Chapter 21 – Comparing Two Means

21.1 This is a matched pairs design. Each couple is a matched pair.

21.2 This involves two independent samples.

21.3 This involves a single sample.

21.4 This involves two independent samples (because the results for the new battery are independent of the results for the prototype battery).

21.5 STATE: Does playing with a Nintendo Wii™ improve the laparoscopic abilities of medical students? PLAN: We test H_0: $\mu_{\text{Wii}} = \mu_{\text{NoWii}}$ versus H_a: $\mu_{\text{Wii}} > \mu_{\text{NoWii}}$, where μ_{Wii} is the mean improvement time to perform a virtual gall bladder operation for those who used the Wii™, and μ_{NoWii} is the mean improvement time for those who did not use the Wii™. We use a one-sided alternative because movements with the Wii™ are similar to those needed to perform the surgery, so practice with the Wii™ should result in more improvement than just performing the same operation again. SOLVE: These data came from participants in a randomized experiment, so the two groups are independent. The provided stemplots suggest some deviation from Normality and a possible high outlier for the No Wii™ group. Boxplots (not shown) indicate no outliers and a relatively symmetric distribution for the Wii™ group, but both the −88 and 229 are outliers for the No Wii™ group. We proceed with the *t* test for two samples appealing to robustness (especially good with equal sample sizes). With $\bar{x}_{\text{Wii}}$ = 132.71, $\bar{x}_{\text{NoWii}}$ = 59.67, s_{Wii} = 98.44, s_{NoWii} = 63.04, n_{Wii} = 21, and n_{NoWii} = 21, $SE = \sqrt{\frac{s_{\text{Wii}}^2}{n_{\text{Wii}}} + \frac{s_{\text{NoWii}}^2}{n_{\text{NoWii}}}}$ = 25.509 and $t = \frac{\bar{x}_{\text{Wii}} - \bar{x}_{\text{NoWii}}}{SE}$ = 2.86. Using df as the smaller of $21 - 1$ and $21 - 1$, we have df = 20, and $0.0025 < P < 0.005$. Using software, df = 34.04 and $P = 0.0036$. CONCLUDE: There is very strong evidence that playing with a Nintendo Wii™ does help improve the skills of student doctors, at least in terms of the mean time to complete a virtual gall bladder operation.

Wii		No Wii
	−0	8
111	−0	21
	0	223444
877775	0	566678899
43322	1	14
8	1	
4421	2	2
9	2	
3	3	

21.6 STATE: Does the average time lying down differ between obese people and lean people? PLAN: We test H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 \neq \mu_2$, where μ_1 is the mean time spent lying down for the lean group, and μ_2 is the mean time for the obese group. SOLVE: We assume that the data come from SRSs of the two populations. See Example 21.2 for a discussion of conditions for inference applied to this problem. The provided stemplots do not indicate non-Normal data. We proceed with the t test for two samples. With $\bar{x}_1$ = 501.6461, $\bar{x}_2$ = 491.7426, s_1 = 52.0449, s_2 = 46.5932, n_1 = 10, and n_2 = 10: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ = 22.0898 and $t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$ = 0.448. Using df as the smaller of $10 - 1$ and $10 - 1$, we have df = 9 and $P > 0.5$. Using software, df = 17.8 and P = 0.6593. CONCLUDE: There is no evidence to support a conclusion that lean people spend a different amount of time lying down (on average) than obese people.

Lean		Obese
9	3	
	4	1
	4	
5	4	44
6	4	6
8	4	
10	5	001
33	5	23
5	5	
6	5	6

21.7 From Exercise 21.5, we have $\bar{x}_{\text{Wii}}$ = 132.71, $\bar{x}_{\text{NoWii}}$ = 59.67, $n_{\text{Wii}} = n_{\text{NoWii}}$ = 21, and SE = 25.509. A 90% confidence interval for the mean difference in improvement in time to complete the virtual gall bladder operation is $(\bar{x}_{\text{Wii}} - \bar{x}_{\text{NoWii}}) \pm t^* SE$ = 73.04 ± 1.725(25.509) = 29.037 to 117.043 seconds. Software uses df = 34.04 and gives an interval of 29.904 to 116.176 seconds.

21.8 In this study, men underestimated their average life expectancy by 19.50%, whereas women underestimated their average life expectancy by 12.71%. If these samples can be viewed as SRSs, then under H_0: $\mu_1 = \mu_2$, a difference in sample means as great as the one observed (12.71% − 19.50%) is 2.177 standard errors below expected (t = −2.177), and a more extreme difference would have occurred by chance alone about 5.28% of the time under repeated sampling (P = 0.05281). There is somewhat strong evidence that men and women differ in their views on their own longevity; we would reject H_0 at the 10% level but not at the 5% level.

21.9 (a) Back-to-back stemplots of the time data are shown below. They appear to be reasonably Normal, and the discussion in the exercise justifies our treating the data as independent SRSs, so we can use the t procedures. We wish to test H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 < \mu_2$, where μ_1 is the population mean time in the restaurant with no scent, and μ_2 is the mean time in the restaurant with a lavender odor. Here, with $\bar{x}_1$ = 91.27, $\bar{x}_2$ = 105.7, s_1 = 14.93, s_2 = 13.105, and $n_1 = n_2$ = 30: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ =

3.627 and $t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = -3.98$. Using software, df = 57.041 and $P = 0.0001$. Using the more conservative df = 29 (lesser of $30 - 1$ and $30 - 1$) and Table C, $P < 0.0005$. There is very strong evidence that customers spend more time on average in the restaurant when the lavender scent is present.

```
   Time in restaurant
No scent |    | Lavender
      98 |  6 |
     322 |  7 |
     965 |  7 | 6
      44 |  8 |
    7765 |  8 | 89
   32221 |  9 | 234
      86 |  9 | 578
      31 | 10 | 1234
    9776 | 10 | 5566788999
         | 11 | 4
      85 | 11 | 6
       1 | 12 | 14
         | 12 | 69
         | 13 |
         | 13 | 7
```

(b) Back-to-back stemplots of the spending data are shown. The distributions are skewed and have many gaps. We wish to test $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 < \mu_2$, where μ_1 is the population mean amount spent in the restaurant with no scent, and μ_2 is the mean amount spent in the restaurant with the lavender odor. Here, with $\bar{x}_1$ = €17.5133, $\bar{x}_2$ = €21.1233, s_1 = €2.3588, s_2 = €2.345, and $n_1 = n_2 = 30$: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ = €0.6073 and $t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = -5.94$. Using software, df = 57.998 and $P < 0.0001$. Using the more conservative df = 29 and Table C, $P < 0.0005$. There is very strong evidence that customers spend more money on average when the lavender scent is present.

```
      Amount spent (euros)
   No scent     |    | Lavender
              9 | 12 |
                | 13 |
                | 14 |
99999999999999  | 15 |
                | 16 |
                | 17 |
  555555555555  | 18 | 55555555555
                | 19 |
              5 | 20 | 7
              9 | 21 | 5599999999
                | 22 | 3558
                | 23 |
                | 24 | 99
              5 | 25 | 59
```

21.10 (a) The provided back-to-back stemplots show the stated right-skewness of both the box-office hits and failures. Additionally, the outlier in the failure group can be seen easily.

hits		failures
	-2	750
	-1	83
466	-0	70
	0	9
	1	1
	2	2
20	3	
	4	
4	5	
	6	2
2	7	

(b) We wish to test H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 > \mu_2$, where μ_1 is the population mean brain activity for box-office hits, and μ_2 is the population mean brain activity for box-office failures. Here, with $\bar{x}_1 = 0.246$, $\bar{x}_2 = -0.068$, $s_1 = 0.313$, $s_2 = 0.167$, $n_1 = 7$, and $n_2 = 10$: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.13$ and $t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = 2.419$. Using software, df = 8.413 and $P = 0.0202$. Using the more conservative df = 6 (lesser of 7 − 1 and 10 − 1) and Table C, $0.025 < P < 0.05$. There is evidence that the population mean brain activity for box-office hits is larger than that for box-office failures.

21.11 We have two small samples ($n_1 = n_2 = 4$), so the t procedures are not reliable unless both distributions are Normal.

21.12 From Exercise 21.10, we have $\bar{x}_1 = 0.246$, $\bar{x}_2 = -0.068$, $n_1 = 7$, $n_2 = 10$, and $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.13$. If we use software, df = 8.413, so $t^* = 1.848$. If we use the more conservative df = 6, then $t^* = 1.943$. A 90% confidence interval for the mean difference in brain activity for box-office hits and box-office failures is $(\bar{x}_1 - \bar{x}_2) \pm t^* SE = 0.074$ to 0.554 (if df = 8.413) or 0.061 to 0.567 (if df = 6).

21.13 Here are the details of the calculations: $SE_{\text{Con}} = 1.10/\sqrt{37} = 0.18$, $SE_{\text{NotCon}} = 0.97/\sqrt{36} = 0.16$, $SE = \sqrt{SE_{\text{Con}}^2 + SE_{\text{NotCon}}^2} = 0.2408$, df $= \frac{(SE_{\text{Con}}^2 + SE_{\text{NotCon}}^2)^2}{\frac{1}{36}(SE_{\text{Con}}^2)^2 + \frac{1}{35}(SE_{\text{NotCon}}^2)^2} = 70.331$, and $t = \frac{5.83 - 5.27}{0.2408} = 2.309$.

21.14 Let μ_1 denote the mean for men and μ_2 denote the mean for women. According to the output, $\bar{x}_1 = -19.50$, $\bar{x}_2 = -12.71$, $s_1 = 5.612$, and $s_2 = 5.589$. With $n_1 =$

6 and $n_2 = 7$, $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{-19.50 - (-12.71)}{\sqrt{\dfrac{5.612^2}{6} + \dfrac{5.589^2}{7}}} = -2.179$ and df $= \dfrac{\left(\dfrac{5.612^2}{6} + \dfrac{5.589^2}{7}\right)^2}{\dfrac{1}{6-1}\left(\dfrac{5.612^2}{6}\right) + \dfrac{1}{7-1}\left(\dfrac{5.589^2}{7}\right)} = 10.68$, rounded to two decimal places.

21.15 Reading from the software output shown in the statement of Exercise 21.13, we find that there is a significant difference in mean appreciation ratings for gifts that are congruent with the giver and ratings for gifts that are not congruent with the giver ($t = 2.309$, df = 70.3, and $P < 0.02$). Because larger scores indicate a greater appreciation, it appears that gifts are more appreciated when they are congruent with the giver.

21.16 (c) the one-sample *t* interval. There is one sample, and only one score comes from each member of the sample.

21.17 (a) the two-sample *t* test. We have two independent populations: females and males.

21.18 (b) the matched pairs *t* test. Two measurements (one for each variety) are taken at each of the plots.

21.19 (b) confidence levels and *P*-values from the *t* procedures are quite accurate even if the population distribution is not exactly Normal.

21.20 (a) 20. Using Option 2 here, df is the lesser of (21 − 1) and (21 − 1).

21.21 (b) 3.05. $t = \dfrac{15.84 - 9.64}{\sqrt{\dfrac{8.65^2}{21} + \dfrac{3.43^2}{21}}} = 3.05$.

21.22 (c) Yes: the SRS condition is OK and large sample sizes make the Normality condition unnecessary. This is because the students were randomly assigned to one of the two groups, and the samples are large enough to overcome problems of potential non-Normality.

21.23 (a) $H_0: \mu_{70} = \mu_{30}$ versus $H_a: \mu_{70} > \mu_{30}$. We suspect that higher chance-of-winning predictions will be judged to be more accurate than lower chance-of-winning predictions. Thus, because the accuracy scale uses higher scores to indicate greater accuracy, this one-sided alternative is used.

21.24 (a) $0.001 < P < 0.005$. Using the conservative df = 79 (the lesser of 80 − 1 and 81 − 1), the *P*-value from software is 0.0011.

21.25 (a) To test the belief that women talk more than men, we use a one-sided alternative: H_0: $\mu_F = \mu_M$ versus H_a: $\mu_F > \mu_M$. **(b)** to **(d)** The small table below provides a summary of t statistics, degrees of freedom, and P-values for both studies. The two-sample t statistic is computed as $t = \frac{\bar{x}_F - \bar{x}_M}{\sqrt{\frac{s_F^2}{n_F} + \frac{s_M^2}{n_M}}}$, and we take the conservative approach for computing df as the smaller sample size minus 1.

Study	t	df	Table C values	P-value
1	−0.248	55	$\|t\| < 0.679$	$P > 0.25$
2	1.507	19	$1.328 < t < 1.729$	$0.05 < P < 0.10$

Note that, for Study 1, we reference df = 50 in Table C. **(e)** The first study gives no support to the belief that women talk more than men; the second study gives weak support, and it is significant only at a relatively high significance level (say $\alpha = 0.10$).

21.26 (a) Call group 1 the Alcohol group and group 2 the Placebo group. Then, because $SEM = s/\sqrt{n}$, we have $s = SEM\sqrt{n}$. So, $s_1 = 0.05\sqrt{25} = 0.25$ and $s_2 = 0.03\sqrt{25} = 0.15$. **(b)** Using the conservative Option 2, df = 24 (the lesser of 25 − 1 and 25 −1). **(c)** Here, with $n_1 = n_2 = 25$, $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.0583$. With df = 24, we have t^*= 1.711, and a 90% confidence interval for the mean difference in proportions is given by $(0.25 - 0.12) \pm 1.711(0.0583) = 0.13 \pm 0.10 = 0.03$ to 0.23.

21.27 (a) The standard errors are $SE_P = 2.05/\sqrt{104} = 0.201$ and $SE_N = 1.74/\sqrt{104} = 0.171$, for the positive mood and neutral mood groups, respectively. **(b)** Using conservative Option 2, df = 103 (the lesser of 104 − 1 and 104 − 1). **(c)** We test H_0: $\mu_P = \mu_N$ versus H_a: $\mu_P \neq \mu_N$, where μ_P is the mean attitude score for the positive mood group and μ_N is the mean score for the neutral mood group. The test statistics is $t = \frac{4.30 - 5.50}{\sqrt{\frac{2.05^2}{104} + \frac{1.74^2}{104}}} = -4.551$ and, with df = 103 (rounded to 100), Table C shows $P < 0.001$. There is overwhelming evidence that the mean attitude towards the indulgent food was different for those who read the happy story (positive mood) than for those who did not read the story (neutral mood).

21.28 (a) Parents who choose a Montessori school probably have different attitudes about education than other parents. **(b)** Over 55% of Montessori parents (30 out of 54) participated in the study, compared with about 22% of the other parents (25 out of 112). **(c)** We test H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 \neq \mu_2$, where μ_1 is the mean math score for Montessori children, and μ_2 is the mean math score for non-Montessori children. With $\bar{x}_1 = 19$, $\bar{x}_2$= 17, s_1= 3.11, s_2= 4.19, n_1= 30 and n_2 = 25: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ = 1.0122 and $t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = 1.976$. Using df = 24 under Option 2, $0.05 < P < 0.10$. Using software, df = 43.5 and $P = 0.0545$. There is moderate evidence of a difference in

mean math scores between these two groups, but not quite enough evidence to reach such a conclusion at the 5% significance level.

21.29 (a) We test $H_0: \mu_{1975} = \mu_{2006}$ versus $H_a: \mu_{1975} > \mu_{2006}$. $SE = \sqrt{\frac{0.81^2}{1165} + \frac{0.80^2}{2177}} = 0.02928$, so the two-sample test statistic is $t = \frac{3.37 - 3.32}{0.0293} = 1.708$. This is significant at the 5% level: $P = 0.0439$ (df = 2353.38) or $0.025 < P < 0.05$ (df = 1000). There is good evidence that mean job satisfaction decreased from 1975 to 2006. **(b)** The difference is barely significant at the 0.05 level (most likely due to the large sample sizes). Knowing that 1975 had the highest mean job satisfaction score in this time period casts doubt about whether this is actually decreasing. Also, a difference of 0.05 in the means may not be of practical importance.

21.30 (a) Using the conservative two-sample procedures, df = 22 (the lesser of 23 − 1 and 23 − 1). **(b)** Using the data summary provided in the problem description, $t = \frac{4.61 - 6.68}{\sqrt{\frac{3.08^2}{23} + \frac{3.45^2}{23}}} = -2.147$. **(c)** We test $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$, where μ_1 is the mean job preference score for the lowest floor group, and μ_2 is the mean job preference score for the highest floor group. Using Table C, with df = 22, $0.04 < P < 0.05$. There is some evidence of a difference in mean job preference scores between the lowest floor and highest floor groups.

21.31 (a) Let μ_C be the mean brain size for players who have had concussions, and let μ_{NC} be the mean for those who have not had concussions. We test $H_0: \mu_C = \mu_{NC}$ versus $H_a: \mu_C \neq \mu_{NC}$. This is a two-sided test, because we simply want to know if there is a difference in mean brain size. $SE = \sqrt{\frac{609.3^2}{25} + \frac{815.4^2}{25}} = 203.5803$, and $t = \frac{5784 - 6489}{203.5803} = -3.463$. Using the conservative version for df (Option 2), df = 24 and $0.002 < P < 0.005$. Using software, df = 44.43 and $P = 0.0012$. There is strong evidence that the mean brain size is different for football players who have had concussions as opposed to those who have not had concussions. **(b)** The fact that these were consecutive cases indicates that they are not a random sample of all football players who have or have not had concussions. That could weaken or negate the results of the test. We'd need more information about how and why these players were referred to the institute.

21.32 (a) The appropriate test is the matched pairs test, because a student's score on Try 1 is certainly dependent on his or her score on Try 2. Using the differences, we have $\bar{x} = 29$ and $s = 59$. **(b)** To test $H_0: \mu = 0$ versus $H_a: \mu > 0$, we compute $t = \frac{29 - 0}{59/\sqrt{427}} = 10.16$ with df = 426. This is certainly significant, with $P < 0.0005$. It appears that coached students improve their scores, on average. **(c)** Table C gives $t^* = 2.626$ for df = 100, while software gives $t^* = 2.587$ for df = 426. The confidence interval is $\bar{x} \pm t^*(s/\sqrt{n})$. Using the conservative value of t^*, this yields $29 \pm 2.626\frac{59}{\sqrt{427}} = 29 \pm$

7.5 = 21.5 to 36.5 points. Using software, the confidence interval is 21.61 to 36.39.

21.33 (a) The hypotheses are H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 > \mu_2$, where μ_1 is the mean gain among all coached students, and μ_2 is the mean gain among uncoached students. We find $SE = \sqrt{\frac{59^2}{427} + \frac{52^2}{2733}} = 3.0235$, and $t = \frac{29 - 21}{3.0235} = 2.646$. Using the conservative approach, df = 426 is rounded down to df = 100 in Table C, and we obtain $0.0025 < P < 0.005$. Using software, df = 534.45 and $P = 0.0042$. There is evidence that coached students had a greater average increase than uncoached students. **(b)** The 99% confidence interval is $8 \pm t^*(3.0235)$, where t^* equals 2.626 (using df = 100 with Table C) or 2.585 (using df = 534.45 with software). This gives either 0.06 to 15.94 points or 0.184 to 15.816 points, respectively. **(c)** Increasing one's score by 0 to 16 points is not likely to make a difference in being granted admission or scholarships from any colleges.

21.34 This was an observational study, not an experiment. The students (or their parents) chose whether or not to be coached; students who choose coaching might have other motivating factors that help them do better the second time. For example, perhaps students who choose coaching have some personality trait that also compels them to try harder the second time.

21.35 (a) Call this year "Year 1" and last year "Year 2." Then $\bar{x}_1 = 41$, $\bar{x}_2 = 38$, $s_1 = 11$, $s_2 = 13$, $n_1 = 50$ and $n_2 = 52$ yield $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.381$. The 95% confidence interval is $(\bar{x}_1 - \bar{x}_2) \pm t^* SE$, where t^* is 2.021 (using df = 49 rounded down to 40 with Table C) or 1.984 (using df = 98.427 with software). This gives either -1.812 to 7.812 units or -1.724 to 7.724 units, respectively. **(b)** No matter which Option is used to calculate the df, the 95% confidence interval contains 0. This means it is possible that there is not a significant difference in the mean number of units sold from this year to last year. In fact, because the intervals contain negative numbers, it is possible that the mean number of units sold last year was greater than the mean number of units sold this year.

21.36 (a) The hypotheses are H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 \neq \mu_2$. We find $SE = \sqrt{\frac{261^2}{100} + \frac{274^2}{100}} = \$37.841$, and $t = \frac{1319 - 1372}{37.841} = -1.401$. Using the conservative approach, df = 99 is rounded down to df = 80 in Table C, and we obtain $0.10 < P < 0.20$. There is not evidence of a significant difference between the mean amounts charged by customers offered the two proposed plans. **(b)** With $n_1 = n_2 = 100$, the samples are large enough to overcome problems of potential non-Normality.

21.37 (a) Stemplots for both data sets are shown. Neither sample histogram suggests a strong skew or the presence of strong outliers; *t* procedures are reasonable here.

Good Weather Tips

Stem and Leaf

Stem	Leaf	Count
27	0	1
26		
25		
24	099	3
23	4	1
22	0122378	7
21	29	2
20	3568	4
19	9	1
18	7	1

18|7 represents 18.7

Bad Weather Tips

Stem and Leaf

Stem	Leaf	Count
23	2	1
22		
21		
20	02	2
19	0124	4
18	0024588	7
17	05	2
16	18	2
15		
14	0	1
13	6	1

13|6 represents 13.6

(b) Let μ_1 be the mean tip percent when the forecast is good, and let μ_2 be the mean tip percent when the forecast is bad. We have $\bar{x}_1 = 22.22$, $\bar{x}_2 = 18.19$, $s_1 = 1.955$, $s_2 = 2.105$, $n_1 = 20$, and $n_2 = 20$. We test $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$. Here, $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.642$ and $t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = 6.274$. Using df = 19 (the conservative Option 2) and Table C, we have $P < 0.001$. Using software, df = 37.8 and $P < 0.00001$. There is overwhelming evidence that the mean tip percentage differs between the two types of forecasts presented to patrons.

21.38 (a) Based on the provided stemplots, the *t* procedures should be safe. Both the stemplots and the means suggest that customers stayed (very slightly) longer when there was no odor.

No odor		Lemon
	5	6
	6	03
98	6	
322	7	34
965	7	58
44	8	33
7765	8	8889
32221	9	0144
86	9	677
31	10	14
9776	10	5688
	11	23
85	11	
1	12	

(b) Testing $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$, where μ_1 is the mean time in the restaurant with no odor, and μ_2 is the mean time in the restaurant with lemon odor. Now, with $\bar{x}_1 = 91.2667$, $\bar{x}_2 = 89.7857$, $s_1 = 14.9296$, $s_2 = 15.4377$, $n_1 = 30$, and $n_2 = 28$, we find $SE = \sqrt{\frac{14.9296^2}{30} + \frac{15.4377^2}{28}} = 3.9927$ and $t = \frac{91.2667 - 89.7857}{3.9927} = 0.371$. This is not at all significant: $P > 0.5$ (df = 27, using Table C with Option 2 for conservative

df) or $P = 0.7121$ (df = 55.4, using software). We cannot conclude that mean time in the restaurant is different when the lemon odor is present.

21.39 Refer to results in Exercise 21.37. Using df = 19, t^*= 2.093 and the 95% confidence interval for the difference in mean tip percents between these two populations is $(22.22 - 18.19) \pm 2.093(0.642) = 4.03 \pm 1.34 = 2.69\%$ to 5.37%. Using df = 37.8 with software, $t^* = 2.025$ and the corresponding 95% confidence interval is 2.73% to 5.33%.

21.40 (a) Summary statistics describing the two samples are given in the table. The sample means suggest (surprisingly) that the Dow Jones Industrial Average performs better after a negative article.

	n	$\bar{x}$	s
Positive article	14	−33.93	188.52
Negative article	12	52.08	246.90

(b) Stemplots are given. The sample sizes are rather small, so the irregular look of the stemplots is fairly typical. Boxplots (not shown) indicate no outliers and no strong skew.

Positive		Negative
2	−3	7
0225	−2	2
	−1	77
27	−0	
7522	0	0
	1	2257
520	2	2
	3	0
	4	7

(c) We find $SE = \sqrt{\frac{188.52^2}{14} + \frac{246.90^2}{12}} = 87.2842$ and $t = \frac{-33.93 - 52.08}{87.2842} = -0.985$. Using df = 11 (the smaller of 14 − 1 and 12 − 1), $0.30 < P < 0.40$ (software gives $P = 0.336$ using df = 20.42). We find no evidence of any impact of a positive or negative article in *USA TODAY* on the Dow Jones Industrial Average. The lack of finding a significant difference is most likely due to small sample sizes and large variability in each sample. **(d)** The data do not support the contention that negative articles contribute to poor performance of the DJIA. The Dow actually performed somewhat better (on average) after a negative article, but not enough so to be significant.

21.41 (a) The summary table shows that the mean rating for those with a positive attitude toward Mitt is larger than the mean for those with a negative attitude; the standard deviations are relatively large, however.

	n	$\bar{x}$	s
Positive	29	3.9172	0.7960
Negative	29	3.6103	0.9127

(b) Shown are back-to-back stemplots for the two groups. The distribution of ratings for those with positive attitudes toward Mitt is somewhat right-skewed (but not extremely so). The distribution of ratings for those with negative attitudes toward Mitt is fairly symmetric. A check with a boxplot (not shown) indicates the two lowest ratings are not outliers.

Positive		Negative
	1	78
	2	
866	2	55569
433220	3	33333
999987776	3	5688999
221	4	1233334
9665	4	7
300	5	34
7	5	

(c) Test $H_0: \mu_1 = \mu_2$ versus $H_0: \mu_1 > \mu_2$, where μ_1 is the mean Mitt trustworthiness rating for students having a positive attitude toward Mitt Romney (as compared with Barack Obama), and μ_2 is the mean Romney trustworthiness rating for students having a negative attitude toward Romney. We find $SE = \sqrt{\frac{0.7960^2}{29} + \frac{0.9127^2}{29}} = 0.2249$ and $t = \frac{3.9172 - 3.6103}{0.2249} = 1.3646$, for which the P-value is $0.05 < P < 0.10$ (using df = 28) or 0.0890 (using software, with df = 54.98). There is not strong evidence that students with a positive attitude toward Mitt Romney give a larger mean trustworthiness rating of his face than students with a negative attitude toward him.

21.42 Summary statistics and background work are done in Exercise 21.40. The 90% confidence interval is $(\bar{x}_1 - \bar{x}_2) \pm t^* SE$, where $t^* = 1.796$ (df = 11) or $t^* = 1.723$ (df = 20.42). This gives either $86.01 \pm 156.762 = -70.752$ to 242.772 points (with df = 11) or $86.01 \pm 150.391 = -64.381$ to 236.401 points (with df = 20.42).

21.43 (a) Stemplots are shown. Each data value in the stemplot is rounded to the nearest thousand, and stems are in units of ten thousand. So, for example, the row "30 | 2 | 2" represents 3 people: 2 women who spoke about 23,000 and 20,000 words, and 1 man who spoke about 22,000 words. The stemplots suggest that there is some skew in both populations, but the sample sizes should be large enough to overcome this problem.

Women		Men
98876	0	4456789
43321000	1	00111333
998765	1	68
30	2	2
76555	2	8
	3	
	3	8
0	4	

(b) With subscripts as assigned in the statement of the problem (Group 1 = women), we test $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 > \mu_2$. We have $\bar{x}_1 = 16{,}496.1$, $\bar{x}_2 = 12{,}866.7$, $s_1 = 7{,}914.35$, $s_2 = 8{,}342.47$, $n_1 = 27$, and $n_2 = 20$; we find $SE = \sqrt{\frac{7914.35^2}{27} + \frac{8342.47^2}{20}} = 2408.26$, and $t = \frac{16{,}496.1 - 12{,}866.7}{2{,}408.26} = 1.51$. With df = 39.8 (using software), $P = 0.07$. Using Table C with the more conservative df = 19, $0.05 < P < 0.10$. There is some evidence that, on average, women say more words than men, but the evidence is not particularly strong.

21.44 This is a two-sample t statistic, comparing two independent groups (supplemented and control). Using the conservative df = 5, $t = -1.05$ would have a P-value between 0.30 and 0.40, which (as the report said) is not significant. The test statistic, $t = -1.05$, would not be significant for any value of df.

21.45 We test $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$, where μ_1 is the mean days behind caterpillar peak for the control group, and μ_2 is the mean days for the supplemented group. Now, with $\bar{x}_1 = 4.0$, $\bar{x}_2 = 11.3$, $s_1 = 3.10934$, $s_2 = 3.92556$, $n_1 = 6$, and $n_2 = 7$, we find $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1.95263$ and $t = \frac{4.0 - 11.3}{SE} = -3.74$. The two-sided P-value is either $0.01 < P < 0.02$ (using df = 5) or 0.0033 (using df = 10.96, with software), which agrees with the stated conclusion (a significant difference).

21.46 These are paired t statistics: for each bird, the number of days behind the caterpillar peak was observed, and the t values were computed based on the pairwise differences between the first and second years. For the control group, df = 5, and for the supplemented group, df = 6. The control t is not significant (so the birds in that group did not "advance their laying date in the second year"), whereas the supplemented group t is significant with one-sided $P = 0.0195$ (so those birds did change their laying date).

21.47 STATE: Does thinking about money lead people to be more reluctant to ask for help? PLAN: We test $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 > \mu_2$, where μ_1 is the mean time for the treatment group, and μ_2 is the mean time for the control group. The alternative hypothesis is one-sided because the researcher suspects that the treatment group will wait longer before asking for help. SOLVE: We must assume that the data come from an SRS of the intended population; we cannot check this

with the data. The provided back-to-back stemplot shows some irregularity in the treatment times and skewness in the control times. We hope that our equal and moderately large sample sizes will overcome any deviation from Normality. With $\bar{x}_1 = 314.0588$, $\bar{x}_2 = 186.1176$, $s_1 = 172.7898$, $s_2 = 118.0926$, $n_1 = 17$, and $n_2 = 17$, we find $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 50.7602$ and $t = \frac{314.0588 - 186.1176}{SE} = 2.521$, for which $0.01 < P < 0.02$ (df = 16) or $P = 0.0088$ (df = 28.27). CONCLUDE: There is strong evidence that the treatment group waited longer to ask for help on average.

Treatment		Control
65	0	5689
3	1	012444
976	1	58
44	2	
5	2	79
	3	
6	3	7
44	4	01
876	4	
3	5	
	5	
0	6	

21.48 STATE: Is gastric banding surgery more effective than lifestyle intervention in helping overweight teens lose weight? PLAN: We test $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 > \mu_2$, where μ_1 is the mean weight loss for adolescents in the gastric banding group, and μ_2 is the mean time for the lifestyle intervention group. The alternative hypothesis is one-sided because the researcher suspects that gastric banding leads to greater average weight loss than lifestyle modification. SOLVE: We must assume that the data come from an SRS of the intended population; we cannot check this with the data. The provided stemplot for each group shows no heavy skew and no outliers. With $\bar{x}_1 = 34.87$, $\bar{x}_2 = 3.01$, $s_1 = 18.12$, $s_2 = 13.22$, $n_1 = 24$, and $n_2 = 18$ (note that not all subjects completed the study), we find $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 4.84$ and $t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = 6.59$, for which $P < 0.0005$ (df = 17) or $P < 0.00001$ (df = 39.98, using software). CONCLUDE: There is strong evidence that adolescents who undergo gastric banding lose more weight on average than those who use lifestyle modification.

Gastric Banding		Lifestyle Intervention
	−1	762
5	−0	44331
	0	12466
953	1	155
97420	2	0
97653221	3	4
931	4	
73	5	
4	6	
	7	
1	8	

21.49 STATE: Does a painful experience in a small group lead to higher average bonding scores for group members than sharing a similar non-painful experience? PLAN: We test H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 > \mu_2$, where μ_1 is the mean bonding score for the pain group, and μ_2 is the mean bonding score for the no-pain group. SOLVE: We must assume that the data come from an SRS of the intended populations; we cannot check this with the data. The provided back-to-back stemplot shows that both groups are slightly skewed left. Also, using the 1.5 × *IQR* criterion, the pain group has two low outliers (1.29 and 1.43). We will remove these outliers, and hope that our moderately large sample sizes will overcome any deviation from Normality. With $\bar{x}_1 = 3.903$, $\bar{x}_2 = 3.138$, $s_1 = 0.7734$, $s_2 = 1.0876$, $n_1 = 25$, and $n_2 = 27$, we find $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.2603$ and $t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = 2.94$, for which $0.0025 < P < 0.005$ (df = 24) or $P = 0.0025$ (df = 46.9755, using software). CONCLUDE: There is strong evidence that a painful experience in a small group leads to higher average bonding scores for group members than sharing a similar non-painful experience.

pain		no pain
	1	033
	1	77
331	2	14
	2	79
44	3	00011144
966666	3	779
44433110	4	11344
977766	4	79

21.50 STATE: Does an encouraging subliminal message help students learn math better? PLAN: Compare mean length by testing H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 > \mu_2$ and by finding a 90% confidence interval for $\mu_1 - \mu_2$, where μ_1 is the mean for the treatment population, and μ_2 is the mean for the control population. SOLVE: We must assume that we have two SRSs, and that the distributions of score improvements are Normal. The provided back-to-back stemplots of the differences (after − before) for the two groups seem to indicate skew (especially for the control group), but the samples are too small to really assess Normality. There are no outliers. With $\bar{x}_1 = 11.4$, $\bar{x}_2 = 8.25$, $s_1 = 3.1693$, $s_2 = 3.6936$, $n_1 = 10$, and $n_2 = 8$, we find

$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1.646$ and $t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = 1.914$. With either df = 7, $0.025 < P < 0.05$. With df =13.92 (using software), $P = 0.0382$. The 90% confidence interval is (11.4 – 8.25) ± t^*SE, where $t^* = 1.895$ (df = 7) or $t^* = 1.762$ (df = 13.92), and either 0.03 to 6.27 points or 0.25 to 6.05 points. CONCLUDE: We have fairly strong evidence that the encouraging subliminal message led to a greater improvement in math scores, on average. We are 90% confident that this increase is between 0.03 and 6.27 points (or 0.25 and 6.05 points).

Treatment		Control
	0	455
76	0	7
	0	8
110	1	1
332	1	2
5	1	4
6	1	

21.51 (a) Refer to Exercise 21.49 for details. The 90% confidence interval for the difference in the mean bonding score for students in the no-pain and pain groups is (3.903 – 3.138) ± t^*(0.2603), where $t^* = 1.711$ (using df = 24) or $t^* = 1.678$ (using df = 46.9755). This interval is then 0.32 to 1.21 (df = 24) or 0.33 to 1.20 (df = 46.9755). **(b)** Using the notation from Exercise 21.49, the 90% confidence interval for the mean bonding score of students in the pain group is $\bar{x}_1 \pm t^* \left(\frac{s_1}{\sqrt{n_1}}\right) = 3.903 \pm 1.711 \left(\frac{0.7734}{\sqrt{25}}\right) = 3.638$ to 4.168 (where df $= 25 - 1 = 24$ for t^*).

21.52 STATE: Do the lengths of *H. caribaea* red and yellow differ enough to believe that they may have adapted to different hummingbird species? If so, how much is the mean difference in the two species? PLAN: Compare mean length by testing $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$ and by finding a 95% confidence interval for $\mu_1 - \mu_2$, where μ_1 is the mean for the red population, and μ_2 is the mean for the yellow population. SOLVE: We must assume that the data come from an SRS. We also assume that the data are close to Normal. The provided back-to-back stemplots show some skewness in the red lengths, but the *t* procedures should be reasonably safe. With $\bar{x}_1 = 39.7113$, $\bar{x}_2 = 36.18$, $s_1 = 1.7988$, $s_2 = 0.9753$, $n_1 = 23$, and $n_2 = 15$, we find $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.4518$ and $t = \frac{\bar{x}_1 - \bar{x}_2}{SE} = 7.817$. With either df = 14 or df = 35.1, $P < 0.0001$. The 95% confidence interval is (39.711 – 36.180) ± t^*SE, where $t^* = 2.145$ (df = 14) or $t^* = 2.030$ (df = 35.1), which is either 2.562 to 4.500 mm or 2.614 to 4.448 mm. CONCLUDE: We have very strong evidence that the two varieties differ in mean length. We are 95% confident that the mean red length minus the mean yellow length is between 2.562 and 4.500 mm (or 2.614 and 4.448 mm).

Red		Yellow
	34	56
	35	146
	36	0015678
9874	37	01
8722100	38	1
761	39	
65	40	
9964	41	
10	42	
0	43	

21.53 Because this exercise asks for a "complete analysis" without suggesting hypotheses or confidence levels, student responses may vary. This solution gives 95% confidence intervals for the means in parts (a) and (b), and performs a hypothesis test and gives a 95% confidence interval for part (c). Note that the first two problems call for single-sample *t* procedures (Chapter 20), whereas the last uses the Chapter 21 procedures. Student answers should be formatted according to the "four-step process" of the text; these answers are not formatted as such, but can be used to check student results. We begin with summary statistics.

	n	$\bar{x}$	s
Women	95	4.2737	2.1472
Men	81	6.5185	3.3471

A back-to-back stemplot of responses for men and women is shown. This plot reveals that the distribution of claimed drinks per day for women is slightly skewed but has no outliers. For men, the distribution is only slightly skewed but contains four outliers. However, these outliers are not too extreme. In all problems, it seems that the use of *t* procedures is reasonable.

Women		Men
00000000	1	000
5555555500000	2	0000
55550000000000000000	3	0000000
500000000000000000	4	0000000000555
00000000000	5	000000005
50000000	6	00000005
00000000	7	000000005
000	8	000000000
000	9	0000
00	10	00000005
	11	0
	12	05
	13	
	13	
	15	000
	16	0

(a) We construct a 95% confidence interval for μ_w, the mean number of claimed drinks for women. Here, t^* = 1.990 (df = 80 in Table C) or t^* = 1.9855 (df = 94, using software), and $SE = 2.1472/\sqrt{95} = 0.2203$. A 95% confidence interval for μ_w is $4.2737 \pm 1.990(0.2203) = 3.84$ to 4.71 drinks. The interval using software is virtually the same. With 95% confidence, the mean number of claimed drinks for women is between 3.84 and 4.71 drinks. **(b)** We construct a 95% confidence interval for μ_m, the mean number of claimed drinks for men. Here, t^* = 1.990 (df = 80 in Table C or using software), and $SE = 3.3471/\sqrt{81} = 0.3719$. A 95% confidence interval for μ_m is $6.5185 \pm 1.990(0.3719) = 5.78$ to 7.26 drinks. With 95% confidence, the mean number of claimed drinks for men is between 5.78 and 7.26 drinks. **(c)** We test $H_0: \mu_m = \mu_w$ versus $H_a: \mu_m \neq \mu_w$. We have $SE = \sqrt{\frac{2.1472^2}{95} + \frac{3.3471^2}{81}} = 0.4322$ and $t = \frac{4.2737 - 6.5185}{SE} = -5.193$. Regardless of the choice of df (80 or 132.15), this is highly significant ($P < 0.001$). We have very strong evidence that the claimed number of drinks is different for men and women. To construct a 95% confidence interval for $\mu_m - \mu_w$, we use t^* = 1.990 (df = 80) or t^* = 1.9781 (df = 132.15). Using $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, we obtain either 2.2448 ± 0.8601 or 2.2448 ± 0.8549. After rounding either interval, we report that, with 95% confidence, on average, sophomore men who drink claim an additional 1.4 to 3.1 drinks per day compared with sophomore women who drink.

21.54 and **21.55** are Web-based exercises.