The Variance of a Random Variable

The mean is a measure of the center of a distribution. Even the most basic numerical description requires in addition a measure of the spread or variability of the distribution. The variance and the standard deviation are the measures of spread that accompany the choice of the mean to measure center. Just as for the mean, we need a distinct symbol to distinguish the variance of a random variable from the variance s^2 of a data set. We write the variance of a random variable X as σ_X^2. Once again the subscript reminds us which variable we have in mind. The definition of the variance σ_X^2 of a random variable is similar to the definition of the sample variance s^2 given in Chapter 1. That is, the variance is an average of the squared deviation $(X - \mu_X)^2$ of the variable X from its mean μ_X. As for the mean, the average we use is a weighted average in which each outcome is weighted by its probability in order to take account of outcomes that are not equally likely. Calculating this weighted average is straightforward for discrete random variables but requires advanced mathematics in the continuous case. Here is the definition.

Variance of a Discrete Random Variable

Suppose that X is a discrete random variable whose distribution is

Value of X:	x_1	x_2	x_3	...	x_k
Probability:	p_1	p_2	p_3	...	p_k

and that μ is the mean of X. The variance of X is

$$\begin{aligned}\sigma_X^2 &= (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \ldots + (x_k - \mu_X)^2 p_k \\ &= \sum (x_i - \mu_X)^2 p_i\end{aligned}$$

The standard deviation σ_X of X is the square root of the variance.

Example 7.7 *Linda sells cars*
Variance of a random variable

DO-NOW

what section in our book is this from?

Linda is a sales associate at a large auto dealership. She motivates herself by using probability estimates of her sales. For a sunny Saturday in April, she estimates her car sales as follows:

Cars sold:	0	1	2	3
Probability:	0.3	0.4	0.2	0.1

We can find the mean and variance of X by arranging the calculation in the form of a table. Both μ_X and σ_X^2 are sums of columns in this table.

x_i	p_i	$x_i p_i$	$(x_i - \mu_X)^2 p_i$
0	0.3	0.0	$(0 - 1.1)^2(0.3) = 0.363$
1	0.4	0.4	$(1 - 1.1)^2(0.4) = 0.004$
2	0.2	0.4	$(2 - 1.1)^2(0.2) = 0.162$
3	0.1	0.3	$(3 - 1.1)^2(0.1) = 0.361$
		$\mu_X = 1.1$	$\sigma_X^2 = 0.890$

invalid signatures in a sample of size 1000 have (approximately) a binomial distribution? Explain.

7.54 A coin is to be tossed 25 times. Let x = the number of tosses that result in heads (H). Consider the following rule for deciding whether or not the coin is fair:

Judge the coin to be fair if $8 \le x \le 17$.
Judge it to be biased if either $x \le 7$ or $x \ge 18$.

a. What is the probability of judging the coin to be biased when it is actually fair?
b. What is the probability of judging the coin to be fair when $P(\text{H}) = .9$, so that there is a substantial bias? Repeat for $P(\text{H}) = .1$.
c. What is the probability of judging the coin to be fair when $P(\text{H}) = .6$? When $P(\text{H}) = .4$? Why are the probabilities so large compared to the probabilities in part **b**?
d. What happens to the "error probabilities" of parts **a** and **b** if the decision rule is changed so that the coin is judged fair if $7 \le x \le 18$ and unfair otherwise? Is this a better rule than the one first proposed?

7.55 A city ordinance requires that a smoke detector be installed in all residential housing. There is concern that too many residences are still without detectors, so a costly inspection program is being contemplated. Let π = the proportion of all residences that have a detector. A random sample of 25 residences will be selected. If the sample strongly suggests that $\pi < .80$ (fewer than 80% have detectors), as opposed to $\pi \ge .80$, the program will be implemented. Let x = the number of residences among the 25 that have a detector, and consider the following decision rule:

Reject the claim that $\pi \ge .8$ and implement the program if $x \le 15$.

a. What is the probability that the program is implemented when $\pi = .80$?
b. What is the probability that the program is not implemented if $\pi = .70$? If $\pi = .60$?
c. How do the "error probabilities" of parts **a** and **b** change if the value 15 in the decision rule is changed to 14?

7.56 Exit polling has been a controversial practice in recent elections, since early release of the resulting information appears to affect whether or not those who have not yet voted will do so. Suppose that 90% of all registered California voters favor banning the release of information from exit polls in presidential elections until after the polls in California close. A random sample of 25 California voters is selected.

a. What is the probability that more than 20 favor the ban?
b. What is the probability that at least 20 favor the ban?
c. What are the mean value and standard deviation of the number who favor the ban?
d. If fewer than 20 in the sample favor the ban, is this at odds with the assertion that (at least) 90% of the populace favors the ban? (*Hint*: Consider $P(x < 20)$ when $\pi = .9$.)

7.57 Sophie is a dog who loves to play catch. Unfortunately, she isn't very good, and the probability that she catches a ball is only .1. Let x = number of tosses required until Sophie catches a ball.

a. Does x have a binomial or a geometric distribution?
b. What is the probability that it will take exactly two tosses for Sophie to catch a ball?
c. What is the probability that more than three tosses will be required?

7.58 Selected boxes of a breakfast cereal contain a prize. Suppose that 5% of the boxes contain the prize and the other 95% contain the message "Sorry, try again." A consumer determined to find a prize decides to continue to buy boxes of cereal until a prize is found. Consider the random variable x, where x = number of boxes purchased until a prize is found.

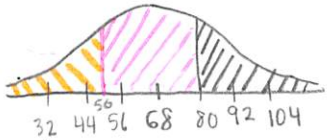
a. What is the probability that at most 2 boxes must be purchased?
b. What is the probability that exactly four boxes must be purchased?
c. What is the probability that more than four boxes must be purchased?

SUN

Math 210 Intro to Prob and Statistics

1. Weights of the Pacific yellowfin tuna follow a normal distribution with mean weight 68 pounds and standard deviation 12 pounds. For a randomly caught Pacific yellowfin tuna, what is the probability that the weight is

   a. less than 50 pounds?
   b. more than 80 pounds?
   c. between 50 and 80 pounds?



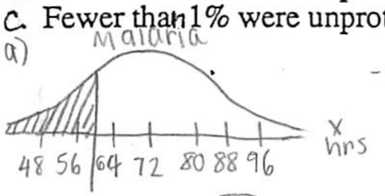
b) pencil
$z = \frac{80-68}{12} = 1$
$P(x > 80 lbs) = 1 - 0.8413 = 0.1587$
∃x a 15.87% chance that tuna will be greater than 80 lbs.

a) orange
$z = \frac{50-68}{12} = -1.5$
$P(z < -1.5) = 0.0668$
∃x a 6.68% chance that tuna will be less than 50 lbs.

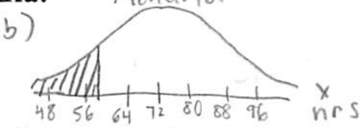
c) pink
$z = \frac{50-68}{12} = -1.5$
$P(-1.5 < z < 1) = 0.8413 - 0.0668 = 0.7745$
∃x a 77.45% chance that the tuna will weigh between 50 and 80 lbs.

2. A malaria prevention pill was developed to protect U.S. soldiers in the South Pacific during World War II. The pill had a number of mildly unpleasant side effects, so most soldiers wanted to take as few pills as possible. After extensive medical work using blood tests, it was found that for a single pill the duration of protection times was normally distributed, with mean $\mu = 72$ hours and standard deviation $\sigma = 8$ hours. If each soldier in a battalion was given a pill at breakfast mess (and ordered to take it), after how many hours should another pill be issued so that

   a. Fewer than 10% of the soldiers in the battalion were unprotected from malaria?
   b. Fewer than 5% were unprotected from malaria?
   c. Fewer than 1% were unprotected from malaria?

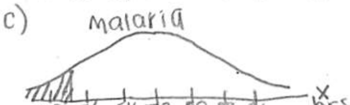
a)


$-1.28 = \frac{x-72}{8}$
$x = 61.76$ hrs
≈61 hrs

b)


$z = -1.645$
$-1.645 = \frac{x-72}{8}$
$x = 58.84$ hrs
≈ 58 hrs

c)

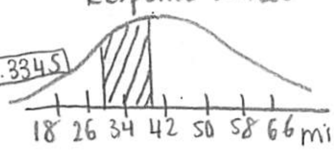
$z = -2.33$
$-2.33 = \frac{x-72}{8}$
$x = 53.36 \approx 53$ hrs

3. The Fight For Life emergency helicopter service is available for medical emergencies occurring from 15 to 90 miles from the hospital. Emergencies that occur closer to the hospital can be handled effectively by ambulance service. A long-term study of the service shows that the response time from receipt of the dispatch call to arrival at the scene of the emergency is normally distributed with mean 42 minutes and standard deviation 8 minutes. For a randomly received call, what is the probability that the response time will be

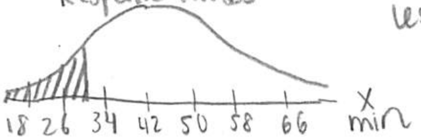
   a. between 30 and 40 minutes?
   b. less than 30 minutes?
   c. more than 60 minutes?

a) $z = \frac{40-42}{8} = -0.25$ $\quad z = \frac{30-42}{8} = -1.5$

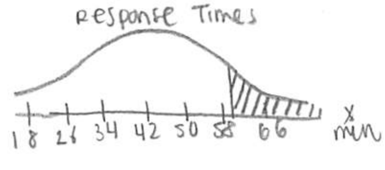
$P(-1.5 < z < -0.25) = 0.4013 - 0.0668 = 0.3345$
∃x a 33.45% chance that response time is between 30 and 40 min.



b) $z = -1.5$
$P(z < -1.5) = 0.0668$ ∃x a 6.68% chance that response time is less than 30 min.



c) $z = \frac{60-42}{8} = 2.25$
$P(z > 2.25) = 1 - 0.9878 = 0.0122$
∃x a 1.22% chance that response time is more than 60 minutes.



SUN

8.41 a) P(no one wins) = $[(0.50)^2\ 0.50] \times 2 = 0.25 = q$

$\exists$x a 25% chance that no one will win on a given coin toss.

b) P = 0.75

c) X = number of coin tosses until someone wins

Geometric, b/c we're interested in tosses required UNTIL win.

d)

X	1	2	3	4	5
pdf P(X)	0.75	0.1875	0.04688	0.01172	0.002930
cdf P(X	0.75	0.9375	0.9844	0.9961	0.9990

e) $P(X \leq 2) = 0.9375$ $\exists$x a 93.75% chance that it takes no more than 4 rounds for someone to win.

f) $P(X > 4) = 0.25^4 = 0.003906$ $\exists$x a 0.3906% chance that it takes more than four rounds for someone to win.

g) $\mu = 1/p = 1/0.25 = 4$ tosses

8.44 a) $P(X=1) = 0.325$ $\exists$x a 32.5% chance that he will hit on his first at-bat.

b) $P(X \leq 3) = 0.6925$ $\exists$x a 69.25% chance that it will take him at most 3 at-bats to get his first hit.

c) $P(X > 4) = (1-0.325)^4 = 0.2076$ $\exists$x a 20.76% chance that it will take him more than 4 at-bats to get his first hit.

SUN

8.24 a) Geometric, success: tail and failure: heads
A trial is a coin flip, $p: 0.50$, $q: 0.50$

b) Not geometric, because the probability of success can change based on pressure and confidence and thus, there may be some dependence. Also, the variable of interest is not the number of trials until success, which is crucial for a geometric distribution.

c) Geometric, success: jack, failure: any other card, a trial is drawing a card from the deck, $p = 4/52 = 1/13$, $q = 12/13$

until we win, no set n, independent

d) Geometric, success: winning lottery, failure: losing lottery, a trial is each time the player plays match 6. $p = 720/5082517440 = 1.42 \times 10^{-7} \approx 0$

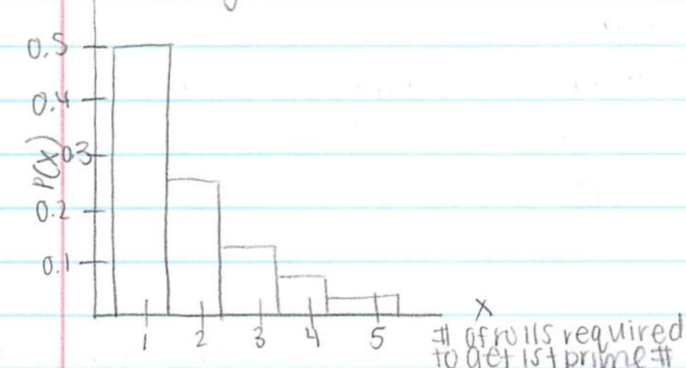
e) Not geometric because since there are no replacements, the probability of success changes with each observations and they are dependent

8.25 a) it is geometric because a success is rolling a prime number and a failure is every other trial, p is the same (1/2), and it is independent. we are also looking for the number of rolls needed for the 1st prime number.

b)

X	1	2	3	4	5
P(X)	0.5	0.25	0.125	0.0625	0.03125

c) Rolling a prime #

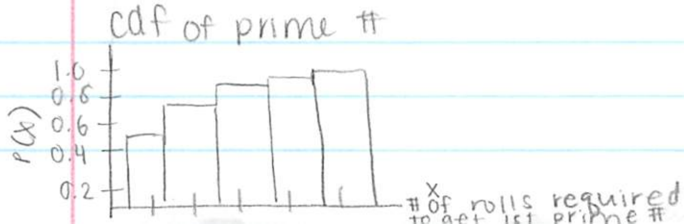


d)

X	1	2	3	4	5
cdf, P(X)	0.5	0.75	0.875	0.9375	0.9688

e) sum $= \frac{a}{1-r}$
$= \frac{0.5}{1-0.5} = 1$

cdf of prime #



8.26 a) It is geometric because it is either defective or not (2 outcomes), $p = 0.03$ and is constant, all observations are independent, and no set n
X = number of units tested required to get first defective hard drive
b) $P(X=5) = (0.97)^4\, 0.03 = 0.02656$
∃X a 2.06% chance that the first defective hard drive is the fifth unit tested.
c)

X	1	2	3	4	5
P(X)	0.03	0.0291	0.02823	0.02738	0.02656

8.32 a) S: gets question correct
F = gets question wrong
$p = 0.2$
X = number of questions before 1st correct answer
b) $P(x=5) = (0.80)^4\, 0.20$
$= 0.08192$ ∃X an 8.192% chance that Carla's first correct answer occurs on problem 5.
c) $P(X > 4) = (1-0.20)^4$
$= 0.4096$ ∃X a 40.96% chance that it will take more than 4 questions before Carla answers one correctly.
d)

X	1	2	3	4	5
P(X)	0.2	0.16	0.128	0.1024	0.08192

e) $\mu = 1/p = 1/0.2 = 5$ questions

8.37 $p = 346/869 \approx 0.3982$
$q = 523/869 \approx 0.6018$
$\mu = (25)(0.3982) = 9.955$
$\sigma = \sqrt{\mu q} = \sqrt{9.955 \times 0.6018} = 2.45$
$P(X \le 5) = 0.03058$
∃X a 3.058% chance that a group will have no more than 5 alcohol-related road fatalities.

Section 5.3 Counting Techniques (Optional)

FREE-RESPONSE QUESTIONS

Open-Ended Questions (page 112)

1. $5 \cdot 7 \cdot 4 = 140$
2. $26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 = 17{,}576{,}000$
3. $41 \cdot 41 \cdot 41 = 68{,}921$
4. $9 \cdot 5 \cdot 10 \cdot 5 = 2{,}250$. *Note:* Only 9 numbers are possible for the 1st digit.
5. $4 \cdot 3 \cdot 2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 576$
6. a. 10
   b. 10
   c. 210
   d. 210
   e. 2,598,960
   f. 2,598,960
7. a. $\binom{100}{96} = \binom{100}{4} = 3{,}921{,}225$
   b. $\binom{250}{247} = \binom{250}{3} = 2{,}573{,}000$
   c. $\binom{n}{r} = \binom{n}{n-r}$
8. a. $\binom{10}{4} = 210$
   b. Order matters since each order will give a different set of officers. Therefore, we have ${}_{10}P_4 = 5{,}040$.
9. $\binom{10}{3}\binom{8}{2}\binom{5}{1} = 16{,}800$
10. $\dfrac{\binom{150}{0}\binom{40+35+20}{12}}{\binom{245}{12}} = 7.36 \times 10^{-6}$, very unlikely.
11. $\dfrac{\binom{4}{3}\binom{4}{2}\binom{44}{0}}{\binom{52}{5}} = 9.23 \times 10^{-6}$

Section 5.4 Binomial and Geometric Probability

MULTIPLE-CHOICE QUESTIONS (page 118)

1. **D** For 8 games, this situation follows *Binomial*(8, .4).
2. **A or C** We are counting the number of arrangements for a specific order.
3. **B** This is a geometric situation.
4. **C** This is only a characteristic of a binomial.
5. **C** $nPr = r! \cdot nCr$

FREE-RESPONSE QUESTIONS

Open-Ended Questions (page 118)

1. a. Binomial: probability is the same on each trial; each trial is independent; fixed number of trials.
   b. Let X = the number of homeruns in the 50 at-bats.
   c. $P(X) = \frac{2}{11}$; number of trials = 50
   d. $\{0, 1, 2, \ldots, 50\}$
2. a. Let X = number of boxes until you get a picture of Babe Ruth
   b. Probability on each trial is the same; independent trials; X = number of trials.
   c. $\{1, 2, \ldots\}$

It is recommended that these probabilities be calculated directly as indicated and checked using the functions on the *DISTR* menu of the TI-83.

3. $\binom{10}{7}(.6^7)(.4^3) = .2150$
4. $\binom{30}{15}(.79^{15})(.21^{15}) = .00031$
5. No, because p may not be the same on all trials; each trial may not be independent as well.
6. $\binom{5}{3}(.5^3)(.5^2) = .3125$
7. $\binom{6}{0}(.75^0)(.25^6) + \binom{6}{1}(.75^1)(.25^5) + \ldots + \binom{6}{5}(.75^5)(.25^1)$
   $= binomcdf(6, .75, 5)$
   $= 1 - \binom{6}{6}(.75^6)(.25^0)$
   $= .8220$
8. $\binom{10}{8}(.7^8)(.3^2) + \binom{10}{9}(.7^9)(.3^1) + \binom{10}{10}(.7^{10})(.3^0)$
   $= binomcdf(10, .7, 10) - binomcdf(10, .7, 6) = .6496$
9. a. Geometric; $.8(.2) = .16$
   b. Geometric; $(.8^9)(.2) = .0268$
   c. P(getting Babe Ruth) = .2; P(getting Mickey Mantle) = .2. Therefore, P(getting neither Babe nor Mickey) = .6
   Therefore, $p = .2 + (.6)(.2) + (.6^2)(.2) + \ldots$
   $= \sum_{i=1}^{\infty} .2(.6)^{i-1}$ that is, an infinite geometric series
   $= \dfrac{.2}{1 - .6} = .5$
   Or note that since the probability of getting either star is the same, there is a probability of 50% that one will occur before the other assuming that there are equal numbers of pictures available.

1a) Binomial, b/c of the set n

b) Random variable = # of home runs

c) It is the # of home runs out of 50 at-bats

d) $0 \leq X \leq 50$

2a) Number of boxes opened before Babe Ruth picture is in box

b) It is the number of trials <u>UNTIL</u> first success

c) $1 < x < \infty$

3. $P(r=7) = C_{10,7}\,(0.6)^7(0.4)^3$

$= 0.2150$

∃x a 21.50% chance that they will win 7 of next 10 games.

4. $P(r=15) = C_{30,15}(0.79)^{15}(0.21)^{15}$

$= 3.08 \times 10^{-4}$

0.0003

∃x a 0.0308% chance that he will complete 15 out of 30 passes.

5. No, because the trials may be dependent or the probability of success can change.

6) $P(r=3) = C_{5,3}\,(0.5)^3(0.5)^2$

$= 0.3125$

∃x a 31.25% chance that there will be exactly 3 girls.

7. $P(r \leq 5) = 1 - P(r=6)$

$= 1 - C_{6,6}\,(0.75)^6(0.25)^0$

$= 0.8220$

∃x an 82.2% chance that player will score on at most 5 of next 6 shots.

8. $P(r \geq 7) = 1 -$ binomcdf $(10, 0.7, 6)$

$= 0.6496$

∃x a 64.96% chance that all students will pass at least 7 of next 10 tests.

9a) $p = 0.20$

$P(x=2) = 0.80^1\, 0.20^1$

$= 0.16$

∃x a 16% chance that pic of Babe Ruth is in 2nd box.

b) $P(x=10) = 0.80^9\,(0.20)^1$

$= 0.02684$ ∃x a 2.684% chance that pic of Babe Ruth is in 10th box.

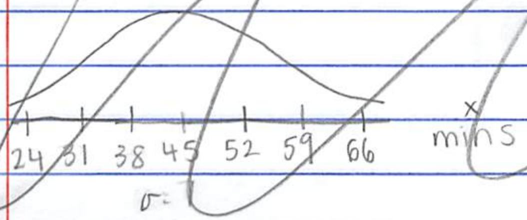
c) $\mu_{Babe} = 1/0.20 = 5$

$\mu_{Mantle} = 1/0.20 = 5$

50%, since expected value is same.

SUN

YOU GOT A 10 ON THE QUIZ 10

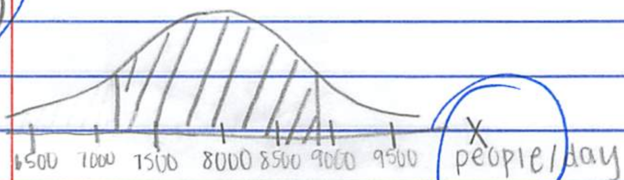


$\frac{78-45}{7} = 4.71$

$P(Z > 4.71) = 1 - 0.9999 = 0.0001$

Chapter 7 Focus Problem

30 c)



$z = \frac{8900-8000}{500} = 1.8$

$z = \frac{7200-8000}{500} = -1.6$

$P(7200 < x < 8900) = 0.9641 - 0.0548 = 0.9093$

∃x a 90.93% chance that between 7200 and 8900 people show up.

01

The central limit theorem allows us to use normal probability calculations to answer questions about sample means from many observations even when the population distribution is not normal.

The time that a technician requires to perform preventive maintenance on an air-conditioning unit is governed by the exponential distribution whose density curve appears in Figure 9.11(a). The mean time is $\mu = 1$ hour and the standard deviation is $\sigma = 1$ hour. Your company operates 70 of these units. What is the probability that their average maintenance time exceeds 50 minutes?

The central limit theorem says that the sample mean time $\bar{x}$ (in hours) spent working on 70 units has approximately the normal distribution with mean equal to the population mean $\mu = 1$ hour and standard deviation

$$\frac{\sigma}{\sqrt{70}} = \frac{1}{\sqrt{70}} = .12 \text{ hour}$$

The distribution of $\bar{x}$ is therefore approximately $N(1, 0.12)$. Figure 9.12 shows this normal curve (solid) and also the actual density curve of $\bar{x}$ (dashed).

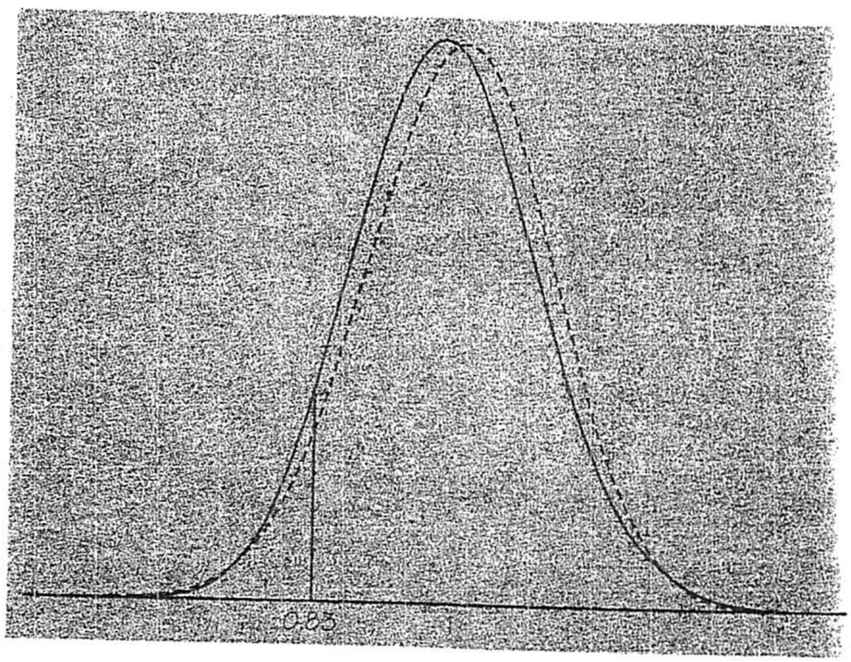


FIGURE 9.12 The exact distribution (*dashed*) and the normal approximation from the central limit theorem (*solid*) for the average time needed to maintain an air conditioner, for Example 9.9.

Because 50 minutes is 50/60 of an hour, or 0.83 hour, the probability we want is

$$P(\bar{x} > .83) = P\left(\frac{\bar{x} - 1}{.12} > \frac{.83 - 1}{.12}\right)$$

$$= P(Z > -1.42) = .9222$$

This is the area to the right of 0.83 under the solid normal curve in Figure 9.12. The exactly correct probability is the area under the dashed density curve in the figure. It is 0.9294. The central limit theorem normal approximation is off by only about 0.007.

EXERCISES

9.30 The scores of students on the ACT college entrance examination in a recent year had the normal distribution with mean $\mu = 18.6$ and standard deviation $\sigma = 5.9$.

(a) What is the probability that a single student randomly chosen from all those taking the test scores 21 or higher?

(b) Now take an SRS of 50 students who took the test. What is the probability that the mean score $\bar{x}$ of these students is 21 or higher?

9.31 A bottling company uses a filling machine to fill plastic bottles with cola. The bottles are supposed to contain 300 milliliters (ml). In fact, the contents vary according to a normal distribution with mean $\mu = 298$ ml and standard deviation $\sigma = 3$ ml.

(a) What is the probability that an individual bottle contains less than 295 ml?

(b) What is the probability that the mean contents of the bottles in a six-pack is less than 295 ml?

This famous fact of probability is called the *central limit theorem*. It is much more useful than the fact that the distribution of $\bar{x}$ is exactly normal if the population is exactly normal.

CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population whatsoever with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is close to the normal distribution $N(\mu, \sigma/\sqrt{n})$ with mean μ and standard deviation $\sigma/\sqrt{n}$.

How large a sample size n is needed for $\bar{x}$ to be close to normal depends on the population distribution. More observations are required if the shape of the population distribution is far from normal.

EXAMPLE 9.8

Figure 9.11 shows the central limit theorem in action for a very nonnormal population. Figure 9.11(a) displays the density curve for the distribution of the population. The distribution is strongly right-skewed, and the most probable outcomes are near 0 at one end of the range of possible values. The mean μ of this distribution is 1 and its standard deviation σ is also 1. This particular distribution is called an *exponential distribution* from the shape of its density curve. Exponential distributions are used to describe the lifetime in service of electronic components and the time required to serve a customer or repair a machine.

Figures 9.11(b), (c), and (d) are the density curves of the sample means of 2, 10, and 25 observations from this population. As n increases, the shape becomes more normal. The mean remains at $\mu = 1$ and the standard deviation decreases, taking the value $1/\sqrt{n}$. The density curve for 10 observations is still somewhat skewed to the right but already resembles a normal curve with $\mu = 1$ and $\sigma = 1/\sqrt{10} = .32$. The density curve for $n = 25$ is yet more normal. The contrast between the shape of the population distribution and the distribution of the mean of 10 or 25 observations is striking.

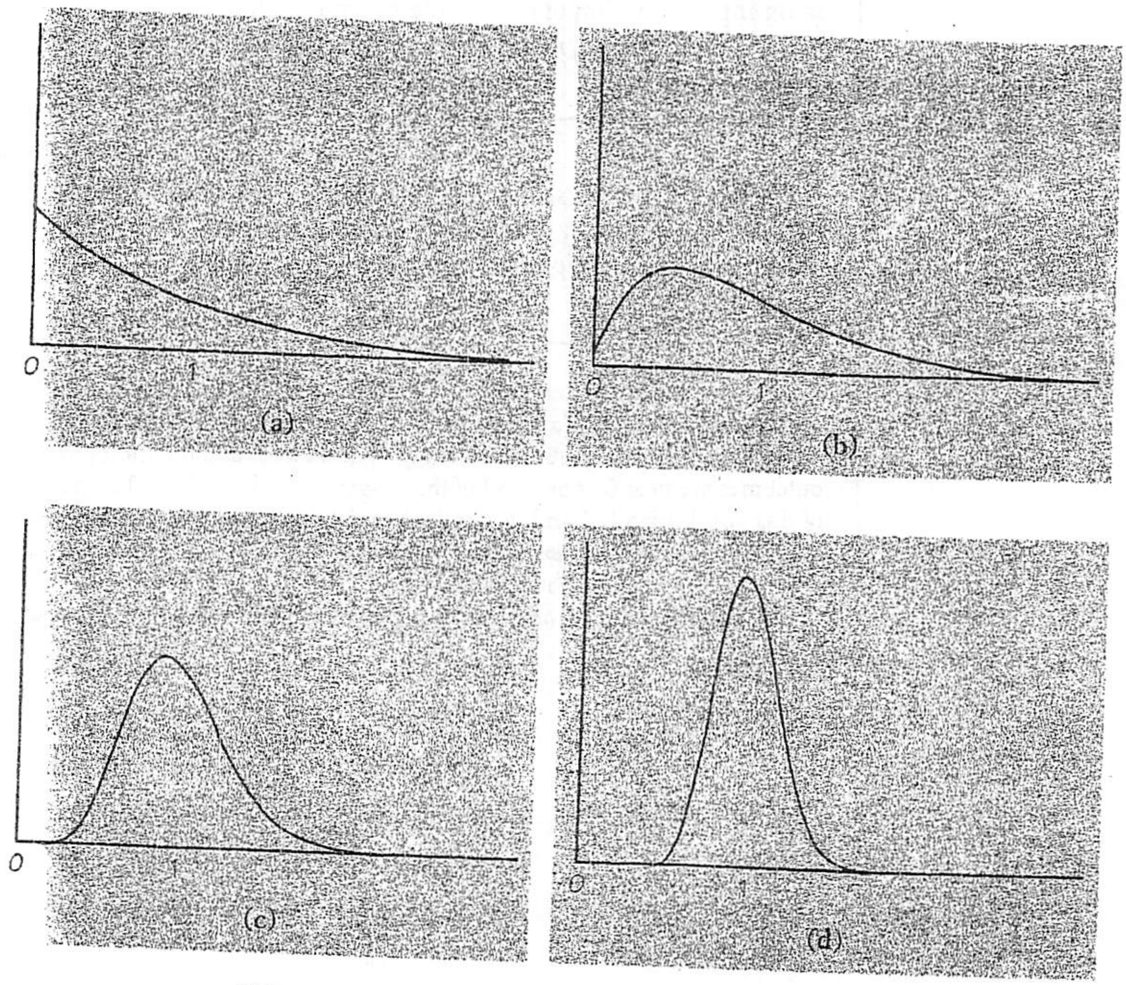


FIGURE 9.11 The central limit theorem in action: the distribution of sample means $\bar{x}$ from a strongly nonnormal population becomes more normal as the sample size increases. (*a*) The distribution of 1 observation. (*b*) The distribution of $\bar{x}$ for 2 observations. (*c*) The distribution of $\bar{x}$ for 10 observations. (*d*) The distribution of $\bar{x}$ for 25 observations.

CHAPTER REVIEW

The previous chapter introduced discrete and continuous random variables and described methods for finding means and variances, as well as rules for means and variances. This chapter focused on two important classes of discrete random variables, each of which involves two outcomes or events of interest. Both require independent trials and the same probability of success on each trial. The **binomial** random variable requires a fixed number of trials; the **geometric** random variable has the property that the number of trials varies. Both the binomial and the geometric settings occur sufficiently often in applications that they deserve special attention. Here is a checklist of the major skills you should have acquired by studying this chapter.

A. BINOMIAL

1. Identify a random variable as binomial by verifying four conditions: two outcomes (success and failure); fixed number of trials; independent trials; and the same probability of success for each trial.
2. Use TI-83 or the formula to determine binomial probabilities and construct probability distribution tables and histograms.
3. Calculate cumulative distribution functions for binomial random variables and construct cumulative distribution tables and histograms.
4. Calculate means (expected values) and standard deviations of binomial random variables.

B. GEOMETRIC

1. Identify a random variable as geometric by verifying four conditions: two outcomes (success and failure); the same probability of success for each trial; independent trials; and the count of interest is the number of trials required to get the first success.
2. Use formulas or a TI-83 to determine geometric probabilities and construct probability distribution tables and histograms.
3. Calculate cumulative distribution functions for geometric random variables and construct cumulative distribution tables and histograms.
4. Calculate expected values of geometric random variables.

CHAPTER REVIEW EXERCISES

8.37 In 1996 there were 869 road fatalities in Virginia, according to the Virginia Department of Motor Vehicles. Of these, 346 were alcohol-related. A DMV analyst wants to randomly select several groups of 25 road fatalities for further study. Find the mean and standard deviation for the number of alcohol-related road fatalities in such groups of 25. What is the probability that such a group will have no more than 5 alcohol-related road fatalities?

8.41 Three friends each toss a coin. The odd man wins; that is, if one coin comes up different from the other two, that person wins that round. If the coins all match, then no one wins and they toss again. We're interested in the number of times the players will have to toss the coins until someone wins.

(a) What is the probability that no one will win on a given coin toss?

(b) Define a success as "someone wins on a given coin toss." What is the probability of a success?

(c) Define the random variable of interest: X = number of ________. Is X binomial? Geometric? Justify your answer.

(d) Construct a probability distribution table for X. Then extend your table by the addition of cumulative probabilities in a third row.

(e) What is the probability that it takes no more than 2 rounds for someone to win?

(f) What is the probability that it takes more than 4 rounds for someone to win?

(g) What is the expected number of tosses needed for someone to win?

(h) Use the `randInt` function on your TI-83 to simulate 25 rounds of play. Then calculate the relative frequencies for X = 1, 2, 3, ... Compare the results of your simulation with the theoretical probabilities you calculated in (d).

8.42 This exercise provides visual reinforcement of the relationship between the probability of success and the mean (expected value) of a geometric random variable.

(a) Begin by completing the table below, where X = probability of success and Y = expected value.

X	.10	.20	.30	.40	.50	.60	.70	.80	.90
Y									

(b) Make a scatterplot of the points (X, Y).

(c) Enter the data into your TI-83 and perform power regression (STAT / CALC / A:PwrReg) on the data. Notice the *r*-value, and remember that the calculator transforms the data into a linear association and finds the correlation between the *transformed* values.

8.44 Suppose that Roberto, a well-known major league baseball player, finished last season with a .325 batting average. He wants to calculate the probability that he will get his first hit of this new season in his first at-bat. You define a success as getting a hit and define the random variable X = number of at-bats until Roberto gets his first hit.

(a) What is the probability that Roberto will get a hit on his first at-bat (i.e., that X = 1)?

(b) What is the probability that it will take him at most 3 at-bats to get his first hit?

(c) What is the probability that it will take him more than 4 at-bats to get his first hit?

We summarize as follows:

$P(X > n)$

The probability that it takes *more* than n trials to see the first success is

$$P(X > n) = (1 - p)^n$$

Before we had the `geometcdf` function on the TI-83, we would habitually use this result to answer questions of the form $P(X > n)$. Although the importance of this result is somewhat diminished in an age of ready access to computers and graphing calculators, it is still quite useful.

EXERCISES

8.27 Consider the following experiment: flip a coin until a head appears.

(a) Use the TI-83's `geometpdf(` command to construct the p.d.f. table for this experiment. Then have the calculator plot the probability histogram.

(b) Use the techniques described in this section for plotting the p.d.f. to compute the c.d.f. and plot its histogram.

8.28 (a) Plot the cumulative distribution histogram for the die-rolling experiment described in Example 8.13 with the p.d.f. table in Example 8.15.

(b) Find the probability that it takes more than 6 rolls to observe a 3.

(c) Find the smallest positive integer k for which $P(X \leq k) > .99$.

8.29 A basketball player makes 80% of her free throws. We put her on the free throw line and ask her to shoot free throws until she misses one. Let X = the number of free throws the player takes until she misses.

(a) What assumption do you need to make in order for the geometric model to apply? With this assumption, verify that X has a geometric distribution. What action constitutes "success" in this context?

(b) What is the probability that the player will make 5 shots before she misses?

(c) What is the probability that she will make at most 5 shots before she misses?

SUMMARY

A count X of successes has a **geometric distribution** in the geometric setting if the following are satisfied: each observation results in a success or a failure; each observation has the same probability p of success; observations are independent; and X counts the number of trials required to obtain the first success. A geometric random variable differs from a binomial variable because in the geometric setting the number of trials varies and the desired number of defined successes (1) is fixed in advance.

If X has the geometric distribution with probability of success p, the possible values of X are the positive integers $1, 2, 3, \ldots$. The **geometric probability** that X takes any value is

$$P(X = n) = (1 - p)^{n-1}p$$

The **mean** (expected value) of a geometric count X is $1/p$.

The probability that it takes *more* than n trials to see the first success is

$$P(X > n) = (1 - p)^n$$

SECTION 8.2 EXERCISES

8.32 Carla makes random guesses on a multiple choice test that has five choices for each question. We want to know how many questions Carla answers until she gets one correct.

(a) Define a success in this context, and define the random variable X of interest. What is the probability of success?

(b) What is the probability that Carla's first correct answer occurs on problem 5?

(c) What is the probability that it takes more than 4 questions before Carla answers one correctly?

(d) Construct a probability distribution table for X.

(e) If Carla took a test like this test many times and randomly guessed at each question, what would be the average number of questions she would have to answer before she answered one correctly?

8.33 In some cultures, it is considered very important to have a son to carry on the family name. Suppose that a couple in one of these cultures plans to have children until they have exactly one son.

(a) Find the average number of children per family in such a culture.

(b) What is the expected number of girls in this family?

(c) Describe a simulation that could be used to find approximate answers to the questions in (a) and (b).

The expected value and other noteworthy properties of the geometric random variable

If you're flipping a fair coin, how many times would you expect to have to flip the coin in order to observe the first head? If you're rolling a die, how many times would you expect to have to roll the die in order to observe the first 3? If you said 2 coin tosses and 6 rolls of the die, then your intuition is serving you well. Here is the principle.

THE MEAN OF A GEOMETRIC RANDOM VARIABLE

If X is a geometric random variable with probability of success p on each trial, then the **mean**, or **expected value**, of the random variable, that is, the expected number of trials required to get the first success, is $\mu = 1/p$.

The demonstration of the preceding fact proceeds as follows. The notation will be simplified if we let p = probability of success and let q = probability of failure. Then $q = 1 - p$ and the probability distribution table looks like this:

X	1	2	3	4	...
P(X)	p	pq	pq^2	pq^3	...

The mean (expected value) of X is calculated as follows:

$$\begin{aligned}
\mu &= 1(p) + 2(pq) + 3(pq^2) + 4(pq^3) + \cdots \\
&= p(1 + 2q + 3q^2 + 4q^3 + \cdots) \\
&= p\left(\frac{1}{1 - 2q + q^2}\right) \\
&= p\left[\frac{1}{(1-q)^2}\right] \\
&= p\left(\frac{1}{p^2}\right) \\
&= \frac{1}{p}
\end{aligned}$$

There is another interesting result that relates to the probability that it takes *more* than a certain number of trials to achieve success. Here are the steps:

$$\begin{aligned}
P(X > n) &= 1 - P(X \le n) \\
&= 1 - (p + qp + q^2p + \cdots + q^{n-1}p) \\
&= 1 - p(1 + q + q^2 + \cdots + q^{n-1}) \\
&= 1 - p\left(\frac{1 - q^n}{1 - q}\right) \\
&= 1 - p\left(\frac{1 - q^n}{p}\right) \\
&= 1 - (1 - q^n) \\
&= q^n = (1 - p)^n
\end{aligned}$$

Using the setting of Example 8.13, let's calculate some probabilities:

$X = 1: \quad P(X = 1) = P(\text{success of first roll}) = 1/6$

$$
\begin{aligned}
X = 2: \quad P(X = 2) &= P(\text{success of second roll}) \\
&= P(\text{failure on first roll and success on second roll}) \\
&= P(\text{failure on first roll}) \times P(\text{success on second roll}) \\
&= (5/6) \times (1/6)
\end{aligned}
$$

(since trials are independent).

$$
\begin{aligned}
X = 3: \quad P(X = 3) &= P(\text{failure on first roll}) \times P(\text{failure on second roll}) \\
&\quad \times P(\text{success on third roll}) \\
&= (5/6) \times (5/6) \times (1/6)
\end{aligned}
$$

Continue the process. The pattern suggests that a general formula for the variable X is

$$P(X = n) = (5/6)^{n-1}(1/6)$$

Now we can state the following principle:

RULE FOR CALCULATING GEOMETRIC PROBABILITIES

If X has a geometric distribution with probability p of success and $(1 - p)$ of failure on each observation, the possible values of X are 1, 2, 3, ... If n is any one of these values, then the probability that the first success occurs on the nth trial is

$$P(X = n) = (1 - p)^{n-1}p$$

Although the setting for the geometric distribution is very similar to the binomial setting, there are some striking differences. In rolling a die, for example, it is possible that you will have to roll the die many times before you roll a 3. In fact, it is theoretically possible to roll the die forever without rolling a 3 (although the probability gets closer and closer to 0 the longer you roll the die without getting a 3). The probability of observing the first 3 on the fiftieth roll of the die is $P(X = 50) = .0000$.

A probability distribution table for the geometric random variable is strange indeed because it never ends; that is, the number of table entries is infinite. The rule for calculating geometric probabilities shown above can be used to construct the table:

EXERCISES

8.24 For each of the following, determine if the experiment describes a geometric distribution. If it does, describe the two events of interest (success and failure), what constitutes a trial, and the probability of success on one trial. If the random variable is not geometric, identify a condition of the geometric setting that is not satisfied.

(a) Flip a coin until you observe a tail.

(b) Record the number of times a player makes both shots in a one-and-one foul-shooting situation. (In this situation, you get to attempt a second shot only if you make your first shot.)

(c) Draw a card from a deck, observe the card, and replace the card within the deck. Count the number of times you draw a card in this manner until you observe a jack.

(d) Buy a "match 6" lottery ticket every day until you win the lottery. (In a "match 6" lottery, a player chooses 6 different numbers from the set $\{1, 2, 3, \ldots, 44\}$. A lottery representative draws 6 different numbers from this set. To win, the player must match all 6 numbers, in any order.)

(e) There are 10 red marbles and 5 blue marbles in a jar. You reach in and, without looking, select a marble. You want to know how many marbles you will have to draw (without replacement), on average, in order to be sure that you have 3 red marbles.

8.25 An experiment consists of rolling a die until a prime number (2, 3, or 5) is observed. Let X = number of rolls required to get the first prime number.

(a) Verify that X has a geometric distribution.

(b) Construct a probability distribution table to include at least 5 entries for the probabilities of X. Record probabilities to four decimal places.

(c) Construct a graph of the p.d.f. of X.

(d) Compute the c.d.f. of X and plot its histogram.

(e) Use the formula for the sum of a geometric sequence to show that the probabilities in the p.d.f. table of X add to 1.

8.26 Suppose we have data that suggest that 3% of a company's hard disk drives are defective. You have been asked to determine the probability that the first defective hard drive is the fifth unit tested.

(a) Verify that this is a geometric setting. Identify the random variable; that is, write X = number of ________ and fill in the blank. What constitutes a success in this situation?

(b) Answer the original question: What is the probability that the first defective hard drive is the fifth unit tested?

(c) Find the first four entries in the table of the p.d.f. for the random variable X.

Exploring geometric distributions with the TI-83

The TI-83 command `geometpdf` (under 2nd / DISTR) takes two arguments: the probability p of success and the number of the trial on which the first success occurs. Consider the roll of a die of Example 8.13. The probability of rolling a 3 (success) is 1/6. So it should come as no surprise that `geometpdf(1/6,1)` gives the answer 0.1666666667, or 1/6. The next entry in the table is `geometpdf(1/6,2)`, which returns 0.1388888889, or 5/36. The second argument can also be a list, such as `geometpdf(.5,{1,2,3,4,5})` or, if you have values of X entered into list L_1, `geometpdf(.5,L₁)`.

Here is an efficient way to quickly construct a p.d.f. table and plot the result as a histogram. From the Home screen, enter the value of X into L_1 with the command `seq(X,X,1,10,1)→L₁`. (We can't list all of the terms; we arbitrarily stop at 10.) Next, enter the probabilities into list L_2 with the command `geometpdf(1/6,L₁)→L₂`.

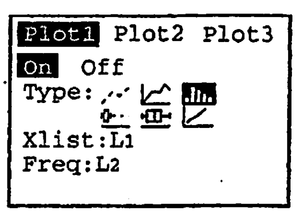
L1	L2	L3 2
1	.16667	------
2	.13889	
3	.11574	
4	.09645	
5	.08038	
6	.06698	
7	.05582	

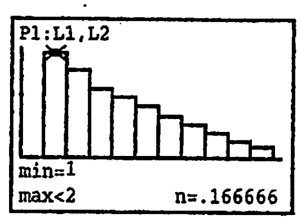
L2(1)=.1666666666...

Before you plot the probability histogram, you will want to specify the dimensions of an appropriate viewing window. Scanning the list of values gives you insight into reasonable dimensions for the window. The following appear to be good choices: X[-1, 11]$_1$ and Y[-.05, .2]$_{.1}$.

```
WINDOW
 Xmin=-1
 Xmax=11
 Xscl=1
 Ymin=-.05
 Ymax=.2
 Yscl=.1
 Xres=1
```

When you specify a histogram for the STAT / PLOT, specify list L_1 as Xlist, and specify list L_2 for the frequency. The resulting plot shows that the distribution is strongly right-skewed.





Suppose we are interested in finding the probability that it would take at most 6 rolls of the die to produce a 3. The c.d.f. can be used to answer questions like this. Recall that if $F(X)$ is the c.d.f. for the die experiment and X_0 is a positive integer, then $F(X_0)$ is defined as the sum of the probabilities of all positive integers less than or equal to X_0. The TI-83 command `geometcdf(1/6,6)` calculates the cumulative probability $F(6)$ for the first 6 values of X and reports the result as 0.6651020233.

X	1	2	3	4	5	6	7	...
P(X)	p	$(1-p)p$	$(1-p)^2p$	$(1-p)^3p$	$(1-p)^4p$	$(1-p)^5p$	$(1-p)^6p$	...

The probabilities (i.e., the entries in the second row) are the terms of a *geometric sequence* (hence the name for this random variable). You may recall from your study of algebra that the general form for a geometric sequence is

$$a,\ ar,\ ar^2,\ ar^3,\ \ldots,\ ar^{n-1},\ \ldots$$

where a is the first term, r is the ratio of one term in the sequence to the next, and the nth term is ar^{n-1}. You may also recall that even though the sequence continues forever, and even though you could never finish adding the terms, the sequence does have a sum (one of the implausible truths of the infinite!). This sum is

$$\frac{a}{1-r}$$

In order for the geometric random variable to have a valid p.d.f., the probabilities in the second row of the table must add to 1. Using the formula for the sum of a geometric sequence, we have

$$\sum_{i=1}^{\infty} P(x_i) = p + (1-p)p + (1-p)^2p + \ldots$$

$$= \frac{p}{1-(1-p)} = \frac{p}{p} = 1$$

EXAMPLE 8.15

The rule for calculating geometric probabilities can be used to construct a probability distribution table for X = number of rolls of a die until a 3 occurs:

X	1	2	3	4	5	6	7	...
P(X)	.1667	.1389	.1157	.0965	.0804	.0670	.0558	...

Here's one way to find these probabilities with the TI-83:

1. Enter the probability of success, 1/6. Press ENTER.
2. Enter ∗(5/6) and press ENTER.
3. Continue to press ENTER repeatedly.

```
1/6
         .1666666667
Ans*(5/6)
         .1388888889
         .1157407407
         .0964506173
         .0803755144
```

If you'd like to see the probabilities as fractions, modify step 2: enter *(5/6)▶FRAC and press ENTER. Verify that the entries in the second row are as shown:

X	1	2	3	4	...
P(X)	1/6	5/36	25/216	125/1296	...

Figure 8.3 is a graph of the distribution of X. As you might expect, the probability distribution histogram is strongly skewed to the right with a peak at the leftmost value, 1. It is easy to see why this must be so, since the height of each bar after the first is the height of the previous bar times the probability of failure $1 - p$. Since you're multiplying the height of each bar by a number less than 1, each new bar will be shorter than the previous bar, and hence the histogram will be right-skewed. Always.

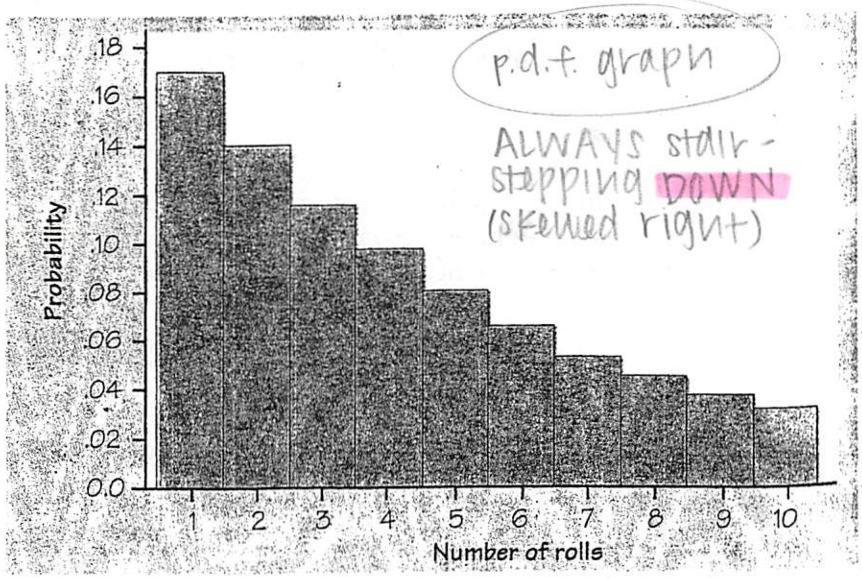


FIGURE 8.3 Probability histogram for the geometric distribution in Example 8.15.

A random variable X is geometric provided that the following conditions are met.

THE GEOMETRIC SETTING

1. Each observation falls into one of just two categories, which for convenience we call "success" or "failure."
2. The probability of a success, call it p, is the same for each observation.
3. The observations are all **independent**.
4. The variable of interest is the number of trials required to obtain the first success.

EXAMPLE 8.13

An experiment consists of rolling a single die. The event of interest is rolling a 3; this event is called a success. The random variable is defined as X = the number of trials until a 3 occurs. To verify that this is a geometric setting, note that rolling a 3 will represent a success, and rolling any other number will represent a failure. The probability of rolling a 3 on each roll is the same: 1/6. The observations are independent. A trial consists of rolling the die once. We roll the die until a 3 appears. Since all of the requirements are satisfied, this experiment describes a geometric setting.

EXAMPLE 8.14

Suppose you repeatedly draw cards without replacement from a deck of 52 cards until you draw an ace. There are two categories of interest: ace = success; not ace = failure. But is the probability of success the same for each trial? No. The probability of an ace on the first card is 4/52. If you don't draw an ace on the first card, then the probability of an ace on the second card is 4/51. Since the result of the first draw affects probabilities on the second draw (and on all successive draws required), the trials are not independent. So this is not a geometric setting.

Using the setting of Example 8.13, let's calculate some probabilities.

$X = 1:\quad P(X = 1) = P(\text{success of first roll}) = 1/6$

$X = 2:\quad P(X = 2) = P(\text{success of second roll})$
$= P(\text{failure on first roll and success on second roll})$
$= P(\text{failure on first roll}) \times P(\text{success on second roll})$
$= (5/6) \times (1/6)$

(since trials are independent).

$X = 3:\quad P(X = 3) = P(\text{failure on first roll}) \times P(\text{failure on second roll}) \times P(\text{success on third roll})$
$= (5/6) \times (5/6) \times (1/6)$

Continue the process. The pattern suggests that a general formula for the variable X is

$$P(X = n) = (5/6)^{n-1}(1/6)$$

Now we can state the following principle:

RULE FOR CALCULATING GEOMETRIC PROBABILITIES

If X has a geometric distribution with probability p of success and $(1 - p)$ of failure on each observation, the possible values of X are 1, 2, 3, ... If n is any one of these values, then the probability that the first success occurs on the nth trial is

$$P(X = n) = (1 - p)^{n-1}p$$

↑
Q

Although the setting for the geometric distribution is very similar to the binomial setting, there are some striking differences. In rolling a die, for example, it is possible that you will have to roll the die many times before you roll a 3. In fact, it is theoretically possible to roll the die forever without rolling a 3 (although the probability gets closer and closer to 0 the longer you roll the die without getting a 3). The probability of observing the first 3 on the fiftieth roll of the die is $P(X = 50) = .0000$.

A probability distribution table for the geometric random variable is strange indeed because it never ends; that is, the number of table entries is infinite. The rule for calculating geometric probabilities shown above can be used to construct the table: