STAT 240 Homework 2

Rebecca Barter, Andrew Do, and Kellie Ottoboni

March 4, 2015

Chapter 26

Review exercise 2

Part (a)

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.

Part (b)

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)

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

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)

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

Part (a)

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.

Part (b)

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.

Part (c)

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

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

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.

Review exercise 10

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 there are only 3 days which do have less than 357 births, and so the probability is:

$$P(\text{3 randomly chosen days have 357 births or less}) = \frac{3}{25}\frac{2}{24}\frac{1}{23} = \frac{1}{2,300}$$

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:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}} = \frac{357 - 436}{\sqrt{\frac{1649.29}{25}}} = -79$$

It is clear that P is so small that it is essentially zero. Thus we can be fairly certain that the difference cannot be explained by chance. However, a possible alternative explanation is that many births were done by cesarean section, and the doctors did not want to schedule them on Sundays.

Since the statistic of interest is the average of three days from the month, I believe the variance should be 1649.29/9. See the .R file. In any case, the math isn't right - plugging in what you have gives $z = -9.726321$. I haven't looked over this too much, but shouldn't that first probability be

$$P(\text{3 randomly chosen days have an } \textit{average} \text{ of 357 births or less})$$

in which case you'd have to consider all triplets that average to 357. The only non-Sunday thing that stands out to me is that Saturday with 377.

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?

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?

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?

Chapter 27

Exercise set D question 4

Note that we have $N = 48835$ 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

(a)

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)

Now, our null hypothesis says that the percentage of respondents that had a "great deal or quite a lot of" confidence in Congress is the same as the percentage of respondents that had a "great deal or quite a lot of confidence" in organized labor. The alternative hypothesis says that the percentage of respondents that had a "great deal or quite a lot of" confidence in Congress is less than the percentage of respondents that had a "great deal or quite a lot of confidence" in organized labor. 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 Congress) is estimated to be

$$\sqrt{1000} \times \sqrt{0.22 \times 0.78} = 13.10$$

and so the SE for the percentage is

$$\frac{13.10}{1000} \times 100\% = 1.31\%$$

On the other hand, the SE for the organized labor is

$$\sqrt{1000} \times \sqrt{0.24 \times 0.76} = 13.51$$

so the SE for the percentage is

$$\frac{13.51}{1000} \times 100\% = 1.35$$

Our null hypothesis tells us that the expected difference is 0%, whereas our observed difference is $22\% - 24\% = -2\%$

Thus our z test is

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE for difference}} = \frac{-2\% - 0\%}{\sqrt{1.35^2 + 1.31^2}} = -1.06$$

which is not large enough to Imply that the difference is real. Using the normal approximation, this corresponds to a p-value of 0.14, which is not large enough to imply that we have evidence against the hypothesis that there is no difference between the percentage or respondents that had a "great deal or quite a lot of" confidence in Congress and the percentage of respondents that had a "great deal or quite a lot of confidence" in organized labor.

My issue with this question, is what is our box model? It's not an experiment with tmt and ctrl responses on either side, and it's not two boxes with independent populations since it's the same responders for each question I agree - you can crank out the math but it doesn't make sense because the question is comparing apples and oranges. Since the question asks "can you tell from the given information?", I'd just give an explanation of why we can't compare the percentages. I think the following box model represents the Gallup poll: there's a box of tickets representing the entire American population in it. On the tickets is whether or not that person has confidence in Congress and the same information for organized labor, so there are four types of tickets, [0|0], [0|1], [1|0], and [1|1]. The null hypothesis would be that the proportion of tickets with 1's on the left side of the tickets equals the proportion with 1's on the right side. We can't answer the question with the aggregated summary—we need to know percentages on how confidence was paired.

Review exercise 8

Part (a)

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)

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.

Part (c)

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.

Review exercise 9

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

$$z = \frac{\frac{28}{53} - \frac{8}{54}}{\sqrt{\frac{28\times25}{53^2}\frac{1}{53} + \frac{8\times46}{54^2}\frac{1}{54}}} \approx \frac{0.5283 - 0.1481}{\sqrt{\frac{0.249199}{53} + \frac{0.1262}{54}}} = 4.5317$$

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.

Review exercise 10

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.

Chapter 29

Exercise set B question 9

In larger studies, we may be
Potential ideas:

- Large studies are likely better funded than smaller studies. More funding leads to both a bigger sample size and better quality of care for patients. A better standard of care could result in lower fatality.

- Publication bias - large studies have better power to detect a significant difference. Journals tend to prefer publishing studies with "statistically significant" results, which are only possible with large effect sizes or large sample sizes. When fatality rates are low overall, a large sample size is needed for sufficient power to detect an effect. (seems likely this is the "correct" answer because the problem comes right after the section on data snooping)

Addendum:

- More on publication bias - And when the sample size is not large, getting a high fatality rate is more probable, increasing the likelihood that in small studies, doctors would see an alarmingly high patient mortality rate worthy of alerting the rest of the medical community.

Exercise set C 6

Part (a)

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)

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

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

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

Part (c)

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)

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.

Part (e)

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

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.