Take-home final

Name:

Instructions: For questions that require R code, upload your code in an `.R` file (such as `problem1.R`), along with a pdf / jpeg file containing your written answer + screenshot of the output from your program. You are able to upload multiple files on gradescope for each question (by holding down the Ctrl key when selecting). Please only upload the right files for each problem. For questions that require running a simulation n times, see the `sapply` function from `r_example_4.R`.

Problem 1. (10 pts) Least squares hypothesis testing.

Assume we have a random sample of points $(x_1, Y_1), \ldots, (x_{20}, Y_{20})$ where

- $x_i = i$ for all $i = 1, \ldots, 20$
- $Y_i = \alpha_1 + \alpha_2 x_i + \alpha_3 x^2 + \alpha_4 \ln x_i + \epsilon_i$ are random variables where $\epsilon_i \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ and the ϵ_i are i.i.d.
- $\alpha_1 = 2, \alpha_2 = 1, \alpha_3 = 1, \alpha_4 = 2, \sigma_\epsilon^2 = 4$ are constants.

Let $(\hat{\alpha}_i)_{i=1}^4$ be the least square estimators for $(\alpha_i)_{i=1}^4$.

Plot the approximated pdf of $\sum_{i=1}^4 \widehat{\alpha}_i$ in R. Approximate its mean and variance.

Instruction: Use R to generate the vector $(Y_i)_{i=1}^{20}$ and then derive $\sum_{i=1}^4 \widehat{\alpha}_i$. Do this at least $(10)^5$ times.

Bonus (10 pts): use actual maths to give the exact formulas for the mean and the variance of $\sum_{i=1}^4 \widehat{\alpha}_i$, and then evaluate them.

Problem 2. (20 pts) Assume we have a random sample of points $(x_1, Y_1), \ldots, (x_n, Y_n)$ where the $x_i = i$ and $n > 10$. We assume that there is a general relationship

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

where the β_i are some real constants and ϵ_i are i.i.d. standard normal (before observation). Let $\hat{\beta}_i$ be the least-squares estimators for β_i, and $\hat{Y}_i := \widehat{\beta_0} + \widehat{\beta_1} x_i + \widehat{\beta_2} x_i^2$ be the theoretical regression model. Then we can still define SSR as

$$\text{SSR} = \sum_i \left(Y_i - \hat{Y}_i \right)^2$$

Use actual maths to find the exact pdf for SSR.

Hint: show that $\text{SSR} = \left\| M\vec{Y} \right\|^2$ where M is a projection matrix ($M^2 = M$) and so is $I - M$. Do not blindly assume linear regression formulas hold for general regression (but linear algebra still works).

Problem 3. (20 pts) Assume we have a random sample of points $(x_1, Y_1), \ldots, (x_{20}, Y_{20})$ where

- $x_i = i$ for all $i = 1, \ldots, 20$.
- $Y_i = \alpha_1 + \alpha_2 x_i + \epsilon_i$ are random variables where $\epsilon_i \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ and the ϵ_i are i.i.d.
- $\alpha_1 = 2, \alpha_2 = 1, \sigma_\epsilon^2 = 4$ are constants.

1. Plot the approximated pdf of $\hat{\alpha}_1^2 + \hat{\alpha}_2^2$ in R. Approximate its mean and variance.
2. Let r be the sampling correlation coefficient (review lectures). Plot the approximated pdf of r in R. Approximate its mean and variance. Approximate $\mathbb{P}(0.9 < r < 0.95)$.

Instruction: Use R to generate the vector $(Y_i)_{i=1}^{20}$. Do this at least $(10)^5$ times.

Remark. These are actually very complicated random distributions. The Monte Carlo method here luckily avoids all that (the cost being a minor amount of inaccuracy, and inability to handle symbolic values).

Problem 4. (10 pts) Assume we have a random sample of points $(x_1, Y_1), \ldots, (x_{100}, Y_{100})$ where

- Each x_i is a fixed constant.
- $Y_i = \alpha_1 + \alpha_2 x_i + \cdots + \alpha_{d+1} x_i^d + \epsilon_i$ are random variables where $\epsilon_i \sim \mathcal{N}(0, 100)$ and the ϵ_i are i.i.d.
- d is unknown. Also the coefficients $\alpha_1, \ldots, \alpha_{d+1}$ are unknown.

We run the experiment and observe the values of $(Y_i)_{i=1}^{100}$. See the `final_problem4_data.R` file for the values of $(x_i)_{i=1}^{100}$ and an observed value of $(Y_i)_{i=1}^{100}$.

1. Figure out the value of d.
   Hint: Keep raising d until you start getting near-zero estimates for the leading coefficient.

   Remark. You should notice that the `solve(...)` function in R quickly starts to malfunction (`system is computationally singular`) as numbers get smaller. Fortunately you will find the answer before hitting that limit.

2. With that value of d, test against the null hypothesis $H_0 : \alpha_1 = 0$, and calculate the 2-sided p-value. Give the 95% confidence interval for α_1.

Problem 5. (10 pts) How are the eigenvalues of a symmetric random normal matrix distributed?

Let M be a symmetric 120×120 matrix, where each entry $M_{i,j}$ (where $i \leq j$) in the upper triangle is an independent standard normal random variable.

Recall that all eigenvalues of a symmetric matrix are real (by the spectral theorem).

Let X be a number randomly picked from all the eigenvalues of M (all choices equally likely).

1. Plot the approximated pdf of X in R.
2. Approximate its mean and variance. Guess what the actual variance should be. Confirm this guess by changing the dimension of the matrix M.

Instruction: it is enough to run the simulation $(10)^3$ times and put all the eigenvalues into a vector. Look up the `eigen(...)$values` function in R.

Remark. This is **Wigner's semicircle law**, one of the most celebrated results in statistical / quantum / nuclear physics. It is analogous to the central limit theorem, but in the setting of non-commutative probability theory.

Understanding why the "guess" (in the second half of the problem) holds true is a big rabbit hole for mathematics and physics. See Tao's notes on random marix theory.

Problem 6. (10 pts) Assume that on a certain road, after each car crash, the next car crash is T months away where T is an exponential random variable with rate parameter $\frac{1}{3}$. Approximate the expected number of car crashes in 10 months.

Instruction: generate in R the times for the first 1000 car crashes (you might need to look up the `cumsum(...)` function), then count how many happen before 10 months. Do this at least $(10)^5$ times and take the average.

Bonus (10 pts): use maths to give the exact value for the expected number of car crashes.

Hint: Let $(X_n)_{n\geq 1}$ be an i.i.d. sequence with $X_i \sim \text{Exp}(\lambda)$ for each i (here $\lambda = \frac{1}{3}$). Let $S_n := X_1 + \cdots + X_n$ be the time of the n-th crash. Then $N_{10} := \max\{n : S_n < 10\}$ is the number of crashes in 10 months. Use a 2D integral to calculate $\mathbb{P}(N_{10} = k)$ for each $k \in \mathbb{N}$, and then get $\mathbb{E}(N_{10})$.

Problem 7. (10 pts) Throw two darts randomly into the 2D square $[-1,1] \times [-1,1]$ (uniformly random over the square). What is the probability that the distance between the two darts is less than 1? Simulate in R over $(10)^6$ tries for an approximation of the answer.

Bonus (10 pts): find the exact answer with maths.

Hint: first figure out the pdf of $X_1 - X_2$ where each $X_i \overset{\text{i.i.d.}}{\sim} U([-1,1])$ (uniform over the interval).

Problem 8. (10 pts) Let $X_1 = 0, X_2 = 1$ and $X_k \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}\left(\frac{1}{\ln(k)}\right)$ for any natural number $k \geq 3$. Then $\pi(n) := X_1 + \cdots + X_n$ is the probabilistic prime counting function (how many primes up to n).

Let $\text{Li}(n) = \int_2^n \frac{dx}{\ln(x)}$ be the (Eulerian) logarithmic integral.

The weak form of the probabilistic Riemann hypothesis is that for any $\delta > 0, \epsilon > 0$:

$$\mathbb{P}\left(\left|\frac{\pi(n) - \text{Li}(n)}{n^{\frac{1}{2}+\delta}}\right| > \epsilon\right) \xrightarrow{n\to\infty} 0$$

Confirm this fact experimentally in R with $\epsilon = 0.2, \delta = 0.2, n = (10)^4$, and over $(10)^4$ tries for sampling $\pi(n)$.

Bonus (30 pts)**:** show it with maths.

Hint: where else in the course did you see a similar-looking limit? What did we use then?

Remark. The Bernoulli parameter $\frac{1}{\ln(k)}$ is from the prime number theorem but you do not need to (or should) use that theorem here. In fact no special knowledge about number theory or complex analysis is even in the solution. This is way easier than the actual Riemann hypothesis.