Probability and Statistics: MA6.101

Tutorial 9

Topics Covered: CLT, Random Vectors

Q1: During each day, the probability that an athlete misses training due to illness is 5%, independent of every other day. Find the probability that the athlete will attend training on at least 45 out of the next 50 days using the Central Limit Theorem (CLT) and otherwise.[Gopal]

Solution

Let X_i be a Bernoulli random variable that equals 1 if the athlete attends training on day i, and 0 otherwise. The probability of attending training on a single day is:

$$P(\text{attend}) = 1 - P(\text{miss}) = 1 - 0.05 = 0.95$$

Thus, $X_i \sim \text{Bernoulli}(0.95)$.

We are interested in the total number of training days attended over the next 50 days, $Y = X_1 + X_2 + \cdots + X_{50}$. Since Y is a sum of independent Bernoulli random variables, it follows a binomial distribution:

$$Y \sim \text{Binomial}(50, 0.95)$$

We are asked to find:

$$P(Y \geq 45)$$

Using Exact Binomial Distribution

The exact probability can be calculated by summing up the binomial probabilities from 45 to 50. The probability mass function for a binomial random variable is:

$$P(Y = k) = \binom{50}{k}(0.95)^k(0.05)^{50-k}$$

However, calculating this exactly can be tedious, so we proceed with an approximation using the Central Limit Theorem.

Using Central Limit Theorem (CLT)

By the Central Limit Theorem, for large n, a binomial distribution can be approximated by a normal distribution:

$$Y \sim N(\mu, \sigma^2)$$

where the mean μ and variance σ^2 are given by:

$$\mu = np = 50 \times 0.95 = 47.5$$
$$\sigma^2 = np(1-p) = 50 \times 0.95 \times 0.05 = 2.375$$
$$\sigma = \sqrt{2.375} \approx 1.541$$

We want to find $P(Y \geq 45)$. Using the normal approximation:

$$P(Y \geq 45) \approx P\left(Z \geq \frac{45 - 47.5}{1.541}\right) = P\left(Z \geq \frac{-2.5}{1.541}\right) = P(Z \geq -1.622)$$

Using the standard normal table:

$$P(Z \geq -1.622) \approx 0.947$$

Thus, the probability that the athlete will attend training on at least 45 out of the next 50 days is approximately 94.7% using the CLT approximation.

Q2: If 10 fair dice are rolled, find the approximate probability that the sum obtained is between 30 and 40. (Given $\Phi(\sqrt{\frac{6}{7}}) = 0.82$

A:
Let X_i represent the outcome of the i-th die roll, where each X_i is uniformly distributed between 1 and 6. The sum of 10 dice rolls is:

$$S = X_1 + X_2 + \cdots + X_{10}$$

For a single fair die:

$$\mathbb{E}[X_i] = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\text{Var}(X_i) = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = \frac{91}{6} - 3.5^2 = \frac{35}{12}$$

Using Normal Approximation based on CLT:
Let $S_n = X_1 + \cdots + X_n$ where X_i are i.i.d. with mean μ and variance σ^2. If n is large, the cumulative distribution function (CDF) of S_n can be approximated as follows:

$$P(S_n < c) \approx \Phi\left(\frac{c - n\mu}{\sigma\sqrt{n}}\right)$$

where $\Phi(z)$ denotes the CDF of the standard normal distribution.
Firstly, let's consider $P(S \leq 40)$.

$$P(S < 40) \approx \Phi\left(\frac{40 - 10 \times 3.5}{\sqrt{\frac{35}{12}} \times \sqrt{10}}\right)$$

$$P(S < 40) \approx \Phi\left(\sqrt{\frac{6}{7}}\right)$$

$$P(S < 40) \approx 0.82$$

Now,

$$P(S \leq 30) \approx \Phi\left(\frac{30 - 35}{\sqrt{\frac{350}{12}}}\right)$$

$$P(S \leq 30) \approx \Phi\left(-\sqrt{\frac{6}{7}}\right)$$

By the symmetry of the normal distribution, $\Phi\left(-\sqrt{\frac{6}{7}}\right) = 1 - \Phi\left(\sqrt{\frac{6}{7}}\right)$, so:

$$P(S \leq 30) \approx 1 - 0.82 = 0.18$$

So,

$$P(30 \leq S \leq 40) = P(S \leq 40) - P(S \leq 30)$$

$$P(30 \leq S \leq 40) \approx 0.82 - 0.18 = 0.64$$

Q3: The quadratic form of a random vector X is given by $X^T AX$. Find the expectation of the quadratic form. [Kavin]

A: Since $X^T AX$ is a scalar

$$E[X^T AX] = E[tr(X^T AX)]$$

Using the property of the trace operator

$$E[tr(X^T AX)] = E[tr(AXX^T)]$$

We can interchange the expectation and trace operators

$$E[tr(AXX^T)] = tr(AE[XX^T])$$

We know that $E[XX^T] = \Sigma + \mu\mu^T$

$$\begin{aligned} E[X^T AX] &= tr(A[\Sigma + \mu\mu^T]) \\ &= tr(A\Sigma) + tr(A\mu\mu^T)]) \\ &= tr(A\Sigma) + tr(\mu^T A\mu)]) \\ &= \mu^T A\mu + tr(A\Sigma) \end{aligned}$$

Q4: Using the central limit theorem show that

$$\lim_{n\to\infty} e^{-n} \sum_{k=0}^{n} \frac{n^k}{k!} = \frac{1}{2}$$

Hint: Let S_n be Poisson with mean n. Use the central limit theorem to show that $P\{S_n \leq n\} \to \frac{1}{2}$.

A:
Let $X_1, X_2, \dots$ i.i.d. with $X_i \sim \text{Poisson}(1)$. Then $\mu = E[X_i] = 1 = \text{Var}[X_i] = \sigma^2$. Further let

$$S_n = X_1 + \cdots + X_n,$$

thus $S_n \sim \text{Poisson}(n)$.

$$\begin{aligned}
e^{-n} \sum_{k=0}^{n} \frac{n^k}{k!} &= \sum_{k=0}^{n} e^{-n} \frac{n^k}{k!} \\
&= \sum_{k=0}^{n} \Pr(S_n = k) \\
&= \Pr(S_n \leq n) \\
&= \Pr\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{n - n\mu}{\sigma\sqrt{n}}\right] \\
&= \Pr\left[\frac{S_n - n\mu}{\sqrt{n}} \leq 0\right]
\end{aligned}$$

Using the central limit theorem, we approximate $\frac{S_n - n\mu}{\sqrt{n}}$ with $N(0,1)$, so

$$\Pr\left[\frac{S_n - n\mu}{\sqrt{n}} \leq 0\right] = \frac{1}{2}$$

Q5: Let $\mathbf{X} = [X_1, X_2]^T$ be a two-dimensional zero-mean Gaussian random vector with covariance matrix $\mathbf{C}$ given by:

$$\mathbf{C} = \begin{bmatrix} 1 & r \\ r & 2 \end{bmatrix}$$

1. Give an expression for $f_{X_2}(x_2)$. 2. Determine the conditional pdf, conditional mean, and conditional variance of X_1 given $X_2 = x_2$.

Solution

1. Marginal Distribution of X_2

We are given that the random vector $\mathbf{X}$ follows a multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{C}$. The marginal distribution of X_2 is also Gaussian, with mean 0 and variance given by the corresponding entry in the covariance matrix. Therefore:

$$X_2 \sim \mathcal{N}(0, 2)$$

The marginal pdf of X_2 is:

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi \cdot 2}} \exp\left(-\frac{x_2^2}{2 \cdot 2}\right) = \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{x_2^2}{4}\right)$$

2. Conditional Distribution of X_1 Given $X_2 = x_2$

The conditional distribution of X_1 given $X_2 = x_2$ for a multivariate Gaussian distribution can be derived using the following properties of conditional distributions:

For a bivariate normal distribution, the conditional distribution of one variable given the other is still a normal distribution

$$X_1|(X_2 = x_2 \sim \mathcal{N})\left(\mu_{X_1|X_2}, \sigma^2_{X_1|X_2}\right)$$

where:

- The conditional mean is given by:

$$\mu_{X_1|X_2} = \mathbb{E}[X_1|X_2 = x_2] = \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_2)} x_2 = \frac{r}{2} x_2$$

- The conditional variance is:

$$\sigma^2_{X_1|X_2} = \text{Var}(X_1|X_2) = \text{Var}(X_1) - \frac{\text{Cov}^2(X_1, X_2)}{\text{Var}(X_2)} = 1 - \frac{r^2}{2}$$

Thus, the conditional pdf of X_1 given $X_2 = x_2$ is:

$$f_{X_1|X_2}(x_1|x_2) = \frac{1}{\sqrt{2\pi\sigma^2_{X_1|X_2}}} \exp\left(-\frac{(x_1 - \mu_{X_1|X_2})^2}{2\sigma^2_{X_1|X_2}}\right)$$

Substituting the values for the conditional mean and variance, we get:

$$f_{X_1|X_2}(x_1|x_2) = \frac{1}{\sqrt{2\pi(1 - \frac{r^2}{2})}} \exp\left(-\frac{(x_1 - \frac{r}{2}x_2)^2}{2(1 - \frac{r^2}{2})}\right)$$

Q6: You are given the random vector $Y' = [Y_1, Y_2, Y_3, Y_4]$ with mean vector

$$\mu_Y = [5, -1, 4, -3]$$

and variance-covariance matrix

$$\Sigma_Y = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}.$$

Let

$$B = \begin{bmatrix} 2 & -1 & 0 & 1 \\ -1 & 2 & -1 & 0 \\ 0 & 1 & 1 & -2 \end{bmatrix}.$$

(a) Find $E(BY)$, the mean of BY.

(b) Find $\mathrm{Cov}(BY)$, the variances and covariances of BY.

(c) Which pairs of linear combinations have zero covariances?

A: Consider the (4×1)-dimensional random vector

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix},$$

whereby

$$\mu_Y = E(Y) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ E(Y_3) \\ E(Y_4) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} = \begin{bmatrix} 5 \\ -1 \\ 4 \\ -3 \end{bmatrix}$$

and

$$\Sigma_Y = \mathrm{Cov}(Y) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}.$$

Let

$$B = \begin{bmatrix} 2 & -1 & 0 & 1 \\ -1 & 2 & -1 & 0 \\ 0 & 1 & 1 & -2 \end{bmatrix}.$$

So,

$$E(BY) = B\mu_Y \quad \text{and} \quad \mathrm{Cov}(BY) = B\Sigma_Y B'.$$

Hence,

(a)

$$E(BY) = \begin{bmatrix} 2 & -1 & 0 & 1 \\ -1 & 2 & -1 & 0 \\ 0 & 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} 5 \\ -1 \\ 4 \\ -3 \end{bmatrix} = \begin{bmatrix} 16 \\ 9 \\ 1 \end{bmatrix}.$$

(b)

$$\mathrm{Cov}(BY) = \begin{bmatrix} 2 & -1 & 0 & 1 \\ -1 & 2 & -1 & 0 \\ 0 & 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & -1 & 1 \\ 1 & 0 & -2 \end{bmatrix}.$$

Calculating this yields:

$$\mathrm{Cov}(BY) = \begin{bmatrix} 21 & -6 & -10 \\ -6 & 13 & 8 \\ -10 & 8 & 21 \end{bmatrix}.$$

(c) Define $Z = BY = \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix}$, where

$$Z_1 = 2Y_1 - Y_2 + Y_4, \quad Z_2 = -Y_1 + 2Y_2 - Y_3, \quad Z_3 = Y_2 + Y_3 - 2Y_4.$$

Then, clearly,

$$\mathrm{Cov}(BY) = \Sigma_Z,$$

and therefore, according to (b),

$$\mathrm{Cov}(Z_i, Z_j) = 0 \quad \text{for } i \neq j, \text{ where } i, j = 1, 2, 3.$$