DSA 8301 - Statistical Inference for Big Data
Spring 2025
Exercise # 1 April 1, 2025

1) Let $X_1, ..., X_n$ be *i.i.d* Poisson(λ).

a) Show that the maximum likelihood estimator of λ is $\overline{X}$.

b) Let X equal the number of flaws per 100 feet of a used computer tape. Forty observations of X yielded flaws as summarized in the following table:

Number of Flaws per Item	0	1	2	3	4	5	6
Frequency	5	7	12	9	5	1	1

Assuming that the flaw counts have the Poisson(λ) distribution, compute the maximum likelihood estimate of λ.

c) Give the model-based estimate of the population variance, and compare it with the sample variance.

d) Assuming the Poisson model correctly describes the population distribution, which of the two estimates would you prefer and why?

2) The probability density function of the Rayleigh distribution is

$$f(x) = \frac{x}{\theta^2} e^{-\frac{x^2}{2\theta^2}},$$

where θ is a positive-valued parameter. It can be shown that the mean and variance of the Rayleigh distribution are

$$\mu = \theta\sqrt{\pi/2}$$

and

$$\sigma^2 = \theta^2(4-\pi)/2.$$

Let $X_1, .., X_n$ be a random sample from a Rayleigh distribution.

a) Construct the method of moments estimator of θ. Is it unbiased?

b) Construct a model-based estimator of the population variance. Is it unbiased?

3) A company manufacturing bike helmets wants to estimate the proportion p of helmets with a certain type of flaw. They decide to keep inspecting helmets until they find $r = 5$ flawed ones. Let X denote the number of helmets that were not flawed among those examined.

a) Write the log-likelihood function and find the MLE of p.

b) Find the method of moments estimator of p.

c) If $X = 47$, give a numerical value to your estimators in (a) and (b) above.

4) Researchers studying healthy body composition recorded various measurements of 252 male subjects. Height, weight, age, and other measurements were collected on each subject. Open the data file *BodyMeasurementsCorrected* (you can used the code below to read the file into *R*). For this exercise we will look at the height of the subjects compared with the weight. Researchers want to predict the weight of a man given his height. Use this information to answer questions (a) through (e) below:

```
body = read.csv("https://github.com/byuistats/data/raw/master/
BodyMeasurementsCorrected/BodyMeasurementsCorrected.csv", header = TRUE,
stringsAsFactors = FALSE)
```

a) If we consider the relationship between the height and weight of these men, which variable should go on the X-axis? The Y-axis? Justify your answer.

b) Create and attach a scatterplot of these two variables. Include the linear regression line on your plot.

c) Find the equation of the linear regression line used to predict the weight of a man given his height.

d) Interpret the slope and intercept of the regression line, if appropriate. If it is not appropriate to make an interpretation, explain why not.

e) Predict the weight of a man who is 76 inches tall in two ways: using the equation for the regression line AND using *R*.