

STAT:2010/4200, Statistical Methods and Computing  
 Spring 2016, Instructor: Cowles  
 Midterm 3

Show your work on any problems that involve calculations.

Name: Solutions Course no. (2010 or 4200) \_\_\_\_\_

1. CNN exit polls from the New York Republican primary showed the following results broken down by gender. <http://www.cnn.com/election/primaries/polls/ny/Rep>

Candidate	Females	Males	
Cruz	63	75	138
Kasich	118	123	241
Trump	240	338	578
	421	536	957

If the respondents to the exit poll were a simple random sample from the population of voters in the NY Republican primary, we can use these data to test the null hypothesis that the proportion of women is the same in the population of all NY Cruz voters, the population of all NY Kasich voters, and the population of all NY Trump voters. Refer to the SAS output for this problem in answering the following questions.

- (a) Write the null hypothesis using conventional symbols.

$H_0: p_c = p_k = p_t$

- (b) The expected count in cell for females who voted for Trump is 254.27. Show how this number was calculated using other numbers in the table.

Two equivalent ways:  
 $\frac{421}{957} \times 578$  or  $.4399 \times 578$

- (c) Are the rules of thumb met so that we can trust the Chi square test procedure? State each rule of thumb and how the data meets or does not meet it.

Each population is a least 10 times as big as its respective sample: Yes, there were at least 1380 Cruz voters, 2410 Kasich voters, and 5780 Trump voters in the NY Republican primary.

all expected counts are  $\geq 5$ : Yes, 60.7, 77.3, 106.0, 135.0, 257.3, 323.7.

- (d) At significance level  $\alpha = .05$ , does the data provide evidence against the null hypothesis? Explain, citing specific numbers from the SAS output.

No, we cannot reject  $H_0$ . The p-value for the  $\chi^2$  test is 0.1352 >  $\alpha = .05$ .

*13/14 leave out Ho*

2

2

2

*13/14 leave out pop size*

2

2. In August 2010, the *Columbus Dispatch* newspaper tested water samples from 20 state park swimming areas in Ohio for bacteria that may make swimmers ill. Of the 20 swimming areas tested, 13 were found to have unsafe levels of the bacteria. Assume that the swimming areas tested represent a random sample of park swimming areas throughout the state. The newspaper investigators wished to draw conclusions about what proportion of all park swimming areas in the state have unsafe bacteria levels.

- (a) Are the rules of thumb for the normal approximation method of computing confidence intervals met? State each rule of them and how the data does or does not meet it.

3  
 Population size is at least 10 times sample size: probably not met. I doubt that there are 200+ state park swimming areas in OH.

Number of successes  $\geq 10$ : Yes, 13.

Number of failures  $\geq 10$ : No, 7.

- (b) Compute a 90% c.i. using the plus-four method. (Numeric answer; show your work.)

3  

$$\hat{p} = \frac{13+2}{20+4} = .675 \quad \hat{p} \pm 1.645 \sqrt{\frac{\hat{p}(1-\hat{p})}{24}}$$

$$(0.462, 0.788)$$

- (c) What quantity are we 90% confident lies in the interval? (Circle one.)

- i. The proportion of park swimming areas in the sample that were found to have unsafe bacterial levels.  
 ii. The proportion of park swimming areas in the whole state that have unsafe bacteria levels.  
 iii. Neither of the above.

- (d) How large a sample of swimming areas would have been needed to get a 90% confidence interval of width no greater than 0.04? (Numeric answer; show your work.)

If width = .04, then margin of error is .02.

3  
 2 if use  $n=204$   

$$.02 = 1.645 \sqrt{\frac{p^*(1-p^*)}{n}} \rightarrow n = \frac{1.645^2 (p^*)(1-p^*)}{.02^2}$$

10  
 For  $p^*$ , you could use any of  
 0.5 (conservative)  $\rightarrow n = 1691.3$  round up 1692  
 $\frac{p}{2}$  from above  $\rightarrow n = 1539.05 \rightarrow 1540$   
 $\frac{p}{2}$  from above  $\rightarrow n = 1585.6 \rightarrow 1586$

3. An experiment was done to compare the effects of two treatments on plant growth. A total of 30 plants were included in the study. Ten plants were randomly assigned to a control group, ten to treatment 1, and ten to treatment 2. At the end of the study, all the plants were dried. The response variable was the dry weight of each plant. The experimenters were interested in determining whether the population means of dry plant weight are equal in populations receiving treatment 1, treatment 2, or control.

Refer to the attached SAS output in answering the following questions.

2 (a) Why was ANOVA used in this analysis instead of a chi-square test?  
 Because the parameters of interest were means of a quantitative variable. Chi square is for proportions.

2 (b) Write the null hypothesis using conventional statistical symbols.  
 $H_0: \mu_c = \mu_1 = \mu_2$

2 (c) Is there anything in the SAS output that suggests that the results of the ANOVA analysis might not be trustworthy? Explain briefly.  
 Not really. There is a mild outlier in the  $trt=1$  sample, but no extreme outliers or skewness. Largest sample standard dev < twice smallest.

2 (d) At the .05 significance level, can the null hypothesis be rejected? Cite specific SAS output in justifying your answer.  
 Yes.  $p$ -value for the overall F test is  $0.0159 < \alpha = .05$ .

(e) Which population means (if any) does the analysis indicate are significantly different from each other (at the .05 significance level)?

2  $\mu_{trt1} \neq \mu_{trt2}$

