Chapter 2
Introduction to Probability

Learning Objectives

1. Obtain an understanding of the role probability information plays in the decision making process.

2. Understand probability as a numerical measure of the likelihood of occurrence.

3. Be able to use the three methods (classical, relative frequency, and subjective) commonly used for assigning probabilities and understand when they should be used.

4. Be able to use the addition law and be able to compute the probabilities of events using conditional probability and the multiplication law.

5. Be able to use new information to revise initial (prior) probability estimates using Bayes' theorem.

6. Know the definition of the following terms:

experiment	addition law
sample space	mutually exclusive
event	conditional probability
complement	independent events
Venn Diagram	multiplication law
union of events	prior probability
intersection of events	posterior probability
Bayes' theorem	Simpson’s Paradox

Solutions:

1. a. Go to the x-ray department at 9:00 a.m. and record the number of persons waiting.

   b. The experimental outcomes (sample points) are the number of people waiting: 0, 1, 2, 3, and 4.

   Note: While it is theoretically possible for more than 4 people to be waiting, we use what has actually been observed to define the experimental outcomes.

   c.

Number Waiting	Probability
0	.10
1	.25
2	.30
3	.20
4	.15
Total:	1.00

   d. The relative frequency method was used.

2. a. Choose a person at random, have her/ him taste the 4 blends and state a preference.

   b. Assign a probability of 1/4 to each blend. We use the classical method of equally likely outcomes here.

   c.

Blend	Probability
1	.20
2	.30
3	.35
4	.15
Total:	1.00

   The relative frequency method was used.

3. Initially a probability of .20 would be assigned if selection is equally likely. Data does not appear to confirm the belief of equal consumer preference. For example using the relative frequency method we would assign a probability of 5 / 100 = .05 to the design 1 outcome, .15 to design 2, .30 to design 3, .40 to design 4, and .10 to design 5.

4. a. Of the 132,275,830 individual tax returns received by the IRS, 31,675,935were in the 1040A, Income Under $25,000 category. Using the relative frequency approach, the probability a return from the 1040A, Income Under $25,000 category would be chosen at random is 31675935/132275830 = 0.239.

   b. Of the 132,275,830 individual tax returns received by the IRS, 3,376,943 were in the Schedule C, Reciepts Under $25,000 category; 3,867,743 were in the Schedule C, Reciepts $25,000-$100,000 category; and were 2,288,550 in the Schedule C, Reciepts $100,000 & Over category. Therefore, 9,533,236 Schedule Cs were filed in 2006, and the remaining 132,275,830 - 9,533,236 = 122,742,594 individual returns did not use Schedule C. By the relative frequency approach, the probability the chosen return did not use Schedule C is 122742594/132275830 = 0.928.

c. Of the 132,275,830 individual tax returns received by the IRS, 12,893,802 were in the Non 1040A, Income \$100,000 & Over category; 2,288,550 were in the Schedule C, Reciepts \$100,000 & Over category; and 265,612 were in the Schedule F, Reciepts \$100,000 & Over category. By the relative frequency approach, the probability the chosen return reported income/reciepts of \$100,000 and over is (12893802 + 2288550 + 265612)/132275830 = 15447964/132275830 = 0.117.

d. 26,463,973 Non 1040A, Income \$50,000-\$100,000 returns were filed, so assuming examined returns were evenly distributed across the ten categories (i.e., the IRS examined 1% of individual returns in each category), the number of returns from the Non 1040A, Income \$50,000-\$100,000 category that were examined is 0.01(26463973) = 264,639.73 (or 264,640).

e. The proportion of total returns in the Schedule C, reciepts \$100,000 & Over is 2,288,550/132,275,830 = 0.0173. Therefore, if we assume the recommended additional taxes are evenly distributed across the ten categories, the amount of recommended additional taxes for the Schedule C, Reciepts \$100,000 & Over category is 0.0173(\$13,045,221,000.00) = \$225,699,891.81.

5. a. No, the probabilities do not sum to one. They sum to 0.85.

b. Owner must revise the probabilities so that they sum to 1.00.

6. a. $P(A) = P(150 - 199) + P(200 \text{ and over})$

$$= \frac{26}{100} + \frac{5}{100}$$

$= 0.31$

b. $P(B) = P(\text{less than } 50) + P(50 - 99) + P(100 - 149)$

$= 0.13 + 0.22 + 0.34$

$= 0.69$

7. a. P(A) = .40, P(B) = .40, P(C) = .60

b. $P(A \cup B) = P(E_1, E_2, E_3, E_4) = .80$. Yes $P(A \cup B) = P(A) + P(B)$.

c. $A^c = \{E_3, E_4, E_5\}$ $C^c = \{E_1, E_4\}$ $P(A^c) = .60$ $P(C^c) = .40$

d. $A \cup B^c = \{E_1, E_2, E_5\}$ $P(A \cup B^c) = .60$

e. $P(B \cup C) = P(E_2, E_3, E_4, E_5) = .80$

8. a. Let P(*A*) be the probability a hospital had a daily inpatient volume of at least 200 and P(*B*) be the probability a hospital had a nurse to patient ratio of at least 3.0. From the list of thirty hospitals, sixteen had a daily inpatient volume of at least 200, so by the relative frequency approach the probability one of these hospitals had a daily inpatient volume of at least 200 is P(*A*) = 16/30 = 0.533, Similarly, since ten (one-third) of the hospitals had a nurse-to-patient ratio of at least 3.0, the probability of a hospital having a nurse-to-patient ratio of at least 3.0 is P(*B*) = 10/30 = 0.333. Finally, since seven of the hospitals had both a daily inpatient volume of at least 200 and a nurse-to-patient ratio of at least 3.0, the probability of a hospital having both a daily inpatient volume of at least 200 and a nurse-to-patient ratio of at least 3.0 is $P(A \cap B) = 7/30 = 0.233$.

b. The probability that a hospital had a daily inpatient volume of at least 200 or a nurse to patient ratio of at least 3.0 or both is $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 16/30 + 10/30 - 7/30 = (16 + 10 - 7)/30 = 19/30 = 0.633$.

c. The probability that a hospital had neither a daily inpatient volume of at least 200 nor a nurse to patient ratio of at least 3.0 is $1 - P(A \cup B) = 1 - 19/30 = 11/30 = 0.367$.

9. Let E = event patient treated experienced eye relief.
S = event patient treated had skin rash clear up.

Given:

$P(E) = 90 / 250 = 0.36$

$P(S) = 135 / 250 = 0.54$

$P(E \cup S) = 45 / 250 = 0.18$

$$
\begin{aligned}
P(E \cup S) &= P(E) + P(S) - P(E \cap S) \\
&= 0.36 + 0.54 - 0.18 \\
&= 0.72
\end{aligned}
$$

10. P(Defective and Minor) = 4/25

P(Defective and Major) = 2/25

P(Defective) = (4/25) + (2/25) = 6/25

P(Major Defect | Defective) = P(Defective and Major) / P(Defective) = (2/25)/(6/25) = 2/6 = 1/3.

11. a. Yes; the person cannot be in an automobile and a bus at the same time.

b. $P(B^c) = 1 - P(B) = 1 - 0.35 = 0.65$

12. a. $P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.40}{0.60} = 0.6667$

b. $P(B|A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{0.40}{0.50} = 0.80$

c. No because $P(A | B) \neq P(A)$

13. a.

	Reason for Applying			
	Quality	Cost/Convenience	Other	Total
Full Time	0.218	0.204	0.039	0.461
Part Time	0.208	0.307	0.024	0.539
Total	0.426	0.511	0.063	1.00

b. It is most likely a student will cite cost or convenience as the first reason: probability = 0.511. School quality is the first reason cited by the second largest number of students: probability = 0.426.

c. P (Quality | full time) = 0.218/0.461 = 0.473

d. P (Quality | part time) = 0.208/0.539 = 0.386

e. P (B) = 0.426 and P (B | A) = 0.473

Since $P(B) \neq P(B \mid A)$, the events are dependent.

14.

	\$0-\$499	\$500-\$999	>=\$1000	
<2 yrs	120	240	90	450
>= 2 yrs	75	275	200	550
	195	515	290	1000

	\$0-\$499	\$500-\$999	>=\$1000	
<2 yrs	0.12	0.24	0.09	0.45
>= 2 yrs	0.075	0.275	0.2	0.55
	0.195	0.515	0.29	1.00

a. P(< 2 yrs) = .45

b. P(>= \$1000) = .29

c. P(2 accounts have > = \$1000) = (.29)(.29) = .0841

d. P(\$500-\$999 | >= 2 yrs) = P(\$500-\$999 and >= 2 yrs) / P(>=2yrs) = .275/.55 = .5

e. P(< 2 yrs and >=\$1000) = .09

f. P(>=2 yrs | \$500-\$999) = .275/.515 = .533981

15. a. A joint probability table for these data looks like this:

		Automobile Insurance Coverage		
		Yes	**No**	**Total**
Age	**18 to 34**	.375	.085	.46
	35 and over	.475	.065	.54
	Total	.850	.150	1.00

For parts (b) through (g):

Let A= 18 to 34 age group
B= 35 and over age group
Y = Has automobile insurance coverage
N = Does not have automobile insurance coverage

b. We have P(A) = .46 and P(B) = .54, so of the population age 18 and over, 46% are ages 18 to 34 and 54% are ages 35 and over.

c. The probability a randomly selected individual does not have automobile insurance coverage is P(N) = .15.

d. If the individual is between the ages of 18 and 34, the probability the individual does not have automobile insurance coverage is

$$P(N \mid A) = \frac{P(N \cap A)}{P(A)} = \frac{.085}{.46} = .1848.$$

e. If the individual is age 35 or over, the probability the individual does not have automobile insurance coverage is

$$P(N|B) = \frac{P(N \cap B)}{P(B)} = \frac{.065}{.54} = .1204.$$

f. If the individual does not have automobile insurance, the probability that the individual is in the 18–34 age group is

$$P(A|N) = \frac{P(A \cap N)}{P(N)} = \frac{.085}{.15} = .5667.$$

g. The probability information tells us that in the US, younger drivers are less likely to have automobile insurance coverage.

16. a. $P(A \cap B) = P(A)P(B) = (0.55)(0.35) = 0.19$

b. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.90 - 0.19 = 0.71$

c. 1 - 0.71 = 0.29

17. a. P(attend multiple games) = 196 / 989 ≈ 19.8%.

b. P(male | attend multiple games) = 177 / 196 ≈ 90.3%.

c. P(male and attend multiple games) = P(male | attend multiple games) × P(attend multiple games) = (177 / 196) × (196 / 989) = 177 / 989 ≈ 17.9%.

d. P(attend multiple games | male) = P(attend multiple games and male) / P(male) = (177 / 989) / (759 / 989) = 177 / 759 ≈ 23.3%.

e. P(male or attend multiple games) = P(male) + P(attend multiple games) – P(male and attend multiple games) = (759 / 989) + (196 / 989) – (177 / 989) = 778 / 989 ≈ 78.7%.

18. a. $P(B) = 0.25$

$P(S|B) = 0.40$
$P(S \cap B) = 0.25(0.40) = 0.10$

b. $$P(B|S) = \frac{P(S \cap B)}{P(S)} = \frac{0.10}{0.40} = 0.25$$

c. B and S are independent. The program appears to have no effect.

19. Let: A = lost time accident in current year
B = lost time accident previous year

∴ Given: $P(B) = 0.06$, $P(A) = 0.05$, $P(A|B) = 0.15$

a. $P(A \cap B) = P(A|B)P(B) = 0.15(0.06) = 0.009$

b. $P(A \cup B) = P(A) + P(B) - P(A| B)$
$= 0.06 + 0.05 - 0.009 = 0.101$ or 10.1%

20. a. $P(B \cap A_1) = P(A_1)P(B \mid A_1) = (0.20)(0.50) = 0.10$

$P(B \cap A_2) = P(A_2)P(B \mid A_2) = (0.50)(0.40) = 0.20$

$P(B \cap A_3) = P(A_3)P(B \mid A_3) = (0.30)(0.30) = 0.09$

b. $P(A_2 \mid B) = \dfrac{0.20}{0.10 + 0.20 + 0.09} = 0.51$

c.

Events	$P(A_i)$	$P(B \mid A_i)$	$P(A_i \cap B)$	$P(A_i \mid B)$
A_1	0.20	0.50	0.10	0.26
A_2	0.50	0.40	0.20	0.51
A_3	0.30	0.30	0.09	0.23
	1.00		0.39	1.00

21. S_1 = successful, S_2 = not successful and B = request received for additional information.

a. $P(S_1) = 0.50$

b. $P(B \mid S_1) = 0.75$

c. $P(S_1 \mid B) = \dfrac{(0.50)(0.75)}{(0.50)(0.75) + (0.50)(0.40)} = \dfrac{0.375}{0.575} = 0.65$

22. a. Let F = female. Using past history as a guide, P(F) = .40

b. Let D = Dillard's

$$P(F \mid D) = \frac{.40(3/4)}{.40(3/4) + .60(1/4)} = \frac{.30}{.30 + .15} = .67$$

The revised (posterior) probability that the visitor is female is .67.

We should display the offer that appeals to female visitors.

23. a. P(Oil) = 0.50 + 0.20 = 0.70

b. Let S = Soil test results

Events	$P(A_i)$	$P(S \mid A_i)$	$P(A_i \cap S)$	$P(A_i \mid S)$
High Quality (A_1)	0.50	0.20	0.10	0.435
Medium Quality (A_2)	0.20	0.20	0.04	0.174
No Oil (A_3)	0.30	0.30	0.09	0.391
	1.00		P(S) = 0.23	1.000

P(Oil) = 0.609 which is good but not as good as estimated prior to the soil test; probabilities also still favor high quality oil.

24. Let S= speeding is reported
S^C= speeding is not reported
F = Accident results in fatality for vehicle occupant

We have P(S) = .129, so $P(S^C)$ = .871. Also P(F|S) = .196 and $P(F|S^C)$ = .05. Using the tabular form of Bayes' Theorem provides:

Events	Prior Probabilities	Conditional Probabilities	Joint Probabilities	Posterior Probabilities
S	.129	.196	.0384	.939
S^C	.871	.050	.0025	.061
	1.000		P(F) = .0409	1.000

P(S | F) = .2195, i.e., if an accident involved a fatality. the probability speeding was reported is 0.939.

25.

Events	$P(A_i)$	$P(D \mid A_i)$	$P(A_i \cap D)$	$P(A_i \mid D)$
Supplier A	0.60	0.0025	0.0015	0.23
Supplier B	0.30	0.0100	0.0030	0.46
Supplier C	0.10	0.0200	0.0020	0.31
	1.00		P(D) = 0.0065	1.00

a. P(D) = 0.0065

b. B is the most likely supplier if a defect is found.

26. a.

Events	$P(D_i)$	$P(S_1 \mid D_i)$	$P(D_i \cap S_1)$	$P(D_i \mid S_1)$
D_1	.60	.15	.090	.2195
D_2	.40	.80	.320	.7805
	1.00		$P(S_1)$ = .410	1.0000

$P(D_1 \mid S_1)$ = .2195

$P(D_2 \mid S_1)$ = .7805

b.

Events	$P(D_i)$	$P(S_2 \mid D_i)$	$P(D_i \cap S_2)$	$P(D_i \mid S_2)$
D_1	.60	.10	.060	.500
D_2	.40	.15	.060	.500
	1.00		$P(S_2)$ = .120	1.000

$P(D_1 \mid S_2)$ = .50

$P(D_2 \mid S_2)$ = .50

c.

Events	$P(D_i)$	$P(S_3 \mid D_i)$	$P(D_i \cap S_3)$	$P(D_i \mid S_3)$
D_1	.60	.15	.090	.8824
D_2	.40	.03	.012	.1176
	1.00		$P(S_3)$ = .102	1.0000

$P(D_1 \mid S_3) = .8824$

$P(D_2 \mid S_3) = .1176$

d. Use the posterior probabilities from part (a) as the prior probabilities here.

Events	$P(D_i)$	$P(S_2 \mid D_i)$	$P(D_i \cap S_2)$	$P(D_i \mid S_2)$
D_1	.2195	.10	.0220	.1582
D_2	.7805	.15	.1171	.8418
	1.0000		.1391	1.0000

$P(D_1 \mid S_1 \text{ and } S_2) = .1582$

$P(D_2 \mid S_1 \text{ and } S_2) = .8418$

27. a. Let A = age 65 or older

$$P(A) = 1 - .835 = .165$$

b. Let D = takes drugs regularly

$$P(A \mid D) = \frac{P(A)P(D \mid A)}{P(A)P(D \mid A) + P(A^C)P(D \mid A^C)}$$

$$= \frac{.165(.82)}{.165(.82) + .835(.49)}$$

$$= \frac{.1353}{.1353 + .4092} = .2485$$

28. a. $P(A_1) = .095$

$P(A_2) = .905$

$P(W \mid A_1) = .60$

$P(W \mid A_2) = .49$

b.

Events	$P(A_i)$	$P(W \mid A_i)$	$P(A_i \cap W)$	$P(A_i \mid W)$
A_1	0.095	0.60	0.05700	0.1139
A_2	0.905	0.49	0.44345	0.8861
			$P(W)$ = 0.50045	1.0000

$P(A_1 \mid W) = .1139$

c.

Events	$P(A_i)$	$P(M \mid A_i)$	$P(A_i \cap M)$	$P(A_i \mid M)$
A_1	0.095	0.40	0.03800	0.0761
A_2	0.905	0.51	0.46155	0.9239
			$P(M)$ = 0.49995	1.0000

$P(A_1 \mid M) = .0761$

d. $P(W) = .50045$

$P(M) = .49965$

29. a.

Gender	**Too Fast**	**Acceptable**
Male Golfers	35	65
Female Golfers	40	60

The proportion of male golfers who say the greens are too fast is 35/(35 + 65) = 0.35, while the proportion of female golfers who say the greens are too fast is 40/(40 + 60) = 0.40. There is a higher percentage of female golfers who say the greens are too fast.

b. There are 50 male golfers with low handicaps, and 10 of these golfers say the greens are too fast, so for male golfers the proportion with low handicaps who say the greens are too fast is 10/50 = 0.20. On the other hand, there are 10 female golfers with low handicaps, and 1 of these golfers says the greens are too fast, so for female golfers the proportion with low handicaps who say the greens are too fast is 1/10 = 0.10.

c. There are 50 male golfers with higher handicaps, and 25 of these golfers say the greens are too fast, so for male golfers the proportion with higher handicaps who say the greens are too fast is 25/50 = 0.50. On the other hand, there are 90 female golfers with higher handicaps, and 39 of these golfers says the greens are too fast, so for female golfers the proportion with higher handicaps who say the greens are too fast is 39/90 = 0.43.

d. When the data are aggregated across handicap categories, the proportion of female golfers who say the greens are too fast exceeds the proportion of male golfers who say the greens are too fast. However, when we introduce a third variable, handicap, we see different results. When sorted by handicap categories, we see that the proportion of male golfers who find the greens too fast is higher than female golfers for both low and high handicap categories. This is an example of Simpson's paradox.

30. a.

	Male Applicants	Female Applicants
Accept	70	40
Deny	90	80

After combining these two crosstabulations into a single crosstabulation with Accept and Deny as the row labels and Male and Female as the column labels, we see that the rate of acceptance for males across the university is 70/(70+90) = .4375 or approximately 44%, while the rate of acceptance for females across the university is 40/(40+80) = .33 or 33%.

b. If we focus solely on the overall data, we would conclude the university's admission process is biased in favor of male applicant. However, this occurs because most females apply to the College of Business (which has a far lower rate of acceptance that the College of Engineering). When we look at each college's acceptance rate by gender, we see the acceptance rate of males and females are equal in the College of Engineering (75%) and the acceptance rate of males and females are equal in the College of Business (33%). The data do not support the accusation that the university favors male applicants in its admissions process.