Case 3, where one estimate lies within the confidence interval for the other mean, but the second estimate lies outside the first interval, is difficult to judge. But in *Case 4*, where the best estimate of each mean lies inside the confidence interval for the other, there is no evidence of any difference.

Finally, it must be mentioned that the discussion of this section applies to the comparison of two confidence intervals only. If there are more than two confidence intervals, then it may be quite misleading to compare the two most disparate ones, unless some adjustment for multiple comparisons is made. This topic is discussed in the next chapter.

5.7 SUMMARY

The term *analysis of variance* is often initially confusing as it seems to imply a comparison of variances. It is most definitely a method for comparing means, however, and the name derives from the approach for doing so—assessing variability from several sources. The analysis of variance F-test is used for assessing equality of several means.

Another point of confusion arises from the mistaken belief—due to the prevalence of the F-test in textbooks and computer programs—that the F-test necessarily plays a central role in the analysis of several samples. Usually it does not. It offers a convenient approach for detection of group differences, but it does not ordinarily provide answers to particular questions of interest. Tests and confidence intervals for pairs of means or linear combinations of means (discussed in the next chapter) provide much more specific information.

Analysis of data in several samples begins with a graphical display, like side-by-side box plots. Transformations of the data should be considered. The need for transformation and the presence of outliers is often better indicated by a residual plot—a plot of residuals versus fitted values. A funnel shape indicates the need for a transformation like the log, for example, for positive data. The analysis of variance table provides the numerical components of the F-test for equality of means. It also contains the within-groups mean square, which exactly equals the pooled estimate of variance, the best estimate of σ^2. Confidence intervals and t-tests for pairs of means should use this pooled estimate of variance from all groups.

Diet Restriction and Longevity Study

In this study, the questions of interest called for five specific pairwise comparisons among the groups. It might be tempting to perform five two-sample t-tests, but it is a much more efficient use of the data to perform t-tests using a pooled estimate of variance from all the groups. The analysis begins with examination of side-by-side box plots. Although there is some skewness in the data, it is not enough to warrant concern—the tests are sufficiently robust against this type of departure from normality—and no transformation is suggested. A closer look at possible problems is available through a residual plot, but the spreads appear to be approximately equal and there are no serious outliers. The analysis proceeds, therefore, with t-tests

and confidence intervals in the usual way, but using the pooled estimate of variance from all groups.

Spock Trial Study

The stem-and-leaf plots in Display 5.4 and box plots in Display 5.5 are useful for suggesting some answers to the question of interest and for indicating the appropriateness of the tools based on the standard one-way classification model. An analysis of variance F-test confirms the strong evidence of some differences between means. An application of the extra-sums-of-squares F-test for comparing equality of the six other judges shows no evidence of a difference in mean percentages of women on their venires. Assuming that the six other means are equal, a further F-test shows overwhelming evidence that the Spock judge mean is different from the mean of the other six. Since this is a test for equality of two means, a t-test could be used. In fact, the F-test is equivalent to a two-sided t-test when $I = 2$. The actual p-value reported in the summary of statistical findings comes from a different test, not based on the assumption of equal means among the other six judges, and is discussed in the next chapter.

5.8 EXERCISES

Conceptual Exercises

1. Spock Trial. Why is it important to obtain a pooled estimate of variance in the Spock trial study? Is it ever a mistake to obtain a pooled estimate of variance in a comparison involving several groups?

2. Four methods of growing wheat are to be compared on five farms. Four plots are used on each farm and each method is applied to one of the plots. Five measurements are therefore obtained on yield per acre for each of the four growing methods. Are the methods of this chapter appropriate for analyzing the results?

3. Diet Restriction. Is there any explanation for why the distribution of lifetimes of mice in Display 5.1 are all negatively skewed?

4. Diet Restriction. For comparing group 3 to group 2, explain why it is better to use the t-tools presented in Section 5.2.3 (using s_p from all six groups) than to use the Chapter 2 t-tools (using s_p from only the two groups involved).

5. Spock Trial. Should Spock's accusers question the defense on how the venires were selected for their study?

6. Spock Trial. Why is it useful to test whether the six judges other than Spock's have equal mean percentages of women on their venires?

7. Why is s_p^2 not simply taken as the average of the I sample variances?

8. Diet Restriction. If the longevity study was a planned experiment, why are the sample sizes different?

9. If s_p is zero, what must be true about the residuals?

10. Explain the role of degrees of freedom of the F-distribution associated with the F-statistic. How are degrees of freedom related to how far the F-statistic is likely to be from 1?

11. What does it mean if the F-statistic is so *small* that the chance of getting an F-statistic that small or smaller is only, say, 0.0001?

12. Flycatcher Species Identification. One of the most challenging field identification problems for North American ornithologists is to distinguish the 10 species of flycatchers in the genus *Empidonax*. Many articles have appeared in popular and scientific journals suggesting different morphological clues to proper identification. F. Rowland ("Identifying *Empidonax* Flycatchers: The Ratio Approach," 2009, *Birding* 41 (2): 30–38) asserted that the relative size of wing length to tail length is the appropriate physical characteristic for distinguishing the species in the field. This conclusion was based on the average values of the wing length minus the tail length for 24 birds in each species, as shown in the following table.

Species:	Yellow-bellied	Acadian	Alder	Willow	Least	Hammond's	Gray	Dusky	Pacific-slope	Cordilleran
Average wing–tail (mm; n=24)	13.6	15.4	14.7	12.4	9.2	13.7	10.3	7.0	9.5	9.5

Explain why a conclusion that this measurement tends to differ in the 10 species cannot be made from the averages alone. What additional piece of information is needed to test for group differences and to evaluate the extent to which individuals from different species can be distinguished?

Computational Exercises

13. Spock Trial. By examining Display 5.8, answer the following:

(a) What is the average percentage of women from all 46 venires?

(b) For how many of the 9 Spock judge's venires is the percentage of women less than the grand average from all 46 venires?

(c) For how many of the 9 Spock judge venires is the percentage of women less than the Spock judge's average?

14. Spock Trial. Use the following summary statistics to (a) compute the pooled estimate of the standard deviation and (b) carry out a t-test for the hypothesis that the Spock judge's mean is equal to the mean for judge A.

Judge:	Spock	A	B	C	D	E	F
Average % women:	14.62	34.12	33.61	29.10	27.00	26.97	26.80
SD of % women:	5.039	11.942	6.582	4.593	3.818	9.010	5.969
Sample size:	9	5	6	9	2	6	9

15. Spock Trial. (a) Use a calculator or statistical package to get the sample variance for the percentage of women on all 46 venires treated as one sample. (b) Multiply this by 45 to get the residual sum of squares for the equal-means model. (c) Multiply s_p^2 found in Exercise 14(a) above by (46−7) to get the residual sum of squares for the separate-means model. (d) Use these to construct an analysis of variance table, including the F-statistic for the hypothesis of equal means. Compare the result with Display 5.10.

16. Spock Trial. Use a statistical computer package to obtain the analysis of variance table in Display 5.10.

17. Display 5.20 shows the start of an analysis of variance table. Fill in the whole table from what is given here. How many groups were there? Is there evidence that the group means are different?

DISPLAY 5.20 Incomplete ANOVA table for Exercise 17

Source	d.f.	Sum of squares	Mean square	F-statistic	p-value
Between groups	?	?	?	?	?
Within groups	24	35,088	?		
Total	31	70,907			

18. Fatty Acid. The data in Display 5.21 were obtained from a randomized experiment to estimate the effect of a certain fatty acid (CPFA) on the level of a certain protein in rat livers. Only one level of the CPFA could be investigated in a day's work, so a control group (no CPFA) was investigated each day as well. (Data from Donald A. Pierce.)

DISPLAY 5.21 Levels of protein (×10) found in rat livers

	Treatment					
Day	CPFA 50	CPFA 150	CPFA 300	CPFA 450	CPFA 600	Control
1	154, 177, 174					157, 165, 150
2		164, 192, 159				186, 206, 195
3			157, 159, 124			192, 202, 216
4				160, 152, 141		190, 187, 160
5					147, 152, 158	191, 188, 199

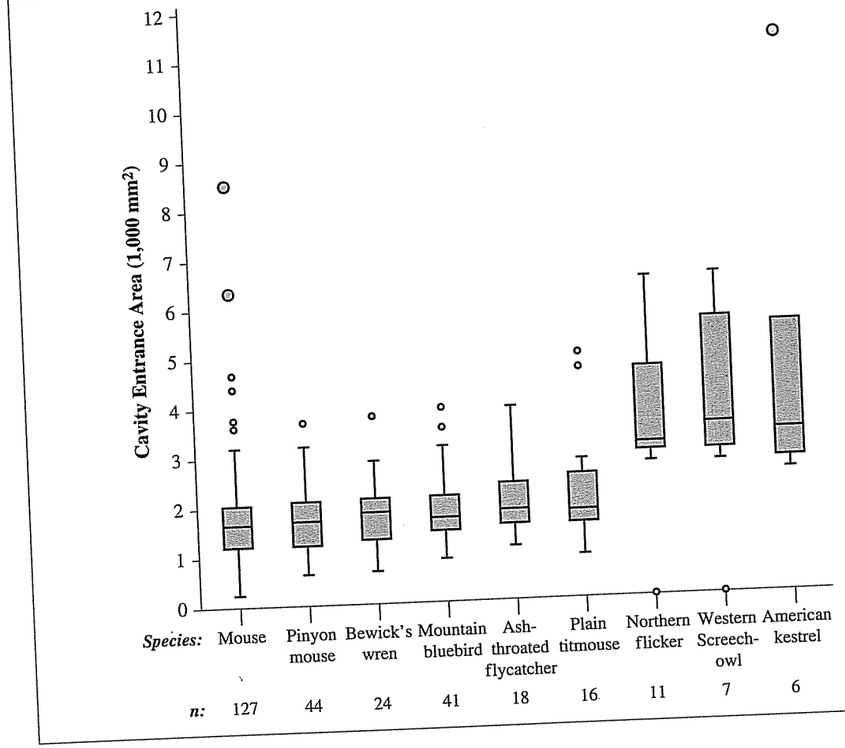
(a) Obtain estimated means for the model with six independent samples, one for each treatment. Determine the residuals and plot them versus the estimated means. Plot the residuals versus the day on which the investigation was conducted. Is there any indication that the methods of this chapter are not appropriate?

(b) Obtain estimated means for the model with 10 independent samples, one from each treatment-day combination. Calculate the ANOVA F-test to see whether these 10 groups have equal means.

(c) Use (a) and (b) and the methods of Section 5.4.1, to test whether the means for the control groups on different days are different. That is, compare the model with 10 different means to the model in which there are 6 different means.

19. Cavity Size and Use. Biologists freely discuss the concept of competition between species, but it is difficult to measure. In one study of competition for nesting cavities in Southeast Colorado, Donald Youkey (Oregon State University Dept. of Fisheries & Wildlife) located nearly 300 cavities occupied by a variety of bird and rodent species. Display 5.22 shows box plots of the entrance area measurements from cavities chosen by nine common nesting species. The general characteristics—positive skewness, larger spreads in the groups with larger means—suggest the need for a transformation. On the logarithmic scale, the spreads are relatively uniform, and the summary statistics appear in Display 5.23. Are the species competing for the same size cavities? Or, are there differences in the cavity sizes selected by animals of different species? It appears that there are two very different sets of species here. The first six selected relatively small cavities while the last three selected larger ones. Is that the only significant difference?

DISPLAY 5.22 Box plots for areas of entrances to cavities used by different species



(a) Compute the pooled estimate of variance.

(b) Construct an analysis of variance table to test for species differences. (The sample standard deviation of all 294 observations as one group is SD = 0.4962.) Perform the F-test.

(c) Verify that the analysis of variance method for calculating the between-group sum of squares yields the same answer as the formula

$$\text{Between-group SS} = \sum_{i=1}^{I} n_i \overline{Y}_i{}^2 - n\overline{Y}^2.$$

(d) Fit an intermediate model in which the first six species have one common mean and the last three species have another common mean. Construct an analysis of variance table with F-tests to compare this model with (i) the equal-means model and (ii) the separate-means model. Perform the F-test.

DISPLAY 5.23 Summary statistics for areas of cavity entrances (logarithmic scale)

Species	n	Mean	Sample SD
Mouse	127	7.347	0.4979
Pinyon mouse	44	7.368	0.4235
Bewick's wren	24	7.418	0.3955
Mountain bluebird	41	7.487	0.3183
Ash-throated flycatcher	18	7.563	0.3111
Plain titmouse	16	7.568	0.4649
Northern flicker	11	8.214	0.2963
Western Screech-owl	7	8.272	0.3242
American kestrel	6	8.297	0.5842

20. Flycatcher Species Identification. Consider the table of averages (of wing lengths minus tail lengths) from 24 birds in each of 10 species of flycatcher in Exercise 12. If it is assumed that the 11 populations all have the same mean and same population standard deviation, what is an estimate of the population standard deviation?

21. A robust test for equality of several population variances is *Levene's test*, which was previously discussed in Section 4.5.3 for the case of two variances. This procedure carries out the usual one-way analysis of variance F-test on the absolute values of the differenes of observations from their group medians. For practice, carry out Levene's test on the Spock data.

22. Equity in Group Learning. Several studies have demonstrated that engaging students in small learning groups increases student performances on subsequent tests. However, N. M. Webb and her colleagues argue that this raises a question of equity: Does the quality of the learning depend on the composition of the group? They chose students from five 7th and 8th grade classes in the Los Angeles school system. Based upon a science and language pretest, they classified each student's ability level as Low, Low-Medium, Medium-High, or High. They formed study groups consisting of three students each. The students were given a problem involving the setting up of two electrical circuits that would produce different brightness in a standard lightbulb. Each group was given the equipment to work with and time to discuss the problem and come to a solution. Afterward, each student was tested on the basics of the problem and its solution.

The table in Display 5.24 shows the results of the scores on this final test of the students whose ability level was Low in pretest. The students are grouped in the table according to the highest level of ability of a member in their study group. (Data from N. M. Webb, K. M. Nemer, A. W. Chizhik, and G. Sugrue "Equity Issues in Collaborative Group Assessment: Group Composition and Performance," *American Educational Research Journal* 35(4): (1998) 607–51.)

DISPLAY 5.24 Achievement test scores of Low ability students who worked in different study groups

	Highest ability level in the study group			
	Low	Low-medium	Medium-high	High
Average:	0.26	0.37	0.36	0.47
St. Dev.:	0.14	0.21	0.17	0.21
n:	17	24	25	14

DISPLAY 5.25 Achievement test scores of High ability students who worked in different study groups

	Lowest ability level in the study group			
	Low	Low-medium	Medium-high	High
Average:	0.75	0.77	0.72	0.85
St. Dev.:	0.16	0.11	0.12	0.10
n:	13	22	42	28

(a) How strong is the evidence that at least one group mean differs from the others?
(b) Display 5.25 shows a companion table. How strong is the evidence from this table that at least one mean differs from the others?
(c) The study groups apparently were not formed using random assignment. How might this affect any conclusions you might draw from the analysis?

Data Problems

23. Was Tyrannosaurus Rex Warm-Blooded? Display 5.26 shows several measurements of the oxygen isotopic composition of bone phosphate in each of 12 bone specimens from a single *Tyrannosaurus rex* skeleton. It is known that the oxygen isotopic composition of vertebrate bone phosphate is related to the body temperature at which the bone forms. Differences in means at different bone sites would indicate nonconstant temperatures throughout the body. Minor temperature differences would be expected in warm-blooded animals. Is there evidence that the means are different for the different bones? (Data from R. E. Barrick, and W. J. Showers, "Thermophysiology of *Tyrannosaurus rex*; Evidence from Oxygen Isotopes," *Science* 265 (1994): 222–24.)

24. IQ and Future Income. Display 5.27 shows the first five rows of a data set with annual incomes in 2005 for 2,584 Americans who were selected in the National Longitudinal Study of Youth 1979, who were available for re-interview in 2006, and who had paying jobs in 2005, along with the quartile of their AFQT (IQ) test score taken in 1981 (see Exercise 2.22). How strong is the evidence that the

DISPLAY 5.26 Measurements of oxygen isotopic composition of vertebrate bone phosphate (per mil deviations from SMOW) in 12 bones of a single *Tyrannosaurus rex* specimen

Bone	Oxygen isotopic composition					
Rib 16	11.10	11.22	11.29	11.49		
Gastralia	11.32	11.40	11.71			
Gastralia	11.60	11.78	12.05			
Dorsal vertebra	10.61	10.88	11.12	11.24	11.43	
Dorsal vertebra	10.92	11.20	11.30	11.62	11.70	
Femur	11.70	11.79	11.91	12.15		
Tibia	11.33	11.41	11.62	12.15	12.30	
Metatarsal	11.32	11.65	11.96	12.15		
Phalange	11.54	11.89	12.04			
Proximal caudal	10.93	11.01	11.08	11.12	11.28	11.37
Mid-caudal	11.35	11.43	11.50	11.57	11.92	
Distal caudal	11.95	12.01	12.25	12.30	12.39	

DISPLAY 5.27 Annual income in 2005 and test score quartile for an IQ test taken in 1981 for 2,584 Americans in the NLSY79 survey, first 5 of 2,584 rows

Subject	IQquartile	Income2005
2	1stQuartile	5,500
6	4thQuartile	65,000
7	2ndQuartile	19,000
8	2ndQuartile	36,000
9	3rdQuartile	65,000

distributions of 2005 annual incomes differ in the four populations? By how many dollars or by what percent does the distribution of 2005 incomes for those within the highest (fourth) quartile of IQ test scores exceed the distribution for the lowest (first) quartile?

25. Education and Future Income. The data file ex0525 contains annual incomes in 2005 of a random sample of 2,584 Americans who were selected for the National Longitudinal Survey of Youth in 1979 and who had paying jobs in 2005 (see Exercise 22 in Chapter 2). The data set also includes a code for the number of years of education that each individual had completed by 2006: <12, 12, 13–15, 16, and >16. How strong is the evidence that at least one of the five population distributions (corresponding to the different years of education) is different from the others? By how many dollars or by what percent does the mean or median for each of the last four categories exceed that of the next lowest category?

Answers to Conceptual Exercises

1. To make comparisons, one must estimate variation. There are not many venires for any particular judge, so pooling the information gives better precision to the variance estimate. But if the groups have very different spreads, pooling is a bad idea.

2. Not appropriate. You should not expect the measurements from plots on the same farm to be independent of each other.

3. Perhaps there is something like an upper bound, a maximum possible lifetime for each group, and healthy mice all tend to get close to it. Unhealthy mice, however, die off sooner and at very different ages.

4. If the variances in all populations are equal, s_p from all groups uses much more data to estimate σ, resulting in a more precise estimator.

5. Yes. Perhaps these are just as good as random samples of all venires for each judge. If there was any bias in the selection, however—for example, if the nine venires for Spock's judge were chosen because they did not have many women—the results would be misleading.

6. Spock's lawyers will have a stronger case if they can show that Spock's judge is particularly different from *all others* in having low representation of women.

7. It is, if the sample sizes are all equal. Otherwise, it gives more weight to estimates from larger samples.

8. It is unusual for experimenters to purposefully plan on unequal sample sizes. In this study it is likely that the larger number of mice in the N/R50 group was planned, because that was the major experimental group. Inequalities in the other group sample sizes are likely the result of losing mice to factors unrelated to the experiment.

9. All the residuals would have to be identically zero for this to happen.

10. The larger the degrees of freedom in either the numerator or denominator, the less variability there is in their sampling distributions. With smaller degrees of freedom in either, sampling variability can result in an F-ratio which is considerably different from 1, even when the null hypothesis is true.

11. That would suggest that the sample averages are closer to each other than one would expect in the course of natural sampling from identical populations. You may want to check out the independence assumption.

12. There are two important quantities: (1) the within-mean square is s_p^2, and (2) the p-value allows for judging group differences.

CHAPTER 6

Linear Combinations and Multiple Comparisons of Means

The F-test for equality of several means gives a reliable result with any number of groups. Its weakness is that it neither tells which means are different from which others nor accounts for any structure possessed by the groups. Consequently, its role is mainly to act as an initial screening device.

If the groups have a structure, or if the research requires a specific question of interest involving several groups, a particular *linear combination* of the means may address the question of interest. This chapter shows how to make inferences about linear combinations of means and how to choose linear combinations for some important kinds of problems.

When no planned comparison is called for by the questions of interest or the group structure, one may compare all means with each other. The large number of comparisons, however, compounds the statistical uncertainty in the statements of evidence. Some methods of adjustment to account for this *multiple comparisons* problem are provided and discussed here.