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 $X \leq 2$?

Solution:

(a) $\mathbf{Pr(X \geq 10) = 0.5}$
(b) $\mathbf{Pr(X = 10) = 0}$
(c) $\mathbf{X \geq 14 = 1 - \Phi(1) = 0.1587}$
(d) $\mathbf{X \leq 2 = \Phi(-2) = 0.0228}$

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,.... Here is their data:
Notice that $N_0 + N_1 + \; = 500$.
(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,...?

Solution:

(a)

$$\lambda = \frac{\sum_{k=0}^{\infty} k * N_k}{\sum_{k=0}^{\infty} N_k} = 1577/500 = 3.154$$

(b)

k	0	1	2	3	4	5	6	7	8	≥ 9
N_k	22	66	106	115	85	55	28	13	10	0
$k * N_k$	0	66	212	345	340	275	168	91	80	0
Pr	0.04268	0.1346	0.2122	0.2232	0.1760	0.1110	0.0584	0.0263	0.0104	0.0052
E	21.34	67.31	106.14	111.59	87.99	55.51	29.18	13.15	5.18	2.61

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.

Solution:

(a) MLE of the bias: $\mathbf{P = 20/20 = 1}$
(b) MLE of the bias with Laplace smoothing: $\mathbf{P = \frac{20+1}{20+2} = 21/22}$
(c) Probability of $HHTTHH = 21/22 * 21/22 * 1/22 * 1/22 * 21/22 * 21/22 = \mathbf{21^4/22^6}$

5. *Fitting a multinomial.* Fix the following vocabulary: V = {a, rose, is, 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.

Solution:

(a) Vector form of "A rose is a rose is a rose" is: **{3, 3, 2}**
(b) **P("a")= 1/4, P("rose")= 1/4, P("is")= 1/4, P("flower")= 1/4**
(c) **With Laplace Smoothing: P("a")= 2/8, P("rose")= 2/8, P("is")= 2/8, P("flower")= 2/8**

8. Suppose $Z_1, ..., Z_k$ are independent N(0, 1) random variables. Define $X = Z_1^2 + ... + Z_k^2$. 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.

Solution:

(a)

```
import numpy as np
from scipy.stats import chi2
import matplotlib.pyplot as plt

# degree of freedom
df = 10
mean, var, skew, kurt = chi2.stats(df, moments='mvsk')

x = np.linspace(chi2.ppf(0.01, df),
                chi2.ppf(0.99, df), 100)

# plot chi-square density function
fig, ax = plt.subplots(1, 1)
ax.plot(x, chi2.pdf(x, df),
        'r-', lw=3, alpha=0.6, label='chi2 pdf')
```

(b)

```
samples = chi2.rvs(df, size=1000)
estimated_median = np.median(r)
print(estimated_median)
```