DS 412 Inferential Statiatics for Data Science Lab Assignments (Last Updated: 3/12/2018)

A. High-Dimensional Space

References: Foundations of Data Science by Blum, Hoperoft, and Kannan; Simulation by Ross.

1. Consider the probability density function $p(x) = \frac{c}{x^4}$ for $x \geq 1$, where c is a constant. Generate 100 random samples from this distribution and plot a histogram. How close is the average of the samples to the expected value of X?

2. Draw a 2-D plot in which the Y-axis represents $V(d)$, the volume of a d-dimensional unit ball, and the X-axis represents $d = 1, 2, 3, \ldots$. State your observations.

3. Draw a 2-D plot in which the Y-axis represents $S(d)$, the surface area of a d-dimensional unit ball, and the X-axis represents $d = 1, 2, 3, \ldots$. State your observations.

4. Draw a 3-D plot in which the Z-axis represents $V(d)$, the volume of a d-dimensional ball of radius R, and the X-axis represents d, and the Y-axis represents the radius R. State your observations.

5. Draw a 3-D plot in which the Z-axis represents $S(d)$, the surface area of a d-dimensional ball of radius R, and the X-axis represents d, and the Y-axis represents the radius R. State your observations.

6. Generate 20 points uniformly at random on a 900-dimensional sphere of radius 30. Calculate the distance between each pair of points. Then, select a method of projection[1] and project the data onto subspaces of dimension $k = 100, 50, 10, 5, 4, 3, 2, 1$ and calculate the difference between $\sqrt{k}$ times the original distances and the new pair-wise distances. For each value of k what is the maximum difference as a percent of $\sqrt{k}$.

B. Data-Driven Documents (D3)

References: Scott Murry, Interactive Data Visualization, O'Reiley; Yihui Xie, Dynamic Documents with R and knitr, CRC press; Deborah Nolan & Duncan Temple Lang, XML and Web Technologies for Data Sciences with R, Springer.

Create a web page (running on local host) with a button which when clicked would display a histogram with the following specifications.

- The data for the histogram should be a random sample of 20 numbers taken without replacement from $\{1, 2, 3, \ldots, 100\}$
- The each block of histogram should be a shade of blue.
- A block of histogram should turn orange when the mouse is placed on it.
- When the mouse is rolled over a subset of adjacent blocks then they should appear in shades of orange.

[1] Johnson-Lindenstrauss Lemma.

Further customization can be done by asking visitors to the web page for the following inputs and creating the histogram accordingly.

- sample size
- ○ with replacement or ○ without replacement
- Normal with $\mu =$ and $\sigma =$
- Uniform with minimum = and maximum =
- Exponential with $\lambda =$

C. Confidence Interval

Reference: *An introduction to bootstrap methods with applications to R (Chapters 3 and 4) by Chernick, Wiley.*

1. Perform the following steps and comment on the observation.

   Step I. Generate one U(-100,100) random number. Call it m.

   Step II. Generate one U(10,50) random number. Call it s.

   Step III. Generate one U(10,25) random number. Call it n.

   Step IV. Generate 1000 N(m,s) random numbers. Call this the population.

   Step V. Sample n numbers without replacement from the population.

   Step VI. Construct 90%, 95%, and 99% confidence intervals for the population mean.

   Step VII. Construct 90%, 95%, and 99% confidence intervals for the population variance.

   Step VIII. Repeat steps V & VI 100/500/1000 times and count the number of times (and percentage) that the population mean is *captured* by the confidence interval.

   Step IX. Repeat steps V & VII 100/500/1000 times and count the number of times (and percentage) that the population variance is *captured* by the confidence interval.

2. In a filament cut test, a razor blade was tested six different times with ultimate forces corresponding to 8.5, 13.9, 7.4, 10.3, 15.7, and 4.0 g.

   (a) Find a 95% confidence interval on the mean using the standard Student's t-distribution.

   (b) Find a 95% confidence interval on the mean using Efron's percentile method.

   (c) Find a 95% confidence interval on the mean using the BCa method and the ABC method. (See Chernick Section 3.5.)

   (d) Find a 95% confidence interval on the mean using the percentile-t method.

   (e) How do the intervals compare? Which intervals do you trust? What does this tell you about the benefits of parametric methods on small ($n < 30$) samples and the problems of using bootstrap on such samples? What does it tell you about the percentile-t method compared with the other bootstrap methods, at least when a formula for the standard error is known?