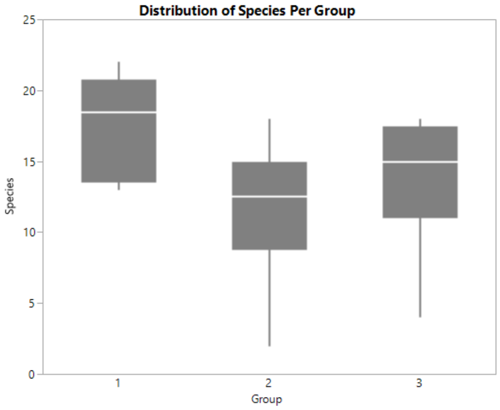
Chapter 27 – One-Way Analysis of Variance: Comparing Several Means

27.1 (a) The null hypothesis is that all age groups have the same (population) mean road-rage measurement, and the alternative is that at least one group has a different mean. **(b)** The F test is quite significant, giving strong evidence that the means are different. The sample means suggest that the degree of road rage decreases with age. (We assume that higher numbers indicate more road rage.)

27.2 (a) The null hypothesis is that the mean political spectrum score is the same for each highest degree earned, and the alternative is that at least one highest-degree earned group has a different mean political spectrum score. **(b)** The F test is significant, providing strong evidence that the mean political spectrum score is not the same for each degree earned. From the graph, it appears the mean political spectrum score decreases as more degrees are earned. In particular, the mean political spectrum score for those with graduate degrees appears to be much smaller than for those with high school or junior college degrees.

27.3 (a) Side-by-side boxplots of species for each group are provided. The boxes overlap, but we see more species in the plots that have never been logged than in the plots that have been logged.



(b) The mean for the group that has never been logged is largest. The mean number of species for the group that was logged 8 years ago is slightly larger than the group logged one year ago. **(c)** The F statistic is $F = 6.02$, and the P-value is $P = 0.006$. There is strong evidence that logging is related to differences in the mean number of species for a forest plot.

27.4 (a) This is an observational study, because prisoners were not selected at random to live in the restrictive environments. Since this is not an experiment, we cannot conclude that living environment causes psychological distress. **(b)** Those sampled and living in the general population had the smallest average psychological distress, and those living in disciplinary segregation has the largest average distress. **(c)** We are testing the hypotheses H_0: $\mu_{\text{G}} = \mu_{\text{AD}} = \mu_{\text{DS}}$ versus H_{a}: not all of

$\mu_{\text{G}}, \mu_{\text{AD}}, \mu_{\text{DS}}$ are the same. The test statistic is $F = 4.86$, and the P-value is $P = 0.016$. There is evidence that the average psychological distress is not the same for prisoners in the three environments.

27.5 (a) Answers will vary due to randomness. **(b)** By moving the middle mean to the same level as the other two, it is possible to reduce *F* to about 0.02, which has a *P*-value very close to the left end of the scale (near 1). **(c)** By moving any mean up or down (or any two means in opposite directions), the value of *F* increases (and *P* decreases) until it moves to the right end of the scale.

27.6 (a) *F* can be made as small as 0.3174, for which *P* > 0.5. **(b)** *F* can be made quite large (and *P* small) by separating the means—for example, by moving two means all the way down and one all the way up.

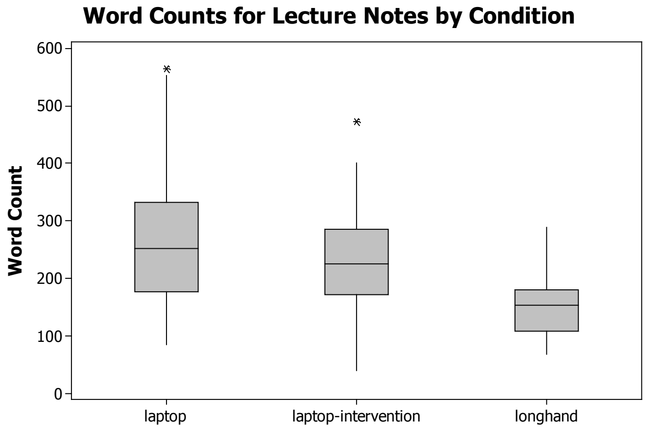
27.7 (a) We have $s_1^2 = 12.45, s_2^2 = 19.11,$ and $s_3^2 = 20.25$. The largest standard deviation is $s_3 = \sqrt{20.25} = 4.5$, and the smallest standard deviation is $s_1 = \sqrt{12.45} = 3.53$. The ratio of largest to smallest is $4.5/3.53 = 1.27$, which is less than 2. Conditions are satisfied. **(b)** The standard deviations are $s_{\text{G}} = 21, s_{\text{DS}} = 8$, and $s_{\text{AD}} = 16$. The ratio of largest to smallest is $\frac{21}{8} = 2.625$, which is at least 2. Conditions are satisfied.

27.8 The standard deviations (0.1201, 0.1472, 0.1134) do not violate our rule of thumb. However, the distributions provided appear to be skewed and have outliers, especially the one-year-ago group.

```
Never logged        1 year ago          8 year ago
   4 | 8               4 | 2               4 |
   5 |                 5 |                 5 |
   6 | 357             6 |                 6 | 8
   7 | 5889            7 | 7               7 | 8
   8 | 111             8 | 3588            8 | 13
   9 | 5               9 | 01123           9 | 34
  10 |                10 | 0              10 | 000
```

27.9 (a) STATE: We want to determine if there is a difference in average word count for people who take notes longhand, with a laptop, or with laptop-intervention. PLAN: We will examine the data by looking at side-by-side boxplots. We will assess if it is safe to use ANOVA to test the hypotheses H_0: all means are the same against H_a: at least one mean is different from the others. SOLVE: Looking at the summary statistics given, those who took longhand notes clearly seem to have written fewer average words than either laptop condition. However, the laptop-intervention group had the smallest minimum. The laptop group clearly seems to have written the most words. We note that the ratio of largest to smallest standard deviation is 118.5/59.64 = 1.99, which is just slightly less than 2. The boxplots graphically display and support the summary statistics. The laptop group shows the most variability, and the longhand group shows the least. The laptop group seems right-skewed (at least more than the other two conditions), and both laptop groups show outliers. However, with the smallest sample size being 48, the central limit theorem says the sample means should be approximately Normal. So it is safe to use the F test. The provided output shows $F = 17.04$ and $P < 0.0005$. CONCLUDE: There is clearly a difference in the number of words written while taking notes with the three methods. It seems fairly obvious from the graphs that those who take notes on a laptop write the most; those taking longhand notes write the least.

```
Variable   condition              N  N*    Mean    SE  Mean  StDev   Minimum   Q1
wordcount  laptop                51   0   260.9       16.6   118.5   85.0      177.0
           laptop-intervention   52   0   229.0       11.8   84.8    85.0      171.8
           Longhand              48   0  155.98       8.61   59.64   68.00     108.25

Variable   condition              Median     Q3  Maximum
wordcount  laptop                  252.0  332.0    566.0
           laptop-intervention     226.0  285.0    473.0
           Longhand               153.50 181.00   289.00
```



One-way ANOVA: wordcount versus condition

```
Source      DF       SS      MS      F      P
condition    2   284599  142300  17.04  0.000
Error      148  1235978    8351
Total      150  1520577

S = 91.38   R-Sq = 18.72%   R-Sq(adj) = 17.62%
```

27.10 (a) The number of populations is $I = 3$; the sample sizes from each population are $n_1 = n_2 = 12$ and $n_3 = 9$; the total sample size is $N = 12 + 12 + 9 = 33$. **(b)** The numerator (between groups) df: $I - 1 = 2$; the denominator (within groups) df: $N - I = 30$.

27.11 (a) $I = 3$ and $N = 96$, so df = 2 and 93. **(b)** $I = 3$ and $N = 90$, so df = 2 and 87.

27.12 (a) Let μ_1 denote the mean length of *H. bihai*, μ_2 the mean length of *H. caribaea yellow*, and μ_3 the mean length of *H. caribaea red.* The three null hypotheses are $H_0\colon \mu_1 = \mu_2$, $H_0\colon \mu_1 = \mu_3$, and $H_0\colon \mu_2 = \mu_3$. **(b)** Use technology to compute the three Tukey simultaneous confidence intervals (provided). None of the intervals contain zero, so there is enough evidence to conclude none of the pairs are equal. That is, the mean lengths for all three groups are different.

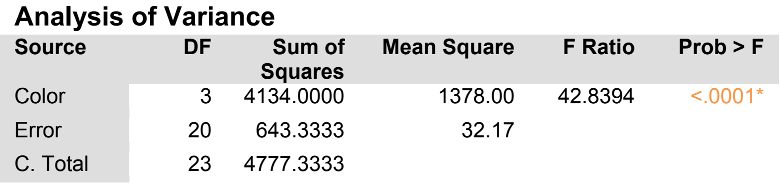
$\mu_1 - \mu_2$	10.32 to 12.51
$\mu_1 - \mu_3$	6.90 to 8.87
$\mu_2 - \mu_3$	−4.54 to −2.52

27.13 (a) Let μ_1 be the mean word count for the laptop group, μ_2 the mean count for the laptop-intervention group, and μ_3 the mean count for the longhand group. The intervals are provided.

$\mu_1 - \mu_2$	61.43 to 148.45
$\mu_1 - \mu_3$	29.73 to 116.35
$\mu_2 - \mu_3$	−10.74 to 74.54

(b) 95% confidence means, prior to taking any samples, there is a 0.95 probability that all of the intervals simultaneous capture the true pairwise differences in mean word count. **(c)** The laptop and laptop-intervention group and the laptop and longhand group significantly differ, because zero is not contained in those two intervals.

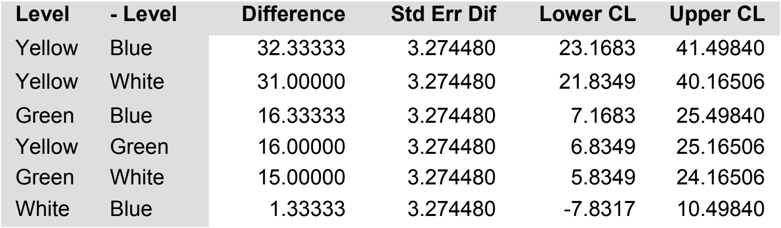
27.14 (a) Let μ_B be the mean number of beetles trapped with the blue board, μ_G with the green board, μ_W with the white, and μ_Y with the yellow. Since the sample size is small for each group, we need to assume the distribution for each group is approximately Normal. The smallest standard deviation is $s_W = 3.76$ and the largest is $s_Y = 6.79$, so the ratio of largest to smallest is less than 2. It is safe to use the ANOVA *F* test of the hypotheses $H_0\colon \mu_B = \mu_G = \mu_W = \mu_Y$ versus H_a: not all of the means are the same. Output from JMP is provided. The test statistic is $F = 42.84$, and the *P*-value is $P < 0.0001$. There is strong evidence that the mean number of beetles is not the same for the four colored boards.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Color	3	4134.0000	1378.00	42.8394	<.0001*
Error	20	643.3333	32.17		
C. Total	23	4777.3333			

(b) There are six pairwise comparisons when there are four groups. **(c)** We can test each pair of means using Tukey simultaneous 95% confidence intervals (provided in the JMP output). The only pair that is not significantly different is the white/blue pair. Since all of the yellow intervals contain only positive values, yellow is significantly better than every other group.



Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL
Yellow	Blue	32.33333	3.274480	23.1683	41.49840
Yellow	White	31.00000	3.274480	21.8349	40.16506
Green	Blue	16.33333	3.274480	7.1683	25.49840
Yellow	Green	16.00000	3.274480	6.8349	25.16506
Green	White	15.00000	3.274480	5.8349	24.16506
White	Blue	1.33333	3.274480	-7.8317	10.49840

27.15 (a) The sample sizes are quite large, and the *F* test is robust against non-Normality with large samples. **(b)** Yes (barely): The ratio is 3.11/1.60 = 1.94, which is slightly less than 2. **(c)** We have *I* = 3 and *N* = 1342. The details of the computations are given; some of the fractional values have been rounded.

$$\bar{x} = 1.31 = \frac{244(2.22) + 734(1.33) + 364(0.66)}{1342}$$
$$\text{SSG} = 356.14 = 244(2.22 - 1.31)^2 + 364(0.66 - 1.31)^2$$
$$\text{MSG} = 5.12 = \frac{356.14}{3 - 1}$$
$$\text{SSE} = 6859.65 = 243(3.11)^2 + 734(1.33 - 1.31)^2 + 364(0.66 - 1.31)^2$$
$$\text{MSE} = 5.12 = \frac{6859.65}{1342 - 3}$$
$$\text{F} = 34.76 = \frac{178.07}{5.12}$$

(d) We compare to an *F* distribution with df = 2 and 1339. We have strong evidence that the means differ among the age groups; specifically, road rage decreases with age.

27.16 (a) Yes, the ratio of the largest and smallest standard deviation is $\frac{1.601}{1.330}$ = 1.2. This is less than 2, so it is safe to use ANOVA. **(b)** The overall mean is $\bar{x} = \frac{1228(4.171) + 179(4.134) + 464(4.002) + 279(3.735)}{1228 + 179 + 464 + 279} = 4.075$. The mean square for groups is $\text{MSG} = \frac{1228(4.171 - 4.075)^2 + 179(4.134 - 4.075)^2 + 464(4.002 - 4.075)^2 + 279(3.735 - 4.075)^2}{4 - 1} = 15.555$, and the mean square error is $\text{MSE} = \frac{1227(1.377^2) + 178(1.330^2) + 463(1.507^2) + 278(1.601^2)}{1228 + 179 + 464 + 279 - 4} = 2.053$. The test statistic is $F = \frac{15.555}{2.053} = 7.577$. **(c)** The *F* distribution has numerator degrees of freedom $I - 1 = 4 - 1 = 3$ and denominator degrees of freedom $N - I = 2146$. The *P*-value is $P < 0.0001$. There is strong evidence that the mean political spectrum score is not the same for the four education groups. Based on the sample means, it appears the groups with more education have lower mean political spectrum scores.

27.17 (a) We have independent samples from the five groups, and the standard deviations easily satisfy our rule of thumb (1.40/1.28 = 1.09 < 2). **(b)** The details of the computations, with *I* = 5 and *N* = 4413, are given.

$$\overline{x} = 2.459 = \frac{(809)(2.57) + (1860)(2.32) + (654)(2.63) + (883)(2.51) + (207)(2.51)}{4413}$$

$$SSG = 67.86 = 809(2.57 - \overline{x})^2 + 654(2.63 - \overline{x})^2 + 883(2.51 - \overline{x})^2 + 207(2.51 - \overline{x})^2$$

$$MSG = 16.97 = \frac{67.86}{5 - 1}$$

$$SSE = 8010.98 = 808(1.40)^2 + 1859(1.39)^2 + 653(1.32)^2 + 882(1.31)^2 + 2.6(1.28)^2$$

$$MSE = 1.82 = \frac{8010.98}{4413 - 5}$$

$$F = 9.34 - \frac{16.97}{1.82}$$

(c) The ANOVA is very significant (P < 0.001), but this is not surprising because the sample sizes were very large. The differences might not have practical importance. (The largest difference is 0.31, which is relatively small on a five-point scale.)

27.18 (c) the means of several populations.

27.19 (b) 2 and 447. There are 3 − 1 = 2 df for groups and 450 − 3 = 447 df for error.

27.20 (c) the mean ADCS-AD inventory scores for the three groups are not all the same. The alternate hypothesis for ANOVA is always that there is some difference in the means (but it does not specify the type of difference).

27.21 (b) a family of distributions that are right-skewed and take only values greater than or equal to 0.

27.22 (a) there is strong evidence ($P = 0.000$) that the mean heart rates are not the same for all three conditions.

27.23 (a) we don't know how confident we can be that all three intervals cover the true differences in means.

27.24 (c) 14.08. $F = \frac{2388/2}{3561/42} = 14.08$.

27.25 (c) the assumption that the data are independent for the three days is unreasonable because the same teams were observed each day. We do not have three independent samples from three populations.

27.26 (c) six. There are six pairwise comparisons.

27.27 (b) just $\mu_1 \neq \mu_2$; there is not enough evidence to draw conclusions about the other pairs of means. There is evidence to conclude μ_1 and μ_2 are different, but there is not evidence to suggest the other pairs are different. In testing, we never conclude the null is true.

27.28 The populations are college students who might view the advertisement with an art image, college students who might view the advertisement with a non-art image, and college students who might view the advertisement with no image. The response variable is student evaluation of the advertisement on the 1 to 7 scale. We test the hypothesis $H_0\colon \mu_1 = \mu_2 = \mu_3$ (all three groups have equal mean advertisement evaluation) versus H_a: not all means are equal. There are I = 3 populations; the samples sizes are $n_1 = n_2 = n_3 = 39$, so there are N = 39 + 39 + 39 = 117 individuals in the total sample. There are then I − 1 = 3 − 1 = 2 and N − I = 117 − 3 = 114 df.

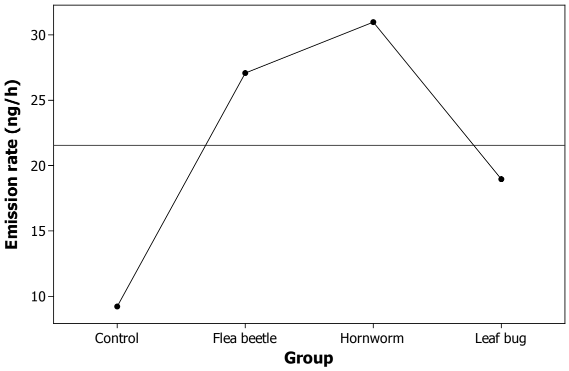
27.29 The populations are students who wear a lemon-flavored mouth guard, students who wear a non-flavored mouth guard, and students who wear no mouth guard. The response is the rating of perceived exertion. $I = 3, N = 43, n_1 = 12, n_2 = 15$, and $n_3 = 16$. The degrees of freedom are 2 and $43 - 3 = 40$.

27.30 There are I = 4 populations: learning-disabled children with each of the three accommodations plus a control group. The response variable is the scores on the state math exam. We test the hypothesis $H_0\colon \mu_1 = \mu_2 = \mu_3 = \mu_4$ (all four groups have equal means) versus H_a: not all means are equal. The sample sizes are $n_1 = n_2 = n_3 = n_4 = 25$, with a total sample size of N = 100. The degrees of freedom are therefore I − 1 = 3 and N − I = 96.

27.31 The response variable is hemoglobin A1c level. We have I = 4 populations; a control (sedentary) population, an aerobic exercise population, a resistance training population, and a combined aerobic and resistance training population. We test hypothesis $H_0\colon \mu_1 = \mu_2 = \mu_3 = \mu_4$ (all four groups have equal mean hemoglobin A1c levels) versus H_a: not all means are equal. Sample sizes are n_1 = 41, n_2= 73, n_3= 72, and n_4= 76. Our total sample size is N = 41 + 73 + 72 + 76 = 262. We have I − 1 = 4 − 1 = 3 and N − I = 262 − 4 = 258 df.

27.32 (a) The ratio of largest-to-smallest standard deviation is 1.50/0.87 = 1.72, which is less than 2. So, ANOVA will be safe to use for comparing means. Comparing the means provided, males seeing a model are clearly more positive in their evaluation of the product. Among female subjects, there is little difference between the impact of model or student confederate. Notice that for both sexes, seeing a model confederate scores higher on average than seeing a student confederate. **(b)** There are I = 5 populations being compared. We have N = 22 + 23 + 24 + 23 + 27 = 119 subjects in total. There are I − 1 = 5 − 1 = 4 and N − I = 119 − 5 = 114 df. With F = 8.30 and using software, P = 0.000007, so there is overwhelming evidence of a difference in population means.

27.33 (a) The graph provided suggests that emissions rise when a plant is attacked because the mean control emission rate is half the smallest of the other rates.



(b) The null hypothesis is that all groups have the same mean emission rate. The alternative is that at least one group has a different mean emission rate. **(c)** The most important piece of additional information would be whether the data are sufficiently close to Normally distributed. (From the description, it seems reasonably safe to assume that these are more or less random samples.) **(d)** The SEM =$s/\sqrt{8}$, so we can find the standard deviations by multiplying by $\sqrt{8}$; they are 16.77, 24.75, 18.78, and 24.38. However, this factor of $\sqrt{8}$ would cancel out in the process of finding the ratio of the largest and smallest standard deviations, so we can simply find this ratio directly from the SEMs: $\frac{8.75}{5.93} = \frac{24.75}{16.77} = 1.48$, which satisfies our rule of thumb that the largest sample standard deviation is no more than twice the smallest sample standard deviation.

27.34 Only Design A would allow use of one-way ANOVA because it produces four independent sets of data. The data resulting from Design B would be dependent (a subject's responses to the first list would be related to that same subject's responses to the other lists), so that ANOVA would not be appropriate for comparison.

27.35 (a) The stemplots are provided, as are the means and standard deviations in the Minitab output. The means suggest that extra water in the spring has the greatest effect on biomass, with a lesser effect from added water in the winter. ANOVA is risky with these data; the standard deviation ratio is nearly 3 (58.77/21.69 = 2.71), and the winter and spring distributions may have skewness or outliers (although it is difficult to judge with such small samples).

```
Winter                Spring                Control
 1 |                   1 |                   1 | 11
 1 |                   1 |                   1 | 2
 1 | 4                 1 |                   1 | 44
 1 | 67                1 |                   1 | 7
 1 | 8                 1 |                   1 |
 2 |                   2 |                   2 |
 2 |                   2 |                   2 |
 2 |                   2 |                   2 |
 2 |                   2 |                   2 |
 2 |                   2 | 889               2 |
 3 |                   3 | 1                 3 |
 3 |                   3 | 2                 3 |
 3 |                   3 |                   3 |
 3 |                   3 |                   3 |
 3 |                   3 | 8                 3 |
```

```
Level     N    Mean   StDev
control   6  136.65   21.69
spring    6  315.39   37.34
winter    6  205.17   58.77
```

(b) We wish to test whether the mean biomass from any group differs from the others: $H_0: \mu_{\mathrm{w}} = \mu_3 = \mu_c$ (all treatments have the same mean) versus H_a: at least one mean is different. **(c)** ANOVA gives a statistically significant result (F = 27.52, df = 2 and 15, P < 0.0005), but as noted in part (a), the conditions for ANOVA are not satisfied. Based on the stemplots and the means, however, we should still be safe in concluding that added water increases biomass.

One-Way ANOVA: Biomass versus Treatment

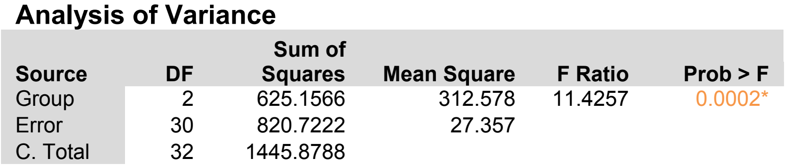
```
Source        DF        SS      MS      F      P
Treatment      2     97583   48792  27.52  0.000
Error         15     26593    1773
Total         17    124176

S = 42.11   R-Sq = 78.58%   R-Sq(adj) = 75.73%
```

27.36 The ANOVA test statistic is F = 4.92 (df = 3 and 92), which has P = 0.003, so there is strong evidence that the means are not all the same. In particular, list 1 seems to be the easiest, and lists 3 and 4 are the most difficult.

27.37 (a) The table is provided. The ratio of the largest and smallest standard deviation is less than 2, so the condition is satisfied. The means reveal that unsurprisingly, logging reduces the number of trees.

Group	Mean	Standard Deviation
1	23.75	5.065
2	14.08	4.98
3	15.78	5.76

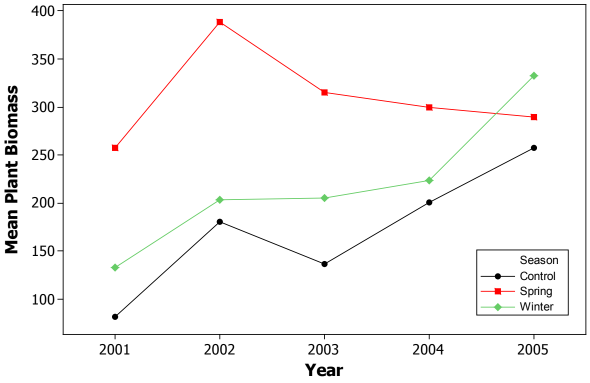
(b) JMP output is provided for the ANOVA test of $H_0: \mu_1 = \mu_2 = \mu_3$ versus H_{a}: the mean number of trees is not the same for all three groups. The F statistic is F = 11.43, and the P-value is P = 0.0002. There is strong evidence that the mean number of trees is not the same for the three groups.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Group	2	625.1566	312.578	11.4257	0.0002*
Error	30	820.7222	27.357		
C. Total	32	1445.8788			

27.38 (a) See the plot provided.



(b) There is a slight increase in growth when water is added in the wet season, but there is a much greater increase when it is added during the dry season. **(c)** The means differ significantly during the first three years. **(d)** The year 2005 is the only one for which the winter biomass was higher than the spring biomass.

27.39 PLAN: We compare the mean biomass of the three groups using a plot of the means and ANOVA, testing H_0: $\mu_{\text{w}} = \mu_3 = \mu_c$ versus H_a: at least one mean is different. SOLVE: It is reasonable to view the samples as SRSs from the three populations, but the standard deviation ratio is high (49.59/11.22 = 4.42), so ANOVA is risky. The Minitab output provided includes a table of the means and a display that is equivalent to a plot of the means. As with the 2003 data, the means suggest that extra water in the spring has the greatest effect on biomass, with a lesser effect from added water in the winter. ANOVA gives a statistically significant result (F = 43.79, df = 2 and 15, P < 0.0001). CONCLUDE: The combination of the (questionable) ANOVA results and the means supports the conclusion that added water in the spring increases biomass. The benefit of additional water in the winter is not so clear, especially when taking the plot of means in the solution to the previous exercise.

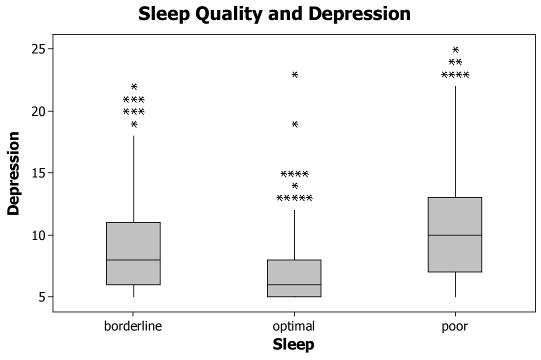
One-Way ANOVA: 2001 versus Trt

```
Source      DF      SS     MS      F      P
Trt          2   98451  49225  43.79  0.000
Error       15   16863   1124
Total       17  115314
```

```
 S = 33.53   R-Sq = 85.38%   R-Sq(adj) = 83.43%

                                  Individual 95% CIs For Mean Based on
                                  Pooled StDev
 Level    N    Mean    StDev     -+---------+---------+---------+--------
 control  6    81.67   28.07     (----*---)
 spring   6    257.69  49.59                                (----*----)
 winter   6    132.58  11.22              (----*----)
                                  -+---------+---------+---------+--------
                                  60       120       180       240

Pooled StDev = 33.53
```

27.40 In addition to a high standard deviation ratio (117.18/35.57 = 3.29), the spring biomass distribution has a high outlier.

27.41 (a) STATE: Does sleep quality affect depression? PLAN: We have data on sleep quality and depression scores for 898 students at a large Midwestern university. We'll have to assume these students are close to a random sample of college students and that the observations (students) are independent of one another. SOLVE: With such large sample sizes, we'll use side-by-side boxplots to examine the distributions. All three groups show outliers at the high end of the depression score range, but with such large samples (the smallest is 246), it is reasonable to believe the sample means have Normal distributions. The condition on standard deviations is satisfied, because 4.719/2.560 = 1.84 < 2. We have F = 75.52 with df = 2 and 895, giving P = 0.000 (to three decimal places). CONCLUDE: The mean depression scores for the three levels of sleep quality are not the same. From the given output and graphs, it appears the mean depression score for poor sleepers is highest; the mean depression score for optimal sleepers is lowest.

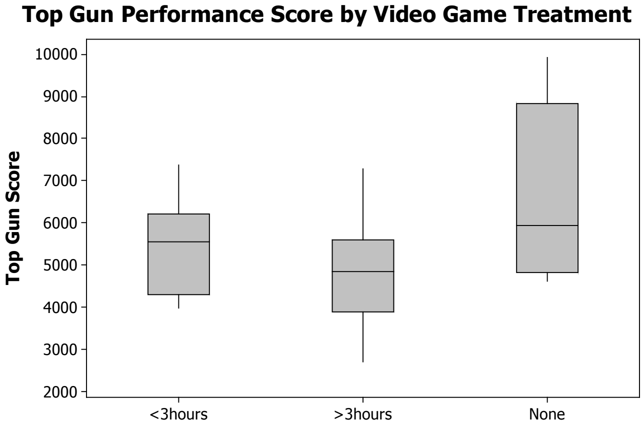


```
 Source        DF        SS       MS       F       P
 Sleep          2    2162.3   1081.1   72.52   0.000
 Error        895   13343.7     14.9
 Total        897   15506.0

 Level          N    Mean   StDev
 borderline   246   8.764   3.892
```

```
Level          N    Mean   StDev
optimal      309   7.013   2.560
poor         343  10.656   4.719
```

(b) Assuming the students were randomly selected, the large sample size would lead us to believe these students are most likely representative of other college students. **(c)** Students were not randomly assigned to sleep conditions. Explanations about causation may vary, but this might well be a case of one condition (poor sleep) feeding the other (depression) in a vicious cycle.

27.42 Side-by-side boxplots show that the scores for both video game groups are fairly symmetric, whereas the distribution for those who have never played video games is right-skewed and more variable than the others. Because lower scores indicate better performance, it seems that playing video games may help with laparoscopic surgery skills. As seen in the results provided, the standard deviations meet our rule of thumb because 1947/1106 = 1.76 < 2. The ANOVA results are also provided. We have $F = 4.92$ and $P = 0.014$, so we have very good evidence to reject the null hypothesis of no difference in means. This study supports the hypothesis that greater prior video game experience might help surgeons learn laparoscopic skills more easily.



```
Variable   category   N  N*  Mean    SE  Mean  StDev  Minimum  Q1       Median    Q3
topgun     <3hours    9   0   5420        369   1106   3968     4308       5540  6204
           >3hours    9   0   4787        438   1313   2703     3884       4845  5596
           None      15   0   6793        503   1947   4605     4828       5947  8837
```

One-way ANOVA: topgun versus category

```
Source      DF         SS        MS     F      P
category     2   25157257  12578628  4.92  0.014
Error       30   76645322   2554844
Total       32  101802578

S = 1598   R-Sq = 24.71%   R-Sq(adj) = 19.69%
```

27.43 (a) We can be 99% confident that all three of these intervals capture the true difference between pairs of population means. **(b)** Combining the results from the *F* test and the multiple comparisons, we can conclude that on average, depression is greatest for those with poor sleep quality and depression is the lowest for those with optimal sleep quality.

27.44 (a) We can be 95% confident that all three intervals capture the true difference between pairs of population means. **(b)** Combining the results from the *F* test and multiple comparisons, we conclude video game experience is related to scores. More specifically, the average score for those with no experience is higher (worse) than the average score for those with at least three hours of experience. The difference between no experience and less than three hours is not significant, neither is the difference between less than three hours and at least three hours.

27.45 (a) Stemplots are provided. There is some degree of left-skew in the data corresponding to lemon odor, but it is not strong. At least there is not strong evidence of non-Normality in any of the population distributions based on these stemplots. There are no real outliers.

```
 Lavender
  5 |
  6 |
  7 | 6
  8 | 89
  9 | 234578
 10 | 12345566788999
 11 | 46
 12 | 1469
 13 | 7
```

```
 Lemon
  5 | 6
  6 | 03
  7 | 3458
  8 | 338889
  9 | 014677
 10 | 145688
 11 | 23
 12 |
 13 |
```

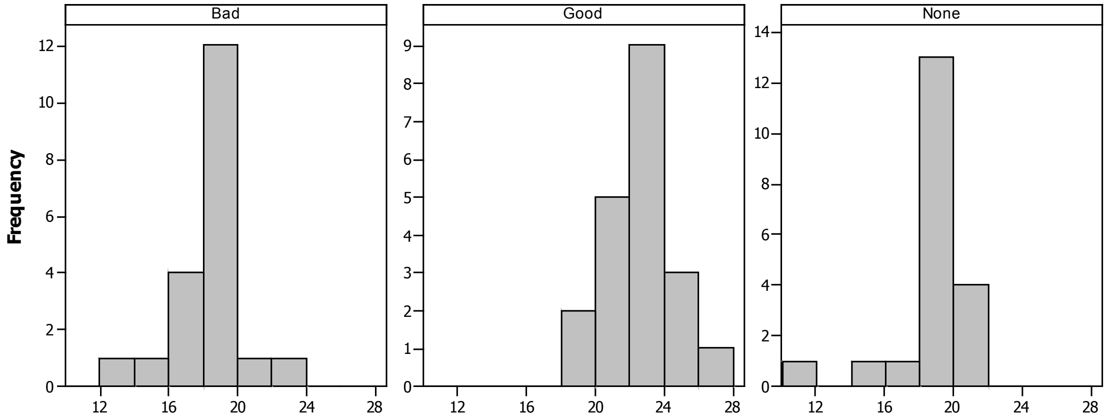
```
 No odor
  5 |
  6 | 89
  7 | 223569
  8 | 445677
  9 | 1222368
 10 | 136779
 11 | 58
 12 | 1
 13 |
```

(b) STATE: Do customer times differ on average depending on the odor present? PLAN: We will compare mean times spent in the restaurant by using ANOVA. SOLVE: As discussed in part (a), there is little evidence of non-Normality in any of the three distributions. Also, the three standard deviations are reasonably close: The ratio of largest standard deviation to smallest standard deviation is 15.44/13.10 = 1.18, which is less than 2. It is safe to apply ANOVA procedures. The Minitab output is provided. We have *F* = 10.861 with df = 2 and 85, yielding *P* = 0.000. There is overwhelming evidence of a difference in the mean amount of time that customers spend in the restaurant, depending on the odor present. Lavender odor yields the longest mean time, while lemon odor reduces time spent on average, compared with no odor at all.

One-way ANOVA: Time versus Odor

```
Source      DF     SS      MS      F      P     Level      N    Mean   StDev
Odor         2   4569    2285  10.86  0.000     lavender  30  105.70   13.10
Error       85  17879     210                   lemon     28   89.79   15.44
Total       87  22448                           noodor    30   91.27   14.93

S = 14.50   R-Sq = 20.35%   R-Sq(adj) = 18.48%
```

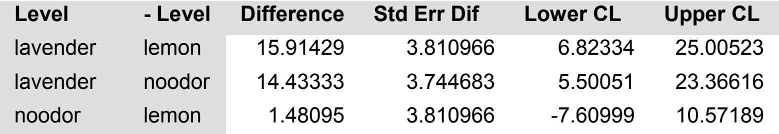
27.46 STATE: Are the mean tip percents constant for all types of weather forecasts (no forecast, good forecast, bad forecast)? PLAN: We will perform an ANOVA test for the equality of means. SOLVE: First, we see that the ratio of largest standard deviation to smallest standard deviation is 2.388/1.959 = 1.22, which is less than 2. Histograms of the samples are provided. There is some evidence of non-Normality and perhaps one outlier in the "no weather report" group. We proceed, because the samples are reasonably large. From the given output, we have F = 20.679 with 3 − 1 = 2 and 60 − 3 = 57 df, with P = 0.000. CONCLUDE: There is overwhelming evidence that the mean tip percents are not the same for all three groups. Examination of the summary statistics and the histograms provided suggests that while the mean tip for the bad forecast group is similar to that of the no forecast group, the mean tip for the good forecast is higher.



One-way ANOVA: Percent versus Report

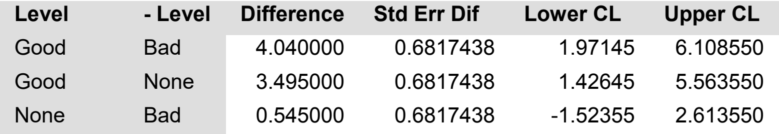
```
Source      DF      SS     MS      F      P     Level     N    Mean   StDev
Report       2  192.22  96.11  20.68  0.000     Bad      20  18.180   2.098
Error       57  264.92   4.65                   Good     20  22.220   1.959
Total       59  457.15                          None     20  18.725   2.388

S = 2.156   R-Sq = 42.05%   R-Sq(adj) = 40.02%
```

27.47 (a) Let μ_{L} denote the mean time with lavender scent, μ_{Le} the mean for lemon, and μ_{N} the mean with no odor. We are testing the three hypotheses $H_0\colon \mu_{\mathrm{L}} = \mu_{\mathrm{Le}}$, $H_0\colon \mu_{\mathrm{L}} = \mu_{\mathrm{N}}$, and $H_0\colon \mu_{\mathrm{Le}} = \mu_{\mathrm{N}}$. **(b)** JMP output for the Tukey pairwise comparisons is provided. In Exercise 27.45, we concluded that the means of the three groups are not all the same. Based on the pairwise comparisons, there is a significant difference between lavender and both lemon and no odor. There is not a significant difference between lemon and no odor. We can conclude that the smell of lavender is related to a higher mean time than the other two smells, and lemon and no odor are similar with respect to the mean time.



Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL
lavender	lemon	15.91429	3.810966	6.82334	25.00523
lavender	noodor	14.43333	3.744683	5.50051	23.36616
noodor	lemon	1.48095	3.810966	-7.60999	10.57189

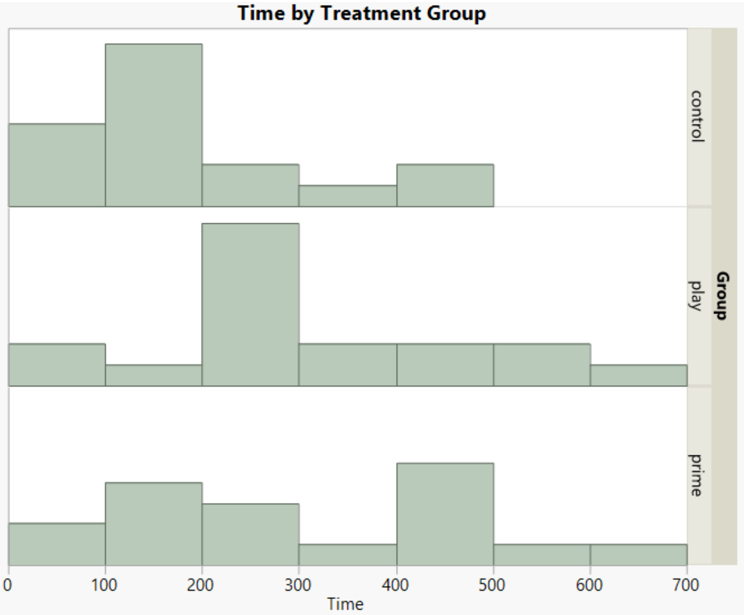
27.48 (a) The Tukey simultaneous 99% confidence intervals are given in the JMP output.

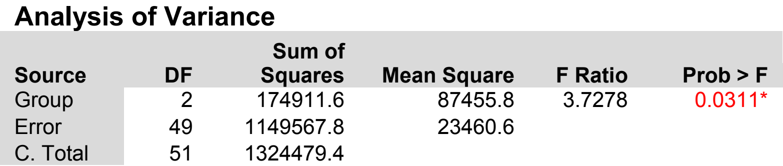


Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL
Good	Bad	4.040000	0.6817438	1.97145	6.108550
Good	None	3.495000	0.6817438	1.42645	5.563550
None	Bad	0.545000	0.6817438	-1.52355	2.613550

(b) 99% confidence means, prior to collecting any data, there is a 0.99 probability that all three intervals will contain the true difference in mean tipping percent. **(c)** At the 0.01 significant levels, the Good and Bad and the Good and None groups significantly differ. Based on the intervals, the true mean tipping percent is significantly higher for the Good group than for either of the other groups. There is not a significant difference in mean tipping percent for None and Bad.

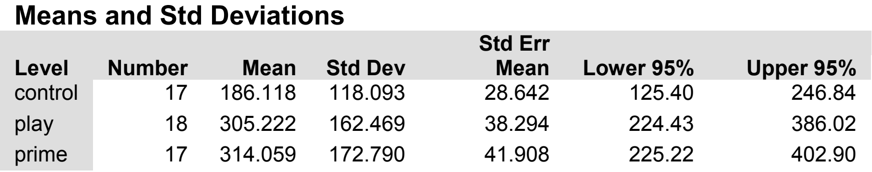
27.49 STATE: We want to know if the average time to ask for help differs significantly for the three treatment groups. PLAN: We will examine the data to compare the effect of the treatments and determine if we can use ANOVA to test the significant of the observed differences in mean times to ask for help. SOLVE: Histograms of the data for each of the three groups and JMP output are provided. The histograms are not symmetric, but with 17 or 18 observations in each group, the sample size is likely large enough for Normality to hold since there is not strong skewness nor are there outliers. The standard deviations satisfy the rule of thumb, and so it is safe to use ANOVA. Using ANOVA to test the null hypothesis that the mean time for the three groups is the same against the null hypothesis that it is not the same for all three groups, we get a test statistic of $F = 3.73$, and a P-value of $P = 0.031$. CONCLUDE: There is good evidence that the mean time to ask for help is not the same for the three groups. Based on the histograms and summary statistics, it appears being reminded about money is related to people taking longer to ask for help.





Analysis of Variance

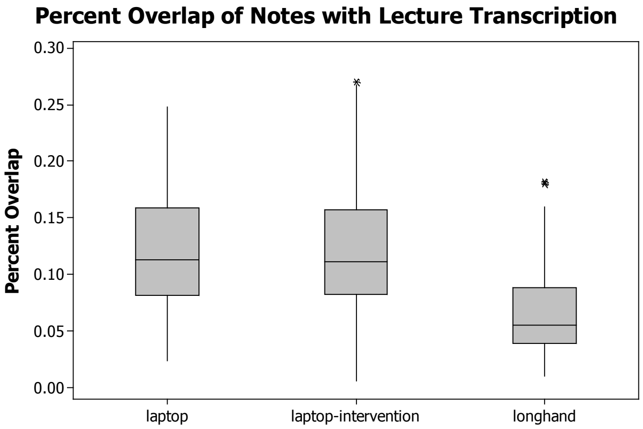
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Group	2	174911.6	87455.8	3.7278	0.0311*
Error	49	1149567.8	23460.6		
C. Total	51	1324479.4			



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
control	17	186.118	118.093	28.642	125.40	246.84
play	18	305.222	162.469	38.294	224.43	386.02
prime	17	314.059	172.790	41.908	225.22	402.90

27.50 STATE: We want to know if those who take notes with a laptop are more likely to take verbatim notes, as measured by the percent of matches of three-word chunks transcribed from the lecture. PLAN: Examine side-by-side boxplots for shape. (Are these reasonably symmetric distributions, even though sample sizes are fairly large?) Check the standard deviation condition, and if appropriate, continue to ANOVA. SOLVE: The provided boxplots of the data are somewhat right-skewed, especially for the intervention and longhand groups. Those two groups also have mild high outliers. The mean and standard deviations are also provided. We note that the means for both laptop groups are almost equal and about twice the value of the mean for the longhand group (the same relationship is seen with the medians in the boxplots). The ANOVA finds $F = 16.63$ with $P < 0.0005$. CONCLUDE: There is a clear difference in the amount of overlap with the actual lecture. The ANOVA finds a highly significant difference; looking at the treatment group means, those who take longhand notes have the least amount of overlap with the actual lecture. This could explain why those students had a greater understanding of the concepts—they may have been more likely to have digested the lecture as it was happening, as indicated by using their own words and phrases in their notes.

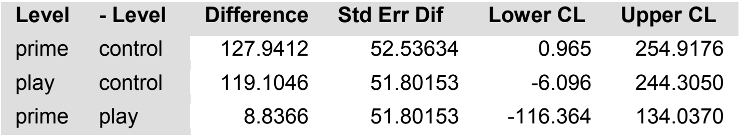


```
                  Mean    StDev
laptop         0.12109  0.05045
laptop-inter   0.12067  0.06007
longhand       0.06881  0.04219
```

One-way ANOVA: percent overlap versus condition

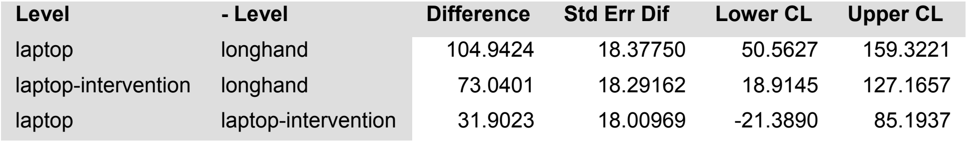
```
Source       DF        SS        MS      F      P
condition     2   0.08879   0.04439  16.63  0.000
Error       148   0.39496   0.00267
Total       150   0.48375

S = 0.05166  R-Sq = 18.35%   R-Sq(adj) = 17.25%
```

27.51 The 95% Tukey pairwise comparison intervals are in the provided JMP output. Our conclusion from the ANOVA *F* test was that not all three groups had the same mean time to ask for help. Based on the multiple comparisons, it appears there is only a significance difference in mean time for the prime group and the control group, and that the other two comparisons are not significantly different at the 0.05 significance level.



Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL
prime	control	127.9412	52.53634	0.965	254.9176
play	control	119.1046	51.80153	-6.096	244.3050
prime	play	8.8366	51.80153	-116.364	134.0370

27.52 The 99% Tukey multiple comparison intervals are provided in the JMP output. The result of the ANOVA concluded the mean word count was not the same for all three groups. Based on the multiple comparisons, it appears there is a significant difference in mean word count when using a laptop versus longhand and the laptop-intervention versus longhand. More specifically, the mean word count is significant lower (at the 0.01 level) when using longhand versus one of the two laptop methods. There is not a significant difference in word count for the two laptop methods of taking notes.



Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL
laptop	longhand	104.9424	18.37750	50.5627	159.3221
laptop-intervention	longhand	73.0401	18.29162	18.9145	127.1657
laptop	laptop-intervention	31.9023	18.00969	-21.3890	85.1937

27.53 (a) This is a comparison of two means, so it requires a two-sample *t* test. **(b)** This is a comparison of three means, so it requires ANOVA. **(c)** This is a comparison of three proportions, so it requires a chi-square test of homogeneity.

27.54 and **27.55** are Web-based exercises.