BIOSTATISTICS
TOPIC 5: SAMPLING DISTRIBUTION II
THE NORMAL DISTRIBUTION

The normal distribution occupies the central position in statistical theory and practice. The distribution is remarkable and of great importance, not only because most naturally occurring phenomena with continuous random variables follow it exactly, and not because it is a useful model in all but abnormal circumstances. The importance of the distribution lie in its convenient mathematical properties leading directly to much of the theory of statistics available as a basis for practice, in its availability as an approximation to other distributions, in its direct relationship to sample means from virtually any distribution, and in its application to many random variables that either are approximately normally distributed or can be easily transformed to approximate variables.

The word "normal" as used in describing the normal distribution should not be construed as meaning "usual" or "typical", "physiological" or "most common". In particular, a distribution that does not follow this distribution should be named "non-normal distribution" rather than "abnormal distribution". This problem of terminology has led many authors to refer to the distribution as Gaussian distribution, but this substitutes for a historical inaccuracy. In 1718, De Moivre, a great French mathematician, had derived a mathematical expression for the normal density in his 1718 tract *Doctrine of Chances*. Like Poisson's previous work, De Moivre's theorem did not initially attract the attention it deserved; it did however finally catch the eye of Pierre-Simon Marquis de Laplace (another great French mathematician and philosopher), who generalised it and included in his influential *Theorie Analytique des Probabilites* published in 1812. Carl F. Gauss, a great German mathematician, was the one who had developed the mathematical properties and shown the applicability of the De Moivre's distribution to many natural "error" phenomena, hence the distribution is sometimes referred to as Gaussian distribution.

So, how does the distribution work? The normal distribution was originally stated in the following way. Suppose that 1000 people use the same scale to weigh a package that actually weighs 1.00 kg, there will be values above and below 1.00 kg; if the probability of an error on either side of the true value is 0.5, a frequency plot of observed weights will have a strong tendency around 1.00 kg (Figure 1). The *error* about the true value may be defined as a random variable *X* which is continuous over the range $-\infty$ to $+\infty$. The probability distribution of the errors was called the *error distribution.* However, since the distribution was found to describe many other natural and physical phenomena, it is now generally known as the *normal* distribution. We will, therefore, use the term "normal" rather than De Moivre or Gaussian distribution.

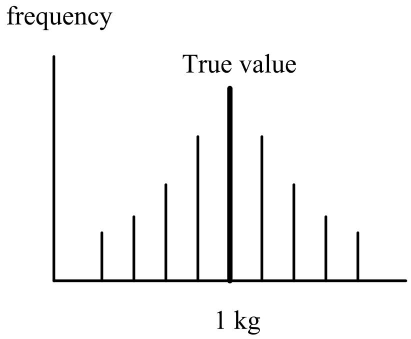


Figure 1: Plot of central tendency of observe weights around true mean of 1 kg.

I. CHARACTERISTICS OF RANDOM VARIABLES

Let us take the following cases.

Example 1: (a) Dr X has followed Mrs W for many years and found that her BMD was measured by DPX-L fluctuated around a mean of 1.10 g/cm^2 and standard deviation of 0.07 g/cm^2. At a recent assessment, her BMD was 1.05 g/cm^2. Is it reasonable to put her on a treatment?

(b) Mrs P has entered a clinical trial involving the evaluation of a drug treatment for osteoporosis. At baseline, multiple measurements of BMD (g/cm^2) was taken and the results are as follows:

0.95, 0.93, 0.97

After 6 months of treatment, the BMD was remeasured and found to be:

1.02, 1.05, 1.10, 1.03

She, however, complained that the medicine has made her slightly weak and other problems. Should you advise her to continue with the trial ?

We know that BMD or any other quantitative measurements are subject to random errors. But how much error was attributable to chance fluctuation and how much was due to systematic variation is a crucial issue. So, before answering this question (from a statistical point of view) properly, we will consider a fundamental distribution in statistics - the normal distribution.

The normal random variable is a continuous variable X that may take on any value between $-\infty$ to $+\infty$(while real world phenomena are bounded in magnitude), and the probabilities associated with X can be described in the following probability distribution function (pdf):

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

where μ and σ^2 are the mean and variance, respectively. These are, of course, parameters, and since they are the only quantities that must be specified in order to calculate the value of the probability.

For example, if μ = 50 and σ^2 = 100, we can calculate various probabilities as follows:

x	$\frac{1}{\sigma\sqrt{2\pi}}$	$\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	$f(x)$
20	0.03989	0.011109	0.00044
30	0.03989	0.135335	0.00540
40	0.03989	0.606531	0.02420
50	**0.03989**	**1.000000**	**0.03989**
60	0.03989	0.606531	0.02420
70	0.03989	0.135335	0.00540
80	0.03989	0.011109	0.00044

A plot of $f(x)$ and x resembles the bell-shape (Figure 2)

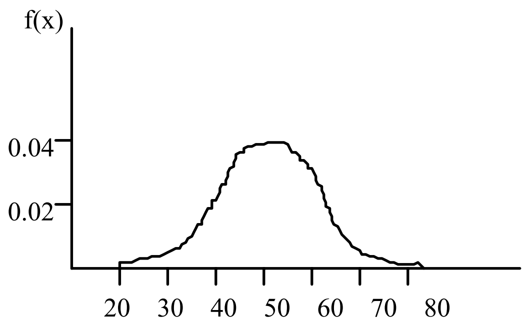


Figure 2: Graph of a normal distribution with mean = 50 and variance = 100.

It could be seen from this distribution that, the normal has the following properties:

(a) The probability function f(x) is non-negative.
(b) The area under the curve given by the function is equal to 1.
(c) The probability that the value *X* take on any value between x_1 and x_2 is represented by the area under the curve between the two points (Figure 3)

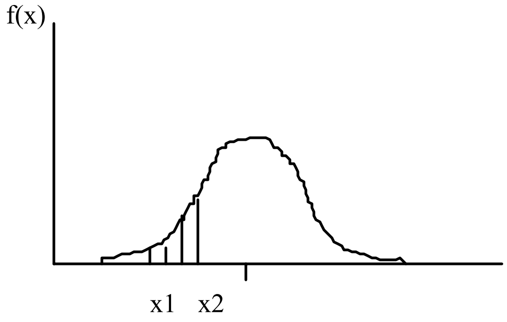
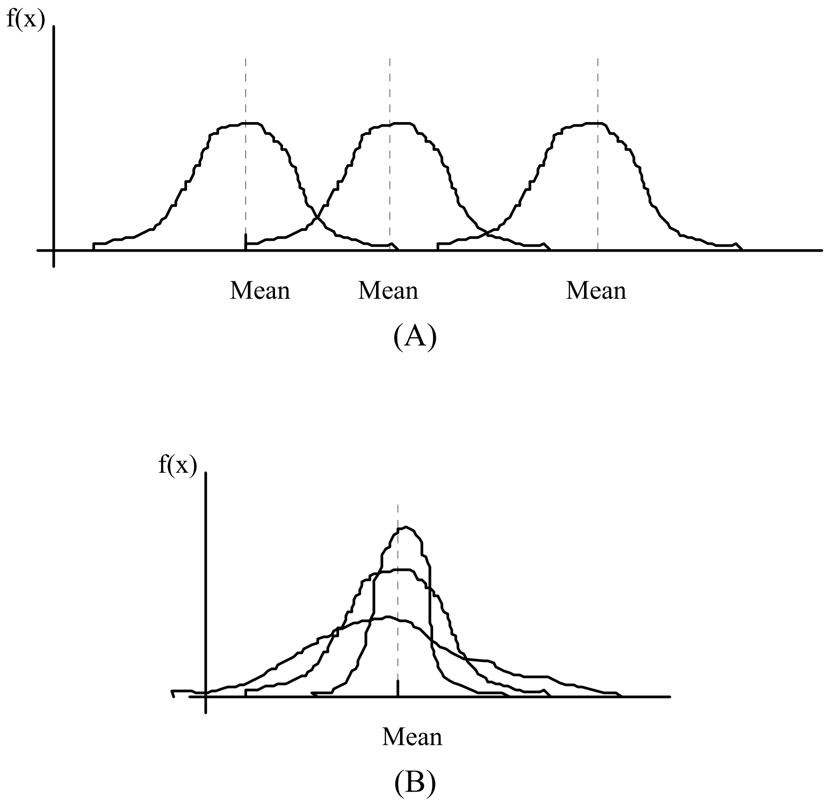


Figure 3: The probability that *X* takes value between x_1 and x_2.

(A) Effect of the Mean and Variance

We mentioned earlier that the normal probability distribution function (pdf) is determined by two parameters, namely, the mean (μ) and variance (σ^2). We can observe the effect of changing the value of either of these parameters. Since the mean describes the central tendency of a distribution, a change in the mean value have the effect of shifting the whole curve intact to the right or left a distance corresponding to the amount of change (Figure 4A). On the other hand, for a fixed value of μ, changing in the variance σ^2 has effect of locating the inflexion points closer to or farther from the mean, and since the total area under the curve is still equal to 1, this results in values clustered more closely or less closely about the mean (Figure 4B; please excuse my drawing!).



(A)

(B)

Figure 4 (A): The effect of changing in mean and (B) in standard deviation.

(B) Mean and Variance of A Normal Random Variable

It could be shown (by calculus) that the expected value (mean) and variance of the normal random variable are μ and σ^2, respectively. For brevity we write $X \sim N(\mu,\ \sigma^2)$ to mean that "X is normally distributed with mean μ and variance σ^2".

II. The Standard Normal Distribution

The normal distribution is, as we have noted, really a large family of distributions corresponding to the many different values of μ and σ^2. In attempting to tabulate the normal probabilities for various parameter values some transformation is necessary.

We have already seen in Topic 2 what happens to the mean and variance of any variable (say *Y*) when we make the transformation

$$Z = \frac{Y - \mu}{\sigma} \;;$$

we obtain a new variable *Z* with mean zero and variance 1. This also holds true for a normal variable; in fact, we obtain an even better result by such a transformation, as follows:

THEOREM: *If X is normally distributed with mean* μ and σ^2, *the transformation* $Z = \dfrac{X - \mu}{\sigma}$ *results in a variable Z which is also normally distributed, but with mean zero and variance* 1*; that is:*

Given: $X \sim N(\mu,\, \sigma^2)$

Transformation: $Z = \dfrac{X - \mu}{\sigma}$

Result: $Z \sim N(0,\, 1)$ [2]

In other words: $f(z) = \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{z^2}{2} \right)$ [3]

Geometrically, this transformation is a conversion the basic scale of *x* values in order that we measure on a standard scale with mean value corresponding to μ and with a measurement of 1 standard deviation. In other words, the standardised normal variable represent the **measurements in the numbers of standard deviation units above or below the mean**. (Figure 5)

This result is not to be taken lightly - it is very important result. For many types of probability distribution functions, analogous results can also be held. In fact, whatever the distribution of a random variable *X* - normal or non-normal, continuous or discrete - the z-transformation will simplify to the transformed variable to have a zero mean and unit variance.

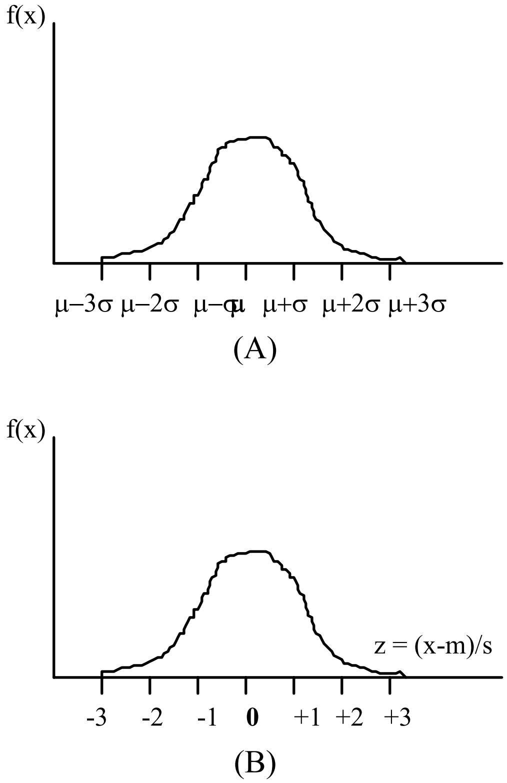


Figure 5 (A) Normal random variable with original scale and (B) its corresponding standardised normal variable with scale as the number of standard deviation units.

III. The Use of Tables for The Standard Normal Distribution

If $Z \sim (0, 1)$, then we have the following results:

(a) the area under the curve (AUC) between points located 1 standard deviation (SD) in each direction from the mean is 0.6826.

(b) the AUC between points located 2 SD in each direction from the mean is 0.9546;

(c) the AUC between points located 3 SD in each direction from the mean is 0.9974

These results are shown in Figure 6.

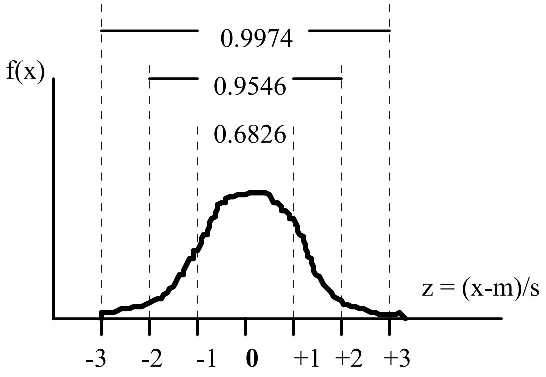


Figure 6: Area under the standardised normal distribution curve

The probabilities (AUC) for various values of *z* are tabulated in several statistical texts. I reproduce here one of such table for your reference and working purpose. In the following examples (and exercises), use of this Table is required.

DETERMINING PROBABILITIES

Example 2: Use the table of the normal distribution to find the following probabilities:

(a) $P(z < 1.75)$	(b) $P(z < -2.76)$	(c) $P(z > -1.15)$
(d) $P(0.78 < z < 1.32)$	(e) $P(-1.18 < z < 1.46)$	(f) $P(-1.56 < z < -0.68)$

Answer: (a) $P(z < 1.75) = 0.9599$.

(b) $P(z < -2.76) = 0.0029$.

(c) $P(z > -1.15) = 1 - P(z \leq 1.15) = 1 - 0.1251 = 0.8749$.

(d) $P(0.78 < z < 1.32) = P(z < 1.32) - \mathrm{P}(z < 0.78)$

$= 0.9066 - 0.7823$

$= 0.1243$.

(e) $P(-1.18 < z < 1.46) = P(z < 1.46) - \mathrm{P}(z < -1.18)$

$= 0.9278 - 0.1190 = 0.8088$.

(f) $P(-1.56 < z < -0.68) = P(z < -0.68) - P(z < -1.56)$

$= 0.2482 - 0.0594 = 0.1888$.

Example 3: The mean and standard deviation of lumbar spine BMD (among elderly women) in a community is 1.026 g/cm^2 and 0.19 g^2/cm^4, respectively.

(a) What is the probability that a woman selected randomly from this community would have a BMD less than 0.9 g/cm^2.

(b) If 100 women are to be selected from this community, how many women would have BMD (i) less than 0.9 g/cm^2 or greater than 1.1 g/cm^2;

(ii) between 0.8 g/cm^2 and 1.20 g/cm^2.

In order to answer these questions, we need to use the standardised normal distribution (eg z-transformation). Now the $Z = (x - \mu)/\sigma$ for question (a) would be $Z = (0.9 - 1.026)/0.19$ = -0.66, therefore:

P(LSBMD < 0.9) = $P(Z < -0.66)$ = 0.2546 or 25.46%.

(See Figure 7A)

(b) Similarly P(LSBMD > 1.1) = $P(Z > 0.39)$

= 1 - $P(Z < 0.39)$

= 1 - 0.652 = 0.348 or 34.8%.

So the probability that lumbar spine BMD less than 0.9 g/cm^2 or greater than 1.1g/cm^2 is the sum of 25.4 + 34.8 = 60.2%; it follows that if 100 women were selected, 60 women would have BMD in the range (Figure 7B).

Part (ii) of question (b), by using the standardised normal distribution, we have:

P(LSBMD>0.8) = $P(Z > -1.19)$

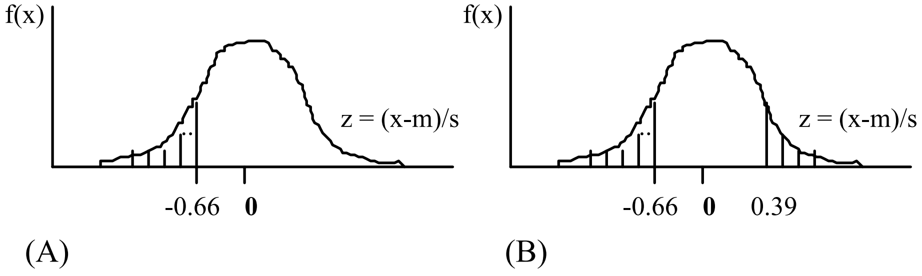
= 1 - $P(Z < -1.19)$

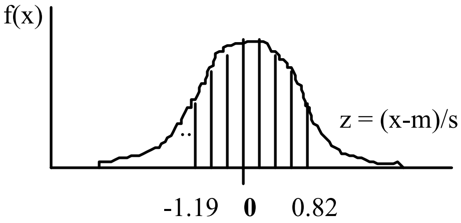
= 1 - 0.117

= 0.883

and P(LSBMD<1.2) = P(Z < 0.92) = 0.179,

then, the probability that LSBMD lies between 0.8 g/cm^2 and 1.20 g/cm^2 is simply 0.883 - 0.179 = 0.704 or 70.4%. In 100 randomly selected women, we would expect to see 70 women with BMD in this range (Figure 7C). //





(C)

Figure 7 Shaded are represent the probability that (A) P(Z<-0.66), (B) P(Z<-0.66 or Z>0.39) and (C) P(-1.19 < Z < 0.82).

DETERMINING THE PERCENTILES.

Example 4: Suppose that the mean and variance of BMD is 1.026 g/cm^2 and 0.19 g^2/cm^4, respectively. What is the 1st and 99th percentiles of BMD?

We can use the Table of the Standardised Normal Distribution (SND) to solve this problem. We see from this table that the 99th percentile of the SND is $z(0.99) = 2.33$ and $z(0.01) = -2.33$. (Note that these numbers are only approximate, the actual numbers are 2.326 and -2.326, respectively, but for now it is sufficient for our purpose). What this means is that the BMD limits are therefore located 2.33 standard deviation on either side of the mean, i.e. at the BMD:

$$1.026 - 2.33(\sqrt{0.19}) = 0.01 \text{ g/cm}^2$$

and $$1.026 + 2.33(\sqrt{0.19}) = 2.04 \text{ g/cm}^2 .$$

In other words, $P(0.01 \leq \text{BMD} \leq 2.04) = 0.98$. (Figure 8)

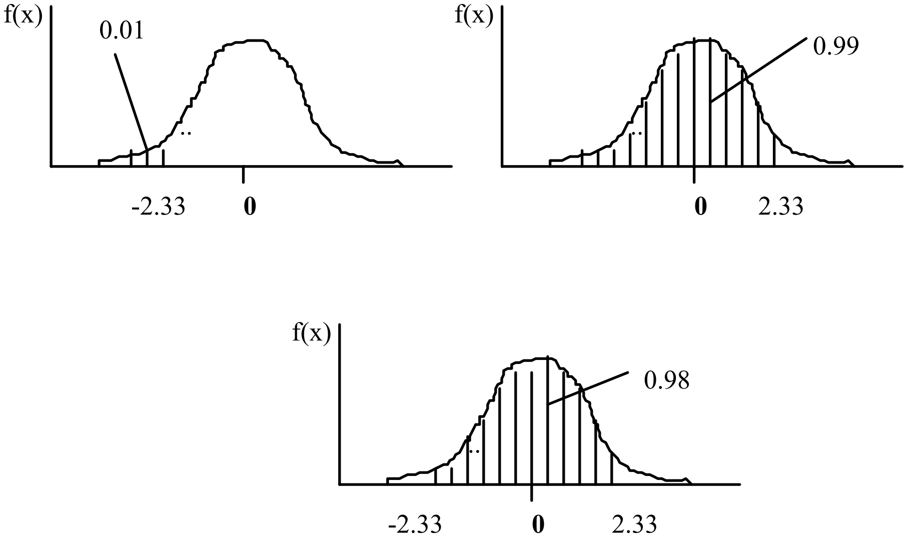


Figure 8.

We mentioned earlier that these are only approximation, the actual values can be more accurately computed. Listed below are exact values of *z* for some common percentiles:

SELECTED PERCENTILES:
Entry is $z(a)$ where $P[Z < z(a)] = a$

a	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.001**
$z(a)$	-1.282	-1.645	-1.960	-2.054	-2.326	-2.576	-3.090
***a*:**	**0.90**	**0.95**	**0.975**	**0.98**	**0.99**	**0.995**	**0.999**
$z(a)$	1.282	1.645	1.960	2.054	2.326	2.576	3.090

IV. THE CENTRAL LIMIT THEOREM AND THE EXACT DISTRIBUTION OF $\overline{X}$.

Some of the most important properties which make much of statistical inference possible are expressed in the central limit theorem (CLT). This section discusses the meaning and implications of this great theorem.

Most of the statistical inference and estimation are techniques are based on the normal distribution. However, since the samples used in these techniques are taken from the real world, they have a distribution far from normal. The CLT allows us to use normal distribution theory to infer about the population from a nonnormal sampling distribution. To do this, we work with the mean of sample data, not the individual values.

The CLT may be stated as follows:

The population may have any unknown distribution with a mean μ and a finite variance σ^2. Take sample of size *n* from the population. As the size of *n* increases, the distribution of sample means will approach a normal distribution with mean μ and a finite variance σ^2/n. .

Because the mathematical proof for this statement is quite "heavy", we adopt a procedural approach to illustrate the theorem. Assume there is a population X which has some distribution with mean μ and variance σ^2. The CLT may be illustrated by the following steps:

(a) Determine n;

(b) Take a random sample of size n and calculate the sample mean $\bar{x}$;

(c) Plot $\bar{x}$ on a histogram of $\bar{x}$ values;

(d) Repeat steps (b) and (c) for k samples;

(e) Calculate the mean and standard deviation of the $\bar{x}$ histogram. Call these $\bar{\bar{x}}$ and $s_{\bar{x}}$;

(f) Compare $\bar{\bar{x}}$ and $s_{\bar{x}}$ with μ and $\sigma/\sqrt{n}$;

(g) Determine a larger n value and repeat steps (b) to (f);

(h) Compare the shapes of the $\bar{x}$ histogram to notice the tendency toward a normal distribution.

(See also Figure 8)

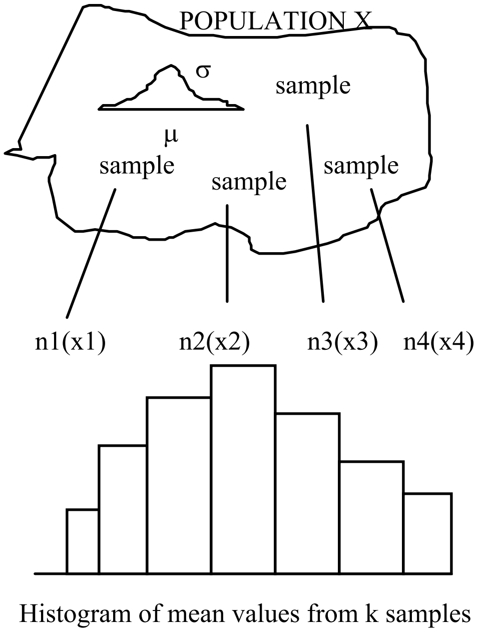


Figure 8: The CLT is illustrated by taking samples of size *n* and plotting means to observe the tendency toward the normal probability distribution function.

Several researchers mistakenly understand that the CLT theorem will apply in any data set with significant size. This is not true. The most important thing to remember when using the results of the CLT us that *we are working with the distribution of sample means,* $\bar{x}$*, not the original X population.* The standard normal distribution transformation is used with $\mu = \bar{x}$ and $\sigma_{\bar{x}} = \sigma / \sqrt{n}$. The form is: $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$.

The Distribution of $\bar{x}$.

In practice, the CLT means that if we have a population with mean μ and variance σ^2, and that we randomly select a sample of *n* subjects from this population and find the mean and standard deviation of this sample to be $\bar{x}$ and *s*, then it could be reasoned that the mean and variance $\bar{x}$ (not *X*) are:

mean of $\bar{x} = \mu$
and variance of $\bar{x} = s^2 / n$
i.e. S.D of $\bar{x} = s / \sqrt{n}$.

This relation may be used either to calculate probabilities for observed mean values or to determine the required sample size such that the observed $\bar{x}$ is within a specified range around the true population mean μ.

Example 6: Suppose that a paediatric population in which systolic blood pressure was normally distributed with mean μ = 115 and variance σ^2 = 15. If a random sample of size 25 is selected from this population, find P(110 < $\bar{x}$ < 120), where $\bar{x}$ is the sample mean.

According to the CLT, the sample mean $\bar{x}$ is normally distributed with mean 115 and standard deviation of $\sigma / \sqrt{n} = 15 / \sqrt{25} = 3$. The z-value corresponding to 110 and 120 are -1.67 and +1.67, respectively. The required probability is 0.9051. //

V. Applications of The Normal Distribution.

(A) Test of hypothesis

(a) We are now using the normal distribution theory to tackle two questions in Example 1. In question (a) we are given "population" mean and standard deviation of BMD of Mrs W as 1.1 g/cm^2 and 0.07 g/cm^2, respectively. Since BMD is normally distributed, under normal circumstances, we would expect that 95% of the times, her BMD would lie between (1.1 - 0.07×2 =) 0.96 g/cm^2 and (1.1 + 0.07×2 =) 1.24 g/cm^2. Therefore, a measurement of 1.05 g/cm^2 lies well within this expected range. Put it other way, a BMD of 1.05 is equivalent to a z value of $(1.05-1.10)/0.07=-0.71$; hence, the probability that her BMD is less than 1.05 g/cm^2 is equivalent to P(Z < -0.71) which is equal to 0.24. That is, there is a 24% chance that her BMD would be less than 1.05 g/cm^2, so from a statistical viewpoint, it may be not necessary to put her on a drug treatment.

(b) In question (b), if the treatment had no effect, then we would expect the BMD in the two occasions would be similar, i.e. the difference would be centred around 0. However, The mean baseline BMD for Mrs P is:

$$\bar{x}_1=\frac{0.95+0.93+0.97}{3}=0.95\text{ g/cm}^2$$

and her follow-up mean is:

$$\bar{x}_2=\frac{1.02+1.05+1.1+1.03}{4}=1.05\text{ g/cm}^2$$

So, an improvement of 1.05 - 0.95 = 0.10 g/cm^2 was observed. Now, BMD measurements are subject to random errors, it is reasonable to ask whether this is a real improvement or just due to chance. If the former is true case, we probably would advise her to continue with the treatment; however if the latter is the case, then a discontinuation of treatment would probably be appropriate.

In Topic 2, we mentioned briefly a general idea that $\bar{x}_1$ and $\bar{x}_2$ are two means of size n_1 and n_2 , respectively, from populations with means μ_1 and μ_2 and standard deviation σ_1 and σ_2, then: $\bar{x}_1$ - $\bar{x}_2$ is approximately normally distributed with mean μ_1 - μ_2 and standard deviation $\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$. If $\sigma_1^2=\sigma_2^2=\sigma^2$ then this reduces to $\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$.

In our problem the baseline and follow-up measurements could be considered as x_1 and x_2. We already see that $\bar{x}_1=0.95$ g/cm^2 and $\bar{x}_2=1.05$ g/cm^2. We could assume that the variance of two occasions are the same, so we could estimate the pooled variance as follows:

$$s_p^2 = \frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

$$= \frac{\left[(1.02 - 1.05)^2 + ... + (1.03 - 1.05)^2\right] + \left[(0.95 - 0.95)^2 + ... + (0.97 - 0.95)^2\right]}{4 + 3 - 2}$$

$$= 0.00092$$

and the standard deviation of the difference is:

$$s = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{0.00092\left(\frac{1}{4} + \frac{1}{3}\right)}$$

$$= 0.023$$

Under the theory of the normal distribution, the probability that there is a 95% chance that her true improvement in BMD varies between 0.1-0.023(2) = 0.054 g/cm^2 to 0.1+0.023(2) = 0.146 g/cm^2. We note that 0 is not in the interval, so it is unlikely that the improvement of 0.10 g/cm^2 was due to chance. This means that we are confident that Mrs P's BMD has been improved significantly. She should probably be advised to continue with the treatment.

We will return to deal with this kind of tests in a later topic.

(B) THE NORMAL APPROXIMATION TO BINOMIAL DISTRIBUTION

The normal distribution is an exact distribution for continuous data which can take on any value from $-\infty$ to $+\infty$. Since not many problems can assume all these values (especially not below 0) most uses are approximations to other discrete or continuous variables. The most common is the normal approximation to the discrete binomial. It can be shown (by De Moivre in 1733) that if X~$B(x; n, p)$; that is:

mean $\mu = np$
and variance $\sigma^2 = npq$ (i.e. standard deviation = $\sqrt{npq}$),

then the variable

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{npq}}$$

has a limit of the standardised normal distribution (SND) as *n* increases. Thus, *Z*~*N*(0, 1). In other words, the binomial asymptotically approaches the SND as *n* increases. The approximation is very accurate when *p* is close to 0.5 because of the symmetry of the binomial distribution. As *p* deviates from 0.5, *n* must be larger for good approximation.

Since there is an asymptotic relation between the binomial and Poisson distributions (Topic 4) and between the binomial and normal distributions, there is one between the Poisson and normal distribution. If *X* is a Poisson variable with mean and variance equal to λ, the transformation $Z = \frac{X - \lambda}{\sqrt{\lambda}}$ is approximately a SND.

<u>Example 5</u>: The rate of operative complications in a vascular surgery is 20%. This includes all complications ranging from wound separation of infection to death. In a series of 50 such procedures, what is the probability that there will be at most 5 patients with operative complication ?

We assume that there is no systematic variation in the pattern of occurrence and non-occurrence of complications. Then for 50 procedures we would expect to have a mean of $50 \times 0.2 = 10$ complications with variance $50 \times 0.2 \times (1 - 0.2) = 8$, i.e. standard deviation $\sqrt{8} = 2.8284$.

Now the probability that there will be at most 5 patients with complication (P($X \leq 5$)) can be found be using the z-transformation:

$$z = \frac{X - \mu}{\sigma} = \frac{5 - 10}{2.8284} = -1.59$$

So: P($X \leq 5$) = P(z $\leq$ -1.59)
= 0.0559 or 5.6%.

whereas the exact value (by using the binomial probability formula) is:

$$P(X \leq 5) = \sum_{x=0}^{5} C_x^{50} (0.2)^{x} (0.8)^{50-x} = \qquad //$$

VI. How to Fit a Normal Distribution

Example 6: Suppose that we have a set of data on weight from a group of 195 students as follows:

Weight (Interval)	Midpoint	No. of students (Frequency)
62-63	62.5	2
64-65	64.5	16
66-67	66.5	30
68-69	68.5	48
70-71	70.5	48
72-73	72.5	39
74-75	74.5	11
76.7	76.5	1

Is the distribution of weight in this group of students normally distributed ?

The question is simple, yet the answer requires somewhat laborious solution. The idea is that to know whether the distribution is normal, we need to calculated the **expected frequencies** of the number of subjects under the hypothesis of the normal distribution. If the expected frequencies do not differ significantly from the observed frequencies, then it is reasonable to conclude that the data are normally distributed; otherwise, not normally distributed.

Now, the mean weight calculated from the grouped data is 69.47 kg and the standard deviation (SD) is 2.8164 kg. In order to calculate the expected frequencies for the normal distribution with this mean and SD, we need to determine the area or probability under the normal curve for each interval (by using the midpoint); this probability is present in column 4 of the following table. The expected number of students in each interval is then equal to the product of this probability and the sample size (n=195); the expected frequencies are given in column 5 of the table below.

Weight	Midpoint	*z*	P(*z*<*x*)	**No. of students**	
(Interval)	(*x*)	value		Expected	Observed
(1)	(2)	(3)	(4)	(5)	(6)
62-63	62.5	-2.12	0.017	3	2
64-65	64.5	-1.41	0.062	12	16
66-67	66.5	-0.70	0.162	32	30
68-69	68.5	0.01	0.262	51	48
70-71	70.5	0.72	0.260	50	48
72-73	72.5	1.43	0.159	31	39
74-75	74.5	2.14	0.060	12	11
76.7	76.5		0.016	3	1

As can be seen from this table, there is a close agreement between observed and expected frequencies. There is a formal test whether the differences are statistically significant, which we will introduce in the next few topics, however, for now it is reasonable to conclude that the data are normally distributed.

VII. NORMAL- RELATED DISTRIBUTIONS

In the last few sections, we have been primarily concerned with using the standard normal distribution - mainly because we needed to make probability statements about the sample mean, set of confidence intervals, and test hypotheses about the sample mean when the variance is assumed to be known. Primarily because of the CLT, we have used the sample mean as our basic sample statistic.

Now, many times, we wish to make probability statements about a statistic, construct confidence intervals, and test hypotheses concerning a parameter by using a statistic for which we must know the sampling distribution. Generally, when we must construct a confidence interval for or test a hypothesis about an unknown parameter we must find an appropriate pivotal quantity; a primary requirement for such an entry is that we must know the characteristics of a distribution.

In this section, we only learn about the relationship between the normal distribution and its related distributions such as the Chi square, F, and t distributions - we will not dwell into the theory or examples these distributions.

(A) THE CHI SQUARE DISTRIBUTION.

In Example 6, we remarked that the observed and expected frequencies distribution of weight in 195 students was quite close and hence justifies for a conclusion of normal distribution of weight. We did this without any formal test. Chi square (χ^2) distribution can be used for such a test. In fact, χ^2 is one of the most important distributions in statistics. It can also be used for conducting tests of independence and set confidence interval for the variance of a normal population, which we will explore in a next topic.

DEFINITION: *Given a sequence of k independent random variables* $Z_1, Z_2, ..., Z_k$ *such that each is normally distributed with mean zero and variance of 1, we define the chi square variable with k degrees of freedom as* $U = Z_1^2 + Z_2^2 + ... + Z_k^2$ *and write* $U \sim \chi_k^2$.

In other words, a chi square variable with *k* degrees of freedom is the sum of squares of *k* independent standard normal variables.

What do we mean by **degrees of freedom (df)**? A rather strict interpretation is that the number of df associated with a chi square variable is the number of independent (standard normal) random variables that conceptually go into the make-up of the variable. For a more intuitive understanding of the term, let us compare two ways of estimating the variance of a population by taking a sample of size *n* - first when we know the value of the population mean μ, and then when we do not know μ.

In the first instance, we estimate the variance by $\sum_{i=1}^{n} (x_i - \mu)^2 / n$; here, the *n* terms $x_i - \mu$ are all independent, hence each makes an independent contribution to the estimation of the variance. Thus we do not lose any degrees of freedom in estimating the variance.

In the second instance, we do not μ, we must replace it by the sample mean $\bar{x}$ and estimate the variance by $\sum_{i=1}^{n}(x_i - \bar{x})^2 / n$. Now recall that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. This means that the n terms $x_i - \bar{x}$ are not independent because, as soon as we know n-1 of the terms, the value of the remaining term is fixed. This fact, resulting from our use of an estimate of μ (which is $\bar{x}$) rather than μ itself, causes us to lose one degree of freedom in estimating the variance. Ultimately, we will see that, in the general problem of estimation, we lose a df for each parameter that is replaced by a sample estimate.

Conceptually, the Chi square distribution with k df could be generated as follows:

(a) Take one observation from each of the k independent standard normal distributed samples: $z_1, z_2, \ldots, z_{ik}$

(b) Square each observation and compute a single observation from a chi square distribution as: $U_i = Z_{i1}^2 + Z_{i2}^2 + \ldots. + Z_{ik}^2$

(c) Repeat steps (a) and (b) for an infinite number of samples, that is, for i = 1, 2, . . . ∝

(d) Compile the probability distribution of the U_i. The result will be the probability distribution of U, a chi square variable with k df.

Consider the following problem: we have a series of values x_i, i = 1, 2, . . ., n, with sample mean $\bar{x}$ and variance s^2. We know that variance of this whole population (in which the sample was drawn from) is σ^2. It is interesting to see that:

$$\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$$

Since $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, therefore the above expression becomes:

$$\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2 = \frac{ns^2}{\sigma^2}$$

But the unbiased estimate of σ^2 is $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$, hence:

$$\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2 = \frac{ns^2}{\sigma^2} = \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \qquad [5]$$

This variable is distributed according to the Chi square distribution with n-1 df.

This important result shows that if we know $\hat{\sigma}^2$ (the estimate sample variance) then we can use the Chi square distribution to test whether $\hat{\sigma}^2$ is equal to a population variance σ^2.

Example 7: A sample of 10 subjects show that the variance of lumbar spine BMD is 0.19 g^2/cm^4. It was however known that the variance of LSBMD in the general population was 0.15 g^2/cm^4. Is there evidence that the sample was biased?

Using [5], we have U = $\frac{9 \times 0.19}{0.15}$ = 11.4. Now at the significance level of 5% and 9 df, we would expect the chi square value to be 16.92. The observed value of 11.4 is well below this critical value, we therefore have reasonable evidence to believe that there was no bias in the sampling scheme. //

(B) THE F DISTRIBUTION.

We are concerned here with another important distribution which was named after an eminent statistician Sir Ronald A. Fisher - the F distribution.

DEFINITION: *If U and V are independently distributed chi-square variables with m and n degrees of freedom (df), respectively, then the ratio* $W = \frac{U/m}{V/n}$ *is distributed according to the F distribution with m and n df.*

Conceptually, an F distribution with m and n df would result if we were able to perform the following processes:

(a) Take one observation (say u_i) from the variable U and one observation (v_i) from the variable V;

(b) Compute a single observation from an F distribution with m and n df as: $w_i = \frac{u_i / m}{v_i / n}$.

(c) Repeat steps (a) and (b) for an infinite number of samples (i = 1, 2, . . . , ∞)

(d) Compile the probability distribution of the w_i. The result is the probability distribution of W, an F distribution with *m* and *n* df.

If *X* follows an F distribution with *m* and *n* df, it is symbolically written as: $X \sim F_{m,n}$. Mathematically, it can be shown that if $X \sim F_{m,n}$, then $\frac{1}{X} \sim F_{n,m}$.

In the previous section we stated that if U and V are independently distributed Chi square variables with $n_1 - 1$ and $n_2 - 1$ df, respectively, then:

$$U = \frac{\sum_{i=1}^{n_1}\left(X_{1j} - \overline{X}_1\right)^2}{\sigma_1^2} \sim \chi^2_{n_1 - 1}$$

and $$V = \frac{\sum_{i=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2}{\sigma_2^2} \sim \chi^2_{n_2 - 1}$$

Now, let $m = n_1 - 1$ and $n = n_2 - 1$, according to the definition of the F distribution, we have:

$$\frac{U/m}{V/n} = \frac{\frac{\sum_{i=1}^{n_1}\left(X_{1j} - \overline{X}_1\right)^2}{\sigma_1^2} / \left(n_1 - 1\right)}{\frac{\sum_{i=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2}{\sigma_2^2} / \left(n_2 - 1\right)} \sim F_{n_1 - 1, n_2 - 1}.$$

Rearranging the right-hand term and substituting the sample values for two specific samples, we obtain the formula for computing an observed value of the above statistic, that is:

$$\frac{\frac{\sum_{i=1}^{n_1}\left(X_{1j} - \overline{X}_1\right)^2}{\sigma_1^2\left(n_1 - 1\right)}}{\frac{\sum_{i=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2}{\sigma_2^2\left(n_2 - 1\right)}} = \frac{\hat{\sigma}_1^2 / \sigma_1^2}{\hat{\sigma}_2^2 / \sigma_2^2} \qquad [6]$$

where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the unbiased estimates of the population variances for population 1 and 2, respectively. Thus, [6] is a function of σ_1^2 and σ_2^2 (the unknown variances). The distribution however holds regardless of the true values of σ_1^2 and σ_2^2 . Therefore, under the unique condition (and only such condition) that $\sigma_1^2 = \sigma_2^2$, [5] can be written as:

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \qquad [7]$$

This result ([7]) is often used to test for the equality of two variances.

Example 8: A sample of 10 subjects show that the variances of lumbar spine and femoral neck BMD are 0.19 g^2/cm^4 and 0.12 g^2/cm^4. Is there evidence that the two variances are different ?

We use the F statistic: F = 0.19/0.12 = 1.58, now this statistic is distributed with 9 numerator df and 9 denominator df . The critical value at 5% level for $F_{9,9} = 3.18$. Since the observed F value is below the expected value (of 3.18), we conclude that there is evidence suggesting the equality of two variances.

(C) THE T DISTRIBUTION.

In most of the discussions so far, we have assumed that either the mean or the variance of a variable is known. If, however, either of the above assumptions is not satisfied, we must determine other ways of making probability statements. We can determine what happens when one assumption is met and the other is not. This is precisely what was done by W. S. Gossett, a statistician who, while working for a tobacco company in England, wrote under the pseudonym "Student".

Gossett derived the exact distribution of the statistic

$$\frac{\overline{X} - \mu}{\hat{\sigma} / \sqrt{n}} = \frac{\overline{X} - \mu}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}$$

for situations in which a sample of any size *n* is selected from a normal population having an unknown variance. This distribution is also known as the "Student's distribution".

DEFINITION: *If Z and U are independent random variables such that Z is distributed normally with mean 0 and variance 1, and U is distributed according to the Chi square distribution with k df, then the statistic* $W = Z / \sqrt{U / k}$ *is distributed according to the t distribution with k df.*

Conceptually, an F distribution with *m* and *n* df would result if we were able to perform the following processes:

(a) Take one observation (say z_i) from the variable Z and one observation (u_i) from the variable U;

(b) Compute a single observation from a t distribution with *k*a df as: $w_i = z_i / \sqrt{u_i / k}$.

(c) Repeat steps (a) and (b) for an infinite number of samples (i = 1, 2, . . . , ∝)

(d) Compile the probability distribution of the w_i. The result is the probability distribution of W, a t distribution with *k* df.

In sample statistic, we could infer from the above definition: If $\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}$ and $\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$, then:

$$\frac{\dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}}{\sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}\left(\dfrac{X_i - \overline{X}}{\sigma}\right)^2}} \sim t_{n-1}$$

This formula can be simplied to obtain:

$$\frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}\dfrac{1}{\sigma}\sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}} = \frac{\overline{X} - \mu}{\sqrt{\dfrac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \overline{X})^2}} \sim t_{n-1} \qquad [8]$$

This relation provides us immediately with a pivotal quantity for problems involving a normal distributed population with unknown variance. Thus, the essential steps for making a one-tail, 100α percent significance test concerning the mean μ of a normal population with unknown variance can be carried out with sound theoretical background.

We do not give example in this sub-section as we will deal with this distribution extensively in the next topic.

VIII. Exercises

1. The distribution of lumbar spine BMD in a NSW population is as follows: for males, mean = 1.24 g/cm^2 and standard deviation = 0.21 g/cm^2 ; for females, mean = 1.02 g/cm^2 and standard deviation = 0.19 g/cm^2. Write the complete probability distribution function of BMD for males and females.

2. Use the normal probability distribution function in [1] and the idea of function (which you have learned in Topic 1) to determine the value of f(x) for the following cases:
   (a) $\mu = 0$, $\sigma = 0.5$ and $x = 0.5$
   (b) $\mu = -5\sigma = 2$ and $x = -8$
   (c $\mu = 2050 = 158$ and $x = 2130$.

3. Given that Z is a standard normal variable, determine the following probabilities:
   (a) $P(Z \geq 1.78)$ (b) $P(Z \leq 1.25)$
   (c) $P(Z \geq -1.20)$ (d) $P(Z \leq -2.58)$
   (e) $P(1.29 \leq Z \leq 2.15)$ (e) $P(-2.74 \leq Z \leq -1.40)$
   (f) $P(-1.3 \leq Z \leq 1.3)$ (g) $P(-1.45 \leq Z \leq 2.01)$

4. Suppose that weight (denoted by X) of a group of boys is normally distributed with a mean of 44 kg and standard deviation of 5 kg. Find:
   (a) $P(40 < Z < 48)$
   (b) $P(Z < 42)$
   (c) $P(Z > 45)$
   (d) Between what two values does the middle 90% of weights lie ?
   (e) Your son (also in this age group) weighs 38 kg. Should you fear that he is abnormally light and doomed never to become a football player ?

5. For the weight in question 1, a random sample of 10 boys are selected and weighed. Let the sample mean be $\bar{x}$. Find:
   (a) $P(42 < \bar{x} < 46)$
   (b) $P(\bar{x} < 40)$

(c) P($\bar{x}$ > 48)

(d) Between what two values does the middle 95% lies ?

(e) If $\bar{x}$ = 38, would this indicate an unusual sample of boys ?

6. Mr WP is started on treatment. He has the following blood pressures (BP) at his next 4 visits: 86, 92, 82 and 84.

   (a) Assuming that the standard deviation of his blood pressure is 5, about average, compute the 80% and 95% confidence intervals for his mean blood pressure. What is your confidence that his mean BP is below 90 mmHg.

   (b) Use the measurements to estimate his standard deviation (*s*).

   (c) Compute the 80% and 95% confidence limits for his mean blood pressure using *s, n*.

7. Mr WP is followed and his average BP over many visits is 85 mmHg. Suppose that his true standard deviation for individual measurements is 6 mmHg.

   (a) How often would you expect a reading of 95 mmHg or higher ? 100 or higher ?

   (b) On the next visit, his BP is 95 mmHg. How would you settle whether his average BP is no longer below the goal of 90 mmHg ?

8. The probability that an individual with a rare disease will be cured is 1%. A random sample of 600 persons with the disease is selected; find the probability that 1 person is cured, using (a) Binomial distribution theory and (b) Normal approximation.

9. The following statement was found in a popular medical journals: "As the sample size increases, the distribution of the data becomes approximately normal, by virtue of the Central Limit Theorem". Explain what is wrong with the statement?

10. A surgeon wants to conduct a clinical trial to estimate the average time to recovery for patients benefiting from a new therapy for advanced breast cancer. For the standard therapy, the time to recovery is 110 weeks, and the variation among respondents is such that the standard deviation is 24 weeks. How many patients are needed in the trial, if the surgeon is to be 95% confident of estimating the average time to recovery to within 10 weeks? Assume that the variation among patients is comparable to the standard therapy.

11. The acidity of human blood measured on the pH scale is normal random variable with mean 7.2. Determine the standard deviation if the probability that the pH level is greater than 7.47 is 0.0359.