Applied Probability II

Section 9: The Normal Distribution

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 9: The Normal Distribution

The normal or Gaussian distribution

The normal (or Gaussian) distribution has a central place in statistics, largely as a result of the central limit theorem.

In this Section we will examine various aspects of the normal distribution.

Section 9.1: Assessing normality

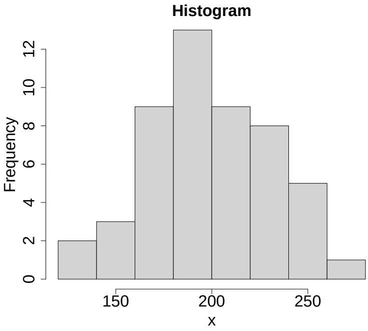
The univariate normal distribution

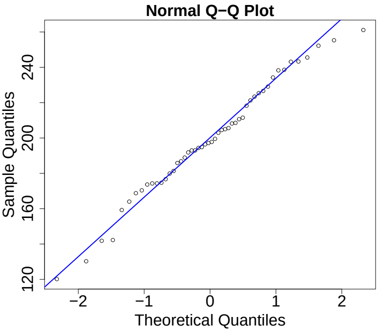
In some sections so far in this module, we have carried out statistical analyses or modelling where a normal distribution was assumed.

To validate such an assumption, we can assess the normality of the data for which the assumption is made. This can be done by observing a histogram of the data (which is an approximation of the probability density function), or we can use a quantile-quantile (QQ) plot, which is a little bit more formal, but also subjective.

We have already used QQ plots to assess normality in other Sections. In this sub-section, we will examine in more detail how to assess QQ plots.

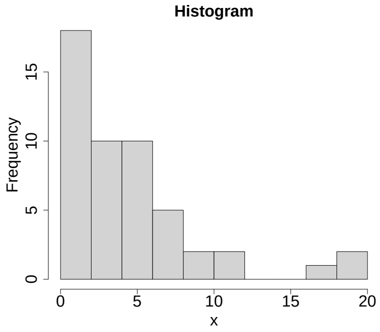
Example 1





- Normality seems to be a reasonable assumption

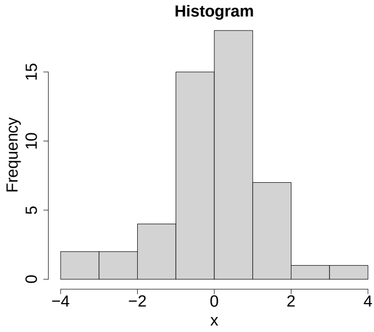
Example 2

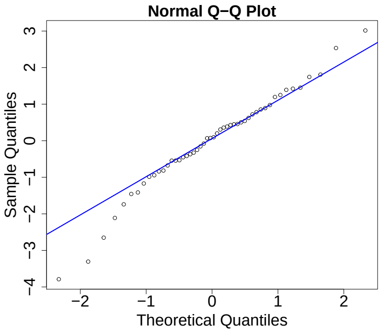


In the QQ plot:

- Short tail shown at lower end with points above the line (points would be below the line here if long tailed in this direction).
- Long tail shown at upper end with points above the line (points would be below the line here if short tailed in this direction).

Example 3





In the QQ plot:

- Long tail shown at lower end with points below the line (points would be above the line here if short tailed in this direction).
- Slight long tail shown at upper end with points above the line (points would be below the line here if short tailed in this direction).

Assessing normality summary

We have assessed the normality of a sample of data using the following methods

- Histogram: we observe if the histogram follows the bell shaped curve typical of the normal distribution.
- QQ plots: we plot the sample quantiles against the theoretical quantiles of the normal distribution and see if they follow a straight line.

There are also other plots that can be used (e.g., PP plots), and there are formal hypothesis tests that can assess normality (e.g., the Shapiro-Wilks test).

Section 9.2: The bivariate normal distribution

Joint discrete random variables

If X and Y are discrete random variables we can define their joint probability mass function (pmf) as:

$$p(x, y) = P(X = x, Y = y)$$

The marginal probability mass function of X is:

$$p_X(x) = P(X = x) = \sum_y P(X = x, Y = y)$$

Joint discrete random variables - Example

Toss a coin three times. Let $X =$ the number of heads on the first toss. Let $Y =$ the total number of heads. There are 8 equally likely outcomes:

S	X	Y
hhh	1	3
hht	1	2
hth	1	2
thh	0	2
htt	1	1
tht	0	1
tth	0	1
ttt	0	0

We can tabulate the joint pmf of X and Y:

		Y				
		0	1	2	3	
X	0	1/8	2/8	1/8	0	0.5
	1	0	1/8	2/8	1/8	0.5
		1/8	3/8	3/8	1/8	1

The entries in the final column and row give $p_x(x)$ and $p_x(y)$, the marginal probability functions of X and Y respectively.

Joint continous distributions

Let X and Y be continuous random variables with joint cumulative distribution function (cdf) $F(x, y)$. They are jointly continuous if there is a function $f(x, y) \geq 0$ such that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

f is called the joint probability density function (pdf) of X and Y.

The marginal cumulative distribution function (cdf) F_x of X can be obtained as:

$$F_x(a) = P(X \leq a, Y \leq \infty) = \int_{-\infty}^{a} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx$$

Thus, the marginal density f_x of X is

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

Similarly, the marginal density f_y of Y is

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

The bivariate normal distribution

The univariate normal distribution probability density function is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

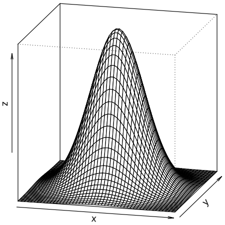
The bivariate normal density is given by:

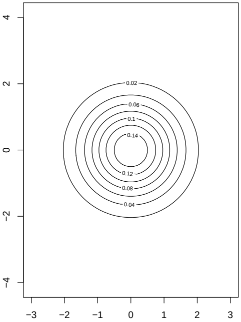
$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-\rho^2)}\right]$$

The bivariate normal distribution - Example

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\dfrac{(x-\mu_x)^2}{\sigma_x^2} - \dfrac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \dfrac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-\rho^2)} \right]$$

Normal with $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$ and $\rho = 0$:

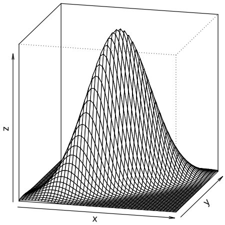


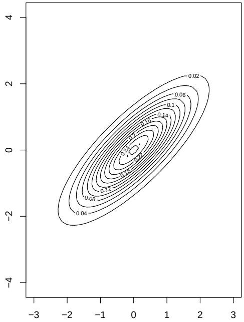


The bivariate normal distribution - Example

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-\rho^2)}\right]$$

Normal with $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$ and $\rho = .8$:

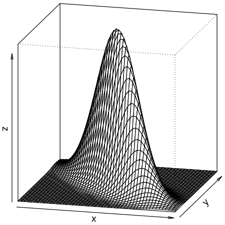


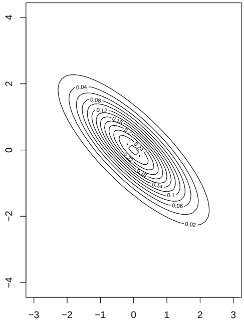


The bivariate normal distribution - Example

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\dfrac{(x-\mu_x)^2}{\sigma_x^2} - \dfrac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \dfrac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-\rho^2)} \right]$$

Normal with $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$ and $\rho = -.8$:





The marginal distributions of the bivariate normal

The joint pdf is:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-\rho^2)} \right]$$

The marginal distributions of X is:

$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$$

Which we can show is equal to:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{1}{2} \frac{(x-\mu_x)^2}{\sigma_x^2} \right] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{(y-b_x)^2}{2\sigma_Y^2(1-\rho^2)} \right] dy$$

where

$$b_x = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x-\mu_X)$$

The marginal distributions of the bivariate normal

When we examine:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{(y-b_x)^2}{2\sigma_Y^2(1-\rho^2)}\right] dy$$

We can see that the right hand side is the integral of a normal probability density function with parameters: $N(b_x, \sigma_Y^2(1-\rho^2))$, and thus integrates to 1.

We arrive at:

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right)$$

Therefore $X \sim N(\mu_x, \sigma_x^2)$.

It can be shown similarly that the marginal distribution of Y is $N(\mu_y, \sigma_y^2)$.

Independence

In general, if two random variables are independent, we know that the correlation between them (ρ) equals 0. But a zero correlation does not imply independence.

However, if X and Y follow a bivariate normal distribution, then X and Y are independent *if and only if* ρ equals 0.

Section 9.3: The multivariate normal distribution

The multivariate normal distribution

Consider a set of n independent and identically distributed (IID) standard normal random variables

$$Z_i \underset{IID}{\sim} N(0,1)$$

The covariance matrix for $\mathbf{Z}$ is $\mathbf{I_n}$ and $\mathrm{E}[\mathbf{Z}] = 0$.

Let $\mathbf{B}$ be an $m \times n$ matrix of fixed coefficients and $\boldsymbol{\mu}$ be an m-vector of fixed coefficients.

Then, the m-vector $\mathbf{X} = \mathbf{BZ} + \boldsymbol{\mu}$ is said to have a multivariate normal distribution.

The mean of $\mathbf{X}$ is: $E[\mathbf{X}] = \boldsymbol{\mu}$.

The covariance matrix of $\mathbf{X}$ is: $Var[\mathbf{X}] = \mathbf{BB^T} = \boldsymbol{\Sigma}$.

Or, we can say that

$$\mathbf{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The multivariate normal distribution pdf

The bivariate normal probability density function that we looked at in the last sub-section generalises to the multivariate case.

If

$$\mathbf{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}|}} exp \left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right)$$

for $\mathbf{x} \in \mathbb{R}^m$

If you study the module Multivariate Linear Analysis next year (and Data Analytics the following year), you will come across the MVN distribution again.