Applied Statistic with R

Fall 2019, ASDS, YSU

Homework No. 04

Due time/date: 3:30 PM, 26 October, 2019

Note: Please use **R** only in the case the statement of the problem contains (R) at the beginning. Otherwise, show your calculations on the paper. Supplementary Problems will not be graded, but you are very advised to solve them and to discuss later with TA or Instructor.

Problem 1, Probability Refresher, RVs

a.

Let $X \sim Pois(2)$ and $Y \sim Exp(3)$. Calculate

1. $\mathbb{P}(X \geq 2)$;
2. $\mathbb{E}(X^2)$;

Hint: You can use the Variance!

3. $\mathbb{P}(Y < 3)$
4. Assuming X and Y are independent, calculate $\mathbb{E}(XY)$.

b. (R)

Assume I made an Ad on FB and I want to model the number of clicks during a day on my Ad. I have calculated that the average number of clicks in a day is 34.3.

1. Suggest a model for the number of clicks
2. Calculate the probability that I will have more than 40 clicks tomorrow.
3. Generate a possible scenario for the number of clicks for each day of the next week.

Problem 2. Convergence of r.v.s

a.

Assume X_n, $n \geq 3$, is a r.v. with the following PMF:

Values of X_n	$-\dfrac{1}{n}$	$3 + \dfrac{n+1}{n^2+1}$
$\mathbb{P}(X_n = x)$	$\dfrac{1}{n}$	$1 - \dfrac{1}{n}$

Check, using only the definitions, if X_n converges to some limit in three senses: in Probability, in Quadratic Mean and in Distributions.

b.

Assume $X_n \sim Exp(\frac{1}{n})$ and $Y_n \sim Exp(n)$ (assume also that all r.v.s are defined on the same Probability Space). Check, using only the definitions, if

1. $X_n \xrightarrow{\mathbb{P}} 0$ and $Y_n \xrightarrow{\mathbb{P}} 0$
2. $X_n \xrightarrow{qm} 0$ and $Y_n \xrightarrow{qm} 0$
3. $X_n \xrightarrow{D} 0$ and $Y_n \xrightarrow{D} 0$

Note: You can "prove" or "disprove" the convergence in Distributions using graphs in **R**: say, you can plot the CDFs for different values of n to see the dynamics.

c. (R)

Assume $X_n \sim \mathcal{N}(0, \frac{1}{n})$. Guess the limit in Distributions of X_n, and "prove" that X_n indeed tends to your guess, in the Distributions sense, geometrically. To that end, you need to plot the CDFs for different increasing values of n, on the same graph, and also the CDF of the limit. Use different colors/line types for different n-s and the limit. Add also the legend (explanation which line is for which CDF).

Note: In the `plot` function, you can change the line type by using the `lty` parameter. Try, for example `lty=1`, `lty=2`, `lty=3`,... . To add a legend to a graph, use the `legend` function, see, e.g., this link.

Problem 3: LLN, CLT and Co

a.

Assume we have a sequence of IID r.v.s X_k with

$$X_k \sim Exp(2).$$

Calculate the following limits in Probability:

1. $\lim\limits_{n\to\infty} \dfrac{X_1^2 + X_2^2 + ... + X_n^2}{n}$
2. $\lim\limits_{n\to\infty} \Big(X_1 \cdot X_2 \cdot ... \cdot X_n\Big)^{1/n}$

Note1: Do some transformations to bring to the form containing

$$\frac{Y_1 + Y_2 + ... + Y_n}{n}$$

with some Y_k-s

Note2: You can use the fact that the integral

$$\int_0^{+\infty} e^{-x}\ln(x)dx = -\gamma,$$

where $\gamma \approx 0.57721$ is the Euler-Mascheroni constant[1].

b.

Let $\Phi(x)$ be the CDF of the Standard Normal Distribution. Assume $X_1, X_2, ..., X_n$ are IID r.v. with the Mean μ and variance σ^2, and let

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

Express, approximately, the Probability

$$\mathbb{P}\left(\overline{X}_n > A\right)$$

in terms of $\Phi(x)$ and the given parameters μ, σ, n, A.

c. (R)

Here we want to get the approximate Distribution of

$$Y_n = \frac{X_1^2 + X_2^2 + ... + X_n^2}{n}$$

for a sequence of IID r.v. X_k with $X_k \sim Unif[0,1]$, using the Monte-Carlo Simulations method. Say, let us fix $n = 50$ and find the approximate Distribution of Y_{50}.

- Take $n = 50$ and choose a (large) number of iteration (observations) *iter*;
- generate $n * iter$ size random numbers from $Unif[0,1]$ and keep in the variable x;
- calculate the squares of elements of x and split into a DataFrame - each column of the DataFrame need to contain $n = 50$ squares (one observation for $X_1^2, ..., X_{50}^2$), and you need to have *iter* columns;
- use `sapply` to calculate the mean of each column, and keep the resulted means vector (of size *iter*) in the variable y;
- plot the Density Histogram of y, using 20 bins;

Now we need to add the Theoretical Approximate Distribution obtained from the CLT:

- Find the Approximate Distribution of Y_{50}, using the CLT
- Plot the graph of the PDF of that Distribution over the Histogram obtained above
- Enjoy! (If you have done it correctly, of course $\ddot\smile$)

[1] See https://en.wikipedia.org/wiki/Euler-Mascheroni_constant

Problem 4: Statistical Models, Estimators and Estimates

a.

Write Statistical Models for the following studies, describe the Statistical Problem you want to solve (i.e., describe the Parameter(s) you want to Estimate, and based on what you will Estimate). Give 2 different Estimators for your Parameters. So you need to give:

- A Parametric Family of Distributions, with indication of the Parameter Set;
- Description of the Random Sample;
- What you want to do, what is the Problem you want to solve;
- The Method of your Solution, i.e., using which Estimators you will use.

Give also some reasoning behind your choice. Note that here, at this stage, it is not necessary to give some very nice Estimators, just give *some* Estimators.

1. We want to study (Model) the number of AUA Caffeteria visitors during the 12PM-1PM period. We want to use 14-day information to build the Model.
2. We want to Model how much money a person is spending daily in AUA Caffeteria. We want to ask 300 persons to obtain the Model.
3. We want to see which percentage of AUA Caffeteria visitors are vegetarians. We want to ask 100 persons to get the Model.
4. We want to model the time between two Khachapuri orders in AUA Caffeteria. We will use 50 observations of "time-to-the next Khachapuri order".

b.

1. We consider the following Problem: We have the observation

$$3, 2, 0, 1, 2, 1, 3, 4, 5, 1, 0, 1$$

   coming, supposedly, from the Poisson Distribution, and we want to Estimate the Parameter of our Poisson Distribution.

- Build the Model for this Problem, with a Random Sample;
- use the Sample Mean as an Estimator for the Parameter;
- Calculate the Estimate of the Parameter.

2. We consider the Problem of Estimation of the Parameter p for the Geometric Distribution using the Random Sample

$$X_1, X_2, ..., X_n \sim Geom(p).$$

Which one of the followings are Estimators for p:

$$\hat{p}_1 = X_2 + X_4; \qquad \hat{p}_2 = \frac{X_1 + X_2 + ... + X_n}{n}; \qquad \hat{p}_3 = \frac{1}{3},$$

$$\hat{p}_4 = \frac{p + X_n}{2}; \qquad \hat{p}_5 = 1000 + X_n; \qquad \hat{p}_6 = \frac{X_{(1)} + X_{(n)}}{2}, \qquad \hat{p}_7 = \ln(X_1 \cdot X_2 \cdot ... \cdot X_n).$$

For each Estimator you will find, calculate the Estimate for p, if we have the following observation ($n = 4$):

$$3, 5, 1, 1.$$

Problem 5: Bias, MSE, Comparison of Estimators, and the Standard Error

a.

Assume we use the Random Sample

$$X_1, X_2, ..., X_n \sim Pois(\lambda)$$

to estimate the unknown Parameter $\lambda > 0$.

To that end, we consider the following Estimator:

$$\hat{\lambda} = \frac{X_1 + 2 \cdot X_2 + 3 \cdot X_3 + ... + n \cdot X_n}{1 + 2 + 3 + ... + n}.$$

Calculate the Bias and the Risk (MSE) when estimating λ using $\hat{\lambda}$.

b.

Let $X_1, X_2, ..., X_n$ be IID from a Family of Distributions with unknown Mean μ and known Variance σ^2. We want to estimate μ, based on the observation from $X_1, ..., X_n$.

1. To estimate μ, we consider two Estimators:

   $$\hat{\mu}_1 = \frac{3}{4}X_1 + \frac{1}{4}X_n \qquad \text{and} \qquad \hat{\mu}_2 = \frac{1}{7}X_1 + \frac{6}{7}X_n.$$

   Show that these two Estimators are Unbiased, and choose the preferable one from these two.

2. Among all Estimators of the form

   $$\hat{\mu}_a = aX_1 + (1 - a)X_n, \qquad a \in [0, 1],$$

   find the one with the minimal Risk.

c.

In the Problem 2b-1, Calculate the Standard Error and the Estimated Standard Error.

Problem 6: Properties of Estimators: Unbiasedness and Consistency

a.

Assume we have a Random Sample from one of the Distributions of the family $\{\mathcal{N}(2, \sigma^2) : \sigma^2 \in (0, +\infty)\}$:

$$X_1, X_2, ..., X_n \sim \mathcal{N}(2, \sigma^2),$$

and we want to estimate σ^2.

1. We take
$$\widehat{\sigma^2} = \frac{1}{n-1} \cdot \sum_{k=1}^{n} (X_k - \overline{X})^2,$$
where
$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}.$$
Is $\widehat{\sigma^2}$ Unbiased/Asymptotically Unbiased? Is it Consistent?

2. Now we take
$$\widehat{\sigma^2} = \frac{1}{n-1} \cdot \sum_{k=1}^{n} (X_k - 2)^2,$$
Note the difference: instead of the sample mean $\overline{X}$ here we have the theoretical mean $\mu = 2$, which we know.

   Is $\widehat{\sigma^2}$ Unbiased/Asymptotically Unbiased? What if we will take $\frac{1}{n}$ instead of $\frac{1}{n-1}$? Is this Estimator Consistent?

3. Next, we want to Estimate σ, and we take
$$\widehat{\sigma} = \sqrt{\frac{1}{n} \cdot \sum_{k=1}^{n} (X_k - 2)^2}$$
as an Estimator. Show that this Estimator is Consistent.

   (Supplementary) What about the Unbiasedness?

b.

Assume we want to Estimate λ in the Exponential Model: we take a Random Sample
$$X_1, X_2, ..., X_n \sim Exp(\lambda),$$
and use the Estimator
$$\hat{\lambda}_n = \frac{n}{X_1 + X_2 + ... + X_n}$$
to Estimate λ.

1. Show that $\hat{\lambda}_n$ is a Consistent Estimator for λ;
2. (Supplementary) Show that $\hat{\lambda}_n$ is a Biased Estimator for λ.

c. (R)

We want to check the results of the **Problem 4b**, using **R**. We consider two subproblems: first one is about Consistency. To see the Consistency,

- Fix some value of λ;
- Take large n;
- Generate a random sample of size n from the $Exp(\lambda)$ Distribution;

- For each $k = 1, 2, 3, ..., n$, calculate $\hat{\lambda}_k$, and plot these points on the graph against k, joining the points with lines (use the `type = "l"` as a parameter in `plot` command);
- Now, do the 3-rd step several times, and plot on the previous plot, each time using a different color;
- Finally, add a horizontal line $y = \lambda$, passing through the real value of λ.
- Explain the results

Now, about the Biasedness:

- Fix some value of λ;
- Take $n = 20$;
- Take a number of iterations *iter*;
- Generate a random sample of size $n * iter$ from the $Exp(\lambda)$ Distribution;
- Put the sample in the DataFrame, so that we will have Samples of size n in each column, and the number of columns will be *iter* (so *iter* times we are generating an observation of size n);
- Calculate $\hat{\lambda}$ for each observation; you need to obtain a vector of $\hat{\lambda}$-s of size *iter*;
- Calculate the Mean of $\hat{\lambda}$-s minus the actual value of λ.
- increase the number of iterations *iter* and do your calculations again to see if the difference between the mean of $\hat{\lambda}$-s and λ is getting closer to 0
- Explain the results
- (Supplementary) Find a way to visualize the fact that when $iter \to \infty$, the difference between the mean of $\hat{\lambda}$-s and λ is not tending to zero.

Note: By calculating the mean of $\hat{\lambda}$-s for large *iter*, we are approximating $\mathbb{E}(\hat{\lambda})$. So if *iter* is large, then the difference between the mean of $\hat{\lambda}$-s and λ is an approximation for the $Bias(\hat{\lambda}, \lambda)$.

d. (R)

Assume we want to test if the coin is fair. To this end we want to use the Estimator

$$\hat{p} = \frac{X_1 + X_2 + ... + X_{10}}{10}.$$

The attached Excel file *Hw06p4d.xlsx* contains results for 150 experiments of tossing a coin 10 times (150 realizations/experiments for X_1,...,X_{10}, 0 means Tails, 1 means Heads, the first column is the number of the experiment). Construct the distribution (histogram and KDE) of $\hat{p}$ using **R**, calculate the mean of all estimates $\hat{p}$. What do you think, was the coin fair? Explain your reasoning.

Problem 7, Supplementary: Sampling Distributions

a.

Assume $X_1, X_2, ..., X_n$ are IID from some Distribution with CDF $F(x)$.

a. Find the CDF for order statistics $X_{(1)}$ and $X_{(n)}$. Here, as for datasets,

$$X_{(1)} = \min\{X_1, ..., X_n\} \qquad \text{and} \qquad X_{(n)} = \max\{X_1, ..., X_n\}$$

b. (**R**) Here we want to find the distributions of $X_{(1)}$ and $X_{(n)}$ by simulations. Assuming $X_k \sim Unif[-1, 1]$, generate observations $x_1, ..., x_n$ many times and calculate the corresponding values of $X_{(1)}$ and $X_{(n)}$. Then draw the density histogram, KDE, and the theoretical PDF obtained from the above calculations on the same plot, both for $X_{(1)}$ and $X_{(n)}$.