Stats 10 Lab 5: Statistical Inference!

Do not post, share, or distribute anywhere or with anyone without explicit permission.

Confidence Interval & Hypothesis Testing

Objectives:

1. Learn how to construct confidence intervals for one proportion.
2. Learn how to conduct hypothesis tests for one proportion.

Section 1. Simple Random Sampling and Confidence Interval

The sample() function can be used to conduct simple random sampling from a population by sampling from the row numbers of a data frame.
This is done in two steps. We use the dataset - NCbirths as an example:

1. We will use function `sample()` to randomly select n numbers between 1 and 1992 (the number of babies in the NCbirths data frame). This represents choosing the babies based on ID numbers. Note that we typically use the default argument replace = FALSE to ensure we get n unique ID's.
2. We then use the selected numbers from Step 1 as an index for the rows of observations in the NCbirths data frame that we want to extract as our sample.

As an example, try out the code below, which takes a simple random sample of size 5 from the NCbirths data frame.

```
# Set the seed for reproduceability
set.seed(123)
# Select 5 numbers from 1 to 1992.
sample_index <- sample(1992, size = 5)
# Display the indices we sampled.
sample_index
## [1] 573 1570 814 1757 1870
# Extract the rows in NCbirths that correspond to sample_index.
NCbirths[sample_index, ]
```

Exercise 1

The Sweetums candy factory in Pawnee, Indiana, is under investigation for violating EPA regulations. Factory workers have improperly disposed of arsenic and sulfur waste from the candy-making process, and the contamination has reached the local water supply! We have data for arsenic and sulfur levels from the water in all houses within a 2-mile radius of the factory. Download the “pawnee.csv” file from the course website, then read it into RStudio with the following line:

```
pawnee <- read.csv("yourpath/pawnee.csv", header = TRUE)
```

Some important variables include:
- Arsenic: arsenic levels for each home in ppm
- Sulfur: sulfur levels for each home in ppm
- New_hlth_issue: Indicates “Y” if someone living at the home has experienced a major health issue after the date of contamination, “N” if no new health complications.

a. Use the head() function to print out the first few rows of this data. Then, use the dim() function to print out the number of rows and columns of this data frame.

b. Set the seed to 1337 and take a simple random sample of size 30 from the entire pawnee data frame. Save the random sample as a separate R object, and print the first few lines to make sure you saved it correctly.

c. Report the proportion of households experiencing a major health issue from your sample. Also report the population proportion of all households which experienced a new major health issue.

d. Generate confidence intervals for our sample proportion using the sample results. Produce 90%, 95%, and 99% confidence intervals for the true population proportion. Consult your lecture materials if you are unsure how to do this. You can use R and/or a calculator for this question, but please include code or calculations to show your work.

Section 2. The One-Sample z-Test for proportions

To do a (theory-based) hypothesis test (test of significance) for the proportions, we follow several steps:

1. State the null and alternative hypotheses about the population proportion, p0.
2. Check the validity conditions to apply the Central Limit Theorem.
3. From a sample of data, compute the sample proportion $\hat{p}$ and its standardized z-statistic $z = \frac{\hat{p}-p_0}{SE}$, where SE is calculated as $\sqrt{\frac{p_0(1-p_0)}{n}}$, where n is the sample size.
4. If the validity conditions hold, compute the p-value by comparing the observed z-statistic to a standard normal distribution [using normal tables or pnorm()].

Exercise 2 – Hypothesis testing with one proportion.

We will be working with a Flint dataset, which can be found on the course website. Please download the file and read it into R. You may recall that lead levels were considered dangerous if the result was greater than or equal to 15PPB. We are interested in determining if the proportion of dangerous lead levels in Flint is greater than 10%. Assume the Flint data is a random sample used to address this research question.

a. We will conduct a hypothesis test for this research question. What are the null and alternative hypotheses? Is this a one-sided or a two-sided test?

b. Calculate the sample proportion and sample standard deviation of the sample proportion of dangerous lead levels.

c. Now, calculate the SE of sample proportions, and the z-value for this test. Consult the above instructions and/or the lecture materials for guidance.

d. Using the z-statistic in (c), calculate the p-value associated with this test. You may use R's pnorm() function or a normal table, but please show all work.

e. Using a significance level of 0.05, do you reject the null hypothesis?

f. If greater than 10% of households in Flint contain dangerous lead levels, the EPA requires remediation action to be taken. Based on your results, what should you tell the EPA?