Computational Statistics and Data Analysis (MVComp2)

Solutions to exercise 3

Lecturer Tristan Bereau

Semester Wi23/24
Due Nov. 9, 2023, 23:59

1 Polluted water (2 points)

A particular concentration of a chemical found in polluted water has been found to be lethal to 20% of the fish that are exposed to the concentration for 24 hours. Twenty fish are placed in a tank containing this concentration of chemical in water.

(a) Find the probability that at least 14 survive.

(b) Find the mean and variance of the number that survive.

1.1 Solution

(a)

First note that

$$P(X \geq 14) = 1 - P(X < 14) = 1 - P(X \leq 13)$$

For $P(X \leq 13)$ use the binomial cumulative distribution function with survival probability 0.8:

$$P(X \leq 13) = \sum_{k=0}^{13} \binom{20}{13} 0.8^{13} 0.2^{20-13}.$$

This yields a probability $P(X \geq 14) \approx 91.3\%$.

(b)

Mean

$$\mu = N\pi = 16$$

Variance :

$$\mathrm{Var}[X] = N\pi(1-\pi) = 3.2$$

2 Distribution means (4 points)

(a) If X is a random variable with a geometric distribution $P(X=x) = (1-p)^{k-1}p$, prove that the mean is given by

$$E[X] = \frac{1}{p}$$

(b) If X is a random variable with a Poisson distribution $P(X=x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}$, prove that the mean is given by

$$E[X] = \lambda$$

2.1 Solution

(a)

$$E[X] = \sum_{y=1}^{\infty} y q^{y-1} p = p \frac{\mathrm{d}}{\mathrm{d}q}\left(\sum_{y=1}^{\infty} q^y\right) = p \frac{\mathrm{d}}{\mathrm{d}q}\left(\frac{q}{1-q}\right) = p\left[\frac{1}{(1-q)^2}\right] = \frac{1}{p}$$

(b)

$$E[X] = \sum_{i=0}^{\infty} i \frac{\lambda^i}{i!}\mathrm{e}^{-\lambda} = \sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!}\mathrm{e}^{-\lambda} = \lambda \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}\mathrm{e}^{-\lambda} = \lambda \sum_{i=0}^{\infty} \frac{\lambda^l}{l!}\mathrm{e}^{-\lambda} = \lambda$$

3 Conjugate priors (2 points)

Consider the Beta distribution with parameters α and β

$$\mathrm{Beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

and the Binomial distribution for N observations

$$P(X=k) = \binom{N}{k} \theta^k (1-\theta)^{N-k}.$$

Show that the Beta distribution is a conjugate prior to the Binomial distribution. What are the parameters of the resulting distribution?

3.1 Solution

Posterior from Bayes' rule:

$$\begin{aligned} p(\theta|X=k) &= p(X=k|\theta)p(\theta) \\ &= \binom{N}{k} \theta^k (1-\theta)^{N-k} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+k-1}(1-\theta)^{\beta+N-k-1}. \end{aligned}$$

The posterior indeed follows a Beta distribution, with parameters $\alpha' = \alpha + k$ and $\beta' = \beta + N - k$.

4 The de Moivre–Laplace theorem (2 points)

The de Moivre–Laplace theorem states that the normal distribution may be used as an approximation to the Binomial distribution under certain conditions.[1] Let's find out!

Consider a random variable, X. The probability of getting k successes in n independent Bernoulli trials with probability p is given by the Binomial

$$P(X = k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The thoerem states that the Binomial probability will converge to the normal probability density function, $\mathcal{N}(np, \sqrt{np(1-p)})$, as n becomes large and for p away from 0 or 1.

We will verify that the standardized random variable, $\frac{X-np}{\sqrt{np(1-p)}}$, approaches the standard normal, $\mathcal{N}(0, 1)$, as n grows larger. Consider $p = 0.5$ and the cases $n = \{2, 5, 10, 50, 100\}$. Plot all curves on the same graph to show that they overlap.

4.1 Solution

```
import numpy as np
from scipy.stats import binom, norm
import matplotlib.pyplot as plt

fig, ax = plt.subplots(1, 1)

plt.xlim([-4., 4.])

x_norm = np.linspace(*plt.xlim(), num=100)
ax.plot(
  x_norm, norm.pdf(x_norm, loc=0., scale=1.), "--", label="unit Gaussian"
)

p = 0.50
ns = [2, 5, 10, 50, 100]
for n in ns:
  mean = n * p
  std = np.sqrt(n * p * (1.-p))

  # Generate a range for x (from 0 to n) and calculate the binomial probability
  x = np.arange(0, n+1)
  y_binom = binom.pmf(x, n, p)

  # Standardize the x values (convert to z-scores)
  z_scores = (x - mean) / std

  # Convert probability to a PDF-like form by scaling with std
```

[1]This is a special case of the central limit thoerem. A famous physical realization of the de Moivre–Laplace theorem is the Galton box, also called bean machine.

```
    y_scaled = y_binom * std

    # Plot the binomial distributions as step functions
    ax.plot(z_scores, y_scaled, drawstyle='steps-mid', label=f'Binomial n={n}')

plt.grid()
plt.legend()
plt.xlabel(r"$z$")
plt.ylabel(r"$P(z)$")
plt.show()
```