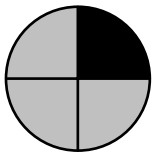
# Chapter 3

# 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.

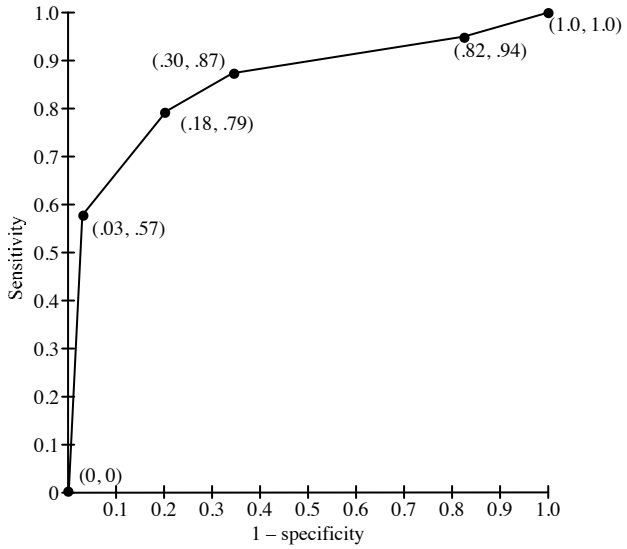
For example, in one study, 4 participating readers used two different types of film, PACS film and plain film to evaluate abnormality based on radiographic images. There was a 5 point rating scale with a lower score indicating abnormality. The issue is what cutpoint to use to designate abnormality. The 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 $\leq 1$ to designate abnormal and $\geq 2$ 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 |
|---|---|---|
| $\leq 0$ | .00 | 1.00 |
| $\leq 1$ | .57 | .97 |
| $\leq 2$ | .79 | .82 |
| $\leq 3$ | .87 | .70 |
| $\leq 4$ | .94 | .18 |
| $\leq 5$ | 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),

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

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%.

## SOLUTIONS

**3.1** Yes

**3.2** No

**3.3** .3

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

**3.5** $A \cap C =$ {$250 \leq$ serum cholesterol $\leq 280$}

**3.6** $B \cup C =$ {serum cholesterol $\leq 280$ or $\geq 300$}

**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 $B_1, \ldots, B_6$.