

# Module 3

## Probability & Distributions

**Objective:** In this module you will learn some basic concepts of probability and some common probability distributions which are important in deciding on the statistical methods to be used for a data analysis. Further, in this module you will learn statistical methods to quantify uncertainty in the sample estimate of parameters in the population.

### 3.1 Introduction

In this module we discuss some very basic concepts of probability and statistical theoretical distributions. The idea of probability is not a new one in medical research. Health professionals very often use the term ‘probability’ in their daily communication. For example, we may hear a cardiac surgeon say that the patient has a 70% chance of surviving if he/she undergoes a Coronary Artery Bypass Graft (CABG) surgery. A clinician may say that two out of twenty patients were lost to follow up in a cancer study. A general practitioner (GP) may say that she is 90% sure that regular exercise lowers the blood glucose level of type II diabetic patients. Thus, everyone has an idea about probability. Probability itself is a vast area of statistics and hence a detailed discussion on it is beyond the scope of this module – in this module, we only concentrate on some very basic probability concepts.

A distribution of a random variable may be expressed in the form of a table, a graph, or a formula (mathematical functional form). Clinicians and researchers should have sound knowledge of the distribution of a random variable. Such knowledge provides them with a powerful tool for summarising and describing a data set. This also helps clinicians and researchers in reaching conclusions about a population of data on the basis of a sample of data drawn from the population. Most importantly the choice of statistical methods to be used for data analyses largely depends on the distribution of the data. In this module we discuss some commonly used statistical probability distributions such as Bernoulli, Binomial, Normal and t distributions. Bernoulli and Binomial distributions are appropriate for categorical binary data (only have two possible outcomes of the variable, e.g., current smoking status: yes/no, sex of patients in a cancer study: male/female, etc.) and for continuous data the Normal and t distributions are appropriate (e.g., age, height, weight, blood glucose level, cholesterol level, etc.).

In this module we aim to discuss on the following topics:

- Basic concepts of probability
- Inference about a population based on a sample
- Statistical distributions: Bernoulli, Binomial, Normal and t-distribution
- Sampling distribution for
  - the sample mean

- the difference in two sample means
- the sample proportion
- the difference in two sample proportions
- Standard Error (SE) and difference between SD and SE
- Central limit theorem

## 3.2 Basic Concepts of Probability

In the previous modules the idea of variation among the observations through descriptive measures, for example, the standard deviation was introduced. Variation in data is very common in many aspects of health sciences; variation comes from within patient (case/subject) variability, between patient variability, measurement error, and other sources. Inferential statistics is concerned with measuring the degree of uncertainty which can be quantified using the concept of probability. Knowledge of probability is the fundamental of statistical analyses of data however detailed discussion on probability is beyond the scope of this module.

Probability can mainly be defined in the following two ways: relative frequency probability and bayesian probability. In this module, however, we provide only the relative frequency definition of probability. This definition of probability depends on some process or experiment that occurs repeatedly under identical conditions, the number of times the experiment was repeated and also the number of times some outcomes of interest occurs. Let us assume that an experiment is repeated  $n$  times, and out of  $n$  times an event of interest occurs  $m$  times. Then the probability that the event occurs is just the relative frequency of occurrence of the event. Thus the relative frequency definition of probability is given by

$$\text{Probability (Pr) of the event occurs} = \frac{m}{n}$$

Note, however, that  $\frac{m}{n}$  is only an estimated probability of the event of interest.

For example, consider a retrospective case control study where 10000 ( $n = 10000$ , the number of smokers) smokers were studied to identify those who developed lung cancer in the recent years. If the study report shows that 54 ( $m = 54$ , number of patients who developed lung cancer) out of 10000 smokers experienced lung cancer, then the probability that a smoker developed lung cancer in the sample is as follows: Probability of developing lung cancer among smokers =  $\frac{m}{n} = \frac{54}{10000} = 0.0054$  or 0.54%.

Consider another example where we want to determine the probability of having a baby girl in a particular community. Let us assume that we observe 1000 births in a community and calculate the proportion of female births. If out of 1000 births, 490 were baby girls, then the probability that an expectant mother in this community will have a baby girl is  $490/1000 = 0.49$  or 49%.

### Some Common Definitions:

- **Experiment:** An experiment is the process by which an observation (measurement) is obtained. For example, observing 1000 births in the above example is an experiment.
- **Event:** An event is any set of outcomes (the number of outcomes could be single or multiple). In the baby birth example if we are interested in birth of baby girl, then birth of a baby girl is our event of interest.
- **Mutually Exclusive Event:** Two or more events are said to be mutually exclusive if they cannot occur simultaneously (one excludes the other). Consider the current smoking status (yes/no) of a group of diabetic patients. The two categories are mutually exclusive because if a diabetic patient is smoker, at the same time he/she can not be non-smoker.
- **Independent Events:** Two events are independent if the knowledge of one event can not be used to predict another event. Let us assume that we are interested in predicting the probability that a couple will have two girls, the event of interest in this example is the birth of a girl. Since the sex of future children are not influenced by the sex of previous children, we can assume the events are independent.

**Multiplicative Rule of Probability:** The multiplicative rule of probability states that the probability of two or more independent events occurring together is the multiplication of the probability of each event occurring. If the probability of having a girl is 0.49, the probability of having two girls in the next two births is:  $0.49 \times 0.49 = 0.24$  or 24% (corrected to two decimal places). Similarly, the probability of having three girls in the next three births is:  $(0.49)^3 = 0.117649 = 0.118$  or 11.8% (corrected to three decimal places).

**Additive Rule of Probability:** Using the additive rule of probability we can calculate the probability of mutually exclusive events. This is calculated as the sum of probabilities of observing each event. Assume that the probability of observing a blood type group O, A, B, and AB are 0.46, 0.43, 0.08 and 0.03 respectively. Then the probability of an individual having blood type A or B is  $0.43 + 0.08 = 0.52$ . The probability of being any blood type is the sum of all probabilities which is 1.0 ( $0.46 + 0.43 + 0.08 + 0.03 = 1.0$ ). This intuitively makes sense because a person has to have a blood type!

**NOTE:** If you find difficulty to understand the above examples, please consider coin and /or die tossing experiments – I am sure you are familiar with these experiments from GRADE 9 and 10 maths.

**Properties of probability:** Mainly there are three properties of probability. However, in this module we mention only two of them. They are as follows.

- Any event must have a probability greater than or equal to zero and less than or equal to one.
- If an experiment results in several mutually exclusive outcomes, then the sum of the probabilities of these outcomes is one (see additive rule above).

### 3.3 Parameter and its Sample Estimate

As was discussed in Module 2, the mean of a data set gives information about the central value of the data and the standard deviation conveys the information regarding the amount of variability present in the data. In fact, the mean measures the central value and standard deviation measures the spread or scatter of the observations about their mean. The mean and the standard deviations are two summary measures which have extensive use in all areas of applied research.

Consider the body mass index (BMI) of all patients who had cardiac surgery at Victorian public hospitals from 2001 to 2006 (a broader group of cardiac patients or population of cardiac patients). If all the patients in the population are taken into account to calculate the mean BMI, it is called the true mean or the population mean BMI and is denoted by the Greek letter  $\mu$  (Mu). Similarly, if all the patients in the population are considered for calculating the standard deviation, it is known as the true or population standard deviation and is denoted by the Greek letter  $\sigma$  (Sigma). Furthermore, if we calculate the proportion of patients in the population who are obese ( $\text{BMI} > 30 \text{ kg/m}^2$ ), the proportion is called population proportion or true proportion – the population proportion is denoted by the Greek letter  $\pi$ . The true mean, standard deviation and proportion are known as the **parameters** for the distribution of BMI data in the population.

In practice, the parameters are unknown and difficult to calculate particularly for a large population but can be estimated from the sample data. Therefore we use data in the sample to make inference about the parameters (e.g., true mean, standard deviation and proportion). If we take a sample from a population and calculate the sample mean, standard deviation and proportion, then they are known as the sample mean, standard deviation and proportion respectively. The sample mean, standard deviation and proportion are denoted by  $\bar{x}$ ,  $s$  and  $\mathcal{P}$  respectively. The sample mean, standard deviation and proportion are also known as the estimate of the true mean, standard deviation and proportion respectively.

As discussed earlier, calculation of parameters requires the collection of data from every patient in the population – this involves a lot of money, longer time and a lot of people to help in data collection. However, time and money constraints do not allow us to collect data for all patients in the population, hence we depend on sample data for all most all medical studies. In practice, we take a random sample from the population of interest and collect data from the patients in the sample. The number of patients in the sample should not be too large or too small – an optimum number of patient can be determined by using some statistical methods known as the sample size calculation (will not be covered in this subject).

For example, let us assume that we a random sample of 10 patients from the cardiac surgery population discussed above – the BMI for the patients in the sample are: 32.4, 33.4, 25.0, 27.5, 33.9, 26.6, 34.1, 26.0, 30.3 and 21.6 kg/m<sup>2</sup>. The mean and standard deviation for this sample data are respectively 28.35 kg/m<sup>2</sup> and 4.78 kg/m<sup>2</sup> and the proportion of patients with BMI above 30 kg/m<sup>2</sup> is 50% (5 out of 10 patients). Hence, 28.35 kg/m<sup>2</sup>, 4.78 kg/m<sup>2</sup> and 50% are respectively the sample estimate of the true mean, standard deviation and proportion of patients who are obese (we assume that patients with BMI>30 kg/m<sup>2</sup> are obese). Thus the data in a random sample help us to make inferences about the population from which the sample has been drawn.

## 3.4 Statistical Distributions

Most of the data collected in medical research follow some specific patterns. A pre-specified pattern of a data set is known as the distribution of the data. There are many statistical distributions, however, there are data sets those do not follow any of the distributions available in statistical science. A data of this type is known as the distribution free data or alternately we say that the data follows a non-parametric distribution. In this module we will discuss the Bernoulli, Binomial, Normal and t distributions. The Bernoulli and Binomial distributions are appropriate for discrete data and the Normal and t distributions are useful for continuous data. Brief discussions on these four distributions are given below.

### 3.4.1 The Bernoulli Distribution

The Bernoulli distribution is an example of a discrete distribution, it is the simplest discrete distribution in statistics and is the foundation of other discrete distributions such as the binomial distribution (which will be discussed in the next section). The Bernoulli distribution describes the probability of only two possible mutually exclusive outcomes; the distribution considers only one patient in a sample. For example, the probability that an individual smoker has or has not developed lung cancer. These outcomes are often described as ‘success’ (yes) often denoted by 1, and a ‘failure’ (no) denoted by 0. Success is usually used to describe an individual having a disease. The Bernoulli distribution has only one parameter which is denoted by a Greek letter  $\pi$  (pi). The value of this parameter is the probability of observing a success. The probability of failure is then  $1 - \pi$ . In the lung cancer study discussed in Section 3.2, the probability of success (developing lung cancer) is 0.0054 or 0.54% ( $= \pi$ ) and the probability of failure is 0.9946 or 99.46% ( $= 1 - \pi$ ).

### 3.4.2 The Binomial Distribution

The binomial distribution is one of the most widely used probability distributions in the health sciences. This distribution is an extension of the Bernoulli distribution. As discussed earlier, the Bernoulli distribution describes the probability of a success in a sample with one patient (observation/trial) only. The Binomial distribution describes the probability of observing a number of successes, with more than one patient.

Let us assume that two smokers have the lung cancer screening test. The binomial distribution gives the following probabilities: (a) neither of the smokers have lung cancer, (b) one of the smokers has lung cancer, or (c) both of them have lung cancer. This can be extended for any number of patients. The Binomial distribution has two parameters, namely:

- the probability of a success  $\pi$  ( $P_i$ ) and
- the sample size  $n$  (the number of patients).

If the values of these two parameters are known, we can calculate the probability of observing a specified number of successes. We can also calculate the expected number of success in a sample (expected number of success = sample size  $\times$  probability of success). Consider the lung cancer study discussed earlier where 0.54% of the smokers develop lung cancer – assume that this claim is true. Let us draw a random sample of 500 smokers from the population, then the expected number of smokers who will develop lung cancer in this sample is  $500 \times 0.54\% = 2.7$  or 3 (approximately).

Using binomial distribution we can calculate, for example, the probabilities that more than 3 smokers will develop lung cancer, none will develop lung cancer, less than 3 will develop lung cancer, exactly 5 will develop lung cancer, etc. These binomial probabilities can be calculated using binomial distribution probability table or a computer. However, discussion on the calculation of probabilities using the binomial table is beyond the scope of this module.

Binomial distribution assumes the following conditions:

- There is a fixed number of patients (smokers).
- There are exactly two possible outcomes for each patient, e.g., lung cancer status of a patient in the lung cancer study – yes/no.
- The patients are independent (the patients are selected randomly).
- The probability of either outcome is the same for each patient, e.g., consider the lung cancer study – the probability of developing lung cancer is the same as not developing lung cancer (50%/50%).

The binomial distribution is useful when we are interested in the occurrence of an event, not in its magnitude. For instance, in a clinical trial, a patient may survive or die; here we study the number of survivors, and not how long the patient survives after treatment. Another example is whether a person is ambitious or not. Here, the binomial distribution describes the number of ambitious persons, and not how ambitious they are.

### 3.4.3 The Normal Distribution

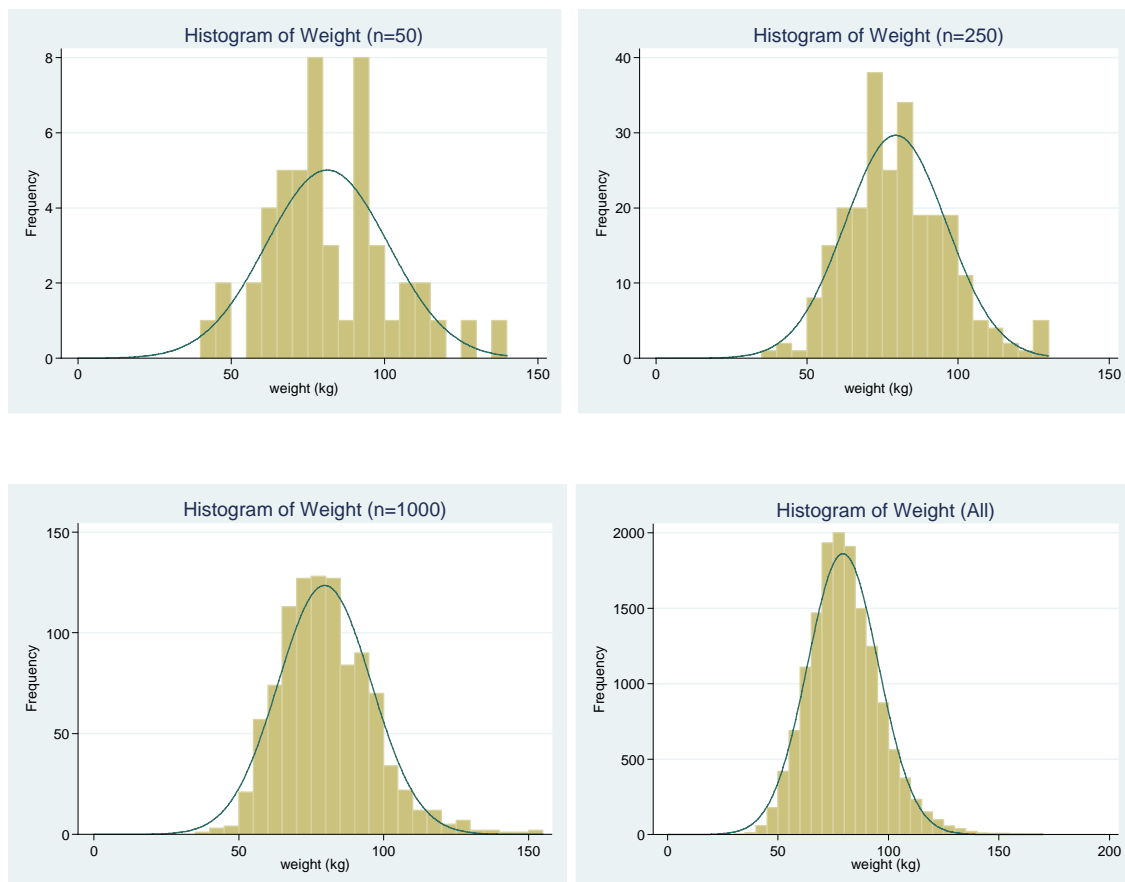
The normal distribution is the most useful distribution in statistical science. Many statistical theories have been developed based on the assumption that the data follows the normal distribution. The normal distribution is also known as the Gaussian distribution, the most commonly used distribution in statistics. Many measurements in biological science are normally or approximately normally distributed, examples includes: blood

pressure, pulse rates, height, weight, cholesterol level, etc. The normal distribution is a continuous distribution which is specified by a true mean ( $\mu$ ) and a standard deviation ( $\sigma$ ). The mean and standard deviation of the normal distribution are usually unknown and they are **called the parameters of the normal distribution**.

To help us understand the nature of the distribution of a continuous random variable, let us consider the frequency distribution of the weight for a sample of 50 cardiac surgery patients presented in Figure 3.1. Figure 3.1 shows that although the weights pile up towards the middle of the range, they do not do so very evenly, the curve is right skewed and contains peaks and valleys. If we increase the number of patients in our sample from 50 to 250 and summarize the weight in a histogram, they would then show the distribution of weight of 250 patients.

The histograms in Figure 3.1 clearly show that for larger samples the shape of the data approaches symmetry (this happens only when the population distribution of the variable is symmetric). We could add more and more patients in our sample, for example, number of patients 1000. As the sample size increases, the “peak and valley” effect disappears slowly and the curve become smoother and smoother, it becomes symmetric and bell shaped.

**Figure 3.1:** Distribution of weight for cardiac patients

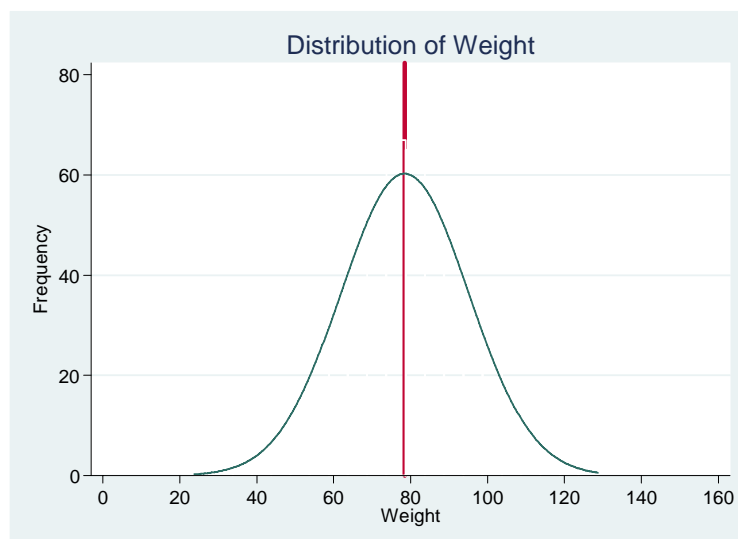


In general, as the number of observations in the sample becomes very large, and the width of class interval becomes very small, we would approach a histogram that is

symmetric and bell shaped. If the mid-points of all blocks are joined with a smooth line and the blocks are deleted, we will have a smooth curve such as the one shown in Figure 3.2. The curve is bell shaped and perfectly symmetric; the shape of a distribution of a data set like this is known as the normal distribution. In other words, the data follows a normal distribution.

The total area under the curve above the horizontal axis is one square unit because the area under the curve is the cumulative sum of relative frequencies (%s) of a relative frequency distribution table – the relative frequencies always add up to the value of one. In Figure 3.2, draw a perpendicular line from the peak of the curve (the highest point) to the horizontal line at the bottom. The point where the perpendicular meets the horizontal line is called the centre or mean of the normal distribution. Because of symmetry, 50% of the area is to the right of a perpendicular erected at the mean, and 50% is to the left. **The mean, median and mode of a normal distribution are the same.** From the Figure 3.2, the mean, median and mode for weight of cardiac surgery patients in the population is approximately 78.5kg.

**Figure 3.2:** Population distribution of weight (all patients)



### 3.4.3.1 Checking Normality of a Data Set

Steps for checking the normality of a data set are as follows:

- Construct a frequency or relative frequency table for the data
- Draw the histogram for the frequency/relative frequency table
- If the histogram is symmetric or approximately symmetric and bell shaped, the data is normally or approximately normally distributed.

An alternative way to check the normality of a data set is:

- Calculate the mean, median and mode for the data (USE Excel)
- If the mean, median and mode are approximately the same or closer to each other, the data is approximately normally distributed.

We can also check the normality by calculating skewness and kurtosis for the data but details about skewness and kurtosis is beyond the scope of this module.

### 3.4.3.2 Reference Ranges

If a data set follows the normal distribution with true mean  $\mu$  and true standard deviation  $\sigma$  then:

- 68% of the observations fall within one standard deviation of the true mean i.e. between  $\mu - \sigma$  and  $\mu + \sigma$ , or  $\mu \pm \sigma$ .
- 95% of the observations fall within 1.96 (or approximately 2) standard deviations of the true mean or between  $\mu - 2\sigma$  and  $\mu + 2\sigma$ , or  $\mu \pm 2\sigma$
- 99% of the observations fall within 2.58 (or approximately 3) standard deviations of the true mean or between  $\mu - 3\sigma$  and  $\mu + 3\sigma$ , or  $\mu \pm 3\sigma$

Consider weight of all patients in the cardiac surgery population – Figures 3.1-3.2 show that the distribution of weight follows the normal distribution. The mean and standard deviation for weight in the population are 78.5kg and 15.56kg respectively. Then,

- 68% of patients have weight between 62.94 and 94.06kg (i.e., probability that the weight of a randomly selected patient lies between 62.94 and 94.06kg is 68%).
- 95% of patients have weight between 47.38 and 109.62kg (i.e., probability that the weight of a randomly selected patient lies between 47.38 and 109.62kg is 95%)
- 99% of patients have weight between 31.82 and 125.18kg (i.e., probability that the weight of a randomly selected patient lies between 31.82 and 125.18kg is 99%).

Let us consider another example where the body mass index (BMI) of the cardiac surgery patients in the population follows a normal distribution with true mean of  $\mu = 27.94 \text{ kg/m}^2$  and true standard deviation of  $\sigma = 4.85 \text{ kg/m}^2$ . Fill up the gaps below:

- 68% of patients have BMI between \_\_\_\_ and \_\_\_\_ (i.e., probability that the BMI of a randomly selected patient lies between \_\_\_\_ and \_\_\_\_ is 68%).
- 95% of patients have BMI between \_\_\_\_ and \_\_\_\_ (i.e., probability that the BMI of a randomly selected patient lies between \_\_\_\_ and \_\_\_\_ is 95%).
- 99% of patients have BMI between \_\_\_\_ and \_\_\_\_ (i.e., probability that the BMI of a randomly selected patient lies between \_\_\_\_ and \_\_\_\_ is 99%).

### 3.4.3.3 Probabilities for Other Ranges

The probabilities for other reference ranges can not be calculated easily as was done for reference ranges discussed above. However, a simple transformation of the variable (e.g., BMI) can be used to calculate the probability. This transformation requires the knowledge of the true mean and standard deviation of the variable in the population. The transformation is achieved by subtracting the true mean from the observed value and dividing this difference by the true standard deviation of the variable. The whole expression is denoted by  $Z$  and is known as  $Z$ -score or standard score. The  $Z$ -score follows the normal distribution with mean zero and standard deviation one – a normal distribution with mean zero and standard deviation one is called the standard normal distribution. The formula for  $Z$ -score is:

$$\mathbf{Z\text{-score} = (\text{Observation} - \text{True Mean})/(\text{True Standard Deviation})}$$

Let us consider an example where we want to calculate the probability that a randomly selected patient from the cardiac population has BMI greater than 25, that is, we want to find  $\Pr(\text{BMI} > 25)$  [How to read it? – Probability of BMI greater than 25  $\text{kg/m}^2$ ]. Let us assume that the BMI in the population has mean of 27.94  $\text{kg/m}^2$  and the standard deviation of 4.85  $\text{kg/m}^2$ . As discussed above, we need to transform the value of the variable ( $\text{BMI} = 25 \text{ kg/m}^2$ ) to a  $Z$ -score.

For the BMI data, the true mean is 27.94  $\text{kg/m}^2$  and the true standard deviation is 4.85  $\text{kg/m}^2$ . So the transformed variable or the  $Z$ -score is:  $Z = (25 - 27.94) / 4.85 = -0.61$ . This transformed observation on the standard normal scale has an useful explanation. It tells us how many standard deviations the observation is away from the true mean. Thus  $Z = -0.61$  means that the patient's BMI of 25 is 0.61 standard deviations below the true mean of BMI. Usually statistical packages are used to calculate this probability however in this module we will discuss how to calculate it using the table that provides the probability under the standard normal curve (see **Table 3.2** in the Appendix). Please note that understanding of the normal probability table is vital to success in this subject! Once the observation (BMI value of 25) has been transformed to  $Z$ -score, the probability of BMI greater than 25 is the same as the probability of  $Z$ -score greater than -0.61, i.e.  $\Pr(\text{BMI} > 25) = \Pr(Z\text{-score} > -0.61)$ . Now using normal probability table (**Table 3.2**), we can easily calculate the required probability – the use of the normal probability table is discussed below.

The following steps and the instructions provided in the **Table 3.2** will help to calculate the probability using normal table.

- Calculate the  $Z$ -score
  - $Z$ -score = -0.61 for the BMI example.
- Consider the absolute value of the  $Z$ -score and break it into two parts: (a) the whole number and the tenth and (b) the hundredth.
  - The absolute value of the  $Z$ -score is 0.61
  - The whole number and the tenth is 0.6 (the whole number is zero)
  - The hundredth is 0.01

- The whole number and the tenth (0.6) are looked up along first column and the hundredth (0.01) is looked up across the first row in **Table 3.2**.
- The value in the intersection of the row and column is the probability between zero and the absolute value of the Z-score. Thus, from **Table 3.2** the probability between 0 and 0.61 is 0.2291.
  - Note: Because of symmetry of normal distribution, the probability between -0.61 and zero is the same as the probability between zero and 0.61.

We are interested in calculating the probability of Z-score greater than -0.61 which is the sum of the following two probabilities: (a) probability of Z-score from -0.61 to zero which is 0.2291 and (b) probability of upper half of the normal curve which is 0.50. Thus the required probability is 0.7291 ( $0.2291+0.50 = 0.7291$ ). We can also use instructions in **Table 3.3** to calculate probabilities from **Table 3.2**.

**Note:**

- Because of symmetry of the normal distribution, we consider the absolute value of the Z-score. This means even if the Z-score is negative, consider it positive while calculating probability.
- Because of symmetry probability between “-a” and “zero” is the same as the probability between “zero” and “+a”.

**Note:** Many of you are already familiar with calculation of probability using normal probability table – if you know any easier way to find these probabilities please avoid the methods I discussed here. The **Table 3.3** is just a guideline for the students who are not familiar with normal probability calculations.

**HOMEWORK:** Assuming that BMI follows a normal distribution with true mean of 27.94 and true standard deviation of 4.85, find the following probabilities:

- The probability that a patient selected at random has BMI less than 27.94  $\text{kg/m}^2$  (Answer: 50%).
- The probability that a patient selected at has a BMI within two standard deviation of the true mean (Answer: 95%)
- The probability that a patient selected randomly has a BMI level above 30  $\text{kg/m}^2$ , that is, find  $\text{Pr}(\text{BMI}>30)$  (Answer:  $0.5 - 0.1628 = 0.3672$ ; Z-score = 0.42)
- The probability that a patient selected at random has a BMI between 25 and 30  $\text{kg/m}^2$ , that is, find  $\text{Pr}(25<\text{BMI}<30)$  (Answer:  $0.1628 + 0.2291$ ; Z-score for BMI=25  $\text{kg/m}^2$  is -0.61, Z-score for BMI=30  $\text{kg/m}^2$  is 0.42) .

### 3.4.4 The $t$ Distribution

The  $t$  distribution is similar to the normal distribution; both are symmetric and bell shaped. Tails are a bit fatter for the  $t$  distribution compared to the normal distribution. The shape of the  $t$  distribution depends on the sample size. For large sample sizes, the  $t$  distribution is more like the normal distribution. The  $t$ -distribution is appropriate for continuous data.

## 3.5 Sampling Distribution and Quantifying Uncertainty

Medical research often involves acquiring data from a sample of individuals and using the information gathered from the sample to make inferences about a broader group of individuals. This broader group, as mentioned in Module 1, is called the population of interest or target population. We define a population as the complete group of individuals that we are interested in studying, for example, the preoperative creatinine level of the cardiac surgery patients in Victoria between 2001 and 2007. Ideally we would like to measure all individuals in a population directly however this is rarely possible due to time and financial restrictions. Fortunately, by selecting a representative sample from the population, the data from the sample can be used to draw conclusions about aspects of the population.

### 3.5.1 Random Sampling

Selection of a representative sample from a population is best performed using a random sampling procedure, for example, simple random sampling discussed in Module 1. In this procedure a sample of individuals is selected at random from a list of the entire population of interest. Each individual in the population has an equal chance of being selected. A computer generated list of random numbers can be used to select at random from the population list, e.g. the random numbers may tell us to choose the individuals from the list.

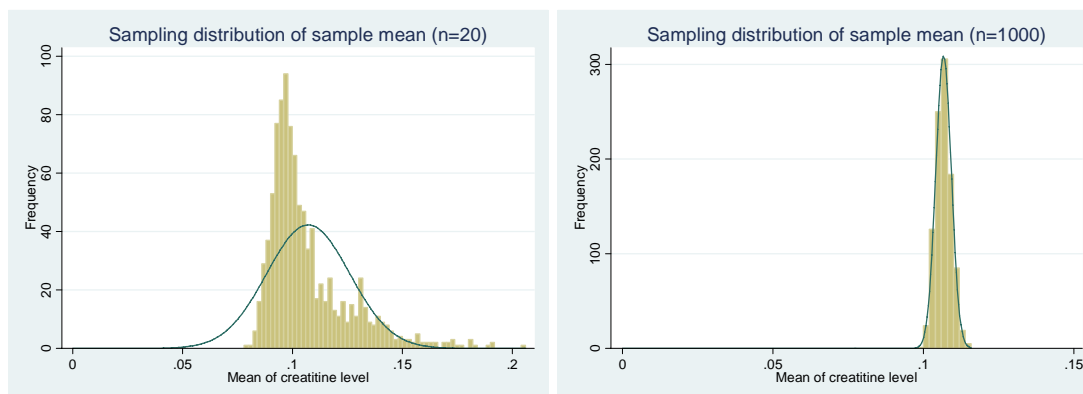
For example, we may be interested in determining the mean creatinine level for the cardiac surgery patients. We could select a sample of, for example, 20 patients from the population. The mean of creatinine level in the sample would be our “estimate” in the population. Let us now make our first attempt at the logic of using data from a sample to draw inference about a population. In the above example, common sense tells us that it is highly unlikely that the mean of surgery patients in the sample will exactly equal the population mean. Furthermore, if we selected another random sample of surgery patients, the sample mean creatinine level of the patients based on this second sample would most likely be slightly different to the mean of the first sample. This variability in results from sample to sample is purely random, i.e., due to chance alone; we call it sampling variability. One aim of statistical theory is to quantify the uncertainty associated with a sample result, e.g., sample mean of creatinine level, when used to estimate the population mean.

### 3.5.2 Sampling Distribution (for Sample Mean)

The method for quantifying uncertainty is based on knowing **how** sample means (or sample proportions, even sample relative risks and odds ratios) vary from sample to sample when the samples are based on the same population. This then gives us an idea about the uncertainty in a single sample, and so, for example, we can work out the uncertainty in a sample mean calculated as an estimate of the population mean.

Let us now illustrate the uncertainty concept with an “experiment” involving taking many samples from the population – one would never do this in real-world research! Let us draw a random sample of size 20 from the cardiac population and calculate the sample mean creatinine level. If we repeat this many times, ultimately we have a list of many sample means from different samples. The sample means can be summarised in a frequency distribution table or histogram. This frequency or histogram is called the sampling distribution for the sample mean. Statistical theory states that sample means arising from repeated samples will cluster around the population mean if the size of the sample is sufficiently large. For a symmetric distribution, the population mean lies in the tallest block of the histogram. Figure 3.3 shows that the sample means are clustered more tightly around the population mean for larger sample size.

**Figure 3.3:** Sampling distribution for sample means



Steps for the construction of a sampling distribution are as follows.

- Step 1: Select the sampled population.
- Step 2: Draw repeated samples of the same size from the population.
- Step 3: Compute the sample mean for each sample.
- Step 4: Construct a frequency table/histogram for the sample means in Step 3. This frequency distribution table or histogram is known as the sampling distribution for the sample mean.

In repeated random samples, the differences among sample means are due to sampling variability. Thus there is uncertainty in the sample estimates of the true mean. This uncertainty cannot be eliminated totally however it can be reduced by increasing the sample size.

### 3.5.3 Quantifying Uncertainty

As the sample size increases the sample mean approaches the true mean (which falls at the centre of the distribution, see Figure 3.3) however the sample means in repeated samples may not be the same. The differences among sample means are due to sampling variability. Thus there is uncertainty in the sample estimate of the population mean and this uncertainty cannot be eliminated totally however it can be reduced by increasing the sample size.

The uncertainty in the sample estimates can be quantified by calculating the standard error, which will discuss now. The sampling distributions for sample means for sample of 20 and then 1000 patients are presented in the histogram in Figure 3.3. We now look at the spread or clustering of these sample means around the population mean. This spread can be measured using the standard deviation of the sample means in the repeated samples and this leads us to a special term: the standard error (SE) of the mean.

Let us continue the creatinine level experiment to consider what happens if we take sample of size 1000 rather than of size 20. Intuitively, because of the larger number (1000 compared to 20) in each sample, we would expect the spread of the individual sample means from samples of size 1000 to be small in comparison with the spread for samples of size 20. This is a key point. From this we can infer that a sample mean from a larger sample is likely to lie closer to the population mean than a sample mean from a smaller sample.

To summarise what we have about the SE:

- It quantifies the anticipated “spread” of sample means across many samples.
- It is a measure of precision of the sample mean from a single sample in estimating the population mean. The smaller the SE, the more precisely the population mean is being estimated.

We can think of the standard error as measuring (in rough terms) the “average amount of error” in a single sample mean. So if the SE is small, then we have some degree of faith that the sample mean is fairly close to the population mean. However, it must be reiterated that uncertainty cannot be eliminated, merely quantified. The result of any sample will always be subject to uncertainty.

### 3.5.4 Calculation of Standard Error

Let us consider the preoperative serum creatinine level of five randomly selected patients from the cardiac surgery population, the creatine levels are: 0.04, 0.11, 0.10, 0.13 and 0.18. The mean and SD for these sample observations are respectively 0.112 and 0.0507 (up to 4 decimal places). Steps for calculating SE:

- Step 1: Draw a random sample
- Step 2: Calculate the SD of the sample observations
- Step 3: Divide the SD by the square root of the sample size; this gives the SE for the sample mean

A computation formula for SE for the sample mean is give by:

$$SE = \frac{\text{Sample SD}}{\sqrt{\text{sample size}}} = \frac{s}{\sqrt{n}}$$

Thus the SE for the sample mean for creatinine level data is 0.0227 ( $\frac{0.0507}{\sqrt{5}} = 0.0227$ ).

**HOME WORK:** Draw three random samples of sizes 10, 20 and 50 from the pulse rate data presented in Table 1.1 of Module 1. Calculate the mean for all pulse rate data and then calculate the mean and the standard error for each of the above samples. Compare the means and then standard errors – provide a conclusion if you have noticed any trend in your results.

You may use Microsoft Excel “ Data Analysis” option to calculate the standard error (Open Excel, click Tools, Data Analysis, Descriptive Statistics, select the data range, select the output range, click on descriptive statistics, click OK. In case you do not have “Data Analysis” option, install it operating the following steps: click “Tools”, “Add-In”, “Analysis ToolPak-VBA”, then OK).

### 3.5.5 Difference Between SE and SD

The following extract outlines the difference between SD and SE (Gardner, M.J. and Altman, D.G (1988): Confidence Intervals rather than p-values: estimation rather than hypothesis testing, British Medical Journal, 292: 746-750).

*“When numerical findings are reported ....., they are often presented with additional statistical information. The distinction between two widely quoted statistics – the standard deviation and the standard error – is however, often misunderstood.*

*The SD is a measure of the variability between individuals in the level of the factor being investigated, such as blood alcohol concentrations in a sample of car drivers, and is thus a descriptive index. By contrast, the SE is a measure of uncertainty in a sample statistic (e.g., sample mean). For example the SE of the mean indicates the uncertainty of the mean blood alcohol concentration among the sample of drivers as an estimate of the mean value among the population of all car drivers. The SD is relevant when variability between individuals is of interest; the SE is relevant to summary statistics such as means, proportions, relative risk, odds ratio, differences, etc.*

*The SE of sample statistics (e.g., mean, proportion, odds ratio, relative risk), which depends on both SD and the sample size, is a recognition that a sample is most unlikely to determine the population value exactly. In fact, if a further sample is taken in identical circumstances almost certainly it will produce a different estimate of the same population value. The sample statistic is therefore imprecise, and the SE is a measure of this imprecision. By itself the SE has limited meaning, but it can be used to produce a confidence interval (will be discussed in Module 4), which does have a useful interpretation.”*

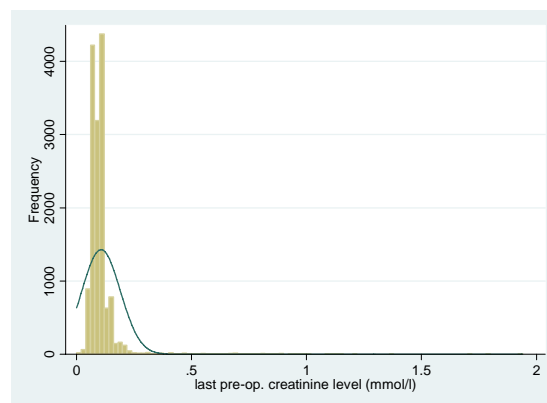
## 3.6 Central Limit Theorem (CLT)

The central limit theory is a very important theorem in statistics that is fundamental to many statistical methods. In this module we will not discuss the mathematical details of this theorem instead we will state the theorem and the implications of its use through practical examples.

**Theorem :** *For a large sample size, the distribution of the sample mean is normally distributed, even when the distribution from which the sample has been drawn is decidedly non-normal, with mean equal to the true mean of the sampled population and standard deviation equal to the standard error of the sample mean.*

Let us consider the distribution of preoperative creatinine level of the cardiac surgery patients presented in Figure 3.4; clearly the distribution of creatinine level in the population is heavily right skewed.

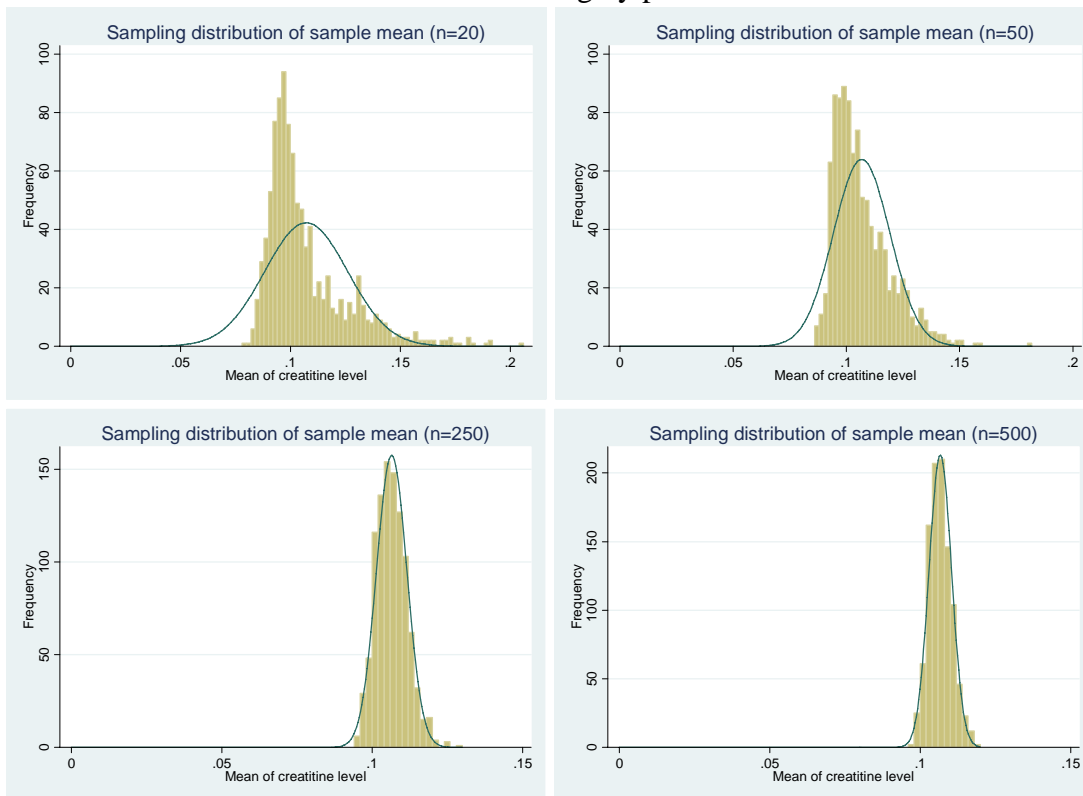
**Figure 3.4:** Preoperative creatinine level of cardiac surgery patients in the population



Let us randomly select **many samples** each of **size 20** from the cardiac surgery population and calculate the mean creatine level for the patients in each sample. These sample means are presented in a histogram in Figure 3.5 (top left). What is the shape of the sampling distribution for sample means? The distribution is not symmetric and is skewed to the right or positively skewed.

Suppose again we randomly select many samples each of size 50, and for each sample we calculate the sample mean. What does the distribution of the sample means presented in Figure 3.5 (top right) look like? How does this distribution of sample mean compare to the sampling distribution of sample mean for sample size 20? The sample means for sample size 50 have less spread than the spread of sample means for sample size 20.

**Figure 3.5:** Sampling distribution for the sample mean for preoperative creatinine level of cardiac surgery patients.



Again draw repeated random samples each of size 250 and for each of the sample calculate the mean. The histogram in Figure 3.5 (bottom left) shows that the distribution of sample mean is symmetric and bell shaped. It should be noted that for symmetric and bell shaped data, the true mean lies in the centre of the distribution. What do we notice from the sampling distributions for the mean for sample sizes 20, 50 and 250 (Figure 3.5)? The distribution of sample means is skewed and spread out when the sample sizes are small. As the sample size increases the sampling distribution of the sample mean becomes more concentrated around the true mean. This is more evident from the sampling distribution for sample of 500 patients (Figure 3.5, bottom left).

In conclusion, even though the population distribution of preoperative creatinine level is non-normal (see Figure 3.4, skewed to the right), the sampling distribution of the sample mean becomes normal for large sample. Hence, the sampling distribution for the sample mean presented in Figure 3.5 support the Central Limit Theorem stated earlier.

From the above discussion we note the following summary:

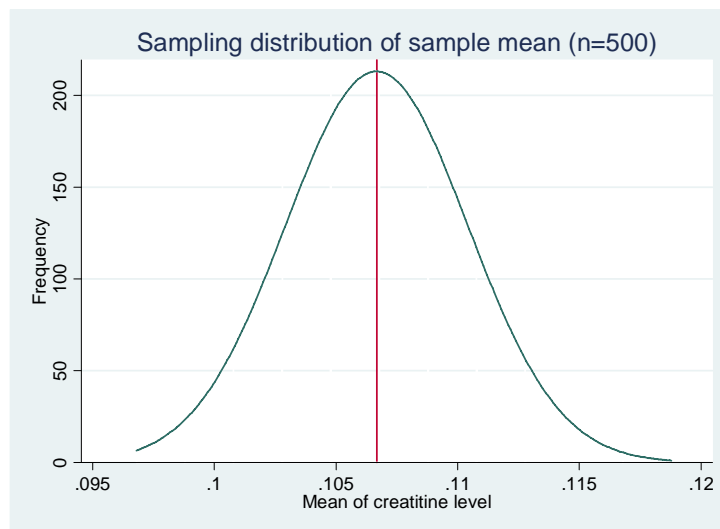
- For small sample sizes the distribution of the sample mean is spread out, and may be skewed.
- The distribution of the sample mean becomes more concentrated around the true mean for larger sample sizes. The degree of clustering also depends on the variability of the data in the population.

### 3.6.1 Probability for Reference Ranges for Sampling Distribution

Let us draw a smooth lined curve by connecting the mid points of the blocks in Figure 3.5 (bottom right, sample size 500); the smooth curve is shown in Figure 3.6. If we draw a perpendicular from the peak to the horizontal axis, the meeting point on the horizontal axis is the true mean of the distribution. Because the distribution is symmetric and bell shaped, the perpendicular divides the total area of the curve into two equal parts (with 50% each). Then according to the normal distribution probability law, in repeated sampling:

- 68% of the sample means are expected to lie within **one standard error** of the true mean, that is, within  $(\mu \pm 1 \times SE)$ .
- 95% of the sample means are expected to lie within 1.96 or approximately **two standard errors** of the true mean, that is, within  $(\mu \pm 2 \times SE)$ .
- 99% of sample means are expected to lie within 2.58 or approximately **three standard errors** of the true mean, that is, within  $(\mu \pm 3 \times SE)$ .

**Figure 3.6: Sampling distribution for the sample mean**



**HOME WORK:** Let us assume that the preoperative creatine level data for the cardiac surgery patients follows normal distribution with mean 0.10113 mmol/L and standard deviation 0.09241 mmol/L. Then,

- 95% of the sample means will be between \_\_\_\_ mmol/L and \_\_\_\_ mmol/L, or in 95% of the cases the true mean will be in the interval \_\_\_\_ mmol/L and \_\_\_\_ mmol/L.
- 99% of the sample means will be between \_\_\_\_ mmol/L and \_\_\_\_ mmol/L, or in 99% of the cases the true mean will be in the interval \_\_\_\_ mmol/L and \_\_\_\_ mmol/L.

### 3.6.2 Probability for other Ranges for Sampling Distribution

Probability for the other ranges can also be calculated using the normal probability table by transforming the sample mean to a Z-score, where the Z-score is calculated by subtracting the true mean from the sample mean and dividing this difference by the standard error of the sample mean. The formula for Z-score is:

$$\text{Z-score} = (\text{Sample Mean} - \text{True Mean}) / \text{Standard Error for the Mean}$$

Let us consider an example where we want to compute the probability that the mean birth-weight from a random sample of 10 infants from the Children's Hospital in Dhaka will be between 98.0 and 126.0 oz (ounces). The population mean birth weight in the hospital is 112 oz with a standard deviation of 20.6oz. Calculate the above probability (Answer: 96.48%). Also calculate the following probabilities.

- Probability that the sample mean birth-weight will be greater than 98.0 oz (Answer: 98.42%; Z-score for the mean weight of 98.0oz is -2.15).
- Probability that the sample mean birth-weight will be less than 126.0 oz (Answer: 98.42%; Z-score for the mean weight of 126.0oz is 2.15).

## 3.7 Sampling Distribution for the Sample Proportion

In the analysis of **categorical binary data**, we are usually concerned with the proportion of times that an event occurs rather than the number of times. Let us assume that we want to determine the proportion of cardiac surgery patients in Victoria who have preoperative creatinine level above 0.133mmol/L. Ideally we should investigate all patients in the population in our study, however as discussed in Module 3, the financial and time constraint does not allow us to deal with all patients in the population – instead we draw inferences about the population on the basis of the sample data. If the objective of the study is to determine the population proportion, we calculate the sample proportion and then use it as the estimate of the population proportion. The population and the sample proportions are denoted by  $\pi$  (probability of success, see Binomial distribution in Module 3) and  $\hat{p}$  respectively.

Consider a random sample of 50 patients from the cardiac surgery patients who have diabetes in Victoria, the data is presented in Table 3.1. The number of patients in the sample with a creatinine level greater than 0.133mmol/L is 8, thus the sample proportion is 0.16 (8/50) which is denoted by the Greek letter  $\hat{p}$ . This is known as the point estimate of the true proportion of patients in the population with preoperative creatinine level above 0.133 mmol/L. Let us draw another sample of the same size from this population and calculate the sample proportion- most likely this sample proportion will be different from 0.16. Thus if we draw repeated random samples each of the same size and calculate the sample proportion for each sample, the proportions are likely to be different from each other.

**Table 3.1:** Preoperative creatinine level (mmol/L) for 50 diabetes

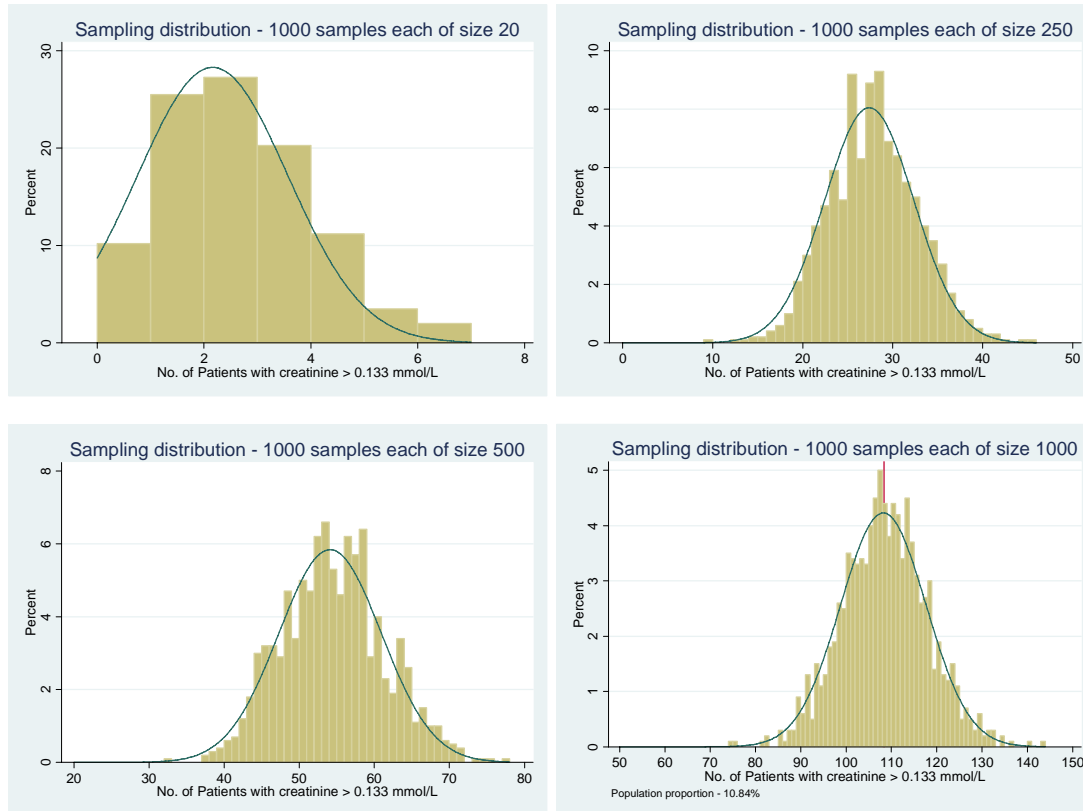
Creatinine > 0.133mmol/L (Yes = 1, No = 0)					
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	0
1	0	0	0	0	0
0	0	0	0	0	0
0	0	0	1	0	
0	1	0	1	0	
0	0	0	1	1	
0	0	0	0	0	

The difference among the sample proportions is due to the sampling variability or by chance alone. As we know from Module 3 that this sampling variability or uncertainty can be quantified by calculating the standard error of the sample estimates. The confidence interval incorporates the sampling variability (or uncertainty in the sample estimate) through inclusion of the standard error in its calculation procedures. So, it is reasonable to calculate an interval for the true proportion with a certain confidence, say 95%. The calculation of the confidence interval for proportions also requires the knowledge of sampling distribution for sample proportions.

Like the sampling distribution for the sample mean, the sampling distribution for sample proportion also follows a normal distribution for large samples. Sampling distributions for sample proportion for various sample sizes are shown in Figure 4.3. Clearly, for a large sample size the sampling distribution is more like a normal distribution. Thus, according to the normal distribution probability law in repeated samples:

- 95% of the sample proportions will be within 1.96 SE of the population proportion, i.e., within  $\pi \pm 1.96 \times SE$ . This means, in repeated sampling, 95% of the times the true proportion will be included by the interval  $p \pm 1.96 \times SE$ .
- 99% of the sample proportion will be within 2.57 (or approximately 3.0) SE of the population proportion, i.e.,  $\pi \pm 2.57 \times SE$ . This means, in repeated sampling, 95% of the times the true proportion will be included by the interval  $p \pm 2.57 \times SE$ .

**Figure 3.7: Sampling distribution for the sample proportion**



### 3.8 Sampling Distribution for Differences of Two Statistics

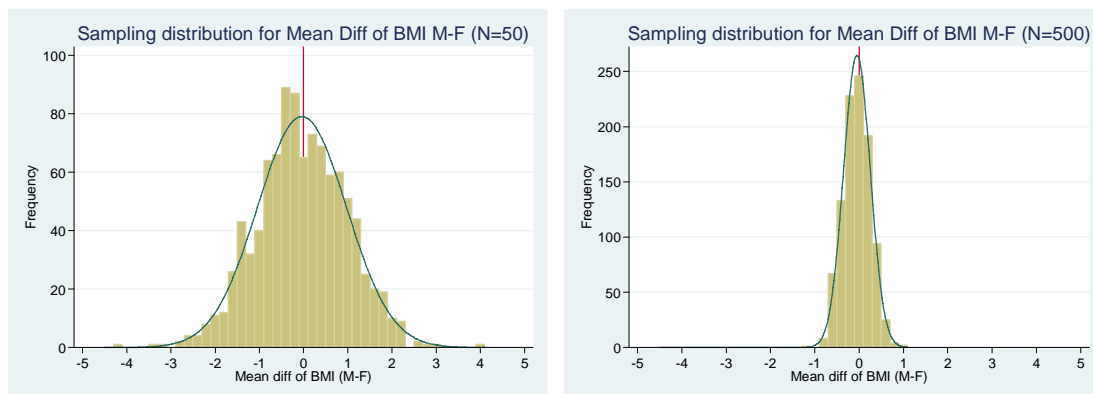
In medical as well as other research, we often require the sampling distribution for the difference of two statistics, e.g., difference of two sample means and difference of two sample proportions. It should be noted that the two samples are drawn independently from two separate populations. For example, creatinine levels of male and female cardiac surgery patients in Victoria – here male and female patients are two different populations within Victorian Cardiac Surgery patients. Another example could be populations of smokers and non-smokers in Australia.

Let us first discuss the sampling distribution for the difference of two sample means. For large sample, the sampling distribution for the difference in two sample means follows normal distribution- we can show this theoretically, however the theoretical discussion may be bit complex to follow. Alternatively, we can explain the concept through empirical sampling distribution for the difference of two sample means.

For example, let us draw two independent samples (samples in each pair are independent) from male and female cardiac surgery patients in Victoria and calculate the difference in sample means of BMI – and repeat this for 100 times. These differences are shown in Figure 4.2; clearly the sampling distribution for the difference in sample means approaches a normal distribution as the sample size increases. Then, according to the normal distribution law in repeated sampling:

- 68% of the differences in sample means fall within one SE of population mean difference, i.e., within “population mean difference  $\pm 1*SE$ ”
- 95% of the differences in sample means fall within two SE of population mean difference, i.e., within “population mean difference  $\pm 2*SE$ ” and
- 99% of the differences fall within three SE of population mean difference.

**Figure 3.8:** Sampling distribution for the difference between sample mean of BMI for male (M) and female (F) cardiac surgery patients (1000 samples).



Similarly, it can be shown that for large samples the sampling distribution for the difference in two sample proportions follows a normal distribution. Hence, according to the probability law of the normal distribution, in repeated sampling:

- 68% of the differences of sample proportions fall within one SE of population proportion difference, i.e., within “population proportion difference  $\pm 1*SE$ ”.
- 95% of the differences of sample proportions fall within two SE of population proportion difference, i.e., within “population proportion difference  $\pm 2*SE$ ”.
- 99% of the differences fall within three SE of population proportion difference, i.e., within “population proportion difference  $\pm 3*SE$ ”.

**Important Note:** If the absolute value of the Z-score is beyond the largest value (3.0) given in the normal distribution table (Table 3.2), then the probability b/w 0 and Z-score is obtained as 0.4990. For example, let us consider Z-score = 4.5, then the probability b/w 0 and 4.5 is 0.4990

## Module Summary:

- Probability- simple definitions
- Probability distributions – Bernoulli, binomial, normal and t distributions
- Reference ranges for the normal distribution probability curve for **observations** of a variable.
- Sampling distribution, quantifying uncertainty, standard error, difference b/w standard deviation and standard error
- **CLT:** For LARGE sample, sampling distribution of any statistic (e.g., mean, proportion, difference in two means, difference in two proportions, etc.) follows normal distribution.
- Reference ranges for the distribution of **sample statistics** (e.g., sample mean, difference in sample means, mean of differences, sample proportions, difference in sample proportions, etc.) – the sampling distributions.
- **Sampling distribution for the sample mean:**
  - In repeated sampling, 95% of the sample means fall within two standard error of the population mean, i.e., within “true mean  $\pm 2* SE$ ”.
    - Thus, if we draw 100 samples and for each sample we calculate the interval “sample mean  $\pm 2*SE$ ”, 95% of these intervals will include the true mean.
- **Sampling distribution for the difference in two sample means**
  - In repeated sampling, 95% of the differences b/w sample means fall within two standard error of the difference of population means, i.e., within the interval “difference in true means  $\pm 2* SE$ ”.
    - Thus, 95% of the times the interval “difference in sample means  $\pm 2*SE$ ” will include the difference of true means.
- **Sampling distribution for the sample proportion:**
  - In repeated sampling, 95% of the sample proportions fall within two standard error of the population proportion, i.e., within “true proportion  $\pm 2* SE$ ”.
    - Thus, if we draw 100 samples and for each sample we calculate the interval “sample proportion  $\pm 2*SE$ ”, 95% of these intervals will include the true proportion.
- **Sampling distribution for the difference in two sample proportions**
  - In repeated sampling, 95% of the differences b/w sample proportions fall within two standard error of the difference of population proportions, i.e., within “difference in true proportions  $\pm 2* SE$ ”.
    - Thus, 95% of the times the interval “difference in sample proportions  $\pm 2*SE$ ” will include the difference of true proportions.

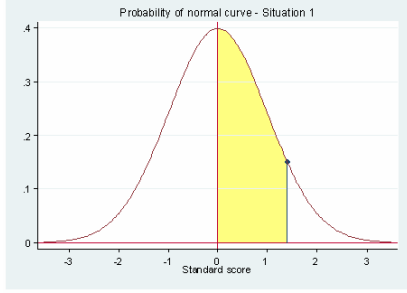
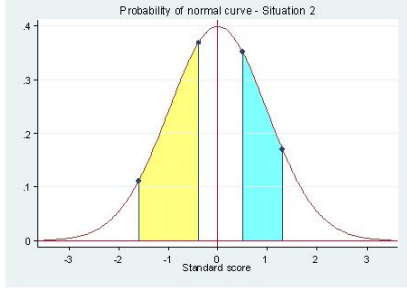
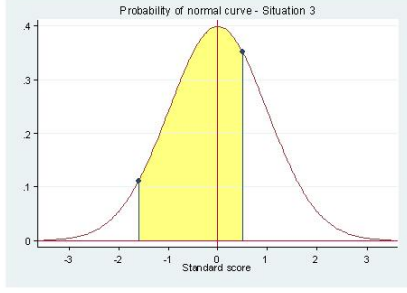
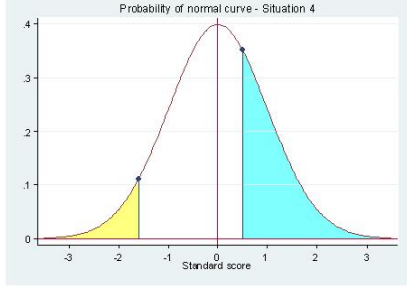
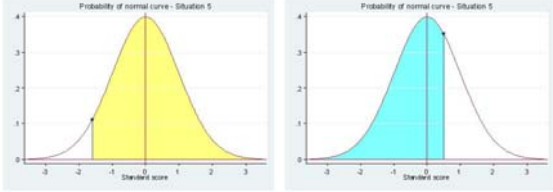
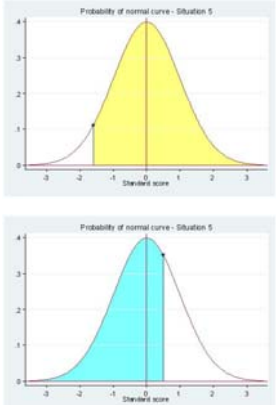
# Normal Probability Distribution



**Table 3.2: Normal Probability Between 0 and Z**

<b>Z</b>	<b>0.00</b>	<b>0.01</b>	<b>0.02</b>	<b>0.03</b>	<b>0.04</b>	<b>0.05</b>	<b>0.06</b>	<b>0.07</b>	<b>0.08</b>	<b>0.09</b>
<b>0.0</b>	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
<b>0.1</b>	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
<b>0.2</b>	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
<b>0.3</b>	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
<b>0.4</b>	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
<b>0.5</b>	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
<b>0.6</b>	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
<b>0.7</b>	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
<b>0.8</b>	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
<b>0.9</b>	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
<b>1.0</b>	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
<b>1.1</b>	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
<b>1.2</b>	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
<b>1.3</b>	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
<b>1.4</b>	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
<b>1.5</b>	0.4332	0.4345	0.4350	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
<b>1.6</b>	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
<b>1.7</b>	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
<b>1.8</b>	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
<b>1.9</b>	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
<b>2.0</b>	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
<b>2.1</b>	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
<b>2.2</b>	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
<b>2.3</b>	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
<b>2.4</b>	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
<b>2.5</b>	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
<b>2.6</b>	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
<b>2.7</b>	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
<b>2.8</b>	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
<b>2.9</b>	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
<b>3.0</b>	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

**Table 3.3: Situations to find normal probabilities**

Situation #.	Probability of Z-score & Instructions	Diagram
1	<p><b>Z Score:</b> Between zero and any number “a”</p> <p><b>Instructions:</b> Look up the area in the table b/w zero and “a” the <b>Table 3.2</b></p>	 <p>Probability of normal curve - Situation 1</p>
2	<p><b>Z Score:</b> Between two positives, i.e. b/w “+a” and “+b”, OR Between two negatives, i.e. b/w “-a” and “-b”</p> <p><b>Instructions:</b> Look up the areas b/w zero to “+a” and zero to “+b” in <b>Table 3.2..</b> Then, subtract the smaller from the larger.</p>	 <p>Probability of normal curve - Situation 2</p>
3	<p><b>Z Score:</b> Between a negative and a positive, i.e. b/w “-a” and “+b”</p> <p><b>Instructions:</b> Look up the areas b/w zero and “+a” and zero and “+b” in <b>Table 3.2</b> and add them together.</p>	 <p>Probability of normal curve - Situation 3</p>
4	<p><b>Z Score:</b> Less than a negative (<math>&lt; -a</math>), or Greater than a positive (<math>&gt; +a</math>)</p> <p><b>Instructions:</b> Look up the area b/w zero and “+a” in <b>Table 3.2</b>, and subtract from 0.5</p>	 <p>Probability of normal curve - Situation 4</p>
5	<p><b>Z Score:</b> Greater than a negative (<math>&gt; -a</math>), or Less than a positive (<math>&lt; +a</math>)</p>  <p><b>Instructions:</b> Look up the area b/w zero to “+a” in <b>Table 3.2</b> and add to 0.5.</p>	 <p>Probability of normal curve - Situation 5</p>