Chapter 5 – The normal distribution

5.1 Probability distributions of continuous random variables

A random variable *X* is called **continuous** if it can assume any of the possible values in some interval i.e. the number of possible values are infinite. In this case the definition of a discrete random variable (list of possible values with their corresponding probabilities) cannot be used (since there are an infinite number of possible values it is not possible to draw up a list of possible values). For this reason probabilities associated with individual values of a continuous random variable X are taken as 0.

The clustering pattern of the values of X over the possible values in the interval is described by a mathematical function f(x) called the **probability density function** (pdf)**.** A high (low) clustering of values will result in high (low) values of this function. For a continuous random variable X, only probabilities associated with ranges of values (e.g. an interval of values from a to b) will be calculated. The probability that the value of *X* will fall between the values a and b is given by the area between a and b under the curve describing the probability density function f(x). For any probability density function the total area under the graph of f(x) is 1.

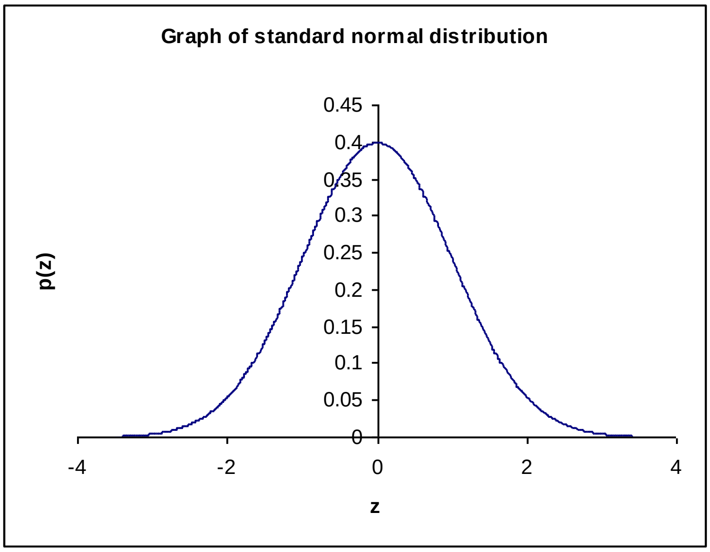
5.2 Normal distribution

A continuous random variable *X* is normally distributed (follows a normal distribution) if the probability density function of X is given by

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

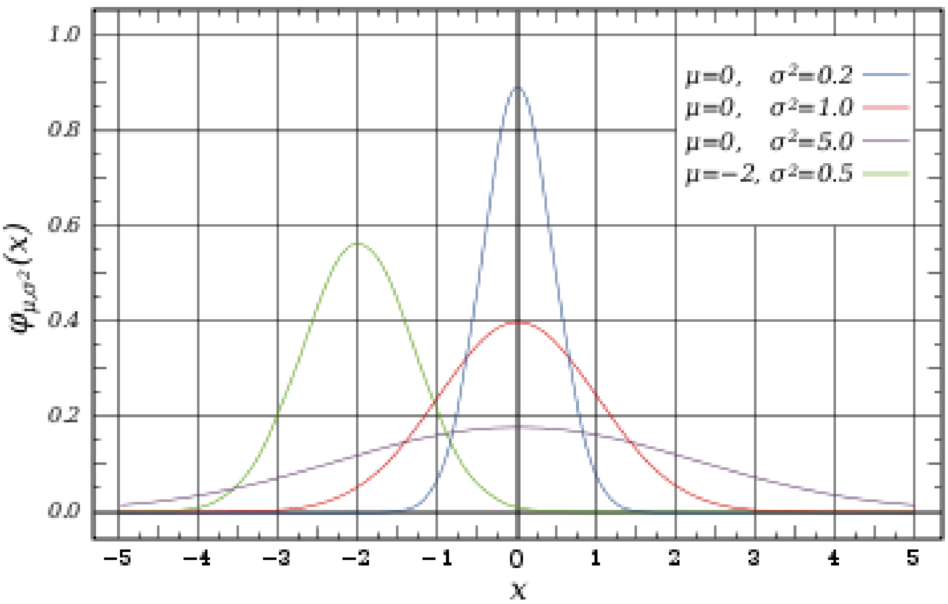
$$for \qquad -\infty < x < \infty$$

The constants μ and σ can be shown to be the mean and standard deviation respectively of *X*. These constants completely specify the density function. A graph of the curve describing the probability function (known as the normal curve) for the case $\mu = 0$ and $\sigma = 1$ is shown on the following page.



5.2.1 Properties of the normal distribution

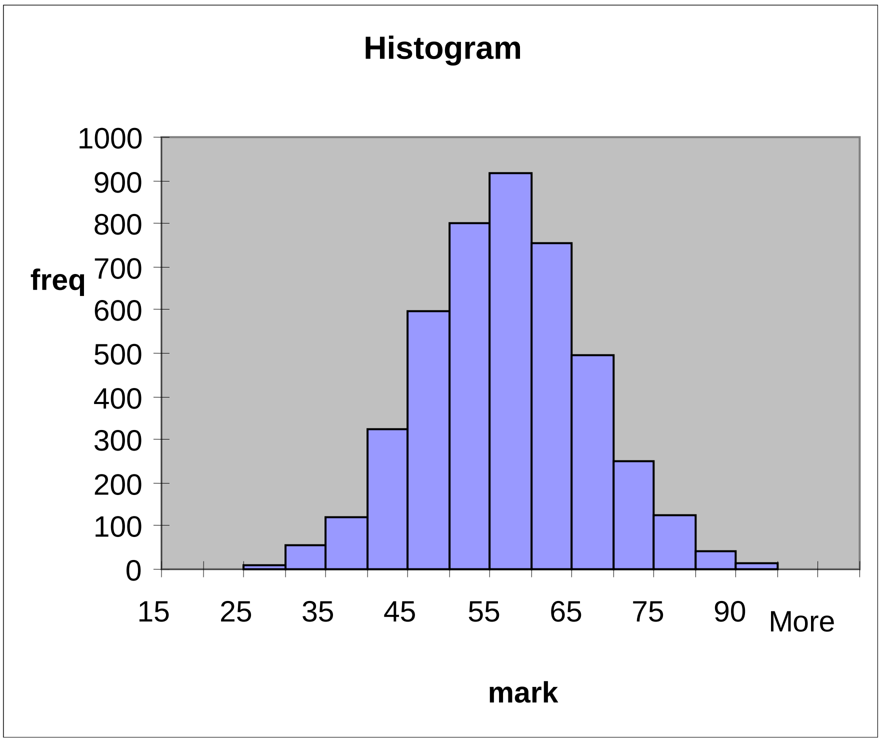
The graph of the function defined above has a symmetric, bell-shaped appearance. The mean μ is located on the horizontal axis where the graph reaches its maximum value. At the two ends of the scale the curve describing the function gets closer and closer to the horizontal axis without actually touching it. Many quantities measured in everyday life have a distribution which closely matches that of a normal random variable e.g. marks in an exam, weights of products, heights of a male population. The parameter μ shows where the distribution is centrally located and σ the spread of the values around μ. A short hand way of referring to a random variable X which follows a normal distribution with mean μ and variance σ^2 is by writing $X \sim N(\mu\,;\sigma^2)$. The next diagram shows graphs of normal distributions for various values of μ and σ^2.



An increase (decrease) in the mean μ results in a shift of the graph to the right (left). An increase (decrease) in the standard deviation σ results in the graph becoming more (less) spread out e.g. compare the curves of the distributions with $\sigma^2 = 0.2, 0.5, 1$ and 5 in the previous diagram.

5.2.2 Empirical example – The normal distribution and a histogram

Consider the scores obtained by 4 500 candidates in a matric mathematics examination.



The histogram of the marks has an appearance that can be described by a normal curve i.e. it has a symmetric, bell-shaped appearance. The mean of the marks is 51.95 and the standard deviation 10.

5.3 The Standard Normal Distribution

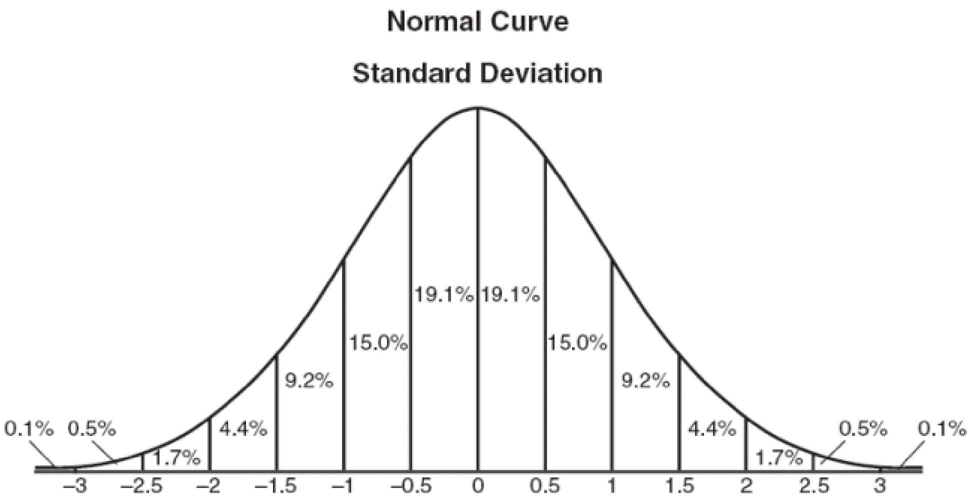
To find probabilities for a normally distributed random variable, we need to be able to calculate the areas under the graph of the normal distribution. Such areas are obtained from a table showing the cumulative distribution of the normal distribution (see appendix). Since the normal distribution is specified by the mean (μ) and standard deviation (σ), there are many possible normal distributions that can occur. It will be impossible to construct a table for each possible mean and standard deviation. This problem is overcome by transforming X, the normal random variable of interest [X ~ N(μ; σ^2)], to a standardized normal random variable

$$Z = \frac{X - \mu}{\sigma}.$$

It can be shown that the transformed random variable is normally distributed with μ = 0 and σ = 1 i.e. Z ~ N(0; 1). The random variable Z can be transformed back to X by using the formula

$$X = \mu + Z\sigma.$$

The normal distribution with mean μ = 0 and standard deviation σ = 1 is called the **standard normal distribution**. The symbol Z is reserved for a random variable with this distribution. The graph of the standard normal distribution appears below.



Various areas under the above normal curve are shown. The standard normal table gives the area under the curve to the left of the value *z*. Other types of areas can be found by combining several of the areas as shown in the next examples.

5.4 Calculating probabilities using the standard normal table

The standard normal table is found at the back of your notes.
The areas shown in the table are those under the standard normal curve to the left of the value of z looked up i.e. P(Z < z)

Note

- For negative values of z less than the minimum value (-3.79) in the table, the probabilities are taken as 0 i.e. $P(Z \le z) = 0$ for $z < -3.79$.
- For positive values of z greater than the maximum value (3.79) in the table, the probabilities are taken as 1 i.e. $P(Z \le z) = 1$ for $z > 3.79$.

<u>Examples</u>

In all the examples that follow, $Z \sim N(0; 1)$.

a) $P(Z < 1.35) = 0.9115$

b) $P(Z > -0.47) = 1 - P(Z \le -0.47)$

$= 1 - 0.3192$

$= 0.6808$

c) $P(-0.47 < Z < 1.35) = P(Z < 1.35) - P(Z < -0.47)$

$= 0.9115 - 0.3192$

$= 0.5923$

d) $P(Z > 0.76) = 1 - P(Z < 0.76)$

$= 1 - 0.7764$

$= 0.2236$

e) $P(0.95 \le Z \le 1.36) = P(Z \le 1.36) - P(Z \le 0.95)$

$= 0.9131 - 0.8289$

$= 0.0842$

f) $P(-1.96 \le Z \le 1.96) = P(Z \le 1.96) - P(Z \le -1.96)$

$= 0.9750 - 0.0250$

$= 0.95$

In all the above examples an area was found for a given value of z. It is also possible to find a value of z when an area to its left is given. This can be written as $P(Z \le z_\alpha) = \alpha$ (α is the greek letter for "a" and is pronounced "alpha"). In this case z_α has to be found where α is the area to its left

<u>Examples</u>

1) Find the value of z that has an area of 0.0344 to its left.

   Search the body of the table for the required area (0.0344) and then read off the value of z corresponding to this area. In this case $z_{0.0344} = -1.82$.

2) Find the value of z that has an area of 0.975 to its left.

   Finding 0.975 in the body of the table and reading off the z value gives $z_{0.975} = 1.96$.

3) Find the value of z that has an area of 0.95.

   When searching the body of the table for 0.95 this value is not found. The z value corresponding to 0.95 can be estimated from the following information obtained from the table.

z	area to left
1.64	0.9495
?	0.95
1.65	0.9505

   Since the required area (0.95) is halfway between the 2 areas obtained from the table, the required z can be taken as the value halfway between the two z values that were obtained

   from the table i.e. z = $\frac{1.64+1.65}{2} = 1.645$.

   <u>Exercise:</u> Using the same approach as above, verify that the z value corresponding to an area of 0.05 to its left is –1.645.

4) Find the value of z that has an area of 0.9841 to its left.

   When searching the body of the table this area is not found. The following information can be found.

z	area to left
2.14	0.9838
?	0.9841
2.15	0.9842

   The area required is not midway between the 2 other areas so the z-value corresponding to the closer area is used i.e. z = 2.15 is used.

At the bottom of the standard normal table selected percentiles z_α are given for different values of α. This means that the area under the normal curve to the left of z_α is α.

Examples:

1 $\alpha = 0.900$, $z_\alpha = 1.282$

means $P(Z < 1.282) = 0.900$.

2 $\alpha = 0.995$, $z_\alpha = 2.576$

means $P(Z < 2.576) = 0.995$.

3 $\alpha = 0.005$, $z_\alpha = -2.576$

means $P(Z < -2.576) = 0.005$.

The standard normal distribution is symmetric with respect to the mean = 0. From this it follows that the area under the normal curve to the right of a positive z entry in the standard normal table is the same as the area to the left of the associated negative entry (– z) i.e.

$$P(Z \geq z) = P(Z \leq -z) .$$

For example, $P(Z \geq 1.96) = 1 - 0.975 = 0.025 = P(Z \leq -1.96)$.

5.5 Calculating probabilities for any normal random variable

Let X be a $N(\mu ; \sigma^2)$ random variable and Z a $N(0 ; 1)$ random variable. Then

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)$$

Example 1

The height *H* (in inches) of a population of women is approximately normally distributed with a mean of $\mu = 63.5$ and a standard deviation of $\sigma = 2.75$ inches. To calculate the probability that a woman is less than 63 inches tall, we first find the z-value that is associated with h = 63 inches. (This z-value is sometimes referred to as the z-score.)

$$z = \frac{63 - 63.5}{2.75} = -0.18$$

Then use $P(H \le 63) = P(Z \le -0.18) = 0.4286$.
This means that 42.86% (a proportion of 0.4286) of women are less than 63 inches tall.

Example 2

The length X (inches) of sardines is a N(4.62 ; 0.0529) random variable. What proportion of sardines is

(a) longer than 5 inches? (b) between 4.35 and 4.85 inches?

(a) $P(X > 5) = P(Z > \frac{5 - 4.62}{0.23})$

$= P(Z > 1.65)$

$= 1 - P(Z \le 1.65)$

$= 1 - 0.9505$

$= 0.0495.$

(b)

$$P(4.35 \le X \le 4.85) = P\left(\frac{4.35 - 4.62}{0.23} \le Z \le \frac{4.85 - 4.62}{0.23}\right)$$

$= P(-1.17 \le Z \le 1)$

$= P(Z \le 1) - P(Z \le -1.17)$

$= 0.8413 - 0.1210$

$= 0.7203.$

5.6 Finding percentiles by using the standard normal table