STAT 240 Homework 2

Rebecca Barter, Andrew Do, and Kellie Ottoboni

March 4, 2015

Chapter 26

Review exercise 2

With a perfectly balanced roulette wheel, in the long run, red numbers should turn up 18 times in 38. To test its wheel, one casino records the results of 3,800 plays, finding 1,890 red numbers. Is that too many reds? Or chance variation?

Part (a)

Formulate the null and alternative hypotheses as statements about a box model.

The null hypothesis is that in a box of 38 tickets, there are 18 that are red. The alternative is that in this box of 38 tickets, there are more than 18 red tickets.
Bec – could be totes wrong, but should this be that the box specifically has 3,800 tickets, of which 1,800 are red? Since we are specifically drawing from the box of the 3,800 plays?
Andrew - I think it's fine if we say the sampling is with replacement?

Part (b)

The null says that the percentage of reds in the box is ???. The alternative says that the percentage of reds in the box is ???. Fill in the blanks.

The null says that the percentage of reds in the box is $\frac{18}{38} \approx 48\%$. The alternative says that the percentage of reds in the box is greater than 48%.

Part (c)

Compute z and P

$$z = \frac{1890 - (\frac{18}{38})(3800)}{\sqrt{(\frac{18}{38})(\frac{20}{38})(3800)}} = \frac{1890 - 1800}{\sqrt{\frac{36000}{38}}} \approx 2.924$$

Under the null hypothesis, z has approximately a standard normal distribution. Thus the p-value for the one-sided test is $P = P(z \geq 2.924) \approx 0.0017$.

Part (d)

Were there too many reds?

At the significance level 0.05, we reject the null hypothesis that there were 18 red tickets in the box. In particular, this means that there were too many reds in the 3800 roulette spins to be due to chance alone.

Review exercise 5

A newspaper article says that on average, college freshmen spend 7.5 hours a week going to parties. One administrator does not believe that the figures apply to her college, which has nearly 3,000 freshmen. She takes a simple random sample of 100 freshmen, and interviews them. On average, they report 6.6 hours a week going to parties, and the SD is 9 hours. Is the difference between 6.6 and 7.5 real?

Part (a)

Formulate the null and alternative hypothesis in terms of a box model

Consider a box containing 3000 tickets, from which we take a random sample of 100 without replacement. The null hypothesis is that the average of the box is 7.5 and the alternative is that the average is something other than 7.5.
Bec - should we say instead, Consider a box containing 3000 tickets corresponding to the 3000 students. Ticket i contains the reported number of hours a week going to parties for student i

Part (b)

Fill in the blanks. The null says that the average of the box is ???. The alternative says that the average of the box is ???.

The null hypothesis is that the average of the box is 7.5 and the alternative is that the average is something other than 7.5.
Bec - should we say "less than 7.5" rather than "something other than 7.5", since at this point in the book we haven't introduced two-sided tests yet?

Part (c)

Now answer the question: is the difference real?

$$T = \frac{7.5 - 6.6}{\frac{9}{\sqrt{100}}\sqrt{\frac{100}{99}}} = \frac{0.9}{0.904534} \approx 0.9950$$

Under the null hypothesis, T has a t distribution with 99 degrees of freedom. Thus the p-value for the two-sided test is $P(t_{99} \geq T) \approx 0.16$. We fail to reject the null hypothesis that the box average is different from 7.5, at the 0.05 significance level.

Review exercise 7

I.S. Wright and associates did a clinical trial on the effect of anticoagulant therapy for coronary heart disease. Eligible patients who were admitted to participating hospitals on odd days of the month were given the therapy; eligible patients admitted on even days were the controls. In total there were 580 patients in the therapy group and 442 controls. An observer says: "since the odd-even assignment to treatment or control is objective and impartial, it is just as good as tossing a coin". Do you agree or disagree? Explain briefly. Assume the trial was done in a month with 30 days

We disagree with the statement that the odd-even assignment to treatment or control is objective and impartial, and thus it is just as good as tossing a coin. The primary evidence stems from the fact that the number of patients in the treatment group is 580 and the number of patients in the control group is 442, which corresponds to a difference of 138, which seems very large. We can in fact conduct a hypothesis test

to see whether or not this study generated a balance that was just as good as randomizing.

Our null hypothesis is that the probability of being assigned to treatment is the same as the probability of being assigned to control.

On the other hand, our alternative hypothesis is that the probability of being assigned to treatment is greater than the probability of being assigned to control. Note that we have observed a difference of $580 - 442 = 138$ but our null hypothesis says that this difference should be 0. Our test statistic can be calculated as follows

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}} = \frac{138 - 0}{\sqrt{1022 \times 0.5 \times 0.5}} = 8.63$$

which is absurdly large. In particular, our P is so tiny that we could not possibly imagine that this difference occurred by pure chance. Thus we reject the null hypothesis that the off-even assignment to treatment or control is just as good as tossing a coin.
Andrew - Since we were given the liberty to use whatever test we'd like, do you think we should do a binomial test here? The normal approximation is probably fine given the large sample size though.

Review exercise 10

On November 9, 1965, the power went out in New York City, and stayed out for a day – the Great Blackout. Nine months later, the newspapers suggested that New York was experiencing a baby boom. The table below shows the number of babies born every day during a 25-day period, centered nine months and ten days after the Great Blackout. These numbers average out to 436. This turns out not to be unusually high for New York. But there is an interesting twist to the data: the 3 Sundays only average 357. How likely is it that the average of 3 days chosen at random from the table will be 357 or less? Is chance a good explanation for the difference betters Sundays and weekdays? If not, how would you explain the difference?

We have observed an average of 436 births over the entire 25-day period, but an average of 357 for the three Sundays in the 25-day period. The first question is how likely is it that the average of 3 days chosen at random from the table is 357 births or less. Note first that the only non-Sunday which could contribute to an average of 357 or less is Saturday the 7th. Thus the possible triples with average 357 or less are:

{Sun 7, Sun 14, Sun 21} , {Sun 7, Sun 14, Sat 6} , {Sun 7, Sat 6, Sun 21} , {Sat 6, Sun 14, Sun 21}

$$\begin{aligned} P(\text{3 randomly chosen days have 357 births or less}) &= \frac{\#\{ \text{ of all triples which average to 357 or less } \}}{\#\{ \text{ all triples } \}} \\ &= \frac{4}{\binom{25}{3}} \\ &= \frac{1}{575} \end{aligned}$$

which is extremely small. Next, we ask if chance is a good explanation for the difference between Sundays and weekdays. Using a hypothesis test, we take our null hypothesis to be that the probability of a birth occurring is the same for each day, and our alternative hypothesis to be that the probability of being born on Sunday within this 25 day period is lower than on other days. We observe an average of 357 births on Sundays, however under our null hypothesis, we expect that we would observe an average of 436 births. Thus, since the variance of the observations is 1649.29 our test statistic is given by:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}} = \frac{357 - 436}{\sqrt{\frac{1649.29}{3}}} = -3.37$$

which corresponds to $P \approx 0.00038$ which is very small. Thus we can be fairly certain that the difference cannot be explained by chance. However, a possible explanation is that many births were done by cesarean section, and the doctors did not want to schedule them on Sundays.

Review exercise 12

(Hard.) Does the psychological environment affect the anatomy of the brain? This question was studied experimentally by Mark Rosenzweig and his associates. The subjects for the study came from a genetically pure strain of rats. From each litter, one rat was selected at random for the treatment group, and one for the control group. Both groups got exactly the same kind of food and drink—as much as they wanted. But each animal in the treatment group lived with 11 others in a large cage, furnished with playthings which were changed daily. Animals in the control group lived in isolation, with no toys. After a month, the experimental animals were killed and dissected.

On the average, the control animals were heavier and had heavier brains, perhaps because they ate more and got less exercise. However, the treatment group had consistently heavier cortexes (the "grey matter," or thinking part of the brain). This experiment was repeated many times; results from the first 5 trials are shown in the table: "T" means the treatment, and "C" is for control. Each line refers to one pair of animals. In the first pair, the animal in treatment had a cortex weighing 689 milligrams; the one in control had a lighter cortex weighing only 657 milligrams. And so on.

Two methods of analyzing the data will be presented in the form of exercises. Both methods take into account the pairing, which is a crucial feature of the data. (The pairing comes from randomization within litter.)

Part (a)

First Analysis. How many pairs were there in all? In how many of these pairs did the treatment animal have a heavier cortex? Suppose treatment had no effect, so each animal of the pair had a 50-50 chance to have the heavier cortex, independently from pair to pair. Under this assumption, how likely is it that an investigator would get as many pairs as Rosenzweig did, or more, with the treatment animal having the heavier cortex? What do you infer?

The treatment animal had a heavier cortex in 52 of 59 pairs. We use a one-tailed binomial test to check the probability of seeing a count of 52 or greater under the null hypothesis that each animal of the pair had a 50-50 chance to have a heavier cortex.

$$\begin{aligned} P(\text{In 52 pairs or more, treatment heavier}) &= P(\text{In 7 pairs or fewer, control heavier}) \\ &= \sum_{k=1}^{7} \binom{59}{k} \left(\frac{1}{2}\right)^{k} \left(\frac{1}{2}\right)^{59-k} \\ &\approx 8.8 \times 10^{-11} \end{aligned}$$

This is a negligible p-value, so we can reject the null hypothesis in favor of the alternative that treatment effected a heavier cortex.

Part (b)

Second Analysis. For each pair of animals, compute the difference in cortex weights "treatment - control." Find the average and SD of all these differences. The null hypothesis says that these differences are like

draws made at random with replacement from a box whose average is 0—the treatment has no effect. Conduct an appropriate hypothesis test. What do you infer?

First we note that the treatment and control averages are dependent since the rats came in pairs from the same litter, meaning if one rat is predisposed to having a heavy cortex, its littermate is also. That being said, we use a box model where there is only one box and the tickets represent the difference between treatment and control pairings (the observational unit is the pair of rats). We perform a t-test with 58 degrees of freedom and begin by calculating the following:

- The average of the differences is 36.2
- The standard deviation for the differences is approximately 31.5 mg.
- The standard error for the average of the differences is $31.5/\sqrt{59} \approx 4.1$ mg
- The t-statistic then is

$$t = \frac{observed - expected}{standard\ error} = \frac{36.2 - 0}{4.1} \approx 8.8$$

This gives a p-value of 1.2×10^{-12}. Thus we reject the null hypothesis in favor of the alternative that the differences in cortex mass is not due to chance variability.

Part (c)

To ensure the validity of the analysis, the following precaution was taken. "The brain dissection and analysis of each set of littermates was done in immediate succession but in a random order and identified only by code number so that the person doing the dissection does not know which cage the rat comes from." Comment briefly on the following: What was the point of this precaution? Was it a good idea?

This precaution reduces detection bias. In principle, the blinding reduces the likelihood that the person performing the dissection does something for the treatment group that he/she does not do for the control group (or vice versa).

Chapter 27

Exercise set D question 4

Many observational studies conclude that low-fat diets protect against cancer and cardiovascular "events". Experimental results, however, are generally negative. In 2006, the Women's Health Initiative published its results. This was a large-scale randomized trial on women who had reached menopause. As one part of the study, 48,835 women were randomized: 19,541 were assigned to the treatment group and put on a low-fat diet. The other 29,294 women were assigned to the control group and ate as they normally would. Subjects were followed for 8 years
Among other things, the investigators found that 1,357 women on the low-fat died experienced at least one cardiovascular event, compared to 2,088 in the control group. Can the difference between the two groups be explained by chance? What do you conclude about the effect of the low fat diet?

Note that we have $N = 48,835$ women who had reached menopause and we are in the situation where our box contains N tickets each of which has two sides t_i and c_i. We have drawn a sample $X_1, ..., X_n$ without replacement ($n = 19,541$) and assigned them to the treatment group (low-fat diet). We also drew a sample $Y_1, ..., Y_m$ without replacement ($m = 29,294$) and assigned them to the control group (eat as normal). We

observed $\sum_i X_i = 1357$ cardiovascular events in the treatment group and $\sum_i Y_i = 2088$ cardiovascular events in the control group. The standard deviation for the treatment group is

$$SD(X) = \sqrt{\frac{1357}{19541} \times \frac{18184}{19541}} = \sqrt{0.06462}$$

which corresponds to a SE of

$$SE\left(\sum_i X_i\right) = 35.535$$

Similarly, the standard deviation for the control group is

$$SD(Y) = \sqrt{\frac{2088}{29294} \times \frac{27206}{29294}} = \sqrt{0.066197}$$

so that the SE is given by

$$SE\left(\sum_i Y_i\right) = 44.036$$

Thus the SE of the difference in the sums is given by

$$SE\left(\sum_i X_i - \sum_i Y_i\right) = \sqrt{34.434^2 + 44.036^2} = 56.584$$

and our test statistic, where our null hypothesis is that there is no difference between the two groups versus our alternate hypothesis is that the number of cardiovascular events is lower in the treatment group (low-fat diet) than in the control group is thus

$$z = -\frac{731}{56.584} = -12.93$$

which is extremely large, and from which it is clear that this difference was not due to chance. In particular, this implies that we have enough evidence to reject the hypothesis that there is no difference in the number of cardiovascular events in the two groups and that the number is lower in the group with the low-fat diet. That is, we conclude that it is unlikely that the difference is due to chance, and so people on a low-fat diet are less likely to suffer from a cardiovascular event.

Exercise set D question 6

Some years, the Gallup Poll asks respondents how much confidence they have in various American institutions. You may assume that results are based on a simple random sample of 1,000 persons each year; the samples are independent from year to year

(a)

In 2005, only 41% of the respondents had "a great deal or quite a lot of" confidence in the supreme court, compared to 50% in 2000. Is the difference real? Or can you tell from the information given?

In this situation, our null hypothesis says that the percentage for the 2005 box is the same as for the 2000 box, whereas the alternative hypothesis says that the percentage for the 2005 box is smaller than the percentage in the 2000 box. The SE for the number of 1's (where a 1 corresponds to a respondent having "a great deal or quite a lot" of confidence in the supreme court) in the 2005 sample is estimated to be

$$\sqrt{1000} \times \sqrt{0.41 \times 0.59} = 15.55$$

and so the SE for the percentage is

$$\frac{15.55}{1000} \times 100\% = 1.56\%$$

On the other hand, the SE for the 2000 percentage is

$$\sqrt{1000} \times \sqrt{0.5 \times 0.5} = 15.81$$

so the SE for the percentage is

$$\frac{15.81}{1000} \times 100\% = 1.58\%$$

Our null hypothesis tells us that the expected difference is 0%, whereas our observed difference is $41\% - 50\% = -9\%$
Thus our z test is

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE for difference}} = \frac{-9\% - 0\%}{\sqrt{1.58^2 + 1.56^2}} = -4.05$$

which is large enough to imply that the difference is real. In particular, this implies that in 2005, people has less confidence in the Supreme Court, compared to in 2000.

(b)

In 2005, only 22% of the respondents had "a great deal or quite a lot of" confidence in Congress, whereas 24% of the respondents had "a great deal or quite a lot of" confidence in organized labor. Is the difference betters 24% and 22% real? Or can you tell from the information given?

In this case, we are comparing responses to two unrelated questions about confidence in Congress and organized labor, respectively. We could examine this question as if it came from a box model where there are four types of tickets [0|0], [0|1], [1|0], and [1|1], where the first entry on the ticket corresponds to whether or not the respondent had "a great deal or quite a lot" of confidence in Congress and the second entry corresponds to whether or not the respondent had "a great deal or quite a lot of" confidence in organized labor. Now, our null hypothesis would be that the proportion of [1|0] and [1|1] tickets (proportion of respondents with confidence in Congress) was the same as the proportion of [0|1] and [0|0] tickets (proportion of respondents with confidence in organized labor). However, we cannot answer the question with the aggregated summaries provided, as we would need to know the proportion of respondents with each of the four confidence combinations. Thus we cannot tell if the difference is real from the information given.

Andrew - I think in terms of the question asked, we're looking for the difference in proportion of [1|1] and [1|0] vs [1|1]*and*[0|1] or equivalently just difference in proportion between [1|0] and [0|1]

Review exercise 8

One experiment contrasted responses to "prediction-request" and to "request-only" treatments, in order to answer two research questions. (i) Can people predict how well they will behave? (ii) Do their predictions influence their behavior? In the prediction-request group, subjects were first asked to predict whether they would agree to do some volunteer work. Then they were requested to do the work. In the request only group, the subjects were requested to do the work; they were not asked to make predictions beforehand. In parts (a-b-c), a two sample z-test may not be legitimate. If it is legitemate, make it. If not, why not?

Part (a)

46 residents of Bloomington, Indiane were chosen at random for the "prediction-request" treatment. They were called and asked to predict whether they would agree to spend 3 hours collecting for the American Cancer Society if contacted over the telephone with such a request. 22 out of the 46 said they would. Another 46 residents of that town were chosen at random for the "request-only" treatment. They were requested to spend the 3 hours collecting for the American Cancer Society. Only 2 out of the 46 agreed to do it. Can the difference be due to chance? What do the data say about research questions (i) and (ii)?

The data tell us nothing about the research questions (i) and (ii); the two proportions measure different things. The fraction 22 of 46 is the proportion of people who *predicted* that they would help whereas 2 of 46 is the proportion of people who actually agreed to help. A z-test is not applicable here.

Part (b)

Three days later, the prediction-request group was called again, and requested to spend 3 hours collecting for the ACS: 14 out of 46 agreed to do so. Can the difference be due to chance? What to the data say about the research questions?

This data can be used to answer question (i), can people predict how well they will behave? However, a two-sample z-test is inappropriate because the groups under comparison are comprised of the same individuals. This introduces correlation between the samples so the standard error of the difference in proportions cannot be computed with the usual formula. We should use a paired z-test instead of a two-sample z-test.

Andrew - I hate the wording in this question. Difference between 14 out of 46 compared and what? We may be able to answer more about (ii) as we can compare the 14 out of 46 in the P-R group to the 2 out of 46 in the P-O group. Both groups were simple random samples drawn independently of each other, which leads me to believe that we have two comparable groups.

Part (c)

Can the difference between 22/46 and 14/46 be due to chance? What do the data say about the research questions?

The proportion of volunteers in each group can be used to answer question (ii), do predicted responses influence behavior? We would conduct a two-sample z-test to compare the proportions.

The null hypothesis is that the rate of volunteers among those who are asked to predict their response is the same as the rate of volunteers among those who are not asked. The alternative hypothesis is that the rates are different. The observed proportion in the "prediction-request" group is 22/46 and the observed proportion in the "request-only" group is 14/46. The test statistic is

$$z = \frac{\frac{22}{46} - \frac{14}{46}}{\sqrt{\frac{22\times 24}{46^2}\frac{1}{46} + \frac{14\times 32}{46^2}\frac{1}{46}}} = \frac{8}{\sqrt{\frac{976}{46}}} \approx 1.7368$$

Under the null, z is approximately standard normally distributed, so the p-value for the two-sided test is

$$P(|Z| \geq |z|) = 2P(Z \geq 1.7368) \approx 0.0824$$

At the 0.05 level, we fail to reject the null hypothesis that the rates are the same.

Andrew - The difference between 22/46 and 14/46 answers research question (i) I think as it's the difference between prediction and what they actually did. In that case we should do a paired t-test

Review exercise 9

A researcher wants to see if the editors of a journal in the field of social work are biased. He makes up two versions of an article, "in which an asthmatic child was temporarily separated from its parents in an effort to relieve the symptoms of an illness that is often psychosomatic". In one version the separation has a positive effect; in the other, negative. The article is submitted to a group of 107 journals; 53 are chosen at random to get the positive version and 54 the negative. The first column of the table says that 28 of the journals getting the positive version accepted it for publication, and 25 rejected it. The second column gives the results for the journals that got the negative version. IS chance a good explanation for the results? If not, what can be concluded about journal publication policy?

The null hypothesis is that the rate of rejections is the same for the paper with positive results and the paper with negative results. The alternative hypothesis is that the rate is higher for the paper with negative results. The observed proportion of rejections for the positive paper is $28/53$ and the observed proportion of rejections for the negative paper is $8/54$. We will conduct a one-sided, two-sample z-test for the difference in proportions. The test statistic is

$$z = \frac{\frac{28}{53} - \frac{8}{54}}{\sqrt{\frac{28\times 25}{53^2}\frac{1}{53} + \frac{8\times 46}{54^2}\frac{1}{54}}} \approx \frac{0.5283 - 0.1481}{\sqrt{\frac{0.249199}{53} + \frac{0.1262}{54}}} = 4.5317$$

The p-value for this test is

$$P(Z \geq 4.5317) = 2.9255 \times 10^{-6}$$

There is strong evidence to reject the null hypothesis that the rejection rates are the same in favor of the alternative hypothesis that the rate of rejections is higher for the paper with negative results.
Bec - Should this be a two-sample test instead, since the question doesn't specify which direction they're testing. Also the proportions reported here are the proportion accepted rather than rejected as stated

Review exercise 10

An investigator wants to show that first-born children score higher on IQ tests than second-borns. He takes a simple random sample of 400 two-child families in a school district, both children being enrolled in elementary school. He gives these children WISC vocabulary test, with the following results: The 400 first-borns average 29 and their SD is 10. The 400 second bots average 28 and their SD is 10. He makes a two-sample z-test. Comment briefly on the use of statistical tests

There is a natural pairing in this experiment: children from the same household are the same with respect to a variety of potential confounders for IQ, including parents' socioeconomic status, parents' education, level of parent involvement, etc. Thus, siblings in this study are highly correlated. It therefore doesn't make sense to use a test which assumes the sample of first-borns, $X_1, \ldots, X_{400}$ is independent of the sample of second-borns, $Y_1, \ldots, Y_{400}$.

In reality, the correlation between the IQ of the first-borns and the IQ of the second-borns is positive, so

$$\begin{aligned} \text{Var}(\bar{X} - \bar{Y}) &= \text{Var}\bar{X} + \text{Var}\bar{Y} - 2\text{Cov}(\bar{X}, \bar{Y}) \\ &< \text{Var}\bar{X} + \text{Var}\bar{Y} \end{aligned}$$

Therefore the estimated standard error, given by $\sqrt{0.5^2 + 0.5^2}$ overestimates the true standard error of the difference in means. Consequently, the test statistic is biased towards 0. The z-test here is inappropriate.

If we had accounted for the covariance between $\bar{X}$ and $\bar{Y}$, the estimated standard error would have been smaller, thus the test statistic z would have been larger. In other words, using a better estimator of the standard error, accounting for the correlation between siblings in the sample, would have given more power to detect a difference in sample means.
Andrew - In fact, the sampling is not random for treatment and control as you didn't pick the ordering of the children.

Chapter 29

Exercise set B question 9

Transfusion of contaminated blood creates a risk of infection. A physician must balance the gain from the transfusion against the risk, and accurate data are important. In a survey of published medial literature on serum hepatitis resulting from transfusions, Chalmers and associates found that the larger studies had lower fatality rates. How can this be explained?

There could be several possible explanations for the observation that larger studies on serum hepatitis resulting from transfusions had lower fatality rates. Firstly, it is likely that larger studies are better funded than smaller studies, which leads both to a bigger sample size and a better quality of care for patients. The higher standard of care in particular could result in lower observed fatality. Next, we could be experiencing some sort of publication bias, whereby we note that journals tend to prefer studies with "statistically significant" results. Statistical significance, however, is more attainable when there are either large effect sizes or large sample sizes. Moreover, it is likely that the only small studies being published are those which achieve statistically significant results, whereas the larger studies probably find it easier to publish their results even if they did not attain statistical significance, simply because the studies themselves are more prestigious.

Exercise set C 6

Before publication in a scholarly journal, papers are reviewed. Is this process fair? To find out, a psychologist makes up two versions of a paper. Both versions describe a study on the effect of rewarding children for classroom performance. The versions are identical except for the data. One dataset shows that rewards help motivate learning, the other, that rewards don't help. Some reviewers were chosen at random to get each version. All the reviewers were associated with a journal whose position was behaviorist: rewards for learning should work. As it turned out, both versions of the paper contained a minor inconsistent in the description of the study. The investigator did a two-sample z-test, concluding that "of the individuals who got the positive version, only 25% found the mistake. Of those who got the negative version, 71.5% found the mistake. By the two-sample z-test, this difference must be considered substantial".

Part (a)

Why is the two-sample z-test legitimate? Or is it?

The two-sample z-test is legitimate here. We have two randomly sampled groups, drawn without replacement from a population of reviewers. The randomization justifies the test.

Part (b)

The standard error for the difference was about ?? percentage points.

We aren't given the sample sizes, but we can recover the standard error using the z-score. The one-sided p-value is approximately 0.02, so the z-score is $z = \Phi^{-1}(1 - 0.02) \approx 2.0537$. Then

$$z = \frac{0.715 - 0.25}{\sqrt{\text{Var}(\bar{X} - \bar{Y})}} = 2.0537$$

Solving gives $\text{SE}(\bar{X} - \bar{Y}) = 0.2264$.

Part (c)

Is the difference substantial? Answer yes or no, and discuss briefly

Yes, the difference between 71.5% and 25% is substantial. Just from a qualitative perspective, this means the rate of rejection with negative results is nearly three times higher than if one had reported positive results.

Part (d)

What to the results of the z-test add to the argument?

In this case, the z-test does not add a great deal of information. It confirms our belief that there is a significant qualitative difference in the rates and gives evidence that the difference is not simply due to chance sampling variability.

Andrew - Adds to the defense against journals saying something like "we were unlucky that too many of our super-critical reviewers got the negative papers"

Part (e)

What to the data say about the fairness of the review process?

The data suggest that the review process is biased to favor positive results. Since reviewers were selected at random to receive either the paper with positive findings or the paper with negative findings, we expect that the two groups are roughly the same with respect to all other variables. Thus, the difference in acceptance rates can be attributed to the difference in the paper's results.

Exercise set D 3

Two researchers studies the relationship between infant mortality and environmental conditions in Dauphin County, Pennsylvania. As a part of the study, the researchers recorded, for each baby born during a six-month period, in what season the baby was born and whether or not the baby died before reaching one year of age. If appropriate, test to see whether infant mortality depends on the season of birth. If a test is not appropriate, explain why not

It is inappropriate to do a statistical test to compare infant mortality rates in this scenario. There is no randomness in the data. Babies weren't randomized to be born in different seasons; their birth simply happened when it happened and we observed the date. Furthermore, we have data for every baby that was born, so there is no sampling variability. Since we observe the entire population of interest and the "treatment" (season of birth) is fixed by Nature, applying a probability model makes no sense.

Review exercise 6

Using election data, investigators make a study of the various factors influencing voting behavior. They estimate that the issue of inflation contributed about 7 percentage points to the Republican vote in a certain election. However, the standard error for this estimate is about 5 percentage points. Therefore, the increase

is not statistically significant. The investigators conclude that "in fact, and contrary to widely held views, inflation has no impact on voting behavior." Does the conclusion follow from the statistical test? Answer yes or no, and explain briefly.

The investigators concluded from a hypothesis test with non-significant p-value where the null hypothesis is that inflation has no effect on the outcome of an election and the alternative hypothesis is that inflation has some effect on the outcome (positive or negative), that "inflation has no impact on voting behavior". This conclusion does not in fact follow from the statistical test, where all that we can conclude is that the data is consistent with the hypothesis that inflation has no impact on voting behavior, and thus that any observed change could be explained from chance variation. We can't "accept" the null hypothesis; we can only fail to reject it.

Review exercise 9

In 1970, 36% of first-year college students thought that "being very well off financially is very important or essential." By 2000, the percentage had increased to 74%. These percentages are based on nationwide multistage cluster samples.

Part (a)

Is the difference important? Or does the question make sense?

The question makes sense and a difference in attitudes gives insight to changes in priorities over time.

Part (b)

Does it make sense to ask if the difference is statistically significant? Can you answer on the basis of the information given?

The box model underlying the question consists of two boxes, one containing the tickets for first-year college students in 1970 and in the other, tickets of their counterparts from 2000. Since they took probability samples, it would make sense to ask if the differences are statistically significant, but because they used cluster sampling, we cannot answer the question with the information given.

Part (c)

Repeat (b), assuming the percentages are based on independent simple random samples of 1,000 first-year college students drawn each year.

The two samples differ by $74 - 36 = 38$ percentage points. The two standard errors for the percentages are

$$\sqrt{\frac{(.74)(1-.74)}{1000}} \times 100\% \approx 1.4\% \text{ and } \sqrt{\frac{(.36)(1-.36)}{1000}} \times 100\% \approx 1.5\%$$

Hence the SE for the difference is

$$\sqrt{1.4^2 + 1.5^2} \approx 2.3$$

percent, so the we get a test statistic

$$t = \frac{38}{2.3} \approx 16.5,$$

which results in a tiny p-value, leading us to rejecting the null hypothesis and concluding that attitudes toward financial status have changed over the years.

Review exercise 11

A market research company interviews a simple random sample of 3,600 persons in a certain town, and asked what they did with their leisure time last year: 39.8% *of the respondents read at least one book, whereas* 39.3% *of them entertained friends or relatives at home. A reporter wants to know whether the difference between the two percentages is statistically significant. Does the question make sense? Can you answer it with the information given?*

While this question makes sense from a marketing standpoint—gaining perspective on what market shares are available—reading books is not mutually exclusive of entertaining friends, so with the given information, it's not a well-formulated in the statistical sense. To answer this question, we would need more specific information: namely what percentage of people read and entertain for pleasure, what percentage just read, what percentage just entertain, and what percentage do neither. The underlying probability model follows a multinomial hypergeometric distribution (multinomial asymptotically). It should also be noted that statistical significance in this setting refers only to whether the difference in proportions is derived from chance and has nothing to do with the practical importance of who to market to.

Special exercise 33

In the U.S., there are two sources of national statistics on crime rates:

i *The FBI's Uniform Crime Reporting Program, which publishes summaries on all crimes reported to police agencies in jurisdictions covering virtually 100% of the population.*

ii *The National Crime Survey, based on interviews with a nationwide probability sample of households.*

In 2001, 3% of the households in the sample told the interviewers they had experienced at least one burglary within the past 12 months. The same year, the FBI reported a burglary rate of 20 per 1,000 households, or 2%. Can this difference be explained as chance error? If not, how would you explain it? You may assume that the Survey is based on a simple random sample of 50,000 households out of 100 million households.

We are interested in the probability that of the $N = 50,000$ interviewed, $1,500$ (3% of $50,000$) experienced burglary when the finite population of $M = 100,000,000$ has $K = 2,000,000$ cases on file. Let $F_{M,K,N}$ be the hypergeometric distribution function with parameters M, K, N as defined above, then the probability p of observing a burglary rate of 3% or higher in a sample size of $50,000$ is given by:

$$\begin{aligned} p &= P(\text{number of households burglarized} \geq 1500) \\ &= 1 - P(\text{number of households burglarized} < 1500) \\ &= 1 - F_{M,K,N}(1499) \\ &= 1 - \sum_{i=0}^{1499} \frac{\binom{2,000,000}{i}\binom{98,000,000}{50000-i}}{\binom{100,000,000}{50,000}} \\ &\approx 4.7 \times 10^{-55} \end{aligned}$$

In other words, the difference between the two rates is almost certainly not due to chance. The most likely explanation is that many burglaries go unreported because people don't want to deal with the hassle of police reports and/or increased insurance rates especially if the stolen items are not particularly valuable.