Lab 10: Inference for numerical data: ANOVA

Template for lab report

```
download.file("http://stat.duke.edu/~cr173/Sta102_Sp14/Lab/lab10.Rmd", destfile = "lab10.Rmd")
```

Write your report, or at least run the code and create the plots, as you go so that if you get errors you can ask your TA to help on the spot. Knit often to more easily determine the source of the error.

Revised inference function

Download the inference function to use in this lab.

```
source("http://bitly.com/dasi_inference")
```

2012 Presidential Campaign Finance

On October 19, 2012, The Guardian Datablog published a post linking to data released by The Federal Election Commission on contributions made to 2012 presidential campaigns, and asked readers to help them explore these data. The original dataset can be found here, and contains information for all contributions, from over 3 million contributors. In this lab we will work with a random sample of 10,000 contributions from these data. You can see The Guardian's visualizations using the entire dataset here.

```
cont = read.csv("http://stat.duke.edu/~cr173/Sta102_Sp14/Lab/contributions.csv")
```

A detailed description of the variables can be found here. Below is the relevant information from this document. While there are 19 variables in the data set, in this lab we'll only focus on a few of them.

1. `no`: values ranging from 1 to 10,000.
2. `cmte_id`: committee ID, a 9-character alpha-numeric code assigned to a committee by the Federal Election Commission.
3. `cand_id`: candidate ID, a 9-character alpha-numeric code assigned to a candidate by the Federal Election Commission.
4. `cand_nm`: candidate name
5. `contbr_nm`: reported name of the contributor.
6. `contbr_city`: reported city of the contributor.
7. `contbr_st`: reported state of the contributor.
8. `contbr_zip`: reported zip code of the contributor.
9. `contbr_employer`: reported employer of the contributor.
10. `contbr_occupation`: reported occupation of the contributor.
11. `contb_receipt_amt`: reported contribution amount.
12. `contb_receipt_dt`: reported receipt date.
13. `receipt_desc`: additional information reported by the committee about a specific contribution.
14. `memo_cd`: 'X' indicates the reporting committee has provided additional text to describe a specific contribution. See the `memo_text`.
15. `memo_text`: additional information reported by the committee about a specific contribution

16. `form_tp`: indicates what schedule and line number the reporting committee reported a specific transaction.
17. `file_num`: a unique number assigned to a report and all its associated transactions.
18. `tran_id`: a unique identifier permanently associated with each itemization or transaction appearing in an FEC electronic file.
19. `election_tp`: This code indicates the election for which the contribution was made. EYYYY (election plus election year).

- P = Primary
- G = General
- O = Other
- C = Convention
- R = Runoff
- S = Special
- E = Recount

Note that the variable names have been kept as short as possible, while still being meaningful, and contain no spaces. This is good practice, and one you should employ in your projects.

Exercise 1 Who are the presidential candidates for whom we have contribution data? Which candidate received the highest number of contributions? Which received the lowest?

Hint: You can use a frequency table to answer this question.

Next, let's look at the distribution of number of contributions between various elections. We can use a frequency table to answer this question as well.

```
table(cont$election_tp)
```

In this sample there are 2,648 contributions for the general election, and 7,346 for the primary. There are also 5 contributions that are not labeled, and 1 contribution labeled as `P`, we'll ignore these observation in the rest of the analysis.

Republican primary - major candidates

We'll start our analysis with exploring contributions for major candidates for the Republican primary election: Mitt Romney, Ron Paul, Newt Gingrich, and Rick Santorum. We first subset the data to inlcude only these candidates, and in the second step subset further to only include data from contributions made for the primary election. We can use the `subset` function create these subsetted datasets. Remember that `|` means 'or' and `&` means 'and'.

```
# subset for major Republican candidates
rep_mjr = subset(cont, (  cont$cand_nm == "Romney, Mitt"
                        | cont$cand_nm == "Paul, Ron"
                        | cont$cand_nm == "Gingrich, Newt"
                        | cont$cand_nm == "Santorum, Rick"))

# subset for primary election
rep_mjr_pri = subset(rep_mjr, rep_mjr$election_tp == "P2012")
```

Next, we make a new frequency table to obtain the number of contributions made for just these candidates for the primary election.

```
table(rep_mjr_pri$cand_nm)
```

This table still includes names of all candidates, even though it's created using the subsetted dataset. However only the four candidates we're interested in have counts associated with them, the rest are 0s.

This is because R retains the levels of a categorical variable, even if there are no observations at a particular level, i.e. for a particular candidate. We can get rid of these empty levels using the `droplevels` function. We'll save the data with the dropped levels as a new dataset, and give it a shorter name (`pri`) to make it easier to use in the rest of the lab.

```
pri = droplevels(rep_mjr_pri)
```

Make a new table to check that dropping the levels cleaned up the dataset.

```
table(pri$cand_nm)
```

Exercise 2 Make a side-by-side box plot displaying the distributions of contribution amounts for these four candidates in the Republican primary election. Any interesting features?

The box plot reveals that some contributions are negative. We can view the comments associated with such contributions to understand why such values exist. In order to figure out which rows of the data contain negative contributions we use the `which` function, and then we view the descriptions for these spefic rows.

```
neg_index = which(pri$contb_receipt_amt < 0)
pri$receipt_desc[neg_index]
```

It looks like some of these are refunds, and some are redesignations. We'll retain these negative values in the data in order to avoid overestimating the true total and average contributions.

Exercise 3 Which of these four candidates has the highest total contribution?

Hint: Use the `by` and `sum` functions.

Exercise 4 Which of these four candidates has the highest average contribution?

Hint: Use the `by` and `mean` functions.

Next, we'll conduct an ANOVA to determine if the observed differences are in fact statistically significant.

Exercise 5 Write the hypotheses for comparing the average contribution amounts for these candidates for the primary election.

Exercise 6 What are the conditions for ANOVA? Are they met in this analysis?

Hint: Use the following code to make a normal probability plot of Newt Gingrich's contributions:

```
qqnorm(pri$contb_receipt_amt[pri$cand_nm == "Gingrich, Newt"], main = "Gingrich")
qqline(pri$contb_receipt_amt[pri$cand_nm == "Gingrich, Newt"])
```

You will need to make similar plots for the other three candidates as well.

Regardless of whether or not you find that the conditions are met, let's go ahead and run the test using

the `inference` function. Note that the `alternative` hypothesis is set to be `"greater"` since the p-value in ANOVA is defined as the area under the F curve above the observed F statistic.

```
inference(y = pri$contb_receipt_amt, x = pri$cand_nm,
          est = "mean", type = "ht", alternative = "greater", method = "theoretical")
```

The p-value for the F test is very small. Since it's less than 5%, we reject H_0. The data provide evidence of that at least one candidate's true average contribution is different than the others. This conclusion shouldn't come as a surprise since exploratory data analysis revealed that Mitt Romney had a much higher average than the other three candidates. Using the ANOVA output (under `Pairwise tests: t tests with pooled SD`) we can also determine which pairs of candidates have significantly different averages. Note that the values on this table are p-values corresponding to a two-sided test evaluating whether a pair of candidates have different average contribution amounts, e.g. p-value for the test evaluating $\mu_{PaulRyan} \neq \mu_{NewtGingrich}$ is 0.91. Note that the displayed p-values are rounded to four decimals, so that 0.0000 is not actually equal to 0, but it's very small.

Exercise 7 Determine the modified significance level ($\alpha^\star$) that should be used in these pairwise comparisons. Then, using the output from the `inference` function, determine which candidates have significantly different averages.

General Election

Next we'll shift our focus to the general (presidential) election and the contributions for Barack Obama, Mitt Romney, and Gary Earl Johnson. Let's start with making a new data set for just this election.

```
# subset for general elections and Obama, Romney, and Johnson
pres_temp1 = subset(cont, cont$election_tp == "G2012")
pres_temp2 = subset(pres_temp1, (pres_temp1$cand_nm == "Obama, Barack" | pres_temp1$cand_nm == "Romney,
# droplevels
pres = droplevels(pres_temp2)
```

Exercise 8 Conduct ANOVA to compare the average contributions for these three candidates for the presidential election. What is the conclusion? (For this question, you do not need to show how you would check the conditions.)

Exercise 9 How many contributions does this dataset contain for each of these three candidates? Do you think the results of the ANOVA you just conducted are reliable? Why, or why not?

Exercise 10 Create a new dataset that contains information only for Barack Obama and Mitt Romney, and call it `pres2`. You will use this new dataset to answer the following questions.

Exercise 11 Calculate the total contributions for Barack Obama and Mitt Romney. Which candidate has a higher total? Does this candidate also have a higher average? If not, how can you explain what's going on here?

Exercise 12 Which test should we use to compare the average contributions for these two candidates: Z, T, or ANOVA (F)? Why?

Exercise 13 Use the test you chose in the previous question to evaluate if there is a significant difference between the average contributions for Barack Obama and Mitt Romney. Make sure

to interpret your conclusion in context of the data. *Hint:* You can use the `inference` function to answer this question.

Exercise 14 Calculate a 95% confidence interval for the difference, interpret it in context, and comment on whether or not it agrees with your hypothesis test.