others. Selecting an appropriate statistical procedure requires the researcher to evaluate honestly whether hypotheses were clearly stated prior to data collection or whether the data themselves guided the formulation of hypotheses. In the latter case, proper statistical evaluation must acknowledge the data-snooping process.

This chapter demonstrates some statistical tools for assessing uncertainty when a family of inferences is desired. Selecting an appropriate tool requires some introspection about the nature of the family examined. This chapter also introduced a powerful tool—computer simulation—that can help evaluate evidence about more complex hypotheses suggested by the data.

6.7 EXERCISES

Conceptual Exercises

1. Handicap Study. (a) Is it possible that the applicant's handicap in the videotape is confounded with the actor's performance? (b) Is there a way to design the study to avoid this confounding?

2. Mate Preference of Platyfish. If $\mu_1, \mu_2, \ldots, \mu_6$ represent the mean percentage of time spent by females with the yellow-sword male, for the six pairs of males, (a) state the null and alternative hypotheses that are tested by (i) the analysis of variance F-test and (ii) the t-test for the hypothesis that the linear contrast (for the linear effect of male body size) is zero. (b) Say why it is possible that the second test might find evidence that the means are different even if the first does not.

3. Mate Preference of Platyfish. For the test that the mean percent of time females spent with yellow-sword males is 50%, a one-tailed p-value was reported. Why?

4. An experimenter takes 20 samples of bark from each of 10 tree species in order to estimate the differences between fuel potentials. The data give 10 species averages, the lowest being 1.6 Btu/lb and the highest 3.8 Btu/lb for a range of 2.2 Btu/lb. A colleague suggests that another species be included, so the experimenter plans to gather 20 samples from that species and calculate its average potential. Which of the following is true about the range that the 11 species averages will have when the new species is included? (a) The range will equal the old range, 2.2 Btu/lb. (b) The range will be larger than 2.2 Btu/lb. (c) The range will be smaller than 2.2 Btu/lb. (d) The range cannot be smaller than 2.2 Btu/lb. (e) The range cannot be larger than 2.2 Btu/lb. (f) It is not possible to say that any of the above options is true until the average is known.

5. O-Ring Data. The case study in Section 4.1.1 involved the numbers of O-ring events on U.S. space shuttle flights launched at temperatures above and below 65°F. In the context of this chapter, is anything suspicious about that data? (*Hint*: Is there a possibility of data snooping?)

6. Does a confidence interval for the difference between two groups use information about variability from other groups? Why? or Why not?

7. What is the distinction between planned and unplanned comparisons?

8. Does a planned comparison always consist of estimating the difference between the means in two groups?

9. In comparing 10 groups a researcher notices that $\overline{Y}_7$ is the largest and $\overline{Y}_3$ is the smallest, and then tests the hypothesis that $\mu_7 - \mu_3 = 0$. Why should a multiple comparison procedure be used even though there is only one comparison being made?

10. If the analysis of variance screening test shows no significant evidence of any group differences, does that end the issue of there being any differences to report?

DISPLAY 6.10 Test scores for the experimental CAD instruction course

Group	Logo	Teaching method	n	Average	SD
1	L+D	Lecture and discussion	9	30.20	3.82
2	R	Programmed text	9	28.80	5.26
3	R+L	Programmed text with lectures	9	26.20	4.66
4	C	Computer instruction	9	31.10	4.91
5	C+L	Computer instruction with lectures	9	30.20	3.53

11. When choosing coefficients for a contrast, does the choice of $\{C_1, C_2, \ldots, C_I\}$ give a different t-ratio than the choice of $\{3C_1, 3C_2, \ldots, 3C_I\}$?

Computational Exercises

12. Handicap Study. Consider the groups *amputee*, *crutches*, and *wheelchair* to be handicaps of mobility and *hearing* to be a handicap affecting communication. Use the appropriate linear combination to test whether the average of the means for the mobility handicaps is equal to the mean of the communication handicap.

13. Handicap Study. Use the Bonferroni method to construct simultaneous confidence intervals for $\mu_2 - \mu_3, \mu_2 - \mu_5$, and $\mu_3 - \mu_5$ (to see whether there are differences in attitude toward the mobility type of handicaps).

14. Handicap Study. Examine these data with your available statistical computer package. See what multiple comparison procedures are available within the one-way analysis of variance procedure. Verify the 95% confidence interval half-widths in Display 6.6.

15. Comparison of Five Teaching Methods. An article reported the results of a planned experiment contrasting five different teaching methods. Forty-five students were randomly allocated, nine to each method. After completing the experimental course, a one-hour examination was administered. Display 6.10 summarizes the scores on a 10-minute retention test that was given 6 weeks later. (Data from S. W. Tsai and N. F. Pohl, "Computer-Assisted Instruction Augmented with Planned Teacher/Student Contacts," *Journal of Experimental Education*, 49(2) (Winter 1980–81): 120–26.)

(a) Compute the pooled estimate of the standard deviation from these summary statistics.

(b) Determine a set of coefficients that will contrast the methods using programmed text as part of the method (groups 2 and 3) with those that do not use programmed text (1, 4, and 5).

(c) Estimate the contrast in (b) and compute a 95% confidence interval.

16. A study involving 36 subjects randomly assigned six each to six treatment groups gives an ANOVA F-test with p-value $= 0.0850$. What multipliers are used to construct 95% confidence intervals for treatment differences with the following methods: (i) LSD, (ii) F-protected LSD, (iii) Tukey–Kramer, (iv) Bonferroni, and (v) Scheffé?

17. Adder Head Size. Red Riding Hood: "My, what big teeth you have!" Big Bad Wolf: "The better to eat you with, my dear." Are predators morphologically adapted to the size of their prey? A. Forsman studied adders on the Swedish mainland and on groups of islands in the Baltic Sea to determine if there was any relationship between their relative head lengths (RHL) and the body size of their main prey, field voles. (Data from A. Forsman, "Adaptive Variation in Head Size in *Vipera berus* L. Populations," *Biological Journal of the Linnean Society* 43 (1991): 281–96.) Relative head length is head length adjusted for overall body length, determined separately for males and females. Field vole body size is a combined measure of several features, expressed on a standardized scale.

DISPLAY 6.11 Average relative head lengths of adders from seven μ Swedish localities with their distances to the mainland and the body sizes of prey

Locality	Sample size	Average relative head length	Distance (km) to mainland	Field vole body size
Uppsala	21	−6.98	0	−1.75
In-Fredeln	34	−4.24	25.1	
Inre Hamnskär	20	−2.79	13.4	−0.16
Norrpada	25	2.22	14.7	1.31
Kärringboskär	7	1.27	10.0	
Ängskär	82	1.88	22.7	1.67
Svenska Hägarna	48	4.98	39.6	2.17

The data appear in Display 6.11. The pooled estimate of standard deviation of the RHL measurements was 11.72, based on 230 degrees of freedom.

(a) Determine the half-widths of 95% confidence intervals for all 21 pairwise differences among means for the seven localities, using (i) the LSD method and (ii) the Tukey–Kramer method.

(b) Using a linear contrast on the groups for which vole body size is available, test whether the locality means (of relative head length) are equal, against the alternative that they fall on a straight line function of vole body size, with nonzero slope.

(c) Repeat (b) for the pattern of distances to the mainland rather than vole body size.

18. Nest Cavities. Using the nest cavity data in Exercise 5.19, estimate the difference between the average of the mean entry areas for flickers, screech-owls, and kestrels and the average of the mean entry areas for the other six animals (on the transformed scale). Use a contrast of means.

19. Diet Restriction. For the data in Display 5.1 (and the summary statistics in Display 5.2), obtain a 95% confidence interval for the difference $\mu_3 - \mu_2$ using the Tukey–Kramer procedure. How does this interval differ from the LSD interval? Why is the Tukey–Kramer procedure the wrong thing to use for this problem?

20. Equity in Group Learning. [Continuation of Exercise 5.22.] (a) To see if the performance of low-ability students increases steadily with the ability of the best student in the group, form a linear contrast with increasing weights: -3 = Low, -1 = Low–Medium, $+1$ = Medium–High, and $+3$ = High. Estimate the contrast and construct a 95% confidence interval. (b) For the High-ability students, use multiple comparisons to determine which group composition differences are associated with different levels of test performance.

21. Education and Future Income. Reconsider the data problem of Exercise 5.25 concerning the distributions of annual incomes in 2005 for Americans in each of five education categories. (a) Use the Tukey–Kramer procedure to compare every group to every other group. Which pairs of means differ and by how many dollars (or by what percent)? (Use *p*-values and confidence intervals in your answer.) (b) Use the Dunnett procedure to compare every other group to the group with 12 years of education. Which group means apparently differ from the mean for those with 12 years of education and by how many dollars (or by what percent)? (Use *p*-values and confidence intervals in your answer.)

22. Reconsider the measurements of oxygen composition in 12 dinosaur bones from Exercise 5.23. Using a multiple comparisons procedure in a statistical computer package, find 95% confidence intervals for the difference in means for all pairs of bones (a) without adjusting for multiple comparisons and (b) using the Tukey–Kramer adjusted intervals. (c) How many of the unadjusted intervals exclude zero? (d) How many of the Tukey–Kramer adjusted intervals exclude zero? (e) By how much does the width of the adjusted interval exceed the width of the unadjusted interval, as a percentage, for comparing bone 1 to bone 2?

Data Problems

23. Diet Wars. To reduce weight, what diet should one combine with exercise? Most studies of human dieting practices have faced problems with high dropout rates and questionable adherence to diet protocol. Some studies comparing low-fat and low-carbohydrate diets have found that low-carb diets produced weight loss early, but that the loss faded after a short time. In an attempt to exert more control on subject adherence, a team of researchers at Ben-Gurion University in Negev, Israel, conducted a trial within a single workplace where lunch—the main meal of the day—was provided by the employer under the guidance of the research team. The team recruited 322 overweight employees and randomly assigned them to three treatment groups: a low-fat diet, a low-carb diet (similar to the Atkins diet), and a Mediterranean diet. Trans-fats were discouraged in all three diets. Otherwise, the restrictions and recommendations were as follows:

	Low-fat (n = 104)	Mediterranean (n = 109)	Low-carbohydrate (n = 109)
Calorie/day restriction	Women: 1,500 kcal Men: 1,800 kcal	Women: 1,500 kcal Men: 1,800 kcal	(Not Specified)
Percentage of calories from fat	30%	35%	(not specified)
Carbohydrates/day	(not specified)	(not specified)	20 g at start; increasing to 120 g
Percentage of calories from saturated fat	10%	(not specified)	(not specified)
Cholesterol/day	300 mg	(not specified)	(not specified)
Recommended:	low-fat grains vegetables fruits legumes	30–45 g olive oil 5–7 nuts < 20 g vegetables fish and poultry	get fat and protein from vegetables
Discouraged:	added fats	beef and lamb sweets high-fat snacks	

The study ran for two years, with 272 employees completing the entire protocol. Display 6.12 shows some of a data set simulated to match the weight losses (kg) of the participants at the study's conclusion. Is there evidence of differences in average weight loss after two years among these diets? If so, which diets appear to be better than which others? (Notice the consequences of controlling the family-wise confidence level on the widths of 95% confidence intervals.)

24. A Biological Basis for Homosexuality. Is there a physiological basis for sexual preference? Following up on research suggesting that certain cell clusters in the brain govern sexual behavior, Simon LeVay (data from S. LeVay, "A Difference in Hypothalamic Structure Between Heterosexual and Homosexual Men," *Science*, 253 (August 30, 1991): 1034–37) measured the volumes of four cell groups in the interstitial nuclei of the anterior hypothalamus in postmortem tissue from 41 subjects at autopsy from seven metropolitan hospitals in New York and California. The volumes of one cell

DISPLAY 6.12 Partial listing of data from a diet and weight loss experiment, showing subject number (there were 272 subjects), diet treatment group (there were 3), and weight loss (in kg) after 24 months

Subject	Group	WtLoss24
1	Low-Fat	2.2
2	Low-Fat	−4.8
3	Low-Fat	2.9
. . .		
105	Mediterranean	10.8
106	Mediterranean	6.4
107	Mediterranean	−0.3
. . .		
214	Low-Carbohydrate	3.4
215	Low-Carbohydrate	4.8
216	Low-Carbohydrate	10.9

DISPLAY 6.13 Volumes of INAH3 ($1{,}000 \times \text{mm}^3$) cell clusters from 41 human subjects at autopsy, by sex, sexual orientation, and cause of death

Males				Females	
Heterosexual		Homosexual		Heterosexual	
AIDS death	Non-AIDS death	AIDS death		AIDS death	Non-AIDS death
12	20	1	34	12	10
105	37	7	39		19
105	103	12	41		29
118	129	15	46		105
119	135	18	66		155
161	140	18	86		
	161	23	128		
	175	26	142		
	179	29	193		
	209	32			

cluster, INAH3, are re-created in Display 6.13. The numbers are 1,000 times volumes in mm^3. Subjects are classified into five groups according to three factors: gender, sexual orientation, and cause of death. One male classified as a homosexual who died of AIDS (volume 128) was actually bisexual. LeVay used the term *presumed* heterosexual to indicate the possibility of misclassifying some subjects. Do heterosexual males tend to differ from homosexual males in the volume of INAH3? Do heterosexual males tend to differ from heterosexual females? Do heterosexual females tend to differ from homosexual males? Analyze the data and write a brief statistical report including a summary of statistical findings, a graphical display, and a details section describing the details of the particular methods used. Also describe the limitations of inferences that can be made. (*Hint:* What linear combination of the five means can be used to test whether cause of death can be ignored? If cause of death can be ignored, what linear combinations of the resulting three means are appropriate for addressing the questions above?)

Answers to Conceptual Exercises

1. (a) Yes, it is possible that the acting performance of the actor portrayed a more competent worker in the *crutches* role, even though the script was held constant. (b) With two pairs of actors and twice as many groups, the handicap effect could be isolated from the actor effect.

2. (a) (i) Hypothesis: all means are equal; alternative: at least one is different from the others. (ii) Hypothesis: all means are equal: alternative: the means are not equal but fall on a straight line function of male body size (with nonzero slope). (b) The alternative in (ii) is more specific. For a given set of data there is more power in detecting differences if the (correct) specific alternative can be investigated. (This has previously been noticed by the fact that a one-tailed p-value is smaller than a two-tailed.)

3. The researcher had reason to believe, because of other species of fish in the same genus, that the colored tail would be more attractive to the females.

4. If the new species average is somewhere between 1.6 and 3.8 Btu/lb, the range of the set of means is unchanged. If the new species average is either less than 1.6 or greater than 3.8 Btu/lb, the range is larger. Those are the only possibilities. So the answer is (d): the range cannot be smaller than the old range (but it could be larger). The range increases as the number of groups increase.

5. Where did the 65°F cutoff come from? If the analyst chose that cutoff because it produced the most dramatic difference between the two groups, the search procedure should be included in the assessment of evidence.

6. Yes, it does. It is important to pool information about variability because the population SD is difficult to estimate from small samples.

7. A planned comparison is one of a few specific comparisons that is designed to answer a question of interest. An unplanned comparison is one (of a large number) of comparisons that is suggested by the data themselves.

8. No. More complex comparisons can be made by examining linear combinations of group means.

9. The more groups there are, the larger the difference one expects between the smallest and largest averages. To incorporate the selection of this hypothesis on the basis of how the data turned out, the appropriate statistical measure of uncertainty is the same as the one that is appropriate for comparing every mean to every other mean.

10. Not necessarily. If there are no planned comparisons, it may be best to report no evidence of differences (protected LSD procedure). But the evidence about planned comparisons to answer questions of interest should be assessed on its own. It is possible that a planned comparison shows something when the F-test does not.

11. No. The parameter changes from γ to 3γ, the estimate changes from g to $3g$, and the standard error also changes from $\text{SE}(g)$ to $\text{SE}(3g) = 3\text{SE}(g)$. So the t-ratio is not changed at all. This is why one can take the convenient step of multiplying a set of coefficients by a common factor to make all coefficients into integers, if desirable.