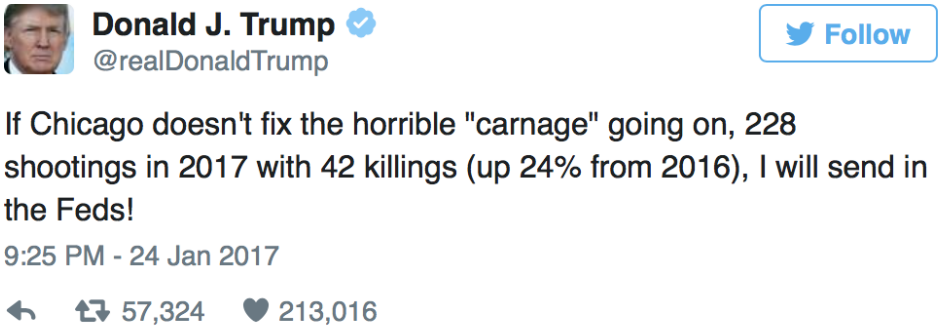
GEORGE MASON UNIVERSITY
Systems Engineering and Operations Research
SYST/OR 568: Applied Predictive Analytics, Spring Semester 2020: Midterm. Due April 3 @ Midnight.

This is a take-home open-book exam.

Honor Code: By signing my name below, I pledge my honor that I have not violated the GMU Honor Code during this examination.

Signature:

1. Chicago Crime Data Analysis (30 pts) On January 24, 2017 Donald Tramp tweeted about "horrible" murder rate in Chicago.



Our goal is to analyze the data and check how statistically significant such a statement. I downloaded Chicago's crime data from the data portal: `data.cityofchicago.org`. This data contains reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified. This data set has 6.3 million records. Each crime incident is categorized using one of the 35 primary crime types: NARCOTICS, THEFT, CRIMINAL TRESPASS, etc.. I filtered incidents of type HOMICIDE into a separate data set stored in `chi_homicide.rds`. Use `chi_crime.R` as a starging script for this problem.

a) Create a heat map for the homicide incidents. In which areas of the city you think houses are very affordable and in which they are not?

b) Create a map by plotting a dot for each of the homicide incidents. You will see similar picture as you saw with the heat plot. Look at the Hyde Park area in the south side Chicago. There is an "island" with no homicide incidents! Can you explain why? Hint: You might want open Google maps in your browser and zoom-in into this area.

c) Though president't tweet is consistent with the data (`goo.gl/VTPzFw`), observing 52 homicides in January is not that unusual. Calculate the total number of homicides for each January. Use bootstrap to estimate 95% confidence interval for the mean μ over January homicides. Is 52 within the interval? Calculate confidence interval using t-ratio. Do you think results from t-ratio based calculations are reliable?

d) The history of 2001-present data is rather short. Chicago tribune provided total number of homicides for Chicago for each month of the 1957-2014 period. Use this data set and calculate the confidence interval for μ using bootstrap and t-ratio. Further answer the following questions: (i) Assuming monthly

homicide rate follows Normal distribution, what is the probability that we observe 52 homicides or more? (ii) Do you think Normality assumption is valid? (iii) Assuming monthly homicide rate follows Poisson distribution, what is the probability that we observe 52 homicides or more?

e) There is a hypothesis that crime rates are related to temperatures (`goo.gl/nPpHwv`). Check this hypothesis using simple regression. Use linear model to regress homicide rate to the average maximum temperature. Does this relation appear significant? Perform residual diagnostics and find outliers and leverage points.

f) There is another hypothesis that rise in murder is related to the pullback in proactive policing that started in November of 2015 as a result of Laquan McDonald video release (`https://goo.gl/7cm1CC`, `https://goo.gl/WcH2uB`). I calculated total number of homicides for each day and split data into two parts: before and after video release. Using t-ratio, check the hypothesis H_0: the homicide rate did not change after video release.

was at least partly related to a pullback in proactive policing since 2015

2. A/B Testing (20 pts) Use dataset from `ab_browser_test.csv`

Here is the definition of the columns:

- `userID`: unique user ID
- `browser`: browser which was used by userID
- `slot`: status of the user (exp = saw modified page, control = saw unmodified page)
- `n_clicks`: number of total clicks user did during as a result of `n_queries`
- `n_queries`: number of queries made by `userID`, who used browser `browser`
- `n_nonclk_queries`: number of queries that did not result in any clicks

Note, that not everyone uses a single browser, so there might be multiple rows with the same `userID`. In this dataset cobination of `userID` and `browser` is the unique row identifier.

a) Count how many users in each group. How much larger (in percent) exp group when compared to control group

b) Using bootstrap, construct 95% confidence interval for mean and median of number of clicks in group exp and group control. Are the mean and median significantly different?

c) Using bootstrap, check if mean of each group has a normal distribution. Generate $B = 1000$ bootstrap samples, calculate mean of each and plot `qqplot`.

d) Use z-ratio for the means, to perform hypothesis testing, with H_0: there is no difference in average number of clicks between 2 groups

e) Mann-Whitney (`http://www.statmethods.net/stats/nonparametric.html`) is another test for comparing means, that does not require normality assumption. Use this test to check hypothesis that means are equal.

f) For each browser type and each of the 2 groups (control and exp) count the percent of queries that did not result in any clicks. You can do it be dividing sum of `n_nonclk_queries` by sum of `n_queries`. Comment your on your results.

3. Russian Parliament Election Fraud (5 pts) On September 28, 2016 United Russia party won a supermajority of seats, which will allow them to change the Constitution without any votes of other parties. Throughout the day there were reports of voting fraud including video purporting to show officials stuffing ballot boxes. Additionally, results in many regions demonstrate that United Russia on many poll stations got anomalously closed results, for example, 62.2% in more than hundred poll stations in Saratov Region.

Using assumption that United Russia's range in Saratov was [57.5%, 67.5%] and results for each poll station are rounded to one decimal point (when measure in percent), calculate probability that in 100 poll stations out of 1800 in Saratov Region the majority party got exactly 62.2%.

Do you think it can happen by a chance?

4. Back cast US Presidential Elections (15 pts) Use data from presidential polls to predict the winner of the elections. We will be using data from `http://www.electoral-vote.com/`. The goal is to use simulations to predict the winning percentage for each of the candidates. Use `election.R` script as the starter.

Report prediction as a 50% confidence interval for each of the candidates.

5. Weather Data (15 pts) I've downloaded high resolution data from the DCA Airport weather station. I am giving you sixteen years of roughly hourly observations. Use `weather.R` as a starter script.

Check out the structure. I used `lubridate::ymd_hms` to **convert** the date and time (DateUTC column) to R date time classes `POSIXct`. I also converted time to local EST time.

I just moved to Fairfax. I would like to understand what the climate is like here. But I normally end up looking at a graph that shows me the average min, mean and max temperature by month, and maybe total precipitation, and that doesn't really help me decide how my day to day life will be affected by the weather. I'm much more interested in things like:

a) How often will I be biking to work in below freezing temperatures (assuming I bike every workday to work)?

b) How many days a year will I see the sun?

c) How many days do I need a raincoat, assuming I'm only outside during my morning and evening commute?

Your task is to come up with a climate metric that you would find useful, and by using the data provided, calculate it and present it. Think how you would quantify and present the variability in your metric.

`dplyr` library provides a convinient `summary` and `filter` functions to calculate the resutls required for this problem. You can read about this library here: `https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html`. However, the syntax migh appear counter-intuitive and you might prefer doing it some other way.

6. Credit Scoring (25 pts) Credit scoring is a classic problem of classification. The goal is to use borrower/loan characteristics and previous defaults to predict performance of potential new loans. In this problem we will use the German loan/default data. It contains borrower and loan characteristics such as job, installments, etc. Using starter code given in `credit.R` do the following

a) Build 3 logistic regression models: credscore, reduced and backward

b) Use AIC and Cross-Validation to choose the best model

c) If we gain \$0.25 on every dollar when loan is payed and loose entire amount when it goes to default, show why do we use 1/5 as the cut-off rule for issuing a loan

d) What is your specificity and sensitivity at $p = 1/2$ and $p = 1/5$?

e) Interpret the ROC curve. Why we choose a cut-off rule that leads to test that is sensitive but not specific?