Chapter 9

Exercises for Section 9.2

In all that follows, the hypotheses relate to values for true (or population) means or proportions. The evidence we have about the truth or otherwise of those hypotheses comes from what is happening in the sample data.

1. Here μ is the population or true mean volume. **(a)** $H_0 : \mu = 750$. **(b)** $H_1 : \mu < 750$. **(c)** We would check whether the sample mean volume, $\overline{x}$, from the 40 bottles tested is too much smaller than 750 for the difference to be explained simply in terms of sampling variation.

2. **(a)** $H_0 : \mu_{white} - \mu_{blue} = 0$. **(b)** $H_1 : \mu_{white} - \mu_{blue} > 0$. **(c)** We would check whether the sample mean blood pressure from the white-collar sample, $\overline{x}_{white}$, is sufficiently much larger than the sample mean from the blue-collar sample, $\overline{x}_{blue}$, that the difference could not be explained simply in terms of sampling variation.

3. **(a)** $H_0 : \mu_{French} - \mu_{math} = 0$. **(b)** $H_1 : \mu_{French} - \mu_{math} > 0$. **(c)** We would check whether the sample mean comprehension mark from the French class, $\overline{x}_{French}$, was sufficiently larger than the sample mean from the Additional Mathematics class, $\overline{x}_{math}$, that the difference could not be explained simply in terms of sampling variation.

4. **(a)** $H_0 : \mu = 10$. **(b)** $H_1 : \mu \neq 10$. **(c)** We would check whether the sample mean diameter, $\overline{x}$, from the 50 ball-bearings tested is too far from 10 (in either direction) for the difference to be explained simply in terms of sampling variation.

5. **(a)** $H_0 : p = 0.5$. **(b)** $H_1 : p \neq 0.5$. **(c)** We would check whether the sample proportion of the 50 ball-bearings tested with a diameter greater than the target, $\widehat{p}$, is too far from 0.5 (in either direction) for the difference to be explained simply in terms of sampling variation.

6. **(a)** $H_0 : p = 0.5$. **(b)** $H_1 : p \neq 0.5$. **(c)** We would check whether the sample proportion of heads, $\widehat{p}$, in the 1000 coin tosses is too far from 0.5 (in either direction) for the difference to be explained simply in terms of sampling variation.

7. $H_0 : p_{giveaway} - p_{none} = 0$. **(b)** $H_1 : p_{giveaway} - p_{none} < 0$, expecting people attracted by free gifts to be less loyal. (Could also argue for a "$\neq$" alternative.) **(c)** We would check whether the sample proportion of the giveaway sample who were still doing business with the bank 5 years later, $\widehat{p}_{giveaway}$, is sufficiently much smaller than the corresponding proportion for the no-giveaway sample, $\widehat{p}_{none}$, that the difference could not be explained simply in terms of sampling variation.

8. **(a)** $H_0 : \mu = 0$. **(b)** $H_1 : \mu \neq 0$. **(c)** We would check whether the sample mean net-earnings per person, $\overline{x}$, for the 1000 customers studied is sufficiently far from zero that the difference could not be explained simply in terms of sampling variation.

Exercises for Section 9.3

1. Let p be the true proportion sucking their left thumbs in the womb.

(a) The research hypothesis is that birth-stress "pushes infants towards left-handedness," and thus there should be fewer "left handers" before birth than there are after birth. Let p be the true proportion of babies who are "left-handed" before birth. We thus want to test the sceptical $H_0 : p = 0.1$ (before birth is the same as after) versus $H_1 : p < 0.1$ (from the research hypothesis).

We have a sample of $n = 224$ babies of which a sample proportion $\widehat{p} = 12/224 = 0.05357$ suck their left thumbs. Now $\text{se}(\widehat{p}) = \sqrt{\frac{0.05357 \times 0.94643}{224}} = 0.01504$ and the t-test statistic is $t_0 = \frac{0.05357 - 0.1}{0.015045} = -3.086$. This tells us that the sample mean from the data is more than 3 standard errors below the value of 0.1 hypothesized for the true mean. The (1-tailed) P-value is $\text{pr}(Z \leq -3.086) = 0.001$. There is very strong evidence against H_0 in favor of H_1, or in terms of p, there is very strong evidence that fewer than 10% of babies suck their left thumbs.

[Warning: the 10% rule gives n to be at least 960, which is not true, so large-sample theory is a little suspect.]

(b) The study premise is that the thumb-sucking behavior of fetuses relates to left and right handedness after birth (apart from some switching due to such things as "birth stress"). We also assume that Belfast left-handedness rates are 10% or more. Our analysis relates to a population from which these babies can be considered a random sample.

2. Let p be the true probability of a person dying in the month before her or his birthday. The research hypothesis is that this probability p should be lower than for other months because of the postponing effect. We will assume that, if such an effect did not exist, the month before the birthday would be just like a randomly chosen month and so the probability of dying in that month would be 1 chance in 12. In these terms, our research hypothesis says that $p < \frac{1}{12}$.

(a) We wish to test the sceptical $H_0 : p = \frac{1}{12}$ (a month like any other) versus $H_1 : p < \frac{1}{12}$ (from the research hypothesis).

We have a sample of $n = 348$ individuals for which the sample proportion dying in the month before the birthday is $\widehat{p} = \frac{16}{348} = .04598$. Now $\text{se}(\widehat{p}) = \sqrt{\frac{0.04598 \times 0.95402}{348}} = 0.011227$ from which we obtain $t_0 = \frac{0.04598 - 0.08333}{0.01123} = -3.327$. This tells us that the sample proportion from the data is more than 3.3 standard errors below the value of $\frac{1}{12}$ hypothesized for the true probability. The (1-tailed) P-value is thus $\text{pr}(Z \leq -3.327) = 0.0004$.
There is very strong evidence against H_0 in favor of H_1, or more concretely, there is very strong evidence in favor of the postponing-death theory.

[Warning: the 10% rule gives n to be at least 960, which is not the case, so large-sample theory is a little suspect.]

(b) These were all "Notable Americans." To generalize we would have to assume that "ordinary" people have the same survival behavior as "notable" people as far as postponing death goes. We assume some sort of uniformity of the birth

and death rates throughout the year. For example, if most births were in the summer and most deaths in the winter for reasons which had nothing to do with "postponing" death, our estimate of $\widehat{p}$ would be small.

3. Let μ be the true mean nicotine content. We will test $H_0 : \mu = 18$ versus $H_1 : \mu > 18$ (as the prior claim is one sided). We have a sample of $n = 12$ cigarettes for which the sample mean nicotine content is $\overline{x} = 19.1$ with a standard deviation of $s = 1.9$. Now $\text{se}(\overline{x}) = \frac{s}{\sqrt{n}} = \frac{1.9}{\sqrt{12}} = 0.54848$. The t-test statistic is thus $t_0 = \frac{19.1-18}{0.54848} = 2.0055$. This tells us that the sample mean from the data is more than 2 standard errors above the value of 18 hypothesized for the true mean.

Using $T \sim \text{Student}(df = n - 1 = 11)$, the (1-tailed) P-value is $\text{pr}(T \geq 2.00555) = 0.035$. There is some evidence against H_0 in favor of H_1, or more concretely, we do have some evidence that the claim is false.

4. Let p_S be the true proportion of smoking mothers with infants getting colic and p_{NS} be the true proportion of non-smoking mothers with infants getting colic. There is not enough information given for us to determine whether the investigators suspected some particular effect of smoking or whether they just thought they noticed something. So we will test the sceptical $H_0 : p_S - p_{NS} = 0$ (smoking makes no difference) versus the 2-sided alternative $H_1 : p_S - p_{NS} \neq 0$. Of a sample of $n_S = 200$ babies of smoking mothers, a sample proportion $\widehat{p}_S = 0.4$ had colic compared with a proportion $\widehat{p}_{NS} = 0.2$ among $n_{NS} = 400$ babies of nonsmoking mothers. We are comparing proportions from independent samples (situation (a) in Fig. 8.5.1), so $\text{se}(\widehat{p}_S - \widehat{p}_{NS}) = \sqrt{\frac{0.4\times0.6}{200} + \frac{0.2\times0.8}{400}} = 0.04$. The test statistic is thus $t_0 = \frac{(0.4-0.2)-0}{0.04} = 5$. This tells us that our estimated difference in proportions from the data is more than 5 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 5) = 0.0000$. There is very strong evidence against H_0. There is very strong evidence that a true difference exists, or more concretely, very strong evidence that smoking mothers are more likely to have colicky babies. (We deduce the direction of the effect from the sample estimates. Later we will state as a rule never to perform a test without also constructing a confidence interval from which we can read off the likely the size of the difference.)

5. Let p_{ES} be the proportion knowing that Christ was resurrected on Easter Sunday and p_{GF} be the proportion knowing that Christ was crucified on Good Friday. We will test $H_0 : p_{ES} - p_{GF} = 0$ (no difference) versus the 2-sided alternative $H_1 : p_{ES} - p_{GF} \neq 0$, as we have no prior reason to expect a difference in one direction of the other.

In our sample of size $n = 1101$ people, the corresponding sample proportions were $\widehat{p}_{ES} = 0.66$ and $\widehat{p}_{GF} = 0.61$, thus suggesting that more people know what Easter Sunday commemorates. This is a situation (c) comparison in Fig. 8.5.1 so $\text{se}(\widehat{p}_{ES} - \widehat{p}_{GF}) = \sqrt{\frac{0.34+0.39-(0.66-0.61)^2}{1101}} = 0.02571$. Our test statistic is thus $t_0 = \frac{(0.66-0.61)-0}{0.02571} = 1.945$. This tells us that our estimated difference in proportions from the data is nearly 2 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 1.945) = 0.052$. We do have some evidence against H_0. We do have some evidence that a real difference exists, or more concretely, that more people know what Easter Sunday commemorates. (The direction of the difference is deduced from the sample estimates.)

6. Let p_B be the probability of accepting if claimed beneficial and p_{NB} be the probability of accepting if claimed not beneficial. We will test the sceptical hypothesis $H_0 : p_B - p_{NB} = 0$ (whether the paper "found" that social-work intervention was beneficial or not makes no difference to the probability of acceptance) versus the 2-sided alternative $H_1 : p_B - p_{NB} \neq 0$ on the grounds that the story did not contain enough information for us to know what Epstein hypothesized before starting the study. [We strongly suspect that his research hypothesis was that articles claiming intervention was beneficial would be *more* likely to be accepted. If this was the case, the alternative hypothesis should be $H_1 : p_B - p_{NB} > 0$.]

Of the $n_B = 70$ articles claiming benefit, a proportion $\widehat{p}_B = 0.53$, were accepted, whereas of $n_{NB} = 70$ claiming no benefit only a proportion $\widehat{p}_{NB} = 0.14$ were accepted. We are comparing proportions from two independent samples so $\text{se}(\widehat{p}_B - \widehat{p}_{NB}) = \sqrt{\frac{0.53\times0.47}{70} + \frac{0.14\times0.86}{70}} = 0.07265$. Our test statistic is thus $t_0 = \frac{0.53-0.14}{0.07265} = 5.368$. This tells us that our estimated difference in proportions from the data is more than 5 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 5.368) = 0.0000$. There is very strong evidence against H_0. There is very strong evidence that a true difference exists, or more concretely, that journals are more likely to accept articles claiming intervention is beneficial. (The direction of the effect is deduced from the data estimates.)

[Warning: The 10% rule require n_B to be at least 11 and n_{NB} to be at least 243, so large-sample theory is a little suspect.]

We are assuming that the 70 journals to get the "beneficial" version were selected at random and the journals made decisions independently, e.g., we do not have the situation where different journals are using the same referees to determine their decisions.

7. Let μ_{HS} be the true mean length in hedge-sparrow nests and μ_{GW} be the true mean length in garden-warbler nests.
We will test the sceptical null hypothesis $H_0 : \mu_{HS} - \mu_{GW} = 0$ (type of nest makes no difference) versus the 2-sided alternative $H_1 : \mu_{HS} - \mu_{GW} \neq 0$ (as there is no information about a direction of difference from a prior research hypothesis).

The $n_{HS} = 58$ eggs in hedge sparrow nests had a sample mean length of $\overline{x}_{HS} = 22.6$ and standard deviation of $s_{HS} = 0.8759$ compared with $\overline{x}_{GW} = 21.9$ and $s_{GW} = 0.7860$ for the $n_{GW} = 91$ eggs in garden warbler nests. Now, $\text{se}(\overline{x}_{HS} - \overline{x}_{GW}) = \sqrt{\frac{0.8759^2}{58} + \frac{0.7860^2}{91}} = 0.14148$ so that $t_0 = \frac{(22.6-21.9)-0}{0.14148} = 4.948$. This tells us that our estimated difference in means from the data is nearly 5 standard errors from zero. Using $T \sim$ Student with $df = \min(n_{HS} - 1, n_{GW} - 1) = 57$, the (2-tailed) P-value is $2 \times \text{pr}(T \geq 4.948) = 0.0000$. There is very strong evidence that a true difference in mean lengths exists, or more concretely, that eggs in hedge-sparrow nests tend to be larger. (The direction of the difference is deduced from the sample estimates.)

We cannot immediately conclude that the type of nest causes the observed differences in size as this is observational data. There may be other mechanisms such as bigger birds tending to select hedge-sparrow nests, or differences (e.g., in food supplies) between habitats containing mainly hedge sparrows or mainly garden warblers.

Review Exercises 9

Throughout the following Review Exercises we continue to abbreviate "confidence interval" to "CI." In choosing the alternative hypothesis for testing we have used the conservative 2-sided alternative unless it is clear that there was a research hypothesis that should determine the null. In many cases the researchers probably did have a research hypothesis and we have a strong suspicion about what that hypothesis would have been. In these cases, we have discussed the consequences of the use of our "suspected" research hypothesis.

1. **(a)** Let p_T and p_C be the respective true probabilities that a person will return a completed questionnaire with or without telephone contact. We will test the sceptical null hypothesis $H_0 : p_T - p_C = 0$ (phone calls make no difference) versus the 2-sided alternative $H_1 : p_T - p_C \neq 0$. [If the very plausible proposition that "a followup telephone call would increase the likelihood of a response" was the research hypothesis, then we should test versus $H_1 : p_T - p_C > 0$.]

Of the sample of $n_T = 239$ people followed up by telephone a proportion $\widehat{p}_T = \frac{134}{239} = 0.5607$ responded, versus a proportion $\widehat{p}_C = \frac{186}{836} = 0.2225$ of the $n_C = 836$ people in the control group.

Since we are comparing proportions from independent samples, $\text{se}(\widehat{p}_T - \widehat{p}_C) = \sqrt{\frac{0.56067 \times 0.43933}{239} + \frac{0.22249 \times 0.77751}{836}} = 0.03518$. Our test statistic is $t_0 = \frac{0.56067 - 0.22249}{0.03518} = 9.613$. This tells us that our estimated difference in proportions from the data is more than 9 standard errors from zero! The P-value is 0 to many more than 4 decimal places whether we do it 1- or 2-tailed. There is very strong evidence that a true difference exists, or more concretely, that phone calls increase the response rate.

(b) A 95% CI for $p_T - p_C$ is [0.27,0.41]. With 95% confidence, calls increase the percentage responding by between 27 and 41 percentage points.

(c) Even though there is substantially less nonresponse in the treatment group, it is still quite high so nonresponse bias would still be a worry. If only people they contacted by phone were sent questionnaires, this could add further bias.

2. [Note that the data gives us standard errors of $\overline{x}$ for each sample, not the sample standard deviations.]

(a) Let μ_{WO} be the true mean number of words per sentence for the old version and μ_{WN} be the true mean average number of words per sentence for the new version. We will test $H_0 : \mu_{WO} - \mu_{WN} = 0$ (no change) versus the 2-sided alternative $H_1 : \mu_{WO} - \mu_{WN} \neq 0$. [If "dumbing down" was Lendvoy's research hypothesis, then we should test versus $H_1 : \mu_{WO} - \mu_{WN} > 0$. The resulting P-value would be half the size of the one quoted below. For this problem, this would have no effect on the conclusions reached.]

From the data we get sample means $\overline{x}_{WO} = 15.31$ and $\overline{x}_{WN} = 11.88$, giving an estimated difference of $15.31 - 11.88 = 3.43$. Because the individual $\text{se}(\overline{x})$ values have been supplied to us and we have independent samples, we obtain the standard error of the difference using $\text{se}(\overline{x}_{WO} - \overline{x}_{WN}) = \sqrt{\text{se}(\overline{x}_{WO})^2 + \text{se}(\overline{x}_{WN})^2} =$

$\sqrt{0.71^2 + 0.65^2} = 0.9626$. From this we obtain $t_0 = \frac{(15.31-11.88)-0}{0.9626} = 3.563$. Using $T \sim$ Student with $df = \min(n_{WN} - 1, n_{WO} - 1) = 99$, the (2-tailed) P-value is $2 \times \text{pr}(T \geq 3.563) = 0.00028$. There is very strong evidence against H_0 in favor of H_1, or more concretely, there is very strong evidence of a real change in average words per sentence.

A 95% CI for $\mu_{WO} - \mu_{WN}$ is given by [1.5, 5.3]. The true mean number of words per sentence is lower in the new version than in it was in the older version by somewhere between approximately 1.5 and 5.3 words per sentence.

(b) We will test $H_0 : \mu_{SO} - \mu_{SW} = 0$ versus $H_1 : \mu_{SO} - \mu_{SW} \neq 0$.

The sample means from the data are $\overline{x}_{SO} = 21.02$ and $\overline{x}_{SW} = 16.34$, giving an estimated difference of 4.68. The standard error of the difference is $\text{se}(\overline{x}_{SO} - \overline{x}_{SW}) = \sqrt{0.97^2 + 0.95^2} = 1.3577$ from which we obtain $t_0 = \frac{4.68}{1.3577} = 3.447$. The (2-tailed) P-value is 0.0004, using Student($df = 99$). Once again there is very strong evidence that a real difference exists.
The 95% CI for $\mu_{SO} - \mu_{SW}$ is [2.0, 7.4]. With 95% confidence, there has been a reduction of between approximately 2 and 7 syllables per sentence on average.

(c) One suggestion is take a simple random sample of pages. For each page selected, number the sentences 1,2, $\cdots$ and choose a single random sample of sentences. Delete uncompleted sentences from previous page and include incomplete sentence of the end of the page. (or vice versa).

3. (a) Let p_{none} be the true probability that a regular purchaser (no incentive) will buy again and p_{coup} be the true probability that a purchaser using a coupon will buy again. We will test $H_0 : p_{none} - p_{coup} = 0$ (no change) versus the 2-sided alternative $H_1 : p_{none} - p_{coup} \neq 0$. [If the very plausible proposition that "people buying using a coupon would be less loyal" was the research hypothesis, then we should test versus $H_1 : p_{none} - p_{coup} > 0$. In this problem, the change has no effect on the conclusions reached.]

From the data we get sample estimates $\widehat{p}_{none} = 0.87$ and $\widehat{p}_{coup} = 0.49$ from samples of size $n_{none} = 23,794$ and $n_{coup} = 671$ respectively.

We are comparing proportions from independent samples so $\text{se}(\widehat{p}_{none} - \widehat{p}_{coup}) = \sqrt{\frac{0.87 \times 0.13}{23794} + \frac{0.49 \times 0.51}{671}} = 0.019421$. From this we obtain test statistic $t_0 = \frac{(0.87-0.49)-0}{0.019421} = 19.566$. This tells us that our estimated difference in proportions from the data is more than 19 standard errors from zero! The P-value is vanishingly small whether we perform the test 1- or 2-tailed. It is clear that there is a true difference, or more concretely, there is very strong evidence that brand loyalty is lower when customers are attracted by inducements.

(b) A 95% CI for $p_{none} - p_{coup}$ is [0.34, 0.42]. With 95% confidence, brand loyalty is lower by between 34 and 42 percentage points when a coupon offer is involved.

(c) Among new customers, the reduction in brand loyalty will probably be higher as new customers may have only switched to the brand during the coupon special.

(d) The real issue here is whether such a promotion attracts sufficient new profits to be cost effective. (It does not matter if only a low proportion of the customers who switched during the promotion stayed with the brand.) To test this, it would

be better to look at sales trends before and after the promotion and analyze these to see if there has been any significant jump in sales. (Why might you expect to see a short term drop in sales immediately after a promotion?)

4. (a) The research hypothesis is that woodpeckers prefer older trees and therefore that cavity trees should be older on average than colony trees. Thus we will test the sceptical null hypothesis $H_0 : \mu_{cav} - \mu_{col} = 0$ (no difference in age) versus $H_1 : \mu_{cav} - \mu_{col} > 0$.

Form the data we have $\overline{x}_{cav} - \overline{x}_{col} = 104.1 - 83.6 = 20.5$ with $\text{se}(\overline{x}_{cav} - \overline{x}_{col}) = \sqrt{\frac{24.1^2}{54} + \frac{38.3^2}{143}} = 4.58407$. Thus, $t_0 = \frac{(104.1-83.6)-0}{4.58407} = 4.472$. This tells us that our estimated difference in means from the data is nearly 4.5 standard errors from zero. Using $T \sim$ Student with $df = \min(n_{cav} - 1, n_{col} - 1) = 53$, the (1-tailed) P-value is $\text{pr}(T \geq 4.58407) = 0.0000$. There is very strong evidence that a true difference exits, or more concretely, that cavity trees are older on average.

A 95% CI for the true difference is [11.3, 29.7], telling us that cavity trees are older on average by somewhere between approximately 11 and 30 years.

(b) Older trees have a longer time period to be visited by woodpeckers. Because of seed dispersal, we may expect an older tree (parent) to surrounded by younger trees (offspring). We need to find out about the current movement of woodpeckers, e.g., using radio tags.

(c) Not strictly independent as trees are in the same neighborhood.

5. (a) Let p_{pay} be the true proportion who would cooperate if the payment is made and p_{con} be the true proportion who would cooperate under control conditions (no payment). We will test $H_0 : p_{pay} - p_{con} = 0$ (payments make no difference to the probability of cooperation) versus the 2-sided alternative $H_1 : p_{pay} - p_{con} \neq 0$. [If the very plausible proposition that "payment would increase the probability of cooperation" was the research hypothesis, then we should test versus $H_1 : p_{pay} - p_{con} > 0$. The resulting P-value would be half the size of the one quoted below. For this problem, this would result in somewhat stronger evidence for the existence of the effect.]

The sample proportions from the data were $\widehat{p}_{pay} = 0.793$ and $\widehat{p}_{con} = 0.672$ from samples of size $n_{pay} = 111$ and $n_{con} = 116$ respectively.
We are comparing proportions from independent samples so $\text{se}(\widehat{p}_{pay} - \widehat{p}_{con}) = \sqrt{\frac{0.793\times0.207}{111} + \frac{0.672\times0.328}{116}} = 0.058129$. Our test statistic is $t_0 = \frac{(0.793-0.672)-0}{0.058129} = 2.0816$. This tells us that our estimated difference in proportions from the data is more than 2 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 2.082) = 0.037$. There is some evidence that a true difference exists, or more concretely, that payments increase cooperation rates.

(b) The 95% CI for $p_{pay} - p_{con}$ is [0.007, 0.235]. In this environment, with 95% confidence, a \$5 payment increases the percentage cooperating by somewhere between 0.7 percentage points (almost no increase) and 24 percentage points.

(c) Paying participants reduces the number of people you will be able to afford to survey. So one of the tradeoffs is response rate versus sample size.

6. Let p_{none}, p_{1-10} and p_{11+} respectively represent the true proportions later becoming schizophrenic among those who had never used marijuana, among those who had used it 1–10 times, and those who had used it 11 or more times.

(a) We will test $H_0 : p_{none} - p_{1-10} = 0$ versus $H_1 : p_{none} - p_{1-10} \neq 0$. The sample proportions are $\widehat{p}_{none} = \frac{197}{41,280} = 0.0047723$ and $\widehat{p}_{1-10} = \frac{18}{2836} = 0.0063470$.
As we are comparing proportions from independent samples (situation (a) in Fig. 8.5.1),
$\text{se}(\widehat{p}_{none} - \widehat{p}_{1-10}) = \sqrt{\frac{0.0047723 \times 0.9952277}{41280} + \frac{0.0063470 \times 0.993653}{2836}} = 0.00152933$. The test statistic is $t_0 = \frac{(0.0047723 - 0.0063470) - 0}{0.00152933} = -1.0297$. This tells us that our estimated difference in proportions from the data is only about 1 standard errors from zero. The P-value is $2 \times \text{pr}(Z \geq 1.0297) = 0.303$, which is quite large. There is no evidence of a real difference. Sampling variation alone would quite often give rise to a difference that is as big or bigger than the difference we see between our sample proportions (it would happen 30% of the time).
The 95% CI for the true difference, $p_{1-10} - p_{none}$, is [-0.0046, 0.0014].
[Warning: there are problems with the minimum sample-size rule here and in (b).]

(b) This time we will test $H_0 : p_{11+} - p_{none} = 0$ versus $p_{11+} - p_{none} \neq 0$. We have samples of size $n_{11+} = 702 + 752 = 1454$ and $n_{none} = 41,280$ from which the sample proportions are $\widehat{p}_{11+} = \frac{31}{1454}$ and $\widehat{p}_{none} = \frac{197}{41,280}$. Following the same steps as in (a) we find $\text{se}(\widehat{p}_{11+} - \widehat{p}_{none}) = 0.003803$ and $t_0 = 4.3509$. This tells us that our estimated difference in proportions from the data is more than 4 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 4.3509) = 0.0000$. There is very strong evidence against H_0, i.e., very strong evidence that a true difference between schizophrenia rates between the two groups exists.
The 95% CI for $p_{11+} - \widehat{p}_{none}$ is [0.009, 0.024]. With 95% confidence, the true percentage contracting schizophrenia is higher in the 11+ group than in the no-marijuana group by somewhere between 0.9 and 2.4 percentage points.

(c) Although there is a relationship between marijuana use and the subsequent development of schizophrenia, this is an observational study so it does not prove causation. It could just be, for example, that the sorts of people who are most likely to develop schizophrenia tend to find drug use attractive. Marijuana-use lifestyles may go along with other behavior patterns, one of which may be the real cause.

(d) We have no information about drug use patterns over the 15 years of followup. For example, some of the no-marijuana people may subsequently have started using the drug. We are relying on memories of usage so some people will be misclassified into the wrong groups. With the data being collected as the soldiers were being inducted into military service, many people may have misrepresented their real usage.

7. Let p_{before} represent the true proportion of those opening counts before the promotion whose accounts were still open 6 months later. Let p_{during} be the corresponding true proportion for accounts opened during the promotion.

(a) We will test $H_0 : p_{before} - p_{during} = 0$ (no difference in loyalty) versus the 2-sided alternative $H_1 : p_{before} - p_{during} \neq 0$. [If the very plausible proposition that "people

opening accounts during the promotion should be less loyal" was the research hypothesis, then we should test versus $H_1 : p_{before} - p_{during} > 0$ resulting in a P-value half the size of the one presented below. In this problem, this would have no real effect on the conclusions reached.]

Our data gives sample proportions of $\widehat{p}_{before} = \frac{178}{200} = 0.89$ and $\widehat{p}_{during} = \frac{158}{200} = 0.79$ from samples of size $n_{before} = 200$ and $n_{during} = 200$, respectively.
Since we are comparing proportions from separate samples (situation (a) in Fig.8.5.1), $\text{se}(\widehat{p}_{before} - \widehat{p}_{during}) = \sqrt{\frac{0.89\times0.11}{200} + \frac{0.79\times0.21}{200}} = 0.036318$. From this we obtain $t_0 = \frac{(0.89-0.79)-0}{0.036318} = 2.7535$. This tells us that our estimated difference in proportions from the data is more than 2.75 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 2.75) = 0.006$. There is strong evidence against H_0, or more concretely, strong evidence that the induced customers are less loyal.

The 95% CI for $p_{before} - p_{during}$ is [0.029, 0.17]. With 95% confidence, the true percentage of those accounts opened during the promotion that "remain loyal" is lower by somewhere between 3 and 17 percentage points than that for accounts opened before the promotion.

(b) The actual number of accounts retained and the value of the accounts to the bank. The cost of the promotion.

8. In this problem we test many hypotheses of the form $H_0 : \mu_1 - \mu_2 = 0$, where μ_1 and μ_2 are the true means, using a test statistic of the form $t_0 = \frac{(\overline{x}_1-\overline{x}_2)-0}{\text{se}(\overline{x}_1-\overline{x}_2)}$. In each case we will be dealing with sample means from independent samples so $\text{se}(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. Degrees of freedom for calculating P-values are obtained using $df = \min(n_1 - 1, n_2 - 1)$. Details of calculation will be given in (a)(i). Thereafter we simply state results.

(a) **(i)** We will test $H_0 : \mu_{A.no} - \mu_{A.sim} = 0$ versus $H_1 : \mu_{A.no} - \mu_{A.sim} \neq 0$. From our data, we have $\overline{x}_{A.no} = 6.20$ and $s_{A.no} = 1.38$ from a sample of $n_{A.no} = 43$; and $\overline{x}_{A.sim} = 6.36$ and $s_{A.sim} = 1.28$ from a sample of $n_{A.sim} = 40$
We have $\text{se}(\overline{x}_{A.no} - \overline{x}_{A.sim}) = \sqrt{\frac{1.38^2}{43} + \frac{1.28^2}{40}} = 0.29197$. The resulting t-test statistic is $t_0 = \frac{(6.20-6.36)-0}{0.29197} = -0.548$. This tells us that our estimated difference in means from the data is only about half a standard error from zero. Using Student's t distribution with $df = \min(n_{A.no}-1, n_{A.sim}-1) = 39$, the (2-tailed) P-value is $2 \times \text{pr}(T \geq 0.548) = 0.59$. There is no evidence that a a real difference exists. [Approximately 60% of the time, sampling variation alone would result in differences as big as or bigger than those we saw in our data.]

The 95% CI for $\mu_{A.no} - \mu_{A.sim}$ is $[-0.74, 0.42]$. With 95% confidence, the true mean attraction score under "no-information" conditions could be anywhere between being 0.74 units smaller than the mean under "similar" conditions and being 0.42 units bigger. This includes the possibility that there is no difference at all.

(ii) We will test $H_0 : \mu_{A.no} - \mu_{A.dis} = 0$ versus $H_1 : \mu_{A.no} - \mu_{A.dis} \neq 0$. Summary statistics from the data are $\overline{x}_{A.dis} = 4.64$, $s_{A.dis} = 1.33$ and $n_{A.dis} = 39$. We find $\text{se}(\overline{x}_{A.no} - \overline{x}_{A.dis}) = 0.2994$ and $t_0 = 5.210$. This tells us that our estimated difference in means from the data is more than 5 standard

errors from zero. Using Student($df = 38$), the (2-tailed) P-value is $2 \times \text{pr}(T \geq 5.210) = 0.0000$. There is very strong evidence against H_0, i.e., very strong evidence that a true difference exists. The 95% CI for the true difference $\mu_{A.no} - \mu_{A.dis}$ is [0.95, 2.17]. With 95% confidence, mean attraction ratings are larger on average when no information is given than when told the stranger is attitudinally dissimilar by somewhere between 0.95 and 2.17 points.

(b) **(i)** We will test $H_0 : \mu_{P.no} - \mu_{P.sim} = 0$ versus, $H_1 : \mu_{P.no} - \mu_{P.sim} \neq 0$. Summary statistics from the data are $\overline{x}_{P.no} = 5.98$, $s_{P.no} = 1.52$ and $n_{P.no} = 43$ $\overline{x}_{P.sim} = 6.60$, $s_{P.sim} = 1.37$ and $n_{P.sim} = 40$. We find $\text{se}(\overline{x}_{P.no} - \overline{x}_{P.sim}) = 0.317258$ and $t_0 = -1.954$. This tells us that our estimated difference in means from the data is nearly 2 standard errors from zero. Using Student($df = 39$), the (2-tailed) P-value is $2 \times \text{pr}(T \geq 1.954) = 0.058$.
There is some evidence against H_0, i.e., there is some evidence that a real difference exists.

The 95% CI for the true difference, $\mu_{P.no} - \mu_{P.sim}$, is $[-1.26, 0.02]$. With 95% confidence the true mean similarity rating under no-information conditions could be anywhere from being 1.26 units smaller than the mean under "stranger is similar" conditions to being very slightly larger.

(ii) We will test $H_0 : \mu_{P.no} - \mu_{P.dis} = 0$ versus $H_1 : \mu_{P.no} - \mu_{P.dis} \neq 0$. Summary statistics from the data are $\overline{x}_{P.dis} = 2.28$, $s_{P.dis} = 1.15$ and $n_{P.dis} = 39$. We find $\text{se}(\overline{x}_{P.no} - \overline{x}_{P.dis}) = 0.29604$ and $t_0 = 12.498$. This tells us that our estimated difference in means from the data is more than 12 standard errors from zero! Using Student($df = 38$), the (2-tailed) P-value is $2 \times \text{pr}(T \geq 12.498) = 0.0000$. There is very strong evidence against H_0, i.e., there is very strong evidence that a real difference exists. The 95% CI for the true difference, $\mu_{P.no} - \mu_{P.dis}$, is [3.10, 4.30]. With 95% confidence, mean similarity ratings are larger on average when no information is given than when told the stranger is attitudinally dissimilar by somewhere between 3.1 and 4.3 points.

(c) The confidence intervals are given and interpreted at the end of the relevant parts of (a) and (b).

(d) The data is consistent with the theory, but does not prove the 2nd part of it. The dissimilarity effect is clearly established and clearly much larger. However, the implication that there is no similarity/attraction effect is too strong. In one case ((b)(i)) we did have some evidence of a similarity/attraction effect. Non-significance is not proof of "no effect."

(e) and **(f)** are for class discussion. Many designs are possible and there other ways of obtaining subjects. We just present one or two ideas. Was only one "stranger" used? If we use other people for the "stranger," do we get similar results? This study uses students from a particular place. Do the results generalize to other people and other places? Are their cultural differences? Can we detect a similarity/attraction effect more strongly with a larger study?

9. **(a)** Let p_{smoke} be true proportion of smokers who then have a stroke and p_{nonsm} be the corresponding true proportion for nonsmokers. We will test $H_0 : p_{smoke} -$

$p_{nonsm} = 0$ (smoking makes no difference) versus the 2-sided alternative H_1 : $p_{smoke} - p_{nonsm} \neq 0$ (as there is no description of a research hypothesis suggesting a particular direction).
Our data gives sample proportions of $\widehat{p}_{smoke} = \frac{171}{3435} = 0.049782$ and $\widehat{p}_{nonsm} = \frac{117}{4437} = 0.026369$ from samples of size $n_{smoke} = 3435$ and $n_{nonsm} = 4437$, respectively. We are comparing proportions from separate samples (situation (a) in Fig 8.5.1) so $\text{se}(\widehat{p}_{smoke} - \widehat{p}_{nonsm}) = \sqrt{\frac{0.049782 \times 0.950218}{3435} + \frac{0.026369 \times 0.973631}{4437}} = 0.0044224$. Thus our test statistic is $t_0 = \frac{(0.049782 - 0.026369) - 0}{0.0044224} = 5.294$. This tells us that our estimated difference in proportions from the data is more than 5 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 5.294) = 0.0000$. There is very strong evidence against H_0, i.e., very strong evidence of a true difference (smokers are more likely to have strokes).

(b) The 95% CI for the true difference, $p_{smoke} - p_{nonsm}$, is [0.015, 0.032]. With 95% confidence, the true percentage of smokers having strokes is higher by between 1.5 and 3.2 percentage points than the percentage for nonsmokers. Put another way, the risk is increased by somewhere between 1.5 and 3.2 chances in 100.

10. (a) Let p_{tv} be true proportion of purchasers who were most affected by TV advertising and p_{mag} be the true proportion who were most affected by magazine advertising. We will test $H_0 : p_{tv} - p_{mag} = 0$ (no difference) versus $H_1 : p_{tv} - p_{mag} \neq 0$ (as there is no description of a research hypothesis suggesting a particular direction). We have a situation (b) comparison in Fig. 8.5.1 so $\text{se}(\widehat{p}_{tv} - \widehat{p}_{mag})$ is given by $\sqrt{\frac{0.21 + 0.18 - (0.21 - 0.18)^2}{500}} = 0.027896$.
The resulting test statistic is $t_0 = \frac{(0.21 - 0.18) - 0}{0.027896} = 1.075$. This tells us that our estimated difference in proportions from the data is only about 1 standard error from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 1.075) = 0.28$. We have no evidence of a true difference. [Approximately 30% of the time, sampling variation alone would result in differences at least as big as those we saw in our data.]

The 95% CI for the true difference, $p_{tv} - p_{mag}$, is $[-0.025, 0.085]$, putting the true percentage most influenced by TV ads somewhere between being 2.5 percentage points smaller and 8.5 percentage points larger than the percentage most influenced by magazine ads.

(b) What mode of advertising has the largest impact on people? The number of people the advertising reaches? Overall, the cost effectiveness of the advertising.

(c) For discussion.

11. (a) Let μ_{morn} and μ_{aft} be the respective population mean pH levels for morning and afternoon patients. We will test $H_0 : \mu_{morn} - \mu_{aft} = 0$ (no difference between morning and afternoon) versus the 2-sided alternative $H_1 : \mu_{morn} - \mu_{aft} \neq 0$ (as there is no description of a research hypothesis suggesting a particular direction).

Summary statistics from the data are $\overline{x}_{morn} = 3.94$ and $s_{morn} = 2.51$ from a sample of size $n_{morn} = 50$, and $\overline{x}_{aft} = 2.93$ and $s_{aft} = 2.39$ from a sample of size $n_{aft} = 49$. Since we are dealing with separate or independent samples, $\text{se}(\overline{x}_{morn} - \overline{x}_{aft}) = \sqrt{\frac{2.51^2}{50} + \frac{2.39^2}{49}} = 0.49252$. The resulting t-test statistic is $t_0 =$

$\frac{(3.94-2.93)-0}{0.49252} = 2.051$. This tells us that our estimated difference in means from the data is more than 2 standard errors from zero. Using Student's t distribution with $df = \min(n_{A.no} - 1, n_{A.sim} - 1) = 48$, the (2-tailed) P-value is $2 \times \text{pr}(T \geq 2.051) = 0.046$. We do have some evidence against H_0, i.e., evidence that a true difference exists (lower pH on average for morning patients).

The 95% CI for the true difference, $\mu_{morn} - \mu_{aft}$, is [0.02, 2.00]. With 95% confidence, the true mean pH level for morning patients is bigger than it is for afternoon patients by somewhere between 0.02 and 2.0 units.

(b) let p_{aft} and p_{morn} be the respective population proportions of morning and afternoon patients with a pH level below 2.5. We will test $H_0 : p_{aft} - p_{morn} = 0$ versus $H_1 : p_{aft} - p_{morn} \neq 0$.
Our data gives sample proportions of $\widehat{p}_{aft} = \frac{31}{49} = 0.63265$ and $\widehat{p}_{morn} = \frac{21}{50} = 0.42$ from samples of size $n_{aft} = 200$ and $n_{morn} = 200$ respectively. Since we are working with proportions from independent samples (situation (a) in Fig. 8.5.1), $\text{se}(\widehat{p}_{aft} - \widehat{p}_{morn}) = \sqrt{\frac{0.63265 \times 0.36735}{49} + \frac{0.42 \times 0.58}{50}} = 0.098056$. The resulting test statistic is $t_0 = \frac{(0.63265-0.42)-0}{0.098056} = 2.169$. This tells us that our estimated difference in proportions from the data is more than 2 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 2.169) = 0.03$. We do have some evidence against H_0, i.e., evidence that a true difference exists (more afternoon patients have a pH below 2.3).
The 95% CI for the true difference, $p_{aft} - p_{morn}$, is [0.02,0.40]. With 95% confidence, the true percentage of afternoon patients with a pH below 2.5 is larger than the corresponding percentage for morning patients by somewhere between 2 and 40 percentage points.

(c) It opens the possibility of biases. One would need to be assured that the allocation to morning or afternoon lists could not depend in any way on the metabolism of the patient.

12. We will use subscripts *aggress* for the "potentially aggressive" group and *less* for the "less aggressive" group. Let $p_{aggress}$ and p_{less} be the true proportions of the two groups who will be reconvicted for violent offenses within 1 year. We will test $H_0 : p_{aggress} - p_{less} = 0$ (the classification has no predictive value) versus the $H_1 : p_{aggress} - p_{less} > 0$ (those deemed potentially aggressive are indeed more likely to be reconvicted of violent offenses).
The "20%" and "80%" must be whole numbers, thus $n_{aggress} = 1542$ and $n_{less} = 6170$. The sample proportions are $\widehat{p}_{aggress} = 0.0031$ and $\widehat{p}_{less} = 0.0028$. For independent proportions, $\text{se}(p_{aggress} - p_{less}) = \sqrt{\frac{0.0031 \times 0.9969}{1542} + \frac{0.0028 \times 0.9972}{6170}} = 0.001567$. The resulting test statistic is $t_0 = \frac{(0.0031-0.0028)-0}{0.001567} = 0.191$. This tells us that our estimated difference in proportions from the data is only about 0.2 standard errors from zero! We know immediately that there is no evidence of a true difference. Continuing with the standard pattern, the (1-tailed)P-value is $\text{pr}(Z \geq 0.191) = 0.42$ and the 95% CI for the true difference, $p_{aggress} - p_{less}$, is $[-0.0028, 0.0034]$.

The classification system has not been demonstrated to have any ability to discriminate as to who is likely to reoffend. It is clearly not good enough to be useful in practice.
[Warning: Sample sizes are too small for the 10% rule.]

13. **(a)** There are substantial proportions of reoffenders in both groups.

(b) We expect bias against the classification system as not paroling the "worst" prisoners should lower the reoffending rate in the high risk group and make the rates in the 2 groups more similar.

(c) Longer followup times would lead to higher proportions reoffending in both groups.

14. **(a)** We will test $H_0 : p = 0.28$ (smoking rates among football-club members are no different from those in the general population) versus the 2-sided alternative $H_1 : p \neq 0.28$ (as there is no description of a research hypothesis suggesting a particular direction). From our sample of size $n = 130$ football-club members, we have $\widehat{p} = 0.32$. This has standard error $\text{se}(\widehat{p}) = \sqrt{\frac{0.32 \times 0.68}{130}} = 0.04091$. The resulting test statistic is $t_0 = \frac{0.32 - 0.28}{0.04091} = 0.978$. The sample proportion for football-club members is less than 1 standard error from the general population figure. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 0.978) = 0.33$. We have no evidence against H_0, or more concretely, we have no evidence that the percentage smokers among football-club members differs from that for the general population. A 95% CI for the true difference is [0.24, 0.40].

(b) Let p_{male} and p_{female} be the true proportions who do not participate in sports etc. for males and females, respectively. We will test $H_0 : p_{male} - p_{female} = 0$ (no sex difference) versus the 2-sided alternative $H_1 : p_{male} - p_{female} \neq 0$ (as there is no description of a research hypothesis suggesting a particular direction). From our data we have sample proportions $\widehat{p}_{male} = 0.116$ and $\widehat{p}_{female} = 0.089$ from samples of size $n_{male} = 1300$ and $n_{female} = 1300$. For independent proportions, $\text{se}(\widehat{p}_{male} - \widehat{p}_{female}) = \sqrt{\frac{0.116 \times 0.884}{1300} + \frac{0.089 \times 0.911}{1300}} = 0.011885$. The resulting test statistic is $t_0 = \frac{(0.116 - 0.089) - 0}{0.011885} = 2.2718$. This tells us that our estimated difference in proportions from the data is more than 2.2 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 2.2718) = 0.023$. We do have some evidence against H_0, i.e., we do have some evidence of a true difference (more male than female participation).
The 95% CI for the true difference, $p_{male} - p_{female}$, is [0.004, 0.050]. With 95% confidence, the true percentage nonparticipation for males is higher than it is for females by somewhere between 0.4 and 5 percentage points.

(c) Let p_{rural} be the true proportion of rural people belonging to sports or recreation clubs and p_{urban} be the corresponding proportion for urban dwellers. We will test $H_0 : p_{rural} - p_{urban} = 0$ versus $H_1 : p_{rural} - p_{urban} \neq 0$ (as there is no description of a research hypothesis suggesting a particular direction).
From our data we have sample proportions $\widehat{p}_{rural} = 0.47$ and $\widehat{p}_{urban} = 0.31$ from samples of size $n_{rural} = 1200$ and $n_{urban} = 1400$. For independent proportions, $\text{se}(\widehat{p}_{rural} - \widehat{p}_{urban}) = \sqrt{\frac{0.47 \times 0.53}{1200} + \frac{0.31 \times 0.69}{1400}} = 0.0189834$. The resulting test statistic is $t_0 = \frac{(0.47 - 0.31) - 0}{0.0189834} 8.4284$. This tells us that our estimated difference in proportions from the data is more than 8 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq = 8.4284) = 0.0000$. We have very strong evidence against H_0, i.e., we have very strong evidence of a true difference (higher participation in rural areas).

The 95% CI for the true difference, $p_{rural} - p_{urban}$, is [0.123, 0.197]. With 95% confidence the true percentage club membership for rural dwellers is higher than it is for urban dwellers by somewhere between 0.4 and 5 percentage points.

15. Let p_{Asian} be the true proportion of Asians voting Republican in 1998 and p_{Hispan} be the corresponding proportion for Hispanics. We will test $H_0 : p_{Asian} - p_{Hispan} = 0$ versus $H_1 : p_{Asian} - p_{Hispan} \neq 0$ (as there is no description of a research hypothesis suggesting a particular direction).
From our data we have sample proportions $\widehat{p}_{Asian} = 0.42$ and $\widehat{p}_{Hispan} = 0.35$ from samples of size $n_{Asian} = 100$ and $n_{Hispan} = 500$. For independent proportions, $\text{se}(\widehat{p}_{Asian} - \widehat{p}_{Hispan}) = \sqrt{\frac{0.42\times0.58}{100} + \frac{0.35\times0.65}{500}} = 0.05376802$.
$t_0 = \frac{(0.42-0.35)-0}{0.05376802} = 1.3019$. This tells us that our estimated difference in proportions from the data is only 1.3 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 1.3019) = 0.193$. We have no evidence of a real difference.
The 95% CI for the true difference, $p_{Asian} - p_{Hispan}$, is $[-0.035, 0.18]$. With 95% confidence, the percent-Republican vote for Asian Americans could be anywhere between 3.5 percentage points lower than it is for Hispanic Americans and 18 percentage points higher.

16. See comments and formulas given at the beginning of our answer to problem 8. We give details of calculations in (a) but in (b) and (c) just report results. We are doing all our tests as 2-tailed. In Review Exercises 8, problem 2, we thought about what should happen in this experiment before we looked at at the results of the experiment. This gives some partial idea of the thinking involved in formulating research hypotheses. If we had obtained data to confirm those prior research hypotheses, we should now be performing 1-tailed tests with the direction taken from the research hypothesis.

Let μ_{enthus} be the true mean team-building score for enthusiastic volunteers, and similarly μ_{reluct} and μ_{nonvol} for reluctant volunteers and nonvolunteers respectively.

(a) We will test $H_0 : \mu_{enthus} - \mu_{reluct} = 0$ versus, $H_1 : \mu_{enthus} - \mu_{reluct} \neq 0$. The sample values from our data are $\overline{x}_{enthus} = 3.52$ and $s_{enthus} = 0.951$ from a sample of size $n_{enthus} = 38$, and $\overline{x}_{reluct} = 3.39$ and $s_{reluct} = 0.831$ from a sample of size $n_{reluct} = 28$. The standard error of the estimated difference, for independent samples, is $\text{se}(\overline{x}_{enthus} - \overline{x}_{reluct}) = \sqrt{\frac{0.951^2}{38} + \frac{0.831^2}{28}} = 0.2201430$. The resulting t-test statistic is $t_0 = \frac{(3.52-3.39)-0}{0.2201430} = 0.5905$. This tells us that our estimated difference in means from the data is only about 0.6 standard errors from zero. Using Student's t distribution with $df = \min(n_{A.no} - 1, n_{A.sim} - 1) = 27$, the (2-tailed) P-value is $2 \times \text{pr}(T \geq 0.5905) = 0.56$. There is no evidence that a real difference exists.
The 95% CI for the true difference, $\mu_{enthus} - \mu_{reluct}$, is $[-0.32, 0.58]$. With 95% confidence, the true mean team-building score for enthusiastic volunteers lies somewhere between being 0.32 units smaller than for reluctant volunteers and being 0.58 units larger.

(b) We will test $H_0 : \mu_{enthus} - \mu_{nonvol} = 0$ versus, $H_1 : \mu_{enthus} - \mu_{nonvol} \neq 0$. We find $\text{se}(\overline{x}_{enthus} - \overline{x}_{nonvol}) = 0.2455952$ and $t_0 = 2.7281$. Using the Student($df = 12$) distribution, the (2-tailed) P-value is 0.018. There is some evidence of a true

difference (enthusiastic volunteers have larger scores on average).

The 95% CI for the true difference, $\mu_{enthus} - \mu_{nonvol}$, is [0.13, 1.21]. With 95% confidence, the true mean team-building score for enthusiastic volunteers is larger than for nonvolunteers by somewhere between 0.13 and 1.21 units.

(c) We will test $H_0 : \mu_{reluct} - \mu_{nonvol} = 0$ versus, $H_1 : \mu_{reluct} - \mu_{nonvol} \neq 0$. We find $\text{se}(\overline{x}_{reluct} - \overline{x}_{nonvol}) = 0.2473457$ and $t_0 = 2.1832$. Using the Student($df = 12$) distribution, the (2-tailed) P-value is 0.05. There is evidence that a real true difference exists.

The 95% CI for the true difference, $\mu_{reluct} - \mu_{nonvol}$, is [0.001, 1.08]. With 95% confidence, the true mean team-building score for reluctant volunteers is larger than for nonvolunteers by somewhere between 0.001 points (or almost nothing) and 1.08 points.

17. (a) Let p_{sinpar} be the true proportion of single parents who are stressed by relationships with parents and let p_{alone} be the corresponding proportion for those living alone. We will test $H_0 : p_{sinpar} - p_{alone} = 0$ versus $H_1 : p_{sinpar} - p_{alone} \neq 0$.
From the data we have the sample proportions $\widehat{p}_{sinpar} = 0.129$ from a sample of size $n_{sinpar} = 575$ and $\widehat{p}_{alone} = 0.103$ from a sample of size $n_{alone} = 875$. As we are comparing proportions from two independent samples (situation (a) in Fig. 8.5.1), we have $\text{se}(\widehat{p}_{sinpar} - \widehat{p}_{alone}) = \sqrt{\frac{0.129 \times 0.871}{575} + \frac{0.103 \times 0.897}{875}} = 0.017349$. The resulting test statistic is $t_0 = \frac{(0.129 - 0.103) - 0}{0.017349} = 1.4986$. This tells us that the estimated difference in proportions from our data is about 1.5 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 1.4986) = 0.13$. We have no evidence that a true difference exists.

The 95% CI for the true difference, $p_{sinpar} - p_{alone}$, is $[-0.008, 0.060]$. With 95% confidence, the true percentage stressed by relationships with parents for single parents is somewhere between being about 1 percentage point smaller than for those living alone and being 6 percentage points larger.

(b) Here we are only looking at those living as single parents. Let p_{smoke} be the true proportion of them who smoke and let $p_{unhealthy}$ be the true proportion with unhealthy eating practices. We will test $H_0 : p_{smoke} - p_{unhealthy} = 0$ versus $H_1 : p_{smoke} - p_{unhealthy} \neq 0$.
We have data on a sample of size $n = 496$ for which the sample proportions are $\widehat{p}_{smoke} = 0.541$ and $\widehat{p}_{unhealthy} = 0.432$. We are performing a situation (c) comparison from Fig. 8.5.1 so $\text{se}(\widehat{p}_{smoke} - \widehat{p}_{unhealthy}) = \sqrt{\frac{0.541 + 0.432 - (0.541 - 0.432)^2}{575}} = 0.040884$. The resulting test statistic is $t_0 = \frac{(0.541 - 0.432) - 0}{0.040884} = 2.6661$. This tells us that the estimated difference in proportions from our data is more than 2.6 standard errors from zero. The (2-tailed) P-value is $2 \times \text{pr}(Z \geq 2.6661) = 0.008$. We have strong evidence against H_0, i.e., we have strong evidence that a true difference exists (more likely to smoke than have unhealthy eating practices).

The 95% CI for the true difference, $p_{smoke} - p_{unhealthy}$, is [0.03 , 0.19]. With 95% confidence the true percentage who smoke is higher than the percentage who would report unhealthy eating practices by somewhere between 3 and 19 percentage points.

(c) Here we are only looking at those living with a partner and child(ren). Let p_{underw} be the true proportion of them falling into the underweight category and let p_{overw} be the true proportion falling into the overweight category. We will test $H_0 : p_{underw} - p_{overw} = 0$ versus $H_1 : p_{underw} - p_{overw} \neq 0$.
We have data on a sample of size $n = 915$ for which the sample proportions are $\widehat{p}_{underw} = 0.253$ and $\widehat{p}_{overw} = 0.216$. We are performing a situation (b) comparison from Fig. 8.5.1, so $\text{se}(\widehat{p}_{underw} - \widehat{p}_{overw}) = \sqrt{\frac{0.253+0.216-(0.253-0.216)^2}{915}} = 0.022607$. The resulting test statistic is $t_0 = \frac{(0.253-0.216)-0}{0.022607} = 1.6367$. This tells us that the estimated difference in proportions from our data is about 1.6 standard errors from zero. The (2-tailed) P-value is (2-tailed) P-value $= 2 \times \text{pr}(Z \geq 1.6367) = 0.10$. We have only weak evidence of a true difference.

The 95% CI for the true difference, $p_{underw} - p_{overw}$, is $[-0.007, 0.081]$. With 95% confidence, the true percentage who are underweight is somewhere between being 0.7 percentage points smaller than the percentage who are overweight and being 8 percentage points larger.

18. **(a)** True.

(b) **(i)** True (interpreted as statistically significant). **(ii)** No, the gender difference is small compared with person-to-person variability.

(c) **(i)** True (interpreted in terms of statistical significance). **(ii)** False. Nonsignificance does not demonstrate "no effect".

(d) **(i)** True. **(ii)** True.

(e) False. The P-value relates to the probability of a difference *when* chance is the cause, not the probability that chance caused the difference.

(f) True.

(g) True.

(h) **(i)** True. **(ii)** True. **(ii)** False. We can reduce sampling variation, but not the non-sampling errors. The latter are harder to control in very large studies.

19. **(a)** If people were just guessing, the chances of identifying the one of the three slices that was different would be one in three.

(b) We test $H_0 : p = \frac{1}{3}$ versus $H_1 : p > \frac{1}{3}$ (there is some ability to discriminate). We have $\widehat{p} = \frac{16}{27} = 0.5925926$, $\text{se}(\widehat{p}) = \sqrt{\frac{0.5925926 \times 0.4074074}{27}} = 0.094561$, and $t_0 = \frac{0.59259-0.33333}{0.094561} = 2.7417$. The sample proportion of correct identifications is over 2.7 standard errors above $\frac{1}{3}$. The (1-tailed) P-value is $\text{pr}(Z \geq 2.7417) = 0.003$. We have strong evidence against H_0, i.e., strong evidence that the true proportion of correct identifications is greater than "just guessing".
(Warning: The sample size is too small for this large sample theory.)

(c) P-value tells us that we have strong evidence that the identification rate is better than 1/3. The magazine has got it wrong.

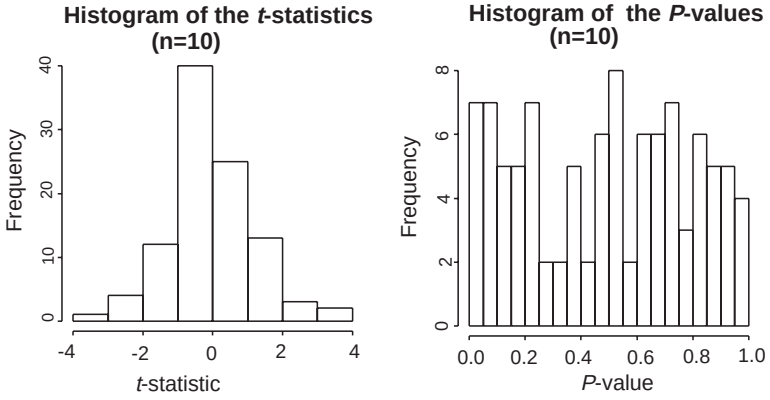
(d) Possible differences in appearance can be catered for by using blindfolds. There is the possibility of learning over the 3 attempts so we could have more people and make only one identification each. Other ideas?

(e) If you use $H_0 : p = 1/2$ the result is not significant.

***20.** **(a)** Using $Y \sim \text{Binomial}(n = 20, p = 0.2)$, $\text{pr}(Y \geq 8) = 0.032$.

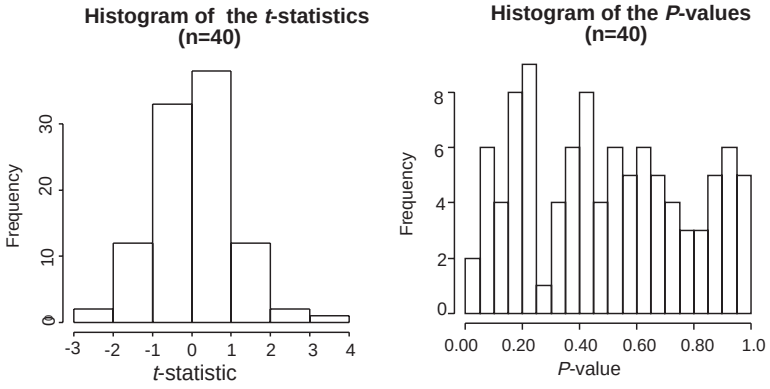
(b) For $Y \sim \text{Binomial}(n = 27, p = 1/3)$, $\text{pr}(Y \geq 16) = 0.005$. This P-value leads us to the same conclusions as we reached in problem 19 using large-sample theory.

21. **(a)** One hundred samples, each of size $n = 10$, were generated under circumstances in which the null hypothesis was true. For each sample the t-statistic and the P-value for testing $H_0 : \mu = 5.517$ were obtained. A histogram of the t_0 values is shown below left and a histogram of the P-values is shown below right.



Our histogram of t_0-values is centered at about 0 (with a reasonably symmetric bell shape). When H_0 is true, P-values less than or equal to 0.05 occur 5% of the time over the long run. The proportion of our 100 P-values that was less than 0.05 was $\frac{7}{100} = 0.07$ or 7%. Your results will be somewhat different.

(b) We repeated (a) using 100 samples each of size $n = 40$. A histogram of the t_0 values is shown below left and a histogram of the P-values is shown below right.

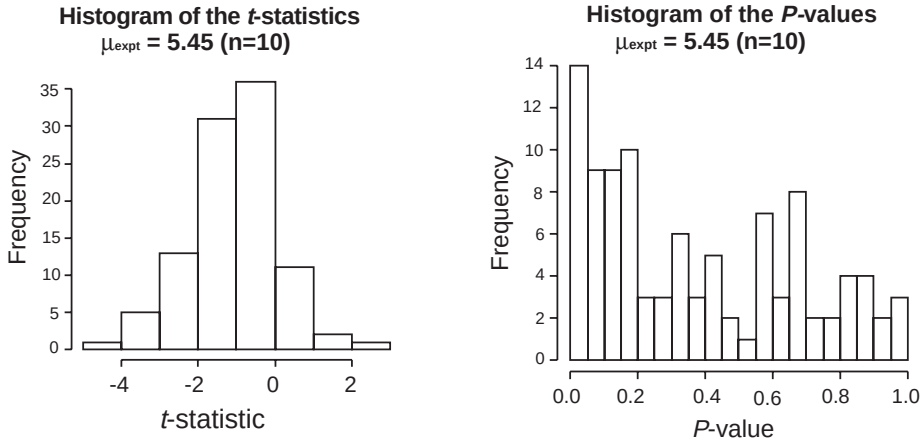


Our histogram of t_0 values is centered at about 0, bell shaped and looks somewhat right skewed. (We might have expected it would be more symmetric – see Note 1 to follow.) The proportion of our 100 P-values less than or equal to 0.05 was $\frac{2}{100} = 0.02$. Your results will be somewhat different.

Notes: We make the folowing points about (a) and (b).

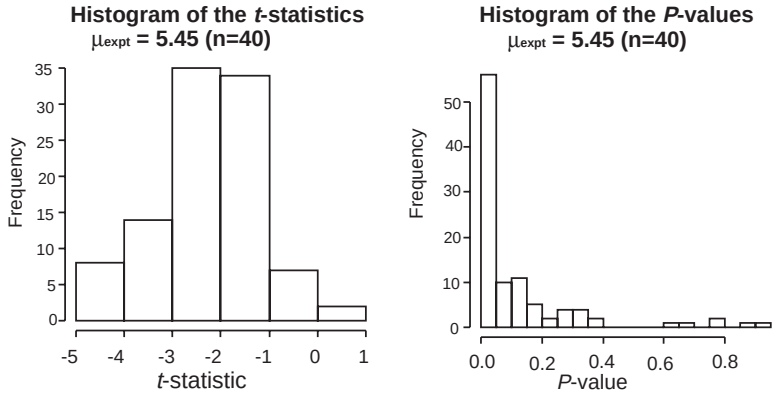
1. The reason our histograms are not necessarily symmetric like Student's t-distribution (your one might be) is that we are only using 100 values and there is quite a bit of variation, from histogram to histogram, in histograms of 100 values. (Some are given at the end of this set of answers for comparative purposes.) If we had used t_0 values from hundreds of thousands of samples, our histogram would look like a t distribution.
2. It can be shown that when H_0 is true, the P-value is equally likely to fall anywhere between 0 and 1 (technically they have a Uniform(0,1) distribution) with 5% of them falling below 0.05 in the long run. Our histograms in (a) and (b) do look like histograms of samples of size 100 from the Uniform distribution. Some are given at the end of this set of answers for comparative purposes.

(c) *Samples of size 10*: 100 samples of size $n = 10$ with $\mu_{\text{expt}} = 5.45$ were generated (i.e., H_0 is false in that the experiment is slightly biased). For each sample the t-statistic and the P-value for testing $H_0 : \mu = 5.517$ were obtained. A histogram of our 100 t_0 values is shown below left and a histogram of our P-values is shown below right.



We see that the distribution of t_0 values is no longer centered at 0, but is now centered at approximately -1. The distribution of P-values is no longer uniform in shape but is now negatively skewed and beginning to stack up against the left hand side of the plot. The proportion of our P-values that were less than or equal to 0.05 is now bigger at $\frac{14}{100} = 0.14$ (cf. 0.05) but still fairly small.

Samples of size 40: We repeated the above experiment using samples of size $n = 40$ under exactly the same conditions. Histograms of the t_0 values (below left) and the P-value (below right) for each sample for testing $H_0 : \mu = 5.517$ follow.

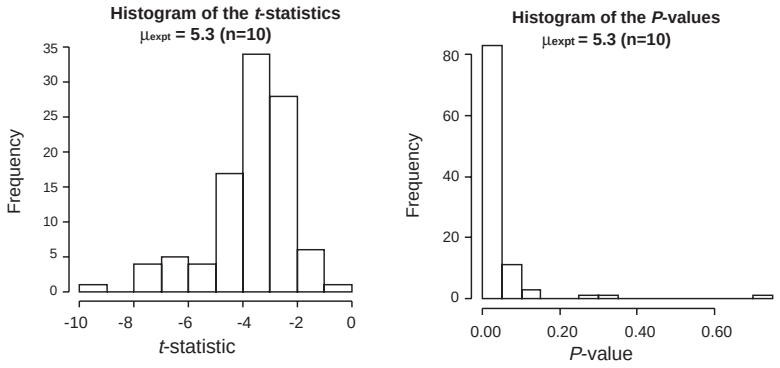


We see that the distribution of t_0 values is now centered at approximately -2.5. Note also how the histogram has become very skewed and stacked up against the left-hand side of the plot. The proportion of the P-values less than or equal to 0.05 is much bigger at $\frac{56}{100} = 0.56$ or nearly 60%.

The intended lesson is that it is easier to detect departures from a null hypothesis with larger samples.

(d) We now shift the true value of μ even further away from the hypothesized value.

Samples of size 10: Histograms of our 100 t_0 values and the P-values from the 100 samples are given below.

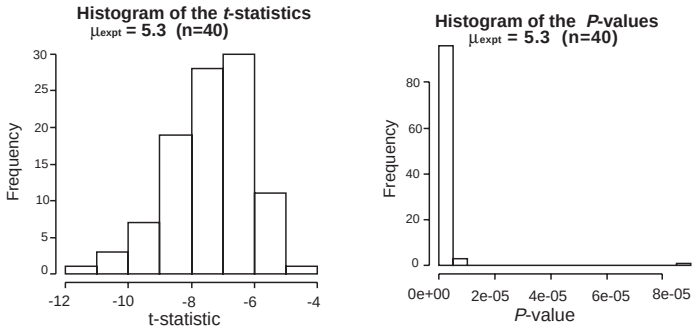


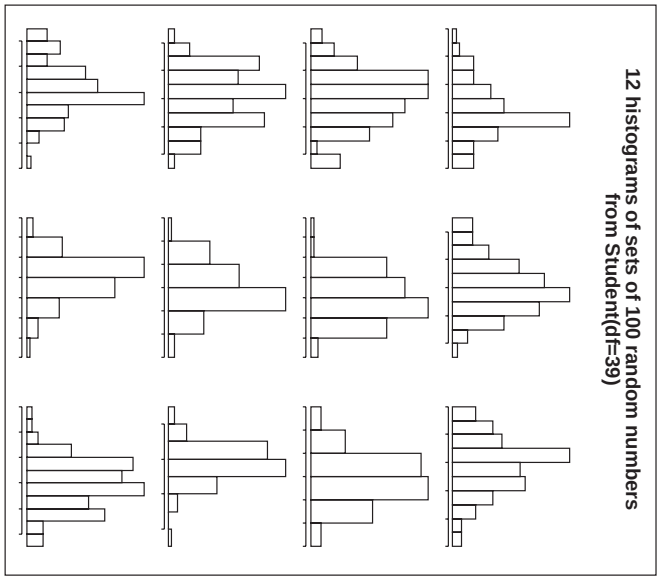
We should compare these plots with our other plots for $n = 10$. The distribution of t_0 values has moved further to the left (now centered around approximately -3.5), the P-value histogram is stacked more strongly against the left-hand side and the proportion of our P-values less than or equal to 0.05 is $\frac{83}{100} = 0.83$.

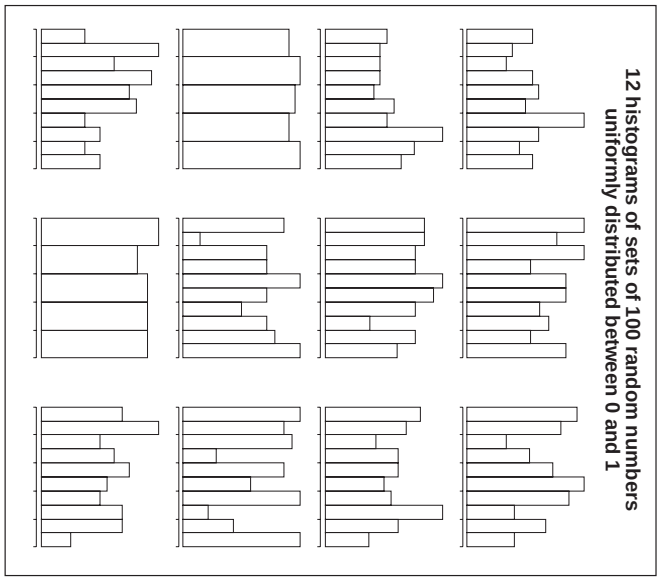
Samples of size 40:

We should compare these plots (given below) with our other plots for $n = 40$. The distribution of t_0 values has moved further to the left (now centered around approximately -7), the P-value histogram is stacked more strongly against the left-hand side and *all* of our P-values were less than or equal to 0.05.

The intended lesson is that it is easier to detect larger departures from a null hypothesis than it is to detect smaller ones. It is also easier with larger samples.


Histogram of the t-statistics
μexpt = 5.3 (n=40)
Frequency
t-statistic
Histogram of the P-values
μexpt = 5.3 (n=40)
Frequency
P-value


12 histograms of sets of 100 random numbers from Student(df=39)

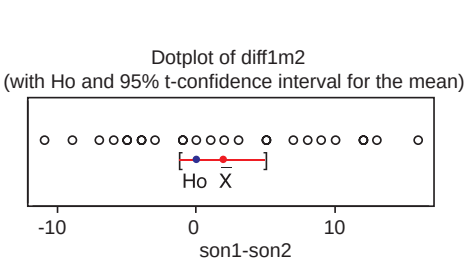

12 histograms of sets of 100 random numbers uniformly distributed between 0 and 1

Chapter 10

All answers in this chapter were computed using Minitab.

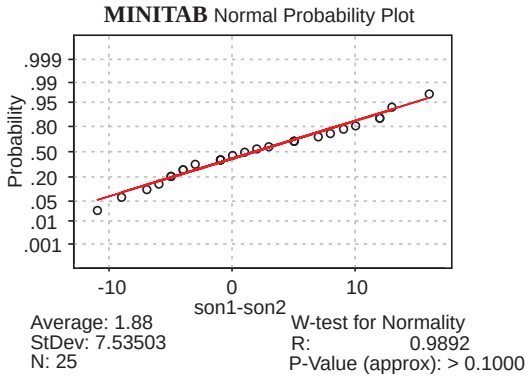
Exercises for Section 10.1.2

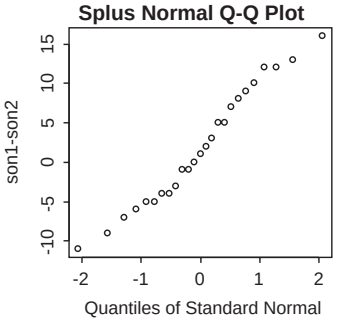
1. **(a)** We plot the differences (son1-son2). The following dot plot or stem-and-leaf plot do not show up any unusual points, though the data tends to be fairly uniformly spread. However, the Normal probability plot is close to a straight line and the W-test shows no evidence of non-Normality (P-value > 0.1).



Character Stem-and-Leaf Display

```
Stem-and-leaf of diff1m2   N  = 25
Leaf Unit = 1.0

 1   -1 1
 6   -0 97655
11   -0 44311
(4)   0 0123
10    0 55789
 5    1 0223
 1    1 6
```





Note: We have included another Normal probability plot (from Splus). Here the data axis is the vertical axis and the Normal distribution axis is the horizontal axis. This is the reverse of the Minitab plot. We have done this to illustrate that there are differences between packages in the way they orient their Normal probability plots. Apart from the choice and labelling of axes they are, however, the same type of plot.]

(b) Let μ_{diff} be the population mean of the differences. We wish to test $H_0 : \mu_{diff} = 0$ versus $H_1 : \mu_{diff} \neq 0$. Using a paired-comparison t-test, $t_0 = 1.25$ and P-value $= 0.22$, i.e., no evidence against H_0. There is no evidence of a difference between the head lengths. Assuming Normality, a 95% t-confidence interval for μ_{diff} is $[-1.23, 4.99]$, so at this level of confidence, the true mean difference could be anywhere between -1.23 (1st sons smaller) and 4.99 (1st sons larger). This information is depicted on the dot plot above.

(c) How were the families selected? How were the measurements taken? Was a standard procedure strictly followed?

2. People would vary in how they administered the procedure. The size of any systematic difference between the two sets of calipers will vary with how the head measurement is taken and from what part of the head it is taken. As the cardboard calipers wear, they will tend to give bigger measurements.

Exercises for Section 10.1.3

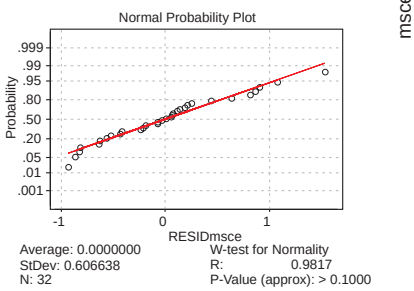
1. Let $\tilde{\mu}_{diff}$ be the population median of the differences. We wish to test $H_0 : \tilde{\mu}_{diff} = 0$ versus $H_1 : \tilde{\mu}_{diff} \neq 0$. Using a sign test we have 13 plus signs, 11 minus signs, and 1 zero. Intuitively such a result is not significant. (Think about tossing a fair coin.) In fact, P-value $= 0.84$. There is no evidence of a difference, i.e., no evidence of a difference in head length. A sign 95% confidence interval for the true median difference $\tilde{\mu}_{diff}$ is $[-3.80, 6.60]$.

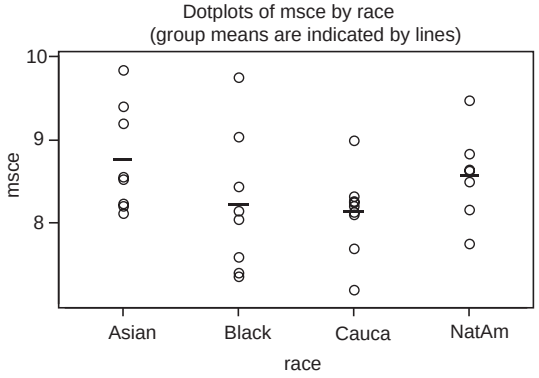
2. Let $\tilde{\mu}$ be the median score. We wish to test $H_0 : \tilde{\mu} = 28$ versus $H_1 : \tilde{\mu} \neq 28$. Using the sign test, we have 10 plus signs, and 4 minus signs with P-value $= 0.18$. We have no evidence against H_0, i.e., no evidence that cyclozocine is an effective treatment. A sign 95% confidence interval for $\tilde{\mu}$ is $[27, 51]$ so that with 95% confidence, the true median is somewhere between 27 and 51. Note that the interval contains the hypothesized value of 28.

Exercises for Section 10.3

1.

```
Analysis of Variance for msce
Source   DF      SS      MS    F    P
race      3   2.136   0.712 1.75 0.180
Error    28  11.408   0.407
Total    31  13.544

Level       N     Mean    StDev
Asian       8   8.7587   0.6397
Black       8   8.2237   0.8389
Cauca       9   8.1311   0.4851
NatAm       7   8.5729   0.5377
```

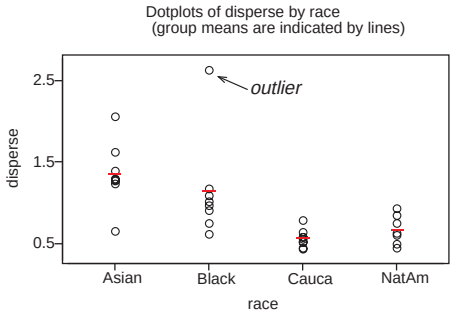


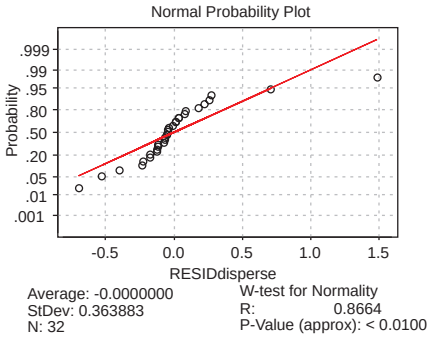


For the MSCE data we test H_0: population means all equal versus H_1: population means not all equal. Using an F-test, $f_0 = 1.75$ and P-value $= 018$ (see the printout above). There is no evidence against H_0, that is, no evidence of racial differences. The dot plots indicate that the four samples have acceptably similar spreads (the standard deviations range from 0.49 to 0.84). The (combined) Normal probability plot of the residuals is closely linear (apart from displaced end points, which is not atypical of Normal plots; see Fig. 10.1.3). The W-test has P-value > 0.1 providing no evidence of non-Normality.

2.

```
Analysis of Variance for disperse
Source     DF        SS        MS        F        P
race        3     3.473     1.158     7.90    0.001
Error      28     4.105     0.147
Total      31     7.578

                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level       N      Mean     StDev  ---+---------+---------+---------+---
Asian       8    1.3563    0.3933                      (------*------)
Black       8    1.1463    0.6259                 (------*------)
Cauca       9    0.5667    0.1063   (-----*------)
NatAm       7    0.6743    0.1774    (-------*------)
                                    ---+---------+---------+---------+---
Pooled StDev =   0.3829              0.40      0.80      1.20      1.60
```





For the DISPERSION data we wish to test H_0: population means all equal versus H_1: population means not all equal. Using an F-test, $f_0 = 7.90$ and P-value $= 0.001$. There is very strong evidence of racial differences. Looking at the 95% confidence intervals for the four individual means in the computer printout above, we see that the Asian confidence interval does not overlap with the Caucasian or Native American confidence intervals, and the Black confidence interval does not overlap with the Caucasian. We will not go any further with this analysis because of the presence of the outlier labelled in the dot plot, and worries about differences in spreads between the groups.

3. The dot plot above shows a high outlier at 2.63. The numerator of the F-test measures how far apart the sample means are. Removing the outlier will reduce the mean of the Black group. This will move three of the means closer together, thus reducing the numerator. However, removing the outlier will substantially reduce the internal variation of the Black data thus reducing the denominator. Since means are less sensitive than standard deviations to outliers, the F-ratio might be expected to increase, though it is hard to tell.

4. We have the following computer printout when the outlier is removed.

```
Analysis of Variance for disperse
Source     DF        SS        MS        F        P
race        3    3.0148    1.0049    17.08    0.000
Error      27    1.5887    0.0588
Total      30    4.6035

                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level       N      Mean     StDev  ---------+---------+---------+-------
Asian       8    1.3563    0.3933                          (----*----)
Black       7    0.9343    0.1943              (-----*----)
Cauca       9    0.5667    0.1063  (----*----)
NatAm       7    0.6743    0.1774     (----*-----)
                                    ---------+---------+---------+-------
Pooled StDev =   0.2426             0.70      1.05      1.40
```

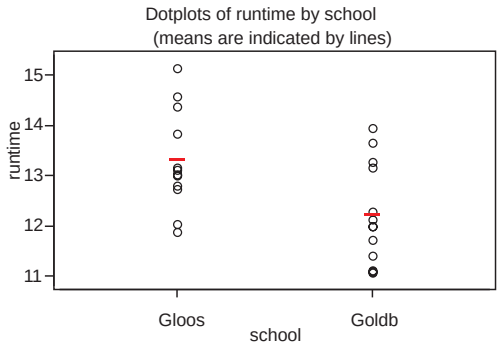
We see that $f_0 = 17.08$ and P-value $= 0.000$, again indicating very strong evidence of racial differences. However, the individual 95% confidence intervals for the Asian and Black groups no longer overlap, so that the Asian group is clearly different from the other three. The value of f_0 has increased, as suggested in 3.

5. The spread for the Asian group is much greater than that for the other three, which are quite similar. The standard deviation for the Asian group is 0.3933 and that for the Caucasion group is 0.1063, a ratio of nearly 4. The F-test and confidence intervals for differences between the means may be of doubtful validity.

[In fact a Levene test for differences in spread was nonsignificant indicating that the apparent differences in spread could have arisen just through sampling variation.]

Review Exercises 10

1. (a)



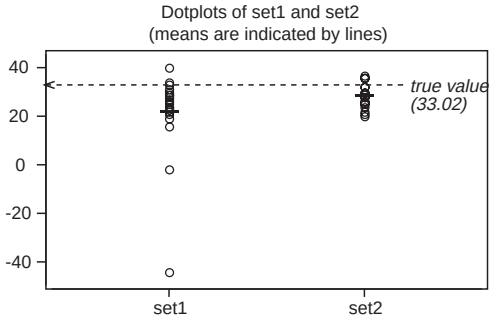
From the dot plots, running times seem longer on average at Glooscap. The spreads look similar. Let μ_{Gloo} and μ_{Cold} be the respective mean running times for Glooscap and Coldbrock. We wish to test $H_0 : \mu_{Gloo} - \mu_{Cold} = 0$ versus $H_0 : \mu_{Gloo} - \mu_{Cold} \neq 0$. Using a Welch two-sample t-test we have P-value $= 0.012$. There is reasonably strong evidence of a difference between the two schools. A 95% confidence interval for the difference in the means is $[0.26, 1.91]$, that is, a difference in true mean running times of between about 0.3 and 1.9 seconds.

The dot plot looks reasonable, the individual Normal probability plots (not shown) look reasonably linear and both groups give P-values > 0.1 on a W-test for Normality. The Normal theory methods appear to be applicable.

***(b)** Let $\tilde{\mu}_{Gloo}$ and $\tilde{\mu}_{Cold}$ be the respective median running times. We wish to test $H_0 : \tilde{\mu}_{Gloo} - \tilde{\mu}_{Cold} = 0$ versus $H_0 : \tilde{\mu}_{Gloo} - \tilde{\mu}_{Cold} \neq 0$. The Mann-Whitney (Wilcoxon) test gives P-value $= 0.036$, which provides some evidence of a school difference. An approximate 95% confidence interval for the difference in true (or population) medians is $[0.16, 1.92]$.

(c) The problem here is that we have an observational study, not an experiment, so that we cannot prove causality, namely, that the coach makes a difference. For example, the better runners might go to Glooscap. (How would you prove that coaching makes a difference?)

2. (a) The dot plots follow. Set 1 appears to have two small outliers which appear to be pulling the mean for the set downwards appreciably.



(b) We wish to test $H_0 : \mu = 33.02$ versus $H_0 : \mu \neq 33.02$. Using a one-sample t-test we have $t_0 = -2.86$ with P-value $= 0.010$, signalling strong evidence against H_0, or strong evidence that the true mean is not 33.02. The 95% confidence interval puts the true mean for this experiment at somewhere between 13.5 and 30.0.

(c) After the two outliers have been removed, $t_0 = -4.61$ and P-value $= 0.000$. H_0 is even more strongly rejected. (Can you think why?) The 95% confidence interval for the true mean is now $[23.8, 29.6]$.

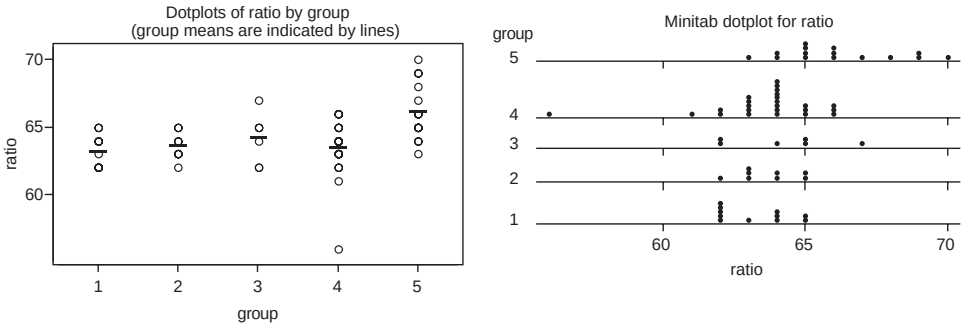
[We note that after the outliers have been removed, the Normal probability plot for set 1 (not shown) looks reasonably linear and the W-test shows no evidence against the Normality assumption.]

(d) We wish to test $H_0 : \mu_1 - \mu_2 = 0$ versus $H_0 : \mu_1 - \mu_2 \neq 0$. A Welch test for the difference of two means gives $t_0 = -1.66$ with P-value $= 0.11$ indicating no evidence of a difference. Without the outliers $t_0 = -1.02$ and P-value $= 0.31$. Again there is no evidence of a true difference.

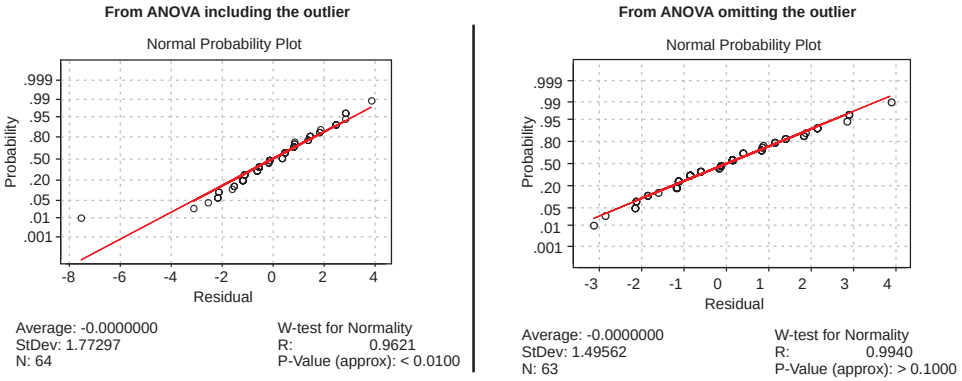
(e) It makes no difference to our conclusions from the test. However it makes a big difference to the 95% confidence interval for the difference in true means, $\mu_1 - \mu_2$, is $[-15.3, 1.7]$ with the outliers included, and $[-5.5, 1.8]$ with the outliers excluded. The upper limit, which tells us how much bigger μ_1 could be than μ_2, has hardly changed. The big change is in the lower confidence limit which tells how much smaller μ_1 could be than μ_2.

3. (a) We have included two sets of dot plots from Minitab. The left-hand set comes from the analysis of variance program and makes no adjustment for overprinting. The right-hand set comes from Minitab's specialist dot plot program and uses

stacking to avoid overprinting. This data is clearly heavily rounded and overprinting is a real problem here. We see that 56 in group 4 is an outlier. Also group 5 has a larger mean and a larger spread than the other groups.



(b) We wish to test H_0 : population group means all equal versus H_1 : population group means not all equal. Using the F-test we have $f_0 = 5.99$ and P-value $= 0.000$. There is very strong evidence of a difference in the group means. The outlier shows up very clearly in the Normal probability plot of the residuals (and is the cause of the significant P-value for the W-test).



Without the outlier we have the following output.

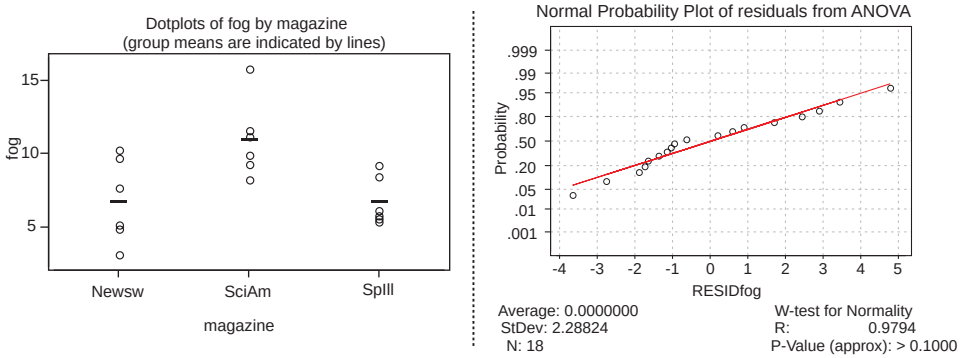
```
Analysis of Variance for ratio
Source     DF        SS        MS        F        P
setup       4     72.17     18.04     7.55    0.000
Error      58    138.69      2.39
Total      62    210.86

                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level       N      Mean     StDev  ------+---------+---------+---------+
1          11    63.182     1.250   (-----*-----)
2           8    63.625     1.061     (------*------)
3           6    64.167     1.941       (--------*-------)
4          23    63.870     1.290         (----*---)
5          15    66.133     2.066                    (----*----)
                                  ------+---------+---------+---------
```

We see that $f_0 = 7.55$ with P-value $= 0.000$. The conclusion that real differences exist between the true means is unchanged. If we leave out the outlier, the 95% confidence interval for the mean of group 5 does not overlap with the other four confidence intervals. The combined Normal probability plot of the residuals

(above) is reasonable, and the maximum ratio of two standard deviations is (just) less than 2.

(c) Without the outlier, Fisher's pairwise comparisons are:
$1-2:[-1.88, 0.10]$, $1-3:[-2.56, 0.59]$, $1-4:[-1.82, 0.45]$, and $1-5:[-4.18, -1.72]$; $2-3:[-2.21, 1.13]$, $2-4:[-1.52, 1.03]$, and $2-5:[-3.86, -1.15]$; $3-4:[-1.12, 1.72]$, $3-5:[-3.46, -0.47]$, and $4-5:[-3.29, -1.24]$.
The group 5 mean is clearly different from the other 4 means. The intervals for differences between the other means contain zero so we cannot demonstrate the existence of real differences. As the confidence intervals show, however, we also cannot rule out the possibility of quite large differences in either direction.

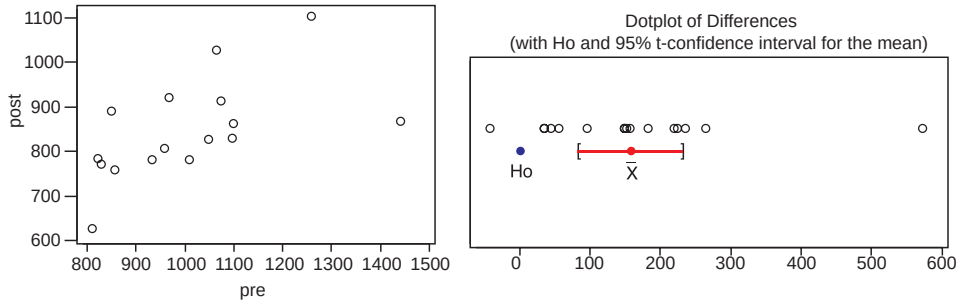
4. The dot plots are given below. Fog indices seem to be greater on average in Scientific American ads than in the other two, which appear roughly similar. The dot plots are reasonable though there is a hint of skewness in Sports Illustrated. However, as seen below, a combined Normal probability plot of the residuals is reasonably linear and the W-test for Normality has P-value > 0.1 showing no evidence of non-Normality.




```
Analysis of Variance for fog
Source     DF        SS        MS        F        P
magazine    2     70.93     35.46     5.98    0.012
Error      15     89.01      5.93
Total      17    159.94

                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level       N      Mean     StDev  --+---------+---------+---------+----
Newsw       6     6.777     2.850   (-------*--------)
SciAm       6    10.968     2.647                    (--------*-------)
SpIll       6     6.738     1.636  (--------*-------)
                                  --+---------+---------+---------+----
Pooled StDev =    2.436            5.0       7.5      10.0      12.5
```

The three standard deviations are reasonably similar, so we can use the F-test for testing H_0: population means all equal versus H_1: population means not all equal. We see that $f_0 = 5.98$ and P-value $= 0.012$, giving strong evidence of a difference in the three magazines. The three 95% confidence intervals for the individual means suggest that Scientific American is different from the other two, which are similar. This is confirmed by Fisher's pairwise comparison intervals $SciAm - Newsweek : [1.20, 7.19]$ and $SciAm - SpIll : [1.23, 7.23]$. We can read the latter interval, for example, as telling us with 95% confidence, that the mean fog index for Scientific American ads is larger than that for Sports Illustrated ads by somewhere between approximately 1.2 and 7.2.

5. (a)

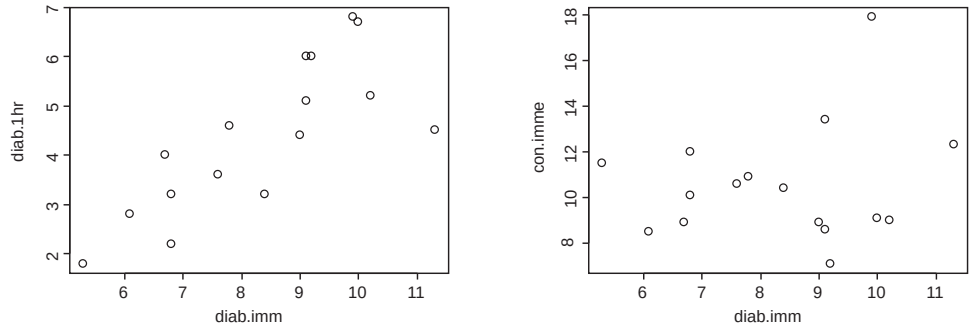


Using a scatter plot, we see that the poststerilization-factor-V level tends to get larger as the presterilization-factor-V level gets larger. There is a definite upward trend.

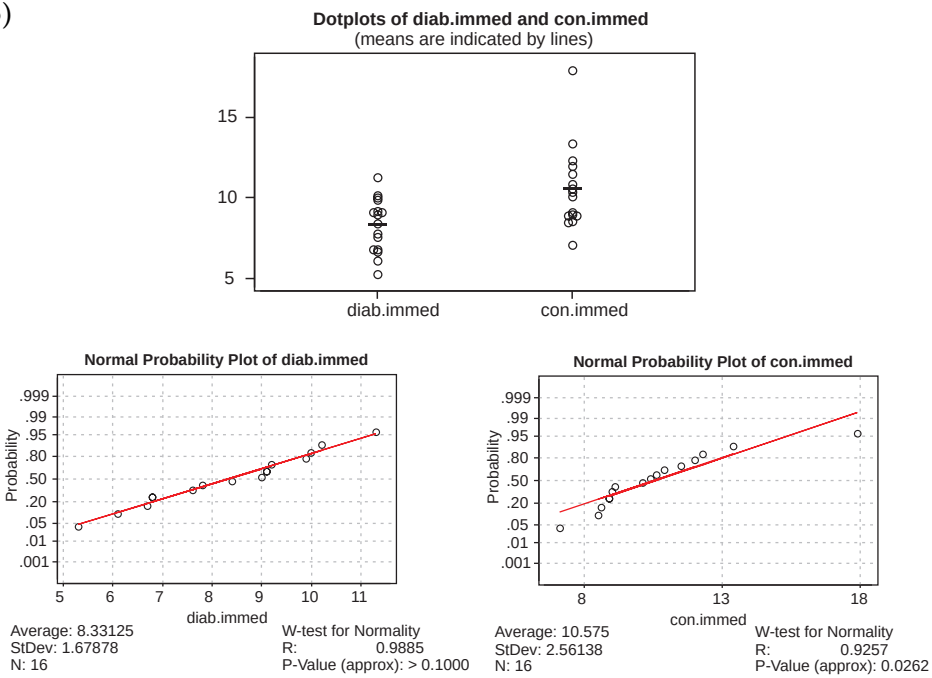
(b) We use the paired comparison method. If $diff = pre - post$ we wish to test $H_0 : \mu_{diff} = 0$ versus $H_1 : \mu_{diff} \neq 0$. We use a two-sided test as there is no suggestion that there was research hypothesis that predicted a direction of difference. Using a t-test, $t_0 = 4.50$ with P-value $= 0.000$. There is very strong evidence that sterilization makes a difference. A 95% confidence interval for the true mean difference, μ_{diff}, is $[82.9, 232.2]$. Since $diff = pre - post$ gives the reduction in factor V with sterilization, we can say with 95% confidence that sterilization decreases factor V levels by somewhere between 83 and 230 units, on average. At least that would be our conclusion if we were happy with the way the data looked.

(c) A dot plot of the differences (above) indicates an outlier. It also shows up very clearly in the Normal probability plot (not shown) which has a P-value of approximately 0.01. After removing the outlier (donor number 16) the Normal probability plot and W-test become satisfactory (not shown). Retesting without the outlier gives us $t_0 = 5.62$ with P-value $= 0.000$, so that there is no change in our conclusion about the existence of a difference. However, the confidence interval for the true difference is now $[80.4, 179.6]$, which is a lot shorter. We can say with 95% confidence that sterilization decreases factor V levels by somewhere between 80 and 180 units.

6. (a) We would expect a relationship between columns 2 and 3 as they are measurements on the same person. The scatter plot (below left) shows this trend. However we would expect the measurements in columns 2 and 5 to be independent as they come from different people. The corresponding scatter plot (below right) shows little evidence of a trend. (We do, however, see an apparent outlier).



(b)



From the dot plot, it appears that measurements taken immediately are larger on average for control subjects than diabetics. The spreads appear similar and we note a large outlier in the control group, which also shows up in the Normal probability plot and W-test (above, left). There are no indications of any problems with the diabetic group.

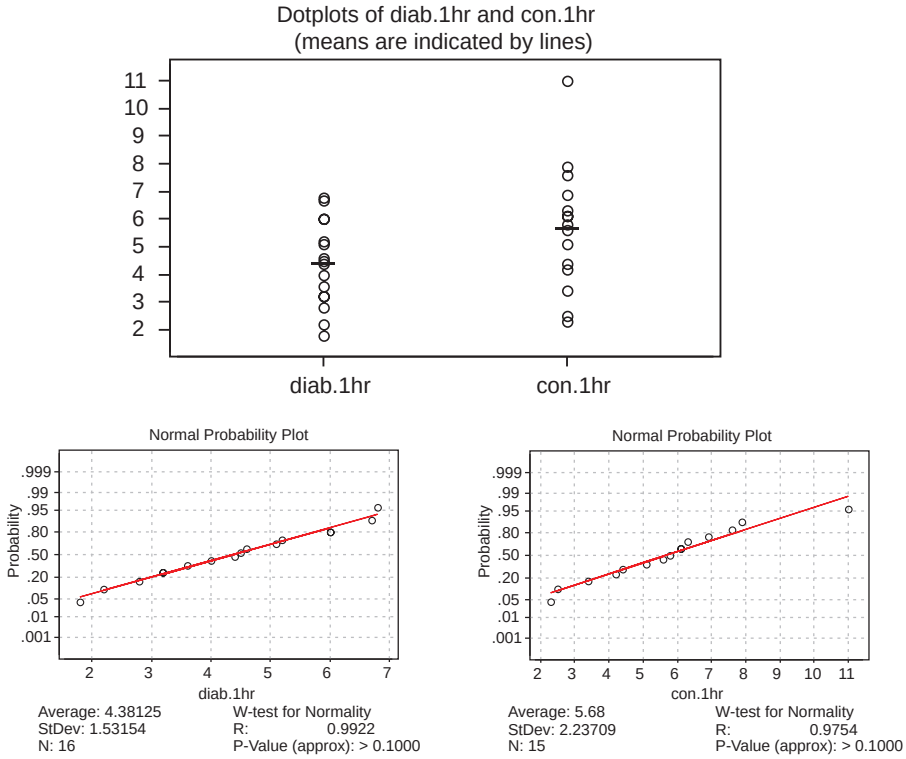
A Welch test for a difference in true means immediately after the removal of the lenses, i.e., of $H_0 : \mu_{diab.immed} - \mu_{con.immed} = 0$ versus $H_1 : \mu_{diab.immed} - \mu_{con.immed} \neq 0$, gives $t_0 = -2.93$ with P-value $= 0.0071$. There is strong evidence of a difference. A 95% confidence interval for the true difference in means is $[-3.82, -0.67]$.

Removing the outlier does not greatly affect the outcome of the test or the confidence interval. For the t-test, $t_0 = -2.88$ with P-value $= 0.0076$, and the 95% confidence interval for the true difference, $\mu_{diab.immed} - \mu_{con.immed}$, is $[-3.01, -0.51]$. With 95% confidence, immediately after removal of the lenses,

true mean percent-swelling is smaller for diabetics than for the control population by somewhere between 0.5 and 3.0 percentage points.

***(c)** Using a Mann-Whitney test, P-value $= 0.013$, and we have an approximate 95% confidence interval of $[-3.3, -0.5]$. The conclusions do not change.

(d) From the following dot plot there seems to be an outlier in the control group showing up both in the dot plot and the Normal probability plot.



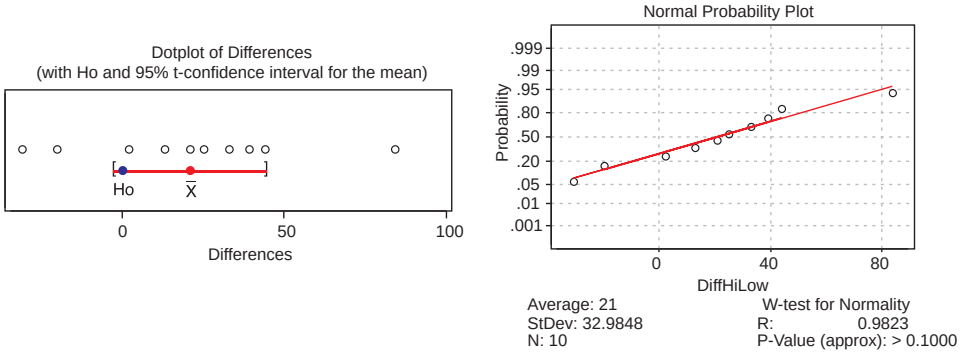
After omitting the outlier, both Normal probability plots are now satisfactory, and the Welch test of $H_0 : \mu_{diab.1hr} - \mu_{con.1hr} = 0$ versus $H_1 : \mu_{diab.1hr} - \mu_{con.1hr} \neq 0$, yields $t_0 = -1.87$ with P-value $= 0.073$. There is evidence for a true difference is rather weak. The 95% confidence interval for the difference in true means, $\mu_{diab.1hr} - \mu_{con.1hr}$, is $[-2.73, 0.13]$. With 95% confidence, the mean percent-swelling for diabetics after 1 hour could be anywhere between being 2.7 percentage points smaller than the mean for controls and being 0.13 percentage points larger. (It might be considerably smaller but is unlikely to be much larger.)

(e) Method of paired comparisons.

(f) It could be the effect of normal variation and/or be due to measurement error. Alternatively, their eyes may have been slightly swollen at the beginning of the experiment.

7. (a) We use the method of paired comparisons, as we have measurements on the same brand. Let $diff = high - low$. We wish to test $H_0 : \mu_{diff} = 0$ versus $H_1 : \mu_{diff} > 0$. Using a one-sample t-test, $t_0 = 2.01$ with a (one-sided) P-value of 0.037. There

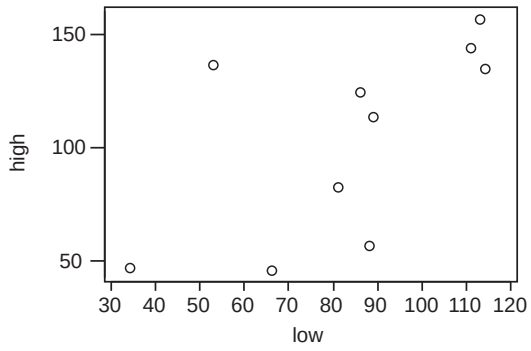
is some evidence against H_0, i.e., or some evidence that high-recall commercials do tend to have higher activity indices.



The dot plot, Normal probability plot and the W-test indicate that Normality is a reasonable assumption.

(b) No, as all that is established is that a difference in mean activity levels between high and low-recall commercials exists. This does not establish that the relationship between activity and recall is very close. We note that two brands actually had negative differences.

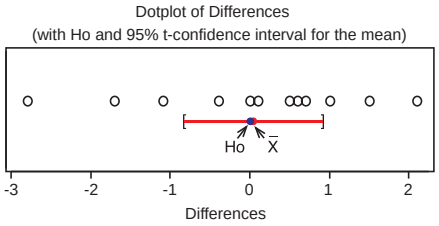
(c) The following scatter plot shows that there is a weak relationship (upwards trend) which seems to be almost nonexistent for the seven observations closest to the origin.

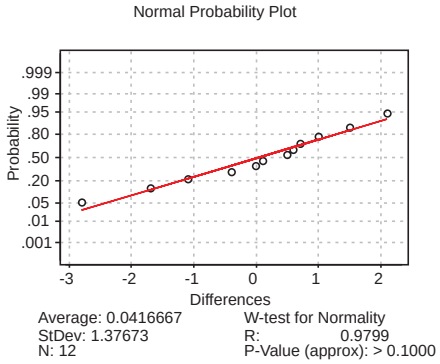


(d) Through randomization one can try and eliminate any systematic bias due to the order in which the ads are seen, e.g., effects due to experimental subjects becoming more tired or inattentive over time.

8. (a) Let *dec.air* be the decrease over the 4 minutes when air is breathed, and let *dec.ox* be the corresponding decrease with pure oxygen. The values of *dec.air* are obtained by taking column 2 minus column 3, and the values of *dec.ox* are obtained by taking column 4 minus column 5. We now apply the paired comparison method to the difference. Let $diff = dec.air - dec.ox$. Then we wish to test $H_0 : \mu_{diff} = 0$ versus $H_0 : \mu_{diff} \neq 0$. The one-sample t-test gives $t_0 = 0.1$ with P-value $= 0.92$. There is no evidence of a difference. The following dot plot,

Normal probability plot and the W-test of Normality indicate that the Normal assumption is reasonable.

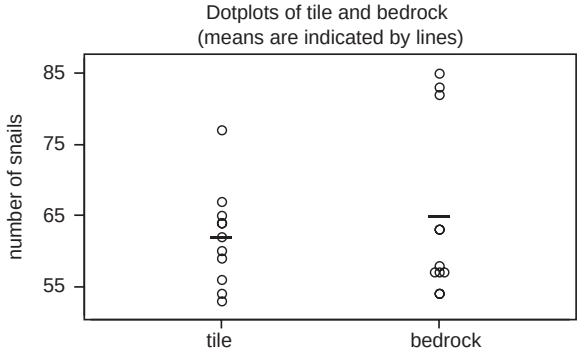




(b) Not quite. There is no evidence of a difference. It doesn't mean, however, that one does not exist. A 95% confidence interval for the true difference is given by $[-0.83, 0.92]$.

(c) It would be of interest to choose periods other than just 4 minutes. Also other athletes could be used.

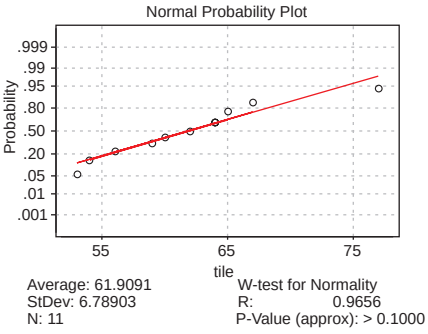
9. (a)

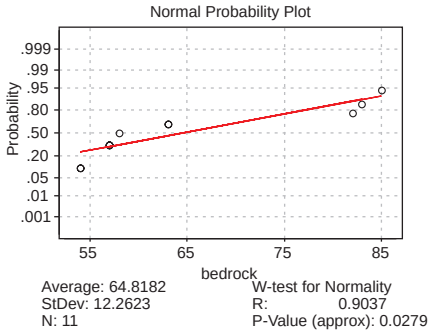


The dot plot suggests that average numbers of snails on bedrock might possibly be greater than on tile.

Using a Welsh two-sample t-test to test $H_0 : \mu_{tiles} - \mu_{bedr} = 0$ versus $H_1 : \mu_{tiles} - \mu_{bedr} \neq 0$, we obtain $t_0 = -0.69$ and P-value $= 0.5$ providing no evidence of a true difference.

From the dot plot, there is a hint of an outlier in the tile sample. However, the following Normal probability plot and W-test (below left) are supportive of the Normality assumption. The bedrock sample looks a little strange in the dot plot (we have done some staggering to cope with overprinting). There are 8 points below the sample mean, a large gap and then 3 larger observations. We see under the Normal probability plot (below right) that W-test has P-value $= 0.03$ indicating significant departures from Normality.

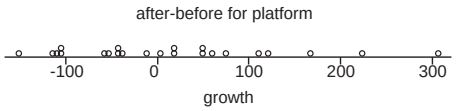


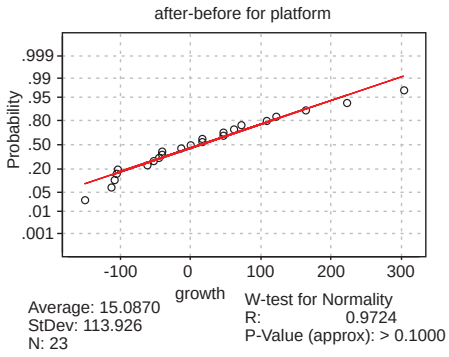


(b) A Mann-Whitney (Wilcoxon) test has P-value $= 1.000$! The reason for this strange result is that it uses a t_0-statistic that takes the value of zero; this has as one-sided P-value of 0.5, which is doubled. An approximate 95% confidence interval for the difference in the medians is $[-10, 7]$. There is clearly no evidence of a difference. However, we need to be careful about the bedrock sample. The Mann-Whitney test is strictly a test to see if two independent samples come from the same distribution, and, although we don't have significance, the two samples are very different looking.

(c) You would need to randomize the placing of the tiles and the selection of bedrock samples to avoid any systematic bias.

10. (a) This is paired data so we should be looking at differences. Let *plat* $=$ *after* $-$ *before*. A dot plot and Normal probability plot of these differences follow. Visually, the center of the sample appears to be to the right of 0, thus suggesting that average areas on the platforms are larger, on average, after 2 weeks.

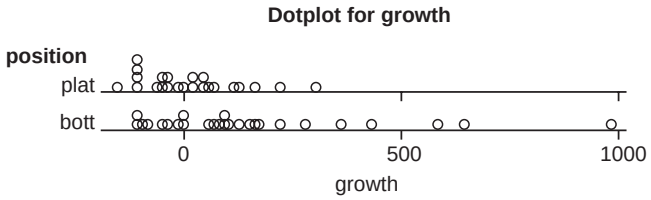




A paired comparison t-test for testing $H_0 : \mu_{plat} = 0$ versus $H_0 : \mu_{plat} \neq 0$ gives $t_0 = 0.64$ and P-value $= 0.53$, indicating no evidence of a real change.

There is some mild skewness in the dot plot (above left), and this is confirmed by slight curvature in the Normal probability plot (above right). However, the W-test provides no evidence against Normality, so what we are seeing can be explained in terms of sampling variation. Also we have a large enough sample for the t-tests and intervals to cope with mild skewness.

(b)



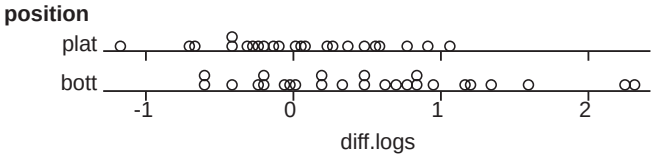
The variable called *growth* in the dot plot above left is the difference between area after and area before *growth* = *after* − *before*. We have compared growth for platform and bottom sites. The bottom group appears to be shifted to the right of the platform group suggesting greater areas covered, on average, on bottom sites than on platform sites.

Applying a Welch two-sample t-test to *growth* to test $H_0 : \mu_{bot} - \mu_{plat} = 0$ versus $H_0 : \mu_{bot} - \mu_{plat} > 0$, we obtain $t_0 = 2.54$ and P-value $= 0.0079$ giving strong evidence of a true difference. An approximate 95% confidence interval for the true difference in the means, $\mu_{bot} - \mu_{plat}$, is given by $[29, 257]$. With 95% confidence, the true mean canopy area is greater on bottom sites than on platform sites by somewhere between 29 and 257 units.

(c) The dot plot on the left in (b) indicates skewness for both groups. The "bottom" data are strongly skewed. The question arises as to if we are operating beyond the robustness limits of the 2-sample t-test. We have 26 observations in the bottom group and 23 in the platform. The difference in (a) is still highly significant if we use a 2-tailed test (double the P-value). We are fairly confident in the results of the t procedures but will try to confirm the basic conclusions using other methods.

We could use a Mann-Whitney test, though the two distributions have noticeably different shapes and spreads, so significance from the Mann-Whitney might come as much from changes in those features as change in location.

Extension: A better approach is to transform the original data using logarithms and then work with the differences of the logarithms for both sets of data as in the dot plot below. Again the center for the bottom group seems to be shifted towards the right.



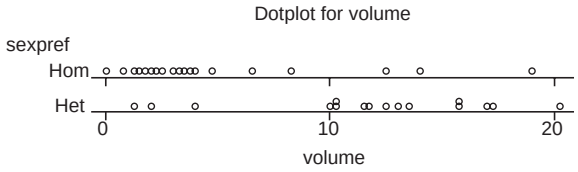
As seen from the dot plots for the logged data, we have got rid of most of the skewness. Working with the differences in the logarithms is equivalent to working with the logarithms of the ratios, a common approach in dealing with growth data. Applying a one-sided Welch test to the difference of the two means for the two sets of transformed data we get $t_0 = 2.6$ with P-value $= 0.0063$ leading us to the same basic conclusion as before.

(d) It would be a good idea to have one tile of each type reasonably close together so that the environmental conditions were fairly similar for each pair. Clearly the

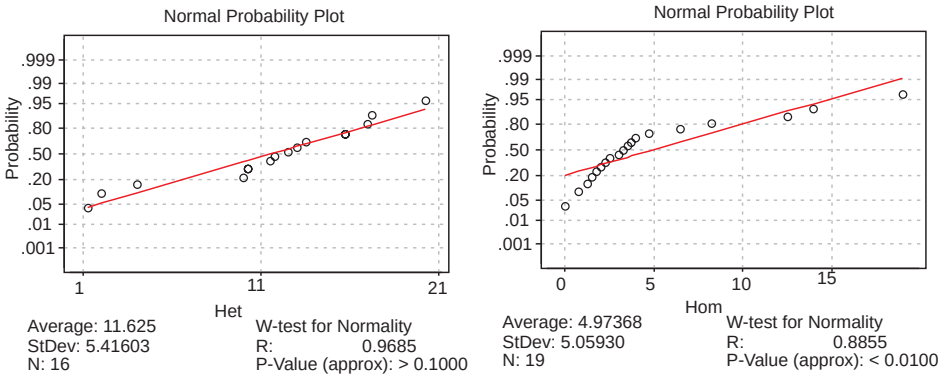
river bottom could vary a lot with regard to snail density so that some form of random placement is necessary.

(e) One concern is that the control and treatment tiles are treated differently so that there may be some factor other than snails, and related to height, which could cause the difference. Perhaps another set of tiles at a different height could be used.

11. (a) From the following dot plots, it appears that average INAH-3 volume is larger for heterosexuals than homosexuals. (The question remaining to be answered in (b) is whether the shift we are seeing might just be due to sampling variation.) We see that the heterosexual data appear slightly skewed while the homosexual data are more strongly skewed, but in the opposite direction.



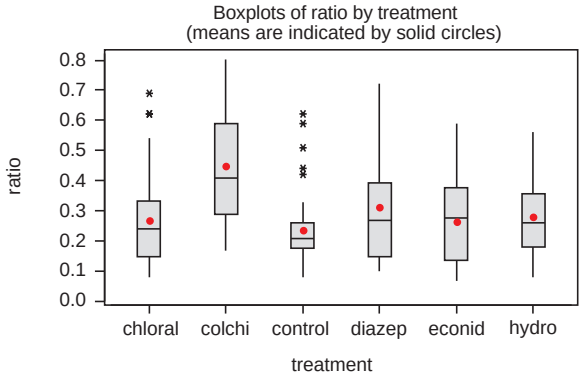
The Normal probability plot and the W-test for the heterosexual data provide no evidence against the Normality assumption (below left), while the plot for the homosexual data emphasizes the skewness and the W-test shows significant non-Normality.



(b) We use a two-sample t-test to test $H_0 : \mu_{het} - \mu_{hom} = 0$ versus $H_1 : \mu_{het} - \mu_{hom} \neq 0$. The Welch test gives $t_0 = 3.73$ with P-value $= 0.0008$, giving very strong evidence for a difference. A 95% confidence interval for $\mu_{het} - \mu_{hom}$ is given by $[3.0, 10.3]$.

(c) This is an observational study, so we cannot prove causality. The samples are not random, as a very high percentage (about 38%) of the heterosexual men died of AIDS.

12. (a) We have given box plots below.



The colchicine sample has a distinctly higher average level than any of the others, while the average level for the control sample looks a little smaller. The spreads for 4 groups appear similar. The control sample appears to have a smaller spread, but with several large outside values. The colchicine spread seems larger. Because of the presence of large outside values, longer upper whiskers, and means usually being bigger than medians, there seems to be a general right-skewness in the data. The stem-and-leaf plots yield similar information.

(b) We wish to test H_0 : population means all equal versus H_1 : population means not all equal. The printout for the F-test and individual confidence intervals follow:

```
One-way Analysis of Variance

Analysis of Variance for ratio
Source     DF        SS        MS        F        P
treatmen    5    1.4345    0.2869    13.82    0.000
Error     294    6.1053    0.0208
Total     299    7.5398

                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level       N      Mean     StDev  ------+---------+---------+---------+
chloral    50    0.2686    0.1406       (----*----)
colchi     50    0.4482    0.1755                            (----*----)
control    50    0.2366    0.1124   (----*----)
diazep     50    0.3116    0.1762             (----*----)
econid     50    0.2646    0.1258      (----*----)
hydro      50    0.2812    0.1205        (----*----)
                                    ------+---------+---------+---------+
Pooled StDev =   0.1441              0.240     0.320     0.400     0.480
```

We see that $f_0 = 13.82$ with P-value $= 0.000$, which indicates very strong evidence against H_0, i.e, we have strong evidence that real differences exist between at least some of the true mean levels. The 95% confidence intervals for the individual means show that the interval for colchicine does not overlap with any of the others, suggesting that this may be the main reason for rejecting H_0.

(c) **(i)** $f_0 = 1.14$ with P-value $= 0.34$. There is no evidence of true differences in mean levels between chloral hydrate, hydroquinone, diazepam and econidazole.

(ii) $f_0 = 1.98$ with P-value $= 0.098$. There is at best very weak evidence of differences if we include the control sample 1.

The main conclusion is that sample 6 (colchicine) is different from the others.

(d) Fisher's pairwise comparisons are given below:

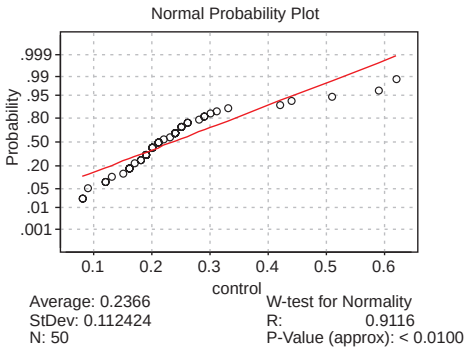
```
Fisher's pairwise comparisons

    Family error rate = 0.361
Individual error rate = 0.0500

Critical value = 1.968

Intervals for (column level mean) - (row level mean)

             chloral      colchi     control      diazep      econid

colchi       -0.2363
             -0.1229

control      -0.0247      0.1549
              0.0887      0.2683

diazep       -0.0997      0.0799     -0.1317
              0.0137      0.1933     -0.0183

econid       -0.0527      0.1269     -0.0847     -0.0097
              0.0607      0.2403      0.0287      0.1037

hydro        -0.0693      0.1103     -0.1013     -0.0263     -0.0733
              0.0441      0.2237      0.0121      0.0871      0.0401
```

All of the confidence intervals contain zero except those involving colchicine, namely, *colch.* − *control* : $[0.15, 0.27]$; *colch.* − *diazep.* : $[0.08, 0.19]$; *colchi.* − *econid.* : $[0.13, 0.24]$; and *colchi.* − *hydro* : $[0.11, 0.22]$.

(e) Most of the samples appear to be positively skewed.

(f) Use stem-and-leaf plots. The control sample is skewed with a steep mode.

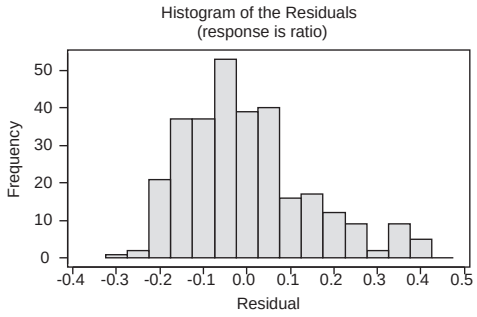
```
Character Stem-and-Leaf Display


Stem-and-leaf of control   N  = 50
Leaf Unit = 0.010


    3    0 889
    6    1 223
   19    1 5666678899999
  (15)   2 000011112344444
   16    2 55566899
    8    3 013
    5    3
    5    4 24
    3    4
    3    5 1
    2    5 9
    1    6 2
```

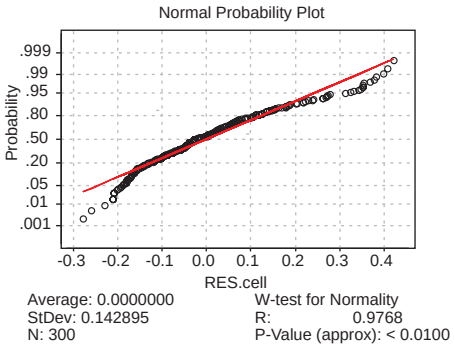


This is supported by the Normal probability plot given (above right).

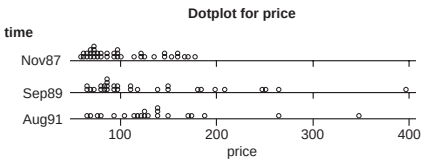
(The plots for the other samples are informative, revealing several modes in some samples. These modes do not show up with box plots.)

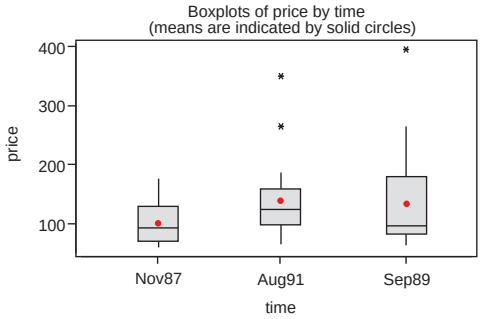
(g) The histogram of the residuals is skewed, and this is reflected in the nonlinear Normal probability plot and the significant W-test. (The plots follow.) We therefore have clear evidence of non-Normality. (Various tests also indicate that the standard deviations are significantly different.) However, the sample sizes are large and equal, and the standard deviations are within reasonable bounds. We would therefore trust the F-test, but have less faith in our confidence intervals. (Working with the logarithms of the data leads to similar standard deviations and a good Normal probability plot. There is no change in the conclusions about significance, however.)





13. **(a)** Dot plots and box plots follow.





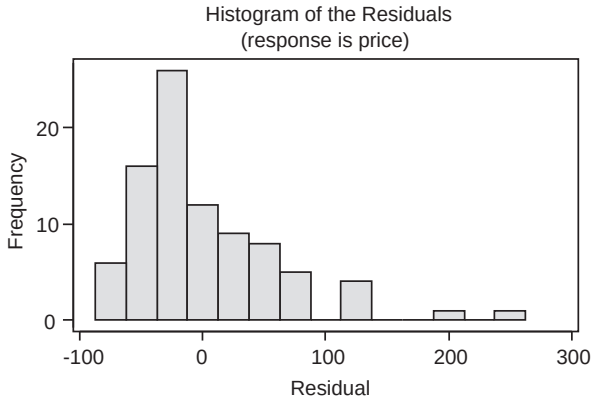
The means and standard deviations are: $\overline{x}_{87} = 101.59$, $s_{87} = 36.11$; $\overline{x}_{89} = 134.37$, $s_{89} = 76.89$; and $\overline{x}_{91} = 139.33$, $s_{91} = 66.19$. There is a substantial increase in the sample mean from November 1987 to September 1989 and almost no difference between September 1989 and August 1991. There is also a substantial increase in the spread after 1987. From the dot plots we see that this is in part due a few more expensive homes in 1989 and 1991. The box plots show similar trends, though the differences don't appear to be so obvious because of the compressed vertical scale.

(b) We wish to test H_0: three means equal versus H_1: three means not all equal. The printout for the F-test follows. We see that $f_0 = 3.65$ with P-value $= 0.030$, yielding some evidence against H_0. The individual 95% confidence intervals overlap so we cannot immediately conclude that 1987 is different.

```
One-way Analysis of Variance

Analysis of Variance for price
Source     DF        SS        MS        F        P
time        2     26299     13149     3.65    0.030
Error      85    306033      3600
Total      87    332331

                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level       N      Mean     StDev  --------+---------+---------+--------
Aug91      21    139.33     66.19              (-----------*---------)
Nov87      37    101.59     36.11   (-------*------)
Sep89      30    134.37     76.89              (--------*-------)
                                    --------+---------+---------+--------
Pooled StDev =    60.00             100       125       150
```

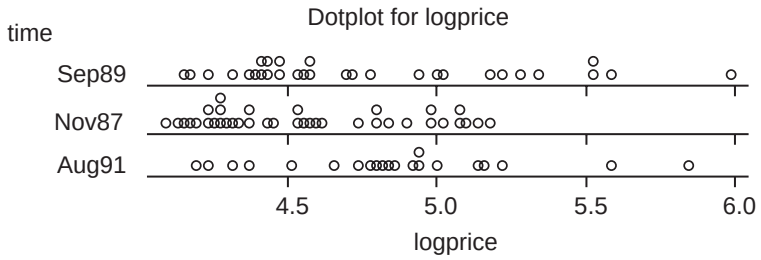
(c) Using Fisher's pairwise comparisons, we have $91 - 87 : [5.1, 70.3]$, $89 - 87 : [3.5, 62.1]$, $91 - 89 : [-29.0, 38.9]$. The intervals are quite wide, indicating a fairly large degree of uncertainty about the differences between the true means. For example, with 95% confidence, the true 1987 mean was smaller than that for 1989 by somewhere between \$3500 and \$62,000.

(d) We see that $s_{89} > 2s_{87}$. There are outliers present, and the 1989 data are clearly skewed. The histogram of the residuals is skewed.

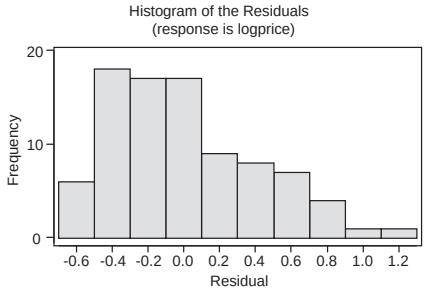


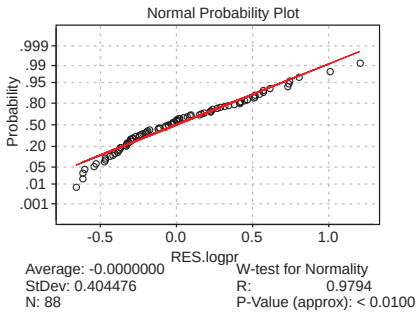
The above analysis is therefore suspect.

(e) Except for possible outliers, the dot plots indicate that some of the skewness seems to have been removed, and the spreads are now more similar.



(f) Using the F-test with the logarithmic data, we get $f_0 = 4.00$ and P-value $= 0.022$, so that our conclusion is unchanged. The data are still skewed, as seen from the histogram of the residuals (below left) and the slight curvature in their Normal probability plot (below right).



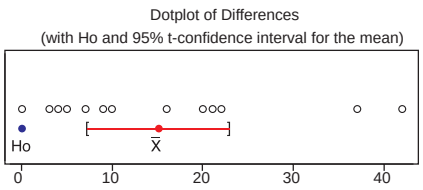


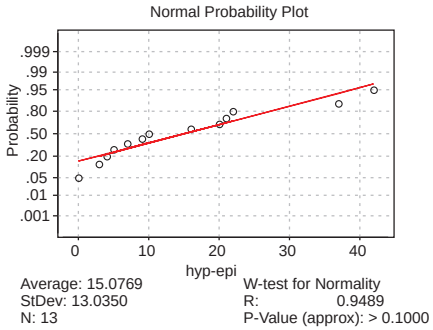
We find that the standard deviations are now similar. We again conclude that there is a significant increase in house prices from 1987 to 1989, and no evidence of a change from 1989 to 1991.

(g) An increase in a mean house price does not imply that all individual house prices go up; some will go down as well. The top end of the market may tend to rise or fall while the bottom end stays fairly static. Furthermore, any increase in the average may be due to just a few expensive houses being sold. These comments would apply to all houses. We would need to look at houses sold more than once or, if there are few in this category, compare houses with similar valuations.

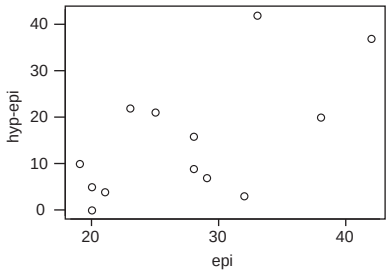
14. (a) This is a paired comparison experiment. Let $diff = hypo - epi$, then we wish to test $H_0 : \mu_{diff} = 0$ versus $H_0 : \mu_{diff} \neq 0$. Using a one-sample t-test, $t_0 = 4.17$ with P-value $= 0.001$. There is very strong evidence of a difference. The hyplimnion values are clearly larger. A 95% CI puts the true mean difference at somewhere between 7.20 units and 22.95 units.

(b) The dot plot (below left) looks reasonable, though there appears to be two outliers. However, the Normal probability plot (below right) looks satisfactory for a small sample, and the W-test for Normality is not significant indicating that the problems we think we are seeing in the dot plot could just be due to sampling variation.





(c) The following scatter plot indicates an increasing trend. Hypolimnion and epilimnion values taken at the same time are clearly related. They tend to increase together in a linear way.



15. **(a)** Select a random sample of 10 out of 20, and assign them to the standard treatment.

(b) You could use a paired-comparison method based on the differences.

(c) No, as we have two independent samples.

(d) You can again use a paired-comparison method, though there are more complicated methods of analyzing design III.

(e) To allow for any carry-over effect or changes over time.

(f) Design III. Any carry-over effect will be balanced out: half of the subjects will get treatment 1 first and the other half treatment 2 first. This is in contrast to design II, where you may not get 10 subjects with each ordering.

16. **(a)** True.

(b) False. The alternative is that the means are not all equal so that some, but not all, could be equal.

(c) False. A small *P-value* is evidence of a difference.

(d) False. We have an observational study. Also the F-test relates to sample means and not to individual women.

17. **(a)** **(i)** One sample. Confidence interval (as we are not told what "effective" means).

(ii) The percentages are approximately Normal with equal standard deviations.

(iii) No placebo is used for a comparison. Also, some patients will have more headaches than others so that the (Binomial) percentages will have different standard deviations. Generalizability: How similar are the people under study to those the treatment will be marketed to?

(b) **(i)** (We would need random assignment of plots to A or B, i.e., a completely randomized design.) Two independent samples. Confidence interval.

(ii) The data set for each method is Normally distributed and the sets are independent.

(iii) Variability in the fertility, for example, of the plots, which may become confounded with the method difference. Generalizability: How similar is the land in the experiment to that potatoes will ultimately be grown on?

(c) **(i)** More than two independent samples. Confidence intervals.

(ii) Assume that the numbers trapped for each color are Normally distributed and that the four standard deviations are all equal. Also, assume that the four samples are independent. (We need to have some randomized method, such as a randomized block design, for allocating the color to each board.)

(iii) There may be a variation in the numbers of beetles in different parts of the field.

(d) **(i)** Paired data. Hypothesis test.

(ii) Differences Normally distributed with the same standard deviation.

(iii) There may be a carry-over learning effect. The order of using the thread needs to be randomized so that half the students use the right-hand thread first and the other half use the left-hand thread first.

18. **(a)** One would need to ensure that the control group was free of infection and that some form of blocking was used along with random allocation.

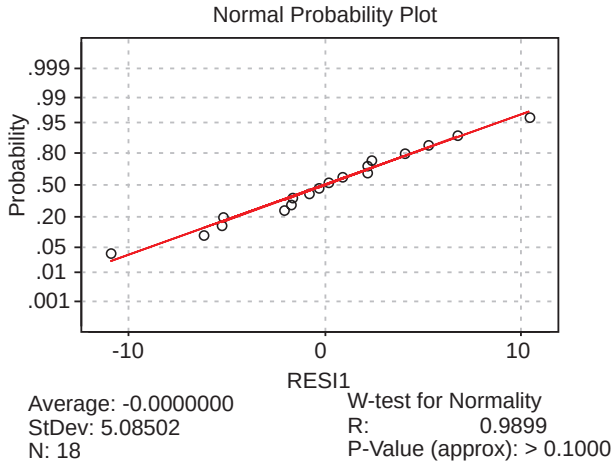
(b) If there are temperature differences, we really have four populations for each of the control and fungal treatments, and not just one of each as required for a two-sample test. Another way of looking at this is that, as far as comparing treatment to control is concerned, observations taken at the same temperature are related rather than independent as they tend to be similar. (This is clear from Fig. 1(a) in the book).

(c) A one-way analysis of variance on all plants to compare temperatures would not be applicable because responses from plants subject to the same treatment are likely to be related rather than independent (unless there is no treatment effect).

(d) The yield tends to increase with temperature for both the control and fungal data up to a certain temperature. The fungal data seem to have a marginal reduction in the growth. There is considerable variability in the spreads.

(e) Yes, provided the assumptions are satisfied for a one-way *ANOVA*. We are now looking at the effect of temperature on one type of plant.

(f) We wish to test H_0: the population means for 18°, 22° and 26° are all equal versus H_1: the population means are not all equal. Applying the F-test we get $f_0 = 4.28$ with P-value $= 0.034$. There is some evidence of a temperature difference. (Note that the combined Normal probability plot for the residuals is satisfactory.



However, $s_{18} = 2.8$ while $s_{22} = 7.6$, which is more than double; the F-test may not be valid. Certainly the individual confidence intervals for the means based on a pooled standard deviation are not appropriate.)

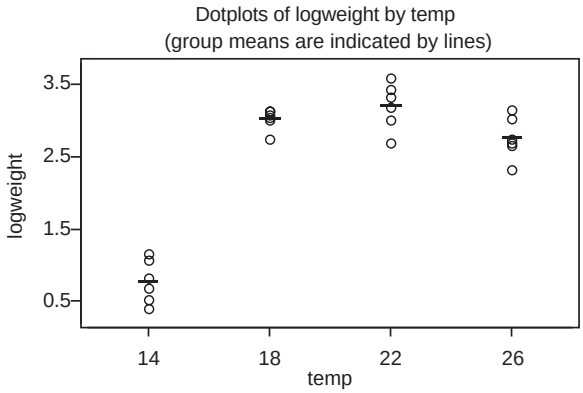
(g) Using Welch's method we get the following confidence intervals. $\mu_{22} - \mu_{26}$: $[0.7, 17.6]$ and $\mu_{22} - \mu_{18}$: $[-3.2, 13.0]$.

(h) The spreads are too different.

(i) Yes.

(j) We use Welch's test for *control* − *fungal*.
For 18°, $t_0 = 3.27$, P-value $= 0.011$ so we have strong evidence of a treatment effect. A 95% confidence interval for the true difference in means, $\mu_{control} - \mu_{fungal}$, is given by $[2.1, 12.3]$. The mean weight for control grass is bigger than that for paspalum by somewhere between 2 and 12 g.
For 22°, $t_0 = 0.53$, P-value $= 0.61$ showing no evidence of that a real difference exists. The 95% confidence interval for the true difference in means, $\mu_{control} - \mu_{fungal}$, is given by $[-7.1, 11.4]$.

(k) Yes, for all but one of the data sets (control, 18°).



(l) Here $f_0 = 101.03$ with P-value $= 0.000$. There is a big difference between the yields at 14° and the other temperatures. This is confirmed by Fisher's pairwise comparisons given below, which also indicate a difference between the yields at 22° and 26°.

```
Fisher's pairwise comparisons

    Family error rate = 0.192
Individual error rate = 0.0500

Critical value = 2.086

Intervals for (column level mean) - (row level mean)

                  14          18          22

      18     -2.5656
             -1.9079

      22     -2.7469     -0.5101
             -2.0891      0.1476

      26     -2.3109     -0.0742      0.1071
             -1.6532      0.5835      0.7648
```

(m) The fungal treatment does not look very useful as the effect is small, if any. It may also vary with the temperature.