Exam 2 Sample
KEY

STA209-04: Applied Statistics

April 5, 2019

Formulas

Statistic	Standard Error
$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$
$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$
$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$\bar{x}_d$	$\frac{\sigma_d}{\sqrt{n_d}}$

Other Formula(s)

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

Confidence Level	80%	90%	95%	99%
z	1.282	1.645	1.960	2.576
$t_{df=5}$	1.476	2.015	2.571	4.030
$t_{df=10}$	1.372	1.812	2.228	2.764
$t_{df=15}$	1.341	1.753	2.131	2.602
$\chi^2_{df=1}$	1.640	2.710	3.840	6.630
$\chi^2_{df=2}$	3.220	4.610	5.990	9.210
$\chi^2_{df=3}$	4.640	6.250	7.810	11.340
$\chi^2_{df=4}$	5.990	7.780	9.490	13.280

1) [40 pts] In the late 1980s, the University of California recruited over 3,000 school-aged children for a study on the effects of ground-level ozone on the prevalence of asthma. Recruited children had no history of asthma, and were recruited from schools in 12 different southern California communities. Researchers followed the recruited children for five years and recorded those children who were medically diagnosed with asthma.

The following tables summarize demographic data for 1571 children from 5 of the 12 communities sampled.

Table 1: Observed counts of white children, male children and children from families with income greater than or equal to $50,000 within each community.

Community	**Total**	**N White**	**N male**	**N with family income $\geq$ $50,000**
Alpine	298	250	148	112
Long Beach	325	123	156	101
Riverside	369	167	174	79
Santa Maria	300	139	144	39
Upland	279	194	138	183

Table 2: Average and standard deviation of daily ozone concentration measurements within each community. The last column provides the number of measurements over which the average or standard deviation was computed.

Community	**Average ozone (ppb)**	**Standard deviation**	**N days measured**
Alpine	48.7	10.4	15
Long Beach	18.3	6.3	16
Riverside	34.0	6.7	16
Santa Maria	18.4	5.6	15
Upland	31.5	8.9	16

Table 3: Asthma diagnoses by sex for each community.

Community	**Asthma (male)**	**Asthma (female)**
Alpine	16	9
Long Beach	12	6
Riverside	17	10
Santa Maria	15	14
Upland	18	12

a) The US Environmental Protection Agency (EPA) states that an average ozone exposure of 40 ppb may have a detrimental effect on one's health. According to Table 1, the average ozone exposure in Alpine over a 15 day period was 48.7. Since this estimate is above the EPA threshold, can we conclude that the health of Alpine children is at risk? If so, why? If not, what can we do (statistically) to make this determination?

We cannot reliably conclude that the health of Alpine children is at risk using only the point estimate of 48.7. Instead, we can make this determination using either a confidence interval or one sample t-test of the mean. With the confidence interval approach, we would check whether the interval contained 40, and with the t-test approach, we would test whether the sample mean is different from 40.

b) Suppose that we were interested in computing a confidence interval for the average ozone concentration in Riverside. Assuming a fixed confidence level, how would the width of the interval compare between treating the provided standard deviation as a sample estimate versus the true population parameter?

If we were to assume that the provided standard deviation is the true population parameter, our confidence interval would be based on the standard normal distribution. In contrast, when treating the standard deviation as a sample estimate, the confidence interval would be based on a t-distribution with $df = 15$ (since the sample size provided in Table 2 is 16). This considered, the first case (true parameter standard deviation) would yield a narrower interval than the second case (sample estimate standard deviation).

c) Among the five communities, Riverside and Santa Maria appear to be the least affluent in that they have the lowest proportions of children from families with incomes greater than $50,000. Is it reasonable to conclude that the these two communities are equal in their lack of wealth? Explain and justify your response.

To answer this question we should test whether the proportion of affluent families is the same in both communities:

$$H_0 : p_r - p_s = 0; \quad H_A : p_r - p_s \neq 0$$

$$z_{test} = \frac{(\hat{p}_r - \hat{p}_s) - 0}{\sqrt{\frac{p_{pooled}(1-p_{pooled})}{n_r} + \frac{p_{pooled}(1-p_{pooled})}{n_s}}} = \frac{79/369 - 39/300}{\sqrt{\frac{118/669(1-118/669)}{369} + \frac{118/669(1-118/669)}{300}}} = 2.84$$

Since $2.84 > 1.96$, we reject the null hypothesis at the $\alpha = 0.05$ level and conclude that these two communities are not equal in their lack of wealth. Santa Maria is less affluent than Riverside.

d) Regardless of your answer to the previous question, are the ozone levels between Riverside and Santa Maria different from one another? Explain and justify your response.

To answer this question, we should perform a two sample t-test:

$$H_0 : \mu_r - \mu_s = 0; \quad H_A : \mu_r - \mu_s \neq 0$$

$$t_{test} = \frac{(\bar{x}_r - \bar{x}_s) - 0}{\sqrt{\frac{s_r^2}{n_r} + \frac{s_s^2}{n_s}}} = \frac{34 - 18.4}{\sqrt{\frac{6.7^2}{16} + \frac{5.6^2}{15}}} = 7.05$$

Since $7.05 > 2.228$ and $7.05 > 2.131$, we reject the null hypothesis at the $\alpha = 0.05$ level and conclude that the ozone levels between Riverside and Santa Maria are different from on another. Riverside has higher ozone concentrations than Santa Maria.

e) Suppose that we were interested in comparing the average ozone concentrations between each of the twelve communities sampled for this study. One approach to this kind of analysis would be to perform a statistical test of the difference in averages for each possible pair of communities. What concerns, if any, would you have with this approach?

If we were to test each pairwise comparison of means among 12 groups, we would be performing 66 tests! (I don't expect you to know how I got this amount, just know that a lot of tests would be performed) With this many tests, we are almost guaranteed to make at least one type one error.

f) Construct a table describing the counts of children with and without asthma within each community at the end of the five-year followup period.

Using Tables 1 and 3:

g) Is there a relationship between community and asthma diagnoses? Explain and justify your response.

To answer this question, we should perform a χ^2 test for association:

$$H_0 : \text{Community and Asthma Diagnosis are not associated}$$

$$H_A : \text{Community and Asthma Diagnosis are associated}$$

Community	**No Asthma**	**Asthma**
Alpine	273	25
Long Beach	307	18
Riverside	342	27
Santa Maria	271	29
Upland	249	30

	Observed		Expected	
Community	**No Asthma**	**Asthma**	**No Asthma**	**Asthma**
Alpine	273	25	273.53	24.47
Long Beach	307	18	298.31	26.67
Riverside	342	27	338.70	30.30
Santa Maria	271	29	275.37	24.63
Upland	249	30	256.09	22.91

$$\chi^2 = 6.72$$

Since $6.72 < 9.49$, we fail to reject the null hypothesis. There is not enough evidence against the claim that community and asthma diagnoses are not associated.

h) Based on your response to the previous question, does the affluence of a community appear to have a role?

Comparing the observed and expected counts for each community, it seems as if there are slightly fewer asthma diagnoses than expected in Alpine, Santa Maria, and Upland. Looking at the degree of affluence in each of these communities, it seems that there is a mixed bag: Santa Maria is the least affluent and Upland and Alpine are among the most affluent. Given this observation, it would not appear that community affluence is associated with asthma diagnoses.