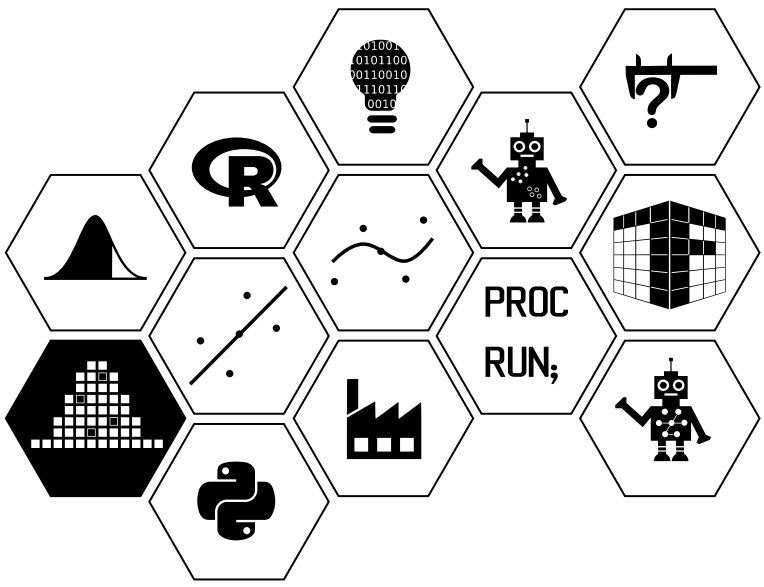
Probability and Sampling Fundamentals

Week 8: The Multivariate Normal Distribution and Large Sample Theory



DATA ANALYTICS
GLASGOW

The multivariate normal distribution and large sample theory

In Week 7 we learned about bivariate **continuous** random variables with a brief discussion on how to extend these ideas to a continuous random vector. The first part of this week's material will introduce the most commonly encountered standard distribution for a continuous random vector, the multivariate normal distribution, with a particular focus on the bivariate case. The second part will introduce some large sample theory including the law of large numbers and one of the most important theorems in probability, the central limit theorem.

Week 8 learning material aims

The material in week 8 covers:

- the multivariate normal distribution;
- calculating marginal and conditional distributions for the bivariate normal;
- the weak law of large numbers;
- the central limit theorem;
- the normal approximation to the binomial and Poisson distribution.

Vector-matrix notation for expected value and variance

In Week 4 and Week 7 we have looked at the expected value ("mean"), variance and covariance of random variables. In this section we will introduce the vector-matrix notation for the mean and (co)variance.

Suppose we have two random variables X_1 and X_2 and their expected values are given by $\mu_1 = E(X_1)$ and $\mu_2 = E(X_2)$.

We can think of X_1 and X_2 as being part of a vector $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. Similarly, we can arrange the expected values into a vector

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}.$$

We can arrange the variances $\sigma_1^2 = \mathrm{Var}(X_1)$ and $\sigma_2^2 = \mathrm{Var}(X_2)$ as well as the covariance $\sigma_{12} = \mathrm{Cov}(X_1, X_2)$ into a symmetric matrix, called the variance matrix or covariance matrix (or even sometimes variance-covariance matrix):

$$\text{Cov}(\mathbf{X}) = \mathbf{\Sigma} = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Notice that the values in the top right and bottom left of the matrix are the same, this is because the order of arguments does not matter for covariance, i.e. $\sigma_{12} = \sigma_{21}$.

This vector-matrix notation comes in especially handy when we look at linear transformations of random vectors.

In Week 4, we have seen that for univariate random variables X and Y, the linear function $Y = aX + b$ has expected value and variance

$$\begin{aligned} E(Y) &= E(aX + b) = aE(X) + b = a\mu + b \\ \text{Var}(Y) &= \text{Var}(aX + b) = a^2\text{Var}(X) = a^2\sigma^2 = a\sigma^2 a \end{aligned}$$

where $\mu = E(X)$ and $\sigma^2 = \text{Var}(X)$. You might, at this stage be wondering, why we have written the variance of Y as $a\sigma^2 a$, but we will see that the formula for the multivariate case will have that form.

Let's go back to the vector $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ and define a matrix $\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ and a vector $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$.

We can then define a random vector $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ as a linear function

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$$

or, equivalently

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} A_{11}X_1 + A_{12}X_2 + b_1 \\ A_{21}X_1 + A_{22}X_2 + b_2 \end{bmatrix}.$$

Then, the expected value and variance are given by

$$\begin{aligned} E(\mathbf{Y}) &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \\ \mathbf{Cov}(\mathbf{Y}) &= \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top \end{aligned}$$

The above formulae hold for vectors $\mathbf{X}$ and $\mathbf{Y}$ of any dimension, not just bivariate vectors.

Note that matrix multiplication is not commutative (i.e. the order in the multiplication matters), i.e. we *cannot* write the covariance matrix as $\mathbf{A}^2\mathbf{\Sigma}$, which would be more similar to the formula for the univariate case.

Supplement 1

Derivation of the formulae

We can work out the expected value, variances and covariance of Y_1 and Y_2 using the rules we have learned in Week 7.

For the expected values,

$$E(Y_1) = E(A_{11}X_1 + A_{12}X_2 + b_1) = A_{11}E(X_1) + A_{12}E(X_2) + b_1 = A_{11}\mu_1 + A_{12}\mu_2 + b_1$$
$$E(Y_2) = E(A_{21}X_1 + A_{22}X_2 + b_2) = A_{21}E(X_1) + A_{22}E(X_2) + b_2 = A_{21}\mu_1 + A_{22}\mu_2 + b_2$$

We can recognise that this is the same as

$$\begin{bmatrix} E(Y_1) \\ E(Y_2) \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

which means nothing other than

$$E(\mathbf{Y}) = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}.$$

Let's now turn to the variances and covariances. These are a lot more complicated, so let's only work out one variance.

$$\begin{aligned} \mathrm{Var}(Y_1) &= \mathrm{Var}(A_{11}X_1 + A_{12}X_2 + b) \\ &= \mathrm{Var}(A_{11}X_1) + \mathrm{Var}(A_{12}X_2) + 2\mathrm{Cov}(A_{11}X_1, A_{12}X_2) \\ &= A_{11}^2\mathrm{Var}(X_1) + A_{12}^2\mathrm{Var}(X_2) + 2A_{11}A_{12}\mathrm{Cov}(X_1, X_2) \\ &= A_{11}^2\sigma_1^2 + A_{12}^2\sigma_2^2 + 2A_{11}A_{12}\sigma_{12}. \end{aligned}$$

If we compare this to the top-left entry of

$$\begin{aligned} \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix} \\ &= \begin{bmatrix} A_{11}\sigma_1^2 + A_{12}\sigma_{12} & A_{11}\sigma_{12} + A_{12}\sigma_2^2 \\ A_{21}\sigma_1^2 + A_{22}\sigma_{12} & A_{21}\sigma_{12} + A_{22}\sigma_2^2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix} \\ &= \begin{bmatrix} A_{11}^2\sigma_1^2 + A_{12}^2\sigma_2^2 + 2A_{11}A_{12}\sigma_{12} & \dots \\ \dots & \dots \end{bmatrix} \end{aligned}$$

we can observe that these are the same. Hence,

$$\mathrm{Cov}(\mathbf{Y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top$$

The multivariate normal distribution

In Week 6 we learned about the univariate normal distribution. There is also a multivariate family of normal distributions. To begin with, let's consider the bivariate case. A univariate normal distribution has one random variable; a bivariate normal distribution is made up of two random variables. The two variables in the bivariate normal are both normally distributed, and they have a normal distribution when they are added together. Let's consider the bivariate normal distribution using the following example.

Example 1

Karl Pearson, a very famous statistician, analysed 1078 pairs of heights of fathers, and their adult sons in inches. Let

$$X_1 = \{\text{heights of fathers}\} \quad \text{and} \quad X_2 = \{\text{heights of adult sons}\}.$$

Let's assume that a bivariate normal distribution is appropriate for these data and the corresponding parameters are

$$\mu_1 = 67.7, \quad \mu_2 = 68.7, \quad \sigma_1 = 2.74, \quad \sigma_2 = 2.81.$$

where μ_1 is the mean height for fathers, μ_2 is the mean height for adult sons, σ_1 is the standard deviation for the height of fathers and σ_2 is the standard deviation for the height of adult sons.

In general, to characterise the bivariate normal distribution, we need the following parameters:

- the mean and variance for X_1 and X_2. These can be denoted as $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and
- the covariance between X_1 and X_2, denoted as σ_{12}.

So we need a total of 5 parameters, however only one of these parameters, the covariance σ_{12}, is needed to specify the **dependence** between the two random variables.

Rather than list all these parameters separately, it is more convenient and useful for calculations to write these in the vector matrix notation we have just seen, where we have a *mean vector* $\boldsymbol{\mu}$ and a *covariance matrix* $\boldsymbol{\Sigma}$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

The covariance matrix $\boldsymbol{\Sigma}$ can also be written in terms of the correlation ρ. In Week 4 we defined correlation as

$$\rho(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}}.$$

Which is equivalent to

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}.$$

We can rearrange this to be

$$\sigma_{12} = \rho_{12}\sigma_1\sigma_2.$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

We can now write that

$$\boldsymbol{X} \sim \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

If the random vector $\boldsymbol{X}$ follows a bivariate (or multivariate normal) with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Example 2

In Example 1, say we are also told that $\rho_{12} = 0.50$. So the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ in this example would be

$$\boldsymbol{\mu} = \begin{pmatrix} 67.7 \\ 68.7 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2.74^2 & 3.85 \\ 3.85 & 2.81^2 \end{pmatrix}$$

Since

$$\begin{aligned}\sigma_{12} &= \rho_{12}\sigma_1\sigma_2 \\ &= 0.50 \cdot 2.74 \cdot 2.81 \\ &= 3.85.\end{aligned}$$

The vector matrix notation shown above makes it easier to generalise to more than two random variables. In general, the **multivariate normal distribution** (MVN) is made up of k random variables and has some generic $k-$dimensional mean vector $\boldsymbol{\mu}$ and $k \times k$ covariance matrix $\boldsymbol{\Sigma}$.

For example, if $k = 3$ with random variables X_1, X_2, X_3 we have

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}$$

In Week 6 we discussed the characteristic bell-shaped curve of the normal density curve. The contour lines of the joint density of multivariate normal distributions have a characteristic *elliptical* shape to them. Below are some examples of bivariate random variables where X_1 and X_2 both follow standard normal distributions ($\mathrm{N}(0, 1)$) with varying amounts of correlation. The contours on the plot are in fact ellipses (for a two-dimensional MVN) centered on $\boldsymbol{\mu} = (0, 0)$ (in red). The elliptical regions moving outwards from the centre contain, respectively, 50%, 90%, 95% and 99% of the total probability.

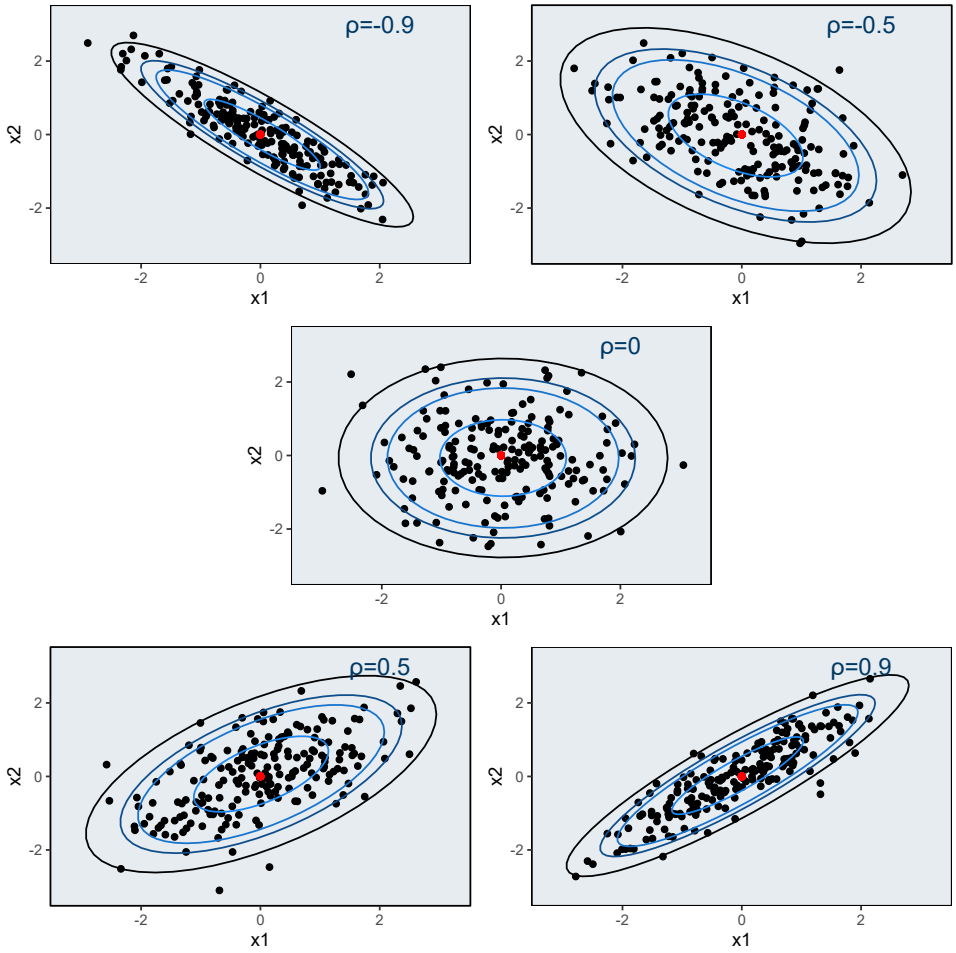


Figure 1

This video discusses the bivariate normal distribution and how changing the correlation parameter ρ affects the shape of the distribution.

Video

The bivariate normal distribution

Duration 3:58

Example 3

Visualising bivariate normal distributions

If we plot the data in Example 1 we can see the characteristic elliptical shape. Here, there is clearly a positive relationship between the father's and son's heights.

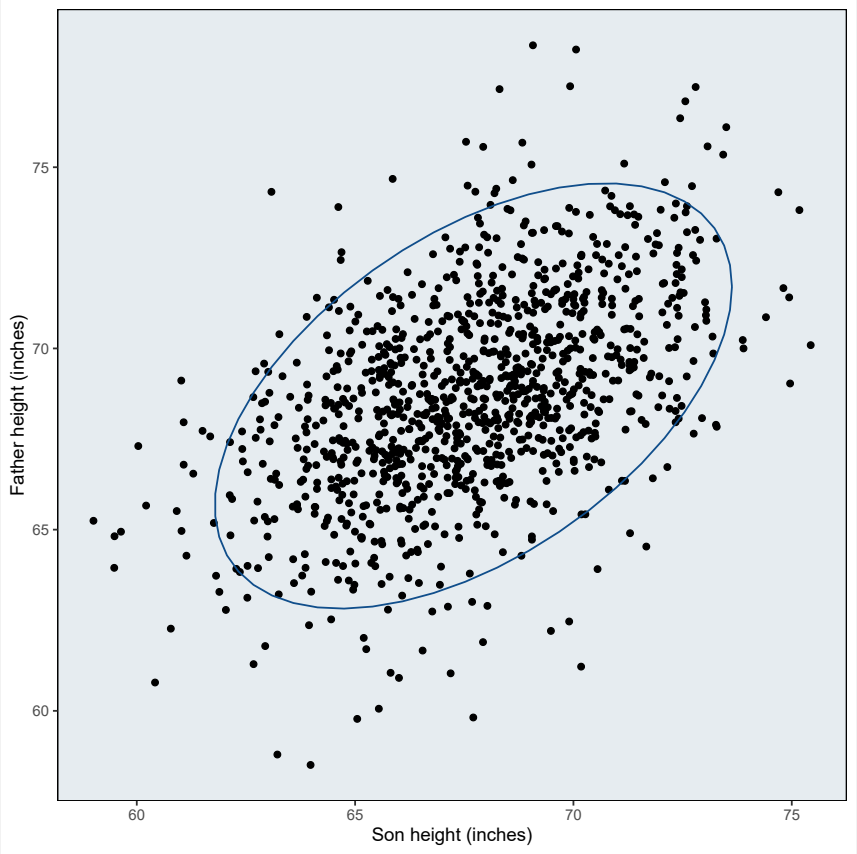


Figure 2

Probability density function of multivariate normal

Let's now define the probability density function for a multivariate normal distribution.

Definition 1

Multivariate normal p.d.f.

Suppose that the random variable $\boldsymbol{X}$ can take any real value and that $\boldsymbol{X}$ has the p.d.f.

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)$$

for all $\boldsymbol{x} \in \mathbb{R}^p$, then $\boldsymbol{X}$ is said to have a **multivariate normal** distribution, with mean $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and (co)variance matrix $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, written

$$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Note:

- $|\boldsymbol{\Sigma}|$ corresponds to the **determinant** of $\boldsymbol{\Sigma}$ and
- $\boldsymbol{\Sigma}^{-1}$ refers to the **inverse** of $\boldsymbol{\Sigma}$.

Example 4

Suppose that $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix}$$

What is the p.d.f. of $\boldsymbol{X}$?

Answer:

First, $|\boldsymbol{\Sigma}| = 5 \times 4 - (2) \times (2) = 16$ and

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{16} \begin{pmatrix} 4 & -2 \\ -2 & 5 \end{pmatrix}.$$

Then

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{X}) &= (2\pi)^{-1}(16)^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2} \begin{pmatrix} x_1 - 1 & x_2 - 2 \end{pmatrix} \tfrac{1}{16} \begin{pmatrix} 4 & -2 \\ -2 & 5 \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 2 \end{pmatrix}\right] \\ &= \tfrac{1}{8\pi} \exp\left[-\tfrac{1}{32}(4[x_1 - 1]^2 + 5[x_2 - 2]^2 - 4[x_1 - 1][x_2 - 2])\right] \\ &= \tfrac{1}{8\pi} \exp\left[-\tfrac{1}{32}(4x_1^2 + 5x_2^2 - 16x_2 - 4x_1x_2 + 16)\right]. \end{aligned}$$

Linear functions

In Week 6 we have seen that if X has a univariate normal distribution, $X \sim \mathrm{N}(\mu, \sigma^2)$ then $aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$. A similar property holds for the multivariate normal distribution.

Proposition 1

Linear functions of MVN

Suppose $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} \sim \mathrm{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\top}).$$

In other words, all this means is that linear functions of a normally distributed random vector are again normally distributed, just like in the univariate case.

Again just like in the univariate case, we can standardise a multivariate normal distribution.

If we choose $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ to be the inverse matrix square root of $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}\top}$ then if $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we can standardise $\mathbf{X}$ as follows,

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathsf{N}(\mathbf{0}, \mathbf{I}).$$

Task 1

Suppose that the continuous random vector

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \right).$$

Identify the distributions of:

(i) $X_1 + X_2$,

(ii) $X_1 - X_2$,

(iii) $X_2 - X_1$,

(iv) $3X_1 - 2X_2 + 1$.

Marginal distributions

The marginal distribution of a subset of variables in a MVN can be found by simply taking the relevant subsets of means, and the relevant subset of the covariance matrix for the variables you are interested in.

Proposition 2

Marginal distributions for bivariate normal

Let X_1 and X_2 be bivariate normal random variables, and suppose

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right).$$

Then

$X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$,

$X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$.

An important consequence of this property is that the marginal distribution of every single variable of a multivariate normal random vector is again normal.

Example 5

In Example 1, calculate the marginal distribution of X and Y.

Answer:

We know

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 67.7 \\ 68.7 \end{bmatrix}, \begin{bmatrix} 2.74^2 & 3.85 \\ 3.85 & 2.81^2 \end{bmatrix} \right).$$

Therefore

$$X_1 \sim \mathrm{N}(67.7, 2.74^2) \quad \text{and} \quad X_2 \sim \mathrm{N}(68.7, 2.81^2).$$

If we plot these variables separately we can see that both variables have the typical bell-shaped curve as we would expect for data which follows a normal distribution.

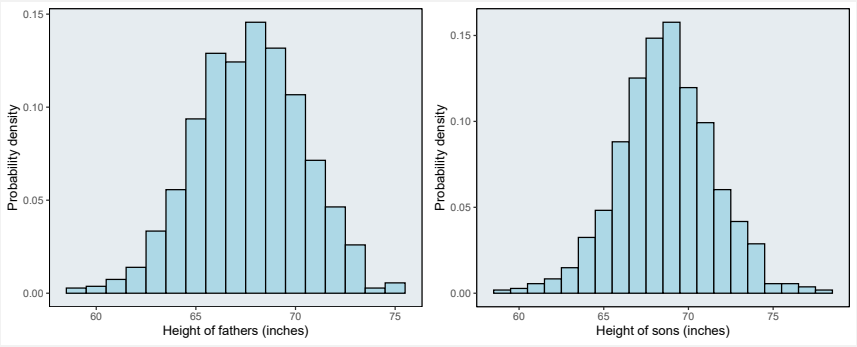


Figure 3

Task 2

Suppose that the continuous random vector

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \right).$$

COVARIANCE

1. Identify the marginal distributions of X_1 and X_2.
2. Find

(i) $\mathrm{E}(X_1)$ and $\mathrm{Var}(X_1)$,

(ii) $\mathrm{E}(X_2)$ and $\mathrm{Var}(X_2)$,

(iii) $\mathrm{Cov}(X_1, X_2)$ and $\rho(X_1, X_2)$.

Conditional distributions

Another important property of the MVN distribution is that if X_1 and X_2 have a multivariate normal distribution, then the conditional distribution of X_1 given that $X_2 = \boldsymbol{x}_2$ also has a normal distribution.

Proposition 3

Conditional distributions for bivariate normal

Let X_1 and X_2 be bivariate normal random variables, and suppose

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right).$$

Then

$$X_1 | X_2 = x_2 \sim \mathrm{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{12}(x_2 - \mu_2), \quad (1-\rho_{12}^2)\sigma_1^2 \right).$$

and

$$X_2 | X_1 = x_1 \sim \mathrm{N}\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{12}(x_1 - \mu_1), \quad (1-\rho_{12}^2)\sigma_2^2 \right).$$

Let's take a moment to try and understand what is going on here, focusing on the conditional probability of $X_1 | X_2 = x_2$.

- the conditional mean is equal to the mean of X_1 (μ_1) plus a constant which will be positive if the value observed for x_2 is larger than the mean for x_2, or negative if the value observed for x_2 is smaller than the mean for x_2 (assuming ρ_{12} is positive, if ρ_{12} is negative the opposite is true).
- the conditional variance ($(1-\rho_{12}^2)\sigma_1^2$) is smaller than the marginal variance (σ_1^2), and gets smaller as the correlation increases.

Example 6

In Example 1, calculate the conditional distribution of fathers' heights given that a son's height is equal to 65 inches.

Answer:

We know

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 67.7 \\ 68.7 \end{bmatrix}, \begin{bmatrix} 2.74^2 & 3.85 \\ 3.85 & 2.81^2 \end{bmatrix} \right).$$

We want the conditional distribution of $X_1|X_2 = 65$. Looking at Proposition Proposition 3 we can pull out all of the relevant pieces of information we need to calculate the conditional mean and variance.

$$\mu_1 = 67.7, \quad \mu_2 = 68.7, \quad \sigma_1 = 2.74, \quad \sigma_2 = 2.81, \quad \rho_{12} = 0.50.$$

We can then substitute these values into the formulae from Proposition Proposition 3 to get

$$\begin{aligned} \mathrm{E}(X_1|X_2 = 65) &= \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{12}(x_2 - \mu_2) \\ &= 67.7 + \frac{2.74}{2.81} \cdot 0.50 \cdot (65 - 68.7) \\ &= 65.90 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(X_1|X_2 = 65) &= (1 - \rho_{12}^2)\sigma_1^2 \\ &= (1 - 0.50^2) \cdot 2.74^2 \\ &= 5.63 \end{aligned}$$

Therefore, $X_1|X_2 = 65 \sim \mathrm{N}(65.90, 5.63)$

So the conditional mean (65.90) is smaller than the mean for fathers ($\mu_1 = 67.7$) since we know that the son's height is smaller than the mean height for sons ($\mu_2 = 68.7$). We can also see that the conditional variance (5.63) is smaller than the marginal variance for fathers ($2.74^2 = 7.51$).

This video discusses the bivariate normal distribution using Example 1. Apologies for the poor sound quality.

Video

The bivariate normal distribution - a worked example

Duration 12:32

Task 3

Suppose that the continuous random vector

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \right).$$

Identify the conditional distribution of X_1 given $X_2 = x_2$.

Supplement 2

Properties of the multivariate normal

The results above for the marginal and conditional distributions are for the bivariate case. These results can be generalised to the multivariate normal as shown below.

Marginal distributions

Let the random vector X be split into two blocks, X_1 and X_2, and suppose

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

Then

$\boldsymbol{X}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$,

$\boldsymbol{X}_2 \sim \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

Conditional distributions

Let the random vector $\boldsymbol{X}$ be split into two blocks, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, and suppose

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

Then

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim \mathrm{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^{\top} \right).$$

and

$$\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1 \sim \mathrm{N}\left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right).$$

Notice that

- the conditional mean is linear in x, it passes through the mean (μ_1, μ_2), and has a steeper slope with higher correlation.
- the conditional variance is smaller than the marginal variance, and gets smaller as the correlation increases.

Independence

We have seen that uncorrelated random variables are not necessarily independent: their relationship might be entirely non-linear.

The multivariate normal distribution is an exception to this. For the multivariate normal distribution absence of correlation and independence are one and the same thing. The reason for this is that the multivariate normal distribution only allows for linear dependency between its components, as we have seen when we have looked at the conditional distributions.

Large sample theory

In probability, we study limits to understand the long-term behaviour of random processes and sequences of random variables. In general, a **limit** tells us the value that a function approaches as that function's inputs get closer and closer to some number (often infinity). This may not, on the face of it, seem particularly useful. However, studying limits can often lead to simplified formulas for otherwise unsolvable probability models, which can lead to insights into complex problems.

In Week 1, we discussed the concept of relative frequency when interpreting a probability, which is an intuitive way of interpreting a probability as simply the frequency with which that outcome occurs in the long run, when the experiment is repeated a large number of times. This idea is illustrated in the example below.

Example 7

Real-world example

John Kerrich's famous experiment

Whilst visiting relatives in Copenhagen in 1940, John Kerrich, a British mathematician, was caught up in the Nazi invasion and interned in a prisoner of war camp. During his time in

the camp, Kerrich conducted an experiment tossing a coin 10,000 times and recording the number of heads obtained. The following graph shows the proportion of heads for 0 - 2000 tosses using the data recorded by Kerrich.

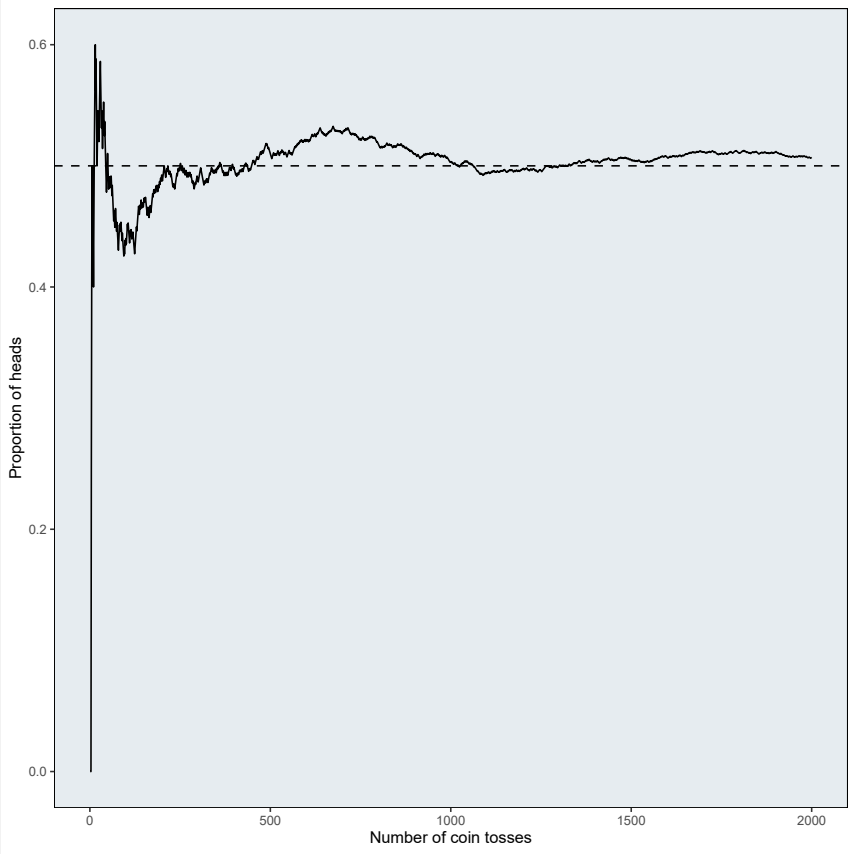


Figure 4

The figure shows wide fluctuations in the proportion of heads at the beginning of the experiment which eventually settle down close to the proportion we would expect of 0.5.

This example illustrates the **Law of Large Numbers**, which justifies the use of simulation to approximate the probability $\mathrm{P}(A)$ of an event A occurring. A consequence of the Law of Large Numbers is that in repeated trials of a random experiment the proportion of trials in which A occurs converges to $\mathrm{P}(A)$.

Let $X_1, X_2, \ldots$ be an independent and identically distributed sequence of random variables with finite expectation μ. For $n = 1, 2, \ldots$, let

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

Then the Law of Large Numbers says that the average, or the cumulative mean, $\bar{X}_n = S_n/n$, converges to μ, as $n \to \infty$.

Example 8

In the Kerrich experiment, A is the event

$$A = \{\text{coin tossed is a head}\}$$

and using what we have learned so far we can identify this experiment as a series of Bernoulli distributions where the probability of tossing a head is equal to $\frac{1}{2}$.

$$A \sim \text{Ber}\left(\frac{1}{2}\right)$$

Therefore, $\mu = \text{E}(A) = \frac{1}{2}$.

Let

$$X_k = \begin{cases} 1, & \text{if } A \text{ occurs on the } k\text{th experiment} \\ 0, & \text{otherwise.} \end{cases}$$

From the figure we can see that as the number of trials increases, the average $\bar{X} = S_n/n = (X_1 + \cdots + X_n)/n$, which is just proportion of trials in which A occurs tends towards $\text{P}(A) = \mu = 1/2$.

That is, the proportion of n trials in which A (heads) occurs converges to $P(A)$ as $n \to \infty$.

To discuss the Law of Large Numbers more formally, let's define what **convergence in probability** means.

Definition 2

Convergence in probability

Let $X_1, X_2, \ldots$ be a sequence of random variables defined on a sample space S. The sequence X_n is said to **converge in probability** to a constant μ if, for every $\epsilon > 0$,

$$\text{P}(|X_n - \mu| < \epsilon) \to 1 \quad \text{as } n \to \infty.$$

For most of the results we will study in this chapter, they key quantity of interest for us will be the cumulative mean, i.e. we study the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

as the sample size n increases to infinity, i.e. $n \to \infty$.

The figures below illustrate this definition. It shows the distribution of $\bar{X}_n$ for $n = 1$, $n = 10$ and $n = 100$. We are considering the probability that $\bar{X}_n$ is within a "tube" of width 2ϵ around μ. We can see that as we increase n the higher becomes the probability that $\bar{X}_n$ is in the interval $(\mu - \epsilon, \mu + \epsilon)$. If we were to keep increasing n then this probability would become 1, as mandated by the definition. This can be seen by noting that in the final figure, the whole distribution is contained within an ϵ distance from the actual mean μ.

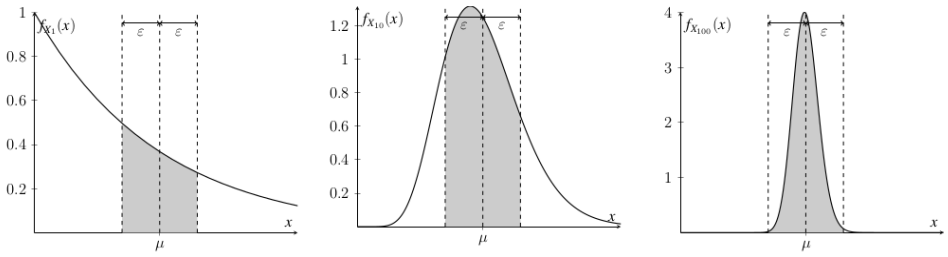


Figure 5

Example 9

In Example 7, we were interested in the average $\bar{X}_n = S_n/n = (X_1 + \cdots + X_n)/n$, which was just the proportion of trials in which the tossed coin resulted in heads, as the number of tosses increased.

Let's now define the Weak Law of Large Numbers which shows that the sample mean of an independent sample drawn from any arbitrary distribution (as long as this distribution does not have too heavy tails) of size n is increasingly concentrated around its mean.

Theorem 1

The weak law of large numbers

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables, each with finite expected value μ. For $n = 1, 2, \ldots$, let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then for any $\epsilon > 0$,

$$\mathrm{P}(|\bar{X}_n - \mu| < \epsilon) \to 1 \qquad \text{as } n \to \infty.$$

In other words, the probability that the absolute value of the difference between the sample mean, $\bar{X}_n$ and the expected value μ is less than some very small number ϵ tends towards 1 as n goes to infinity. The proof of this theorem is beyond the scope of this course but is provided as supplementary material.

Supplement 3

A proof of the weak law

One proof requires a result called **Chebyshev's inequality** (although it only works when the variance exists too).

A key part of the proof of the Weak Law of Large Numbers is the so-called **Chebyshev's inequality**.

Let X be a random variable with *finite* expected value μ. If c is a real constant such that $\mathrm{E}[(X-c)^2]$ is finite, then for any value $\epsilon > 0$

$$\mathrm{P}(|X - c| < \epsilon) \geq 1 - \frac{1}{\epsilon^2}\mathrm{E}[(X-c)^2].$$

In particular, if X has *finite* variance, $\sigma^2 = \mathrm{E}[(X-\mu)^2]$, then for any value $\epsilon > 0$

$$\mathrm{P}(|X - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{\epsilon^2}.$$

This result is easily proved when X is a continuous random variable with probability density function $f_X(x)$. In this case,

$$
\begin{aligned}
\mathrm{E}[(X-c)^2] &= \int_{-\infty}^{\infty} (x-c)^2 f_X(x)\,dx \\
&\geq \int_{-\infty}^{c-\epsilon} (x-c)^2 f_X(x)\,dx + \int_{c+\epsilon}^{\infty} (x-c)^2 f_X(x)\,dx \\
&\geq \int_{-\infty}^{c-\epsilon} \epsilon^2 f_X(x)\,dx + \int_{c+\epsilon}^{\infty} \epsilon^2 f_X(x)\,dx \\
&= \epsilon^2 \left[\int_{-\infty}^{c-\epsilon} f_X(x)\,dx + \int_{c+\epsilon}^{\infty} f_X(x)\,dx \right] \\
&= \epsilon^2 \left[1 - \mathrm{P}(|X-c| < \epsilon) \right] \\
\implies \mathrm{P}(|X-c| < \epsilon) &\geq 1 - \frac{1}{\epsilon^2}\mathrm{E}[(X-c)^2].
\end{aligned}
$$

The first line is a definition. The second line removes a non-negative contribution to the integral from $c-\epsilon$ to $c+\epsilon$. The third line replaces $x-c$ by its smallest value and $x+c$ by its smallest value. The fourth line tidies up. The fifth line identifies the integrals with the probability of a particular event and the last line rearranges.

The second part of the theorem follows immediately from the first by putting $c = \mu$.

We can now prove the Weak Law of Large Numbers. We will prove this result for the simple case where the random variables have *finite* variance σ^2; however this assumption is not required for the weak law to hold.

For all $n \geq 1$, $\bar{X}_n$ has expected value μ and variance σ^2/n . By Chebyshev's inequality, for any $\epsilon > 0$,

$$
\mathrm{P}(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2/n}{\epsilon^2} \to 1 \quad \text{as } n \to \infty.
$$

Informally, the laws of large numbers (there is also a Strong Law too, which involves another form of stochastic convergence, which we will say no more about) tell us that the probability distribution of $\bar{X}_n$ becomes more and more concentrated at its expected value μ as $n \to \infty$. Although interesting and important, this does not help us to calculate probabilities of interest associated with $\bar{X}_n$ since it does not tell us how close $\bar{X}_n$ is to μ for a given value of n. The **central limit theorem** (CLT) provides a means of doing this, at least approximately.

The central limit theorem

Theorem 2

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, each with a finite mean μ and a finite variance σ^2. Then for sufficiently large n we

have that

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \to \mathrm{N}(0,1),$$

in the sense that the cdf of the left-hand side tends to the cdf of the standard normal distribution.

The central limit theorem is often used in one of the following two equivalent forms, which can be obtained by re-arranging the terms.

1. $\sum_{i=1}^{n} X_i$ approximately follows the $\mathrm{N}(n\mu, n\sigma^2)$ distribution for 'sufficiently large' n.
2. $\bar{X}_n$ approximately follows the $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ distribution for 'sufficiently large' n.

Let's focus on 2. for now. Whatever the value of n, the rules for the expected value and the variance of a linear functions tell us that $E(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$. What the central limit theorem tells us is that the *shape* of the distribution of $\bar{X}_n$ tends to the normal distribution. The useful and counter-intuitive thing about the central limit theorem is that this happens no matter what the shape of the original distribution is (unless it has too heavy tails and no finite variance). For most distributions, a normal distribution is approached very quickly as n increases.

Task 4

Verify that $E(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$.

The simplest way to illustrate the central limit theorem is using a graphical example.

Example 10

Exponential to normal

Suppose the random variable $X \sim \mathrm{Exp}(1.5)$. The figure below shows a sample of 100 points from this distribution. You can see very clearly from this plot that this is a highly skewed distribution and therefore non-normal.

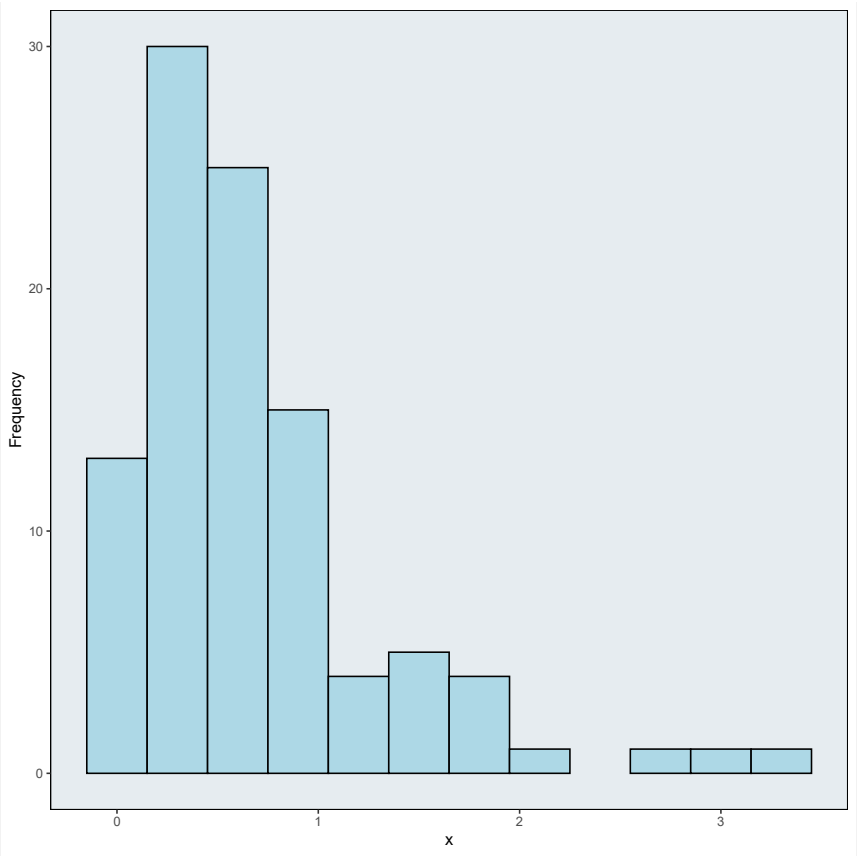


Figure 6

A further simulation was carried out from the same model. This time, a sample of $n = 5$ simulated values was obtained and the sample mean calculated. This was done 1,000 times and the sample means are displayed in histogram (i) below. Though skewed, the distribution of sample means is a lot less skewed than that of the original data. The remaining histograms repeat the simulation for even larger sample sizes, (ii) $n = 10$, (iii) $n = 25$ and (iv) $n = 100$. Clearly as n increases, the distribution of the sample means become more symmetric and looks more and more like the bell-shaped curve of the normal distribution. It can also be seen that as n increases the spread of the distribution decreases (note that the scale on the horizontal axes differs between these plots).

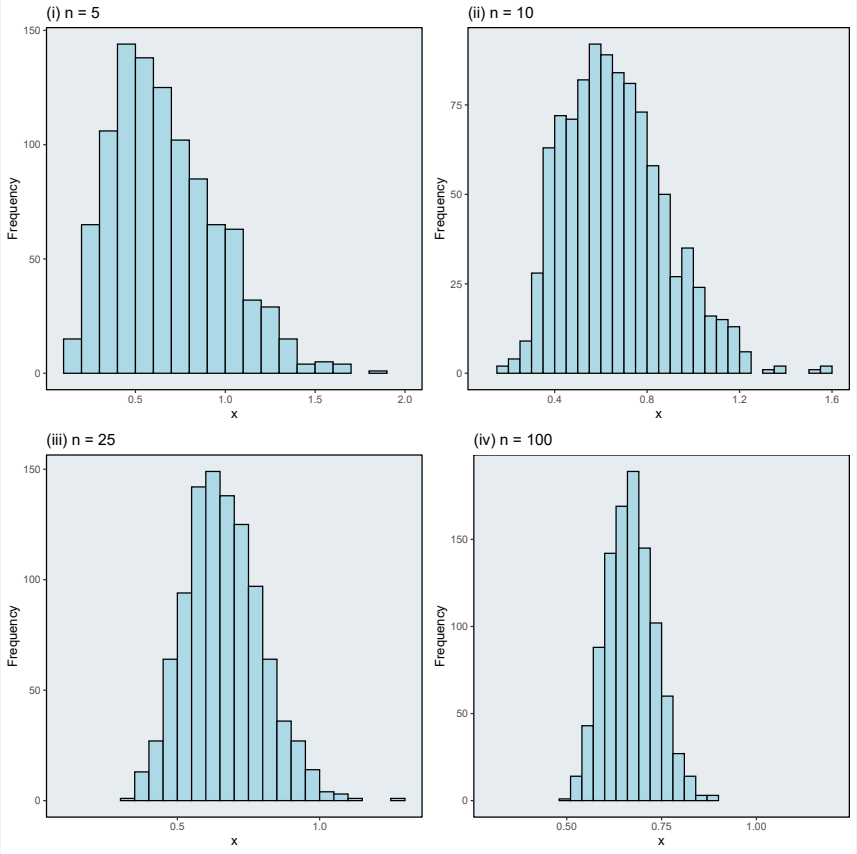


Figure 7

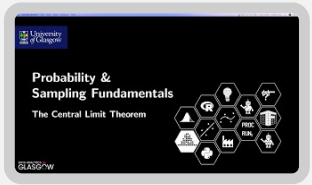
Take a look at this Shiny App to visualise how the distribution of sample means approaches a normal distribution as n increases for a number of different distributions.

This video introduces the central limit theorem by discussing the example above.

Video

The central limit theorem

Duration 2:55



Example 11

Small change

It is usual for even very large financial transactions, to the value of hundreds of thousands of pounds, to be settled to fractions of pence. Suppose, instead, that financial institutions agreed to round all settlements of transactions between them to the nearest whole £1. In one year, a certain institution makes 1500 transactions. What is the probability that this institution will lose more than £5 over the course of the year?

To answer this question, let's begin by defining X_i to be the difference in cost in £ between the computed cost of the i-th transaction and its true cost $(i = 1, 2, \ldots, 1500)$. For each transaction the most that an institution can lose is 50 pence (or £0.5) and the amount that they can gain is also 50 pence (or £0.5), since a transaction will either be rounded up to the nearest pound, or rounded down to the nearest pound. We can therefore assume that $X_i \sim \text{Un}(-0.5, 0.5)$.

We are not interested in the amount lost for a single transaction, but rather the total amount lost in the year over all 1500 transactions. The difference between the total cost of the 1500 transactions and the computed cost is

$$S_{1500} = X_1 + \ldots + X_{1500}.$$

Although we know the distribution for each X_i, we don't know the distribution of S_{1500}. We can however estimate this using the central limit theorem.

Using the results from Week 6 we can calculate $\mu = \text{E}(X_i) = \frac{-0.5+0.5}{2} = 0$ and $\sigma^2 = \text{Var}(X_i) = \frac{0.5-(-0.5)}{12} = \frac{1}{12}$.

Then, using the central limit theorem

$$S_n = \sum_{i=1}^{n} X_i \overset{\text{approx}}{\sim} \text{N}\left(n\mu, n\sigma^2\right).$$

$$S_n \overset{\text{approx}}{\sim} \text{N}\left(0, \frac{1500}{12}\right).$$

We can then follow the same process as in Week 6 to find the probability that the institution loses at least £5 in total is

$$\begin{aligned} P(S_{1500} < -5) &= 1 - P(S_{1500} < 5) \\ &= 1 - P\left(\frac{S_{1500} - 0}{\sqrt{1500/12}} < \frac{5 - 0}{\sqrt{1500/12}}\right) \\ &= 1 - P(Z < 0.45) \\ &= 1 - \Phi(0.45) \\ &= 1 - 0.674 \\ &= 0.326 \end{aligned}$$

Task 5

The life-time of video projector light bulbs are known to follow an exponential distribution with a mean life-time of $\frac{1}{\lambda} = 90$. The university uses projectors for 8500 hours per semester. What is the probability that 100 light bulbs will be sufficient for the semester?

Normal approximation to the binomial

Consider the Binomial distribution $X \sim \text{Bi}(1000, 0.4)$ and suppose you wish to calculate $\text{P}(X > 661)$. The shortest way to calculate this is:

$$\text{P}(X > 661) = \text{P}(X = 662) + \text{P}(X = 663) + \ldots + \text{P}(X = 1000),$$

which involves 340 separate calculations! The central limit theorem allows us to make an approximation using a normal distribution.

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identical $\text{Bern}(\theta)$ random variables. From Week 3 we know that

- $\text{E}(X_i) = \theta$ and $\text{Var}(X_i) = \theta(1-\theta)$.
- The sum of these variables $X = \sum_{i=1}^{n} X_i \sim \text{Bi}(n, \theta)$.

The central limit theorem tells us that $X = \sum_{i=1}^{n} X_i \approx \text{N}(n\theta,\ n\theta(1-\theta))$, which in turn means that

$$\text{Bi}(n, \theta) \approx \text{N}(n\theta,\ n\theta(1-\theta)),$$

providing n is large enough and θ is not too close to zero or one. Therefore, to calculate $\text{P}(X > 661)$ we simply approximate the binomial distribution with a normal and use the normal tables to calculate the probability.

Continuity correction

However in moving from a discrete binomial distribution to a continuous normal approximation we encounter the following problem.

Let $X \sim \text{Bi}(100, 0.3)$, and consider calculating

1. $\text{P}(X < 50)$.
2. $\text{P}(X \leq 50)$.

As the binomial is a discrete distribution these two probabilities are different. However, if we apply the central limit theorem and approximate $X \sim \text{Bi}(100, 0.3)$ with $X \sim \text{N}(100 \cdot 0.3, 100 \cdot 0.3 \cdot 0.7) = \text{N}(30, 21)$ (its normal approximation), we have a problem. This is because using the normal approximation and calculating $\text{P}(X < 50)$ and $\text{P}(X \leq 50)$ will give the same probability, because for a continuous distribution $\text{P}(X = 50) = 0$ (its a single outcome). However, as X is a discrete distribution $\text{P}(X = 50) > 0$.

Therefore each time we approximate a discrete distribution with a continuous one we make the following **continuity correction**. The correction works by adding or subtracting 0.5 to the outcome as follows:

- $\text{P}(X > x)$ is replaced with $\text{P}(X > x + 0.5)$.
- $\text{P}(X \geq x)$ is replaced with $\text{P}(X \geq x - 0.5)$.
- $\text{P}(X < x)$ is replaced with $\text{P}(X < x - 0.5)$.
- $\text{P}(X \leq x)$ is replaced with $\text{P}(X \leq x + 0.5)$.

Essentially if the probability to be calculated has a $<$ or $>$ sign, we need to add or subtract 0.5 to x so that the probability you calculate is smaller than it would have been. In contrast, if the probability to be calculated has a $\leq$ or $\geq$ sign, then we need to add or subtract 0.5 to x so that the probability you calculate is bigger than it would have been.

Example 12

Let $X \sim \text{Bi}(10000, 0.005)$. Using the normal distribution calculate $\text{P}(X < 70)$.

Answer:

As $X \sim \text{Bi}(10000, 0.005)$ then $X \overset{\text{approx}}{\sim} \text{N}(50, 49.75)$ so that $\sigma = \sqrt{49.75}$. Therefore

$$\begin{aligned} \text{P}(X < 70) &= \text{P}\left(\frac{X - 50}{\sqrt{49.5}} < \frac{69.5 - 50}{\sqrt{49.5}}\right) \\ &= \text{P}(Z < 2.77) \quad \text{where } Z \sim \text{N}(0, 1) \\ &= \Phi(2.77) = 0.9972. \end{aligned}$$

Task 6

Let $X \sim \text{Bi}(100, 0.4)$. Using the normal distribution calculate

(a) $\text{P}(X < 51)$,

(b) $\text{P}(X \geq 33)$,

(c) $\text{P}(35 \leq X \leq 41)$,

(d) $\text{P}(X = 38)$.

Normal approximation to the Poisson

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identical Pois(1) random variables. From Week 3 we know that

$$\text{E}(X_i) = 1 \quad \text{and} \quad \text{Var}(X_i) = 1.$$

An important property of the Poisson distribution is that the sum of independent Poisson random variables has a Poisson distribution:

$$\text{If } X \sim \text{Pois}(\lambda) \text{ and } Y \sim \text{Pois}(\mu) \text{ then } X + Y \sim \text{Pois}(\lambda + \mu).$$

So if $X \sim \text{Pois}(1)$ and $Y \sim \text{Pois}(1)$ then $X + Y \sim \text{Pois}(2)$.

Therefore $X = \sum_{i=1}^{n} X_i \sim \text{Pois}(n)$.

Now the central limit theorem tells us that $X = \sum_{i=1}^{n} X_i \approx \text{N}(n,\ n)$, which in turn means

$$\text{Po}(n) \approx \text{N}(n,\ n),$$

providing n is large enough. As with the binomial approximation we have to do a continuity correction as we are moving from a discrete to a continuous distribution.

Example 13

Let $X \sim \text{Pois}(50)$ and calculate

(a) $\text{P}(X < 60)$,

(b) $\mathrm{P}(50 \leq X < 60)$.

Answer: As $X \sim \mathrm{Pois}(50)$ then $X \overset{\text{approx}}{\sim} \mathrm{N}(50, 50)$ so that $\sigma = \sqrt{50}$.

(a)

$$\begin{aligned}\mathrm{P}(X < 60) &= \mathrm{P}\left(\frac{X - 50}{\sqrt{50}} < \frac{59.5 - 50}{\sqrt{50}}\right) \\ &= \mathrm{P}(Z < 1.34) = 0.9099.\end{aligned}$$

(b)

$$\begin{aligned}\mathrm{P}(50 \leq X < 60) &= \mathrm{P}\left(\frac{49.5 - 50}{\sqrt{50}} \leq \frac{X - 50}{\sqrt{50}} < \frac{59.5 - 50}{\sqrt{50}}\right) \\ &= \mathrm{P}(-0.07 \leq Z < 1.34) \\ &= \Phi(1.34) - \Phi(-0.07) = 0.9099 - (1 - 0.5279) = 0.4378.\end{aligned}$$

Learning outcomes for week 8

By the end of week 8, you should be able to:

- calculate linear functions of the multivariate normal distribution;
- calculate marginal and conditional distributions of the multivariate normal (for the bivariate case);
- state and use the central limit theorem;
- calculate normal approximations to the binomial and Poisson distributions.

A summary of the most important concepts and written answers to all tasks are provided overleaf.

Week 8 summary

The multivariate normal distribution

Linear functions of MVN

Suppose $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b} \sim \mathrm{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\top}).$$

Probability density function

Suppose that the random variable $\boldsymbol{X}$ can take any real value and that $\boldsymbol{X}$ has the p.d.f.

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)$$

for all $\boldsymbol{x} \in \mathbb{R}^p$, then $\boldsymbol{X}$ is said to have a **multivariate normal** distribution, with mean $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and (co)variance matrix $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, written

$$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Marginal distributions

Let X_1 and X_2 be bivariate normal random variables, and suppose

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right).$$

Then

$X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$,

$X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$.

Conditional distributions

Let X_1 and X_2 be bivariate normal random variables, and suppose

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right).$$

Then

$$X_1|X_2 = x_2 \sim \mathrm{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{12}(x_2 - \mu_2), \quad (1-\rho_{12}^2)\sigma_1^2\right).$$

and

$$X_2|X_1 = x_1 \sim \mathrm{N}\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho_{12}(x_1 - \mu_1), \quad (1-\rho_{12}^2)\sigma_2^2\right).$$

Large sample theory

The weak law of large numbers

Let X_1, $X_2, \ldots$ be a sequence of i.i.d. random variables, each with finite expected value μ. For $n = 1, 2, \ldots$, let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then for any $\epsilon > 0$,

$$\mathrm{P}(|\bar{X}_n - \mu| < \epsilon) \to 1 \qquad \text{as } n \to \infty.$$

The central limit theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, each with a finite mean μ and a finite variance σ^2. Then for sufficiently large n we have that

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \to \mathrm{N}(0,1),$$

in the sense that the cdf of the left-hand side tends to the cdf of the standard normal distribution.

The central limit theorem is often used in one of the following two equivalent forms:

1. $\sum_{i=1}^{n} X_i$ approximately follows the $\mathrm{N}(n\mu, n\sigma^2)$ distribution for 'sufficiently large' n.
2. $\bar{X}$ approximately follows the $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ distribution for 'sufficiently large' n.

Answer 1

(i) $X_1 + X_2 = \begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}X_1 \\ X_2\end{bmatrix}$

$$
\begin{aligned}
\mathrm{E}(X_1 + X_2) &= \begin{bmatrix} 1 & 1 \end{bmatrix} \boldsymbol{\mu} \\
&= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(X_1 + X_2) &= \begin{bmatrix} 1 & 1 \end{bmatrix} \boldsymbol{\Sigma} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
&= 1
\end{aligned}
$$

Using **Proposition 1**, then, $X_1 + X_2 \sim \mathrm{N}(0, 1)$.

(ii) $X_1 - X_2 = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$

$$
\begin{aligned}
\mathrm{E}(X_1 - X_2) &= \begin{bmatrix} 1 & -1 \end{bmatrix} \boldsymbol{\mu} \\
&= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(X_1 - X_2) &= \begin{bmatrix} 1 & -1 \end{bmatrix} \boldsymbol{\Sigma} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\
&= 3
\end{aligned}
$$

Using **Proposition 1**, then, $X_1 - X_2 \sim \mathrm{N}(0, 3)$.

(iii) $X_2 - X_1 = \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$

$$
\begin{aligned}
\mathrm{E}(X_2 - X_1) &= \begin{bmatrix} -1 & 1 \end{bmatrix} \boldsymbol{\mu} \\
&= \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(X_2 - X_1) &= \begin{bmatrix} -1 & 1 \end{bmatrix} \boldsymbol{\Sigma} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\
&= 3
\end{aligned}
$$

Using **Proposition 1**, then, $X_2 - X_1 \sim \mathrm{N}(0, 3)$.

(iv) $3X_1 - 2X_2 + 1 = \begin{bmatrix} 3 & -2 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + 1$

$$\begin{aligned} \mathrm{E}(3X_1 - 2X_2 + 1) &= \begin{bmatrix} 3 & -2 \end{bmatrix} \boldsymbol{\mu} + 1 \\ &= \begin{bmatrix} 3 & -2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + 1 \\ &= 1 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(3X_1 - 2X_2 + 1) &= \begin{bmatrix} 3 & -2 \end{bmatrix} \boldsymbol{\Sigma} \begin{bmatrix} 3 \\ -2 \end{bmatrix} \\ &= \begin{bmatrix} 3 & -2 \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \end{bmatrix} \\ &= 19 \end{aligned}$$

Using Proposition 1, then, $3X_1 - 2X_2 + 1 \sim \mathrm{N}(1, 19)$.

Answer 2

1. $X_1 \sim \mathrm{N}(0,1)$ and $X_2 \sim \mathrm{N}(0,1)$
2.

(i) $\mathrm{E}(X_1) = 0, \quad \mathrm{Var}(X_1) = 1,$

(ii) $\mathrm{E}(X_2) = 0, \quad \mathrm{Var}(X_2) = 1,$

(iii) $\mathrm{Cov}(X_1, X_2) = -\frac{1}{2}$ and

$\rho(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1) \cdot \mathrm{Var}(X_2)}} = \frac{-\frac{1}{2}}{\sqrt{1 \cdot 1}} = -\frac{1}{2}.$

Answer 3

We have

$$\mu_1 = 0, \quad \mu_2 = 0, \quad \sigma_1 = 1, \quad \sigma_2 = 1, \quad \rho_{12} = -\frac{1}{2}.$$

We can then substitute these values into the formulae from Proposition Proposition 3 to get

$$
\begin{aligned}
\mathrm{E}(X_1|X_2 = x_2) &= \mu_1 + \frac{\sigma_1}{\sigma_2}\rho_{12}(x_2 - \mu_2) \\
&= 0 + 1 \cdot (-\frac{1}{2}) \cdot (x_2 - 0) \\
&= -\frac{x_2}{2}
\end{aligned}
$$

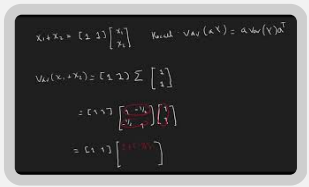
$$
\begin{aligned}
\mathrm{Var}(X_1|X_2 = x_2) &= (1 - \rho_{12}^2)\sigma_1^2 \\
&= (1 - \frac{1}{2}^2) \cdot 1 \\
&= \frac{3}{4}
\end{aligned}
$$

Here is a video worked solution for all of Task 1.

Video

Week 8 - Task 1

Duration 11:04



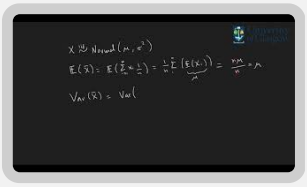
Answer 4

Using that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and that the X_i are independent,

$$
\begin{aligned}
E(\bar{X}_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{\underbrace{n}_{=n\mu}}\sum_{i=1}^{n}\underbrace{E(X_i)}_{=\mu} \\
&= \mu \\
\mathrm{Var}(\bar{X}_n) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{\underbrace{n^2}_{=n\sigma^2}}\sum_{i=1}^{n}\underbrace{\mathrm{Var}(X_i)}_{=\sigma^2} \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Here is a video worked solution.

Video



Week 8 - Task 2

Duration 1:26

Answer 5

Let X_i be the life time of the i-th light bulb.

then $S_{100} = X_1 + \cdots X_{100}$.

Since the life-time of a light bulb is exponentially distributed, the mean and standard deviation of individual life-time is $\mu = \sigma = 90$.

Then, using the central limit theorem

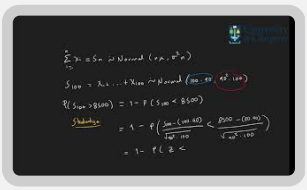
$$S_n \overset{\text{approx}}{\sim} \mathrm{N}\left(100 \cdot 90, 100 \cdot 90^2\right).$$

We are looking for

$$\begin{aligned} P(S_{100} > 8500) &= 1 - P(S_{100} < 8500) \\ &= 1 - P\left(\frac{S_{100} - 9000}{\sqrt{90^2 \cdot 100}} < \frac{8500 - 9000}{\sqrt{90^2 \cdot 100}}\right) \\ &= 1 - P(Z < -0.56) \\ &= 1 - (1 - P(Z < 0.56)) \\ &= 0.7123. \end{aligned}$$

Here is a video worked solution.

Video



Week 8 - Task 3

Duration 3:17

Answer 6

As $X \sim \text{Bi}(100, 0.4)$ then $X \overset{\text{approx}}{\sim} \text{N}(40, 24)$ so that $\sigma = \sqrt{24}$.

(a)

$$\begin{aligned} \text{P}(X < 51) &= \text{P}\left(\frac{X - 40}{\sqrt{24}} \leq \frac{50.5 - 40}{\sqrt{24}}\right) \\ &= \text{P}(Z \leq 2.14) \quad \text{where } Z \sim \text{N}(0, 1) \\ &= \Phi(2.14) = 0.9834. \end{aligned}$$

(b)

$$\begin{aligned} \text{P}(X \geq 33) &= \text{P}\left(\frac{X - 40}{\sqrt{24}} \geq \frac{32.5 - 40}{\sqrt{24}}\right) \\ &= \text{P}(Z \geq -1.53) \\ &= \text{P}(Z \leq 1.53) = 0.937. \end{aligned}$$

(c)

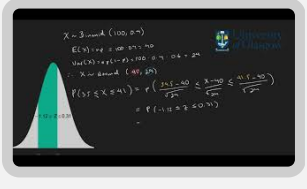
$$\begin{aligned} \text{P}(35 \leq X \leq 41) &= \text{P}\left(\frac{34.5 - 40}{\sqrt{24}} \leq \frac{X - 40}{\sqrt{24}} \leq \frac{41.5 - 40}{\sqrt{24}}\right) \\ &= \text{P}(-1.12 \leq Z \leq 0.31) \\ &= \Phi(0.31) - \Phi(-1.12) = 0.6217 - 0.1314 = 0.4903. \end{aligned}$$

(d)

$$\begin{aligned} \text{P}(X = 38) &= \text{P}\left(\frac{37.5 - 40}{\sqrt{24}} \leq \frac{X - 40}{\sqrt{24}} \leq \frac{38.5 - 40}{\sqrt{24}}\right) \\ &= \text{P}(-0.51 \leq Z \leq -0.31) \\ &= \Phi(-0.31) - \Phi(-0.51) = 0.3783 - 0.3050 = 0.0733. \end{aligned}$$

Here is a video worked solution.

Video



Week 8 - Task 4

Duration 6:59