs. nonsmokers} = \frac{\frac{16}{654}}{\frac{15}{656}} = \frac{.0245}{.0229} = 1.1$$

3.20 Relative risk of possible infarction for

$$\text{Ex-smokers vs. nonsmokers} = \frac{\frac{7}{826}}{\frac{3}{656}} = \frac{.0085}{.0046} = 1.9$$

$$\text{Light smokers vs. nonsmokers} = \frac{\frac{8}{955}}{\frac{3}{656}} = \frac{.0084}{.0046} = 1.8$$

$$\text{Heavy smokers vs. nonsmokers} = \frac{\frac{6}{654}}{\frac{3}{656}} = \frac{.0092}{.0046} = 2.0$$

3.21 We have that

$$PV+ = \frac{(x)(\text{sensitivity})}{(x)(\text{sensitivity}) + (1-x)(1-\text{specificity})}$$

where x = prevalence. Thus, the $PV+$ is given by

$$PV+ = \frac{.20 \times .95}{.20 \times .95 + .80 \times .25} = \frac{.19}{.39} = .487$$

It means that if a patient has a depressed arterial oxygen tension, then there is approximately a 50% chance that she will have a pulmonary embolism.

3.22 We have that

$$PV- = \frac{(1-x)(\text{specificity})}{(1-x)(\text{specificity}) + (x)(1-\text{sensitivity})} = \frac{.80 \times .75}{.80 \times .75 + .20 \times .05} = \frac{.60}{.61} = .984$$

It means that if a patient does not have a depressed arterial oxygen tension, then there is a 98.4% chance that she will *not* have a pulmonary embolism.

3.23 We have

$$PV+ = \frac{.80 \times .95}{.80 \times .95 + .20 \times .25} = \frac{.76}{.81} = .938$$

3.24 We have

$$PV- = \frac{.20 \times .75}{.20 \times .75 + .80 \times .05} = \frac{.15}{.19} = .789$$

Thus, with a higher prevalance, the $PV+$ increases while the $PV-$ decreases.

3.25 Let A = {child gets lead poisoning}

B_1 = {child lives $\leq$ 2 km from the smelter}
B_2 = {child lives $>$ 2 but $\leq$ 5 km from the smelter}
B_3 = {child lives $>$ 5 km from the smelter}

We can write

$$\begin{aligned} Pr(A) &= Pr(A \cap B_1) + Pr(A \cap B_2) + Pr(A \cap B_3) \\ &= Pr(A|B_1)Pr(B_1) + Pr(A|B_2)Pr(B_2) \\ &\quad + Pr(A|B_3)Pr(B_3) \end{aligned}$$

We are given that

$$Pr(A|B_1) = \frac{50}{10^5},\ Pr(A|B_2) = \frac{20}{10^5},\ Pr(A|B_3) = \frac{5}{10^5},$$
$$Pr(B_1) = .05,\ Pr(B_2) = .15 \text{ and } Pr(B_3) = .80.$$

Therefore, it follows that

$$\begin{aligned} Pr(A) &= \left(\frac{50}{10^5}\right)(.05) + \left(\frac{20}{10^5}\right)(.15) + \left(\frac{5}{10^5}\right)(.80) \\ &= \frac{9.5}{10^5} = .000095 \end{aligned}$$

3.26 Let A = {diabetes},

B_1 = {20–39-year-old male},
B_2 = {40–54-year-old male},
B_3 = {55+-year-old male},
B_4 = {20–39-year-old female},
B_5 = {40–54-year-old female},
B_6 = {55+-year-old female}.

We are given

$$Pr(A|B_1) = .005,\ Pr(A|B_2) = .023,\ Pr(A|B_3) = .057,$$
$$Pr(A|B_4) = .007,\ Pr(A|B_5) = .031,\ Pr(A|B_6) = .089.$$

First, we compute the probabilities of the events $B_1, \ldots, B_6$.

$$Pr(B_1) = Pr(20-39|\text{male}) \times Pr(\text{male}) = .40 \times .48 = .192$$
$$Pr(B_2) = Pr(40-54|\text{male}) \times Pr(\text{male}) = .32 \times .48 = .154$$
$$Pr(B_3) = Pr(55+|\text{male}) \times Pr(\text{male}) = .28 \times .48 = .134$$
$$Pr(B_4) = Pr(20-39|\text{female}) \times Pr(\text{female}) = .44 \times .52 = .229$$
$$Pr(B_5) = Pr(40-54|\text{female}) \times Pr(\text{female}) = .37 \times .52 = .192$$
$$Pr(B_6) = Pr(55+|\text{female}) \times Pr(\text{female}) = .19 \times .52 = .099$$

We now use the total probability rule as follows:

$$Pr(A) = \sum_{i=1}^{6} Pr(A|B_i) \times Pr(B_i) = .005 \times .192 + \cdots + .089 \times .099 = .029$$

Thus, the expected prevalence in the new study is 2.9%.

3.27 20–39 M: .034; 40–54 M: .124; 55+ M: .269; 20–39 F: .056; 40–54 F: .209; 55+ F: .308. They are obtained by dividing $Pr(A \cap B_i)$ by $Pr(A)$ for each age-sex group $(i = 1, \ldots, 6)$. For example, for 20–39 year old males, $Pr(A \cap B)$,

$$Pr(B_i|A) = \frac{Pr(B_i \cap A)}{Pr(A)} = \frac{Pr(A|B_i)Pr(B_i)}{Pr(A)} = \frac{.005(.192)}{.029} = .034, \text{ etc.}$$

3.28 $137.5/10^5 = .001375$

3.29
$$Pr(\text{not developing colon cancer over 5 years}) = Pr(\text{not developing colon cancer over the 1st year}) \times Pr(\text{not developing colon cancer over the 2nd year}|\text{disease free after 1 year}) \times \cdots \times Pr(\text{not developing colon cancer over the 5th year}|\text{disease free after 4 years}) = \left(1 - \frac{60.3}{10^5}\right)^5 = .99699$$

Thus,

$$Pr(\text{developing colon cancer over 5 years}) = 1 - .99699 = .00301 = 301 \text{ per } 10^5$$

3.30 By the same rationale as in Problem 3.29,

$$Pr(\text{not developing colon cancer over 15 years}) = Pr(\text{not developing colon cancer from age 50 to 54}) \times Pr(\text{not developing colon cancer from age 55 to 59}|\text{disease free after 5 years}) \times Pr(\text{not developing colon cancer from age 60 to 64}|\text{disease free after 10 years}) = (1 - 35.7/10^5)^5 \times (1 - 60.3/10^5) \times (1 - 98.9/10^5)^5 = .99822 \times .99699 \times .995065 = .99030$$

Therefore, $Pr(\text{developing colon cancer over 15 years}) = 1 - .99030 = .00970$. Thus, the expected number of colon cancers over 15 years among 1000 50-year-old men is $1000 \times .00970 = 9.7$.

3.31 Using methods similar to those used in Problem 3.30, we see that

$$Pr(\text{not developing lung cancer over 15 years}) = (1 - 76.1/10^5)^5 \times (1 - 137.5/10^5)^5 \times (1 - 231.7/10^5)^5 = .996201 \times .993144 \times .988469 = .97796$$

Thus, $Pr(\text{developing lung cancer over 15 years}) = 1 - .97796 = .02204$ and the expected number of lung cancers over 15 years among 1000 50-year-old men is $1000 \times .02204 = 22.0$.

3.32 Using similar methods as in Problem 3.30, we see that

$$Pr(\text{not developing stomach cancer over 15 years}) = (1 - 20.8/10^5)^5 \times (1 - 39.1/10^5)^5 \times (1 - 46.0/10^5)^5 = .998960 \times .998047 \times .997702 = .99472$$

Thus, $Pr(\text{developing stomach cancer over 15 years}) = 1 - .99472 = .00528$ and the expected number of stomach cancers over 15 years among 1000 50-year-old men is $1000 \times .00528 = 5.3$.

3.33 We wish to compute $Pr(\text{test} + |\text{died})$. We use Bayes' theorem to solve this problem as follows:

$$Pr(\text{test} + |\text{dead}) = \frac{Pr(\text{dead}|\text{test} +)Pr(\text{test} +)}{Pr(\text{dead})}$$

We have from Table 3.4 that

$$Pr(\text{dead}|\text{test} +) = \frac{13.2}{100} = .132.$$

Furthermore, we are given that

$$Pr(\text{test} +) = Pr(\text{exercise - test heart rate} > 127) = .20.$$

To compute $Pr(\text{dead})$, we use the relation

$$\begin{aligned} &Pr(\text{dead}) = \\ &Pr(\text{dead}|\text{heart rate} \le 105)Pr(\text{heart rate} \le 105) \\ &+\cdots+ Pr(\text{dead}|\text{heart rate} > 127)Pr(\text{heart rate} > 127) \\ &= .091(.20) + .087(.30) + .116(.30) + .132(.20) \\ &= .1055 \end{aligned}$$

Therefore, we have that

$$Pr(\text{test} + |\text{dead}) = \frac{.132(.20)}{.1055} = .250$$

Thus, there is a 25% probability that people who die of coronary disease over a 20-year follow-up period will have a positive test at baseline. This is the *sensitivity* of the test.

3.34 We wish to compute $Pr(\text{test} + |\text{alive})$. From Bayes' theorem, we have

$$Pr(\text{test} + |\text{alive}) = \frac{Pr(\text{alive}|\text{test} +)Pr(\text{test} +)}{Pr(\text{alive})}$$

From Problem 3.33, we note that

$$\begin{aligned} Pr(\text{alive}|\text{test} +) &= 1 - Pr(\text{dead}|\text{test} +) \\ &= 1 - .132 = .868 \end{aligned}.$$

Furthermore, $Pr(\text{test} +) = .20$,

$$Pr(\text{alive}) = 1 - Pr(\text{dead}) = 1 - .1055 = .8945.$$

Thus, we have

$$Pr(\text{test} + |\text{alive}) = \frac{.868(.20)}{.8945} = .194$$

Thus, there is a 19.4% probability that people who do not die of coronary disease over a 20-year follow-up period will have a positive test at baseline. This probability $= 1 - \text{specificity}$.

3.35 We want

$$Pr(\text{death}|\text{negative test}) = Pr(\text{dead}|\text{heart rate} \le 127).$$

This is given by

$$\begin{aligned} &\frac{Pr\left(\text{death} \cap \text{heart rate} \le 127\right)}{Pr\left(\text{heart rate} \le 127\right)} \\ &= \frac{\begin{array}{c} Pr\left(\text{death} \cap \le 105\right) + Pr\left(\text{death} \cap 106 - 115\right) \\ +Pr\left(\text{death} \cap 116 - 127\right) \end{array}}{Pr\left(\text{heart rate} \le 127\right)} \\ &= \frac{\left[\begin{array}{l} Pr\left(\text{death} \middle| \le 105\right) Pr\left(\le 105\right) \\ +Pr\left(\text{death} \middle| 106 - 115\right) Pr\left(106 - 115\right) \\ +Pr\left(\text{death} \middle| 116 - 127\right) Pr\left(116 - 127\right) \end{array}\right]}{.80} = \frac{.0791}{.80} = .099 \end{aligned}$$

Thus, there is a 9.9% probability of death over 20 years among men with a negative test. This is $1 - \textit{predictive value negative}$. Men who have a negative test, but who subsequently die of coronary heart disease are referred to as *false negatives*. The false-negative rate = 9.9%.

3.36 The sensitivity = the probability that the exercise test will be positive given that a patient has a positive angiogram.

3.37 The specificity = the probability that the exercise test will be negative given that a patient has a negative angiogram.

3.38 We use Bayes' theorem here. Let A = exercise-test positive, B = patient has disease. We wish to compute

$$Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A|B)Pr(B) + Pr\left(A\middle|\overline{B}\right)Pr\left(\overline{B}\right)}$$

where $Pr(A|B) = .79$,

$$Pr\left(A\middle|\overline{B}\right) = 1 - Pr\left(\overline{A}\middle|\overline{B}\right) = 1 - .68 = .32,$$

$Pr(B) = .20$. Therefore, we have

$$Pr(B|A) = \frac{.79 \times .20}{.79 \times .20 + .32 \times .80} = \frac{.158}{.414} = .382$$

Therefore there is a 38.2% chance that the patient has disease given that she has a positive exercise test.

3.39 We wish to compute $Pr\left(B\middle|\overline{A}\right)$. We have from Bayes' theorem that

$$\begin{aligned} Pr\left(B\middle|\overline{A}\right) &= \frac{Pr\left(\overline{A}|B\right)Pr(B)}{Pr\left(\overline{A}|B\right)Pr(B) + Pr\left(\overline{A}\middle|\overline{B}\right)Pr\left(\overline{B}\right)} \\ &= \frac{(1 - .79)(.20)}{(1 - .79)(.20) + .68(.80)} = \frac{.042}{.042 + .544} \\ &= \frac{.042}{.586} = .072 \end{aligned}$$

Thus there is only a 7.2% chance that the patient has disease if her exercise test is negative.

3.40 The sensitivity = 138/156 = .885.

3.41 The specificity = 16/17 = .941.

3.42 The predictive value positive 138/139 = .993.

3.43 The predictive value negative 16/34 = .471.

3.44 We use the formula

$$PV+ = \frac{\text{prevalence} \times \text{sensitivity}}{\text{prevalence} \times \text{sensitivity} + (1 - \text{prevalence})(1 - \text{specificity})}$$
$$= \frac{.8(.885)}{.8(.885) + .2(1 - .941)} = \frac{.708}{.719} = .984$$

3.45 Since both twins must have the same outcome, the probability that both twins are affected = 1/4, the probability that one twin is affected = 0, the probability that neither twin is affected = 3/4. The outcomes are dependent events because

$$\frac{1}{4} = Pr(A_1 \cap A_2) = Pr(A_1)$$
$$= Pr(A_2) \neq Pr(A_1) \times Pr(A_2) = \frac{1}{16}$$

3.46 The outcomes for the two *DZ* twins are independent events. We have $Pr(\text{both affected}) = (1/4)^2 = 1/16$, $Pr(\text{one affected}) = 2(1/4)(3/4) = 6/16 = 3/8$, $Pr(\text{neither affected}) = (3/4)^2 = 9/16$.

3.47 We use the total probability rule. Let B_2, B_1, B_0 be the events that 2, 1, and 0 twins are affected. We have

$$Pr(B_2) = Pr(B_2|MZ)Pr(MZ) + Pr(B_2|DZ)Pr(DZ)$$
$$= \left(\frac{1}{4}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{16}\right)\left(\frac{2}{3}\right) = \frac{1}{12} + \frac{2}{48} = \frac{6}{48} = \frac{1}{8}$$

Similarly,

$$Pr(B_1) = 0\left(\frac{1}{3}\right) + \left(\frac{3}{8}\right)\left(\frac{2}{3}\right) = \frac{1}{4}$$
$$Pr(B_0) = \left(\frac{3}{4}\right)\left(\frac{1}{3}\right) + \left(\frac{9}{16}\right)\left(\frac{2}{3}\right) = \frac{1}{4} + \frac{3}{8} = \frac{5}{8}$$

3.48 We use Bayes' theorem. We wish to compute $Pr(MZ|B_2)$. We have

$$Pr(MZ|B_2) = \frac{Pr(B_2|MZ)Pr(MZ)}{Pr(B_2|MZ)Pr(MZ) + Pr(B_2|DZ)Pr(DZ)}$$

From Problems 3.45 and 3.46, we have

$$Pr(B_2|MZ) = \frac{1}{4},\ Pr(B_2|DZ) = \frac{1}{16}.$$

Also from Problem 3.47,

$$Pr(MZ) = \frac{1}{3}$$
$$Pr(DZ) = \frac{2}{3}.$$

Thus

$$Pr(MZ|B_2) = \frac{(\frac{1}{4})(\frac{1}{3})}{(\frac{1}{4})(\frac{1}{3}) + (\frac{1}{16})(\frac{2}{3})} = \frac{\frac{1}{12}}{\frac{1}{12} + \frac{1}{24}} = \frac{2}{3}$$

3.49 An incidence rate of 48.9 strokes per 1000 person-years among 70–79-year-olds with AF means that if a group of 1000 70–79-year-old people with AF are followed for 1 year, then on average 48.9 of them will develop stroke over the next year.

3.50 The relative risk (*RR*) = incidence rate of stroke among people with AF/incidence rate of stroke among people without AF. Thus,

$$RR_{60-69} = \frac{21.2}{4.5} = 4.7$$
$$RR_{70-79} = \frac{48.9}{9.0} = 5.4$$
$$RR_{80-89} = \frac{71.4}{14.3} = 5.0$$

Thus, the relative risk is approximately 5 for each age group. This means that for persons without stroke in a given age group, those with AF are about 5 times more likely to develop a new case of stroke than those without AF.

3.51 We use the total probability rule. First, we find the incidence rate of stroke within a given age group using the formula

Incidence rate of stroke = prevalence of AF × incidence rate of stroke for people with AF + (1 – prevalence of AF) × incidence rate of stroke for people without AF

Thus we have

$$I_{60-69} = \text{Incidence rate}_{60-69} = .018\left(\frac{21.2}{1000}\right) + .982\left(\frac{4.5}{1000}\right)$$
$$= 4.80 \text{ per } 1000$$
$$I_{70-79} = \text{Incidence rate}_{70-79} = .047\left(\frac{48.9}{1000}\right) + .953\left(\frac{9.0}{1000}\right)$$
$$= 10.88 \text{ per } 1000$$
$$I_{80-89} = \text{Incidence rate}_{80-89} = .102\left(\frac{71.4}{1000}\right) + .898\left(\frac{14.3}{1000}\right)$$
$$= 20.12 \text{ per } 1000$$

The incidence rate of stroke roughly doubles in successive decades of life. We again use the total probability rule to obtain the overall incidence as follows:

$$\text{Incidence rate} = \frac{200 \times I_{60-69} + 200 \times I_{70-79} + 100 \times I_{80-89}}{500}$$
$$= \frac{200\left(\frac{4.80}{1000}\right) + 200\left(\frac{10.88}{1000}\right) + 100\left(\frac{20.12}{1000}\right)}{500}$$
$$= \frac{5.148}{500} = .0103 \approx 10.3 \text{ per } 1000$$

3.52 If n subjects are enrolled, then the expected number of strokes over 1 year = nI, where I is the 1-year incidence rate. Thus, we can solve for n and obtain

$$n = \frac{50}{.0103} = 4856.6 .$$

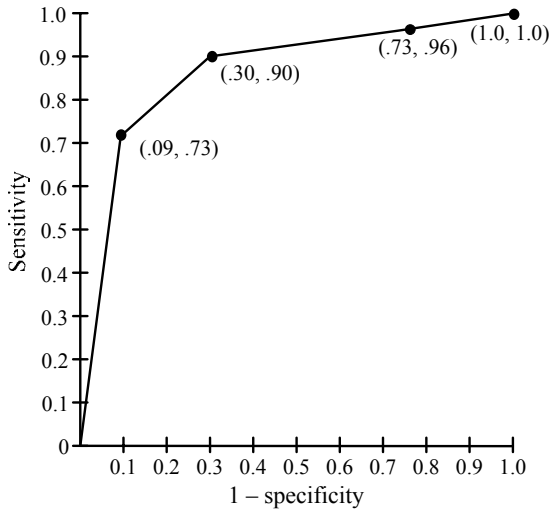
Thus we need to enroll 4857 subjects to ensure that the expected number of strokes over 1 year = 50. An alternative design would be to enroll fewer subjects but follow the subjects over a longer period of time.

3.53 For a cutpoint of ≤ 2 for positivity, the sensitivity = 49/67 = .73. Sensitivity for the other cutpoints are given in the table below.

3.54 For a cutpoint of ≤ 2 for positivity, the specificity = 30/33 = .91. Specificity for the other cutpoints are given in the table below.

Cutpoint for Positivity	Sensitivity	Specificity
<= 1	.00	1.00
<= 2	.73	.91
<= 3	.90	.70
<= 4	.96	.27
<= 5	1.00	.00

3.55 The ROC curve is a plot of sensitivity on the y-axis vs. 1-specificity on the x-axis. This is given below.



3.56 The area under the ROC curve for Reader 2 is

$$\text{area} = \frac{1}{2}\left[\begin{array}{l}(.09-0)(0+.73)+(.30-.09)(.73+.90) \\ +(.73-.30)(.90+.96)+(1.0-.73)(.96+1.0)\end{array}\right]$$
$$= .87.$$

Thus, a randomly selected affected person will have a lower score than a randomly affected normal person about 87% of the time. This is similar to the accuracy for Reader 1 (86%).

REFERENCES

[1] Lee, M. L. T., & Rosner, B. A. (1999). Analysis for areas under correlated ROC curves; a nonparametric approach. Submitted for publication.

[2] Tenkanen, L., Teppo, L., & Hakulinen, T (1987). Smoking and cardiac symptoms as predictors of lung cancer. *Journal of Chronic Disease*, **40**(12), 1121–1128.

[3] Butler, W. J., Ostrander, L. D., Jr., Carman, W. J., & Lamphiear, D. E. (1982). Diabetes mellitus in Tecumseh, Michigan: Prevalence, incidence and associated conditions. *American Journal of Epidemiology*, **116**(6), 971–980.

[4] Doll, R., Muir, C., & Waterhouse, J. (Eds.). (1970). *Cancer incidence in five continents II*. Berlin: Springer-Verlag.

[5] Slattery, M. L., & Jacobs, D. R., Jr. (1988). Physical fitness and cardiovascular disease mortality: The U.S. railroad study. *American Journal of Epidemiology*, **127**(3). 571–580.

[6] Wolf, P. A., Abbott, R. D., & Kannel, W. B. (1987). Atrial fibrillation: A major contributor to stroke in the elderly. The Framingham Study. *Archives of Internal Medicine*, **147**(9), 1561–1564.