Math 207
Introduction to Statistics

Review Exercises

1. Four hundred draws are made at random with replacement from a box of numbered tickets; 150 are positive. Someone tells you that 50% of the tickets in the box show positive numbers. Do you believe it? Answer yes or no, and explain using a P-value. Then find a 95% confidence interval for the percent of positive tickets in the box.

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 400 plays, finding 200 red numbers. Is that too many reds? Or chance variation? Formulate the null and alternative hypotheses as statements about a box model. Compute z and P.

3. Tuddenham and Snyder obtained the following results for 66 California children at ages 6 and 18.

average height at age 6	$\approx$	3 feet 10 inches,	SD	$\approx$	1.7 inches	
average height at age 18	$\approx$	5 feet 10 inches,	SD	$\approx$	2.5 inches	$r \approx 0.80$

The scatter diagram is football-shaped.

(a) Find the regression line and RMS error for predicting height at 18 from height at 6.

(b) Find the regression line and RMS error for predicting height at 6 from height at 18.

(c) Of children who were 4 feet at age 6, 95% of those had heights at age 18 in what range?

4. Laser altimeters can measure elevation to within a few inches, without bias, and with no trend or pattern to the measurements. As part of an experiment, 25 readings were made on the elevation of a mountain peak. These averaged out to 81,411 inches, and their SD was 30 inches.

a) The elevation of the mountain peak is estimated as ________; this estimate is likely to be off by ________ or so.

b) (T/F) 81,411$\pm$ 12 inches is a 95%-confidence interval for the elevation of the mountain peak.

c) (T/F) 81,411$\pm$ 12 inches is a 95%-confidence interval for average of the 25 readings.

d) (T/F) There is about a 95% chance that the next reading will be in the range 81,411 $\pm$ 12 inches.

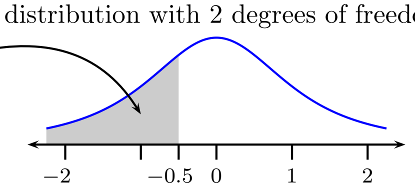
e) (T/F) About 95% of the readings were in the range 81,411 $\pm$ 12 inches.

f) If another 25 readings are made, there is about a 95% chance that their average will be in the range 81,411 $\pm$ 12 inches.

5. A coin is to be flipped 100 times. What is the chance of getting exactly 97 heads? Simplify your answer as much as possible. What is the chance of getting at least 97 heads?

6. A spectrophotometer is correctly calibrated if its average measurement is 70 ppm. Three test measurements are taken. The values are 71, 83, and 74. Use a t-test to see whether the instrument is correctly calibrated. State the null and alternative hypothesis. Calculate the P value. (Hint: $\sqrt{39} \approx 6$.) Use the table and plot below to calculate areas under t-distributions.

Degrees of Freedom	t			
	-2.0	-1.5	-1.0	-0.5
1	0.148	0.187	0.250	0.352
2	0.092	0.136	0.211	0.333
3	0.070	0.115	0.196	0.326



t distribution with 2 degrees of freedom

7. A study reports that freshmen at public universities work 10.2 hours a week for pay, on average, and the SD is 8.5 hours; at private universities, the average is 8.1 hours and the SD is 6.9 hours. Assume these data are based on two independent simple random samples, each of size 1,000. Is the difference between the averages due to chance? If not, what else might explain it?

8. Suppose that you take a large number of independent random samples from a box of numbered tickets. Fill in the blanks.

a) The histogram of the averages of the samples will will look approximately like a ________________.

b) The histogram of the sums of the samples will will look approximately like a ________________.

c) Suppose that the tickets are all 0s and 1s. The histogram of the percent of 1s in the samples will will look approximately like a ________________.

9. One hundred draws will be made at random with replacement from the box

| 0 0 2 2 2 6 |

a) What is the smallest that the sum can be?

b) What is the largest that the sum can be?

c) The sum of the draws will be around ______________, give or take ______________ or so.

d) The chance that the sum will be less than 175 is around ______________ %. Justify your answer.

10. An experiment consists of rolling a die 900 times and recording how often the die shows 1 or 6.

a) Write down a box model to represent the experiment.

b) Compute the mean and SD of the box.

c) What is the expected number of times the die will show a 1 or 6?

d) What expected percentage of times the die will show a 1 or 6?

e) The number of times that the die will show a 1 or a six will be around ______________,

give or take ______________ or so.

f) The chance that the percent of times will be between 31% and 35% is about ______________ %.

Solutions

1. The observed fraction is $150/400 = 37.5\%$. The null hypothesis is that 50% of the tickets are positive and that the difference between the expected and observed values is due to chance. If the null hypothesis is true, then the SD of the box is $\sqrt{0.5 \times 0.5} = 0.5$. So, the expected value of the percent of positives is 50% and the SE is $0.5/\sqrt{400} = 2.5\%$. That gives a z-score of $(37.5 - 50)/2.5 = -5$. The p-value is $\text{pnorm}(-5) \approx 0.000003\%$. That supports the alternative hypothesis. So, I don't believe the 50% claim.

To get a 95% confidence interval on the true percent of positives, use 37.5% to compute $\text{SD} = \sqrt{0.375 \times (1 - 0.375)} = 0.47$ and $\text{SE} = 0.47/\sqrt{400} = 2.35\%$. So, the confidence interval is $37.5\% \pm 2 \times 2.35\% = 37.5\% \pm 4.7\%$.

2. The observed percent of reds is $200/400 = 50\%$. The box model has 18 ones (representing the reds) and 20 zeros (representing the other possibilities). The average of the box is $18/38 = 47.37\%$ and its SD is $\sqrt{.4737 \times (1 - .4737)} \approx 0.5$. The expected value for the percent of reds is 47.37% and the SE for 400 spins is $0.5/\sqrt{400} = 2.5\%$. The null hypothesis is that the observed difference (50% vs 47.37%) is due to chance. The alternative is that the difference is not due to chance. The z-score is $(50 - 47.37)/2.5 = 1.052$ and the p-value is $1 - \text{pnorm}(1.052) = 14.6\%$. This supports the null hypothesis. The roulette wheel seems to be working correctly.

3. a) Use the age 18 data for y and the age 6 data for x. The regression line is $y - 70 = 0.8 \times (2.5/1.7) \times (x - 46)$ That is $y - 70 = 1.18\,(x - 46)$. The RMS error is $\text{SD}_y\sqrt{1 - r^2} = 2.5\,\sqrt{1 - 0.8^2} = 1.5$ inches.

b) Use the age 6 data for y and the age 18 data for x. The regression line is $y - 46 = 0.8 \times (1.7/2.5) \times (x - 70)$ That is $y - 46 = 0.544\,(x - 70)$. The RMS error is $\text{SD}_y\sqrt{1 - r^2} = 1.7\,\sqrt{1 - 0.8^2} = 1.02$ inches.

c) Use the formulas from a). If $x = 48$ inches, then the predicted y is 72.36 inches. 95% of the data should fall in the range $72.36 \pm 2 \times 1.5$ inches.

4. a) First blank is the observed value: 81,411 inches. The second blank is the standard error: $30/\sqrt{25} = 6$ inches.

b) True. observed value $\pm$ 2 standard errors.

c) False. If we have 25 measurements, we can take their average. There is no $\pm$.

d) False. The SD for individual readings is 30 inches. So, this answer should be ± 60 inches.

e) False. Same reason as d).

f) False. We are 95% sure that the true value will be in a confidence interval. We don't have that confidence that the next average reading will be in our previous confidence interval.

5. For exactly 97 heads, $\dfrac{100!}{97!\,3!}\left(\dfrac{1}{2}\right)^{97}\left(\dfrac{1}{2}\right)^{3} = \dfrac{100 \cdot 99 \cdot 98}{3 \cdot 2}\left(\dfrac{1}{2}\right)^{100} = 161{,}700\left(\dfrac{1}{2}\right)^{100}$ It's a really small number. In R you can compute it with dbinom(97, 100, 0.5).

For 97 or more heads, $\dfrac{100!}{97!\,3!}\left(\dfrac{1}{2}\right)^{97}\left(\dfrac{1}{2}\right)^{3} + \dfrac{100!}{98!\,2!}\left(\dfrac{1}{2}\right)^{98}\left(\dfrac{1}{2}\right)^{2} + \dfrac{100!}{99!\,1!}\left(\dfrac{1}{2}\right)^{99}\left(\dfrac{1}{2}\right)^{1} + \dfrac{100!}{100!\,0!}\left(\dfrac{1}{2}\right)^{100}\left(\dfrac{1}{2}\right)^{0}$ which is $(161{,}700 + 4950 + 100 + 1)\,(1/2)^{100}$.

6. The observed average is $(71 + 83 + 74)/3 = 76$. The observed sd is $\sqrt{\frac{(71-76)^2+(83-76)^2+(74-76)^2}{2}} = 6.2$. The standard error (SE) is $6.2/\sqrt{3} = 3.6$. The null hypothesis is that the difference between the observed (76) and expected (70) averages is due to chance. The alternative is that the difference is not due to chance. The test statistic is $(76 - 70)/3.6 = 1.7$. The p-value is $1 - \text{pt}(1.7, \text{df} = 2) = 11.6\%$. This supports the null hypothesis.

7. This is a two-sample z-test. Box A contains work hours for freshmen at public universities. Box B contains work hours for freshmen at private universities. The sample from box A had average 10.2 hours and SD of 8.5 hours. The sample from box B had average 8.1 hours and SD of 6.9 hours. The SEs for the two boxes are estimated to be $8.5/\sqrt{1000} = 0.27$ hours and $6.9/\sqrt{1000} = 0.22$ hours. The SE for the difference is $\sqrt{0.27^2 + 0.22^2} = 0.35$ hours. This gives a z-score of $(10.2 - 8.1)/0.35 = 6$. The p-value is $1 - \text{pnorm}(6) \approx 0$. So, the difference is probably not due to chance.

8. By the Central Limit Theorem, the answer for each blank is *normal curve*.

9. a) $100 \times 0 = 0$.

b) $100 \times 6 = 600$.

c) The mean and SD of the box are both 2. The expected value of the sum is $100 \times 2 = 200$. The standard error for the sum is $\sqrt{100} \times 2 = 20$.

d) Convert 175 to a z-score: $(175 - 200)/20 = -1.25$. So the answer is pnorm$(-1.25) = 10.6\%$.

10. a) The box has 4 zeros (representing 2, 3, 4 and 5) and 2 ones (representing 1 and 6).

b) The mean is $1/3$ and the SD is $\sqrt{2}/3 = 0.47$.

c) The expected value of the sum is $900 \times (1/3) = 300$.

d) The expected value of the percent is 33.3%.

e) The SE for the sum is $\sqrt{900}\,\sqrt{2}/3 = 30\,\sqrt{2}/3 = 10\sqrt{2} = 14.1$. So, the answer is 300 ± 14.1.

f) The SE for the percent is $0.47/\sqrt{900} = 1.6\%$. Convert 31% and 35% to z-scores: $(31 - 33.3)/1.6 = -1.4$ and $(35 - 33.3)/1.6 = 1.1$. So, the answer is pnorm(1.1) – pnorm$(-1.4) = 78.4\%$.