Problem Set 1

Issued: Friday, Sept 10

Part 1 Due: Monday, Sept 20, 11:59PM ET **Part 2 Due:** Monday, Sept 27, 11:59PM ET

Part 1

Problem 1.1: [10 pts] The Salk Vaccine Field Trial

The first polio epidemic hit the United States in 1916. By the 1950s several vaccines against the disease had been discovered. The one developed by Jonas Salk seemed the most promising in laboratory trials. By 1954, the National Foundation for Infantile Paralysis (NFIP) was ready to try the vaccine in the real world. They ran a controlled experiment to analyze the effectiveness of the vaccine. The data is shown in the table below (grade refers to the educational stage). The experiment was later repeated as a randomized controlled double-blind experiment. This data is shown in the second table below.

NFIP study		
	Size	Polio rate per 100,000
Grade 2 (vaccine)	225,000	25
Grades 1 and 3 (no vaccine)	725,000	54
Grade 2 (no consent)	125,000	44

Randomized controlled double-blind experiment		
	Size	Polio rate per 100,000
Treatment (vaccine)	200,000	28
Control (salt injection)	200,000	71
No consent	350,000	46

(a) Describe each of the two studies (e.g., their design) and comment on the differences between them. For each study, explain whether it helps measure what was intended to be estimated.

(b) Which numbers show the effectiveness of the vaccine? Explain why.

(c) In the two studies neither the control groups nor the no-consent groups got the vaccine. Yet the no-consent groups had a lower rate of polio. What could be some of the underlying reasons?

(d) Polio is an infectious disease. The NFIP study was not done blind. Could this bias the results? If yes, to what extent?

(e) In the randomized controlled trial the children whose parents refused to participate in the trial got polio at the rate of 46 per 100,000. On the other hand, the children whose parents consented to participate got polio at a slighter higher rate of 49 per 100,000 (treatment group and control group taken together). On the basis of these numbers, in the following year some parents refused to allow their children to participate in the experiment and be exposed to this higher risk of polio. Were they right? And why so? Please explain your reasoning process.

Problem 1.2: [25 pts] NASA Compton Gamma Ray Observatory Data (source: Rice, Ch.8)

The file `gamma-ray` contains a small quantity of data collected from the Compton Gamma Ray Observatory, a satellite launched by NASA in 1991 (http://cossc.gsfc.nasa.gov/). For each of 100 sequential time intervals of variable lengths (given in seconds), the number of gamma rays originating in a particular area of the sky was recorded. You would like to check the assumption that the emission rate is constant.

a) What is a good model for such data?

b) Describe the null hypothesis H_0 and the alternative H_A.

c) What is(are) the most plausible parameter value(s) for the null model given the observations? Calculate the maximum likelihood estimate(s) (MLE) of the parameter(s). Compute the estimator(s) for these parameter(s) from the data and report the resulting value(s).

d) What is(are) the most plausible parameter value(s) for the alternative model given the observations? Calculate the MLE(s). Compute the estimator(s) for the parameter(s) from the data (you do not need to provide the value(s)). *Hint: You should carefully define the space for the parameter(s) of your model.*

e) Define a test statistic and plot its distribution under H_0 using the software of your choice.

f) Determine the rejection region at a significance level of 0.05. Depict it in the previous plot.

g) Also show the value of the test statistic in the previous plot. What is its p-value? Based on the data collected by the observatory and the analysis that you have conducted, does the emission rate appear to be constant?

Problem 1.3: [10 pts] P-values

Read the statement by the American Statistical Association about p-values (Wasserstein and Lazar: The ASA's statement on p-values: context, process, and purpose) and respond to the following scenarios.

(a) A friend looking at your notes from the first lecture saw that there's a p-value of 0.0012 for the HIP study. They ask you, does that mean there's a 99.88% chance that offering a mammography decreases the risk of death from breast cancer? Explain to your friend exactly what this p-value means, including any assumptions that were made.

(b) Your colleague in education studies cares about what can improve the education outcome in early childhood. He thinks the ideal planning should be to include as many variables as possible and regress children's educational outcome on the set. Then we select the variables that are shown to be statistically significant and inform the policy makers. Is this approach likely to produce the intended good policies? What other approach to this problem could you suggest?

(c) An economist collects data on many nationwide variables and surprisingly finds that if she runs a regression between chocolate consumption and number of Nobel prize laureates, the coefficient is statistically significant. Should she conclude that there exists a relationship between Nobel prize and chocolate consumption? Explain why.

(d) Your lab collects individual-level data on 50,000 humans for 100 features, including IQ and chocolate consumption. They find that their initial hypothesis about the relation between chocolate consumption and IQ has a p-value higher than 0.05. However, they find that there are other variables in the data set that have p-value less than 0.05, namely, a subject's family income and number of siblings. They therefore decide to not write about chocolate consumption, but rather, report these statistically significant results in their paper, and provide possible explanations. Is this sound scientific practice? Please discuss.

(e) A neuroscience lab lab runs a randomized experiment on 100 mice by adding chocolate in half of the mice's diet and another food of the equivalent calories in another half's diet. They find that the difference between the two groups' time in solving a maze puzzle has p-value lower then 0.05. Should they conclude that chocolate consumption leads to improved cognitive power in mice? Explain why.

(f) Should p-values be banned from scientific papers? Provide at least one argument for and one argument against this proposal.

Problem 1.4: [10 pts] Published research findings are false (*optional for undergraduates*)

Read the paper by Ioannidis on why most published research findings are false (PLoS Medicine, 2005) and summarize the paper in your own words. What is the most important lesson you learned from reading this paper? Explain the computations going into Table 1, Table 2, and Table 3. How does Ioannidis get to the conclusion that a research finding is more likely true than false if $(1 - \beta)R > \alpha$ (at the beginning of page 697)? What does this mean?

Part 2

Problem 1.5: [15 pts] Detecting Leukemia types

The data set `golub` consists of the expression levels of 3,051 genes for 38 tumor mRNA samples. Each tumor mRNA sample comes from one patient (i.e., 38 patients total): 27 of these tumor samples correspond to acute lymphoblastic leukemia (ALL), while the remaining 11 correspond to acute myeloid leukemia (AML). How many genes are associated with the different tumor types (meaning that their expression level differs between the two tumor types) using (i) the uncorrected p-values, (ii) the Holm-Bonferroni corrected p-values, and (iii) the Benjamini-Hochberg corrected p-values? Feel free to use existing libraries for multiple hypothesis testing, in `R` or `python`. You can use $\alpha = 0.05$ for the significance threshold.

Source of data: Golub et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, Vol. 286:531-537.

Problem 1.6: [20 pts] Regression and Gradient Descent

In this problem, we will look at OLS and the gradient descent algorithm.

a) Read in the synthetic data matrix `syn_X.csv` and the vector `syn_y.csv` of "observations". Compute the OLS estimator $\hat{\beta}$ by matrix inversion.

b) Implement gradient descent for the least squares problem and run it on the synthetic test data loaded in the previous question. As a function of the iteration t, plot the mean squared error (MSE) and the distance $\|\beta^t - \hat{\beta}\|$ of your current iterates (in separate plots). Play with different initializations β^0 and different step sizes. What do you observe? Explain. Based on your observations, what would be an optimal step size?

Next, we look at some real data. General Motors collected data (found in `mortality.csv`) from 60 US cities to study the contribution of air pollution to mortality. The dependent variable is the age-adjusted mortality (`Mortality`). The data include variables measuring climate characteristics (`JanTemp, JulyTemp, RelHum, Rain`), variables measuring demographic characteristics of the cities (`Educ, Dens, NonWhite, WhiteCollar, Pop, House, Income`), and variables recording the pollution potential of three different air pollutants (`HC, NOx, SO2`).

c) Get an overview of the data and account for possible problems. Which cities stand out? Which of the variables need to be transformed? When applicable, which transformations would you apply?

d) Run your gradient descent algorithm for least squares on the raw data and on the transformed data, with different step sizes as before. What do you observe?

e) Carry out a multiple linear regression containing all variables with the necessary transformations (with gradient descent as in d) or with matrix inversion). Does the model fit well? Check the residuals and comment on what you observe.

f) *Gradient descent for other functions:* A popular regression model for binary observations y is given by the following estimator:

$$\hat{\beta} = \arg\min_{\beta} \sum_i \log(1 + \exp(-y_i \beta^T x_i))$$

How would you solve this via gradient descent? Derive the corresponding gradient and write down the steps of the algorithm.

Problem 1.7: [10 pts] Computational Aspects of Regression

In this problem, we will consider some computational challenges that arise in practice when performing linear regression.

a) Suppose you have a problem in which the feature matrix, X, has 100 million rows and 200 columns. What challenge will arise when you try to apply either the matrix inversion method or the gradient descent method to compute the regression coefficients as in the previous problem? *Hint: if each entry is a 64-bit float, how much memory will be required to store X?*

b) Suggest one method that will allow you to compute the linear regression coefficients for this problem. Be specific. Discuss pros and cons of your proposed approach.

Now, suppose we are in a setting in which the number of data points, n, is much smaller than the number of variables, p, i.e., X has many more columns than rows. This situation occurs often in biological applications, for example, in which the features may represent the expression levels of various genes. This is often referred to as the "high-dimensional" regime. (Assume X is small enough that it can fit in memory.)

c) Can we run gradient descent to compute the regression coefficients? What do you think about the solution? Why? *Hint: what is the maximum rank of the matrix $X^T X$?*

d) *(optional for all)* Load the data from the previous question, `syn_X.csv` and `syn_Y.csv`. Compute the regression coefficients by solving the LASSO problem for various values of λ. What happens to the solution as λ increases? Choose λ such that only one component of the coefficient vector is nonzero. What is the value of λ? Which coefficient is it? (For this problem, feel free to use a package that performs LASSO regularization, such as `scikit-learn` in `python` or `glmnet` in `R`.)

Problem 1.8: Likelihood ratio test for a Gaussian model (*optional for all*)

In this problem, we will analyze the likelihood ratio statistic for the model $X \sim \mathcal{N}(\mu, \sigma^2)$ with unknown mean and unknown variance and null hypothesis $H_0 : \mu = 0$ versus alternative hypothesis $H_A : \mu \neq 0$.

(a) What is the likelihood function for n iid (independent and identically distributed) Gaussian random variables (with mean μ and variance σ^2)?

(b) What is the likelihood ratio statistic for the hypothesis test specified above? (You should simplify the statistic so it only involves the realizations $x_1, \ldots, x_n$.)

(c) What is the rejection region for a one-sample (two-sided) t-test for the same hypothesis test?

(d) How is the likelihood ratio test related to the one-sample t-test? Show that the exact rejection region of the likelihood ratio test (without approximation by the χ^2-distribution) has the same form as the rejection region of the t-test.

(e) Analyze either by simulation or by computation how large the error is if you use the asymptotic distribution of the likelihood ratio statistic versus the exact distribution as in (d).