Stat 217 – Case Study Solutions

Example: Suppose the state wishes to do a survey on the impact of state park closures in San Luis Obispo County. There are 5 parks of interest: Cayucos State Beach, Los Osos Oaks State Reserve, Montana de Oro State Park, Morro Bay State Park, and Pismo State Beach. Suppose they randomly selected visitors as they entered the parks between the dates of May 1 and May 30, 2009 and data were collected on the following variables: age, yearly income, gender, ethnicity, distance traveled to get to the park, # of people in group, # of visits in the last year, length of stay in the park for this visit, and a question about whether the individual would support a pay-to-play option if it were to keep the park open. Additionally, the state has attendance records for 30 randomly selected days from last summer (2008).

The data can be found in ParkSurvey.xls. For the following research questions,
(i) define the parameter(s) of interest
(ii) identify which statistical inference procedure you would use to answer the research question, and
(iii) carry out the analysis (using an applet, you may need to rearrange the columns in Excel before you paste them into the applet…)

(a) The data in columns 1 and 2 specify the length of stay and the respondent’s age group for 130 visitors. Determine whether there is significance evidence of a difference in the average length of stay for all visitors to the parks in these 2 age groups.

The goal is to compare the mean length of stay for these two groups.
(i) Let \( \mu_{18} - \mu_{30} \) represent the difference in the mean length of stay for the population of all 18-29 year old visitors and the population of all 30-39 year old visitors.
(ii) \([H_0: \mu_{18} - \mu_{30} = 0 vs. H_a: \mu_{18} - \mu_{30} \neq 0] \)
We want to compare multiple means. We could use the multiple means applet or, as the sample sizes are above 20 (both equal 65), we would use two sample \( t \)-test or ANOVA (since the alternative is two sided, these are essentially equivalent procedures)
(iii)
Winter, 2014

The simulation-based p-value and both theory-based p-values are similar, about .23.

Or we can go straight to the theory-based inference applet and compare the two means there.

[Notes: So we would fail to reject the null hypothesis (.23 > .05) and conclude that we do not have evidence of a difference average length of stay for all visitors to the parks in these two age populations. These procedures should be valid since these are independent random samples from the respective populations and both samples sizes (65 and 65) exceed 30, though we do see a bit of skewness and outliers if we look at dotplots…]
(b) Of the 65 visitors in the 18-29 year old group, 18 stayed longer than two days, compared to 8 of the 65 visitors in the 30-39 year old age group. Are the younger visitors significantly more likely to stay for longer visits?

The goal is to compare the proportion of “long” stays between these two groups

(i) Let $\pi_{18} - \pi_{30}$ represent the difference in the population proportions of a longer stay for all 18-29 visitors and all 30-39 year old visitors.

(ii) $[H_0: \pi_{18} - \pi_{30} = 0$ vs $H_a: \pi_{18} - \pi_{30} > 0]$ 

We want to compare multiple proportions. We could use the Analyzing Two-way Tables applet. We could consider the two-sample z-procedure or the Chi-squared test. The z procedure is less preferred because we do not have at least 10 successes among the 30-39-year olds. Because the alternative is one-sided, the Chi-squared test is less preferred.

Using the Analyzing Two-way Tables applet (entering the corresponding two-way table)

We get a one-sided p-value of .0220. We also see that the normal approximation would not be so useful here (and so neither would the Chi-squared distribution).

[Notes: So we would reject the null hypothesis at the 5% level (.0141 < .05) but not at the 1% level (.0141 > .01). So we have moderate evidence against the null hypothesis. Therefore, there is some evidence that the 18-29 year old population is more likely to stay more than 2 days than the 30-39 year old population.]

(c) The data in columns 3 and 4 list the parks and the number of visitors to the park on 30 (different) randomly sampled days in the summer of 2008. Determine whether there is significant evidence of any difference in the average attendance among these 5 parks in the summer of 2008.

The goal is to compare the average attendance across 5 populations. Because we want to compare more than 2 means, this calls for multiple means. We could use the Multiple Means applet or ANOVA.
(i) Let \( \mu_1 \) represent the mean number of visitors per day for Cayucos that summer. Similarly for \( \mu_2, \mu_3, \mu_4, \) and \( \mu_5. \)

(ii) To test \( H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \) (the population mean attendance is the same at the 5 parks),

vs. \( H_a: \) not all the population means are equal

Using the F statistic to summarize the differences among the groups:

The F distribution is a reasonable approximation, but either way, the p-value is miniscule!

[This is an extremely statistically significant result. We would reject the null hypothesis and conclude that we have very strong evidence that the mean attendance at the five parks that summer was not the same.]

Dotplots reveal that the number of visitors to Cayucos and Los Osos are similar to each other and much lower than the number of visitors to the other parks which hover between 175 and 275. These dotplots also reveal that the distributions are reasonably symmetric and with similar standard deviations, so that, combined with the independent random samples, the ANOVA procedure should be valid. In fact, the sample standard deviations are: 14.88, 14.13, 28.36, 18.27, 20.86 so we might worry a bit about the outliers in the Montana de Oro data.

Computing the follow-up confidence intervals:

- Compute 95% confidence interval(s)
  - Cayucos - Los Osos: (-14.66, 5.73)
  - Cayucos - Montana de Oro: (-154.99, -134.61)*
  - Cayucos - Morro Bay: (-156.76, -136.37)*
  - Cayucos - Pismo: (-169.23, -148.84)*
  - Los Osos - Montana de Oro: (-150.53, -130.14)*
  - Los Osos - Morro Bay: (-152.29, -131.91)*
  - Los Osos - Pismo: (-164.76, -144.37)*
  - Montana de Oro - Morro Bay: (-11.96, 8.43)
  - Montana de Oro - Pismo: (-24.43, -4.04)*
  - Morro Bay - Pismo: (-22.66, -2.27)*
We would conclude the Cayucos differs from all the parks except Los Osos, Los Osos differs from all the rest, and Pismo differs from Montana de Oro and Morro Bay. Basically, only Cayucos and Los Osos and Montana de Oro and Morro Bay do not show significant differences.

(d) The data in columns 5 and 6 are the number of visits to the park and the distance the visitor lives from the park, for 100 visitors. Determine whether there is a statistically significant relationship between the number of visits and the distance from the park.

We want to examine the relationship between number of visits (quantitative) and distance from park (quantitative).

(i) Let $\beta$ represent the slope between number of visits and distance from park for all visitors in the population.

(ii) To test $H_0: \beta = 0$ (no association) and $H_a: \beta \neq 0$ (is a relationship), we would use regression.

Using the Corr/Regression applet (swapping the explanatory and response variables), we get the following output.

[With a test statistic of $t = -14.94$ and a p-value < .001, we strongly reject the null hypothesis. We have very strong evidence that there is a relationship between distance and number of visits in the population. In fact, because the slope is negative, each additional mile from the park predicts a decrease of 0.2 in the number of visits. Looking at the scatterplot, we might worry about how influential the data from the visitors from large distances are.]
(e) Ten of these 100 visitors stated they traveled more than 50 miles. Estimate the proportion of all park visitors that travel more than 50 miles.

Because we just want to estimate the population proportion, not compare it to some hypothesized value, we need a confidence interval.

(i) Let $\pi$ represent the proportion of all park visitors that travel more than 50 miles.

(ii) We want to analyze **One Proportion**. Because we have 10 successes (barely!) and 90 failures, we just barely pass the sample size check for using the one-sample z-interval for proportions. So if we use the Theory-Based Inference applet:

By hand: $\hat{p} \pm 1.96 \sqrt{\hat{p}(1-\hat{p})/n} = .10 \pm 1.96 \sqrt{(0.1(.9)/100)} = (.041, .159)$

[We are 95% confident that between 4.1% and 15.9% of all park visitors travel more than 50 miles.]

If we wanted to use a simulation-based approach, we could generate a null distribution assuming a value of $\pi$ (e.g., .10 or .50) and then get the standard deviation

$\hat{p} \pm 2(.031) = .10 \pm .062 = (.038, .162)$, very similar to the theory-based confidence interval.