Stat 140: Examples of Inference for a Population Proportion

Cat Ownership

In an online survey of 50,347 volunteer households in the United States conducted in December 2011, approximately 15305 of the households, or about 30.4%, owned a pet cat. (2012 U.S. Pet Ownership & Demographics Sourcebook, American Veterinary Medical Association, see also http://news.vin.com/vinnews.aspx?articleId=31369).

(a) What is the population of interest?

(b) Is the number 0.304 a population parameter or a sample statistic? What symbol would you use for the population parameter and what symbol would you use for the sample statistic?

(c) Are the conditions met for performing inference about the proportion of all US households who own cats based on this sample?

Despite your answer above, let's proceed anyways and see what happens.

(d) The proportion of US households who own cats is something I think about a lot, and before I read about this survey I always believed that 31% of US households owned cats. Conduct a hypothesis test of whether the sample data provide evidence that the population proportion who own a pet cat differs from 31%. State the hypotheses, report the p-value, and draw a conclusion in the context of this study using an $\alpha = 0.01$ significance level. You may use the R output below:

```r
binom.test(15305, 50347, p = .31)
```

```
##
##
##
## data:  15305 out of 50347
## number of successes = 15000, number of trials = 50000, p-value =
## 0.004
## alternative hypothesis: true probability of success is not equal to 0.31
## 95 percent confidence interval:
##  0.300 0.308
## sample estimates:
## probability of success
##                  0.304
```

(e) You found a statistically significant result in the previous section. Is the difference between 31% (the value from the null hypothesis) and 30.4% (the observed percentage in the sample) practically significant (does it matter)?

(f) Interpret the 95% confidence interval for the proportion of households in the US who own cats in context.

(g) What does the phrase "95% confident" mean?

(h) Do you believe any of the results from the hypothesis test and confidence intervals above?

Cancer in the Slater School

This example has been discussed in Brodeur (1992) and Lavine (1999). The Slater school is an elementary school in Fresno, California where teachers and staff were "concerned about the presence of two high-voltage transmission lines that ran past the school. . . " (Lavine). They were particularly concerned about possibly increased cancer rates: out of 145 teachers, teachers' aides, and staff, 8 developed invasive cancer.

To address their concern, Dr. Raymond Neutra of the California Department of Health Services' Special Epidemiological Studies Program conducted a statistical analysis on the "eight cases of invasive cancer, . . . , the total years of employment of the hundred and forty-five teachers, teachers' aides, and staff members, . . . , [and] the number of person-years in terms of National Cancer Institute statistics showing the annual rate of invasive cancer in American women between the ages of forty and forty-four — the age group encompassing the average age of the teachers and staff at Slater — [which] enabled him to calculate that 4.2 cases of cancer could have been expected to occur among the Slater teachers and staff members . . . ."

Another way of phrasing this is that nationally, the probability of an individual in a similar demographic group to the school teachers and staff developing cancer was about $4.2/145 = 0.029$.

There are two public health questions at hand:

- What can we say about the risk of cancer among employees at the Slater school?
- Is this risk different from the risk among similar people nationwide?

(a) What is the population parameter in this example? What is the population?

(b) Are the conditions necessary for conducting inference about the population parameter met in this example?

(c) State the null and alternative hypotheses for testing whether the probability of developing cancer is the same at the Slater school as it is nationally.

(d) Run the necessary R code to conduct the hypothesis test you set up in part (c) and find a confidence interval. This has been set up for you in Lab 12 on Gryd.

(e) How much evidence do these data provide that there is an increased risk of cancer at the Slater school relative to the national level? What are the results of a formal hypothesis test conducted at the $\alpha = 0.05$ significance level?

(f) You should have failed to reject the null hypothesis in the previous part. Does that prove that the null hypothesis was correct, and the probability of developing cancer at the Slater school was exactly the same as it is nationally?

(g) State a 95% confidence interval for the population parameter in the context of this problem. (No need to describe what the term "95% confident" means, although of course you should know that.)

(h) Sometimes people use the results of hypothesis tests to make important decisions like whether a pharmaceutical drug will be released to market. In this example, suppose that state officials would have intervened by shutting down the school if you had found statistically significant evidence that there was a higher risk of developing cancer among employees at the school. In this case, the officials would not have intervened since you did not reject the null hypothesis in part (e). An interesting thought experiment is to consider what might have happened if one more person in the sample had developed cancer.

Conduct the hypothesis test and construct the confidence interval again, supposing that 9 people had developed cancer instead of 8. Would your conclusion for the hypothesis test change? Would your confidence interval change much?