# BMI 510 Midterm 1

## Due Monday, February 27, 2023, 5 PM

**1. `tibble` s and t-tests (7/20 points).**

*A.* There is a dataset of clinical variables and self-reported fall history from people with Parkinson's disease available at [https://jlucasmckay.bmi.emory.edu/global/mckay_2021/S1.csv (https://jlucasmckay.bmi.emory.edu/global/mckay_2021/S1.csv)]. Download the data as a `tibble` and select the columns `Age` , `Sex` , and `num_falls_6_mo` . Exclude any rows with missing data, and exclude the record for `Patient` "bat112", who we discovered had a disqualifying comorbidity after enrollment. **(1 point)**

*B.* The variable `num_falls_6_mo` is the number of self-reported falls for each patient in the six months prior to study enrollment. Use `group_by` and `summarize` to calculate the number of observations for women, men, and for the overall sample, **(1/3 point)** to calculate the average and standard deviation of `Age` for women, men, and for the overall sample, **(1/3 point)**, and to calculate the average and standard deviation of `num_falls_6_mo` for women, men, and for the overall sample. **(1/3 point)**

*C.* Evaluate the null hypothesis that the `Age` values for women and men were generated by the same Gaussian normal process; specifically:

$$H_0 : \mu_F = \mu_M$$

against the alternative hypothesis:

$$H_A : \mu_F \neq \mu_M$$

To do this:[1]

1. Calculate the difference in sample means between the two groups $\bar{X_F} - \bar{X_M}$. **(1 point)**
2. Calculate the pooled variance $S_p^2$. **(1 point)**
3. Calculate the appropriate test statistic $T$. **(1 point)**
4. Find the area under the t-density corresponding to $H_0$. Do these data seem likely under the null? **(1 point)**
5. Compare your result to the results of `t.test` with `var.equal = TRUE` . **(1 point)**

**2. Likelihoods. (7/20 points)**

*A.* Using the same dataset from part 1, fit a Poisson distribution to the `num_falls_6_mo` variable. The Poisson distribution is often described by the formula $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$. For simplicity, we will just refer to it as $P(\lambda)$, similar to how we use $N(0, 1)$ to refer to a standard normal distribution. To do this:

1. Use a grid search process to calculate the log likelihood of the `num_falls_6_mo` data under different candidate values of $\lambda$. Try values for $\lambda$ between 1 and 10; use a step size of 0.01. **(1 point)**
2. Plot the log likelihood as a function of $\lambda$. **(1 point)**
3. Identify the maximum likelihood estimate $\lambda_{MLE}$. **(1 point)** We will refer to this value as $\lambda_{MLE_1}$ later in the problem.

*B.* Plot a histogram of the data. Do any datapoints seem like outliers, in the sense that they are very far from the rest of the datapoints? **(1 point)**

*C.* There are two clear outliers - patients with 50 and 180 self-reported falls, respectively. Remove these patients from the dataset and repeat your grid search process from part *A* with the two most extreme values of `num_falls_6_mo` excluded. Try values for $\lambda$ between 0.1 and 10; use a step size of 0.01. Identify the maximum likelihood estimate of $\lambda$; We will refer to this value as $\lambda_{MLE_2}$. **(1 point)**

*D.* Comparing the likelihoods of a sample of data given different candidate values of a distribution parameter will form the basis of the *Likelihood Ratio Test*, which we will cover later.

1. Calculate the log likelihood of the cleaned-up dataset (with the two outliers removed) under $\lambda_{MLE_1}$ and $\lambda_{MLE_2}$. **(1 point)**
2. Which log likelihood is higher? **(1 point)**

**3. Empirical distributions and ranks. (6/20 points)**

Embattled representative George Santos of New York has been scrutinized for various things, including notable abnormalities in his campaign's bookkeeping. According to the New York Times (https://www.nytimes.com/2023/02/13/nyregion/george-santos-campaign-money.html?unlocked_article_code=O0MFy_OuLFzyD_E8PUolZaFZU76wEvQMOfRig4LzwAOZ_N0MVsyYhMjn_NDqqEVz4KEGHj4jsksUzJp8n0QzZQ5lnyDCrJfrogW7nGKD5U0dyoNg3nBVYdJxEDLLTI9gFRRRjZhGVno_ksf7sqsKtJQPb3UWdy95nEokS6QPGF1wHX4nrMORCMEJxxlJ48Y0sRANDraMtgK13kwQTnvDfnR7ttOpy2rgno0NMpy2_usaauzQMLvipAhfiLQFNQb84lS-G2oRixI8uUo1XY9AJm9v14cvRClqMKUiFyQlcrMh-hX58s4SwSCcGTh4OS9VNScsY&smid=nytcore-ios-share&referringSource=articleShare), representative Santos' campaign is missing receipts for **$365,000**:

> Representative George Santos has spent his campaign money in plenty of conspicuous ways, from lavish hotel stays in Las Vegas and Palm Beach, Fla., to an unusual slew of payments for exactly $199.99 — two cents below the threshold where receipts would be required. But deep within Mr. Santos's campaign filings, The New York Times found another eye-catching number: $365,399.08 in unexplained spending, with no record of where it went or for what purpose. The mysterious expenditures, which list no recipient and offer no receipts, account for nearly 12 percent of the Santos campaign's total reported expenses — many times exceeding what is typical for congressional candidates. Fellow New York House members, for example, failed to itemize between zero and 2 percent of their expenses this past cycle.

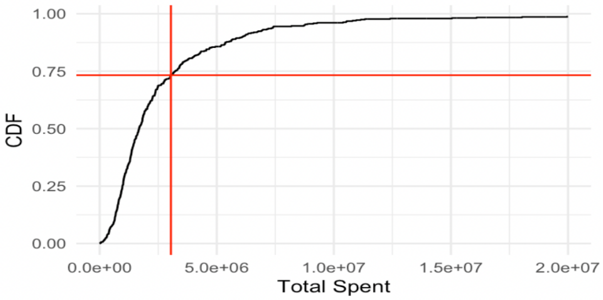
We do not have full numbers for all races in that election cycle, but we do have information on expenditures by incumbents, available here (https://jlucasmckay.bmi.emory.edu/global/bmi510/campaign-spending.csv). For our purposes, you can just examine the columns *Representative* and *Total Spent*, which is the total expenditures in US dollars.

*A.* Assuming that Santos' campaign is missing receipts for $365,000, and that this amount is 12% of Santos' total campaign expenditures, as stated in the article, what were his total campaign expenditures (you can round to the nearest dollar)? **(1 point)**

*B.* Download the dataset (https://jlucasmckay.bmi.emory.edu/global/bmi510/campaign-spending.csv) as a `tibble` called `spending`. Select the first and last columns, and create a new row with `Representative="George Sangos (R-NY)"` and `Total Spent` =your estimate from part A. **(1 point)**

*C.* The r function `rank` will calculate the rank of each value in the vector `Total Spent` and return it as a vector. The representative with the least spending will be rank 1, the representative with the next least will be rank 2, and so on, with the highest spender being rank 434, the total number of representatives. Calculate the rank of each observation of `Total Spent` and save it as a new column in `spending`, `Rank`. How many representatives spent less than Santos? **(1/2 point)** What proportion of the representatives spent less than Santos? **(1/2 point)**

*D.* While `rank` is useful for interpreting a given sample, the r function `ecdf` will create a function representation of the empirical cumulative distribution that you can use on new samples. Instead of returning a vector, `ecdf` returns a *function* that you can use to calculate the relative position (called the quantile) of any arbitrary value within the original data vector. This is very similar to ranking, but accounts for ties between values and other edge cases. Use `ecdf` to plot the empirical cumulative distribution function of `Total Spent`. **(1/2 point)** Highlight Santos' spending on the plot. It should look something like the picture below. **(1/2 point)**



Now that we've seen where Mr. Santos falls with respect to his peers in terms of *overall* spending, let's see what he looks like in terms of *undocumented* spending.

*E.* The article suggests that the most campaigns are missing 0-2% of the documentation for their expenditures. Assuming (charitably) that the rest of the representatives' campaigns were missing exactly 2% of their receipts, use `ecdf` to make a plot similar to part *1D* for the total amount of *undocumented* spending for each campaign. Highlight Santos' spending on the plot. **(1 point)**

*F.* Use the empirical cumulative density functions you have derived to answer the following questions. How many campaigns had more total spending than the Santos campaign? **(1/3 point)** How many campaigns had more undocumented spending than the Santos campaign? **(1/3 point)** Which is more atypical - Santos' total spending or his undocumented spending? **(1/3 point)**

1. See slide 38, one- and two-sample tests lecture for more details.↩