Problems on distributions

(recap on past weeks)

Waiting [20 points]

Two friends plan to meet to go to a nightclub. Each of them arrives at a time uniformly distributed between midnight and 1am and independently of the other. Denote by X (respectively Y) the random variable representing the arrival time of the earlier person (respectively, the later one).

- Formulate the joint PDF and CDF of X and Y
- Find the probability that the earlier person is waiting for her friend for more than 10 minutes.
- Determine the marginal PDFs of X and Y. Check that they are indeed PDFs by integrating them.
- Calculate the means $E(X)$ and $E(Y)$ and the variances $Var(X)$ and $Var(Y)$.
- Find the conditional density of X given that $Y = y$. Check that it is indeed a PDF.

Diagnostics [20 points]

For patients with a certain disease, concentration of the protein SHZR is distributed normally with mean 25 and variance 1. For healthy patients, it is again normal, with mean 20 and variance 1. If you expect 1% of patients to have this syndrome, then what is the probability of syndrome if the level of SHZR equals x? Express it as a function of x and find its values for $x = 20$, $x = 23$ and $x = 25$.

Problems on hypothesis testing.

Each of these problems can be solved in a fully digital environment, such as ipynb. You can submit handwritten solutions as well, but they should be detailed enough to reflect all the important steps.

Model quality [20 points]

I built a new model for music recommendation, and it gave 1235 good recommendations out of 1600. The previous model is known to give 75% of good recommendations. Can I be sure that my model is an improvement? Provide an analytical p-value and a bootstrapped one to test the null hypothesis of no improvement vs the alternative of positive improvement.

Comparing salaries [20 points]

The data are salaries corresponding to two kinds of occupations: (1) creative, media, and marketing and (2) education. Suppose that the datasets are modeled as realizations of normal distributions. Test the null hypothesis that the salary for both occupations is the same at 5% significance level. Don't assume equal variance in two groups.

Counting bombs [20 points]

The table below gives the number of bombs falling into the South of London during WWII. The South of London was divided into n = 576 regions of 0.25 km^2 each. In the table, the number n_k corresponds to the number of domains bombed exactly k times. We want to estimate whether bombs were falling on South of London "at random".

k	0	1	2	3	4	5+
n_k	229	211	93	35	7	1

The total number of bombs is $\sum k n_k = 537$. Can we claim that the number of bombs per region has Poisson distribution? We can answer with the Chi-squared goodness of fit test:

1) Estimate the parameter λ for Poisson distribution from our data, and calculate expected number of regions $\tilde{n}_k$ with number of bombs from 0 to 5 (by multiplying the corresponding probability by 576).
2) If our assumption of Poisson distribution is true, then n_k has approximately normal distribution with mean $\tilde{n}_k$ and variance $\tilde{n}_k$. If so, the statistic $T = \sum_k \frac{\left(n_k - \tilde{n}_k\right)^2}{\tilde{n}_k}$ has approximate χ^2 distribution with 4 degrees of freedom (6 cells, minus 2 restrictions on the counts: total sum and conformity with Poisson distribution). Calculate the sample value of T.
3) Calculate p-value as 1 minus χ^2_4 CDF of T at its sample value. Is it low enough to reject the hypothesis about the Poisson distribution?