# *Supplementary Review Problems for the Final: Solutions*

**1.** There are about 2,700 institutions of higher learning in the United States (including junior colleges and community colleges). In 1976, as part of a continuing study of higher education, the Carnegie Commission took a simple random sample of these institutions. The average enrollment in the 256 sample schools was 3,500, with an SD of 4,000. A histogram for the enrollments was plotted and did not follow the normal curve.

Say whether each of the following statements is *true* or *false*, and explain why.

(a) An approximate 95%-confidence interval for the average enrollment of all 2,700 institutions runs from 3,000 to 4,000.

***True:*** *The standard error (for average) is*

$$SE = \frac{SD}{\sqrt{sample\ size}} = \frac{4000}{\sqrt{256}} = 250$$

*so a* 95%*-confidence interval is*

$$Sample\ average \pm 2SE = 3500 \pm 500.$$

(b) It is estimated that 95% of the institutions of higher learning in the United States enroll between $3,000$ and $4,000$ students.

***False:*** *The '*95%*' in '*95%*-confidence interval' refers to the* ***chance*** *that the interval in question,* 3000 *to* 4000*, contains the true* ***average*** *enrollment at all institutions of higher learning in the U.S.*

(c) There is an approximately 95% chance that the average enrollment at all institutions of higher learning in the United States is between 3000 and 4000 students.

***True:*** See (b).

(d) The normal curve can't be used to figure confidence levels here at all, because the data doesn't follow the normal curve.

***False:*** *The histogram for the (box of all possible)* ***sample averages*** *(for samples of the same size —* 256 *in this case) is well-approximated by the normal curve, even if the histogram for the original data (enrollment numbers) is not (as long as the sample size is big enough).*

**2.** A math instructor at a community college wants to teach his class the benefits of practicing. To do this he divides the class into two groups. Students who have practiced hitting a baseball off a tee for more than a year (in high school or little league, for example) go into group A, and the rest of the students go into group B. The class goes out to the school's practice field and every student hits three balls off a tee. The average distance the ball travels for the 37 students in group A is 151 feet, while for the 68 students in group B the average distance the ball travels is 83 feet. At the end of the afternoon, the instructor says "See — practice makes perfect!"

(a) Is the teacher's study observational or is it a controlled experiment?

*This is an observational study. Even though the teacher divides the students into two groups, he does so based on choices that the students made — students selected themselves for the practice and no-practice groups by choosing to participate in little league or high-school baseball or softball.*

(b) Are there any *confounding* variables?

*Gender is an obvious confounding variable. It is likely that more males than females participated in baseball or softball in high school (or before) and men generally have more upper body strength than women. If there are proportionally more men in group A than in group B, this would explain a lot of the difference between the average dis.*

(c) Does the baseball-hitting study confirm the instructor's conclusion? Explain.

*Not this particular study, at least not without more information about the composition of the two groups. The confounding effect of gender cannot be ignored.*

***Having said that — the more you study, and the more seriously you study, the more you learn.***

**3.** Chapter 14, review question 9.

*The possible outcomes of the two draws are*

$$\left\{\boxed{A1}\boxed{B1}\right\}\ \left\{\boxed{A1}\boxed{B2}\right\}\ \left\{\boxed{A1}\boxed{B3}\right\}\ \left\{\boxed{A1}\boxed{B4}\right\}\ \left\{\boxed{A2}\boxed{B1}\right\}\ \left\{\boxed{A2}\boxed{B2}\right\}$$

$$\left\{\boxed{A2}\boxed{B3}\right\}\ \left\{\boxed{A2}\boxed{B4}\right\}\ \left\{\boxed{A3}\boxed{B1}\right\}\ \left\{\boxed{A3}\boxed{B2}\right\}\ \left\{\boxed{A3}\boxed{B3}\right\}\ \left\{\boxed{A3}\boxed{B4}\right\}$$

*all of which are equally likely.*

*(a) The probability that $A > B$ is $3/12 = 1/4$ because exactly 3 of the 12 outcomes satisfy this condition.*

*(b) The probability that $A = B$ is also $3/12 = 1/4$ because exactly 3 of the 12 outcomes satisfy this condition.*

*(a) The probability that $A < B$ is $6/12 = 1/2$ because exactly 6 of the 12 outcomes satisfy this condition.*

**4.** Chapter 26, exercise set F, problems 4 and 5.

*See the solutions in the back of the book.*

**5.** In a calculus class with 180 students, the final exam score contributed 50% of the course score, the midterm score contributed 30% of the course score and the average homework score contributed 20% of the course score.

**a.** After the course was over, the instructor computed three correlation coefficients based on the class data:

- $r_1$ = correlation between average homework score and midterm exam score,
- $r_2$ = correlation between average homework score and final exam score,
- $r_3$ = correlation between average homework score and score in the class.

The three numbers she computed were 0.3521, 0.5582 and 0.4112, but she forgot to label them. Match each number with the appropriate correlation coefficient and explain your choices.

*Since the homework contributes directly to the score in the class, we expect $r_3$ to be the highest, i.e., $r_3 = 0.5582$. Since only about half the homework was completed by the time of the midterm, the correlation between the average of* all *the homework scores and the midterm is likely to be lower than the correlation between the average on the homework and the final exam, so $r_1 = 0.3521 < 0.4112 = r_2$.*

**b.** The same instructor also divided the class into six groups based on average homework scores, using the ranges

$$(0\% - 49\%),\ (50\% - 59\%),\ (60\% - 69\%),\ (70\% - 79\%),\ (80\% - 89\%) \text{ and } (90\% - 100\%).$$

She the computed the average homework score and average final exam score in each group, and computed the correlation coefficient, $r_4$, for the resulting six data points. Do you think that $r_4$ was higher than, lower than or about the same as $r_2$? Explain your answer.

*This is an example of 'ecological' correlation.*[†] *By grouping the class according to average homework score, then computing the average final exam score in each group, the 'spread' in the data of the final scores around their average in each group goes away. Their will be a lot less spread in the scatter plot for the averages, and as a result the correlation coefficient will be bigger. I.e., we can expect $r_4$ to be bigger than $r_2$.*

**6.** Investigators studying the relationship between cigarette smoking and blood pressure in adult men collected data from 6235 U.S. men aged 20 - 40, and generated the following statistics:

$$\overline{X} = 24 \quad SD_X = 5.5$$
$$\overline{Y} = 135 \quad SD_Y = 9 \quad r = 0.7$$

where $X$ = number of cigarettes per day, and $Y$ = systolic blood pressure, measured in mmHG.

(a) Use the *regression method* to estimate the average systolic blood pressure for U.S. men, aged 20 - 40 who smoke 20 cigarettes per day. *Show your work.*

*20 cigarettes per day is $(24 - 20)/5.5 \approx 0.727$ $SD_x$* ***below*** *average, so the regression method predicts that the average systolic blood pressure of men who smoke 20 cigarettes per day will be*

$$0.727 \times r \times SD_y = 0.727 \times 0.7 \times 9 \approx 4.58 \ mmHG$$

***below*** *the average of 135 mmHG. I.e., the predicted average blood pressure for these men is about 130.42 mmHG.*

(b) What is the predicted systolic blood pressure of a 28-year old man who smokes 30 cigarettes per day? Include a '*give-or-take*' number with your estimate. *Show your work.*

*30 cigarettes per day is $(30 - 24)/5.5 \approx 1.09$ $SD_x$* ***above*** *average. The regression method says that the average blood pressure of men (aged 20 - 40) who smoke this much is predicted to be*

$$1.09 \times 0.7 \times 9 \approx 6.87 \ mmHG$$

***above*** *average, or about 141.87 mmHG.*

*The SER (root mean square error of regression) for predicting blood pressure from cigarettes per day is*

$$SER = \sqrt{1 - 0.7^2} \times SD_y \approx 6.43,$$

*so the blood pressure of an* ***individual*** *28 year old man who smokes 30 cigarettes per day is predicted to be about* $141.87 \pm 6.43$ *mmHG.*

(c) Joseph is a 60-year old man who smokes about 40 cigarettes a day. Is it reasonable to predict that his systolic blood pressure is somewhere between 147 and 160 mmHG, based on the given information? *Explain your answer.*

*This is not a question of whether the calculations were done correctly. The data was collected from men aged 20 - 40, and cannot be used to predict blood pressure for a man whose age is so far outside the range of ages in the study. Especially given the possible effects of age on blood pressure (blood pressure tends to increase with age).*

[†] See Section 4 in Chapter 9.

**7.** John Smith is running for office. One week before the election, his campaign manager hires a Polling firm to survey likely voters. The firm surveyed 2700 likely voters and found that 51% favor Smith. They also found that of the 1250 women in the survey, 54% favor Smith.

You may assume that the survey was based on a simple random sample, that the population is in the millions and that to win the office, the candidate needs to win more than 50% of the votes cast.

**a.** What percentage of the men in the survey favor Smith?

$675 = 54\% \times 1250$ *of the women surveyed favored Smith, and a total of* $51\% \times 2700 = 1377$ *of the people surveyed favored him. So,* $1377 - 675 = 702$ *men surveyed favored Smith. A total of* $2700 - 1250 = 1450$ *men were surveyed, so*

$$\frac{702}{1450} \times 100\% \approx 48.41\%$$

*of the men surveyed favor Smith.*

**b.** Compute 95% confidence intervals for the percentage of women who favor Smith, the percentage of men who favor Smith and the percentage of likely voters who favor Smith.

**Women:** *The observed percentage is 54%, and the standard error is*

$$SE_W = \frac{\sqrt{0.54 \times 0.46}}{\sqrt{1250}} \times 100\% \approx 1.41\%.$$

*The 95% confidence interval for the percentage of women who favor Smith is*

$$(54\% \pm 2SE_W) = (54\% \pm 2.82\%).$$

**Men:** *The observed percentage is 48.41%, and the standard error is*

$$SE_M = \frac{\sqrt{0.4841 \times 0.5159}}{\sqrt{1450}} \times 100\% \approx 1.31\%.$$

*The 95% confidence interval for the percentage of men who favor Smith is*

$$(48.41\% \pm 2SE_M) = (48.41\% \pm 2.62\%).$$

**All:** *The observed percentage is 51%, and the standard error is*

$$SE_A = \frac{\sqrt{0.51 \times 0.49}}{\sqrt{2700}} \times 100\% \approx 0.96\%.$$

*The 95% confidence interval for the percentage of all likely voters who favor Smith is*

$$(51\% \pm 2SE_A) = (51\% \pm 1.92\%).$$

*(*) The following two questions are based on the NAEP (National Assessment of Educational Progress) 2009 survey of science proficiency in 4th, 8th and 12th grades. The scores were scaled so that the national average in all three grade levels was set to 150 (out of 300).*[‡]

**8.** The national averages for females and males in all three grade levels are summarized in the table below (with standard errors in parentheses).

[‡] To serve as a benchmark for future tests.

| Gender | Fourth grade | Eighth grade | 12 grade |
|---|---|---|---|
| Male | 151 (0.3) | 152 (0.4) | 154 (0.9) |
| Female | 149 (0.3) | 148 (0.3) | 146 (0.9) |

**a.** Construct 95%-confidence intervals for the average score in each category in the nation as a whole.

*The Standard errors are given so the confidence intervals are:*

| *Gender* | *Fourth grade* | *Eighth grade* | *12 grade* |
|---|---|---|---|
| *Male* | $(151 \pm 0.6)$ | $(152 \pm 0.8)$ | $(154 \pm 1.8)$ |
| *Female* | $(149 \pm 0.6)$ | $(148 \pm 0.6)$ | $(146 \pm 1.8)$ |

**b.** In which grade is the difference between the average score of males and and average score of females most ***significant***, statistically speaking? Explain you answer.

*The difference is the most* significant *if it is the least likely to be explained by chance, under the assumption that in the nation as a whole there is no difference.*

*To see in which grade the difference is the most significant, we look for the test statistic that is largest (as opposed to the largest observed difference, though it could end up being the same).*

*The test statistic in each grade is given by*

$$z = \frac{\textit{observed difference}}{SE_{\textit{diff}}},$$

*so we have the following z-scores*

**Fourth grade:** $z_4 = \dfrac{151 - 149}{\sqrt{0.3^2 + 0.3^2}} \approx 4.714.$

**Eighth grade:** $z_8 = \dfrac{152 - 148}{\sqrt{0.4^2 + 0.3^2}} = 8.$

**Twelfth grade:** $z_{12} = \dfrac{154 - 146}{\sqrt{0.9^2 + 0.9^2}} \approx 6.285.$

*The differences in all three grades are statistically significant (all the z-scores are greater than* 3*, so they are in fact highly significant), but the difference in eighth grade is most significant (from a statistical point of view), even though the nominal difference in twelfth grade was greater. (There is more variation in both the male and female scores in twelfth grade, making the standard error for the difference bigger and lowering the statistical significance of the difference.)*

**9.** The 2009 NAEP science proficiency survey also breaks down the sample data by race and ethnicity. A researcher sees that the survey reveals that the average score of *white* 12th graders is 159 and the average score of *Asian/Pacific Island* 12th graders is 164.

He concludes that the national *percentage* of Asian/Pacific Islander 12th graders who are proficient or above in science is likely higher than the national percentage of White 12th graders who are proficient or above in science.

To test this conclusion, he breaks down the 2009 survey data for 12th graders by proficiency levels and finds that

- the sample percentage of white 12th grade students who are proficient or above in science is 27%, with a standard error of 0.8%;

- the sample percentage of Asian/Pacific Island 12th grade students who are proficient or above in science is 36%, with a standard error of 3.8%.

**a.** What are the appropriate null and alternative hypotheses the researcher should use to test his belief?

**Null:** *The percentage of Asian/Pacific Islander 12th graders who are proficient or above is the same as the percentage of White 12th graders who are proficient or above.*

**Alternative:** *The percentage of Asian/Pacific Islander 12th graders who are proficient or above is higher than the percentage of White 12th graders who are proficient or above.*

**b.** What is the test statistic?

$$z_{\text{diff}} = \frac{36\% - 27\%}{\sqrt{(3.8\%)^2 + (0.8\%)^2}} \approx 2.3.$$

**c.** What is the p-value?

*P-value* $= P(z \geq z_{diff}) = P(z \geq 2.3) = 1.07\%$ *(This is the area under the normal curve to the right of* 2.3.*)*

**d.** What do you conclude? Are the results significant? Highly significant? Neither?

*The difference is significant, though technically not quite* highly *significant, because the P-value is less than 5% but not less than 1%. Nonetheless, we reject the null hypothesis, and conclude that the observed difference is not due to chance. There is a higher percentage of Asian/Pacific Islander 12th graders than White 12th graders who are proficient or above in science.*

**e.** What, if anything, is wrong with this study/test of significance?

*This test verges on data snooping. We used the same data that led to our hypothesis to test the hypothesis. It would be better if we tested the hypothesis on an independent set of data. That being said, the (limited) conclusion that in the current national population of 12th graders, the percentage of A/PI students who are proficient or above in science is higher than the percentage of W students who are proficient or above in science is very likely to be valid. If we want to test a more general conclusion about the relative proficiency of these two ethnic groups in 12th grade, we should find independent data (independent from the data above).*

**10.** As part of a class project, a statistics student at a large university (15,000 students — 9000 men and 6000 women), went to the central plaza of the campus at noon one day, approached 100 students and asked them where they went to high school. His sample included 51 women and 49 men. Is it likely that the student's sampling procedure was like taking a simple random sample? Justify your answer as precisely as possible (using numbers, probability, etc.).

*The key observation here is that the sample percentage of men is* 49%*, while the population percentage is* 60% *(*$9000/15000 = 0.6$*). This seems unlikely if the student's sampling procedure was like taking a simple random sample. In that (hypothetical) case, it would be like drawing 100 tickets at random,* without *replacement from a 0-1 box of 15000 tickets, where 60% of the tickets are* $\boxed{1}$ *s and 40% are* $\boxed{0}$ *s. The question now becomes:*

***How likely is it to draw 49%*** $\boxed{1}$ ***s (or less) from a box with 60%*** $\boxed{1}$ ***s, in 100 random draws?***

*To answer this question, we use the* ***Normal Approximation****. The SD of the box is*

$$SD = \sqrt{0.6 \times 0.4} \approx 0.49$$

*and the* $SE_{\%}$ *for 100 draws from this box is*

$$SE_{\%} = \frac{0.49}{\sqrt{100}} \cdot 100\% = 4.9\%.$$

*(Technically, the* $SE_{\%}$ *is* ***slightly*** *smaller, because (a) I rounded up in the estimate of the SD and (b) the draws are done without replacement, but since there are 15000 tickets in the box and only 100 are drawn, the correction factor is very close to 1.)*

*According to the normal approximation, the probability of drawing 49% (or fewer)* $\boxed{1}$ *s from this box is about equal to the area under the normal curve to the left of*

$$z = \frac{49\% - 60\%}{4.9\%} \approx -2.245,$$

*which is approximately 1.22% (using the table in the book).*

*To summarize, the probability that a simple random sample of students from this University would have 49% men is about* 0.0122*, and we can conclude that the student's sample in this case was almost certainly* ***not*** *a simple random sample.*

*Indeed, from the description, it is clear that this was a* ***sample of convenience*** *(and biased towards women).*

**11.** According to the 1999 census, the median household income in the city of San Diego was \$46,500. In 2004, a high-end grocery chain hires a statistical research firm to corroborate their marketing consultant's claim that median household income has gone up since 1999. The research firm takes a simple random sample of 600 San Diego households and finds that 55% of the sample households have incomes above \$46,500.

Was the consultant right?

*To answer the question, we use a test of significance.*

- ***Null hypothesis****: The median income has not gone up since 1999. I.e.,* 50% *of the households in San Diego have incomes above \$46,500, (and* 50% *have incomes below this level).*

  ***Alternative hypothesis****: The median income has gone up, so more than* 50% *of the households have incomes above \$46,500.*
- ***Null-hypothetical box model****: A 0-1 box with a* $\boxed{1}$ *for every household in San Diego (in 2004) with income above \$46,500 and a* $\boxed{0}$ *for every household in San Diego (in 2004) with income below \$46,500. The null hypothesis says that* 50% *of the tickets in this box are* $\boxed{1}$ *s.*
- ***Data****: The sample percentage of households with incomes above \$46,500 is* 55%*.*
- ***Test statistic****: The observed percentage is* 55% *and the null-hypothetical expected percentage is* 50%*. Furthermore, the SD of the null-hypothetical box is* $\sqrt{1/2 \times 1/2} = 1/2$*, so the standard error is* $SE_{\%} = \frac{0.5}{\sqrt{600}} \times 100\% \approx 2.04\%$.

  *Hence the test statistic is*

$$z = \frac{\text{observed }\% - \text{expected }\%}{SE_{\%}} = \frac{55\% - 50\%}{2.04\%} \approx 2.45.$$

- ***P-value*** *(observed significance level): The P-value here is the area under the normal curve to the right of* $z = 2.45$ *which is about* $\frac{100\% - 98.57\%}{2} \approx 0.715\%$.
- ***Conclusion:*** *The P-value is very low (less than 1%), so we reject the null hypothesis and conclude that the consultant was right — the median income in 2004 was higher than $46,500.*

**12.** Suppose that a fair die is rolled 600 times.

**a.** What is the expected number of ⚀s?

*Given that the die is fair, the probability of observing a ⚀ on any given roll is* $1/6$*, and the* ***expected number*** *of ⚀s is therefore equal to*

$$\frac{1}{6} \cdot 600 = 100.$$

**b.** What is the probability that a ⚀ is observed between 95 and 105 times?

*The SD of the 'die-box' is* $\sqrt{1/6 \times 5/6} \approx 0.373$*, and the SE for the* **number** *of ⚀s in 600 draws is* $SD \times \sqrt{600} \approx 9.129$*. By the normal approximation, the probability of observing between 95 and 105 ⚀s in 600 draws is therefore approximately equal to the area under the normal curve between*

$$\frac{94.5 - 100}{9.129} \approx -0.60 \quad \textit{and} \quad \frac{105.5 - 100}{9.129} \approx 0.60$$

*which is 45.15%.*

**c.** What is the probability that more than 110 ⚀s are observed?

*Once again, we invoke the normal approximation and conclude that this probability is approximately equal to the area under the normal curve to the right of*

$$\frac{110.5 - 100}{9.129} \approx 1.15$$

*which is*

$$\frac{100\% - 74.99\%}{2} = 12.505\%.$$

**13.** There are about 25,000 high schools in the United States and each high school has a principal. These 25,000 high schools also employ a total of about one million teachers. As part of a national survey of education, a simple random sample of 625 high schools is chosen.

(a) In 505 of the sample high schools the principal has an advanced degree. If possible, find an approximate 95% confidence interval for the percentage of all 25,000 high school principals who have advanced degrees. If this is not possible, explain why not.

*The sample percentage of principals with advanced degrees is* $505/625 \times 100\% = 80.8\%$*. The sample SD in this case is* $\sqrt{0.808 \times 0.192}$*, so the Standard error (for percentage) is*

$$SE_{\%} \approx \frac{\sqrt{0.808 \times 0.192}}{\sqrt{625}} \times 100\% \approx 1.58\%.$$

*Hence a 95% confidence interval for the percentage of high schools whose principal has an advanced degree is*

$$(80.8\% \pm 3.16\%) = (77.64\%, 83.96\%).$$

(b) As it turned out, the 625 sample high schools described above employed a total of 12,000 teachers, of whom 6,500 had advanced degrees. If possible, find an approximate 95% confidence interval for the percentage of all one million high school teachers with advanced degrees. If this is not possible, explain why not.

*The sample of 12,000 teachers in this hypothetical example is **not** a simple random sample of U.S. high school teachers — taking all of the teachers from a random sample of high schools is not the same thing as a random sample of teachers from the whole country — it is a cluster sample. The methods we have been using to find the standard errors (and the related probabilities) do not apply in this case, so we cannot find a 95% confidence interval using these methods.*