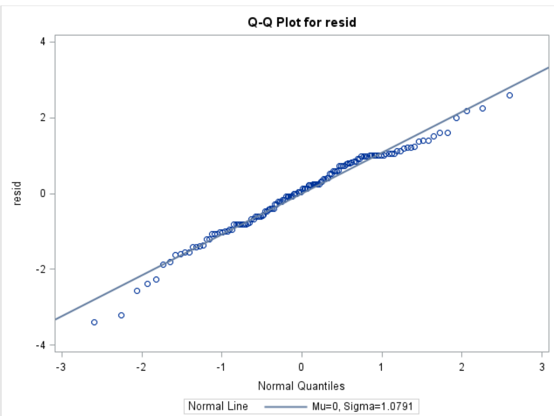
Chapter 12 One-Way Analysis of Variance

12.1. (a) ANOVA tests the null hypothesis that the *population* means are all equal. **(b)** *Experiments* are best for establishing causation. **(c)** ANOVA is used to compare *means* (and assumes that the variances are equal). **(d)** Multiple comparison procedures are used when we wish to determine which means are significantly different, but we do not need specific relations in mind prior to looking at the data.

12.3. $x_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, 2, 3, j = 1, 2, \ldots, 20$. $\varepsilon_{ij} \sim N(0, \sigma)$. There are $I = 3$ groups and $n_i = 20$ subjects per group. The parameters of the model are μ_1, μ_2, μ_3, and σ.

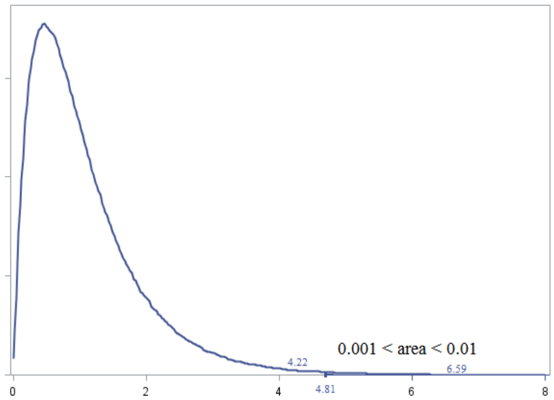
12.5. (a) Yes, the largest s is less than twice the smallest s: $80 < 2(68) = 136$. **(b)** The estimates for μ_1, μ_2, and μ_3 are 279, 245, and 258. $s_p^2 = \frac{(20-1)(78)^2 + (20-1)(68)^2 + (20-1)(80)^2}{(20-1)+(20-1)+(20-1)} = 5702.67.$ The estimate for σ is 75.516.

12.7. The Normal quantile plot shows the residuals are Normally distributed.



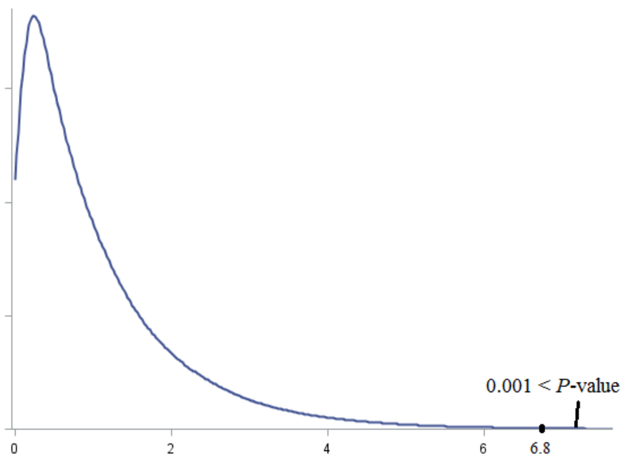
12.9. (a) This sentence describes *between*-group variation. Within-group variation is the variation that occurs by chance among members of the same group. **(b)** The *sums of* squares (not the mean squares) in an ANOVA table will add; that is, SST = SSG + SSE. **(c)** The common population standard deviation σ (not its estimate s_p) is a parameter. **(d)** A small P means the means are not all the same, but the distributions may still overlap quite a bit. (See the "Caution" immediately preceding this exercise in the text.)

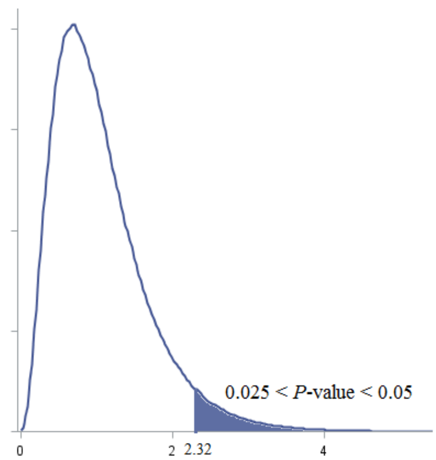
12.11. (a) With $I = 5$ groups and $N = 30$, we have df $I - 1 = 4$ and $N - I = 25$. In Table E, we see that $4.18 < F < 6.49$. **(b)** The sketch on the right shows the observed F value and the critical values from Table E. **(c)** $0.001 < P\text{-value} < 0.01$. **(d)** Because the P-value is small, we reject H_0; however, this does not say that all pairs of group means are different, only that at least one mean is different.



12.13. Generally the one with the largest difference between means and the smallest standard deviation will be the most significant, if this is not clear a ratio between the two can be used. So part **(b)** can be ruled out because it has a larger sigma than part **(a)** with the same means. Of the remaining two, we can see that part **(c)** has a bigger max difference relative to its sigma (10 to 3) than part **(a)** does (5 to 2), so part **(c)** will be the most significant.

12.15. F = MSG/MSE. DFG = $I - 1$, DFE = $N - I$. **(a)** F = 340/50 = 6.8, DFG = 3 − 1 = 2, DFE = 63 − 3 = 60. 0.001 < P-value < 0.01. **(b)** DFG = 8 − 1 = 7, DFE = 48 − 8 = 40. MSG = 77/7 = 11, MSE = 190/40 = 4.75. F = 11/4.75 = 2.32, 0.025 < P-value < 0.05.





12.17. **(a)** Response: egg cholesterol level. Populations: chickens with different diets or drugs. $I = 3$, $n_1 = n_2 = n_3 = 25$, $N = 75$. **(b)** Response: rating on seven-point scale. Populations: the three groups of students. $I = 3$, $n_1 = 31$, $n_2 = 18$, $n_3 = 45$, $N = 94$. **(c)** Response: quiz score. Populations: students in each TA group. $I = 3$, $n_1 = n_2 = n_3 = 14$, $N = 42$.

12.19. For all three situations we have H_0: $\mu_1 = \mu_2 = \mu_3$. H_α: not all of the μ_i are equal. DFG = $I - 1 = 2$, DFE = $N - I$, and DFT = $N - 1$. The degrees of freedom for the F test are DFG and DFE.

Situation	I	N	DFG	DFE	DFT	df for F statistic
(a) Egg cholesterol level	3	75	2	72	74	$F(2, 72)$
(b) Student opinions	3	94	2	91	93	$F(2, 91)$
(c) Teaching assistants	3	42	2	39	41	$F(2, 39)$

12.21. **(a)** This sounds like a fairly well-designed experiment, so the results should at least apply to this farmer's breed of chicken. **(b)** It would be good to know what proportion of the total student body falls in each of these groups—that is, is anyone overrepresented in this sample? **(c)** How well a TA teaches one topic (power calculations) might not reflect that TA's overall effectiveness.

12.23. **(a)** The plot suggests that both drugs cause an increase in activity level, and Drug B appears to have a greater average effect. **(b)** Yes, the largest s is less than twice the smallest s: $\sqrt{17.2} = 4.15 < 2\sqrt{7.75} = 2(2.784) = 5.57$.`12.16 and $s_p = 3.487$. **(c)** DFG = $I - 1 = 4$ and DFE = $N - I = 20$. **(d)** Compared with an $F(4, 20)$ distribution in Table E, we see that $2.25 < F < 2.87$, so $0.05 < P\text{-value} < 0.10$ (software gives 0.0642). We do not have significant evidence of a difference in mean effect.

12.25. **(a)** With $I = 5$ and $N = 183$, DFG = 4 and DFE = 178. **(b)** If the reported df were correct (meaning 32 athletes dropped out of the study), 151 athletes actually participated. **(c)** Answers will vary. For example, the individuals could have been outliers in terms of their ability to withstand the water-bath pain. In either case of low or high outliers, their removal would lessen the standard deviation for their sport and move that sports mean (removing a high outlier would lower the mean, and removing a low outlier would raise the mean).

12.27. To compare males and females, the coefficients are $a_1 = 0.5$, $a_2 = 0.5$, $a_3 = -0.5$, $a_4 = -0.5$.

> Note: *Technically any multiple of the coefficients would also be correct, such as −0.5, −0.5, 0.5, 0.5 or 1, 1, −1, −1.*

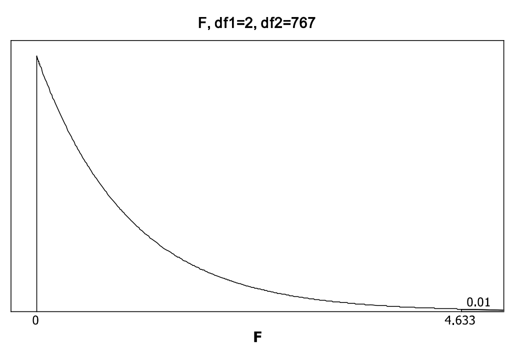
12.29. Because there are only two groups, if the ANOVA test establishes that there are differences in means, then we already know that the two means we have must be different. In this case contrasts and multiple comparison provide no further useful information.

12.31. The power would be larger. For larger differences between alternative means, λ gets bigger, increasing our power to see these differences.

12.33. **(a)** Yes, the largest s is less than twice the smallest s: $0.824 < 2(0.657) = 1.314$.

$$s_p^2 = \frac{(489-1)(0.804)^2 + (69-1)(0.824)^2 + (212-1)(0.657)^2}{489+69+212-3} =$$

0.5902. So $s_p = 0.7683$. **(b)** Comparing $F = 17.66$ with an $F(2, 767)$ distribution, we find $P\text{-value} < 0.001$. In the sketch shown, we see the 1% critical value is 4.633, so the observed value lies well above the bulk of this distribution. **(c)** For the contrast $\psi = \mu_2 - 0.5(\mu_1 + \mu_3)$, we test H_0: $\psi = 0$ versus H_a: $\psi > 0$. We find $c = 0.585$ with $\text{SE}_c = 0.0977$, so $t = 5.99$ with df = 767, and P-value < 0.0001.



12.35. **(a)** $1\mu_2 - 0.5\mu_1 - 0.5\mu_4$. **(b)** $1/3\mu_1 + 1/3\mu_2 + 1/3\mu_4 - 1\mu_3$.

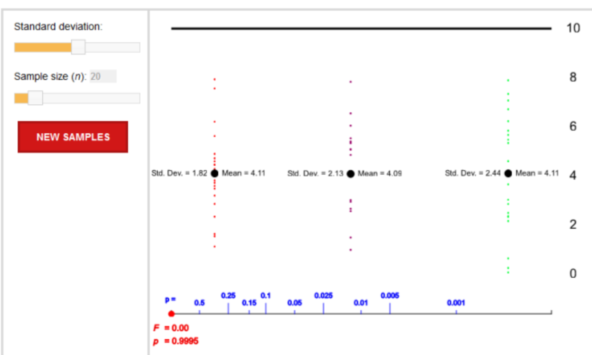
12.37. Let μ_1 be the placebo mean, μ_2 and μ_3 be the means for low and high doses of Drug A, and μ_4 and μ_5 be the means for low and high doses of Drug B. Recall from **Exercise 12.39** that $s_p = 3.487$. **(a)** The first contrast is $\psi_1 = \mu_1 - 1/2(\mu_2 + \mu_4)$; the second is $\psi_2 = \mu_3 - \mu_2 - (\mu_5 - \mu_4)$. **(b)** The estimated contrasts are $c_1 = 11.80 - 0.5(15.25 + 16.15) = -3.9$ and $c_2 = (18.55 - 15.25) - (17.10 - 16.15) = 2.35$. The respective standard errors are $\text{SE}_{c_1} = s_p\sqrt{1/4 + 0.25/4 + 0/4 + 0.25/4 + 0/4} = 2.1353$ and

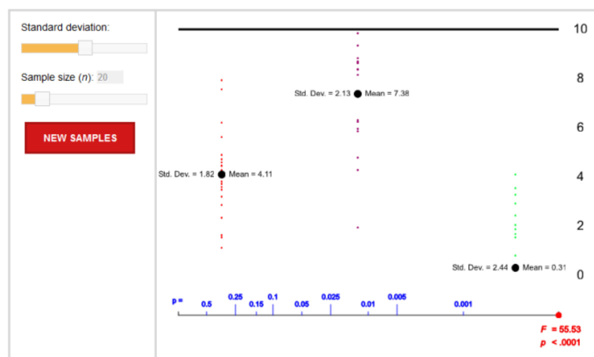
$\text{SE}_{c_2} = s_p\sqrt{0/4+1/4+1/4+1/4+1/4} = s_p = 3.487$. **(c)** The first contrast is significant ($t_1 = -1.826$, one-sided P-value = 0.0414), but the second is not ($t_2 = -0.674$, one-sided P-value 0.2540). We have enough evidence to conclude that low doses increase activity level over a placebo, but we cannot conclude that activity level changes due to increased dosage are different between the two drugs.

12.39. (a) If we believe the average score for the four off-task behaviors is less than the paper-and-pencil control, a contrast would be $\psi = \mu_7 - 0.25(\mu_1 + \mu_2 + \mu_3 + \mu_4)$, keeping the convention that we believe the contrast will be positive. **(b)** H_0: $\psi = 0$ and H_a: $\psi > 0$. **(c)** $\hat{\psi} = 0.62333 - 0.25(0.62667 + 0.57 + 0.53667 + 0.53667) = 0.05583$. The standard error is $\text{SE}_{\hat{\psi}} = 0.1205\sqrt{\frac{1}{21} + (0.25)^2\left(\frac{1}{21} + \frac{1}{20} + \frac{1}{20} + \frac{1}{21}\right)} = 0.0295$. The test statistic is $t = 0.05583/0.0295 = 1.894$ with P-value = 0.0302. We can conclude that students using the paper-and-pencil "on task" behavior do significantly better on quizzes such as these than students who are off-task using digital technologies.

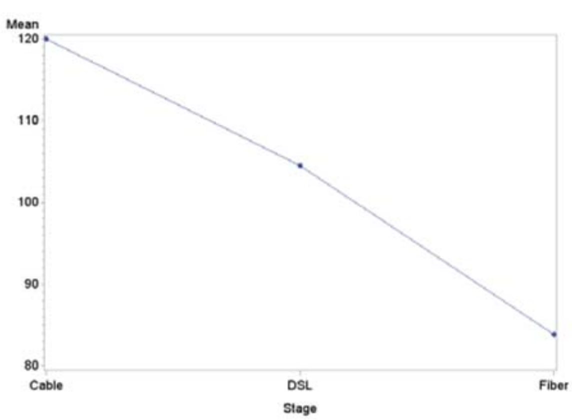
12.41. The power would be larger. For larger differences between alternative means, λ gets bigger, increasing our power to see these differences.

12.43. (a) With roughly equal means, we have $F = 0.00$ and P-value = 0.9995. **(b)** As the means become more different, F increases and P-value decreases. Here, we have $F = 55.53$ and P-value < 0.0001.

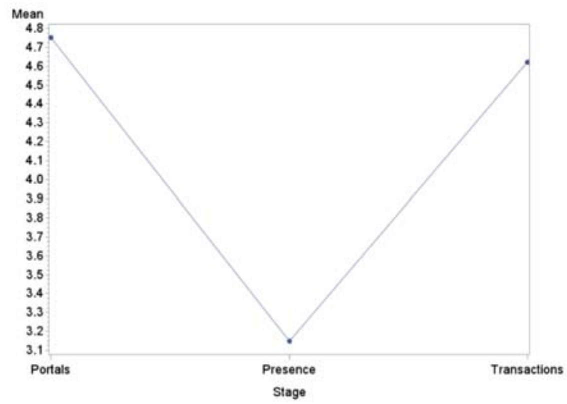




12.45. (a) The pricing among triple-play providers does seem different. Cable has the highest prices followed by DSL. Fiber has the cheapest prices for triple-play. **(b)** Yes, the largest s is less than twice the smallest s; $40.39 < 2(26.09) = 52.18$. **(c)** df = 2, 44, $0.025 < P\text{-value} < 0.05$. There are significant differences in triple-play pricing among the difference provider platforms.



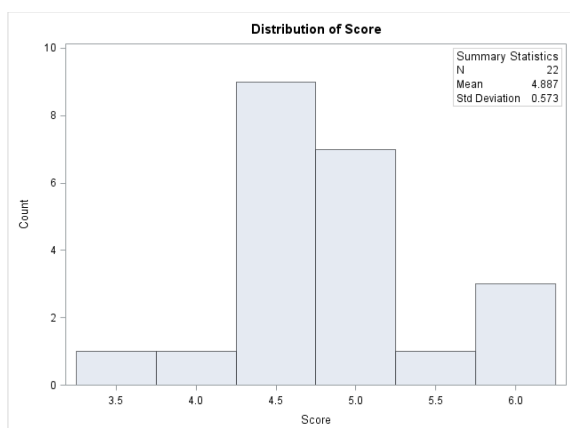
12.47. **(a)** Plot shown below. The Portals and Transactions stages seem to have higher integration features than the Presence stage. **(b)** An observational study. They are not imposing a treatment on the winery. **(c)** Yes, the largest *s* is less than twice the smallest *s*; 2.346 < 2(2.097) = 4.194. **(d)** This shouldn't be a problem because our inference is based on sample means, which will be approximately Normal given the sample sizes. **(e)** $F(2, 190)$; *P*-value < 0.001. There are significant differences in the number of market integration features of the wineries among those with different website stages.

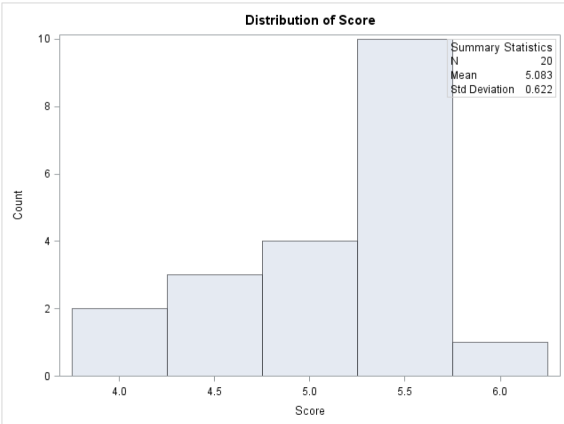


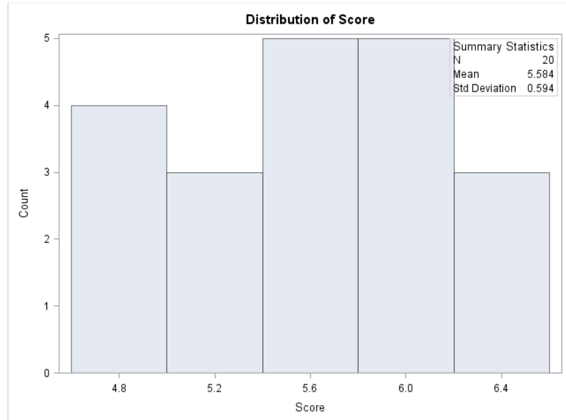
12.49. **(a)** H_0: $\mu_1 = \mu_2 = \mu_3$, H_a: not all of the μ_i are equal, $F = 5.31$, *P*-value = 0.0067. There are significant interval differences among the three groups. **(b)** The Bonferroni procedure shows that Group 2 is not significantly different from either Group 1 or Group 3; however, Group 3 is significantly different (larger) from Group 1. **(c)** This is not appropriate. The regression assumes that Group 2 (coded as 2) would have twice the effect of Group 1 (coded as 1), and Group 3 (coded as 3) would have three times the effect of Group 1, etc. This is likely not true.

12.51. **(a)** Table shown below. Pooling is appropriate; the largest *s* is less than twice the smallest *s*: 0.621669 < 2(0.572914) = 1.146. **(b)** Histograms shown below; while the distributions aren't Normal, there are no outliers or extreme departures from Normality that would invalidate the results. We can likely proceed with the ANOVA.

Level of Food	N N	Score Mean	Score Std Dev
Comfort	**22**	4.8873	0.5729
Control	**20**	5.0825	0.6217
Organic	**20**	5.5835	0.5936



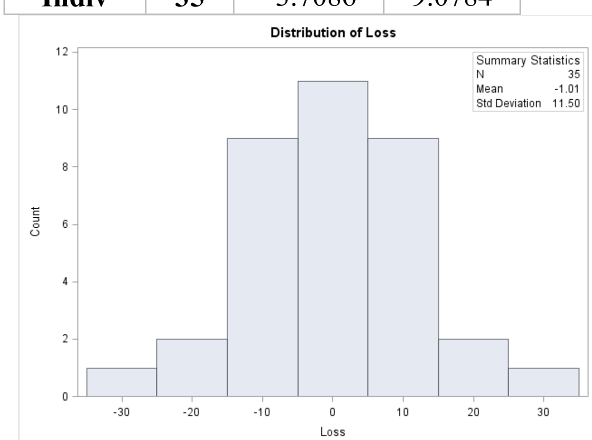


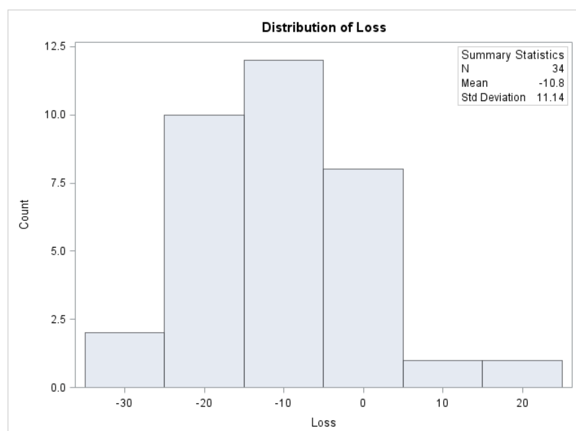


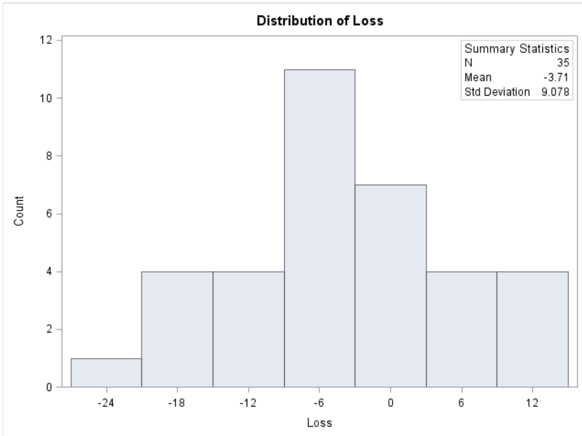
12.53. **(a)** H_0: $\mu_1 = \mu_2 = \mu_3$, H_α: not all of the μ_i are equal, $F = 8.89$, P-value = 0.000. There are significant differences in the number of minutes that the three groups are willing to volunteer. According to the Tukey multiple comparison, the Comfort group is willing to donate significantly more minutes than the Organic group. In other words, the Comfort group shows more prosocial behavior than the Organic group. The Control group is in the middle, not significantly different from either the Comfort or Organic group in the number of minutes they are willing to donate. **(b)** The residual plot shows a slight decrease in variability, suggesting a possible violation of constant variance. The Normal quantile plot looks fine and shows a roughly Normal distribution.

12.55. **(a)** Table shown below. **(b)** Yes, the largest s is less than twice the smallest s; $11.501 < 2(9.078) = 18.156$. **(c)** Histograms shown below. All three distributions are roughly Normal.

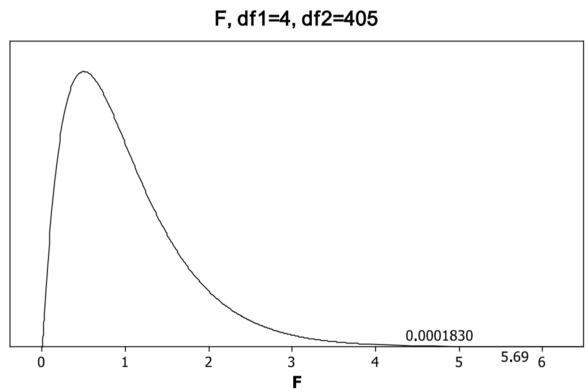
Level of Group	N	Loss Mean	Loss Std Dev
Ctrl	**35**	–1.0086	11.5007
Grp	**34**	–10.7853	11.1392
Indiv	**35**	–3.7086	9.0784





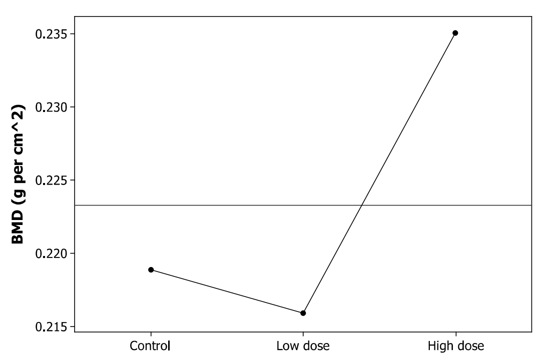


12.57. (a) All weight loss values are divided by 2.2. **(b)** $F = 7.77$, df = 2 and 101, P-value = 0.0007. The results are identical with the transformed data regarding the test statistic, df, and P-value. Transforming the response variable by a fixed amount has no effect on the ANOVA results.



12.59. (a) The variation in sample size is some cause for concern, but there can be no extreme outliers in a 1-to-7 scale, so ANOVA is probably reliable. **(b)** Pooling is appropriate; the largest s is less than twice the smallest s: $1.26 < 2(1.03) = 2.06$. **(c)** With $I = 5$ groups and total sample size $N = 410$, we use an $F(4, 405)$ distribution. We can compare 5.69 with an $F(4, 200)$ distribution in Table E and conclude that P-value < 0.001, or with software determine that P-value = 0.0002. **(d)** Hispanic Americans have the highest emotion scores, Japanese are in the middle, and the other three cultures are the lowest.

12.61. (a) df = 2, 117. **(b)** P-value < 0.001. There are significant average reduction differences among the different groups of bargainers. **(c)** Because the bargainer was the same person each time, there is no way to differentiate if the average reduction was due to race/gender or due to individual ability to bargain. Hence, the results would certainly not be generalizable.



12.63. (a) The means, standard deviations, and standard errors are given (all in grams per cm^2). **(b)** All three distributions appear to be reasonably close to Normal, and the standard deviations are suitable for pooling. **(c)** ANOVA gives $F = 7.72$ (df 2 and 42) and P-value = 0.001, so we conclude that the means are not all the same. **(d)** With df = 42, three comparisons, and $\alpha = 0.05$, the Bonferroni critical value is $t^{**} = 2.4937$. The pooled standard deviation is $s_p = 0.01437$, and the standard error of each difference is

$\text{SE}_D = s_p\sqrt{1/15 + 1/15} = 0.005247$, so two means are significantly different if they differ by $t^{**}\text{SE}_D$ = 0.01308. The high-dose mean is significantly different from the other two. **(e)** Briefly: High doses of kudzu isoflavones increase BMD.

	n	$\bar{x}$	s	$s_{\bar{x}}$
Control	15	0.2189	0.01159	0.002992
Low dose	15	0.2159	0.01151	0.002972
High dose	15	0.2351	0.01877	0.004847

Minitab Output: BMD versus Treatment

```
Source DF SS MS F P
Treatment   2  0.003186  0.001593  7.72  0.001
Error 42  0.008668  0.000206
Total 44  0.011853

S = 0.01437   R-Sq = 26.88%   R-Sq(adj) = 23.39%
```

12.65. (a) Pooling is appropriate; the largest s is less than twice the smallest s: 27.364 < 2(16.594) = 33.188. **(b)** ANOVA gives F = 7.98 (df 2 and 27), for which P-value = 0.002. There are significant differences in bone density between the different jump condition groups; specifically it looks like the high jump group has a much higher mean bone density than the other two groups.

$\bar{x}$	n		s
Control	10	601.1	27.364
Low jump	10	612.5	19.329
High jump	10	638.7	16.594

One-way ANOVA: Density Versus Group

```
Source  DF SS MS F P
Group 2   7434  3717  7.98  0.002
Error   27  12580   466
Total   29  20013

S = 21.58   R-Sq = 37.14%   R-Sq(adj) = 32.49%
```

2.67. $s_p = \sqrt{2.421}$ = 1.556 and the df = 117. **(a)** The contrast is $\psi_{\text{sex}} = \mu_1 - \mu_3$; c_{sex} = 1.055 – 2.310 = –1.255 and $\text{SE}_{c_1} = s_p\sqrt{1/40 + 1/40} = 0.3479,$ so t = –1.255/0.3479 = –3.6. P-value < 0.001. The data shows a difference between the sexes among Hispanics. **(b)** The contrast is $\psi_{\text{nat}} = \mu_1 - \mu_2$; c_{nat} = 1.055 – 1.050 = 0.005 and $\text{SE}_{c_1} = s_p\sqrt{1/40 + 1/40} = 0.3479,$ so t = 0.005/0.3479 = 0.014. P-value > 0.25. The data does not show evidence of a difference between nationalities among males. **(c)** The conclusions from the contrasts were limited to only comparing sex among Hispanics and nationalities among males. Including Anglo females would alleviate this limited inference issue by allowing comparison of sex among both nationalities and comparing nationality among both sexes as well as a possible interaction effect.

12.69. (a) Table shown below. Pooling is not appropriate; the largest s is more than twice the smallest s: $8.6603 > 2(2.8868) = 5.7736$. **(b)** ANOVA gives $F = 137.94$ (df 5 and 12), for which P-value < 0.0005. There are significant Gpi differences among the six types of scaffold material.

	n	*Mean*	s
ECM1	3	65.00%	8.66%
ECM2	3	63.33%	2.89%
ECM3	3	73.33%	2.89%
MAT1	3	23.33%	2.89%
MAT2	3	6.67%	2.89%
MAT3	3	11.67%	2.89%

Minitab Output: Gpi Versus Material

```
Source DF SS MS F P
Material   5  13411.1  2682.2  137.94  0.000
Error 12 233.3 19.4
Total 17  13644.4

S = 4.410   R-Sq = 98.29%   R-Sq(adj) = 97.58%
```

12.71. For convenience, the means and sample sizes are repeated here. In **Exercise 12.26**, we found $s_p = 18.421$. **(a)** We have $\psi_1 = \mu_C - 0.25(\mu_{30\times 1} + \mu_{30\times 2} + \mu_{60\times 1} + \mu_{60\times 2})$, $\psi_2 = 0.5(\mu_{30\times 1} + \mu_{30\times 2}) - 0.5(\mu_{60\times 1} + \mu_{60\times 2})$, $\psi_3 = 0.5(\mu_{60\times 1} - \mu_{60\times 2}) - 0.5(\mu_{30\times 1} - \mu_{30\times 2})$. **(b)** $c_1 = -6.3 - 0.25(-17.4 - 18.4 - 24 - 24) = 14.65$, $c_2 = 0.5(-17.4 - 18.4) - 0.5(-24.0 - 24.0) = 6.1$, $c_3 = 0.5(-24 - 24) - 0.5(-17.4 - 18.4) = -0.5$. $SE_{C1} = 4.209$, $SE_{C2} = SE_{C3} = 3.784$. **(c)** All tests are two-sided (we are interested only in the comparisons; no direction was given). There were df = 114 in the ANOVA for error. The test results are $t_1 = 14.65/4.209 = 3.481$, P-value < 0.001 (from Table D, df = 100) and P-value = 0.0007 (from software); $t_2 = 6.1/3.784 = 1.612$, $0.10 < P\text{-value} < 0.20$ (from Table D, df = 100) and P-value = 0.1097 (from software); $t_3 = -0.5/3.784 = -0.132$, P-value > 0.5 (from Table D, df = 100) and P-value = 0.8952 (from software). The average of the massages relieves arthritis pain in the knee better than conventional treatment; there is not enough evidence to conclude that twice a week relieves pain better than once a week; there is no significant difference in the difference of the averages for 30 and 60 min. It appears that 60 min once or twice a week relieves knee osteoarthritis pain best.

	n	*Mean*
30 min, 1 × /wk	22	−17.4
30 min, 2 × /wk	24	−18.4
60 min, 1 × /wk	24	−24.0
60 min, 2 × /wk	25	−24.0
Usual care (control)	24	−6.3

12.73. (a) The plot shows granularity (due to the fact that the scores are all integers), but otherwise independence is not violated. **(b)** Pooling is appropriate; the largest s is less than twice the smallest s; $1.6 < 2(0.93) = 1.86$. **(c)** The individual normal quantile plots for each shampoo show a lot of granularity due to using integer values but most look at least roughly normal. **(d)** The normal quantile plot for the residuals looks even better and shows that the residuals are indeed normally distributed despite the granularity.