STATS 200: Homework 1

Due Wednesday, October 5, at 5PM

1. **Accounting for voter turnout.** Let N be the number of people in the state of Iowa. Suppose pN of these people support Hillary Clinton, and $(1-p)N$ of them support Donald Trump, for some $p \in (0,1)$. N is known (say $N = 3{,}000{,}000$) and p is unknown.

(a) Suppose that each person in Iowa randomly and independently decides, on election day, whether or not to vote, with probability 1/2 of voting and probability 1/2 of not voting. Let V_{Hillary} be the number of people who vote for Hillary and V_{Donald} be the number of people who vote for Donald. Show that

$$\mathbb{E}[V_{\text{Hillary}}] = \frac{1}{2}pN, \quad \mathbb{E}[V_{\text{Donald}}] = \frac{1}{2}(1-p)N.$$

What are the standard deviations of V_{Hillary} and V_{Donald}, in terms of p and N? Explain why, when N is large, we expect the fraction of voters who vote for Hillary to be very close to p.

(b) Now suppose there are two types of voters—"passive" and "active". Each passive voter votes on election day with probability 1/4 and doesn't vote with probability 3/4, while each active voter votes with probability 3/4 and doesn't vote with probability 1/4. Suppose that a fraction q_H of the people who support Hillary are passive and $1 - q_H$ are active, and a fraction q_D of the people who support Donald are passive and $1 - q_D$ are active. Show that

$$\mathbb{E}[V_{\text{Hillary}}] = \frac{1}{4}q_H pN + \frac{3}{4}(1-q_H)pN, \quad \mathbb{E}[V_{\text{Donald}}] = \frac{1}{4}q_D(1-p)N + \frac{3}{4}(1-q_D)(1-p)N.$$

What are the standard deviations of V_{Hillary} and V_{Donald}, in terms of p, N, q_H, and q_D? If we estimate p by $\hat{p}$ using a simple random sample of $n = 1000$ people from Iowa, as discussed in Lecture 1, explain why $\hat{p}$ might not be a good estimate of the fraction of voters who will vote for Hillary.

(c) We do not know q_H and q_D. However, suppose that in our simple random sample, we can observe whether each person is passive or active, in addition to asking them whether they support Hillary or Donald.[1] Suggest estimators $\hat{V}_{\text{Hillary}}$ and $\hat{V}_{\text{Donald}}$ for $\mathbb{E}[V_{\text{Hillary}}]$ and $\mathbb{E}[V_{\text{Donald}}]$ using this additional information. Show, for your estimators, that

$$\mathbb{E}[\hat{V}_{\text{Hillary}}] = \frac{1}{4}q_H pN + \frac{3}{4}(1-q_H)pN, \quad \mathbb{E}[\hat{V}_{\text{Donald}}] = \frac{1}{4}q_D(1-p)N + \frac{3}{4}(1-q_D)(1-p)N.$$

[1] In reality, "passive" and "active" might correspond to different demographics of people that are known to have different voter turn-out percentages.

2. **Uncorrelated but not independent (Rice 4.59.)** Let (X, Y) be a random point uniformly distributed on the unit disk $\{(x, y) : x^2 + y^2 \leq 1\}$. Show that $\text{Cov}[X, Y] = 0$, but that X and Y are not independent.

3. **Binomial MGF.** Let $X \sim \text{Binomial}(n, p)$. Find the moment-generating function of X in terms of n and p. (Hint: X is the sum of n independent Bernoulli random variables.)

4. **Existence of multivariate normal.**
(a) Suppose $(X_1, \ldots, X_k) \sim \mathcal{N}(0, \Sigma)$ for a covariance matrix $\Sigma \in \mathbb{R}^{k \times k}$. Let $Y_1, \ldots, Y_m$ be linear combinations of $X_1, \ldots, X_k$, given by

$$Y_j = a_{j1}X_1 + \ldots + a_{jk}X_k$$

for each $j = 1, \ldots, m$ and some constants $a_{j1}, \ldots, a_{jk} \in \mathbb{R}$. Consider the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mk} \end{pmatrix}.$$

By computing the means, variances, and covariances of $Y_1, \ldots, Y_m$, show that

$$(Y_1, \ldots, Y_m) \sim \mathcal{N}(0, A\Sigma A^T).$$

(b) Let $A \in \mathbb{R}^{k \times k}$ be any matrix, let $\Sigma = AA^T$, and let $\mu \in \mathbb{R}^k$ be any vector. Show that there exist random variables $Y_1, \ldots, Y_k$ such that $(Y_1, \ldots, Y_k) \sim \mathcal{N}(\mu, \Sigma)$. (Hint: Let $X_1, \ldots, X_k \overset{IID}{\sim} \mathcal{N}(0, 1)$, and let each Y_j be a certain linear combination of $X_1, \ldots, X_k$ plus a certain constant.)

5. **Bivariate normal conditioning.** Let $X, Y \sim \mathcal{N}(0, 1)$ be independent. Compute $\mathbb{P}[X + Y > 0 \mid X > 0]$. (Hint: Use the rotational symmetry of the bivariate normal PDF.)

6. **Simulating a sample median.** Let $X_1, \ldots, X_{99} \overset{IID}{\sim} \mathcal{N}(0, 1)$. The sample median is the 50th largest value among $X_1, \ldots, X_{99}$. Compute the sample medians from 5000 simulations of $X_1, \ldots, X_{99}$. What is the mean of these 5000 sample medians? What is their standard deviation? Plot a histogram of the 5000 values—what does the sampling distribution of the sample median look like? (Include both your code and your histogram in your homework submission.)

If you are new to programming, the following will walk you through how to do this in R. (If you already know how to do this, then you can disregard the below steps.)

(a) Install R from `https://www.r-project.org/`. Launch R and select "New Document" from the "File" drop-down menu. A blank text window should appear. Select "Save" from the "File" drop-down menu, and save this file as `sample_median.R`.

(b) We will write our code in this document `sample_median.R`. First, let's create a numeric vector of length 5000 that will save the results from the 5000 simulations. We'll call it `X.median`:

```
X.median = numeric(5000)
```

(c) To repeat a series of commands 5000 times, we'll use a `for` loop:

```
for (i in 1:5000) {
  ...
}
```

We can fill in any commands in the "`...`" section above, and these will be executed once for each value of `i` from 1 to 5000.

(d) Let's fill in the `...` section. We can simulate 99 independent samples from $\mathcal{N}(0, 1)$ using the `rnorm` function in R, and save it to a temporary vector variable `X`:

```
X=rnorm(99, mean=0, sd=1)
```

(The `mean` and `sd` arguments can be changed above to generate samples from $\mathcal{N}(\mu, \sigma^2)$ for any μ and σ.) We can then use the `median` function in R to compute the sample median of the values in `X`. We will save this as `X.median[i]`, referring to element `i` of the numeric vector we created in step (b):

```
X.median[i] = median(X)
```

(Hence, in the first loop iteration the sample median is saved as `X.median[1]`, in the second iteration it is saved as `X.median[2]`, etc.) Let's put the above two lines of code in the `...` section from part (c).

(e) After the `for` loop in part (c), we can now write some commands that compute and print the mean and standard deviation of the values in `X.median`, and plot a histogram of these values:

```
print(mean(X.median))
print(sd(X.median))
hist(X.median)
```

(f) Let's save our file `sample_median.R`, then go back to the main R Console, and select "Source File..." under the "File" drop-down menu. Select our file `sample_median.R`, and voila! You should see the the mean and standard deviation of the 5000 sample medians printed in the R Console, and a separate plot window displaying the histogram. That's it!

We'll use more built-in functions in R as we go. To see what a function does and how to use it, type `?` followed by the function name in the R console to pull up the help page. For example, entering

```
?median
```

into the R console pulls up a page about the `median` function used above.