retain the penny and the proportion of female students at this university who would vote to retain the penny.

(b) Are the technical conditions for this procedure satisfied? Explain.

(c) Repeat (a) using the Wilson adjustment. How do the intervals compare?

(d) Why should you be cautious about interpreting this interval for the population of all students at this university?

INVESTIGATION 5.1.3 SLEEPLESS DRIVERS

Connor et al. (*British Medical Journal*, May 2002) reported on a study that investigated whether sleeplessness is related to car crashes. The researchers identified all drivers or passengers of eligible light vehicles who were admitted to a hospital or died as a result of a car crash on public roads in the Auckland, New Zealand region between April 1998 and July 1999. Through cluster sampling, they identified a sample of 571 drivers who had been involved in a crash resulting in injury and a sample of 588 drivers who had not been involved in such a crash as representative of people driving on the region's roads during the study period. The researchers asked the individuals if they had a full night's sleep any night during the previous week.

a. Identify the observational units and variables in this study. Which variable would you consider the explanatory variable and which the response variable?

b. Is this an observational study or an experiment? Explain.

c. Would this be considered a case-control, a cohort, or a cross-classified design as defined in Chapter 1? Explain.

d. Is it reasonable to consider these as two independent random samples? If so, from what populations? Explain.

e. Suppose we define the parameter π_1 to be the proportion of drivers who had not experienced a full night's sleep in the previous week that had a car accident and π_2 to be the proportion of drivers who had experienced a full night's sleep that had an accident. Would it be appropriate to estimate $\pi_1 - \pi_2$ from these data? Explain. (*Hint:* Consider some of the study design issues discussed in Section 1.2.)

We cannot estimate this parameter from the data because the distribution of the accident variable was controlled by the researchers. Therefore, we will instead consider the population odds ratio as the parameter of interest. Define τ to be the population odds ratio of a car accident for the sleepless group compared to the "full night sleep" group.

f. State the null and alternative hypotheses for testing whether this odds ratio is greater than 1. State in words what these hypotheses imply about the association between sleeplessness and occurrence of car accidents.

The researchers found that 61 of the 535 "case" drivers who responded to this question (out of 571 identified) and 44 of the 588 "control" drivers had not gotten a full night's sleep in the previous week.

g. Organize these sample data into a two-way table:

	No full night's sleep in past week	At least one full night's sleep in past week	Sample sizes
"Case" drivers (crash)			535
"Control" drivers (no crash)			588

h. Produce and discuss numerical and graphical summaries of these sample data, including the sample odds ratio, denoted by $\hat{\tau}$. What do these summaries reveal? Does the sample odds ratio appear to be extreme?

In order to evaluate whether this sample odds ratio is extreme, we need information about the sampling distribution of the odds ratio when the population proportions are the same (and so the population odds ratio $\tau = 1$). You will use simulation to approximate this sampling distribution when this null hypothesis is true and then assess whether 1.48 is a larger sample odds ratio than would typically occur by chance. The following simulation assumes that 9% (roughly the pooled estimate from the sample data) of drivers in the population did not get a full night's sleep in the previous week (for both the cases and the controls). From this population, we will sample 535 "case" drivers and 588 "control" drivers to mimic the researchers' study, under the hypothesis of no association between the two variables.

i. Use Minitab to randomly generate 1000 observations from a binomial distribution with $\pi = .09$ and $n = 535$ (Calc > Random Data > Binomial), storing the results in C1, and 1000 observations from a binomial distribution with $\pi = .09$ and $n = 588$, storing the results in C2. Then calculate the odds ratio for each row (pair of samples) as follows:

```
MTB> let c3=(c1*(588-c2))/(c2*(535-c1))
```

Produce and discuss numerical and graphical summaries for these simulated odds ratio values. Are they reasonably modeled by a normal distribution? (*Hint:* Examine a normal probability plot.) Is the mean close to what you would have predicted? Explain.

Mean: Standard deviation:

Description:

Normal?

Mean close to prediction?

j. How often did the simulation produce an odds ratio at least as extreme as the 1.59 value observed by the researchers? What conclusion do you draw from this empirical *p*-value?

Study Conclusions

The sample odds ratio of 1.59 indicates that the odds of a sleepless driver having a crash were about 60% higher than those for a well-rested driver in this sample. The empirical *p*-value (less than 5%) provides moderately strong evidence that such an extreme value for the sample odds ratio is unlikely to have arisen by chance alone if the proportion of drivers with sleepless nights was .09 for both the population of "cases" and the population of "controls."

Discussion It would be nice to follow this simulation up with a probability model for the distribution of the sample odds ratio that did not require us to assume a value for the population proportion as we did earlier. However, the distribution of the simulated odds ratios in C3 is clearly skewed to the right and so will not be well modeled by a normal distribution. This makes sense because the range of possible values for the sample odds ratio is not symmetric around 1 (since it is bounded by 0 on the left and unbounded on the right). Thus, to be able to use a probability model, we need either to determine an appropriate probability model for these odds ratios or to transform the odds ratio into another statistic that does follow a common probability distribution. Next you will study the latter strategy.

k. Use Minitab to determine the *log-odds ratio* for your 1000 simulated samples:

```
MTB> let c4=log(c3)
```

(*Note:* Minitab assumes the natural log unless you type "logten(c3)", but either base will suffice here.) Produce and discuss numerical and graphical summaries for these log-odds ratios. Are the log-odds reasonably modeled by a normal distribution? Is the mean close to what you would have predicted? Explain.

Mean: Standard deviation:

Description:

Normal?

Mean close to prediction?

Although the sampling distribution of the sample odds ratio is not normal, the sample log-odds values are approximately normally distributed. Thus, we can conduct a test and construct a confidence interval for the population log-odds ratio using the normal distribution. The standard error of the sample log-odds ratio (using the natural log) is given by the expression:

$$SE(\text{log-odds}) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

where a, b, c, and d are the four table entries.

I. Calculate the log-odds ratio for the sample data in this driver sleepiness study, and then calculate the standard error, $SE(\text{log-odds})$. Verify that the standard deviation of your empirical sampling distribution is close to this value.

m. Construct a 90% confidence interval for the population log-odds based on the sample data. (*Hint:* First calculate the sample value of the log-odds. Then go 1.645 standard errors on either side of that value.)

n. Exponentiate these two endpoints of the interval to get a 90% confidence interval for the population odds ratio. Does your interval contain the value 1? Discuss the implications of the interval containing 1 or not.

Study Conclusions

The proportions of drivers who had not gotten a full night's sleep in the previous week were .114 for the case group of drivers who had been involved in a crash compared to .075 for the control group who had not. Because these proportions are small, and because of the awkward roles of the explanatory and response variables in this study (we would much rather make a statement about the proportion of sleepless drivers who are involved in crashes), the odds ratio is a more meaningful statistic to calculate. The sample odds of having missed out on a full night's sleep were 1.59 times higher for the case group than for the control group. By the invariance of the odds ratio, we can also state that the sample odds of having an accident are 1.59 times higher for those who do not get a full night sleep than those who do. Your initial simulation results found moderately strong evidence (one-sided empirical p-value $\approx .015$) that the population of drivers involved in crashes is more likely to have gone without a full night's sleep. A 90% confidence interval for the population odds ratio extends from 1.13 to 2.24. This interval provides statistically significant evidence that the population odds ratio exceeds 1 and that the odds of having an accident are about 1 to 2 times higher for the sleepless drivers than for well-rested drivers. We cannot attribute this association to a cause-and-effect relationship because this was an observational (case-control) study.

A confidence interval for a population odds ratio τ is

$$\exp\left(\log(\hat{\tau}) \pm z^* \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

where $\hat{\tau}$ denotes the sample odds ratio.