Chapter 12: Analysis of Variance

Section 12-1: One-Way ANOVA

1. a. The chest deceleration measurements are categorized according to the one characteristic of car size.

   b. The terminology of *analysis of variance* refers to the method used to test for equality of the three population means. That method is based on two different estimates of a common population variance.

2. As we increase the number of individual tests of significance, we increase the risk of finding a difference by chance alone (instead of a real difference in the means). The risk of a type I error, finding a difference in one of the pairs when no such difference actually exists, is too high. The method of analysis of variance helps us avoid that particular pitfall (rejecting a true null hypothesis) by using one test for equality of several means, instead of several tests that each compare two means at a time.

3. The test statistic is $F = 3.288$, and the F distribution applies.

4. The P-value is 0.061. Because the P-value is greater than the significance level of 0.05, we fail to reject the null hypothesis of equal means. There is not sufficient evidence to warrant rejection of the claim that the different size categories have the same mean chest deceleration. Based on the available data, the size of a car does not appear to affect chest deceleration.

5. Test statistic: $F = 0.39$; P-value: 0.677; Fail to reject H_0: $\mu_1 = \mu_2 = \mu_3$. There is not sufficient evidence to warrant rejection of the claim that the three categories of blood lead level have the same mean verbal IQ score. Exposure to lead does not appear to have an effect on verbal IQ scores.

6. Test statistic: $F = 2.3034$; P-value: 0.1044; Fail to reject H_0: $\mu_1 = \mu_2 = \mu_3$. There is not sufficient evidence to warrant rejection of the claim that the three categories of blood lead level have the same mean full IQ score. There is not sufficient evidence to conclude that exposure to lead has an effect on full IQ scores.

7. Test statistic: $F = 0.161$; P-value: 0.852; Reject H_0: $\mu_1 = \mu_2 = \mu_3$. There is not sufficient evidence to warrant rejection of the claim that the three size categories have the same mean head injury measurement. Based on the available data, the size of a car does not appear to affect head injuries.

8. Test statistic: $F = 1.1810$; P-value: 0.3167; Fail to reject H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. There is not sufficient evidence to support the claim that the different hospitals have different mean birth weights. It appears that birth weights are about the same in urban and rural areas.

9. Test statistic: $F = 1.304$; P-value: 0.275; Fail to reject H_0: $\mu_1 = \mu_2 = \mu_3$. There is not sufficient evidence to warrant rejection of the claim that males from the three age brackets have the same mean pulse rate. It appears that pulse rates of males are not affected by age bracket.

10. Test statistic: $F = 7.9338$; P-value: 0.0005; Reject H_0: $\mu_1 = \mu_2 = \mu_3$. There is sufficient evidence to warrant rejection of the claim that females from the three age brackets have the same mean pulse rate. It appears that pulse rates of females are affected by age bracket.

11. Test statistic: $F = 0.3476$; P-value: 0.7111; Fail to reject H_0: $\mu_1 = \mu_2 = \mu_3$. There is not sufficient evidence to warrant rejection of the claim that the three size categories have the same mean pelvis injury measurement. The size of a car does not appear to affect pelvis injuries.

12. Test statistic: $F = 22.9477$; P-value: 0.000; Reject H_0: $\mu_1 = \mu_2 = \mu_3$. There is sufficient evidence to warrant rejection of the claim that the three different states have the same mean arsenic content in brown rice. Even though Texas has the highest mean, the results from ANOVA do not allow us to conclude that any one specific population mean is different from the others, so we cannot conclude that brown rice from Texas poses the greatest health problem.

13. Test statistic: $F = 6.1413$; P-value: 0.0056; Reject H_0: $\mu_1 = \mu_2 = \mu_3$. There is sufficient evidence to warrant rejection of the claim that the four treatment categories yield poplar trees with the same mean weight. Although not justified by the results from analysis of variance, the treatment of fertilizer and irrigation appears to be most effective.

14. Test statistic: $F = 0.3801$; P-value: 0.7687; Fail to reject $H_0: \mu_1 = \mu_2 = \mu_3$. There is not sufficient evidence to warrant rejection of the claim that the four treatment categories yield poplar trees with the same mean weight. No treatment appears to be more effective than any other in the sandy and dry region.

15. Test statistic: $F = 18.9931$; P-value: 0.000; Reject $H_0: \mu_1 = \mu_2 = \mu_3$. There is sufficient evidence to warrant rejection of the claim that the three different types of cigarettes have the same mean amount of nicotine. Given that the king-size cigarettes have the largest mean of 1.26 mg per cigarette, compared to the other means of 0.87 mg per cigarette and 0.92 mg per cigarette, it appears that the filters do make a difference, although this conclusion is not justified by the results from analysis of variance.

16. Test statistic: $F = 20.8562$; P-value: 0.000; Reject $H_0: \mu_1 = \mu_2 = \mu_3$. There is sufficient evidence to warrant rejection of the claim that the three samples are from populations with the same mean. It appears that cotinine levels are greater with more exposure to tobacco smoke. (The samples do not appear to be from normally distributed populations, but ANOVA is robust against departures from normality. The samples appear to have very different variances, but the largest variance is roughly five times that of the smallest, and given that the sample sizes are all 40, the requirement of equal variances appears to be satisfied.)

17. The Tukey test results show different P-values, but they are not dramatically different. The Tukey results suggest the same conclusions as the Bonferroni test.

18. a. Test statistic: $F = 6.1413$; P-value: 0.0056; Reject $H_0: \mu_1 = \mu_2 = \mu_3$. There is sufficient evidence to warrant rejection of the claim that the four treatment categories yield poplar trees with the same mean weight.

    b. The displayed Bonferroni results show that with a P-value of 0.039, there is a significant difference between the mean of the no treatment group (group 1) and the mean of the group treated with both fertilizer and irrigation (group 4). (Also, when comparing group 1 and group 2, there is no significant difference between means, and when comparing group 1 and group 3, there is no significant difference between means.)

    c. Test statistic : $t = -4.007$; P-value: $6(0.001018) = 0.00611$; Reject the null hypothesis that the mean weight from the irrigation treatment group is equal to the mean from the group treated with both fertilizer and irrigation.

Section 12-2: Two-Way ANOVA

1. The pulse rates are categorized using two different factors of (1) age bracket and (2) gender.

2. No, to use two individual tests of one-way analysis of variance is to totally ignore the very important feature of the possible effect from an interaction between age bracket and gender. If there is an interaction, it doesn't make sense to consider the effects of one factor without the other.

3. a. An interaction between two factors or variables occurs if the effect of one of the factors changes for different categories of the other factor.

    b. If there is an interaction effect, we should not proceed with individual tests for effects from the row factor and column factor. If there is an interaction, we should not consider the effects of one factor without considering the effects of the other factor.

    c. Because the lines are far from parallel, the two genders have very different effects for the different age brackets, so there does appear to be an interaction between gender and age bracket.

4. Yes, the result is a balanced design because each cell has the same number (10) of values.

5. For interaction, the test statistic is $F = 9.58$ and the P-value is 0.0003, so there is sufficient evidence to warrant rejection of the null hypothesis of no interaction effect. Because there appears to be an interaction between age bracket and gender, we should not proceed with a test for an effect from age bracket and a test for an effect from gender. It appears an interaction between age bracket and gender has an effect on pulse rates. (Remember, these results are based on fabricated data used in one of the cells, so this conclusion does not necessarily correspond to real data.)

6. For interaction, the test statistic is $F = 3.6653$ and the P-value is 0.0322, so there is sufficient evidence to warrant rejection of no interaction effect. Because there appears to be an interaction effect, the tests for effects from age bracket and gender are not conducted. There appears to be sufficient evidence to support a claim that weight appears to be affected by an interaction between age bracket and gender.

7. For interaction, the test statistic is $F = 1.7970$ and the P-value is 0.1756, so there is not sufficient evidence to conclude that there is an interaction effect. For the row variable of age bracket, the test statistic is $F = 2.0403$ and the P-value is 0.1399, so there is not sufficient evidence to conclude that age bracket has an effect on height. For the column variable of gender, the test statistic is $F = 43.4607$ and the P-value is less than 0.0001, so there is sufficient evidence to support the claim that gender has an effect on height.

8. For interaction, the test statistic is $F = 2.88$ and the P-value is 0.076, so there is not sufficient evidence to warrant rejection of no interaction effect. There does not appear to be an interaction between sex and age. For the row variable of sex, the test statistic is $F = 2.01$ and the P-value is 0.169, so there is not sufficient evidence to warrant rejection of the claim of no effect from gender. For the column variable of age, the test statistic is $F = 0.75$ and the P-value is 0.482, so there is not sufficient evidence to conclude that age has an effect on cholesterol levels. It appears that cholesterol levels are not affected by an interaction between sex and age, are not affected by sex, and are not affected by age.

9. For interaction, the test statistic is $F = 3.7332$ and the P-value is 0.0291, so there is sufficient evidence to conclude that there is an interaction effect. The measures of self-esteem appear to be affected by an interaction between the self-esteem of the subject and the self-esteem of the target. Because there appears to be an interaction effect, we should not proceed with individual tests of the row factor (target's self-esteem) and the column factor (subject's self-esteem).

10. For interaction, the test statistic is $F = 0.3328$ and the P-value is 0.5747, so there is not sufficient evidence to warrant rejection of no interaction effect. There does not appear to be an interaction between gender and smoking. For the row variable of gender, the test statistic is $F = 1.3313$ and the P-value is 0.2710, so there is not sufficient evidence to warrant rejection of the claim of no effect from gender. For the column variable of smoking, the test statistic is $F = 1.1186$ and the P-value is 0.3110, so there is not sufficient evidence to conclude that smoking has an effect on body temperatures. It appears that body temperatures are not affected by an interaction between gender and smoking, they are not affected by gender, and they are not affected by smoking.

11. a. Test statistics and P-values do not change.
    b. Test statistics and P-values do not change.
    c. Test statistics and P-values do not change.
    d. An outlier can dramatically affect and change test statistics and P-values.

Chapter Quick Quiz

1. The sample data are partitioned into the three different categories according to the one factor of epoch.

2. H_0: $\mu_1 = \mu_2 = \mu_3$; H_1: At least one of the three population means is different from the others.

3. Test statistic: $F = 4.0497$; Larger test statistics result in smaller P-values.

4. Reject H_0. There is sufficient evidence to support the claim that the different epochs have mean skull breadths that are not all the same.

5. Right-tailed. Yes, all one-way analysis of variance tests are right-tailed.

6. No. The method of analysis of variance does not justify a conclusion that any particular mean is different from the others.

7. With one-way analysis of variance, data from the different samples are categorized using only one factor, but with two-way analysis of variance, the sample data are categorized into different cells determined by two different factors.

8. Test statistic: $F = 1.41$; P-value: 0.281; Fail to reject the null hypothesis of no interaction effect. There is not sufficient evidence to warrant rejection of the claim that head injury measurements are not affected by an interaction between the type of car (foreign, domestic) and size of the car (small, medium, large). There does not appear to be an effect from an interaction between the type of car (foreign or domestic) and whether the car is small, medium, or large.

9. Test statistic: $F = 2.25$ P-value: 0.159; Fail to reject the null hypothesis of no effect from the type of car. There is not sufficient evidence to support the claim that whether the car is foreign or domestic has an effect on head injury measurements.

10. Test statistic: $F = 0.44$; P-value: 0.655; Fail to reject the null hypothesis of no effect from the size of the car. There is not sufficient evidence to support the claim that whether the car is small, medium, or large has an effect on head injury measurements.

Review Exercises

1. a. one (type of diet)

   b. one-way analysis of variance

   c. Because the P-value is high, it appears that the four samples have means that do not differ by significant amounts. It appears that the mean ages of the four treatment groups are about the same.

   d. A small P-value would indicate that at least one of the treatment groups has a mean age that is significantly different from the others, so we would not know if differences from the diet treatments are due to the diets or to differences in age. A small P-value would undermine the effectiveness of the experiment.

2. Test statistic: $F = 42.9436$; P-value: 0.000; Reject H_0: $\mu_1 = \mu_2 = \mu_3$. There is sufficient evidence to warrant rejection of the claim that the three different types of cigarettes have the same mean amount of tar. Given that the king-size cigarettes have the largest mean of 21.1 mg per cigarette, compared to the other means of 12.9 mg per cigarette and 13.2 mg per cigarette, it appears that the filters do make a difference, although this conclusion is not justified by the results from analysis of variance.

3. For interaction, the test statistic is $F = 1.7171$ and the P-value is 0.1940, so there is not sufficient evidence to warrant rejection of no interaction effect. There does not appear to be an interaction between femur and car size. For the row variable of femur, the test statistic is $F = 1.3896$ and the P-value is 0.2462, so there is not sufficient evidence to conclude that whether the femur is right or left has an effect on load. For the column variable of car size, the test statistic is $F = 2.2296$ and the P-value is 0.1222, so there is not sufficient evidence to warrant rejection of the claim of no effect from car size. It appears that the crash test loads are not affected by an interaction between femur and car size, they are not affected by femur, and they are not affected by car size.

4. For interaction, the test statistic is $F = 0.8733$ and the P-value is 0.3685, so there does not appear to be an effect from an interaction between gender and whether the subject smokes. For gender, the test statistic is $F = 0.0178$ and the P-value is 0.8960, so gender does not appear to have an effect on body temperature. For smoking, the test statistic is $F = 3.0119$ and the P-value is 0.1082, so there does not appear to be an effect from smoking on body temperature.

Cumulative Review Exercises

1. a. Presidents: $\bar{x} = 15.5$ years; Popes: $\bar{x} = 13.1$ years; Monarchs: $\bar{x} = 22.7$ years

   b. Presidents: $s = 9.7$ years; Popes: $s = 9.0$ years; Monarchs: $s = 18.6$ years

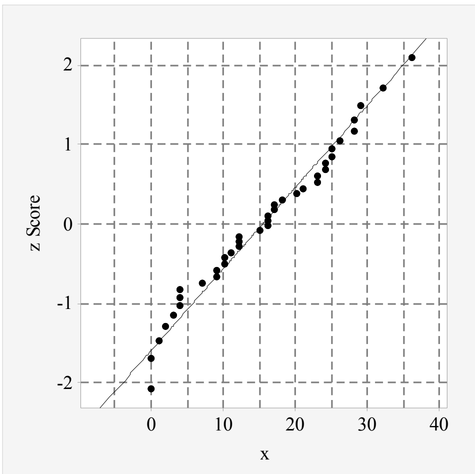
   c. Presidents: $s^2 = 94.5$ years^2; Popes: $s^2 = 80.3$ years^2; Monarchs: $s^2 = 346.1$ years^2

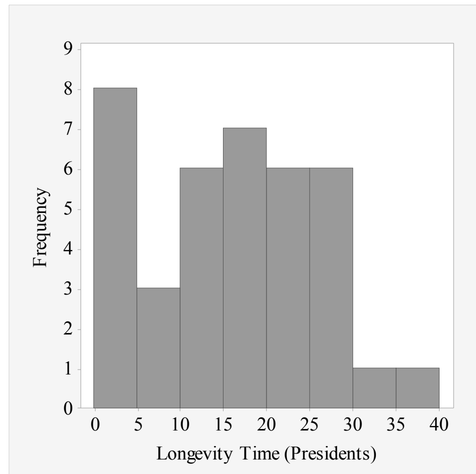
   d. ratio

2. H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$; population$_1$ = presidents, population$_2$ = monarchs;
Test statistic: $t = -1.383$; P-value = 0.1860 (Table: P-value > 0.10); Critical values ($\alpha = 0.05$): $t = \pm 2.123$ Fail to reject H_0. There is not sufficient evidence to support the claim that there is a difference between the means of the two groups.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(15.5 - 22.7) - 0}{\sqrt{\frac{9.7^2}{34} + \frac{18.6^2}{14}}} = -1.383 \text{ (df = 13)}$$

3. Normal, because the histogram is approximately bell-shaped or the points in a normal quantile plot are reasonably close to a straight-line pattern with no other pattern that is not a straight line pattern.





4. The data appear to fit the loose definition of a normally distribution.
95% CI: $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 15.5 \pm 2.035 \cdot \frac{9.7}{\sqrt{34}} \Rightarrow 12.3 \text{ years} < \mu < 18.7 \text{ years}$; We have 95% confidence that the limits of 12.3 years and 18.7 years contain the true value of the population mean.

5. a. H_0: $\mu_1 = \mu_2 = \mu_3$

   b. Because the P-value of 0.051 is greater than the significance level of 0.05, fail to reject the null hypothesis of equal means. There is not sufficient evidence to warrant rejection of the claim that the three means are equal. The three populations do not appear to have means that are significantly different.

6. a. $r = 0.918$; Critical values: $r = \pm 0.707$; P-value = 0.001; There is sufficient evidence to support the claim that there is a linear correlation between September weights and the subsequent April weights.

   b. $\hat{y} = 9.28 + 0.823x$

   c. $9.28 + 0.823(94) = 86.6$ kg, which is not very close to the actual April weight of 105 kg.

7. a. $z_{x=345} = \frac{345 - 280}{65} = 1.00$; which has a probability of $1 - 0.8413 = 0.1587$ to the right.

   b. $z_{x=215} = \frac{215 - 280}{65} = -1.00$ and $z_{x=345} = \frac{345 - 280}{65} = 1.00$; which have a probability of $0.8413 - 0.1587 = 0.6826$ (Tech: 0.6827) between them.

   c. $z_{x=319} = \frac{319 - 280}{65/\sqrt{25}} = 3.00$; which has a probability of 0.9987 to the left.

   d. The z score for the bottom 80% is 0.84, which correspond to a blood platelet count of $0.84 \cdot 65 + 280 = 334.6$ $(1000 \text{ cells}/\mu\text{L})$. (Tech: 334.7)

8. a. $0.20(1000) = 200$

   b. 95% CI: $\hat{p} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}\hat{q}}{n}} = 0.20 \pm 1.96\sqrt{\dfrac{(0.20)(0.80)}{1000}} \Rightarrow 0.175 < p < 0.225$

   c. Yes, the confidence interval shows us that we have 95% confidence that the true population proportion is contained within the limits of 0.175 and 0.225, and $1/4$ is not included within that range.

9. Using normal as approximation to binomial: 0.1020. (Exact result using binomial distribution: 0.0995.) Assuming that one-quarter of all offspring have blue eyes, the probability of getting 19 or fewer offspring with blue eyes is high, so there is not sufficient evidence to conclude that the one-quarter rate is wrong.

   Normal approximation: $np = 100 \cdot 0.25 = 25 \geq 5$ and $nq = 100 \cdot 0.75 = 75 \geq 5$; $z_{x=19.5} = \dfrac{19.5 - 100 \cdot 0.25}{\sqrt{100 \cdot 0.25 \cdot 0.75}} = -1.27$

   Binomial distribution: ${}_{100}C_0 \cdot 0.25^0 \cdot 0.75^{100} + \cdots + {}_{100}C_{19} \cdot 0.25^{19} \cdot 0.75^{81}$

10. H_0: Injury category is independent of type of weapon.

    H_1: Injury category depends on type of weapon.

    Test statistic: $\chi^2 = 2.909$; P-value $= 0.2335$ (Table: P-value > 0.10); Critical value: $\chi^2 = 5.991$; df $= (2-1)(3-1) = 2$; Fail to reject H_0. There is not sufficient evidence to warrant rejection of the claim of independence between injury category and whether the firearm was a handgun or a rifle or shotgun. The type of injury doesn't appear to be affected by whether the firearm is a handgun or a rifle or shotgun.

    $$\chi^2 = \frac{(31-31.8)^2}{31.8} + \frac{(35-30.4)^2}{30.4} + \frac{(162-165.8)^2}{165.8} + \frac{(13-12.2)^2}{12.2} + \frac{(7-11.6)^2}{11.6} + \frac{(67-63.2)^2}{63.2} = 2.909$$