## Green / Stat102

## Answers to Review problems

1 Dichotomous data: a variable that can take on only two distinct values, typically zero and one. Such data arise in the analysis of sample proportions and require statistical techniques slightly different from those applied to continuous data.

Method of least-squares: an estimation technique or algorithm for producing guesses about underlying population parameters. The estimates are chosen because they minimize the sum of squared deviations between actual and fitted values. Widely used to produce regression lines.

Experiments: a technique for discerning cause and effect that involves the random assignment of units of observation to treatment and control conditions. An experiment is informative to the degree that it can be replicated both under controlled conditions and outside the lab.

CLT: Suppose one has an underlying population with mean  $\mu$  and standard deviation  $\sigma$ . The sampling distribution for means drawn from samples of size n, as n becomes large, approaches a *normal* distribution with mean  $\mu$  and standard deviation  $\sigma/sqrt(n)$ . A mean drawn from a normal distribution always has a normal sampling distribution regardless of sample size, but the CLT implies that a mean drawn from a nonnormal distribution takes on a normal sampling distribution when the sample size is large. Of course, how large "large" must be depends on how nonnormal the underlying population is.

Median vs. mean: if X is a random variable sorted in ascending order, the median is the "middle" observation of X. The mean, on the other hand, is the sum of the values of X divided by the number of observations. The median is resistant to extreme observations; the mean is not. Both can be useful statistics when drawing inferences about the mean of an underlying population.

2. Histogram

In forming a "density" histogram, I put my own cutpoints into the Minitab program so that the results would mirror the way the categories were set up in the problem.



Note the way in which the last bar is generated: the area reflects the fact that .15/200=.00075

## 3. $H_0: \pi_1 = \pi_2$ $H_a: \pi_1$ not equal to $\pi_2$

Test statistic = p1 - p2 = .05. The null suggests that this difference should be zero. In terms of standard errors, how far is .05 away from zero? Under the null, the best guess of  $\pi_1 = \pi_2$  is .375 (combining the two samples). To get the standard error of the observed difference we take

$$\sqrt{\frac{\pi_0(1-\pi_0)}{n_1} + \frac{\pi_0(1-\pi_0)}{n_2}} = \sqrt{\frac{.375(1-.375)}{800} + \frac{.375(1-.375)}{800}} = .0242$$

To reject the null with a 2-tailed test at the  $\alpha = .05$  level, we must obtain a test statistic at least +/-1.96 standard errors away from zero. Our test statistic, .05, is (.05/.0242) = 2.06 standard errors away from 0. (Using the normal approximation to the binomial, we find that the probability that a true difference of zero would generate an observed difference of as great as .05 in absolute value is about 4%.) Thus, we reject the null hypothesis of no opinion change at the 5% level of significance.

Variable	X	N	Sample p	
C3	280	800	0.350000	
C4	320	800	0.400000	
Estimate for p	o(C3) - p	c(C4):	-0.05	P-Value = 0.039
95% CI for p(C	C3) - p(C	c4):	(-0.0973799, -0.00262013)	
Test for p(C3)	- p(C4)	= 0	(vs not = 0): Z = -2.07	

4a. Note that 50% is .1/.183=.55 standard normal units away from .4. The probability that the sample proportion is at least this large is 1-.71=.29

b. 35% is -.05/.183=-.27 standard normal units away from .4. The probability that the sample proportion is below this level is .39.

c. To find the proportion between .37 and .46, subtract the proportion below .37 from the proportion below .46. The first proportion is -.03/.183=-.16 standard units below .4, comprising an area of .44 under the normal curve. The second proportion is .06/.183=.33 standard units away from .4, comprising .63 of the area. Subtracting .44 from .63 gives .19.

5. The correlation is -.14. This correlation remains unchanged by changes in units (as long as the transformations are linear). So, for example, one gets the same correlation when one uses housing prices in dollars or in thousands of dollars. The correlation is unaffected because the variables are, in effect, converted to standardized variables for purposes of calculation.

A scatterplot shows a weak relationship.



6. Slope: For each dollar of tax incentives per \$1000 of income, there is an expected increase of \$0.24 in savings per \$1000 of income. Intercept: when there are no tax incentives, savings is expected to be \$7.8 per \$1000 of income. R-square: tax incentives account for 85.7 percent of the observed variance in savings rates *in this sample*.

MTB > Regress 'savings' 1 'taxinc'. The regression equation is savings = 7.80 + 0.240 taxinc Predictor Coef Stdev t-ratio р 7.8000 0.002 Constant 0.6928 11.26 taxinc 0.24000 0.05657 4.24 0.024 s = 0.8944R-sq = 85.7% R-sq(adj) = 81.0%

7.

Pı	Pro-choice		Pro-life	
Voted for Smith	550	69%	350	36%
Voted against Smith	250	31%	630	64%
		100%		100%

Failing to control for partisanship exaggerates the relationship between abortion stance and the vote. Whereas in the three-way table, the gap between pro-choice and pro-life voters was nonexistent *within parties*, the two way table shows it to be rather large. The reason is that party is correlated with liberalism on this issue.

8. P(A) is the probability that A occurs. P(A|B) is the probability that A occurs, given that B occurs. (Don't confuse these two probabilities: The probability that it is raining P(A) is very different from the probability that it is raining given that it is cold outside P(A|B), which in turn is different from the probability that there it is cold outside given that it is raining P(B|A).)

9. Bayes' rule:

10. Naturally, the problem of response bias makes it difficult in this instance to draw a simple random sample of responses. Putting this aside, we can uses the formula for the SD of a proportion:

$$\sqrt{\frac{\pi(1-\pi)}{n}}$$

To calculate the appropriate sample size, we set this formula equal to .05 and solve for N.

$$p(1-p)/n = (.05)^2$$

$$n=p(1-p)/.0025$$
Our answer depends on what value of p we assume; p=.5 is the most conservative guess.  

$$n=.25/.0025=100$$

11. Justices are appointed at a rate of 2/4=.5 per year. The standard deviation is 1.2 per four years or .3 per year. (The standard deviation of aX = a (standard deviation of X).)

We could also have worked out the problem using variances. The rule for calculating the variances of products is: VAR(aX)= $a^2$ VAR(X). Thus, if the variance of x is (1.2)<sup>2</sup>, then the variance of x/4=(1/16)(1.2)<sup>2</sup>=.09. The standard deviation is the square root of .09=.3.