



- (c) Which prediction do you prefer: (a) or (b)? Explain your answer.
- (d) Someone looking at the numerical summaries and not the plots for these analyses says that because both models have very high values of r^2 , they should perform equally well in doing this prediction. Write a response to this comment.
- (e) Discuss the value of graphical summaries and the problems of extrapolation using what you have learned in studying these salary data.

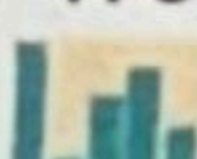
2.136 Faculty salaries. Here are the salaries for a sample of professors in a mathematics department at a large midwestern university for the academic years 2019–2020 and 2020–2021:  **FACULTY**

2019–2020 salary (\$)	2020–2021 salary (\$)	2019–2020 salary (\$)	2020–2021 salary (\$)
160,600	163,700	151,650	154,200
127,700	130,660	147,160	150,140
124,120	126,400	90,290	93,590
113,800	116,900	90,500	93,000
127,000	130,000	100,000	102,900
126,790	130,400	156,850	159,830
118,520	121,700	137,500	140,510
160,050	162,900	130,100	133,100

- (a) Construct a scatterplot with the 2020–2021 salaries on the vertical axis and the 2019–2020 salaries on the horizontal axis.
- (b) Comment on the form, direction, and strength of the relationship in your scatterplot.
- (c) What proportion of the variation in 2020–2021 salaries is explained by 2019–2020 salaries?

2.137 Find the line and examine the residuals. Refer to the previous exercise.  **FACULTY**

- (a) Find the least-squares regression line for predicting 2020–2021 salaries from 2019–2020 salaries.
- (b) Analyze the residuals, paying attention to any outliers or influential observations. Write a summary of your findings.


2.138 Bigger raises for those earning less. Refer to the previous two exercises. The 2019–2020 salaries do an excellent job of predicting the 2020–2021 salaries. Is there anything more that we can learn from these data? In this department, there is a tradition of giving higher-than-average percent raises to those whose salaries are lower. Let's see if we can find evidence to support this idea in the data.  **FACULTY**

- (a) Compute the percent raise for each faculty member. Take the difference between the 2020–2021 salary and the 2019–2020 salary, divide by the 2019–2020 salary, and then multiply by 100. Make a scatterplot with raise as the response variable and the 2019–2020 salary as the explanatory variable. Describe the relationship that you see in your plot.
- (b) Find the least-squares regression line and add it to your plot.

(c) Analyze the residuals. Are there any outliers or influential cases? Make a graphical display and include this in a short summary of your conclusions.



- (d) Is there evidence in the data to support the idea that greater percent raises are given to those with lower salaries? Include numerical and graphical summaries to support your conclusion.

2.139 Firefighters and fire damage. Someone says, "There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage." Explain why this reasoning is wrong.

2.140 Predicting text pages. The editor of a statistics text would like to plan for the next edition. A key variable is the number of pages that will be in the final version. Text files are prepared by the authors using a word processor called LaTeX, and separate files contain figures and tables. For the previous edition of the text, the number of pages in the LaTeX files can easily be determined, as well as the number of pages in the final version of the text. Here are the data:  **TEXTP**

Chapter	1	2	3	4	5	6	7	8	9	10	11	12	13
LaTeX pages	77	73	59	80	45	66	81	45	47	43	31	46	20
Text pages	99	89	61	82	47	68	87	45	53	50	36	52	19

- (a) Plot the data and describe the overall pattern.
- (b) Find the equation of the least-squares regression line and add the line to your plot.
- (c) Find the predicted number of text pages for a chapter in the next edition if the number of LaTeX pages is 52.
- (d) Write a short report for the editor explaining to her how you constructed the regression equation and how she could use it to estimate the number of pages in the next edition of the text.

 **2.141 Plywood strength.** How strong is a building material such as plywood? To be specific, support a 24-inch by 2-inch strip of plywood at both ends and apply force in the middle until the strip breaks. The modulus of rupture (MOR) is the force needed to break the strip. We would like to be able to predict MOR without actually breaking the wood. The modulus of elasticity (MOE) is found by bending the wood without breaking it. Both MOE and MOR are measured in pounds per square inch. Here are data for 32 specimens of the same type of plywood:³⁸  **MOEMOR**

MOE	MOR	MOE	MOR	MOE	MOR	MOE	MOR
2,005,400	11,591	1,774,850	10,541	2,181,910	12,702	1,747,010	11,700
1,166,360	8,542	1,457,020	10,314	1,559,700	11,209	1,791,150	11,400
1,842,180	12,750	1,959,590	11,983	2,372,660	12,799	2,535,170	13,900
2,088,370	14,512	1,720,930	10,232	1,580,930	12,062	1,355,720	9,200
1,615,070	9,244	1,355,960	8,395	1,879,900	11,357	1,646,010	8,800
1,938,440	11,904	1,411,210	10,654	1,594,750	8,889	1,472,310	6,300
2,047,700	11,208	1,842,630	10,223	1,558,770	11,565	1,488,440	9,200
2,037,520	12,004	1,984,690	13,499	2,212,310	15,317	2,349,090	13,600

Can we use MOE to predict MOR accurately? Use the data to write a discussion of this question.

(b) If you toss a fair coin four times and observe the pattern THTH, then the next toss is more likely to be a tail than a head.

(c) The quantity \hat{p} is one of the parameters for a binomial distribution.

(d) The binomial distribution can be used to model the daily number of pedestrian/cyclist near-crash events on campus.

5.28 What's wrong? For each of the following statements, explain what is wrong and why.

(a) In the binomial setting, X is a proportion.

(b) The variance for a binomial count is $\sqrt{p(1-p)/n}$.

(c) The Normal approximation to the binomial distribution is always accurate when n is very large.

(d) The binomial distribution is a good approximation of the sampling distribution of the count X when we draw an SRS of size n students from a population of $5n$ students.

5.29 Should you use the binomial distribution? In each of the following situations, is it reasonable to use a binomial distribution for the random variable X ? Give reasons for your answer in each case. If a binomial distribution applies, give the values of n and p .

(a) A poll of 200 college students asks whether or not you usually feel irritable in the morning. X is the number who reply that they do usually feel irritable in the morning.

(b) You toss a fair coin until a head appears. X is the count of the number of tosses that you make.

(c) Most calls made at random by sample surveys don't succeed in talking with a person. Of calls to New York City, only one-twelfth succeed. A survey calls 500 randomly selected numbers in New York City. X is the number of times that a person is reached.

(d) You deal 10 cards from a shuffled deck of standard playing cards and count the number X of black cards.

5.30 Should you use the binomial distribution? In each of the following situations, is it reasonable to use a binomial distribution for the random variable X ? Give reasons for your answer in each case.

(a) In a random sample of students in a fitness study, X is the mean daily exercise time of the sample.

(b) A manufacturer of running shoes picks a random sample of 20 shoes from the production of shoes each day for a detailed inspection. X is the number of pairs of shoes with a defect.

(c) A nutrition study chooses an SRS of college students. They are asked whether or not they usually eat at least five servings of fruits or vegetables per day. X is the number who say that they do.

(d) X is the number of days during the school year when you skip a class.

5.31 Random digits. Each entry in a table of random digits like Table B has probability 0.1 of being a 0, and digits are independent of each other.

(a) What is the probability that a group of five digits from the table will contain at least one digit greater than 4?

(b) What is the mean number of digits greater than 4 in lines 40 digits long?

5.32 Admitting students to college. A selective college would like to have an entering class of 900 students. Because not all students who are offered admission accept, the college admits more than 900 students. Past experience shows that about 78% of the students admitted will accept. The college decides to admit 1150 students. Assuming that students make their decisions independently, the number who accept has the $B(1150, 0.78)$ distribution. If this number is less than 900, the college will admit students from its waiting list.

(a) What are the mean and the standard deviation of the number X of students who accept?

(b) The college does not want more than 900 students. Use the Normal approximation to find the probability that more than 900 students accept.

(c) If the college decides to decrease the number of admission offers to 1100, what is the probability that more than 900 will accept?

(d) Based on your answers to parts (b) and (c), should the college admit 1100 or 1150 students? Explain your answer.

5.33 Cyberbullying. A survey of 4972 U.S. students aged 12 to 17 years reveals that 25% have received mean or hurtful comments online in the past 30 days.¹⁹ You take a random sample of 15 undergraduates and ask them whether they have received mean or hurtful comments online in the past 30 days. If the rate at your university matches this 25% rate:

(a) What is the distribution of the number of undergraduates who say that they have received hurtful comments online in the past 30 days?

(b) What is the distribution of the number of undergraduates who say that they have not received hurtful comments online in the past 30 days?

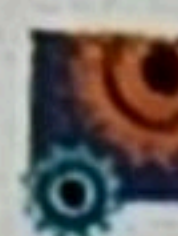
(c) What is the probability that more than 7 of the 15 undergraduates in your sample say that they have received hurtful comments online in the past 30 days?

(d) What is the probability that no more than 7 of the 15 students in your sample say that they have not received hurtful comments online in the past 30 days?

5.34 Genetics of peas. According to genetic theory, the blossom color in the second generation of a certain cross of sweet peas should be red or white in a 3:1 ratio. That is, each plant has probability $3/4$ of having red blossoms, and the blossom colors of separate plants are independent.

(a) What is the probability that exactly 8 out of 10 of these plants have red blossoms?

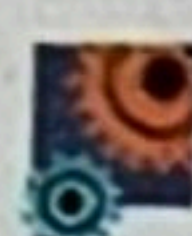
(b) What is the mean number of red-blossomed plants when 130 plants of this type are grown from seeds?

 **5.69 More on watching live television.** Consider the settings of Exercises 5.50 and 5.52.

(a) Using the reported 30% from the survey, what is the largest number m out of $n = 20$ undergraduates such that $P(X \leq m) < 0.05$? This value m (and anything smaller) represents counts that are very unlikely given $p = 0.30$.

(b) Now using the hypothesized rate of 15% and your answer to part (a), what is $P(X \leq m)$? This represents how likely this range of counts occurs when $p = 0.15$.

(c) If you were to increase the sample size from $n = 20$ to $n = 100$ and repeat the calculations of parts (a) and (b), would you expect the probability in part (b) to generally increase or decrease? Explain your answer.

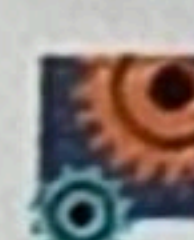
 **5.70 Iron depletion without anemia and physical performance.** Several studies have shown a link

between iron depletion without anemia (IDNA) and physical performance. In one study, the physical performance of 24 female collegiate rowers with IDNA was compared with that of 24 female collegiate rowers with normal iron status.²⁷ Several different measures of physical performance were studied, but we'll focus here on training-session duration. Assume that training-session duration of female rowers with IDNA is Normally distributed, with mean 58 minutes and standard deviation 11 minutes. Training-session duration of female rowers with normal iron status is Normally distributed, with mean 69 minutes and standard deviation 18 minutes.

(a) What is the probability that the mean duration of the 24 rowers with IDNA exceeds 63 minutes?

(b) What is the probability that the mean duration of the 24 rowers with normal iron status is less than 63 minutes?

(c) What is the probability that the mean duration of the 24 rowers with IDNA is greater than the mean duration of the 24 rowers with normal iron status?

 **5.71 Treatment and control groups.** The previous exercise illustrates a common setting for statistical inference. This exercise gives the general form of the sampling distribution needed in this setting. We have a sample of n observations from a treatment group and an independent sample of m observations from a control group. Suppose that the response to the treatment has the $N(\mu_X, \sigma_X)$ distribution and that the response of control subjects has the $N(\mu_Y, \sigma_Y)$ distribution. Inference about the difference $\mu_Y - \mu_X$ between the population means is based on the difference $\bar{y} - \bar{x}$ between the sample means in the two groups.

(a) Under the assumptions given, what is the distribution of \bar{y} ? Of \bar{x} ?

(b) What is the distribution of $\bar{y} - \bar{x}$?

PUTTING IT ALL TOGETHER

5.72 Risks and insurance. The idea of insurance is that we all face risks that are unlikely but carry high cost. Think of a fire destroying your home. So we form a group to share the risk: we all pay a small amount, and the insurance policy pays a large amount to those few of us whose homes burn down. An insurance company looks at the records for millions of homeowners and sees that the mean loss from fire in a year is $\mu = \$600$ per house and that the standard deviation of the loss is $\sigma = \$12,000$. (The distribution of losses is extremely right-skewed: most people have \$0 loss, but a few have large losses.) The company plans to sell fire insurance for \$500 plus enough to cover its costs and profit.

(a) Explain clearly why it would be unwise to sell only 100 policies. Then explain why selling many thousands of such policies is a safe business.

(b) Suppose the company sells the policies for \$700. If the company sells 50,000 policies, what is the approximate probability that the average loss in a year will be greater than \$700?

5.73 Binge drinking. The Centers for Disease Control and Prevention finds that 28% of people aged 18 to 24 years binge drank. Those who binge drank averaged 9.3 drinks per episode and 4.2 episodes per month. The study took a sample of over 18,000 people aged 18 to 24 years, so the population proportion of people who binge drank is very close to $p = 0.28$.²⁸ The administration of your college surveys an SRS of 200 students and finds that 56 binge drink.

(a) What is the sample proportion of students at your college who binge drink?

(b) If, in fact, the proportion of all students on your campus who binge drink is the same as the national 28%, what is the probability that the proportion in an SRS of 200 students is as large as or larger than the result of the administration's sample?

(c) A writer for the student paper says that the percent of students who binge drink is higher on your campus than nationally. Write a short letter to the editor explaining why the survey does not support this conclusion.

5.74 The ideal number of children. "What do you think is the ideal number of children for a family to have?" A Gallup Poll asked this question of 1020 randomly chosen adults. Roughly 41% thought that a total of three or more children was ideal.²⁹ Suppose that $p = 0.41$ is exactly true for the population of all adults. Gallup announced a margin of error of ± 4 percentage points for this poll.