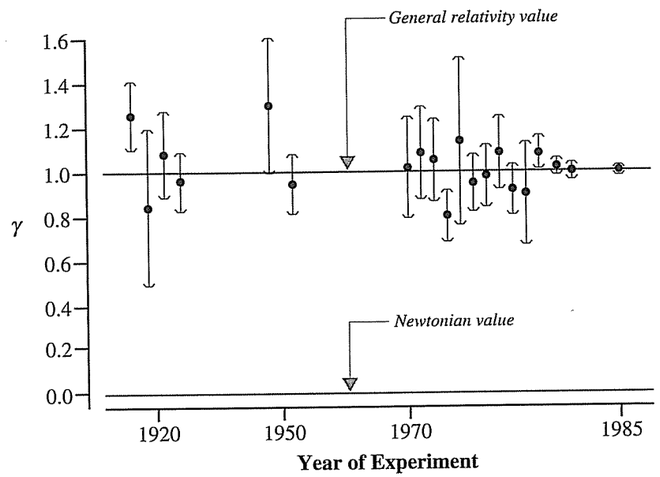
DISPLAY 2.13 Estimates and confidence intervals for γ, the deflection of light around the sun, from 20 experiments



outcomes. Experimental results are typically uncertain. So the fact that some intervals fail to include the value $\gamma = 1$ is not taken to disprove general relativity, but neither would it prove general relativity right if all the intervals did include $\gamma = 1$. When a theory's predictions are consistently denied by a series of experiments—such as the Newtonian prediction of $\gamma = 0$ in this example—scientists agree that the theory is not adequate.

2.5.3 The Rejection Region Approach to Hypothesis Testing

Not long ago, statisticians took a *rejection region* approach to testing hypotheses. A *significance level* of 0.05, say, was selected in advance, and a p-value less than 0.05 called for rejecting the hypothesis at the significance level 0.05; otherwise, the hypothesis was accepted, or more correctly, not rejected. Thus p-values of 0.048 and 0.0001 both lead to rejection at the 0.05 level, even though they supply vastly different degrees of evidence. On the other hand, p-values of 0.049 and 0.051 lead to different conclusions even though they provide virtually identical evidence against the hypothesis. Although important for leading to advances in the theory of statistics, the rejection region approach has largely been discarded for practical applications and p-values are reported instead. P-values give the reader more information for judging the significance of findings.

2.6 SUMMARY

Many research questions can be formulated as comparisons of two population distributions. Comparison of the distributions' centers effectively summarizes the difference between the parameters of interest when the populations have the same variation and general shape. This chapter concentrated on the difference in means, $\mu_2 - \mu_1$, which is estimated by the difference in sample averages.

The statistical problem is to assess the uncertainty associated with the difference between the estimate (the difference in sample averages) and the parameter (the difference in population means). The sampling distribution of an estimate is the key to understanding the uncertainty. It is represented as a histogram of values of the estimate for every possible sample that could have been selected.

Often with fairly large samples, a sampling distribution has a normal shape. A normal sampling distribution is specified by its mean and its standard deviation. When the populations have common standard deviation σ the difference in sample averages has a sampling distribution with mean $\mu_2 - \mu_1$ and standard deviation

$$\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

This could be used to describe uncertainty except that it involves the unknown σ—the common standard deviation in the two populations. In practice, σ is replaced by its best estimate from the data—the pooled standard deviation, s_p, having $n_1 + n_2 - 2$ degrees of freedom. The estimated standard deviation of the sampling distribution is called the standard error.

The standard error alone, however, does not entirely describe the uncertainty in an estimate. More precise statements can be made by using the t-ratio, which has a Student's t-distribution as its sampling distribution (if the ideal normal model applies). This leads directly to confidence intervals and p-values as statistical tools for answering the questions of interest. The confidence interval provides a range of likely values for the parameter, and the confidence level is interpreted as the frequency with which the interval construction procedure gives the right answer. For testing whether a particular hypothesized number could be the unknown parameter, the t-statistic is formed by substituting the hypothesized value for the true value in the t-ratio. The p-value is the chance of getting as extreme or more extreme t-ratios than the t-statistic, and it is interpreted as a measure of the credibility of the hypothesized value.

2.7 EXERCISES

Conceptual Exercises

1. **Finch Beak Data.** Explain why the finch beak study is considered an observational study.

2. For comparing two population means when the population distributions have the same standard deviation, the standard deviation is sometimes referred to as a nuisance parameter. Explain why it might be considered a nuisance.

3. True or false? If a sample size is large, then the shape of a histogram of the sample will be approximately normal, even if the population distribution is not normal.

4. True or false? If a sample size is large, then the shape of the sampling distribution of the average will be approximately normal, even if the population distribution is not normal.

5. Explain the relative merits of 90% and 99% levels of confidence.

6. What is wrong with the hypothesis that $\overline{Y}_2 - \overline{Y}_1$ is 0?

7. In a study of the effects of marijuana use during pregnancy, measurements on babies of mothers who used marijuana during pregnancy were compared to measurements on babies of mothers who did not. (Data from B. Zuckerman et al., "Effects of Maternal Marijuana and Cocaine Use on Fetal Growth," *New England Journal of Medicine* 320(12) (March 1989): 762–68.) A 95% confidence interval for the difference in mean head circumference (nonuse minus use) was 0.61 to 1.19 cm. What can be said from this statement about a p-value for the hypothesis that the mean difference is zero?

8. Suppose the following statement is made in a statistical summary: "A comparison of breathing capacities of individuals in households with low nitrogen dioxide levels and individuals in households with high nitrogen dioxide levels indicated that there is no difference in the means (two-sided p-value $= 0.24$)." What is wrong with this statement?

9. What is the difference between (a) the mean of Y and the mean of $\overline{Y}$? (b) the standard deviation of Y and the standard deviation of $\overline{Y}$? (c) the standard deviation of $\overline{Y}$ and the standard error of $\overline{Y}$? (d) a t-ratio and a t-statistic?

10. Consider blood pressure levels for populations of young women using birth control pills and young women not using birth control pills. A comparison of these two populations through an observational study might be consistent with the theory that the pill elevates blood pressure levels. What tool is appropriate for addressing whether there is a difference between these two populations? What tool is appropriate for addressing the likely size of the difference?

11. The data in Display 2.14 are survival times (in days) of guinea pigs that were randomly assigned either to a control group or to a treatment group that received a dose of tubercle bacilli. (Data from K. Doksum, "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *Annals of Statistics* 2(1974): 267–77.) (a) Why might the additive treatment effect model (introduced in Section 1.3.1) be inappropriate for these data? (b) Why might the ideal normal model with equal spread be an inadequate approximation?

Computational Exercises

12. Marijuana Use During Pregnancy. For the birth weights of babies in two groups, one born of mothers who used marijuana during pregnancy and the other born of mothers who did not (see Exercise 7), the difference in sample averages (nonuser mothers minus user mothers) was 280 grams, and the standard error of the difference was 46.66 grams with 1,095 degrees of freedom. From this information, provide the following: (a) a 95% confidence interval for $\mu_2 - \mu_1$, (b) a 90% confidence interval for $\mu_2 - \mu_1$, and (c) the two-sided p-value for a test of the hypothesis that $\mu_2 - \mu_1 = 0$.

13. Fish Oil and Blood Pressure. Reconsider the changes in blood pressures for men placed on a fish oil diet and for men placed on a regular oil diet, from Chapter 1, Exercise 12. Do the following steps to compare the treatments.

DISPLAY 2.14 Lifetimes of guinea pigs in two treatment groups

Control (n=64)			*Received bacilli (n=58)*
36,18		0	
91,89,87,86,52,50			76,93,97
49,20,19,18,15,14,14,08,02		1	07,08,13,14,19,36,38,39
89,78,73,67,67,66,65,60			52,54,54,60,64,64,66,68,78,79,81,81,83,85,94,98
16,12,09		2	12,13,16,20,25,25,44
92,79,78,73			53,56,59,65,68,70,83,89,91
41		3	11,15,26,26
82,80,67,55			61,73,73,76,97,98
46,32,21,21		4	06
74,63,55			59,66
46,45,05		5	
90,76,69			92,98
41,38,37,34,21,08,07,03		6	
88,85,63,50			
35,25		7	

Legend: | 5 | 98 represents 598 days

(a) Compute the averages and the sample standard deviations for each group separately.
(b) Compute the pooled estimate of standard deviation using the formula in Section 2.3.2.
(c) Compute $\text{SE}(\overline{Y}_2 - \overline{Y}_1)$ using the formula in Section 2.3.2.
(d) What are the degrees of freedom associated with the pooled estimate of standard deviation? What is the 97.5th percentile of the t-distribution with this many degrees of freedom?
(e) Construct a 95% confidence interval for $\mu_2 - \mu_1$ using the formula in Section 2.3.3.
(f) Compute the t-statistic for testing equality as shown in Section 2.3.5.
(g) Find the one-sided p-value (as evidence that the fish oil diet resulted in greater reduction of blood pressure) by comparing the t-statistic in (f) to the percentiles of the appropriate t-distribution (by reading the appropriate percentile from a computer program or calculator).

14. Fish Oil and Blood Pressure. Find the 95% confidence interval and one-sided p-value asked for in Exercise 13(e) and (g) but use a statistical computer package to do so.

15. Auto Exhaust and Lead Concentration in Blood. Researchers took independent random samples from two populations of police officers and measured the level of lead concentration in their blood. The sample of 126 police officers subjected to constant inhalation of automobile exhaust fumes in downtown Cairo had an average blood level concentration of 29.2 μg/dl and an SD of 7.5 μg/dl; a control sample of 50 police officers from the Cairo suburb of Abbasia, with no history of exposure, had an average blood level concentration of 18.2 μg/dl and an SD of 5.8 μg/dl. (Data from A.-A. M. Kamal, S. E. Eldamaty, and R. Faris, "Blood Lead Level of Cairo Traffic Policemen," *Science of the Total Environment* 105(1991): 165–70.) Is there convincing evidence of a difference in the population averages?

16. Motivation and Creativity. Verify the statements made in the summary of statistical findings for the Motivation and Creativity Data (Section 1.1.1) by analyzing the data on the computer.

17. Sex Discrimination. Verify the statements made in the summary of statistical findings for the Sex Discrimination Data (Section 1.1.2) by analyzing the data on the computer.

18. The Grants' Finch Complete Beak Data. The data file ex0218 contains the beak depths (in mm) of all 751 finches captured by Peter and Rosemary Grant in 1976 and all 89 finches captured in 1978 (as described in Section 2.1.1). Use a statistical computer program for parts a–d: (a) Draw side-by-side box plots of the two groups of beak depths. (b) Use the two-sample t-test on these data to find the one-sided p-value for a test of the hypothesis of no difference in means against the alternative that the mean in 1978 is larger. (c) What is the two-sided p-value from the t-test? (d) Provide an estimate and a 95% confidence interval for the amount by which the 1978 mean exceeds the 1976 mean. (e) What is it about the finches in the two populations that might make you question the validity of the independence assumption upon which the two-sample t-test is derived?

19. Fish Oil and Blood Pressure. Reconsider the fish oil and blood pressure data of Chapter 1, Exercise 12. Since the measurements are the reductions in blood pressure for each man, it is of interest to know whether the mean reduction is zero for each group. For the regular oil diet group do the following:

(a) Compute the average and the sample standard deviation. What are the degrees of freedom associated with the sample standard deviation, s_2?
(b) Compute the standard error for the average from this group: $\text{SE}(\overline{Y}_2) = s_2/\sqrt{n_2}$.
(c) Construct a 95% confidence interval for μ_2 as $\overline{Y}_2 + t_d(.975)\text{SE}(\overline{Y}_2)$, where d is the degrees of freedom associated with s_2.
(d) For the hypothesis that μ_2 is zero, construct the t-statistic $\overline{Y}_2/\text{SE}(\overline{Y}_2)$. Find the two-sided p-value as the proportion of values from a t_d-distribution farther from 0 than this value.

20. Fish Oil and Blood Pressure (One-Sample Analysis). Repeat Exercise 19 for the group of men who were given the fish oil diet and then answer these questions: Is there any evidence that the mean reduction for this group is different from zero? What is the typical reduction in blood pressure expected from this type of diet (for individuals like these men)? Provide a 95% confidence interval.

Data Problems

21. Bumpus Natural Selection Data. In 1899, biologist Hermon Bumpus presented as evidence of natural selection a comparison of numerical characteristics of moribund house sparrows that were collected after an uncommonly severe winter storm and which had either perished or survived as a result of their injuries. Display 2.15 shows the length of the humerus (arm bone) in inches for 59 of these sparrows, grouped according to whether they survived or perished. Analyze these data to summarize the evidence that the distribution of humerus lengths differs in the two populations. Write a brief paragraph of statistical conclusion, using the ones in Section 2.1 as a guide, including a graphical display, a conclusion about the degree of evidence of a difference, and a conclusion about the size of the difference in distributions.

DISPLAY 2.15 Humerus lengths of moribund male house sparrows measured by Hermon Bumpus, grouped according to survival status

Humerus Lengths (inches) of 35 Males That *Survived*

0.687, 0.703, 0.709, 0.715, 0.721, 0.723, 0.723, 0.726, 0.728, 0.728, 0.728, 0.729, 0.730, 0.730, 0.733, 0.733, 0.735, 0.736, 0.739, 0.741, 0.741, 0.741, 0.741, 0.743, 0.749, 0.751, 0.752, 0.752, 0.755, 0.756, 0.766, 0.767, 0.769, 0.770, 0.780

Humerus Lengths (inches) of 24 Males That *Perished*

0.659, 0.689, 0.702, 0.703, 0.709, 0.713, 0.720, 0.720, 0.726, 0.726, 0.729, 0.731, 0.736, 0.737, 0.738, 0.738, 0.739, 0.743, 0.744, 0.745, 0.752, 0.752, 0.754, 0.765

22. Male and Female Intelligence. Males and females tend to exhibit different types of intelligence. Although there is substantial variability between individuals of the same gender, males on average tend to perform better at navigational and spatial tasks, and females tend to perform better at verbal fluency and memory tasks. This is not a controversial conclusion. Some researchers, however, ask whether males and females differ, on average, in their overall intelligence, and that *is* controversial because any single intelligence measure must rely on premises about the types of intelligence that are important. Even if researchers don't make a subjective judgment about a type of intelligence being tested, they are constrained by the available tools for measuring intelligence. Mathematical knowledge is easy to test, for example, but wisdom, creativity, practical knowledge, and social skill are not.

Display 2.16 shows the first five rows of a data set with several intelligence test scores for random samples of 1,306 American men and 1,278 American women between the ages of 16 and 24 in 1981. The column labeled AFQT shows the percentile scores on the Armed Forces Qualifying Test, which is designed for evaluating the suitability of military recruits but which is also used by researchers as a general intelligence test. The AFQT score is a combination of scores from four component tests: word knowledge, paragraph comprehension, arithmetic reasoning, and mathematical knowledge. The data set represented in Display 2.16 includes each individual's score on these components. (The overall AFQT score reported here, officially called AFQT89, is based on a nontrivial combination of the component scores)

DISPLAY 2.16 Armed Forces Qualifying Test (AFQT) score percentile and component test scores in arithmetic reasoning, word knowledge, paragraph comprehension, and mathematical knowledge, for 1,278 women and 1,306 men in 1981; first 5 of 2,584 rows

Gender	Arith	Word	Parag	Math	AFQT
male	19	27	14	14	70.3
female	23	34	11	20	60.4
male	30	35	14	25	98.3
female	30	35	13	21	84.7
female	13	30	11	12	44.5

Analyze the data to summarize the evidence of differences in male and female distributions of AFQT scores. Do they differ? By how much do they differ? Also answer these two questions for each of the four component test scores. Write a statistical report that includes graphical displays and statistical conclusions (like those in the case studies of Section 2.1), and a section of details upon which the conclusions were based (such as a listing of the computer output showing the results of two-sample t-tests and confidence intervals).

Notes about the data: Although these are random samples of American men and women between the ages of 16 and 24 in 1981, they are not simple random samples. The data come from the National Longitudinal Study of Youth (NLSY), which used variable probability sampling (see Section 1.5.4). To estimate the means of the larger populations, more advanced techniques are appropriate. For comparing male and female distributions, the naive approach based on random sampling is not likely to be misleading. These data come from the National Longitudinal Survey of Youth, U.S. Bureau of Labor Statistics, http://www.bls.gov/nls/home.htm (May 8, 2008). Rows with missing values of variables, including variables used in related problems in other chapters, have been omitted.

23. Speed Limits and Traffic Fatalities. The National Highway System Designation Act was signed into law in the United States on November 28, 1995. Among other things, the act abolished the federal mandate of 55-mile-per-hour maximum speed limits on roads in the United States and permitted states to establish their own limits. Of the 50 states (plus the District of Columbia), 32 increased their speed limits either at the beginning of 1996 or sometime during 1996. Shown in Display 2.17 are the percentage changes in interstate highway traffic fatalities from 1995 to 1996. What evidence is there that the percentage change was greater in states that increased their speed limits? How much of a difference is there? Write a brief statistical report detailing the answers to these questions. (Data from "Report to Congress: The Effect of Increased Speed Limits in the Post-NMSL Era," National Highway Traffic Safety Administration, February, 1998; available in the reports library at http://www-fars.nhtsa.dot.gov/.)

DISPLAY 2.17 Number of traffic fatalities in 50 U.S. states and the District of Columbia, and status of speed limit change in the state (retained 55 mph limit or increased speed limit); first 5 of 51 rows

State	Fatalities1995	Fatalities1996	PctChange	SpeedLimit
Alabama	1114	1146	2.87	Inc
Alaska	87	81	−6.9	Ret
Arizona	1035	994	−3.96	Inc
Arkansas	631	615	−2.54	Inc
California	4192	3989	−4.84	Inc

Answers to Conceptual Exercises

1. The birds were *observed* to be in the 1976 or 1978 groups, not *assigned* by the researchers.

2. There is rarely any direct interest in the standard deviation, but it must be estimated in order to clear up the picture regarding means.

3. False.

4. True.

5. There is more confidence that a 99% interval contains the parameter of interest, but the extra confidence comes at the price of the interval being larger and therefore less informative about specific likely values.

6. The hypothesis must be about the population means. A hypothesis must be about the value of an *unknown* parameter. The value of a statistic will be known when the data are collected and analyzed.

7. It is less than 0.05.

8. The statement implies that the null hypothesis is accepted as true. It should be worded as, for example, the data are consistent with the hypothesis that there is no difference. (This issue is partly one of semantics, but it is still important to understand the distinction being made.)

9. (a) The mean of Y is the mean in the population of all individual measurements, and the mean of $\overline{Y}$ is the mean of the sampling distribution of the sample mean. With random sampling, the two have the same value μ.

(b) The standard deviation of Y is the standard deviation among all observations in the population, and the standard deviation of $\overline{Y}$ is the standard deviation in the sampling distribution of the sample average. The two are related, but not the same: if the standard deviation of Y is denoted by σ, then the standard deviation of $\overline{Y}$ is $\sigma/\sqrt{n}$.

(c) The standard error is an estimate of the standard deviation in the sampling distribution, obtained by replacing the unknown population standard deviation in the formula by the known sample standard deviation.

(d) The t-ratio is the ratio of the difference between the estimate and the parameter to the standard error of the estimate. It involves the parameter, so you do not generally know what it is. The t-statistic is a trial value of the t-ratio, obtained when a hypothesized value of the parameter is used in place of the actual value.

10. A p-value. A confidence interval.

11. (a) Because the spread of the stem-and-leaf plot is larger for the control group than for the treatment group, it does not appear that the effect of treatment was simply to add a certain number of days onto the lives of every guinea pig. It may have added days for those that would not have lived long anyway, and subtracted days from those that would have lived a long time. (b) The equal variation assumption does not appear to be appropriate.