DSE 210: Probability and Statistics using Python

Worksheet 5 — Fitting distributions to data

1. A real number X is drawn from the Gaussian distribution $N(10, 16)$.
   (a) What is the probability that $X \geq 10$?
   (b) What is the probability that $X = 10$?
   (c) What is the probability (roughly) that $X \geq 14$?
   (d) What is the probability (roughly) that $X \leq 2$?

2. A call center keeps track of the number of phone calls they receive: over a period of 500 hours, they record the number of calls received during every one-hour interval (the number of calls during the first hour, during the second hour, and so on). Let N_k be the number of one-hour intervals during which k calls were received, for $k = 0, 1, 2, \ldots$. Here is their data:

k	0	1	2	3	4	5	6	7	8	≥ 9
N_k	22	66	106	115	85	55	28	13	10	0

   Notice that $N_0 + N_1 + \cdots = 500$.

   (a) You decide to model the number of calls received in an hour by a Poisson(λ) distribution. What value of λ should you choose?
   (b) Under this choice of λ, what are the expected entries in the table above, i.e. the expected number of one-hour intervals (out of 500) during which k calls are received, for $k = 0, 1, \ldots$?

3. Show that the *mode* of the Poisson(λ) distribution—that is, the point with highest probability—is $\lfloor \lambda \rfloor$.

4. *Maximum likelihood and smoothing.* Upon tossing a coin 20 times, you get heads every time.
   (a) How would you estimate the bias (that is, the heads probability) of the coin, using maximum likelihood?
   (b) How would you estimate the bias of the coin, using maximum likelihood *with Laplace smoothing*?
   (c) Under your estimate from part (b), what is the probability of seeing the sequence of tosses $HHTTHH$? You don't need to simplify the numeric expression.

5. *Fitting a multinomial.* Fix the following vocabulary: $V = \{\texttt{a}, \texttt{rose}, \texttt{is}, \texttt{flower}\}$.
   (a) In a bag-of-words representation, what is the vector form of the document `‘‘A rose is a rose is a rose’’`?
   (b) Fit a multinomial model to this document using maximum likelihood but not Laplace smoothing. What is the resulting distribution (give the probability of each word in V)?
   (c) Same as part (b), but with Laplace smoothing.

6. *Fitting an exponential distribution by maximum likelihood.* The exponential distribution with parameter $\lambda > 0$ is a distribution over $(0, \infty)$ with the following density:

$$p(x) = \lambda e^{-\lambda x}.$$

Suppose we observe data $x_1, \ldots, x_n > 0$ and we want to fit an exponential distribution to it. In this problem, we will derive the maximum-likelihood choice of λ.

   (a) Write down the likelihood function $\Pr(\text{data}|\lambda)$.

   (b) Write down the log-likelihood $LL(\lambda) = \ln \Pr(\text{data}|\lambda)$.

   (c) Use calculus to determine the value of λ that maximizes the log-likelihood.

7. For any real number $\lambda > 0$, let U_λ denote the uniform distribution over $[0, \lambda]$.

   (a) Write down the formula for the density of U_λ.

   (b) Given a set of observations $x_1, x_2, \ldots, x_n > 0$, we decide to fit a U_λ distribution to them. What is the maximum-likelihood choice of λ?

8. Suppose $Z_1, \ldots, Z_k$ are independent $N(0, 1)$ random variables. Define $X = Z_1^2 + \cdots + Z_k^2$. The distribution of X is called the chi-squared distribution with k degrees of freedom.

   (a) Plot the density of the chi-squared distribution with 10 degrees of freedom using `scipy.stats.chi2`.

   (b) By generating random samples from this distribution, estimate its median.