Problem Set 1

Issued: Monday, September 10 **Due:** Sunday, September 23, 11:59 PM ET

Problem 1.1: 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 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)	225000	25
Grades 1 and 3 (no vaccine)	725000	54
Grade 2 (no consent)	125000	44

Randomized controlled double-blind experiment		
	Size	Polio rate per 100'000
Treatment (vaccine)	200000	28
Control (salt injection)	200000	71
No consent	350000	46

(a) Compare the two studies and comment on the differences.

(b) Which numbers show the effectiveness of the vaccine?

(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. Why?

(d) Polio is an infectious disease. The NFIP study was not done blind. Could this bias the results?

(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?

Problem 1.2: 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? Compute the MLE(s). Calculate the estimator(s) from the data.

d) What is(are) the most plausible parameter value(s) for the alternative model given the observations? Compute the MLE(s). Calculate the estimator(s) from the data. (You do not need to print the value(s).)

e) Define a test statistic and plot its distribution under H_0.

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? Does the emission rate appear to be constant?

Problem 1.3: 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). Should p-values be banned from scientific papers? Argue for and against this proposal.

Problem 1.4: Detecting Leukemia types

The data set `golub` consists of the expression levels of 3051 genes for 38 tumor mRNA samples. Each tumor mRNA sample comes from one patient (i.e. 38 patients total), and 27 of these tumor samples correspond to acute lymphoblastic leukemia (ALL) and the remaining 11 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 correction, and (iii) the Benjamini-Hochberg correction? Feel free to use libraries for multiple hypothesis testing in `R` or `python`.

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 1.5: Why most 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?

Problem 1.6: 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 our synthetic test data. As a function of the iteration t, plot the mean squared error and the distance $\|\beta^t - \hat{\beta}\|$ of your current iterates (in separate plots). Play with different initializations β^0 and step sizes. What do you observe? Explain. 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 includes 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?

d) Run your gradient descent algorithm for least squares on the raw and 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 matrix inversion). Does the model fit well? Check the residuals.

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 gradient and write down the steps of the algorithm.

Problem 1.7: 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 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 $X^T X$?)

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. Which coefficient is it? (Feel free to use a lasso package such as `scikit learn` or `glmnet` for this problem.)

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 variance and null hypothesis $H_0 : \mu = 0$ versus $H_A : \mu \neq 0$.

(a) What is the likelihood function for n iid (independent and identically distributed) Gaussian random variables (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 $x_1, \ldots, x_n$.)

(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 (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 in (d).