Quiz 6 occurs November 30, in the discussion section. The quiz will be based upon the problems below.

IN CASE THE TA STRIKE OCCURS: then there will be no discussion section, and there will be no quiz (since the TA would be on strike), so you will instead turn in your answers to the problems below on gradescope, as you would if it were a homework.

Quiz 6 Problems

Exercise 1. Using any method you wish to use, find the number of ways that the integer 30 can be written as a sum of 5 distinct positive integers among the elements of $\{1, 2, \ldots, 10\}$. (Hint: you should probably use a computer.)

Exercise 2. Suppose $n = 3, m = 4, n + m = 7$ and we denote Z as the Mann-Whitney statistic. Suppose $Y_1, \ldots, Y_n$ are i.i.d. (treatment outcomes of the n people in the control group) and $X_1, \ldots, X_m$ are i.i.d. (treatment outcomes of the m people in the treatment group). By definition of Z, $1 + 2 + 3 \leq Z \leq 7 + 6 + 5$. The null hypothesis is that the treatment has no effect on people (no "good" effect and no "bad" effect). We should reject the null hypothesis when Z is close to 6 or 18. Consider the (family of) hypothesis tests that rejects the null hypothesis when $|Z - 12| \geq c$ for some constant $c > 0$.

- Compute the p-value for this hypothesis test when $Z = 17$.
- Compute the p-value for this hypothesis test when $Z = 16$.

Now, suppose $n = m = 1000$. Recall that $\mathbf{E}Z = \frac{n(m+n+1)}{2}$. Consider the (family of) hypothesis tests that rejects the null hypothesis when $\left|Z - \frac{n(m+n+1)}{2}\right| \geq c$ for some constant $c > 0$. Using the limiting distribution of Z as an approximation to the distribution of Z itself:

- Approximately compute the p-value for this hypothesis test when $Z = \frac{n(m+n+1)}{2} + 2\sqrt{nm(m+n+1)/12}$.
- Approximately compute the p-value for this hypothesis test when $Z = \frac{n(m+n+1)}{2} + 3\sqrt{nm(m+n+1)/12}$.

Exercise 3. Let $Y_1, \ldots, Y_n$ be i.i.d random variables uniformly distributed in $\{-1, 1\}$. Let

$$W_n := \sum_{i=1}^{n} \max(iY_i, 0).$$

Explicitly write down the distribution of W_6. (Hint: you should probably use a computer.)

Exercise 4. Suppose you are running an experiment to test a new blood pressure drug. The first phase of your trial only involves four people. The following table summarizes the results (for e.g. systolic blood pressure).

Person	Blood pressure before treatment	Blood pressure after treatment
1	120	124
2	140	130
3	130	132
4	110	111

Compute the Wilcoxon ranked sign statistic W_4 for this data.

Now, suppose you run a larger trial, and the data is the following.

Person	Blood pressure before treatment	Blood pressure after treatment
1	120	124
2	140	130
3	130	132
4	110	111
5	122	121
6	143	132
7	131	132
8	140	161
9	100	110
10	100	107
12	100	94
12	140	145
13	130	137
14	110	118
15	100	88
16	135	130
17	136	132
18	113	111
19	129	124
20	145	130

Let $n = 20$. Approximately compute the p-value for this data for the (family of) hypothesis test that reject when $\frac{\left|W_n - \frac{n(n+1)}{4}\right|}{\sqrt{n^3/12}} > c$ for some $c > 0$.

Are you confident in rejecting the null hypothesis?

Exercise 5. Suppose we are presented with data points $(x_1, y_1), \ldots, (x_n, y_n)$. We would like to find the line $y = mx + b$ which lies "closest" to all of these data points. Such a line is known as a **linear regression**. There are many ways to define the "closest" such line. The standard method is to use **least squares minimization**. A line which lies close to all of the data points should make the quantities $(y_i - mx_i - b)$ all very small. We would like to find numbers m, b such that the following quantity is minimized:

$$f(m, b) = \sum_{i=1}^{n} (y_i - mx_i - b)^2.$$

Using the second derivative test (or a convexity argument), show that the minimum value of f is achieved when

$$m = \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} y_j\right) - n\left(\sum_{k=1}^{n} x_k y_k\right)}{\left(\sum_{i=1}^{n} x_i\right)^2 - n\left(\sum_{j=1}^{n} x_j^2\right)} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - m\sum_{j=1}^{n} x_j\right) = \overline{y} - m\overline{x}.$$

Briefly explain why this is actually the minimum value of $f(m, b)$.

Exercise 6. I believe that the number of home runs hit by an MLB baseball player in a single season is linearly related to the number of strikeouts they have in a single season.

Here the data can be found from:

http://seanlahman.com/download-baseball-database/

I recommend using the 2020 Version, comma delimited version. The data is in a zip file, and home run data can be found in `Core` then `batting.csv` then the columns HR and SO.

That is, $(x_1, y_1), (x_2, y_2), \ldots$ is your data, where x_i is the number of home runs hit by the i^{th} person (or i^{th} row) in the data file, and y_i is the number of strikeouts of the i^{th} person (or i^{th} row) in the data file. And we need to find the slope m and intercept b of the line

$$y = mx + b$$

that best fits the data. That is, use the least squares linear regression line.

Plot the data, together with the best fit line. Does the line fit the data well (visually)?

As a test of goodness of fit, compute the quantity

$$\frac{1}{n}\sum_{i=1}^{n} \phi(y_i - [mx_i + b])$$

where $\phi(t) = 1 - e^{-t^2/10}$ for all $t \in \mathbf{R}$, and n is the number of data points. (Note that $\phi(0) = 0$ and $\lim_{t\to\pm\infty} \phi(t) = 1$.) (This quantity is therefore between 0 and 1.) Is this quantity close to 0?

Do all above steps again for the following different statement:

I also believe that the number of stolen bases an MLB baseball player in a single season is linearly related to the number of hits they have in a single season (see the columns SB and H).

Finally, do all above steps again for the following statement:

I also believe that the number of hits of an MLB baseball player in a single season is approximately a constant plus the square of the number of doubles they have in a single season (see the columns 2B and H). That is, $(x_1, y_1), (x_2, y_2), \ldots$ is your data, where y_i is the number of hits by the i^{th} person (or i^{th} row) in the data file, and x_i is the number of doubles of the i^{th} person (or i^{th} row) in the data file. And we need to find the parameters m, b of the parabola

$$y = mx^2 + b$$

that best fits the data. That is, use the least squares linear regression of m, b.

As a test of goodness of fit, compute the quantity

$$\frac{1}{n}\sum_{i=1}^{n}\phi(y_i - [mx_i^2 + b])$$

where $\phi(t) = 1 - e^{-t^2/10}$ for all $t \in \mathbf{R}$, and n is the number of data points. (Note that $\phi(0) = 0$ and $\lim_{t\to\pm\infty}\phi(t) = 1$.) (This quantity is therefore between 0 and 1.) Is this quantity close to 0?

THE PROBLEMS BELOW ARE OPTIONAL. They will not be included on the quiz.

Exercise 7 (Optional)**.** Suppose you have three vegetarian turkeys with numerical "quality" values of

$$Y_i = \beta_1 + \varepsilon_i, \qquad \forall\, 1 \leq i \leq 3$$

Suppose you have three vegetarian turkeys with numerical "quality" values of

$$Y_i = \beta_2 + \varepsilon_i, \qquad \forall\, 4 \leq i \leq 6$$

And suppose you have three vegetarian turkeys with numerical "quality" values of

$$Y_i = \beta_3 + \varepsilon_i, \qquad \forall\, 7 \leq i \leq 9.$$

Here $\varepsilon_1, \ldots, \varepsilon_9$ are i.i.d. Gaussians with mean zero and unknown variance $\sigma^2 > 0$, and $\beta_1, \beta_2, \beta_3 \in \mathbf{R}$ are unknown.

In order to test the null hypothesis that $\beta_1 = \beta_2 = \beta_3$, we perform the F test, i.e. we evaluate

$$F := \sup_{c_1,\ldots,c_3\in\mathbf{R}\colon\ \sum_{i=1}^{3} c_i = 0} \frac{\Big(\sum_{j=1}^{3} c_j\overline{Y}_j - \sum_{j=1}^{3} c_j\beta_j\Big)^2}{S^2\sum_{j=1}^{3}\frac{c_j^2}{3}}$$

Report a p-value for F for the observation that:

$$Y_1 = 5, Y_2 = 6, Y_3 = 7$$
$$Y_4 = 5, Y_5 = 5, Y_6 = 5$$
$$Y_7 = 6, Y_8 = 7, Y_9 = 8.$$

Do you have confidence in accepting the null hypothesis?

Exercise 8 (Optional)**.** Let

$$h(x) := \frac{1}{1 + e^{-x}}, \qquad \forall\, x \in \mathbf{R}.$$

Fix $x \in \mathbf{R}$ and $y \in [0, 1]$. Define $t\colon \mathbf{R}^2 \to \mathbf{R}$ by

$$t(a, b) := \log\Big([h(ax + b)]^y[1 - h(ax + b)]^{1-y}\Big), \qquad \forall\, a, b \in \mathbf{R}.$$

Show that t is concave. Conclude that t has at most one global maximum.

Exercise 9 (Optional)**.** Consider the following table with turkey data. We have 8 (vegetarian) turkeys, with various temperatures x (Fahrenheit), and the status y of each turkey is cooked (corresponding to a value of $y = 1$) or not cooked (corresponding to a value of $y = 0$). Using logistic regression, find $a, b \in \mathbf{R}$, i.e. find a function

$$h(ax + b)$$

that best fits your data, where $h(t) = 1/(1 + e^{-t})$ for all $t \in \mathbf{R}$.

That is, given a temperature x, $h(ax + b)$ should be close to 1 when the turkey is cooked, and $h(ax + b)$ should be close to 0 when the turkey is not cooked.

Turkey	Temperature	Done? Yes or no.
1	140	no
2	145	no
3	150	no
4	155	yes
5	160	no
6	165	yes
7	170	yes
8	175	yes

Exercise 10 (Optional)**.** Suppose $\beta \in \mathbf{R}^m$ is an unknown vector, and A is a known $m \times n$ real matrix. Let $\varepsilon \in \mathbf{R}^n$ be a vector of i.i.d. standard Gaussian random variables. Our observation is $Y := A\beta + \varepsilon$, and the goal is to recover the unknown vector w. In linear least squares regression, we try to determine the best linear relationship w between the rows of A and the observation Y. Assume that $n \leq m$ and the matrix A has full rank (so that $A^T A$ is invertible). Show that the vector $x \in \mathbf{R}^m$ that minimizes the quantity

$$||Y - Ax||^2 := \sum_{i=1}^{n} (y_i - (Ax)_i)^2$$

is

$$x := (A^T A)^{-1} A^T Y. \qquad (*)$$

Equivalently, show that x minimizes

$$\mathbf{E}\,||Y - x||^2$$

over all choices of vectors $x \in \mathbf{R}^m$ such that $x = By$ for some $n \times m$ real matrix B, and such that $\mathbf{E}x = w$. (Since ε is the only random variable here, $\mathbf{E}$ denotes expected value with respect to ε.)