Chapter

13

A Summary of Inference Methods

Objectives

In this chapter we summarize inference methods presented throughout the text. We will

- show the process of choosing an appropriate inference technique from among those presented in earlier chapters.
- consider several examples of choosing an inference method.

13.1 Introduction

In Chapters 2 and 6 through 12 we introduced many statistical methods for visually and numerically summarizing data and for making inferences. Statistics students are often overwhelmed by the number and variety of procedures that have been presented. What a statistician sees as a clearly arranged set of tools for analyzing data can appear as a blur to the novice. In this chapter we present a variety of examples that demonstrate the analysis process from exploration and summary to inference using some of the methods presented in earlier chapters. With the examples, we also provide some guidelines that are useful in deciding how to make an inference from a given set of data.

When presented with a set of data, it is useful to ask a series of questions:

1. *What question were the researchers attempting to answer when they collected these data?* Data analysis is done for a purpose: to extract information and to aid decision making. When looking at data, it helps to bear in mind the purpose for which the data were collected. For example, were the researchers trying to compare groups, perhaps patients given a new drug and patients given a placebo? Were they trying to see how two quantitative variables are related, so that they can use one variable to make predictions of the other? Were they checking whether a hypothesized model gives accurate predictions of the probabilities associated with a categorical variable? A good understanding of why the data were collected often clarifies the next question:
2. *What is the response variable in the study?* For example, if the researchers were concerned with the effect of a medication on blood pressure, then the likely response variable is Y = change in blood pressure of an individual (a continuous numeric variable). If they were concerned with whether or not a medication cures an illness, then the response variable is categorical with two levels: yes if a person is cured, no if a person is not cured, or maybe even categorical with three or more ordered levels: fully cured, improved, no improvement.

3. *What predictor variables, if any, were involved?* For example, if a new drug is being compared to a placebo, then the predictor variable is group membership: A patient is either in the group that gets the new drug or else the patient is in the placebo group. If height is used to predict weight, then height is the predictor (and weight is the response variable). Sometimes there is no predictor variable. For example, a researcher might be interested in the distribution of cholesterol levels in adults. In this case, the response variable is cholesterol level, but there is no predictor variable. (One might argue that there is a predictor: whether or not someone is an adult. If we wished to compare cholesterol levels of adults to those of children, then whether or not someone is an adult would be a predictor. But if there is no comparison to be made, so that everyone in the study is part of the same group (adults), then it is not accurate to speak of a predictor *variable*, since group membership does not vary from person to person.)

The answers to these questions help frame the analysis to be conducted. Sometimes the analysis will be entirely descriptive and will not include any statistical inference, such as when the data are not collected by way of a random sample. Even when a statistical inference is called for, there is generally more than one way to proceed. Two statisticians analyzing the same set of data may use somewhat different methods and draw different conclusions. However, there are commonly used statistical procedures in various situations. The flowchart given in Figure 13.1.1 helps to organize the inference methods that have been presented in this book.

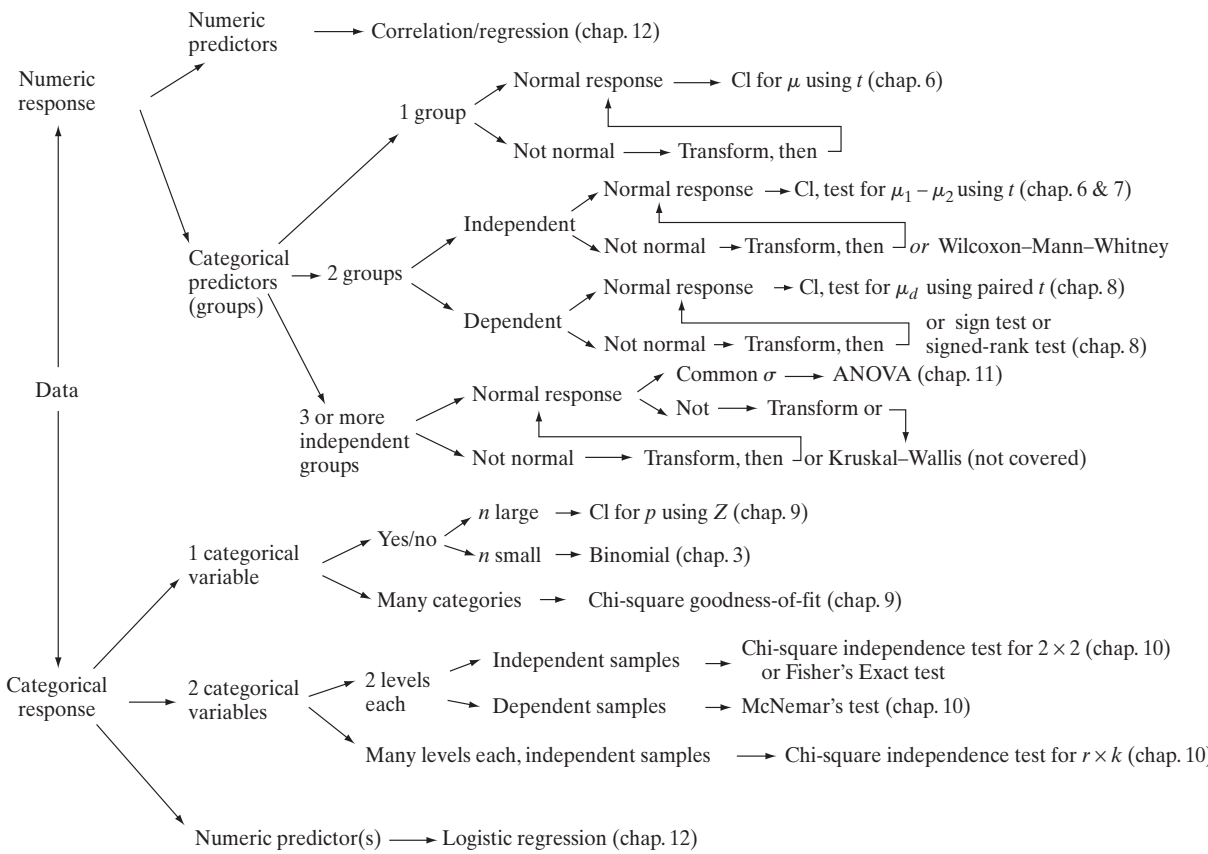


Figure 13.1.1 A flowchart of inference methods

To use this flowchart, we start by asking whether the response variable is quantitative or categorical. We then consider the type of predictor variables in the study and whether the samples collected are independent of one another or are dependent (e.g., matched pairs). Many of the methods, such as the confidence interval for a population mean presented in Chapter 6, depend on the data being from a population that has a normal distribution. (This condition is less important for large samples than it is for small samples, due to the Central Limit Theorem.) Nonnormal data can often be transformed to approximate normality and normal-based methods then applied. If such transformation fails to achieve approximate normality, then nonparametric methods, such as the Wilcoxon-Mann-Whitney test or the Wilcoxon Signed-Rank test, can be used.

Note that the flowchart only directs attention to the collection of inference methods presented in the previous chapters; this is not an exhaustive list. Beware of the Mark Twain fallacy: "When your only tool is a hammer, every problem looks like a nail." Not every statistical inference problem can be addressed with the methods presented here. In particular, these methods center on consideration of parameters, such as a population mean, μ, or proportion, p. Sometimes researchers are interested in other aspects of distributions, such as the 75th percentile. When in doubt about how to proceed in an analysis, consult a statistician.

Exploratory Data Analysis

No matter what type of analysis is being considered, it is always a good idea to start by making one or more graphs of the data. The choice of graphics depends on the type of data being analyzed. For example, when comparing two samples of quantitative data, side-by-side dotplots or boxplots are informative—both as a visual comparison of the two samples and for assessing whether or not the data satisfy the normality condition. When analyzing categorical data, bar charts are useful. When dealing with two quantitative variables, scatterplots are helpful.

Bear in mind that a statistical analysis is intended to help us understand the scientific problem at hand. Thus, conclusions should be stated in the context of the scientific study. In Section 13.2 we present some examples of data sets and the kinds of analyses that might be performed on them.

13.2 Data Analysis Examples

In this section we consider several data sets and the kinds of analyses that are appropriate for each. The three questions stated in Section 13.1 and the flowchart given in Figure 13.1.1 provide a framework for the discussion of the following examples.

Example 13.2.1

Gibberellic Acid Gibberellic acid (GA) is thought to elongate the stems of plants. Researchers conducted an experiment to investigate the effect of GA on a mutant strain of the genus *Brassica* called *ros*. They applied GA to 17 plants and applied water to 15 control plants. After 14 days they measured the growth of each of the 32 plants. For the 15 control plants the average growth was 26.7 mm, with an SD of 37.5 mm. For the 17 plants treated with GA the average growth was 92.6 mm, with an SD of 41.7 mm. The data are given in Table 13.2.1 and are graphed in Figure 13.2.1.[1]

Let us turn to the three questions stated in Section 13.1. (1) In this experiment, the researchers were trying to establish whether GA affects the growth rate of *ros*; (2) the response variable is 14-day growth of *ros*, which is numeric; (3) the predictor variable is group membership (GA group or control group) and is categorical; the two groups are independent of one another.

Table 13.2.1 Growth of *ros* plants (mm) after 14 days

	Control	GA
	3	71
	2	87
	34	117
	13	80
	6	112
	118	66
	14	128
	107	153
	30	131
	9	45
	3	38
	3	137
	49	57
	4	163
	6	47
		108
		35
Mean	26.7	92.6
SD	37.5	41.7

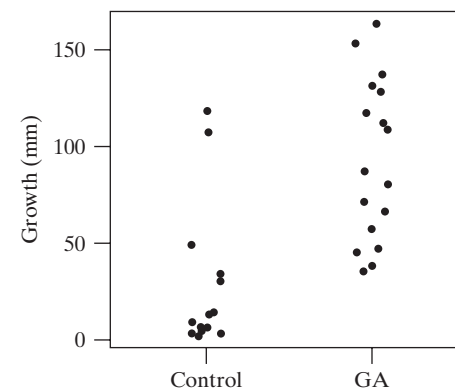
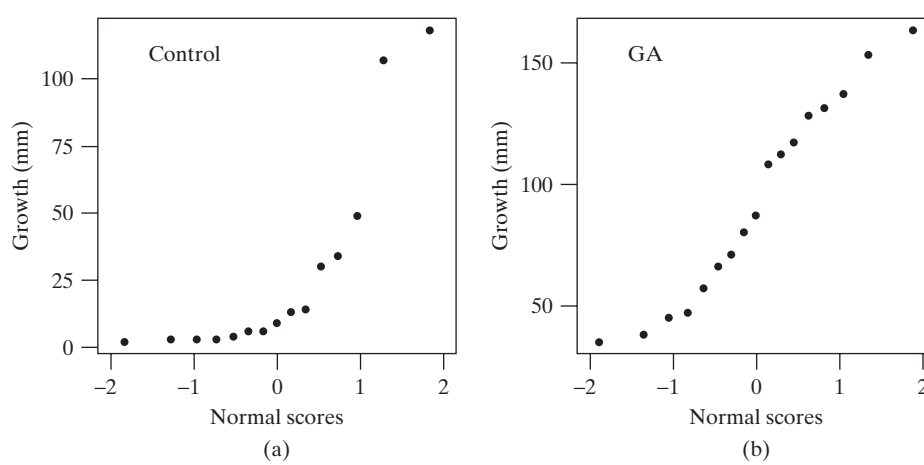


Figure 13.2.1 Dotplots of growth of *ros* plants (mm) after 14 days

The flowchart in Figure 13.1.1 directs us to consider a two-sample t test, if the data are normal or can be transformed to normality, or a Wilcoxon-Mann-Whitney test. Figure 13.2.2 shows that the distribution of the control sample of data is markedly nonnormal; thus, a transformation is called for.

Figure 13.2.2 Normal probability plots of (a) control data and (b) GA data



Taking logarithms of each of the observations produces the dotplots and normal probability plots in Figures 13.2.3 and 13.2.4.

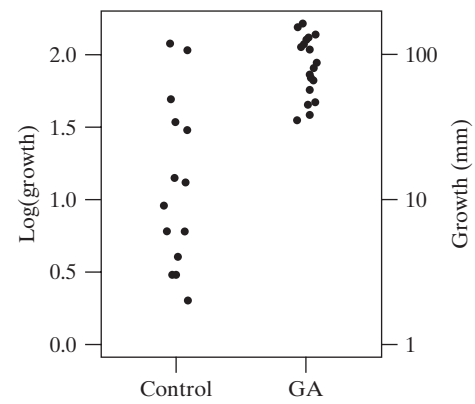


Figure 13.2.3 Dotplots of log(growth) of *ros* plants (mm) after 14 days

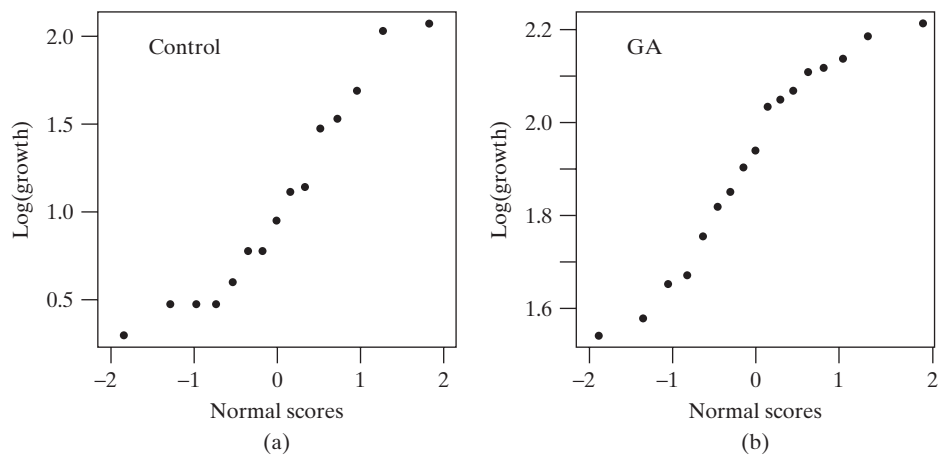


Figure 13.2.4 Normal probability plots of (a) control data and (b) GA data in log scale

In log scale the data do not show marked evidence of abnormality (Shapiro–Wilk P-values for Control and GA are 0.2083 and 0.2296, respectively), so we can proceed with a two-sample t test. The standard deviations of the two samples are clearly quite different, as can be seen from Figure 13.2.3. However, an unpooled t test is still appropriate. The following computer output shows that $t_s = -5.392$ and the P-value is very small. Thus, we have strong evidence that GA increases growth of *ros*. ■

```
Two Sample t-test

data: log10(Growth)
t = −5.3917, df = 17.445, p-value < 0.0001
alt. hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
−1.1943596, −0.5234687
```

Example 13.2.2

Whale Swimming Speed A biologist was interested in the relationship between the velocity at which a beluga whale swims and the tail-beat frequency of the whale. A sample of 19 whales was studied and measurements were made on swimming velocity, measured in units of body lengths of the whale per second (so that a value of 1.0 means that the whale is moving forward by one body length, L, per second) and tail-beat frequency, measured in units of hertz (so that a value of 1.0 means one tail-beat cycle per second).[2] Here are the data:

WHALE	VELOCITY (L/sec)	FREQUENCY (Hz)	WHALE	VELOCITY (L/sec)	FREQUENCY (Hz)
1	0.37	0.62	11	0.68	1.20
2	0.50	0.675	12	0.86	1.38
3	0.35	0.68	13	0.68	1.41
4	0.34	0.71	14	0.73	1.44
5	0.46	0.80	15	0.95	1.49
6	0.44	0.88	16	0.79	1.50
7	0.51	0.88	17	0.84	1.50
8	0.68	0.92	18	1.06	1.56
9	0.51	1.08	19	1.04	1.67
10	0.67	1.14			

It would be natural to ask, "When tails beats faster, do whales travel faster?" but the biologist conducting the study focused on the related question, "Does tail-beat frequency depend on velocity?" For the biologist's question, the response variable, frequency, is numeric, and the predictor is velocity, which is also numeric. Thus, we can consider using regression analysis to study the relationship between velocity and frequency. Figure 13.2.5 is a scatterplot of the data, which shows an increasing trend in frequency as velocity increases.

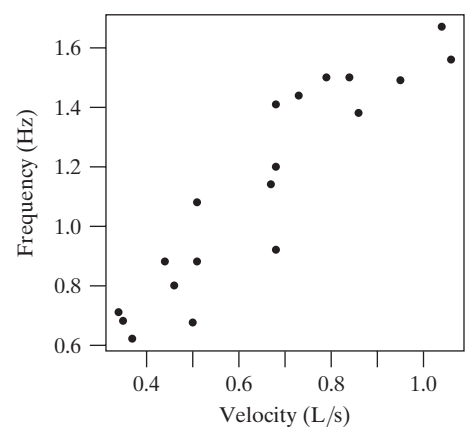


Figure 13.2.5 Scatterplot of frequency versus velocity

A regression model for these data is $Y = \beta_0 + \beta_1 X + \varepsilon$. Fitting the model to the data gives the equation $\hat{y} = 0.19 + 1.439x$, or Frequency $= 0.19 + 1.439 \times$ Velocity, as shown in the following computer output. Figure 13.2.6 shows the residual plot for this fit. The fact that this plot does not have any clear patterns in it supports the use of the regression model.

```
Coefficients:
             Estimate Std. Error t value Pr (>|t|)
(Intercept)    0.1895    0.1004    1.887    0.0763
Velocity       1.4393    0.1451    9.917    1.75e-08

Residual standard error: 0.1396 on 17 degrees of freedom
R-squared: 0.8526
```

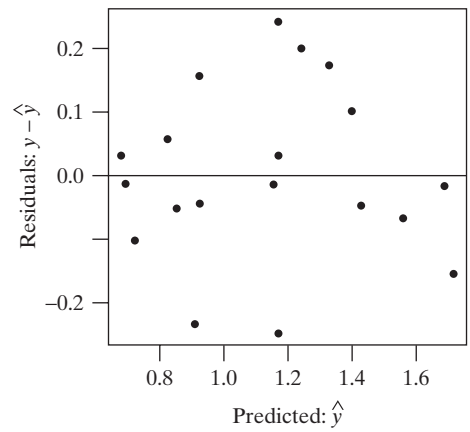


Figure 13.2.6 Residual plot for frequency regression fit

The null hypothesis

$$H_0: \beta_1 = 0$$

is tested with a t test, as shown in the regression output. A normal probability plot of the residuals, given in Figure 13.2.7, supports the use of the t test here, since it indicates that the distribution of the 19 residuals is consistent with what we would expect to see if the random errors came from a normal distribution. The t statistic has 17 degrees of freedom and a P-value of less than 0.0001. Thus, the evidence that frequency is related to velocity is quite strong; we reject the claim that the linear trend in the data arose by chance.

Continuing the analysis, the computer output shows that r^2 is 85.3%. Thus, in the sample 85.3% of the variability in frequency is accounted for by variability in velocity. (This is significantly different from zero, as indicated with the t test for $H_0: \beta_1 = 0$.) ■

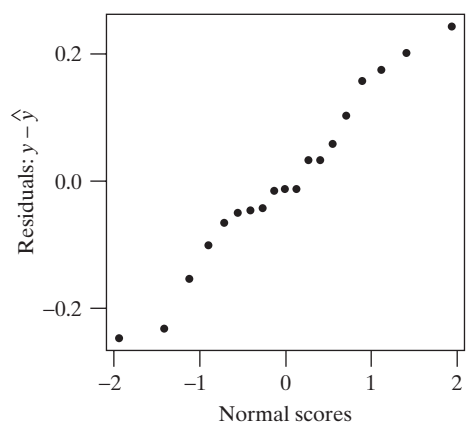


Figure 13.2.7 Normal probability plot of residuals for frequency regression fit

Example 13.2.3

Tamoxifen In a randomized, double-blind experiment the drug tamoxifen was given to 6,681 women and a placebo was given to 6,707 other women. After four years there were 89 cases of breast cancer in the tamoxifen group, compared with 175 in the placebo group.[3]

The purpose of this experiment was to determine whether tamoxifen is effective in preventing cancer. Note that because this was an experiment, and not an observational study, we can talk in terms of a cause–effect relationship. The response variable is whether or not a woman developed cancer. The predictor variable is group membership (i.e., whether or not a woman was given tamoxifen). Figure 13.2.8 is a bar chart of the data, showing that cancer was much more common in the placebo group.

These data can be organized into a 2×2 contingency table, such as Table 13.2.2. A chi-square test of independence yields $\chi^2_s = 28.2$. With 1 degree of freedom, the P-value for this test is nearly zero. There is very strong evidence that tamoxifen reduces the probability of breast cancer.

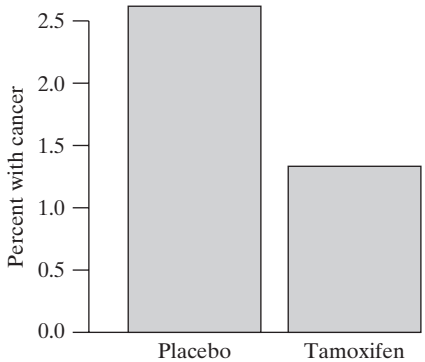


Figure 13.2.8 Bar chart of the tamoxifen data

Table 13.2.2 Tamoxifen data

	Treatment		
	Placebo	Tamoxifen	
Cancer	175	89	264
No cancer	6,532	6,592	13,124
Total	6,707	6,681	13,388

We can also construct a confidence interval with these data. Of placebo patients, $\frac{175}{6707}$ or 2.61% developed cancer so that $\tilde{p}_1 = \frac{175 + 1}{6707 + 2} = 0.0262$. Of tamoxifen patients, $\frac{89}{6681}$ or 1.33% developed cancer so that $\tilde{p}_2 = \frac{89 + 1}{6681 + 2} = 0.0135$. The standard error of the difference is

$$\text{SE}_{(\tilde{P}_1 - \tilde{P}_2)} = \sqrt{\frac{(0.0262)(1 - 0.0262)}{6707 + 2} + \frac{0.0135(1 - 0.0135)}{6681 + 2}} = 0.0024$$

A 95% confidence interval for $p_1 - p_2$ is $(0.0262 - 0.0135) \pm 1.96(0.0024)$ or $(0.0080, 0.0174)$. Thus, we are 95% confident that tamoxifen reduces the probability of breast cancer by between 0.80 and 1.74 percentage points.

We can also calculate the relative risk of cancer. The estimated relative risk is

$$\frac{\Pr\{\text{Cancer} \mid \text{Tamoxifen}\}}{\Pr\{\text{Cancer} \mid \text{Placebo}\}} = \frac{0.0261}{0.0133} = 1.96$$

Thus, we estimate that breast cancer is 1.96 times as likely when taking placebo as when taking tamoxifen. ■

Example 13.2.4

Chromosome Puffs Heat shock proteins (HSPs) are a type of protein produced by some organisms as protection against damage from exposure to high temperature. In the fruit fly *Drosophila melanogaster* the genes that encode HSPs are found on

chromosomes that uncoil and appear to puff out. This chromosome puffing can be seen under a microscope. A biologist counted the number of puffs per chromosomal arm from the salivary glands of 40 *Drosophila* larvae that had been heat shocked at 37 °C for 30 minutes, 40 larvae that had been heat shocked for 60 minutes, and 40 control larvae.

The purpose of this experiment was to determine the effect, if any, of heat shock on the HSPs. The response variable is the number of puffs on a chromosome arm, which is numeric. The predictor variable, group membership (control, 30 minutes, or 60 minutes), is categorical. Dotplots of the data are given in Figure 13.2.9; the data are summarized in Table 13.2.3.[4]

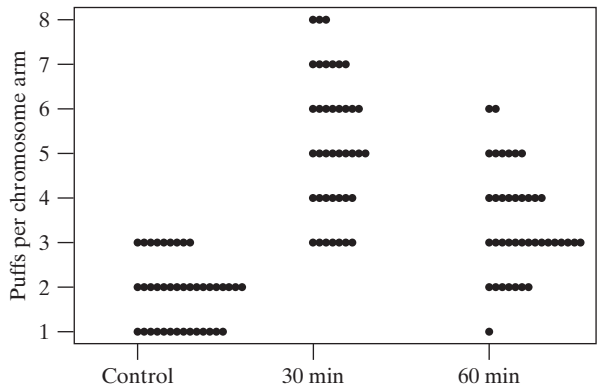


Figure 13.2.9 Dotplots of puffs per chromosome arm for *Drosophila* heat shock experiment

Table 13.2.3 Puffs per chromosome arm for *Drosophila* heat shock experiment

Group	*n*	Mean	SD
Control	40	1.88	0.76
30 min.	40	5.20	1.54
60 min.	40	3.45	1.18

The dotplots suggest an effect due to heat shock (and we can speak of an effect, not just an association, because this was an experiment). This visual impression can be confirmed with an analysis of variance. The plots also show that the distributions take on only a few values each, so that the normality condition for ANOVA is not met. Since, however, the distributions appear to be reasonably symmetric, the sample sizes moderately large and equal, and the SDs are similar among the groups, we can have confidence in the ANOVA *P*-value. The following ANOVA computer output confirms that there is strong evidence against H_0: $\mu_1 = \mu_2 = \mu_3$. We conclude that heat shock does, indeed, increase the number of puffs per chromosome arm.

```
             Df   Sum Sq   Mean Sq   F value    Pr (> F)
Group         2   221.32   110.658   76.757    < 0.0001
Residuals   117   168.68     1.442
Total       110   390.00
```

As an extension of the ANOVA, we could consider a contrast that compares the control mean to the average of the two heat shock means. ■

Example 13.2.5 Therapeutic Touch Therapeutic touch (TT) is a form of alternative medicine in which a practitioner manipulates the human energy field of the patient. However, many persons have questioned the ability of TT practitioners to detect the human energy field—and whether the human energy field even exists. An experimenter tested

the abilities of 28 TT practitioners as follows. A screen was set up between the experimenter and the practitioner, who sat on opposite sides of a table. The practitioner extended his or her hands under the screen and rested them, palms up, on the table. The researcher tossed a coin to choose one of the practitioner's hands. The experimenter then held her right hand, palm down, above the chosen hand of the practitioner. The practitioner was then asked to identify which hand had been chosen, as a test of whether the practitioner could detect a human energy field extending from the hand of the experimenter.

Each of the 28 TT practitioners was tested 10 times. The number of correct detections, "hits," in 10 trials varied from 1 to 8, with an average of 4.4. There were 123 hits in the 280 total trials.[5] Table 13.2.4 shows the distribution of hits among the 28 tested practitioners.

Table 13.2.4 Distribution of hits per ten trials in therapeutic touch experiment

Number of hits	Number of practitioners
0	0
1	1
2	1
3	8
4	5
5	7
6	2
7	3
8	1
9	0
10	0
Total	28

The goal of this experiment was to determine the ability of TT practitioners to detect the human energy field. The response variable is a yes/no (categorical) variable: yes for a hit and no for a miss. There is no predictor variable here, there is just a single group of 28 TT practitioners who were tested.

Let p denote the probability of a hit in one of the trials of the experiment. The natural null hypothesis is $H_0: p = 0.5$. One way to analyze the data would be to conduct a chi-square goodness-of-fit test of H_0 using the 280 total trials, with a directional alternative $H_A: p > 0.5$. The P-value for this test is greater than 0.50, since the data do not deviate from H_0 in the direction specified by H_A.

One might argue that p might be greater than 0.5 for some TT practitioners, but perhaps not for all of them. If p is not the same for each TT practitioner (whether or not p is 0.5 for anyone), then the chi-square goodness-of-fit test using all 280 trials is not appropriate, since the 280 trials are not independent of one another. However, the data for each of the 28 practitioners could be analyzed separately. A binomial model could be used in these analyses, since the sample size of $n = 10$ is rather small. The binomial probabilities are given in Table 13.2.5. The probability of 8 or more hits in 10 trials, for a binomial with $p = 0.5$, is $0.04395 + 0.00977 + 0.00098 = 0.0547$. Thus, if the data from each of the 28 practitioners were analyzed separately, testing $H_0: p = 0.5$ versus $H_A: p > 0.5$, the

Table 13.2.5 Observed and expected numbers (if $p = 0.5$) of hits per ten trials in the therapeutic touch experiment

Number of hits	Binomial probability	Observed number, O	Expected number, E
0	0.00098	0	0.027
1	0.00977	1	0.273
2	0.04395	1	1.231
3	0.11719	8	3.281
4	0.20508	5	5.742
5	0.24609	7	6.891
6	0.20508	2	5.742
7	0.11710	3	3.281
8	0.04395	1	1.231
9	0.00977	0	0.273
10	0.00098	0	0.027
Total	1.00000	28	27.999

smallest of the 28 P-values would be 0.0547, and again provides no significant evidence in support of H_A.*

A different way to conduct the analysis is to investigate whether the 280 observations presented in Table 13.2.4 are consistent with a binomial model. In particular, we can check the model that states that Y has a binomial distribution, with $n = 10$ and $p = 0.5$, where Y is the number of hits in 10 trials. (This is similar to the analysis presented in Section 3.9.) A goodness-of-fit test can be used here. Table 13.2.5 shows the observed numbers (from Table 13.2.4) and expected numbers for each of the 11 possible outcomes. (The expected numbers don't sum to 28 due to round-off error.)

The chi-square statistic is $\chi_s^2 = \Sigma \dfrac{(o_i - e_i)^2}{e_i} = 11.7$. The test statistic has 10 degrees of freedom, since there are 11 categories in the model. The P-value for this test is 0.306, which is quite large. Thus, the data are consistent with a binomial distribution for which $p = 0.5$ (i.e., the TT practitioners might as well have tossed coins to choose a hand, rather than trying to detect the human energy field of the experimenter). (*Note*: These data do not disprove the existence of the human energy field; they only fail to provide evidence for its existence). ■

Brief Examples

We will now consider some examples for which we will identify the type of analysis that is appropriate, but we won't conduct the analysis.

Example 13.2.6

Seastars Researchers measured the length of the longest ray on each of over 200 members of the species *Phataria unifascialis* (a seastar found in the waters of the Gulf of California, Mexico). For a sample of 184 individuals found near Loreto, the

*Considering the material in optional Section 11.9 on multiple comparisons, note that if we were to consider all 28 of these tests, we ought to require a great deal of evidence before rejecting H_0. Using the Bonferroni correction, we would require that an individual P-value be less than $\alpha_{cw} = 0.05/28 = 0.0018$ before rejecting H_0.

average length was 6.78 cm, with an SD of 1.21 cm. For a sample of 77 individuals found near Bahia de Los Angeles, the average length was 8.13 cm, with an SD of 1.33 cm.[6]

The response variable is numeric and there are two independent groups. Thus, a two-sample t test is appropriate, along with a confidence interval for the difference in population means. (*Note*: The normality condition is not essential, since the sample sizes are quite large.) ■

Example 13.2.7

Twins Researchers in Finland studied the physical activity levels of hundreds of sets of same-sex twins. In 1975 they classified subjects into the physical activity categories "exerciser" and "sedentary." They kept track of the health of the subjects through 1994, by which time there were several pairs of twins for whom one twin was alive, but the other had died. In this group there were 49 "sedentary" twins who were living, but whose "exerciser" twin pair was dead. There were 76 "exerciser" twins who were living, but whose "sedentary" twin pair was dead.[7]

The response variable in this observational study is whether or not a subject is alive, a categorical variable. The predictor is also categorical: whether the person is "sedentary" or is an "exerciser." Since the data are paired, McNemar's test is appropriate. ■

Example 13.2.8

Soil Samples Researchers took eight soil samples at each of six locations in Mediterranean pastures. They divided the samples into four pairs and put the soil in pots. One pot from each pair was watered continuously, while the other pot was watered for 13 days, then not watered for 18 days, and then watered again for 30 days. The researchers recorded the number of germinations in each pot during the experiment.[8]

This example is similar to Example 13.2.6, in that there are two samples to be compared and the response variable is numeric. However, the samples here are paired, so a paired analysis (Chapter 8) is called for. If the 24 sample differences show a normal distribution, then a paired t test or confidence interval could be used; if not, a transformation could be tried, or a Wilcoxon signed-rank or sign test could be used. ■

Example 13.2.9

Vaccinations In 1996 there was an outbreak of the disease varicella in a child care center in Georgia. Some of the children had been vaccinated against varicella, but others had not. Varicella occurred in 9 out of 66 vaccinated children and in 72 out of 82 unvaccinated children.[9]

The response and predictor variables in this experiment are both categorical. The data could be arranged into a 2×2 contingency table and analyzed with a chi-square test of independence. The difference in sample proportions is obviously quite large. However, this is an observational study and not an experiment. Thus, we cannot conclude that the difference in proportions is entirely due to the effect of the vaccine, since the effects of other variables, such as economic status, are confounded with the effect of the vaccine. ■

Example 13.2.10

Estrogen and Steroids Plasma estrone plus estradoil (Plasma E_{1+2}) steroid levels were measured in women given estrogen (Premarin) and in a control group of women. The women given estrogen were divided into three treatment groups. One group was given a daily dose of 0.625 mg, one group was given 1.25 mg, and the third group

was given 2.5 mg. The researchers noted that the plasma steroid levels were not normally distributed, but became so after a logarithm transformation was applied. In log scale, the data are given in Table 13.2.6.[10]

Table 13.2.6 Log ng/100 ml plasma E_{1+2} concentration for estrogen study

Group	n	Mean	SD
Control	30	2.01	0.27
0.625	16	2.10	0.31
1.25	24	2.34	0.39
2.5	21	2.20	0.24

The response variable in this experiment, log(plasma E_{1+2} concentration), is numeric. It has already been transformed to normality. There are four independent groups to be compared, so an analysis of variance is appropriate. A contrast that compares the control to the average of the three treatment groups would also be useful. ■

Example 13.2.11

Damselflies A researcher captured male damselflies and randomly assigned them to one of three groups. For those in the first group the sizes of red spots on the wing were artificially enlarged with red ink. For those in the second group the wing spots were enlarged with clear ink. The third group served as a control. The damselflies were then released into a contained area. The numbers surviving in each of the three groups 22 days later were determined. There were 312 damselflies in each of the three groups. After 22 days there were 41 survivors in the "artificially enlarged with red ink" group, 49 survivors in the "enlarged with clear ink" group, and 57 survivors in the control group.[11]

The response variable in this experiment, survival, is categorical, as is the predictor variable, ink status/type. These data could be arranged into a 2×3 contingency table and analyzed with a chi-square test of independence. ■

Example 13.2.12

Tobacco Use Prevention In the Hutchinson Smoking Prevention Project 40 school districts in the state of Washington were formed into 20 pairs on the basis of size, location, and prevalence of high school tobacco use as of the beginning of the study. In each pair, one district was randomly assigned to be in an intervention group and the other was assigned to the control group. If a school district was in the intervention group, then the third-grade students in the district were given a curriculum on preventing tobacco use and the teachers in the district were given special training to help students refrain from smoking. This was repeated one year later with the next new cohort of third-grade students. All the students were then followed for several years. A primary outcome measurement of the study was whether or not students were smoking two years after graduating from high school.

The experimental unit here is an entire school district, so it is natural to use as the response variable the percentage of students from a district who smoke, a numeric variable. The predictor is categorical: intervention group or control group. There are two groups, which are paired together by the design of the experiment. Out of the 20 pairs, there were 13 pairs in which the smoking rate was higher in the control district and 7 pairs in which the smoking rate was higher in the intervention district.[12] A sign test could be used to analyze these data. ■

Exercises 13.2.1–13.2.22

13.2.1 Researchers conducted a randomized, double-blind, clinical trial in which some patients with schizophrenia were given the drug clozapine and others were given haloperidol. After one year 61 of 163 patients in the clozapine group showed clinically important improvement in symptoms, compared with 51 out of 159 in the haloperidol group.[13] Identify the type of statistical method that is appropriate for these data, but do not actually conduct the analysis.

13.2.2 Consider the data of Exercise 13.2.1. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

13.2.3 A biologist collected data on the height (in inches) and peak expiratory flow (PEF—a measure of how much air a person can expire, measured in l/min) for 10 women.[14] Here are the data:

SUBJECT	HEIGHT	PEF	SUBJECT	HEIGHT	PEF
1	63	410	6	62	360
2	63	440	7	67	380
3	66	450	8	64	380
4	65	510	9	65	360
5	64	340	10	67	570

Is PEF related to height? Identify the type of statistical method that is appropriate for these data and this question, but do not actually conduct the analysis.

13.2.4 Consider the data of Exercise 13.2.3. Maria is 1 inch taller than Anika. Using the information from Exercise 13.2.3, how much greater would you predict Maria's PEF to be than Anika's?

13.2.5 A geneticist self-pollinated pink-flowered snapdragon plants and produced 97 progeny with the following colors: 22 red plants, 52 pink plants, and 23 white plants.[15] The purpose of this experiment was to investigate a genetic model that states that the probabilities of red, pink, and white are 0.25, 0.50, and 0.25. Identify the type of statistical method that is appropriate for these data, but do not actually conduct the analysis.

13.2.6 Consider the data of Exercise 13.2.5. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

13.2.7 The effect of diet on heart disease has been widely studied. As part of this general area of investigation, researchers were interested in the short-term effect of diet on endothelial function, such as the effect on triglyceride level. To study this, they designed an experiment in which 20 healthy subjects were given, in random order, a high-fat breakfast and a low-fat breakfast at 8 A.M., following a 12-hour fast, on days one week apart from each other. Serum triglyceride levels were measured on each subject before each breakfast and again four hours after each breakfast.[16] If you had access to all of the measurements collected in this experiment, how would you analyze the data?

13.2.8 Biologists were interested in the distribution of trees in a wooded area. They intended to use the number of trees per 100-square meter plot as their unit of measurement. However, they were concerned that the shapes of the plots might affect the data collection. To investigate the possibility, they counted the numbers of trees in square plots, round plots, and rectangular plots. The data are shown in the following table.[17] What type of analysis is appropriate for these data?

	PLOT SHAPE		
	SQUARE	ROUND	RECTANGULAR
	5	5	10
	5	7	2
	5	5	3
	8	2	12
	8	4	9
	7	4	5
	4	4	3
	9	7	6
	9	7	5
	7	10	3
	5	9	8
	2	2	9
	8	7	3
Mean	6.3	5.6	6.0
SD	2.14	2.47	3.27

13.2.9 Consider the data of Exercise 13.2.8. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

13.2.10 A sample of 15 patients was randomly split into two groups as part of a double-blind experiment to compare two pain relievers.[18] The 7 patients in the first group were given Demerol and reported the following numbers of hours of pain relief:

2, 6, 4, 13, 5, 8, 4

The 8 patients in the second group were given an experimental drug and reported the following numbers of hours of pain relief.

$$0, 8, 1, 4, 2, 2, 1, 3$$

How might these data be analyzed?

13.2.11 Consider the data of Exercise 13.2.10. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

13.2.12 A researcher was interested in the relationship between forearm length and, height. He measured the forearm lengths and heights of a sample of 16 women and obtained the following data.[19] How might these data be (i) visualized and (ii) analyzed?

FOREARM		FOREARM	
HEIGHT (CM)	LENGTH (CM)	HEIGHT (CM)	LENGTH (CM)
163	25.5	157	26
161	26	178	27
151	25	163	24.5
163	25	161	26
166	27.2	173	28
168	26	160	24.5
170	26	158	25
163	26	170	26

13.2.13 A randomized, double-blind, clinical trial was conducted on patients who had coronary angioplasty to compare the drug lovastatin to a placebo. The percentage of stenosis (narrowing of the blood vessels) following angioplasty was measured on 160 patients given lovastatin and on 161 patients given the placebo. For the lovastatin group the average was 46%, with an SD of 20%. For the placebo group the average was 44%, with an SD of 21%.[20] What type of analysis is appropriate for these data?

13.2.14 Consider the data of Exercise 13.2.13.

(a) Conduct an appropriate analysis of the data.

(b) Describe a graphical procedure to visualize these data.

(c) Discuss of how the data likely meet the necessary conditions for validity even though you do not have access to the raw data.

13.2.15 Researchers studied persons who had received intravenous immune globulin (IGIV) to see if they had developed infections of hepatitis C virus (HCV). In part of their analysis, they considered doses of Gammagard (an IGIV product) received by 210 patients. They divided the patients into 4 groups according to the number of doses of "Gammagard made from unscreened or first-generation anti-HCV-screened plasma." Among 48 persons who received 0 to 3 doses, there were 4 cases of HCV infection. There were 2 cases of HCV infection among 45 persons who received 4 to 20 doses, there were 7 cases of HCV infection in the 57 persons who received between 21 and 65 doses, and there were 10 cases of HCV infection among the 51 persons who received more than 65 doses.[21] What type of analysis is appropriate for these data?

13.2.16 Consider the data of Exercise 13.2.15. Conduct an appropriate analysis of the data.

13.2.17 An experiment was conducted to study the effect of tamoxifen on patients with cervical cancer. One of the measurements made, both before and again after tamoxifen was given, was microvessel density (MVD). MVD, which is measured as number of vessels per mm^2, is a measurement that relates to the formation of blood vessels that feed a tumor and allow it to grow and spread. Thus, small values of MVD are better than are large values. Data for 18 patients are shown.[22] How might these data be analyzed?

PATIENT	MVD BEFORE	MVD AFTER	PATIENT	MVD BEFORE	MVD AFTER
1	98	75	10	70	60
2	100	60	11	60	65
3	82	25	12	88	45
4	100	55	13	45	36
5	93	78	14	159	144
6	119	102	15	65	27
7	70	58	16	98	90
8	78	70	17	66	16
9	104	90	18	67	53

13.2.18 Consider the data of Exercise 13.2.17. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

13.2.19 As part of a large experiment, researchers planted 2,400 sweetgum, 2,400 sycamore, and 1,200 green ash seedlings. After 18 years, the survival rates were 93% for the sweetgum trees, 88% for the sycamore trees, and 95% for the green ash trees.[23] What type of analysis is appropriate for these data?

13.2.20 Consider the data of Exercise 13.2.19. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

13.2.21 A group of female college students were divided into three groups according to upper body strength. Their leg strength was tested by measuring how many consecutive times they could leg press 246 pounds before exhaustion. (The subjects were allowed only one second of rest between consecutive lifts.) The data are shown in the following table.[24] What type of analysis is appropriate for these data?

	UPPER BODY STRENGTH GROUP		
	LOW	MIDDLE	HIGH
	55	40	181
	70	200	85
	45	250	416
	246	192	228
	240	117	257
	96	215	316
	225		134
Mean	140	169	231
SD	93	77	112

13.2.22 Consider the data of Exercise 13.2.21. Conduct an appropriate complete analysis of the data that also includes a graphical display and discussion of how the data do or do not meet the necessary conditions for validity.

CHAPTER APPENDICES

Appendix

3.1 More on the Binomial Distribution Formula 566
3.2 Mean and Standard Deviation of the Binomial Distribution 569
4.1 Areas of Indefinitely Extended Regions 570
5.1 Relationship Between Central Limit Theorem and Normal Approximation to Binomial Distibution 572
6.1 Significant Digits 573
7.1 How Power Is Calculated 574
7.2 More on the Wilcoxon-Mann-Whitney Test 576
9.1 More on Confidence Intervals for a Proportion 578
12.1 Least-Squares Formulas 580
12.2 Derivation of Fact 12.3.1 582

Appendix 3.1
More on the Binomial Distribution Formula

In this appendix we explain more about the reasoning behind the binomial distribution formula.

The Binomial Distribution Formula

We begin by deriving the binomial distribution formula for $n = 3$. Suppose that we conduct three independent trials and that each trial results in success (S) or failure (F). Further, suppose that on each trial the probabilities of success and failure are

$$\Pr\{\text{S}\} = p$$

$$\Pr\{\text{F}\} = 1 - p$$

There are eight possible outcomes of the three trials. Reasoning as in Example 3.6.3 shows that the probabilities of these outcomes are as follows:

OUTCOME	NUMBER OF SUCCESSES	NUMBER OF FAILURES	PROBABILITY
FFF	0	3	$(1 - p)^3$
FFS	1	2	$p(1 - p)^2$
FSF	1	2	$p(1 - p)^2$
SFF	1	2	$p(1 - p)^2$
FSS	2	1	$p^2(1 - p)$
SFS	2	1	$p^2(1 - p)$
SSF	2	1	$p^2(1 - p)$
SSS	3	0	p^3

Again by reasoning parallel to Example 3.6.3, these probabilities can be combined to obtain the binomial distribution formula for $n = 3$ as shown in the table:

NUMBER OF		
SUCCESSES, j	FAILURES, $n - j$	PROBABILITY
0	3	$1p^0(1 - p)^3$
1	2	$3p^1(1 - p)^2$
2	1	$3p^2(1 - p)^1$
3	0	$1p^3(1 - p)^0$

This distribution illustrates the origin of the binomial coefficients. The coefficient ${}_3C_1(=3)$ is the number of ways in which 2 S's and 1 F can be arranged; the coefficient ${}_3C_2(=3)$ is the number of ways in which 1 S and 2 F's can be arranged.

An argument similar to this shows that the general formula (for any n) is

$$\Pr\{j \text{ successes and } n - j \text{ failures}\} = {}_nC_j p^j (1 - p)^{n-j}$$

where

${}_nC_j$ = the number of ways in which j S's and $(n - j)$ F's can be arranged.

Combinations The binomial coefficient ${}_nC_j$ is also known as the number of combinations of n items taken j at a time; it is equal to the number of different subsets of size j that can be formed from a set of n items.

The Binomial Coefficients: A Formula

Binomial coefficients can be calculated from the formula

$${}_nC_j = \frac{n!}{j!(n - j)!}$$

where $x!$ ("x-factorial") is defined for any positive integer x by

$$x! = x(x - 1)(x - 2)\cdots(2)(1)$$

and $0! = 1$.

For example, for $n = 7$ and $j = 4$ the formula gives

$${}_7C_4 = \frac{7!}{4!3!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(4 \times 3 \times 2 \times 1)(3 \times 2 \times 1)}$$

$$= 35$$

To see why this is correct, let us consider in detail why the number of ways of rearranging 4 S's and 3 F's should be equal to

$$\frac{7!}{4!3!}$$

Suppose 4 S's and 3 F's were written on cards, like this:

$$\boxed{S_1}\ \boxed{S_2}\ \boxed{S_3}\ \boxed{S_4}\ \boxed{F_1}\ \boxed{F_2}\ \boxed{F_3}$$

Temporarily we put subscripts on the S's and F's to distinguish them. First let us see how many ways there are to arrange the 7 cards in a row:

There are 7 choices for which card goes first;
for each of these, there are 6 choices for which card goes second;
for each of these, there are 5 choices for which card goes third;
for each of these, there are 4 choices for which card goes fourth;
for each of these, there are 3 choices for which card goes fifth;
for each of these, there are 2 choices for which card goes sixth;
for each of these, there is 1 choice for which card goes last.

It follows that there are 7! ways of arranging the 7 cards. Now consider the locations of the 4 S's. There are 4! ways in which the S's can be rearranged among themselves. Likewise, there are 3! ways in which the F's can be rearranged among themselves. If we were to ignore the subscripts on the S's and F's, then some of the 7! ways of arranging the 7 cards would be indistinguishable. Indeed, any rearrangement of the S's *among themselves* leaves the 7 card arrangement looking the same. Similarly, any rearrangement of the F's *among themselves* leaves the 7 card arrangement looking the same. Thus, the number of *distinguishable* arrangements is

$$\frac{7!}{4!3!}$$

Appendix 3.2
Mean and Standard Deviation of the Binomial Distribution

Suppose that Y is a binomial random variable with n trials and p as the probability of success on each trial. Then we can think of Y as the sum of n variables X_1, $X_2, \ldots, X_n$, where each X_i is equal to either 0 or 1 (0 for a failure or 1 for a success). That is, $Y = \sum X_i$, with $\Pr\{X_i = 0\} = 1 - p$ and $\Pr\{X_i = 1\} = p$. The n X_i's are a random sample from a hypothetical population of X's that has average $\mu_X = p$ (since $0 \times (1 - p) + 1 \times p = p$).

Now consider the population standard deviation, σ_X, for the population of X's. Recall, from Section 2.8 that for a variable X the definition of σ is

$$\sigma = \sqrt{\text{Population average value of } (X - \mu)^2}$$

For the population of X's, the mean is $\mu_X = p$. Thus, for this population,

$$\sigma_X = \sqrt{\text{Population average value of } (X - p)^2}.$$

In the population of X's, the quantity $(X - p)^2$ takes on only two possible values:

$$(X - p)^2 = \begin{cases} (0 - p)^2 & \text{if } X = 0 \\ (1 - p)^2 & \text{if } X = 1 \end{cases}$$

Furthermore, these values occur in the proportions $(1 - p)$ and p, respectively, so that the population average value of $(X - p)^2$ is equal to

$$(0 - p)^2 \times (1 - p) + (1 - p)^2 \times p$$

This can be simplified to

$$\begin{aligned} p^2 \times (1 - p) + (1 - p)^2 \times p &= p\{p(1 - p) + (1 - p)^2\} \\ &= p\{p - p^2 + 1 - 2p + p^2\} \\ &= p(1 - p) \end{aligned}$$

Hence, the population average value of $(X - p)^2$ is $p(1 - p)$, so $\sigma_X = \sqrt{p(1 - p)}$.

The binomial random variable Y is $\sum X_i$. To find the mean and standard deviation of Y, we need two facts:

Fact 1: For any collection of random variables $X_1, X_2, \ldots, X_n$ the mean of $\sum X_i = \sum(\text{mean of } X_i)$.

Fact 2: For a collection of independent random variables $X_1, X_2, \ldots, X_n$ the variance of $\sum X_i = \sum(\text{variance of } X_i)$.

(Recall that the variance, σ^2, is the square of the standard deviation, σ.)

Using Fact 1, we see that the mean of Y is the mean of $\sum X_i$, which is $\sum p$. Thus, the mean of Y is $\mu_Y = np$.

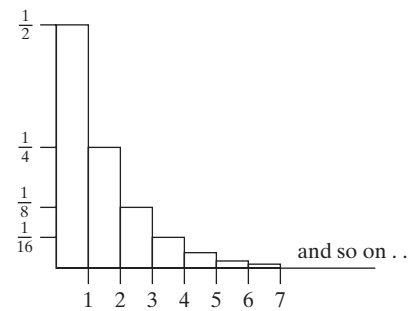
Using Fact 2, the variance of Y is the variance of $\sum X_i$, which equals $\sum(\text{Variance of } X_i)$ or $np(1 - p)$. Thus, the standard deviation of Y is $\sigma_Y = \sqrt{np(1 - p)}$.

Appendix 4.1
Areas of Indefinitely Extended Regions

Consider the region bounded between a normal curve and the horizontal axis. Because the curve never touches the axis, the region extends indefinitely far to the left and to the right. Yet the area of the region is exactly equal to 1.0. How is it possible for an indefinitely extended region to have a finite area?

To gain insight into this paradoxical situation, consider Figure A.4.1, which shows a region that is simpler than that bounded by a normal curve. In this region, the width of each bar is 1.0; the height of the first bar is $\frac{1}{2}$, the second bar is half as high as the first, the third is half as high as the second, and so on. The bars form a region that is indefinitely extended. Nevertheless, we shall see that it makes sense to say that the area of the region is equal to 1.0.

Figure A.4.1



Let us first consider the areas of the individual bars. The area of the first bar is $\frac{1}{2}$, the area of the second bar is $\frac{1}{4}$, the third $\frac{1}{8}$, and so on. Now suppose that we choose a number, say k, and add up the areas of the first k bars, as follows:

BAR	HEIGHT OF BAR	CUMULATIVE TOTAL AREA
1	$\frac{1}{2}$	$\frac{1}{2}$
2	$\frac{1}{4}$	$\frac{3}{4}$
3	$\frac{1}{8}$	$\frac{7}{8}$
4	$\frac{1}{16}$	$\frac{15}{16}$
⋮	⋮	⋮
k	$\frac{1}{2^k}$	$\frac{2^k - 1}{2^k}$

The total area of the first two bars is $\frac{3}{4}$, the total area of the first three bars is $\frac{7}{8}$, and so on. In fact, the total area of the first k bars is equal to

$$\frac{2^k - 1}{2^k} = 1 - \frac{1}{2^k}$$

If k is very large, this area is very close to 1.0. In fact, we can make the area as close to 1.0 as we wish, simply by choosing k large enough. In these circumstances it is reasonable to say that the total area of the entire, indefinitely extended region is equal to exactly 1.0.

The preceding example shows that an indefinitely extended region can have a finite area. Likewise, the total area under the normal curve is 1.0 (but the proof of this fact requires fairly advanced calculus).