Sources of Variation

Section 1.1

1.1.1 **B.**

1.1.2 **B & C.**

1.1.3 **A.**

1.1.4 **C.**

1.1.5 **E.**

1.1.6 **B.**

1.1.7 $\textit{predicted number of uses for items} = \begin{cases} 60.34 \textit{ if rigid librarian} \\ 92.19 \textit{ if eccentric poet} \end{cases}$

1.1.8

a. The inclusion criteria are having a clinical diagnosis of mild to moderate depression without any treatment four weeks prior and during the study.

b. The purpose of randomly assigning subjects to the groups is to make groups very similar except for the one variable (swimming with dolphins or not) that the researchers impose. Volunteering for a group could introduce a confounding variable.

c. It was important that the subjects in the control group swim every day without dolphins so that this control group does everything (including swimming) that the experimental group does except that when they swim they don't do it in the presence of dolphins. Without this we wouldn't know whether just swimming causes the difference in the reduction of depression symptoms.

d. Yes, this is an experiment because the subjects were randomly assigned to the two groups.

1.1.9.

Observed variation in: d. substantial reduction in depression symptoms	**Sources of explained variation**	**Sources of unexplained variation**
Inclusion criteria • b. mild to moderate *depression* • c. no use of antidepressant drugs or psychotherapy four weeks prior to the study *Design* • e. swimming • f. staying on an island for two weeks during the study	a. swimming with dolphins or not	• g. problems in the personal lives of the subjects during the study • h. illness of subjects during the study

1.1.10 Color of a sign is the explanatory variable with white, yellow, and red being the levels.

1.1.11

Observed Variation in: f. whether the student obeyed the sign	**Sources of explained variation**	**Sources of unexplained variation**
Inclusion criteria • c. time of day • e. age of subject	a. color of the sign	b. whether the subject was left-handed or right-handed d. attitude of student e. age of subject

1.1.12

a. The value 6.21 represents the overall mean quiz score, 5.50 represents the group mean quiz score for people who used computer notes, and 6.92 represents the group mean score for people who used paper notes.

b. We look to see how far 6.92 and 5.50 are from one another or from the overall mean of 6.21 to determine whether the note-taking method might affect the score.

c. The number 1.76 represents the typical deviation of an observation from the expected value, in this case, from the overall mean. The number 1.61 represents the typical deviation of an observation after creating a model that takes into account whether the person is using computer or paper notes.

d. Because the standard deviation of the residuals represents the left-over variation, we can see that after including the type of notes as an explanatory variable in our model the unexplained variation has been reduced (down to 1.61 from 1.76). This tells us that knowing the type of note-taking method enables us to better predict scores.

1.1.13 Random assignment should make the two groups very similar with regard to variables like intelligence, previous knowledge, or any other variable and thus likely eliminate possible confounding variables.

1.1.14

a. This table shows us possible confounding variables but then shows that subjects in the two groups are quite similar with regard to these characteristics, thus ruling out these possible confounding variables.

b. We would want the p-values to be large, so we could say that we have little to no evidence that there is a difference in mean age, proportion of males, etc. between the two groups. We want our groups to be very similar going into the study, so a causal conclusion is possible if we find a small p-value after applying the treatment(s).

1.1.15 It is likely that 3- to 5-year-olds might have different preferences when it comes to toy or candy than 12- to 14-year-olds. The older group is probably much more likely to prefer the candy over the toy and the opposite could be true with the younger group. We would not see this difference if the results of all the ages are combined together.

Section 1.2

1.2.1 **B.**

1.2.2 **A, D.**

1.2.3 **C.**

1.2.4 **A.**

1.2.5 **C.**

1.2.6 **D.**

1.2.7 **B.**

1.2.8 Using the *effects model*, because $4.48 + 0.65 = 5.13$ (the mean of the scent group) and $4.48 - 0.65 = 3.83$ (the mean of the non-scent group), the models are equivalent.

1.2.9

a. *SSModel.*

b. *SSError.*

1.2.10

a. $R^2 = SSModel/SSTotal = 0.4651$.

b. $R^2 = 1 - SSError/SSTotal = 0.7111$.

1.2.11

a. 8.

b. $6 - 8 = -2$, $10 - 8 = 2$.

c. 74.

d. 40.

e. 34.

f. 0.5405.

1.2.12

a. The explanatory variable is the type of testing environment; it is categorical.

b. The response variable is the test score; it is quantitative.

c. The two levels are quiet environment and distracting environment.

1.2.13

a. *SSTotal* would probably be larger with these 10 subjects because with the wide variety of ages there would probably be more variability in the test scores.

b. *SSModel* would probably be the same because it would still represent the difference between testing environments.

c. *SSError* would probably be larger because there would probably be more variability in the test scores within each group due to the variability in ages.

1.2.14 The variance of the scores in the distracting environment is 2.5 and the variance of the scores in the distracting environment is 6. The square root of the average of these two variances is $\sqrt{4.25} = 2.06$. The *SSError* is 34, so the standard error of the residuals is $\sqrt{34/8} = 2.06$.

1.2.15

a. The explanatory variably is whether the name of the hurricane is male or female and the response is the perceived risk level.

b. The effect of naming the hurricane Christina is $5.01 - 5.29 = -0.28$ and the effect of naming the hurricane Christopher is $5.57 - 5.29 = 0.28$. The *SSModel* is $142(0.28^2) = 11.1328$.

c. $R^2 = 11.1328/199.62 = 0.0558$. We can interpret this by saying that 5.58% of the variation in the perceived level of risk is explained by whether the name of the hurricane is male or female.

d. $SSError = 199.62 - 11.13 = 188.49$.

e. $\sqrt{188.4872/140} = 1.16$.

f. *predicted hurricane risk rating* $= 5.29 + \begin{cases} 0.28 \text{ if male name} \\ -0.28 \text{ if female name} \end{cases}$,
SE of residuals $= 1.16$.

1.2.16

a. The explanatory variable is the note-taking method and the response variable is the quiz score.

b. The effect of taking notes on paper is 0.71 and the effect of taking notes on the computer is −0.71.

c. $SSModel = 40 \times (0.71^2) = 20.164$.

d. $R^2 = 20.164/120.92 = 0.16675$. We can interpret it by saying that 16.675% of the variation of quiz score is explained by the note-taking method.

e. 120.92 – 20.164 = 100.756.

f. $\sqrt{100.756/38} = 1.628$.

g. *predicted quiz score* $= 6.21 + \begin{cases} 0.71 \text{ if using paper notes} \\ -0.71 \text{ if using computer notes} \end{cases}$.

1.2.17

a. Because the sample sizes of each group are the same, the sample size of each group is just half of the total sample size.

b.
$$\left(\frac{\sum_{all\ obs}(x_i-\bar{x})^2}{\frac{n}{2}-1}+\frac{\sum_{all\ obs}(y_i-\bar{y})^2}{\frac{n}{2}-1}\right)\frac{1}{2}$$
$$=\left(\frac{\sum_{all\ obs}(x_i-\bar{x})^2+\sum_{all\ obs}(y_i-\bar{y})^2}{\frac{n}{2}-1}\right)\frac{1}{2}$$
$$=\left(\frac{\sum_{all\ obs}(x_i-\bar{x})^2+\sum_{all\ obs}(y_i-\bar{y})^2}{n-2}\right)$$

Taking the square root we get $\sqrt{\frac{\sum_{all\ obs}(x_i-\bar{x})^2+\sum_{all\ obs}(y_i-\bar{y})^2}{n-2}}$

Use sum from 1 to n: $\frac{1}{2}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\frac{n}{2}-1}+\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\frac{n}{2}-1}\right)$

$$=\frac{1}{2}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\bar{y})^2}{\frac{n}{2}-1}\right)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-2}$$

Taking the square root, we get $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-2}}$.

Section 1.3

1.3.1 **D.**

1.3.2 **A.**

1.3.3 **D.**

1.3.4 **A.**

1.3.5 **A.**

1.3.6 The validity conditions are not met because the male sample size is small and the distribution of the number of flip-flops owned by the males is quite skewed to the right.

1.3.7

a. $\sqrt{(24.38^2 + 36.99^2)/2} = 31.33$.

b. $t = \frac{92.16 - 60.34}{31.33\sqrt{1/32 + 1/32}} = 4.06$.

c. Yes, there is strong evidence that average creativity is different between "rigid librarians" and "eccentric poets" because the t-statistic is larger than 2.

1.3.8

a. $\sqrt{(24.24^2 + 38.78^2)/2} = 32.34$.

b. $t = \dfrac{69.97 - 85.71}{32.34\sqrt{1/24 + 1/24}} = -1.69$.

c. There is not strong evidence that the average creativity measure is different between biology and theater majors because the absolute value of the t-statistic is larger than 2.

1.3.9 Yes, there is strong evidence that the long-run average game duration differs between replacement and regular referees because the difference in mean game length is 8.03 minutes and that value is way out in the right tail of the null distribution.

1.3.10

a. $t = \dfrac{196.50 - 188.47}{14.47\sqrt{1/43 + 1/48}} = 2.64$.

b. Yes, there is strong evidence that the long-run average game duration differs between replacement and regular referees because the t-statistic is larger than 2.

1.3.11

a. We would need 10 cards.

b. We would write the 10 scores on the cards.

c. After the cards are shuffled, randomly sort them in two piles of 5, labeling one pile D and the other pile Q. Calculate the mean of the numbers on the cards in each pile and find and record the difference in means (e.g., D − Q). Repeat this process many, many times to construct a null distribution of the difference in means.

1.3.12

a. Christopher mean $\bar{x}_{\text{Christopher}} = 5.57$, Christina mean $\bar{x}_{\text{Christina}} = 5.01$, so Christopher tends to be perceived as the riskier name.

b. *predicted hurricane risk*

$$= 5.29 + \begin{cases} -0.28 \textit{ if Christina} \\ 0.28 \textit{ if Christopher} \end{cases}, \textit{SE of residuals} = 1.16.$$

c. Let $\mu_{\text{Christopher}}$ be the population average risk rating for hurricanes given the name Christopher, and similarly for $\mu_{\text{Christina}}$. The hypotheses are H_0: $\mu_{\text{Christopher}} - \mu_{\text{Christina}} = 0$, that is, mean perceived risk ratings are the same regardless of whether the hurricane is named Christopher or Christina name versus H_A: $\mu_{\text{Christopher}} - \mu_{\text{Christina}} \neq 0$, that is, mean perceived risk ratings differ based on whether the hurricane is named Christopher or Christina.

d. The applet shows $t = 2.87$. Because the t-statistic is greater than 2, it looks like the difference in observed mean perceived risk ratings is statistically significant.

e. The t-statistic is far out in the right tail of the simulated null distribution.

f. simulation p-value ≈ 0.006; theory p-value $= 0.0048$.

g. We have very strong evidence that the perceived hurricane threat for the name Christopher is different (more specifically, larger) than the perceived hurricane threat for the name Christina.

1.3.13

a. We are 95% confident that the mean perceived threat rating for the name Christopher is between 0.1747 and 0.9450 points higher than that for the name Christina, in the long run.

b. Yes, because the entire interval (for Christopher minus Christina) is positive it shows the observed mean rating for Christopher is statistically significantly larger than that for Christina.

1.3.14

a. The paper method mean is 6.92 points and the computer method mean is 5.50 points, so the paper method tends to give a higher score.

b. *predicted quiz score* $= 6.21 + \begin{cases} -0.71 \textit{ if computer} \\ 0.71 \textit{ if paper} \end{cases}$,

SE of residuals $= 1.63$.

c. Let μ_{computer} be the population quiz score when notes are taken using a computer, and similarly for μ_{paper}. The hypotheses are H_0: $\mu_{\text{computer}} - \mu_{\text{paper}} = 0$, that is, the long-run mean scores will be the same for both methods of note taking vs. H_a: $\mu_{\text{computer}} - \mu_{\text{paper}} \neq 0$, that is, the mean scores will not be the same for the two methods of note taking.

d. $t = 2.27$. Because this t-statistic is greater than 2, it appears there is a statistically significant difference in the mean quiz scores between the two studying methods.

e. The t-statistic is far in the right tail of the null distribution.

f. Simulation-based p-value ≈ 0.006; theory-based p-value $= 0.0086$.

g. We have very strong evidence that there is a difference in the mean scores on this quiz between taking notes on computer and paper, with the paper method having a higher mean score in the long run.

1.3.15

a. We are 95% confident that the mean score for the paper note-taking method is between 0.3832 to 2.4668 points higher than the computer note-taking method in the long run.

b. Yes. Because the interval is completely positive we have evidence that in the long run the paper-based method population mean is larger than the computer-based method population mean.

1.3.16

a. Let μ_{MusicYes} be the population memory score when people are listening to music and similarly for μ_{MusicNo}. The hypotheses are H_0: $\mu_{\text{MusicYes}} - \mu_{\text{MusicNo}} = 0$, that is, mean memory scores will be the same regardless of whether or not people are listening to music versus H_A: $\mu_{\text{MusicYes}} - \mu_{\text{MusicNo}} < 0$, that is, mean memory scores will be the lower for people who are listening to music compared to those who aren't.

b. There is a lot of overlap between the distribution of the scores between the two groups. It looks like the difference in sample means might not be significant.

c. $t = -1.28$. With $|t| < 2$, there does not appear to be a statistically significant difference in the mean scores between the two groups.

d. The t-statistic is not in the tail of the distribution.

e. Simulation-based p-value ≈ 0.111; Theory-based p-value $= 0.1046$.

f. We do not have strong evidence that listening to music tends to hinder people's abilities to memorize words.

1.3.17

a. Whereas t-statistics and differences in means can be positive or negative, the values of R^2 are never negative. The larger the value of R^2, the bigger the difference between the two samples. Therefore, when we want to find R^2 values that are as extreme as our observed, we always look at those that are equal to or larger than the observed R^2.

b. Using R^2 as the statistic automatically does a two-sided test even though we are looking just in one direction. Therefore, the p-value is about twice as large as it should be for testing whether music *tends* to hinder people's ability to memorize, and we should divide it by 2.

1.3.18

a. Let $\mu_{neutral}$ be the population average amount of chili sauce used by those who play the neutral video game and similarly for $\mu_{violent}$. The hypotheses are H_0: $\mu_{neutral} - \mu_{violent} = 0$, that is, in the long run the average amount of chili sauce used will be the same regardless of which video game is played vs. H_a: $\mu_{neutral} - \mu_{violent} < 0$, those who play the neutral video game will select less chili on average than those who play the violent video game.

b. Yes, the violent condition has some very large chili sauce amounts compared to the neutral condition and their mean is 16.12 vs. a mean of 9.06 for the neutral group.

c. $t = -2.96$. Because $|t| > 2$ there appears to be a significant difference in the amount of chili sauce used by the two groups.

d. The observed t-statistic is far out in left the tail.

e. Simulation-based p-value ≈ 0.004; theory-based p-value $= 0.0019$.

f. We have very strong evidence that people tend to put more chili sauce into the recipe (and thus be more aggressive) after they play a violent video game than when they play a non-violent one.

1.3.19

a. The SD should be around 0.37 which is a bit larger than 0.32.

b. **i.** $\bar{x}_{noscent} = 4.52; \bar{x}_{scent} - \bar{x}_{noscent} = 0.04$.

ii. $\bar{x}_{noscent} = 3.96; \bar{x}_{scent} - \bar{x}_{noscent} = 1.04$.

iii. If the mean of the scent group is unusually large, the mean of the no scent group should be unusually small and the difference in means should be unusually large.

c. If we are forcing some of the simulated differences in means to be unusually large (either positive or negative), we are making the variability of the null distribution (or the SD of the null) a bit larger than in should be compared to what we should get when we are sampling from independent populations.

d. The SD should be around 0.31 which is very close to 0.32.

e. **i.** Through shuffling, you should get two groups that are typically quite similar and hence should have similar means, on average. The difference in these two similar means should then be zero, on average. Therefore, this type of null distribution should be centered on zero.

ii. If we are sampling from two independent populations, we should get two means that are typically close to the two population means. Because our sample means are being used as the estimates for the population means, on average, we should get our two sample means back when we resample. The difference in these should be the difference in our two sample means, on average, or 1.292.

1.3.20

a. Only one combination would produce a result as extreme as −83.77, placing the nine largest times in one group and the nine smallest times in the other group.

b. $C(18,9) = 18!/(9!)^2 = 48{,}620$.

c. $1/48{,}620 \approx 0.0000206$.

d. The simulation-based, theory-based, and exact p-values are all quite similar as the p-values are all extremely small.

Section 1.4

1.4.1 **C.**

1.4.2 **E.**

1.4.3 **B.**

1.4.4 **D.**

1.4.5 **A.**

1.4.6 **D.**

1.4.7

a. C.

b. A.

d. B.

e. A.

1.4.8 **B.**

1.4.9 **A.**

1.4.10 **B.**

1.4.11 **B.**

1.4.12 **C.**

1.4.13

a. The F-statistic will increase and the p-value will decrease.

b. The F-statistic will decrease and the p-value will increase.

1.4.14

a. 4.

b. 93.

c. 0.018.

d. 0.536.

1.4.15 The F-statistic is much larger than 4, so there is strong evidence that the groups are significantly different.

Source of Variation	DF	Sum of Squares	Mean Squares	F
Model	2	35.05	**17.53**	**10.01**
Error	54	94.53	**1.75**	
Total	56	129.58	**19.28**	

1.4.16

a. There were 3 groups.

b. The total sample size was 81.

Source	DF	Sums of squares	Mean squares	F
Model	2	227.63	**113.81**	**7.08**
Error	78	1,253.26	**16.07**	
Total	80	**1,480.89**	**129.88**	

1.4.17

a. The response variable is the amount of money spent on meals and the explanatory variable is the type of music playing. The experimental units are the customers eating at the restaurant during the study.

b. To compute the effects, we compare the group means to the LS mean: (21.69 + 21.91 + 24.13)/3 = 22.576. The effect for no music is –£0.886, for pop music is –£0.666, and for classical music is £1.554. These numbers tell us how much each group mean is above or below the overall mean.

c. *predicted amount of money spent*

$$= £22.58 + \begin{cases} -£0.886 & \textit{if no music} \\ -£0.666 & \textit{if pop music} \\ £1.554 & \textit{if classical music} \end{cases}.$$

1.4.18

a. To compute the sum of squares for the model, we compare the group means to the overall mean: $SSModel = 131(21.69 - 22.52)^2 + 142(21.91 - 22.52)^2 + 120(24.13 - 22.52)^2 = 1454.14$ (computer 451.95); this is a measure of variability between the groups.

b. $SSTotal = SSModel + SSError = 454.14 + 3167.62 = 3{,}621.74$ (computer: 3619.57).

c. $R^2 = 454.14/3{,}621.74 = 0.125$; 12.5% of the variation in spending can be attributed to the type of music playing.

d. $F = (454.14/2)/(3{,}167.6/390) = 28.0$ (computer 27.82); This is the ratio of variation between the groups and the variation within the groups. Because the F-statistic is much larger than 4 these results are significantly significant.

1.4.19

a. Let $\mu_{n/p/c}$ represent the population average amount spend by diners at this restaurant when listening to no, popular, or classical music, respectively. The hypotheses are H_0: $\mu_n = \mu_p = \mu_c$ versus H_a: At least one μ differs from the others.

b. The validity conditions are met because the groups are independent, the sample distributions are fairly symmetric, the sample sizes are each very large, and the SDs are all close to each other, easily within a factor of 2.

c. $F = 27.822$.

d. Both simulation-based and theory-based p-values are about 0.

e. We have strong evidence that at least one population mean amount differs from the others or that the type of music played has an effect on the amount of money spent at the restaurant.

f. We can make a cause-and-effect conclusion because this was an experiment. We can probably generalize to restaurants like the one that was used with customers like those involved in the experiment. It would be difficult to generalize much beyond that.

1.4.20

a. The response variable is the number of uses generated for the items. The explanatory variable is whether they imagined themselves as rigid librarians, eccentric poets, or neither. The experimental units are the 96 subjects involved in the experiment.

b. The effect is −16.45 for the rigid librarians, 15.37 for the eccentric poets, and 1.09 for the control group. These numbers tell us how much each group mean is above or below the overall mean of 76.79.

c. $\textit{predicted number of uses} = 76.79 + \begin{cases} -16.45 & \textit{if rigid librarian} \\ 15.37 & \textit{if eccentric poet} \\ 1.09 & \textit{if control} \end{cases}$.

1.4.21

a. $SSModel = 32(16.45^2) + 32(15.37^2) + 32(1.09^2) = 16256.88$; this is a measure of the variability between the groups.

b. $SSTotal = SSModel + SSError = 109{,}240.9$.

c. $R^2 = SSModel/SSTotal = 0.149$. This tells us that 15% of the variation in the number of uses for the items can be explained by what the subject imagines themselves as.

d. $F = (16{,}256.88/2) / (92{,}984.01/93) = 8.13$. This is the ratio of variation between the groups and the variation within the groups. Because this F-statistic is much larger than 4, we have very strong evidence that at least one of the population mean number of uses for these items is different from the others. (More specifically, the average number of uses for those who imagine themselves eccentric poets is larger than the averages for the librarian and control groups.).

1.4.22

a. From the graphs in the applets, the means of the groups appear roughly the same and there is a lot of overlap between the four groups so there does not appear to be strong evidence that at least one group mean differs from the rest.

b. $F = 0.536$, $R^2 = 0.018$. Although R^2 is very small, it is not a standardized statistic so we can best see that there are not significant results based on the F-statistic which is much smaller than 4.

c. The simulated p-value using either the F-statistic or R^2 is about 0.66. This confirms that there is not much evidence of a difference in the population means.

d. A large p-value is not strong evidence for the null so it is not strong evidence that all the means are the same. It just means we do not have strong evidence that there is at least one mean that is different.

1.4.23

a.

i. $\bar{x}_A = 3.77\%$ ($SD_A = 0.83$).

ii. $\bar{x}_B = 4.08\%$ ($SD_B = 0.52$).

iii. $\bar{x}_C = 5.10\%$ ($SD_C = 0.87$).

iv. $\bar{x}_D = 5.65\%$ ($SD_D = 0.45$).

v. $\bar{x}_E = 5.95\%$ ($SD_E = 1.94$).

b. Group E (>2 times per week) contained the high omega-3 value.

c. The larger mean for group E increases the variability between the groups (thus increasing F). The larger SD of group E will increase the variability within the groups (thus decreasing F). Because the addition of this value will both increase and decrease the F-statistic, it might be hard to determine which will have a greater effect. The new F is 4.467, which is less than the one from Example 1.4, so the increased SD had the greater effect.

d. The new p-value should be about 0.006 and should be a little bit larger than the one from Example 1.4.

e. No, it is not valid to perform a theory-based test because the standard deviations of the different groups are not all within a factor of 2 of each other. In particular, $SD_E/SD_D \approx 4.31$.

f. The theory-based p-value is 0.0081. It is similar to the simulation-based p-value.

g. The high omega-3 value did not make a difference in the conclusions.

1.4.24

a. $R^2 = SSModel/SSTotal$, $1 - R^2 = 1 - (SSModel/SSTotal) = (SSTotal - SSModel)/SSTotal$, so $\left[\frac{R^2}{1-R^2}\right] \times \left[\frac{n-k}{k-1}\right] = (SSModel/SSTotal)/[1 - (SSTotal - SSModel)/SSTotal] \times \left[\frac{n-k}{k-1}\right]$.

b. $\left[\frac{SSModel/SSTotal}{(SSTotal - SSModel)/SSTotal}\right] \times \left[\frac{n-k}{k-1}\right] =$

$\left[\frac{SSModel}{(SSTotal - SSModel)}\right] \times \left[\frac{n-k}{k-1}\right] = \left[\frac{SSModel}{SSError}\right] \times \left[\frac{n-k}{k-1}\right]$

$= \left[\frac{SSModel}{k-1}\right] \times \left[\frac{n-k}{SSError}\right] = \left[\frac{SSModel/(k-1)}{SSError/(n-k}\right]$.

1.4.25 For these data, *MSModel* is 40/1 = 40 and *MSError* = 34/8 = 4.25, so the *F*-statistic = 40/4.25 = 9.41. The *t*-statistic = $\frac{10-6}{\sqrt{4.25}\times\sqrt{\frac{1}{5}+\frac{1}{5}}}$ = 3.068 and $3.068^2 = 9.41$.

1.4.26

a. If we can show that $MSModel = \frac{(\bar{x}_1-\bar{x}_2)^2}{\frac{1}{n_1}+\frac{1}{n_2}}$ then,

$$t^2 = \left(\frac{(\bar{x}_1-\bar{x}_2)}{s_p\left(\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)}\right)^2 = \frac{(\bar{x}_1-\bar{x}_2)^2}{s_p{}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} =$$

$$\frac{(\bar{x}_1-\bar{x}_2)^2}{MSError\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \frac{MSModel}{MSError} = F.$$

b. $MSModel = n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2$

$$= n_1\left(\bar{x}_1 - \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}\right)^2 + n_2\left(\bar{x}_2 - \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}\right)^2$$

$$= n_1\left(\bar{x}_1\frac{(n_1+n_2)}{(n_1+n_2)} - \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}\right)^2 +$$

$$n_2\left(\bar{x}_2\frac{(n_1+n_2)}{(n_1+n_2)} - \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}\right)^2$$

$$= n_1\left(\frac{n_1\bar{x}_1+n_2\bar{x}_1-n_1\bar{x}_1-n_2\bar{x}_2}{n_1+n_2}\right)^2 +$$

$$n_2\left(\frac{n_1\bar{x}_2+n_2\bar{x}_2-n_1\bar{x}_1-n_2\bar{x}_2}{n_1+n_2}\right)^2$$

$$= n_1\left(\frac{n_2\bar{x}_1-n_2\bar{x}_2}{n_1+n_2}\right)^2 + n_2\left(\frac{n_1\bar{x}_2-n_1\bar{x}_1}{n_1+n_2}\right)^2$$

$$= n_1n_2^2\left(\frac{\bar{x}_1-\bar{x}_2}{n_1+n_2}\right)^2 + n_2n_1^2\left(\frac{\bar{x}_2-\bar{x}_1}{n_1+n_2}\right)^2$$

$$= \left(\frac{\bar{x}_1-\bar{x}_2}{n_1+n_2}\right)^2(n_1n_2^2+n_1^2n_2) = \left(\frac{\bar{x}_1-\bar{x}_2}{n_1+n_2}\right)^2(n_1n_2)(n_1+n_2)$$

$$= \frac{(\bar{x}_1-\bar{x}_2)^2}{(n_1+n_2)^2}(n_1n_2)(n_1+n_2) = \frac{(\bar{x}_1-\bar{x}_2)^2}{(n_1+n_2)}(n_1n_2) = \frac{(\bar{x}_1-\bar{x}_2)^2}{\left(\frac{n_1+n_2}{n_1n_2}\right)}.$$

Section 1.5

1.5.1 **D.**

1.5.2 **C.**

1.5.3 **C.**

1.5.4 **C.**

1.5.5 **A, F, H.**

1.5.6 The margin of error is based on a prediction interval. The rangers are not trying to predict the mean time for all future eruptions but are trying to predict the time of the next eruption so that visitors have a high probability of seeing the eruption if they are present during the entire interval.

1.5.7

a. Mean = 7.321 hrs and SD = 1.490 hrs.

b. An approximate prediction interval is $(7.321 \pm 2\times1.49\times\sqrt{1+1/100}) \approx 4.326$ to 10.316 hr; the validity conditions are met because the data are quite symmetric and have no obvious outliers.

c. Ninety-three percent of these data lie within the 95% prediction interval. This is reasonably close to the 95% that we would expect.

1.5.8

a. A 95% confidence interval for the population average score is $6.3 \pm 2\times(12.45/\sqrt{27}) \approx 6.3 \pm 4.79 = (1.51, 11.09)$; the validity conditions are met because the sample size is fairly large.

b. Yes, because the interval is completely positive, there is strong evidence that, on average, people tend to pick a face that is more attractive than their own when they are asked to identify their own face.

1.5.9

a. A 95% prediction interval is $6.3 \pm 2\times(12.45)\times\sqrt{1+1/27} \approx 6.3 \pm 25.36 = (-19.06, 31.66)$; the validity conditions are met because we were told the distribution of the results was fairly symmetric.

b. The prediction interval is trying to capture 95% of the individual results in the long run while the confidence interval is trying to capture the average result in the long run.

1.5.10

a. The applet reports a p-value of 0.0000, so there is strong evidence at least one type of background music results in a different long-run mean amount of money spent; the validity conditions are met because the sample sizes are fairly large, and all four groups have similar SD values.

b. The 95% confidence intervals are Classical–Pop: (£1.52, £2.91), Classical–None: (£1.73, £3.14), and Pop–None: (–£0.4590, £0.8985).

c. We can be 95% confident that, on average, customers will spend between £1.52 and £2.91 more per evening meal when classical music is playing than when pop music is playing at the restaurant.

d. The mean meal cost when classical music is playing is significantly greater than when either pop or no music is playing.

e. Letters plot:

Music	Group Mean	Letters
Classical	£24.13	a
Pop	£21.91	b
None	£21.69	b

1.5.11

a. A 95% confidence interval for the long-run mean cost of a meal when no music is playing is $£21.69 \pm 2 \times £3.38/\sqrt{131} \approx$ (£21.10, £22.28); the validity conditions are met because the sample size is fairly large.

b. A 95% prediction interval for the long run cost of a meal when no music is playing is $£21.69 \pm 2 \times £3.38 \times \sqrt{1+1/131} \approx$ (£14.90, £28.48); the validity conditions are met because the histogram of these data is fairly symmetric and bell-shaped.

c. The prediction interval looks like it contains about 95% of the data (it actually contains 92%), whereas the confidence interval contains a much smaller percentage of the actual data.

1.5.12

a. The p-value = 0.0087, so there is strong evidence that at least one mean is different from the others; the validity conditions are met because the sample sizes are fairly large, and the sample SDs are similar in value.

b. The 95% confidence intervals are Lie − Truth: (−2.68, −0.61), Lie − Control: (−1.97, 0.10), and Truth − Control: (−0.3267, 1.7460).

c. We can be 95% confident that, in the long run, the mean difference in rating for the lie condition is between 0.61 to 2.68 points lower than that for the truth condition.

d. The lie condition has a mean that is significantly less than the truth condition. Nothing else is significantly different.

e. Letters plot:

Condition	Group mean	Letters
Lie	−0.90	A
Control	0.03	AB
Truth	0.74	B

1.5.13

a. A 95% confidence interval for the long run difference in mean ratings between bottled and tap water is $0.03 \pm 2 \times 1.975/\sqrt{31} \approx 0.03 \pm 0.71 = (-0.679, 0.739)$; the validity conditions are met because the sample size is fairly large.

b. A 95% prediction interval for the difference in bottled and tap water ratings is $0.03 \pm 2 \times 1.975 \times \sqrt{1 + 1/31} \approx 0.03 \pm 4.01 = (-3.98, 4.04)$; the validity conditions are met because the dotolot of these data is fairly symmetric and bell-shaped.

c. The prediction interval looks like it contains about 95% of the data (it actually contains 30/31 = 96.8%), whereas the confidence interval contains a much smaller percentage.

1.5.14

a. $\textit{predicted quiz score} = \begin{cases} 5.50 \textit{ if computer} \\ 6.92 \textit{ if paper} \end{cases}$, $\textit{SE of residuals} = 1.63$.

b. A 95% confidence interval for the long-run mean score using paper notes is $6.92 \pm 2 \times 1.07/\sqrt{20} \approx 6.92 \pm 0.384 = (6.44 \text{ to } 7.40)$.

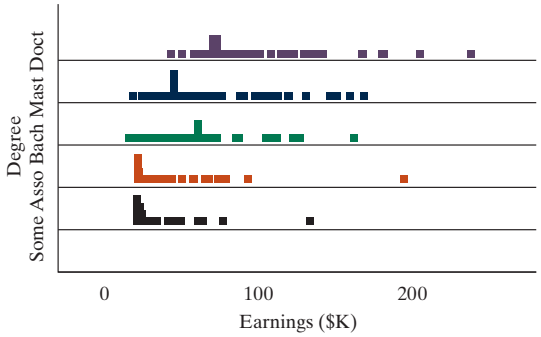
c. $\textit{predicted quiz score} = 6.213 + \begin{cases} -0.713 \textit{ if computer} \\ 0.713 \textit{ if paper} \end{cases}$,

$\textit{SE of residuals} = 1.63$.

d. A 95% confidence interval for the long-run mean effect when using paper notes is $0.71 \pm 2 \times 1.07/\sqrt{20} = 0.71 \pm 0.384 = (0.23, 1.19)$. We can use the same standard deviation because the distribution of effects is the same as the distribution of scores but slid down 6.21 units.

1.5.15

a.



Education level	Group mean	Group SD
Doctorate	$97.40K	$40.50K
Master's	$66.00K	$38.40K
Bachelor's	$55.20K	$32.20K
Associate	$36.80K	$28.50K
Some College	$32.51K	$20.80K

b. The F-statistic is 31.534 and the p-value is < 0.0001, so there is strong evidence of an association between the levels of education and earnings; the validity conditions are met because the sample size is fairly large and the SDs are all within a factor of 2 of each other.

c. Letters plot:

Education level	Estimated group mean	Letters
Doctorate	$97.40K	A
Master's	$66.00K	B
Bachelor's	$55.20K	B
Associate	$36.80K	C
Some College	$32.51K	C

1.5.16

a. A 95% confidence interval for the mean amount earned by those with doctorates is $\$97.4K \pm 2 \times \$40.5K/\sqrt{50} \approx \$97.4K \pm \$11.455K = (\$85.94K, \$108.86K)$; the validity conditions are met because the sample size is fairly large.

b. A 95% prediction interval for the amount earned by those with doctorates is $\$97.4K \pm 2 \times \$40.5K \times \sqrt{1 + 1/50} \approx \$97.4K \pm \$81.81K = (\$15.59K, \$179.21K)$; the validity conditions may not be met in this case because the distribution appears to be skewed to the right with a few large outliers.

c. There are 46/50 or 92% of those with doctoral degrees in this sample contained in the prediction interval.

1.5.17

a. A 95% prediction interval for the mean amount earned by those with associate degrees is $\$36.8K \pm 2 \times \$28.5K \times \sqrt{1 + 1/50} \approx \$36.8K \pm \$57.57K = (-\$20.77K, \$94.37K)$; the validity conditions are suspect because the distribution looks skewed right.

b. There are 49/50 or 98% of these data within the prediction interval.

c. This is such a bad fit because the distribution of salaries is highly skewed to the right. This method is only valid when we have a bell-shaped distribution.

d. The concerns aren't as great for a confidence interval. Even though the distribution is skewed, the associated sampling distribution should be quite symmetric with a sample size as large as 50.

1.5.18

a. Each margin of error is $2s/\sqrt{n}$, so $\bar{y}_2 - \bar{y}_1 = 2 \times 2s/\sqrt{n} = 4s/\sqrt{n}$.

b. The margin of error is $2s\sqrt{1/n + 1/n} = 2s\sqrt{2/n} = 2\sqrt{2}\,s/\sqrt{n}$.

c. With $4 > 2\sqrt{2}$, the answer to part (a) is larger than part (b). Both of the answers represent $\bar{y}_2 - \bar{y}_1$ but only the confidence interval for the difference in means uses the correctly pooled SE in the margin of error expression. If the individual means were just a tiny bit closer together the single mean intervals from part (a) would overlap, however the difference in means interval from part (b) would be completely positive.

Section 1.6

1.6.1 **A, C, D.**

1.6.2 **A.**

1.6.3 **C.**

1.6.4

a Increase alpha level.

b. Increase sample size.

c. Decrease number of groups comparing.

d. Decrease variability within each group.

1.6.5

a. Just under 20.

b. Just under 40.

c. Just under 50.

d. Just under 60.

e. Just over 75.

f. As sample size per group increases, power of the test increases linearly at first, but then plateaus. (The relationship looks logarithmic, or a power between 0 and 1.).

1.6.6

a. This will decrease the power of the test.

b. Power = 0.73.

c. The power would be very close to 1.

d. The relationship between power and sample size looks linear for most values of the sample size, with no plateauing visible even with each sample being even as large as 120.

1.6.7

a. $1 \leq SD \leq 3$.

b. About 5.

c. $4 \leq SD \leq 4.5$.

d. A little more than 5.5.

e. As SD increases, power decreases. For very small SDs the power of the test is one, then it begins to decrease somewhat linearly as the SDs increase.

1.6.8

a. 81%.

b. 59.1%.

c. $0.10 < \alpha < 0.15$.

d. As the level of significance increases, the power of the test increases. Power doesn't increase linearly, but in steps.

1.6.9 A difference between 15 and 20 mL/d.

1.6.10

a. The differences in the group means and overall mean are the same in Scenarios 1 and 2.

b. There is greater variability within the groups in Scenario 1.

c. Scenario 2 will have the larger *F*-statistic.

d. Scenario 2 will be more likely to have a statistically significant result.

e. As the variability within the groups decreases, the *F*-statistic increases, as does the power of the test.

1.6.11

a. The rejection region is any difference in means of 5.9 or greater.

b. A difference in mean heart rates of 7 bpm will be in the rejection region so you would conclude that the two treatment means are significantly different from each other.

c. A difference of 4 pbm is not in the rejection region, so you would conclude it is plausible that the two treatment means do not differ from each other.

d. P(Type I error) = 0.05.

e. ≈ 0.395.

f. ≈ 0.870.

g. As the effect size (difference in mean heart rates) increases, power of the test increases.

Difference in mean heart rates (bpm)	5	10	15	20	25
Power	0.395	0.870	0.995	1.000	1.000

1.6.12

a. Now the rejection region should be a difference in means of about 8.2 or more. In this case, the power is roughly 0.189.

b. Using a significance level of 0.10 would increase the power.

c. Now the rejection region is about ≈ 0.503.

d. As level of significance increases, the power of the test increases.

Level of significance	0.001	0.01	0.05	0.10	0.15
Power	0.028	0.189	0.431	0.503	0.611

1.6.13

a. About 0.238.

b. Increase power.

c. About 0.510.

d. As number per group increases, power of the test increases.

Sample size per group	5	10	15	20	25
Power	0.238	0.395	0.510	0.644	0.672

1.6.14

a. About 0.88.

b. Power will decrease.

c. About 0.253.

d. As SD within each group increases, power of the test decreases.

Standard deviation per group	4	8	12	16	20
Power	0.880	0.427	0.253	0.201	0.153

End of Chapter 1 Exercises

1.CE.1.

a. H_0: $\mu_{regular} = \mu_{filled}$ and H_a: $\mu_{regular} \neq \mu_{filled}$; p-value = 0.003. Because the p-value is less than 0.01, there is very strong evidence against the null and for the alternative that there is an association between the type of soup bowl and the amount of soup consumed with the secretly refilled soup bowl resulting in a higher average consumption of soup (oz).

b. The samples are independent of each other (randomly assigned to bowl type), sample sizes are greater than 20 and there is no strong skewness in the data.

c. Theory-based two-sided p-value is 0.0032, which is very close to the simulation-based p-value. This is expected as validity conditions were met to perform the theory-based test.

d. A 95% confidence interval for $\mu_{regular} - \mu_{filled}$ is (−10.27, −2.20) We are 95% confident that soup eaters who have secretly refilled bowls eat on average 2.2 oz to 10.27 oz more than those eating from the regular soup bowls.

e. The average of the two sample standard deviations is about 7.2 oz. So, the effect size could be reported as $6.233/7.2 \approx 0.866$. Some would consider this a meaningful effect size.

1.CE.2.

a. Yes: $8.44/6.12 = 1.38 < 2$.

b. p-value from ANOVA table is 0.0031, essentially the same as the p-value from the unpooled two-sample *t*-test.

1.CE.3.

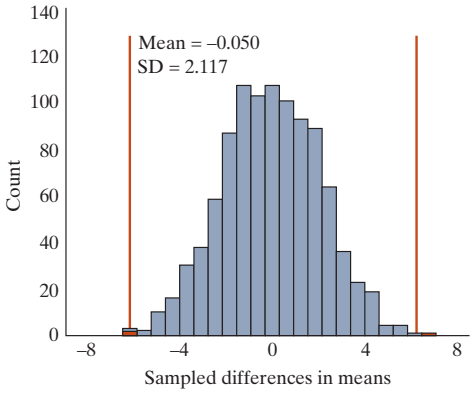
a. −4 was a plausible difference because the confidence interval from question 1.CE.1, part d contained −4. We can draw a cause and effect conclusion because this was a randomized experiment.

b. **i.** To model the refilled bowl subjects consuming four more ounces on average, we could subtract four from all of the observations. Then any differences between the groups are by chance alone. So then rerandomize the responses, any response assigned to the refilled group is given the +4.

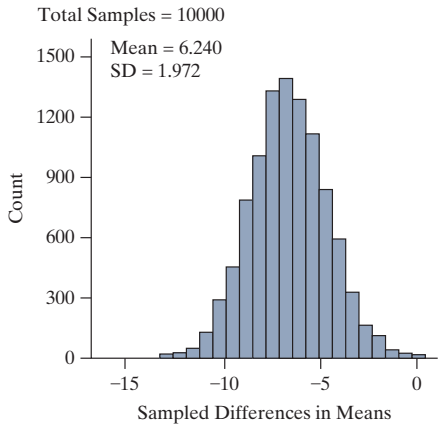
ii. The applet is counting how many of the differences are at least as far from −4 as the observed −6.261. The p-value is not small, indicating, as the confidence interval did, that −4 is a plausible value for the long-run difference in means (regular − refilled).

1.CE.4.

a. The p-value is about 0.003, the same as the shuffled p-value and the theory-based pooled *t*-test p-value. The bootstrapped null distribution has the same bell shape and is centered at 0 like the re-randomized *t*-test null distribution and the SDs are comparable (2.174 for shuffle and 2.117 for bootstrapped).



b. The new bootstrapped null is centered at −6.261, essentially what we assumed the difference in population means to be, and SD = 1.972. So there is also less variation. This is because we are focused on the variation within each group separately, rather than looking at all the data values together (with the shift from the treatment effect).



c. −6.23 + 2(1.972) equivalent to (−10.174, −2.280), which is similar to the 95% *t*-confidence interval.

1.CE.5

a. The MAD will be a larger positive number; the difference in means will be a larger number either positive or negative depending on the direction of the difference; the numerator of the *t*-statistic will be larger, so the *t*-statistic will be larger; the p-value will be smaller; the confidence interval will be the same width, but the midpoint will change shifting in the direction of the difference in means.

b. The MAD and difference in means won't change as the distance between the sample means hasn't changed, but the *t*-statistic will get larger because the sample sizes make the denominator smaller, so the p-value will get smaller and the width of the confidence interval will get smaller as larger sample sizes are less variable (smaller SD of null).

c. The MAD and difference in means won't change as the distance between the sample means hasn't changed, but the *t*-statistic will get smaller because the increase in SDs make the denominator larger, so the p-value will get larger and the width of the confidence interval will get larger as more variability in the data means more variability in the null distribution.

d. Changing the confidence level from 95% to 99% will only increase the multiplier of the SD of the null, thus the margin of error of the confidence internal will get larger. Nothing else will be affected by this change.

1.CE.6.

a. Set #2 will have the larger MAD as the means are father apart.

b. Set #2 will have the larger *F*-statistic because the variability within the groups is the same in set #1 as it is in set #2, but the variability between the groups is larger in set #2 than it is in set #1.

c. The *F*-statistic for Set #1 should be between 0 and 1 as all the group means are very close to the overall mean which would make the *MSGroups* close to zero.

1.CE.7

a. Yes, the validity conditions are met because the groups are independent as the treatments were randomly assigned, and although there are fewer than 20 observations in each group, the corn weights are fairly symmetric, and the SDs are all within a factor of 2.

b. **Step 1: Ask a research question:** Can organic methods be used to control harmful insects and limit their effect on sweet corn growth? **Step 2: Design a study and collect data:** A total of 60 plots were used in the study. In 12 plots of corn a beneficial soil nematode was introduced. In another 12 plots a parasitic wasp was used. Another 12 plots were treated with both the nematode and the wasp. In a fourth set of 12 plots a bacterium was used. Finally, a fifth set of 12 plots of corn acted as a control in which no special treatment was applied. The plots were all randomly assigned to the treatment conditions. Twenty-five ears of corn from each plot were randomly sampled and each was weighed (in ounces). H_0: All population mean weights of corn are equal. H_a: At least one population mean weight of corn is different. **Step 3: Explore the data:** Largest mean weight of corn in ounces is found in the control group and the smallest mean weight of corn in ounces is found in the wasp group. It appears that the control group and possibly the nematode group might have significantly larger mean weights of corn yield than the other treatment groups. **Step 4: Draw inferences:** With $F = 4.49$ (df = 4, 55) and p-value = 0.0033, we have strong evidence (e.g., at the 5% level of significance) that the treatment means are not all equal. This conclusion applies to all plots of sweet corn grown under the same conditions as the experimental plots in this study. **Step 5: Formulate conclusions:** Because random assignment was used, we can say that the treatment was the cause of the differences seen. We can only generalize results to sweet corn in the environments in which it was grown. **Step 6: Look back and ahead:** Answers may vary but should suggest follow-up questions or suggest what can be changed if this study were to be run again.

c. The average weight is largest (best) for control (13.2), then nematode (11.6), bacterium (11.1), nematode + wasp (10.3), and smallest (worst) for wasp (8.5).

d. The control is significantly higher than nematode + wasp and wasp. Wasp also looks significantly lower than bacterium and nematode. The control does not appear to differ from nematode or bacterium. (We can almost separate into two groups: group (1) with control, nematode, and bacterium) and group (2) nematode + wasp and wasp.) See the following letter plot for summary.

Treatment	Mean	Letters
Control	13.2083	a
Nematode	11.5822	ab
Bacterium	11.125	ab
Nematode + Wasp	10.3333	bc
Wasp	8.5	c

e. The overall mean was 10.95 which is not captured in the first and last intervals. So we can say control had a significantly larger than average weight, and wasp had significantly lower than average weights, on average.

Treatment	Mean	95% CI
Control	13.2083	(11.58, 14.84)
Nematode	11.5822	(9.95, 13.22)
Bacterium	11.1250	(9.49, 12.76)
Nematode + Wasp	10.3333	(8.70, 11.97)
Wasp	8.5000	(6.87, 10.13)

Chapter 1 Investigation

1. This was a randomized experiment and it is advantageous because causation may be concluded from this type of study.

2. The experimental units are the Parkinson's disease patients participating in the study.

3. Inclusion criteria are Parkinson's patients (in stages 1, 2, 3, or 4) that have stable medication use and the ability to stand unaided and walk without assistance.

4. The explanatory variable is the type of therapy (tai chi, resistance training, or stretching). The therapy lasted 24 weeks.

5. Functional reach is assessed as the maximal distance (in cm) a participant could reach forward beyond arm's length while standing.

6. Other sources of variation could include a person's sex, age, genetics, prior activity and fitness, and how long they have had Parkinson's.

7.

a. The sources of variation that were not allowed to change is how long the subject participated in the study, having stage 1, 2, 3, or 4 Parkinson's disease, stable medication use, and the ability to stand unaided and walk without assistance.

b. Sources of variation accounted for include the type of therapy used on each patient (tai chi, resistance training, or stretching).

c. Unexplained variation could include person's sex, age, genetics, prior or current activity levels, fitness level, duration of Parkinson's.

d. See below.

Observed variation in: Functional Reach	Sources of explained variation	Sources of unexplained variation
Inclusion criteria • Parkinson's patients (in stages 1, 2, 3, or 4) • stable medication use • ability to stand unaided and walk without assistance *Design* • 24 weeks of therapy	• Therapy type (tai chii, resistance training, stretching)	• person's sex • age • genetics • prior or current activity levels • fitness level • duration of disease

8. Overall mean = 2.697 cm, Overall SD = 5.193 cm, $SSTotal = 5231.38$. *Predicted change in functional reach* = 2.697cm, *SE of residuals* = 5.193 cm.

9. *predicted change in func.reach* =

$$\begin{cases} 4.89\ cm\ if\ tai\ chi \\ 2.34\ cm\ if\ resistance, SE\ of\ residuals = 4.943. \\ 0.86\ cm\ if\ stretching \end{cases}$$

The *SE of residuals* has decreased, making this model better at making predictions about the change in functional reach.

10. The effects are: tai chi = 2.197, resistance training = −0.357, stretching = −1.840.

11. *predicted change in func.reach* =

$$2.697 + \begin{cases} 2.193\ cm\ if\ tai\ chi \\ -0.357\ cm\ if\ resistance\ , SE\ of\ resid. = 4.94. \\ -1.837\ cm\ if\ stretching \end{cases}$$

12. *SSModel* = 542.07 and *SSError* = 4,689.31.

13. $R^2 = 542.37/5{,}231.38 = 0.104$. This means that about 10.4% of the variation in change in functional reach can be attributed to the type of exercise used.

14. This is somewhat subjective. The maximum difference in means is about 4 cm and the SE of the residuals is a bit larger than this, so this difference may not be enough to seem practically significant. However, 4 cm or even less may be the difference between being able to reach down to tie your shoes and not. That would be very practically significant.

15. H_0: There is no association between the type of exercise and change in functional reach. $\mu_{TC} = \mu_R = \mu_S$. H_a: There is an association between the type of exercise and change in functional reach. At least one μ_i is different from the rest.

16. The *F*-statistic is 11.097 (you could have used other statistics). In 1,000 shuffles, an *F* of 11.097 never occurred so the p-value is less than 0.001. Thus we have strong evidence of an association between type of exercise and change in functional reach among Parkinson's patients.

17. The *F*-statistic is 11.097 and the theory-based p-value in the applet is given as 0.0000. With such large sample sizes and SD that are fairly similar (within a factor of 2), the validity conditions are met. Thus we have strong evidence of an association between type of exercise and change in functional reach among Parkinson's patients.

18.

Source	DF	Sums of squares	Mean squares	*F*	p-value
Groups	2	542.07	271.03	11.097	0.0000
Error	192	4689.31	24.42		
Total	194	5231.38			

19. Doing an overall test, like we have done, allows us to keep the type I error rate at 5%.

20. Tai chi – Resistance (0.844, 4.2637), Tai chi – Stretching (2.33, 5.75), Resistance – Stretching (−0.2268, 3.1929); We are 95% confident that the true average increase in functional reach is between 2.33 cm to 5.75 cm larger for those that do tai chi than for those that do stretching. Similarly, we are 95% confident the true average increase in functional reach is between 0.844 cm to 4.26 cm larger for Parkinson's patients who do tai chi than for those who do resistance training.

21.

Exercise type	Sample mean	Letters
Tai Chi	4.89 cm	a
Resistance	2.34 cm	b
Stretching	0.86 cm	b

22. Tai chi has the largest effect of 4.89 cm and stretching has the least at 0.86 cm. It is reasonable to conclude that tai chi significantly increases functional reach because the standardized statistic is $4.89/(4.33/\sqrt{65}) = 9.1$. As that is much greater than 2, there is very strong evidence against the null of no effect of treatment on functional reach.

23. We can be 95% confident that doing tai chi will increase the true average functional reach of Parkinson's patients by between 3.68 cm and 6.10 cm.

We can be 95% confident that doing resistance training will increase the true average functional reach of Parkinson's patients by between 1.13 cm to 3.55 cm.

We can be 95% confident that stretching will change the true average functional reach of Parkinson's patients between a decrease of 0.35 cm up to an increase to 2.07 cm.

24. We predict that 95% of the Parkinson's patients doing tai chi will change their functional reach between a decrease of 4.93 cm to an increase of 14.72 cm.

We predict that 95% of the Parkinson's patients doing resistance training will change their functional reach between an increase of 7.48 cm to 12.16 cm.

We predict that 95% of the Parkinson's patients stretching will change their functional reach between a decrease of 8.97 cm to an increase of 10.68 cm.

25. With a p-value of about 0, there is strong evidence of an association between exercise type and change in functional reach. Even though R^2 was only 0.104 and the maximum difference in means was a bit less than the SE of the residuals, the results would probably be considered practically significant because even a small change in functional reach could be a great benefit.

26. Answers will vary. The addition of a control group would be nice in order to compare the change in functional reach of Parkinson's patients in each treatment group to Parkinson's patients with no intervention. Does their functional reach tend to stay the same, increase, or decrease on average? Follow-up studies could look at other forms of exercise or combinations of exercise. They could also look to see what sort of dose-response there might be. For example if tai chi was done more frequently each week would we get better results?