Homework 1

Requirement

- To be submitted individually. You are allowed to discuss with your friend, but should finish the tasks on your own. Acknowledge your friend if he/she has provided substantial help to you.
- You can finish the homework in either Chinese or English.
- For programming exercises, you can print out the code, results and figures.
- Hand in to the instructor when you attend the lecture, on or before 9 Mar, 2023 (Thursday). Late submissions will be penalized by a 20% deduction in score.

Exercises

Exercise 1:

Prove the bias-variance decomposition relation for the MSE of estimate $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$:

$$\mathrm{MSE}(\hat{\theta}) := \mathbb{E}\left[(\hat{\theta} - \theta^*)^2\right] = \left[\mathbb{E}(\hat{\theta}) - \theta^*\right]^2 + \mathrm{Var}(\hat{\theta}). \tag{1}$$

Exercise 2:

Dirac Delta Function and Empirical Measure

The Dirac delta function, introduced by the famous physicist P. A. M. Dirac, is defined as following

$$\delta_{x_i}(x) = \delta(x - x_i) = \begin{cases} +\infty, & \text{if } x = x_i \\ 0, & \text{if } x \neq x_i \end{cases} \tag{2}$$

$$\text{satisfying } \int_{-\infty}^{\infty} \delta_{x_i}(x)\mathrm{d}x = 1. \tag{3}$$

It can be viewed as the density of a uniform distribution on $[x_i - \frac{\varepsilon}{2}, x_i + \frac{\varepsilon}{2}]$ under the limit $\varepsilon \to 0$ (See Fig. 1). Strictly speaking, $\delta_{x_i}(x)$ is not a function in the classical sense, but a generalized function in the distribution theory, or can be formalized as a measure, denoted as δ_{x_i}.

In physics, it is convenient to use the Dirac delta to represent the density of a point mass (点质量), where the particle have a finite mass but zero volume (thus infinite density).

In statistics, one can use the Dirac delta to define the empirical measure for observed data $\{x_i\}_{i=1}^n$ (assumed $x_i \in \mathbb{R}$) as $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. This notion is useful in theoretical analysis of estimations, non-parametric methods for statistics, optimal transport, etc.

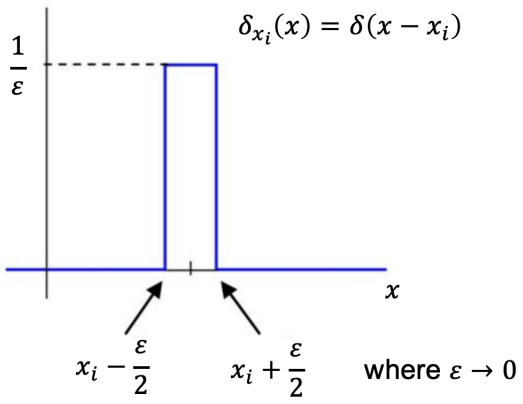

Figure 1: Dirac delta function.

From P_n, we (loosely) define a density of the empirical observations

$$P_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x). \tag{4}$$

Problems:

1. The Heaviside step function is defined as

$$H(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases} \tag{5}$$

   Show that the Heaviside is the integration of the Dirac delta, i.e., $H(x - x_i) = \int_{-\infty}^{x} \delta_{x_i}(x')\mathrm{d}x'$.

2. The empirical cumulative distribution function (CDF) of the data $\{x_i\}$ is

$$F_n(x) = \frac{1}{n}\{\text{no. of } x_i \leq x\}. \tag{6}$$

   Show that it can be expressed as $F_n(x) = \int_{-\infty}^{x} P_n(x')\mathrm{d}x'$.

   You can also find a more explicit expression of the empirical CDF in the extra slide `summarizing_data.pptx` in the homework1 folder.

3. Express the histogram of the data $\{x_i\}_{i=1}^{n}$ with bin edges $(b_1, b_2, ..., b_m)$ using the density $P_n(x)$.

4. In the lecture, we've seen that the negative log-likelihood of the data under the model $P_\theta(X)$ has the form of $-\frac{1}{n}\sum_{i=1}^{n} \log P_\theta(x_i)$. Show that it can be expressed as the the cross entropy between $P_n(x)$ and $P_\theta(x)$.

Exercise 3:
A Random Walk Model for Chromatin (染色质)

(Adapted from Problem 8.10.45, J. A. Rice, MSDA3; see the text therein for a more detailed description of the background).

A human chromosome (染色体) is a very large molecule, about 2 or 3 centimeters long, containing 100 million base pairs (Mbp). The cell nucleus, where the chromosome is contained, is in contrast only about $\frac{1}{1000}$ of a centimeter in diameter. The spatial organization such a large molecule is not well understood.

A series of experiments (Sachs et al., 1995; Yokota et al., 1995) were conducted to learn more about them. Pairs of small DNA sequences (size about 40 kbp) at specified locations on human chromosome 4 were flourescently labeled in a large number of cells. The distances between the members of these pairs were then determined by flourescence microscopy (荧光显微成像). (The distances measured were actually two-dimensional distances between the projections of the paired locations onto a plane.) **The empirical distribution of these distances provides information about the nature of large-scale organization.**

There has long been a tradition in chemistry of modeling the configurations of polymers (聚合物) by the theory of random walks. As a consequence of such a model, the two-dimensional distance should follow a Rayleigh distribution:

$$f(r \mid \theta) = \frac{r}{\theta^2} \exp\left(-\frac{r^2}{2\theta^2}\right).$$

Basically, the reason for this is as follows: The random walk model implies that the joint distribution of the locations of the pair in $\mathbb{R}^3$ is multivariate Gaussian; by properties of the multivariate Gaussian, it can be shown the joint distribution of the locations of the projections onto a plane is bivariate Gaussian. It can be shown that the distance between the points follows a Rayleigh distribution (Section 3.6.2, J. A. Rice, MSDA3).

In this exercise, you will fit the Rayleigh distribution to some of the experimental results. The entire data set comprises 36 experiments in which the separation between the pairs of flourescently tagged locations ranged from 10 Mbp to 192 Mbp. In each such experimental condition, about 100–200 measurements of two-dimensional distances were determined. This exercise will be concerned just with the data from three experiments (short, medium, and long separation). The measurements from these experiments is contained in the files `Chromatin/short`, `Chromatin/medium`, `Chromatin/long`.

Problems:

1. Derive the maximum likelihood estimate (MLE) of θ for data from a Rayleigh distribution.

2. (Optional) Derive the approximate variances of the MLE of θ, in the large sample size limit.

3. Use a computer software to fit the data of the three experiments.

   Do some visualization to compare your fitted model to the data.

   Comment on how good is the fitting.

   In `R`, you can use `fitdistrplus` for this purpose, with only a few commands. The Rayleigh distribution is not supported in `R base`; you can load it from the `extraDistr` package, in which the scale parameter θ is called `sigma`. You may need to supply a starting value for optimization by adding `start=list(sigma = xx)` when fitting.

4. For one of the experiments, use the bootstrap method to construct an approximate 95% confidence interval for θ using $B = 1000$ bootstrap samples.

   (Optional) Compare this interval to that obtained using large sample theory.

Exercise 4:

In the homework1 folder, you will find an extra slide `summarizing_data.pptx` and a data set called `Birthweight_reduced_kg_R.csv`, which contains information on some new born babies and their parents.

Go through the slide `summarizing_data.pptx`, and apply some of the techniques to explore the the birth weight data, in particular, familiarize yourself with box plot, Q-Q plot and scatter plot in `R`. You are also free to do some additional analyses.

A short description of the features of the birth weight data can be found in the following table:

Name	**Variable**	**Data type**
ID	Baby number	
length	Length of baby (cm)	Continuous
Birthweight	Weight of baby (kg)	Continuous
headcirumference	Head Circumference	Continuous
Gestation	Gestation (weeks)	Continuous
smoker	Mother smokes 1 = smoker 0 = non-smoker	Binary
motherage	Maternal age	Continuous
mnocig	Number of cigarettes smoked per day by mother	Count
mheight	Mothers height (cm)	Continuous
mppwt	Mothers pre-pregnancy weight (kg)	Continuous
fage	Father's age	Continuous
fedyrs	Father's years in education	Continuous
fnocig	Number of cigarettes smoked per day by father	Count
fheight	Father's height (cm)	Continuous
lowbwt	Low birth weight, 0 = No and 1 = yes	Binary
mage35	Mother over 35, 0 = No and 1 = yes	Binary

Reading

The following reading assignment is only recommended but not compulsory in this homework.

Exercise 5:

People often use confidence intervals to support the precision of their finding. However, a recent study showed that solely relying on the claimed confidence intervals very often will make people overconfident on the results.

Read the following news article and the study quoted therein and think about the implication on your practice in data analysis. You should be aware that statistical error constitute only one source of error from data (which is almost the only type of error studied in most statistics courses!).

`https://newsroom.haas.berkeley.edu/research/election-polls-are-95-confident-but-only-60-accurate-berkeley-haas-study-finds/`