# Assignment 2a, Due: December 23rd, 2023

**Università della Svizzera italiana**

Probability & Statistics SA 2023-2024

*Assignment's Theoretical Section*

1. [5] **Random Variables + discrete probability.**
Consider the following pseudo-code:

   - `Set the parameters:` $n = 1000$ `(sample size) and` $p = 0.001$ `(probability of success).`
   - `Generate` $5000$ `uniform random numbers` $u$ `between` $0$ `and` $1$`.`
   - `For each random number` $u$ `generated, find the corresponding value` $x$ `value via inverse sampling using a binomial cdf.`
   - `Plot a histogram of the generated` $x$ `values.`
   - `Overlay the true probability mass functions for both the Binomial and Poisson distributions on the histogram.`

   What do you expect from this plot? And why?

2. [5] **Conditional probability + Bayes theorem.**
**Basics of Sentiment Analysis**
The **sentiment random variable** may assume

$$S = \begin{cases} -1 & 0 \leq u \leq 0.4 \quad \text{(Situation 1: Negative Sentiment)} \\ +1 & 0.4 < u \leq 1 \quad \text{(Situation 2: Positive Sentiment)} \end{cases}$$

   Consider the list of all the possible words (countable). The **word random variable** $W$ may assume discrete values $w$ representing the integer position of the word in the list. If a sentence is composed by $n$ words, we observe $n$ realizations of the random variable $W$. You aim to determine the following probabilities:

$$P(S = 1 | W_1 = w_1, ..., W_n = w_n)$$
$$P(S = -1 | W_1 = w_1, ..., W_n = w_n)$$

   for 4 User Reviews of Oppenheimer (I) (2023):

   – Horribly boring film with no action, only people talking and talking about really uninteresting things.
   – There's a terrible soundtrack of melodramatic music to make sure you know how you're supposed to feel, frequent random booming and rumbling to make your liver quiver, and screen-filling explosions that come out of nowhere and seem designed only to startle - sort of a science equivalent of jump scares.

– Half of this movie is something related with quantum physic and the rest all about law and politic. However if you aren't into those things, you can still understand and enjoy the movies.

– Other actors performances were amazing and it's clear that a lot of effort, passion and planning went into this movie.

**Sentiment Dictionary:**

| Word | : | Positive Probability | Negative Probability |
|---|---|---|---|
| horribly | : | 0.1 | 0.9 |
| boring | : | 0.1 | 0.9 |
| film | : | 0.6 | 0.4 |
| action | : | 0.7 | 0.3 |
| people | : | 0.6 | 0.4 |
| talking | : | 0.2 | 0.8 |
| terrible | : | 0.2 | 0.8 |
| soundtrack | : | 0.8 | 0.2 |
| melodramatic | : | 0.2 | 0.8 |
| music | : | 0.5 | 0.5 |
| feel | : | 0.8 | 0.2 |
| frequent | : | 0.6 | 0.4 |
| random | : | 0.7 | 0.3 |
| booming | : | 0.3 | 0.7 |
| rumbling | : | 0.3 | 0.7 |
| liver | : | 0.1 | 0.9 |
| quiver | : | 0.1 | 0.9 |
| screen-filling | : | 0.2 | 0.8 |
| explosions | : | 0.3 | 0.7 |
| science | : | 0.6 | 0.4 |
| equivalent | : | 0.6 | 0.4 |
| jump | : | 0.6 | 0.4 |
| scares | : | 0.3 | 0.7 |
| half | : | 0.5 | 0.5 |
| related | : | 0.7 | 0.3 |
| quantum | : | 0.8 | 0.2 |
| physic | : | 0.8 | 0.2 |
| rest | : | 0.5 | 0.5 |
| law | : | 0.2 | 0.8 |
| politic | : | 0.2 | 0.8 |
| enjoy | : | 0.8 | 0.2 |
| actors | : | 0.8 | 0.2 |
| performances | : | 0.8 | 0.2 |
| amazing | : | 0.9 | 0.1 |
| effort | : | 0.8 | 0.2 |
| passion | : | 0.9 | 0.1 |
| planning | : | 0.8 | 0.2 |

3. $\boxed{5}$ **Continuous distribution + Exponential.**
Let $X$ be a random variable following an Exponential distribution with rate parameter $\lambda > 0$, representing the time until the occurrence of an event. Prove the memoryless property of the exponential distribution by demonstrating that, for any non-negative constants $a$ and $b$, the following holds:

$$P(X > a + b | X > a) = P(X > b)$$

Interpret this result.

4. $\boxed{5}$ **Linear regression + estimation.**
The aim of a study is to find out the effect of exposure to sunlight on the increase of freckles per exposure hour. A study of 100 people find that the effect is $\hat{\beta} = 0.3$.

(a) Interpret the coefficient.

(b) Then a bootstrap is performed. Explain in detail:

i. How the bootstrap works

ii. The purpose of the bootstrap

5. [5] **Logistic regression + prediction.**

(a) Explore the role of logit and logistic function in the logistic regression. Find one function as the inverse of the other.

(b) How do you evaluate the accuracy? Would you evaluate it in your training or test data? Explain why.

6. [5] **Hypothesis testing**
In a recent music festival, organizers measured the duration of unplanned pauses during 100 performances due to suspected technical problems. The mean duration was found to be 1.7 seconds.

(a) State the null and the alternative hypothesis.

(b) The subsequent analysis yielded a p-value of 0.0001. Did the organizers guess correctly?

---

*Assignment's Practical Section, Due: December 23rd, 2023*

7. **You are the CEO of a start-up company....** Just imagine that you are a CEO of a start-up company in medical diagnostics. You are planning to launch a new diagnostic tool to detect diabetes in women using easily available information without invasive testing. You want to convince investors to invest in your company and you need to given them a report with the potential of your test.

   You want to combine easily available values, such as

   (a) **pregnant**: number of times pregnant

   (b) **glucose**: Plasma glucose concentration at 2 hours in an oral glucose tolerance test

   (c) **diastolic**: Diastolic blood pressure (mm Hg)

   (d) **triceps**: Triceps skin fold thickness (mm)

   (e) **insulin**: 2-Hour serum insulin (mu U/ml)

   (f) **bmi**: Body mass index (weight in kg/(height in metres squared))

   (g) **age**: Age (years)

   with a new test developed by your company, called **diabetes**, which measure a score based on the prevalence of diabetes in your family.

   You have performed a clinical trial with 768 women, where you also recorded whether or not the women had diabetes (the outcome of **test**). The data is available on the iCorsi website.

   Write a report covering the following points

(a) [5] **Research problem.** Formulate a research question (Hint: it should contain two elements).

(b) [15] **Exploratory analysis.** Perform an informal analysis of the data (i.e. tables and/or plots of the data) relevant for answering the research question.

i. Evaluate the distribution of each variable, including **pregnant**, **glucose**, **diastolic**, **triceps**, **insulin**, **bmi**, **age**, and **diabetes**. Look for patterns, outliers, or anomalies using histograms, box plots, or other graphical methods.

ii. Check the completeness of the data. Search for missing or blank values in all the variables and consider their potential impact on your analysis.

iii. Investigate the interactions between variables. Use scatter plots or correlation matrices to identify relationships or trends, particularly those that might be crucial in predicting the outcome of the **test**.

(c) [40] **Formal analysis.** Perform a formal analysis of the data based upon several predictors.

i. Split the data into a training set (90%) and a test set (10%) using your student number as seed of the random number generator. Use your own function to split the data.

ii. Fit the test score as a function of the other variables using your own function. Use as input of this function the response and the predictor variables of the training set and as output the estimate of the coefficients.

iii. Interpret the coefficients associated with the variables **pregnant** and **diastolic**.

iv. Perform bootstrap (with B=1000 bootstrap samples) and test whether the coefficients from (ii) are significant at the $\alpha$ =5% significance level.

v. Select the model with only significant variables as the proposed diagnostic test. Fit the new model on the training data and evaluate its performance on the test data in terms of false positive and false negative rates.

(d) [5] **Conclusions.** Answer the initial research problem.

(e) [5] **Discussion.** Provide a brief discussion about some critical aspects in the analysis.

8. [10] **Bonus.** Make a video to accompany your report.