Worksheet 10

Author:

Discussants:

==

This worksheet will give you practice running hypothesis tests on correlation. Please answers the following questions and submit a compiled pdf with your answers to Moodle by 11:59pm on Sunday November 18th. Some useful symbols are: $H_0 :$, ρ, $\neq$. Some useful functions are: cor(), sample(), for(){}.

Part 1: The 1969 draft lottery

In 1969, the United States Selective Service conducted a lottery to decide which young men would be drafted into the armed forces. Each of the 366 birthdays in a year (including February 29) was assigned a draft number. Young men born on days assigned low draft numbers were drafted.

If the draft was completely fair, there should be no correlation between the draft number and the date someone was born. In the following set of exercises we will use hypothesis testing to assess whether indeed there was no correlation.

Step 0: Let's start our analyses, as usual, by visualizing the data. A data frame that contains the draft lottery information can be loaded into R using the code below. This data frame contains two variables (columns). The first column contains sequential days of the year and the second colum contains the draft number associated with that date. Create a scatter plot of the draft number as a function of the sequential date. Does there appear to be any trend in the data?

```r
# load the data into R
load("/home/shared/intro_stats/cs206_data/draft_lottery_data.Rda")


# plot the data
```

Answer:

Exercise 1: Now let's do step 1 of our null hypothesis significance tests (NHSTs) by stating the null and alternative hypotheses in symbols and in words.

Answer:

Exercise 2: Next let's do step 2 of hypothesis testing by calculating the statistic of interest and save it to a variable obs_stat. Describe what this statistic means as clearly as you can (e.g., if the statistic is negative what does that mean in terms of dates and draft numbers?).

Figure 1: Draft Lottery Image

Answer:

Exercise 3: Now let's do step 3 of hypothesis testing by creating a null distribution. Remember that you can randomly shuffle a vector v by using the function sample(v) (this works by sampling every data point without replacement which is the same as shuffling the data). You can calculate one point in the null distribution by shuffling one of the variables and then calculating the correlation. Use a for loop to repeat this process 10,000 times to generate the full null distribution. Plot a histogram of the null distribution and describe its shape. Is the center of the null distribution at a value that makes sense to you?

Answer:

Exercise 4: Now use the vector null_dist and the obs_stat to calculate the p-value by seeing the proportion of points in the null distribution that are *more extreme* than the observed statistic. Is this p-value consistent with there being no correlation between draft numbers and squential dates?

Answer:

Exercise 5: Make a judgement call as to whether you believe the draft lottery was fair. Make sure to justify your answer.

Answer:

Exercise 6: Calculate the confidence interval for the value of the correlation between squential date and draft number. Note that you can sample points in a *data frame* with replacement using: bootstrap_data_frame <- sample(draft_lottery_data, size = 366, replace = TRUE). Does the confidence interval contain 0, and would you expect it to contain 0?

```
# an example of a bootstrapped data frame, i.e,. the rows
# have been sampled with replacement you can use data frames
# like this to generate bootstrap statistics, and
# consequently a bootstrap SE*
one_bootstrap_data_frame <- draft_lottery_data[sample(1:366,
    366, replace = TRUE), ]
```

Answer:

Part 2: Lock5 questions

Lock5 exercise 4.84 (first edition): **Mercury Levels in Fish** Figure 4.26 shows a scatterplot of the acidity (pH) for a sample of n = 53 Florida lakes vs the average mercury level (ppm) found in fish taken from each lake (the full data set, called Floridalakes, can be obtained from the Lock5 website at: http://www.lock5stat.com/datapage.html). There appears to be a negative trend in the scatterplot, and we wish to test whether there is significant evidence of a negative association between pH and mercury levels.

a) What are the null and alternative hypotheses in symbols and words?

b) For these data, a statistical software package produced the following output:

r = -0.575 p-value = 0.000017

Use the p-value to give the conclusion of the test. Include an assessment of the strength of evidence and state your results in terms of rejecting or failing to reject H0 and in terms of pH and mercury.

c) Is the convincing evidence that low pH causes the average mercury level in fish to increase? Why or why not?

d) Bonus (optional) question: use R to verify that the observed correlation value and p-value and close to what are described above.

Answers

a) The hypotheses are H_0: $\rho = 0$ vs $H_a : \rho < 0$, where ρ is the correlation between pH and fish mercury levels in all Florida lakes; i.e., if the pH is higher the fish mercury levels are lower.

b) The very small p-value (0.000017) indicates that we should reject $H_0 : \rho = 0$ in favor of H_a: $\rho < 0$. There is very strong evidence of a negative correlation between mercury content of fish and acidity of Florida lakes.

c) The data are from an observational study and not an experiment, so we can't conclude that low pH causes increased mercury in the fish.

```
# load the data
load(url("http://www.lock5stat.com/datasets/FloridaLakes.rda"))



# calculate the observed statistic and p-value
```

Bonus (optional) question: Lock5 exercise 4.133 (first edition): **Hockey Malevolence** The Lock5 book describes a study that examined a possible relationship between the perceived malevolence of a team's uniforms and penalties called against the team (page 224 first edition, page 263 second edition). The code

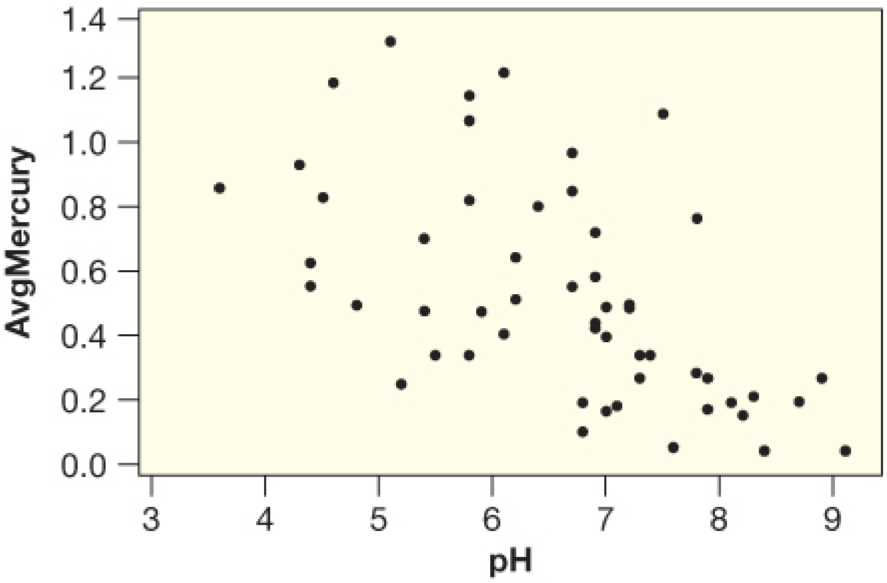

Figure 2: Figure 4.26

below loads data from this study. Please run the 5 steps of a hypothesis test to examine whether there is a statistically significant relationship between perceived malevolence of a team's uniforms and penalties called against the team using a significance level of $\alpha = 0.05$.

```
# load the data
MalevolentUniformsNHL <- read.csv("http://www.lock5stat.com/datasets/MalevolentUniformsNHL.csv")
MalevolentUniformsNHL <- MalevolentUniformsNHL[1:21, ]  # get rid of the last few entries which contain



# Do your analysis here...
```

Answers:

1)

2)

3)

4)

5)