

Chapter 3 Example 1: Age Discrimination?

Try these questions yourself before you use the following solutions to check your answers.

Robert Martin turned 55 in 1991. Earlier in that same year, the Westvaco Corporation, which makes paper products, decided to downsize. They ended up laying off roughly half of the 50 employees in the engineering department where Martin worked, including Martin. Later that year, Martin went to court, claiming that he had been fired because of his age. A major piece of evidence in Martin's case was based on a statistical analysis of the relationship between the ages of the workers and whether they lost their jobs.

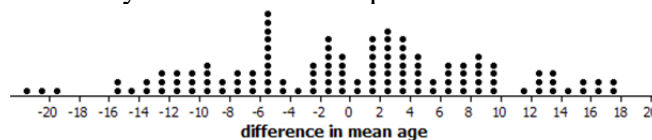
Part of the data analysis presented at his trial concerned the ten hourly workers who were at risk of layoff in the second of five rounds of reductions. At the beginning of Round 2 of layoffs, there were ten employees in this group. Their ages were 25, 33, 35, 38, 48, 53, 55, 55, 56, 64. Three were chosen for layoff: the two 55-year-olds (including Martin) and the 64-year old.

(a) State the null and alternative hypotheses for this study.

(b) Calculate the observed value of the sample statistic.

(c) If we were to carry out an exact randomization test for these data, how many combinations would there be?

(d) Below is the exact randomization distribution. Why is this graph not symmetric? What is the exact p-value? Do you see a short-cut way to determine this p-value?



(e) Simulate a randomization test for these data and state your conclusion at the .01 level of significance.

(f) Carry out a two-sample t -test for these data and hypotheses, and state your conclusion at the .01 level of significance.

(g) Do these analyses reach the same conclusion? If not, which analysis should be used? Explain.

(h) Remind the lawyers what limitations there are to the conclusions that can be made in this study.

Analysis

(a) Let $\mu_{\text{fired}} - \mu_{\text{not fired}}$ represent the difference in the average age of people fired (by the overall process) and the people not fired. (Not worrying too much at this stage about which stage of the firing process we are in. Just keep in mind, we are trying to say something beyond the observed means.)

$H_0: \mu_{\text{fired}} - \mu_{\text{not fired}} = 0$ (no overall difference in the average ages of those getting fired and not)

$H_a: \mu_{\text{fired}} - \mu_{\text{not fired}} > 0$ (those getting fired will tend to have higher ages than those not)

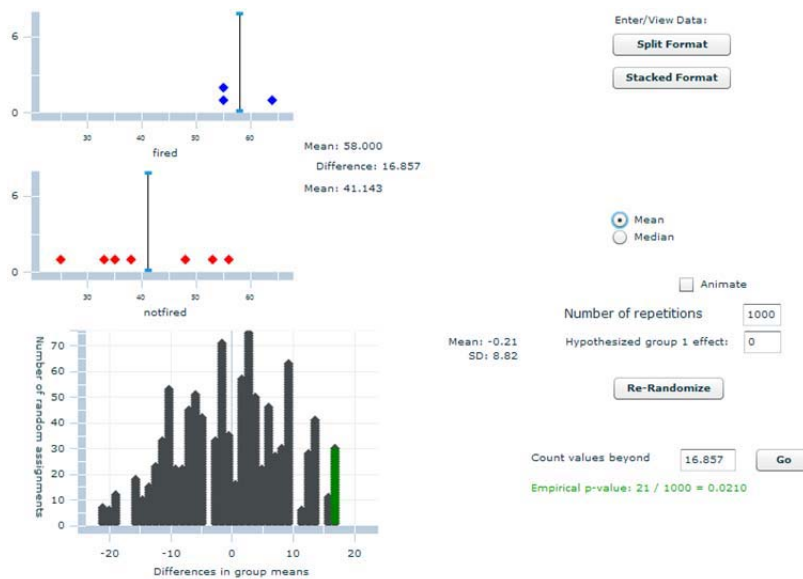
(b) $\bar{x}_1 - \bar{x}_2 = 58.00 - 41.14 = 16.68$ years

(c) $C(10, 3) = 120$, this the number of ways of choosing, completely at random, the three people to be fired from the 10 employees in that round.

(d) Although we suspect the distribution to be centered at zero, the graph is not expected to be symmetric because the group sizes are unequal and we subtracted fired – not fired. Also, the data set is small so there will be “granularity” between the possible values the difference can be. From the graph we can see four differences above 16. Actually three of them are at least 16.68. So the exact p-value = $3/120 = .025$. We can confirm this calculation by seeing that the ages in the fired group are 3 of the top 4 ages. So the only way to get a difference at least as extreme is if the fired group consists of the top 3 ages (55, 56, 64) – and there are 2 ways that can happen – or the observed group of ages (55, 55, 64).

(e) Using the Randomization Test applet (for a quantitative response), the empirical p-value (remembering to match the direction of subtract, which may vary depending on how you pasted the data in), the output below shows an empirical p-value of .021. Because $.021 > .01$, we would not reject the

null hypothesis at the .01 level and include that this firing process was not more likely to fire individuals with larger ages.



Note: You could also carry out this simulation in R

```
> I = 10000; diff = 0
> for (i in 1:I){
  rerandom = sample(age)
  diff[i] = mean(rerandom[1:3]) - mean(rerandom[4:10])
}
```

(f) A two-sample t -test in Minitab or R:

Two-Sample T-Test and CI: age, fired?

Two-sample T for age

fired?	N	Mean	StDev	SE Mean
fired	3	58.00	5.20	3.0
not fired	7	41.1	11.4	4.3

Difference = μ (fired) - μ (not fired)

Estimate for difference: 16.86

95% lower bound for difference: 6.90

T-Test of difference = 0 (vs >): T-Value = 3.21 P-Value = 0.007 DF = 7

Welch Two Sample t-test

data: age by fired.

$t = 3.2064$, $df = 7.763$, $p\text{-value} = 0.006499$

alternative hypothesis: true difference in means is greater than 0

Now $p\text{-value} < .01$, and we would reject the null hypothesis at the .01 level and conclude that this firing process was not more likely to fire individuals with larger ages.

(g) The two-sample t -test yields a much smaller p -value, about .006, which implies much stronger evidence of an underlying age difference between the two populations (fired and not fired). However, because of the small sample sizes (especially with the unbalanced groups), we would have major concerns about using the t -procedures for this study. The empirical p -value from the simulated randomization test is more trustworthy. In fact, in this study, it's probably more work than we need to do, because with only 120 possible combinations it is not unreasonable to calculate the exact p -value.

(h) Even if you find the exact p -value to be convincing evidence of a difference in the average ages of those fired and not by this process, we should not jump to a cause-and-effect conclusion (that age caused the greater chance of being fired) because this was not a randomized experiment but simply an observational study.

Chapter 3 Example 2: Speed Limit Changes

Try these questions yourself before you use the following solutions to check your answers.

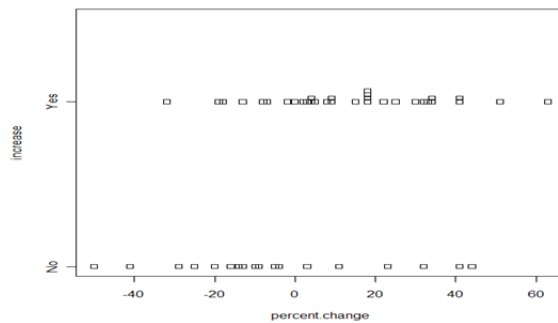
In 1995, the National Highway System Designation Act abolished the federal mandate of 55 miles per hour maximum speed limit and allowed states to establish their own limits. Of the 50 states (plus District of Columbia), 32 increased their speed limits in 1996. The data in `TrafficFatalities.xls` shows the percentage change in interstate highway traffic fatalities from 1995 to 1996 and whether or not the state increased their speed limit. (Data from the National Highway Traffic Safety Administration as reported in Ramsey and Schafer, 2002.)

- (a) Identify the observational units and response variable of interest. Is this a randomized experiment or an observational study?
- (b) Produce numerical and graphical summaries of these data and describe how the two groups compare.
- (c) Are the technical conditions for a two-sample t -test met for this study? Explain.
- (d) Carry out a two-sample t -test to determine whether the average percentage change in interstate highway traffic fatalities is significantly higher in states that increased their speed limit. If you find a significant difference, estimate its magnitude with a confidence interval.
- (e) Discuss what the p -value in (d) measures.

Analysis:

(a) The observational units are the 50 states (and the District of Columbia). The response variable of interest is the percentage change in traffic fatalities from 1995 to 1996 (quantitative). This is an observational study because the researchers did not randomly assign which states would increase their speed limits.

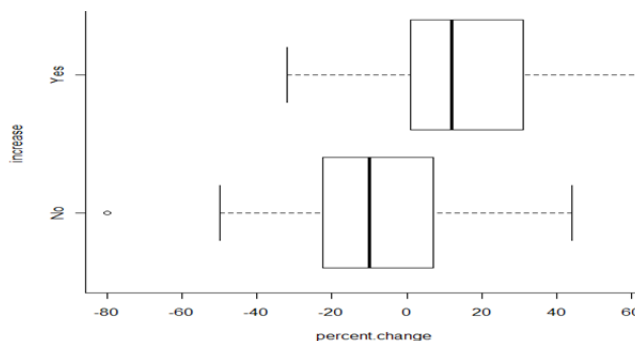
(b) The following graphical display is dotplots of the percentage change in traffic fatalities for each state (and D.C.) in the two groups on the same scale:



Because the distributions are reasonably symmetric, it makes sense to report the means and standard deviations as the numerical summaries:

No increase	$\bar{x}_{\text{no}} = -8.53\%$	$s_{\text{no}} = 31\%$
Increase	$\bar{x}_{\text{yes}} = 13.69\%$	$s_{\text{yes}} = 22\%$

These results indicate that there is a tendency for the percentage change in traffic fatalities to be higher in those states that increase their speed limits. This tendency is also seen in stacked boxplots:



The boxplots also reveal an outlier, the District of Columbia, which did not change its speed limit and had an unusually high decrease in the percentage change of accidents.

These summaries also reveal that the two sample distributions are reasonably similar in shape and spread.

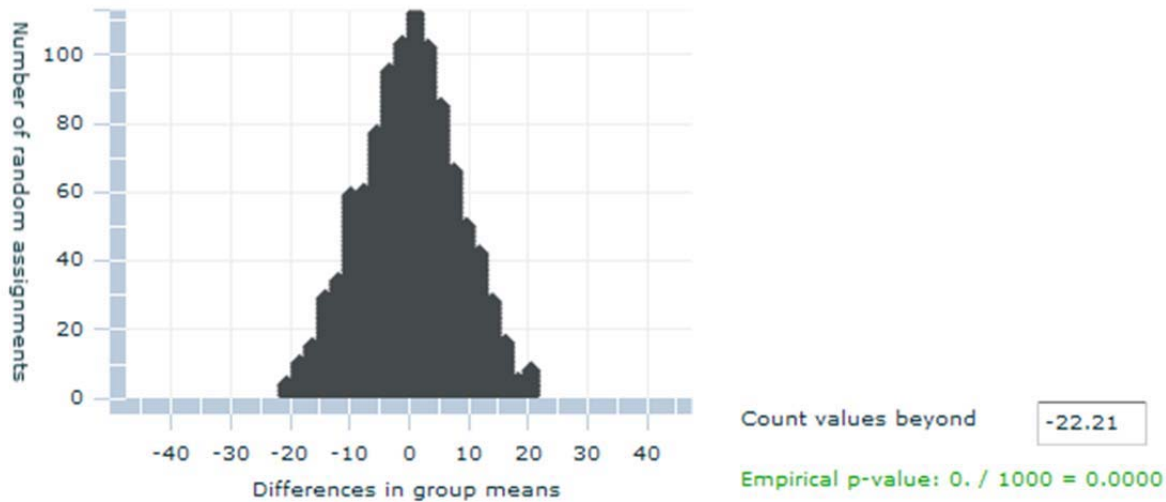
(c) In considering the technical conditions, we see that the sample sizes (19 and 32) are reasonably large. Coupled with the normal shaped sample distributions, the normality/large sample size conditions appears to be satisfied for us to use the t distribution.

(d) Let $\mu_{\text{no}} - \mu_{\text{yes}}$ represent the true “effect” of increasing the speed limit on the traffic fatality rate (states that didn’t change speed limit – states that did change speed limit)

$H_0: \mu_{\text{no}} - \mu_{\text{yes}} = 0$ there is no true effect from increasing the speed limit

$H_a: \mu_{\text{no}} - \mu_{\text{yes}} < 0$ increasing the speed limit leads to an increase in traffic fatalities (higher average percentage change with increase in speed limit)

We can apply a randomization test that would look at what would happen if these groups were mixed up with no difference between the “no” group and the “yes” group.

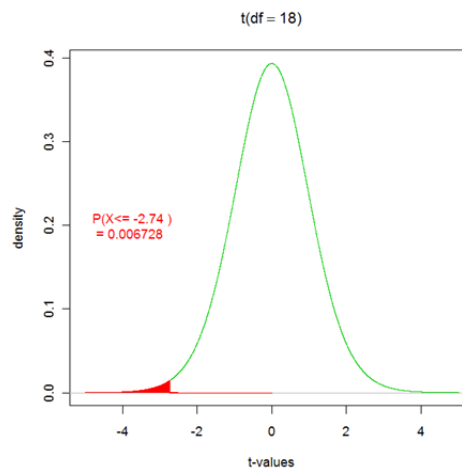


We can also approximate this randomization distribution with the two-sample t procedure.

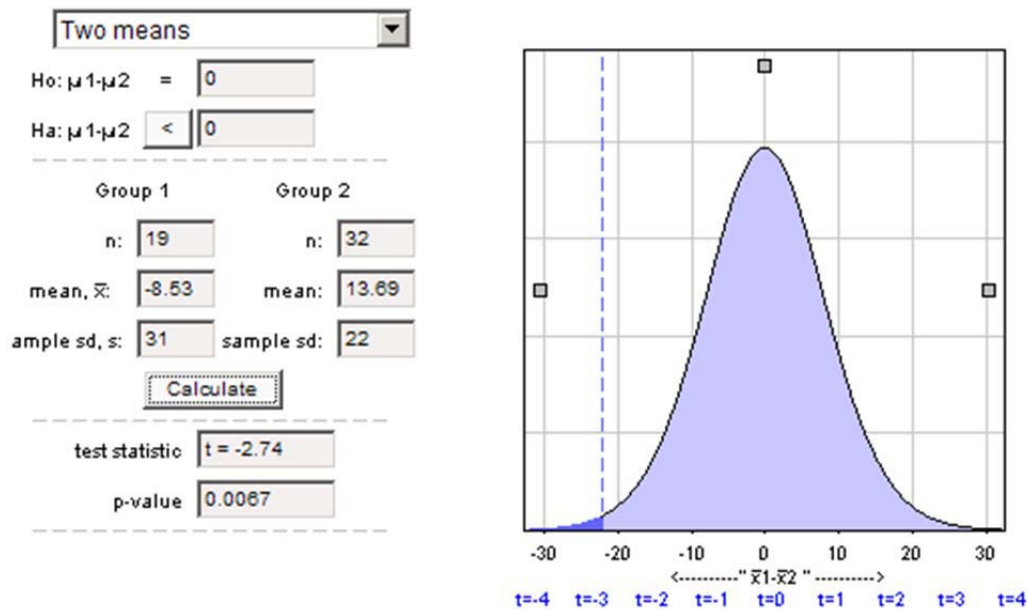
In this case, the (unpooled) test statistic will be $t = \frac{-8.53 - 13.69}{\sqrt{31^2/19 + 22^2/32}} = -2.74$

If we approximate the degrees of freedom by $\min(19-1, 32-1) = 18$, then we find the one-sided p-value in R or Minitab to be:

```
> iscantprob(-2.74, 18, "below")
probability: 0.006728
```



These calculations are confirmed by the Test of Significance Calculator applet and by R:



Welch Two Sample t-test

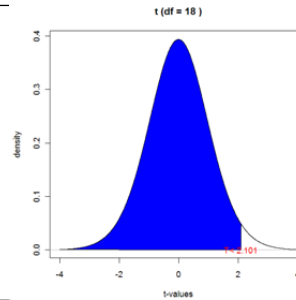
```
data: percent.change by increase
t = -2.7635, df = 28.353, p-value = 0.004968
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf -8.545338
sample estimates:
mean in group No mean in group Yes
 -8.526316          13.687500
```

Note: R uses a more exact method for determining the degrees of freedom. Our “by hand” method (also used in the applet) is conservative in that the p-value found will be larger than the actual p-value as seen here.

Such a small p-value ($.005 < .01$) reveals that we would observe such a large difference in group means by random assignment alone if there was no treatment effect only about 5 times in 1000, convincing us that the observed difference in the group means is larger than what we would expect just from random assignment. We have strong evidence that something other than “random chance” led to this difference. However, we cannot attribute the difference solely to the speed limit change since this was not actually a randomized experiment. Since the states self-selected, there could be confounding variables that help to explain the larger increase in fatality rates in states that increased their speed limit.

Since we rejected the null hypothesis, we are also interested in examining a confidence interval to estimate the size of the treatment effect. We first approximate the t^* critical value for say 95% confidence, again using $\min(19-1, 32-1) = 18$ as the degrees of freedom.

```
> iscaminvt(.975, 18, "below")
the observation with 0.975
probability below is 2.101
```



Then the 95% confidence interval can be calculated,

$$-8.53 - 13.69 \pm 2.10092 \sqrt{31^2/19 + 21^2/32} = -22.2 \pm 16.85$$

We are 95% confident that the true “treatment effect” is in this interval or that the mean percentage increase in traffic fatality rates is between 5.4 percentage points to 39.1 percentage points higher in states that increase their speed limit compared to states that do not increase their speed limit (continuing to be careful not state this as a cause and effect relationship).

Before we complete this analysis, it is worthwhile to investigate the amount of influence that the outlier (the District of Columbia) has on the results, especially since D.C. does have different characteristics from the states in general. The updated R output (two-sided p-value) is below:

Welch Two Sample t-test

```
data: percent.change by increase
t = -2.5079, df = 29.687, p-value = 0.01785
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -33.105399 -3.380712
sample estimates:
 mean in group No mean in group Yes
 -4.555556          13.687500
```

As we might have guessed, the mean increase in fatalities for the “No” group has increased so that the difference in the group means is less extreme. This leads to a less extreme test statistic and a larger p-value (one-sided p-value = $.01785/2 = .0089$) so somewhat weaker evidence against the null hypothesis in favor of the one-sided alternative hypothesis.

(e) The other technical condition is that we have independent random samples or random assignment to groups. We do not have either in this study, because we are examining the population of all states (and D.C.), and the states self-selected whether they changed their speed limit. Thus, any p-value we calculate is in a sense hypothetical. Since we have all the states here, we might ask the question: would the two groups look this different if whether or not they increased their speed limit had been assigned at random?

So the above p-value measures how often we would see a difference in group means at least this large based on random assignment to the two groups if there was no true treatment effect. Even though this p-value is hypothetical, we still have some sense that the difference observed between the groups is larger than we would expect to see “by chance” even in a situation like this where it is not feasible to carry out a true randomized experiment. This gives some information that can be used in policy decisions but we must be careful not to overstate the attribution to the speed limit change.

Chapter 3 Example 3: Distracted Driving?

Researchers asked student volunteers to use a machine that simulated driving situations. At irregular intervals, a target would flash red or green. Participants were instructed to press a “brake button” as soon as possible when they detected a red light. The machine would calculate the mean reaction time to the red flashing targets for each student in milliseconds.

The students were given a warm-up period to familiarize themselves with the driving simulator. Then the researchers had each student use the driving simulation machine while talking on a cell phone about politics to someone in another room and then again with music or a book-on-tape playing in the background (control). The students were randomly assigned as to whether they used the cell phone or the control setting for the first trial. The reaction times (in milliseconds) for 16 students appears below and in the file `driving.txt`.

Subject	A	B	C	D	E	F	G	H
Cell phone reaction	636	623	615	672	601	600	542	554
Control reaction	604	556	540	522	459	544	513	470
Subject	I	J	K	L	M	N	O	P
Cell phone reaction	543	520	609	559	595	565	573	554
Control reaction	556	531	599	537	619	536	554	467

(a) Analyze the *differences* in reaction times (cell phone minus control) for these subjects. Include numerical and graphical summaries of the distribution of differences. Comment on what this descriptive analysis reveals about whether talking on a cell phone tends to produce slower reaction times.

(b) State the appropriate null and alternative hypotheses to be tested, in order to investigate the research question of whether talking on a cell phone tends to produce slower reaction times.

(c) Conduct a simulation analysis of a randomization test for testing these hypotheses. Report the empirical p-value. Summarize the conclusion that you would draw from this analysis.

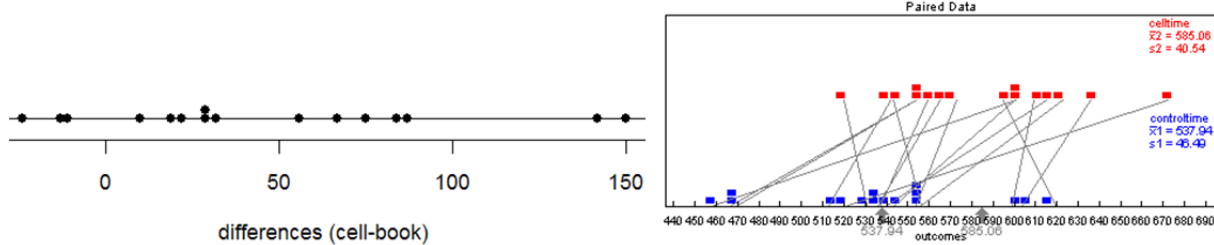
(d) Comment on whether the conditions for applying a paired t -test and t -interval are satisfied for these data.

(e) Conduct a paired t -test of these hypotheses. Report the value of the test statistic and the p-value. Indicate your test decision at the .05 and .01 significance levels, and summarize your conclusion.

(f) Produce and interpret a 95% t -confidence interval for the population mean difference. Also produce and interpret a 95% prediction interval. Comment on how these two intervals compare.

Analysis

(a) Analyzing the differences

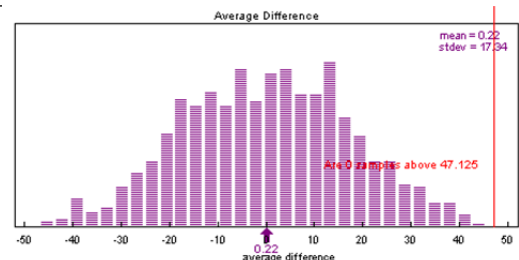


The sample mean difference in reaction times (cell minus control) is $\bar{x}_d = 47.125$ milliseconds, with a standard deviation of $s_d = 51.331$ milliseconds. The dotplot reveals that most of the differences are positive, suggesting that subjects talking on a cell phone tend to take longer to react than subjects listening to a book-on-tape.

(b) The null hypothesis says that the mean reaction time is the same among cell phone users as among book-on-tape listeners ($H_0: \mu_{\text{cell}} - \mu_{\text{control}} = 0$). The alternative says that the mean reaction time is larger among cell phone users than among book-on-tape listeners ($H_a: \mu_{\text{cell}} - \mu_{\text{control}} > 0$).

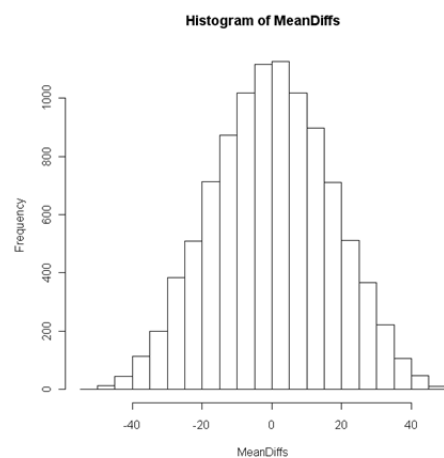
(c) Can carry out the simulation easily with the applet or with R.

Copying and pasting the data into the Matched Pairs applet with 1,000 repetitions



Using R to carry out the simulation instead:

```
MeanDiffs=0
for (i in 1:10000){
  multiplier=sample(c(-1,1), 16,
+  replace=TRUE)
  RandomizedData=differences*multiplier
  MeanDiffs[i]=mean(RandomizedData)
}
```



From R: The empirical p-value is the proportion of these 10,000 repetitions in which the mean difference is 47.125 or more, because 47.125 is the value of the sample mean difference from the actual experimental data. None of the 10,000 repetitions produced such a large mean difference, so the empirical p-value is 0. The simulation therefore shows that we would almost never get a result as

extreme as the actual experiment did, if there were really no difference between reactions to cell phone vs. book-on-tape, so we have extremely strong evidence that the cell phone really does increase reaction times.

(d) Because the sample size (16) is fairly small, the t -procedures are only valid if the population of differences follows a normal distribution. The dotplot of differences from these 16 subjects looks roughly symmetric, so the t -procedures are probably valid to apply here.

(e) The test statistic is $\frac{\bar{x}_d - 0}{s_d/\sqrt{n_d}} = \frac{47.125}{51.331/\sqrt{16}} \approx 3.67$. The p-value is the probability that a t -

distribution with 15 degrees of freedom is 3.67 or larger; R reveals this p-value to be .001137. This p-value is very small, so we would reject the null hypothesis at the .05 and .01 significance levels. The experimental data provide very strong evidence that talking on a cell phone does cause an increase in mean reaction time, as compared to listening to a book-on-tape. The cause/effect conclusion is justified because this is a randomized experiment with a very small p-value.

(f) A 95% confidence interval for the population mean difference μ_d is: $\bar{x}_d \pm t^* s_d/\sqrt{n_d}$, which is $47.125 \pm 2.131(51.331)/\sqrt{16}$, which is 47.125 ± 27.347 , which is (19.778, 74.472). We can be 95% confident that the mean reaction time while talking on a cell phone is between roughly 20 and 75 milliseconds longer than when listening to a book-on-tape.

A 95% *prediction* interval for the difference in reaction times for a particular subject is:

$\bar{x}_d \pm t^* s_d \sqrt{1 + 1/n_d}$, which is $47.125 \pm 2.131(51.331)\sqrt{1 + 1/16}$, which is 47.125 ± 112.753 , which is (-64.628, 159.878). We can be 95% confident that the an individual subject will react anywhere from 65 milliseconds more quickly to 160 milliseconds more slowly talking on a cell phone as compared to listening to a book-on-tape.