MODULE TWO: MEASURING UNCERTAINTY; AND DRAWING CONCLUSIONS ABOUT POPULATIONS BASED ON SAMPLE DATA

TOPIC 5: CONTINUOUS DISTRIBUTIONS AND SAMPLING DISTRIBUTIONS

Learning Objectives

At the completion of this topic, you should be able to:

- calculate probabilities from the normal distribution
- use a normal probability plot to determine whether a set of data is approximately normally distributed
- calculate probabilities from the uniform distribution
- calculate probabilities from the exponential distribution
- calculate probabilities related to the sample mean
- recognise the importance of the Central Limit Theorem
- calculate probabilities related to the sample proportion

+Continuous Probability Distributions

A continuous random variable is a variable that can assume any value on a continuum (can assume an infinite number of values)

These can potentially take on any value, depending only on the ability to measure accurately:

- thickness of an item
- time required to complete a task
- weight, in grams
- height, in centimetres

+Continuous Probability Distributions

In this Unit, we will explore 3 Continuous Probability Distributions:

- Normal distribution
- Uniform distribution
- Exponential distribution

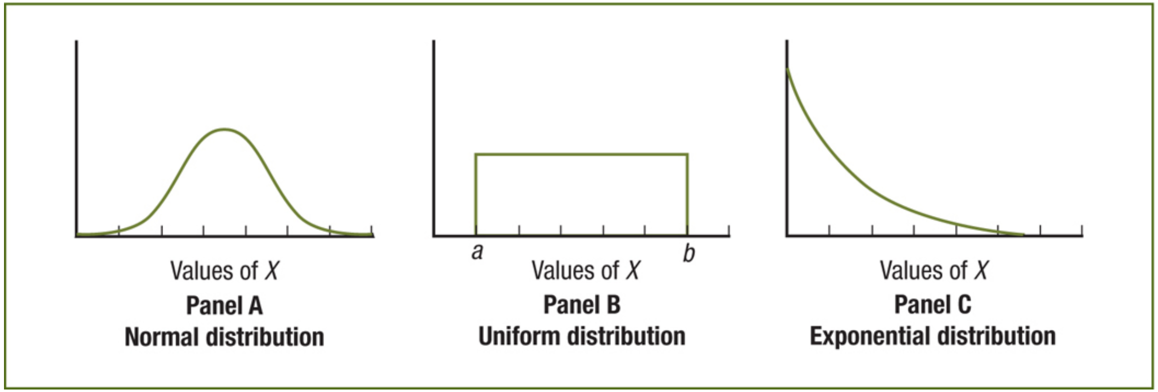
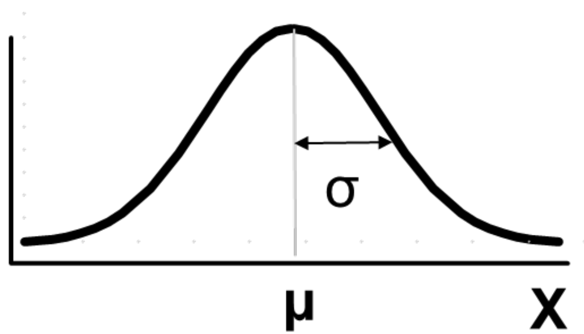


Figure 6.1
Three continuous distributions

+The Normal Distribution (cont)

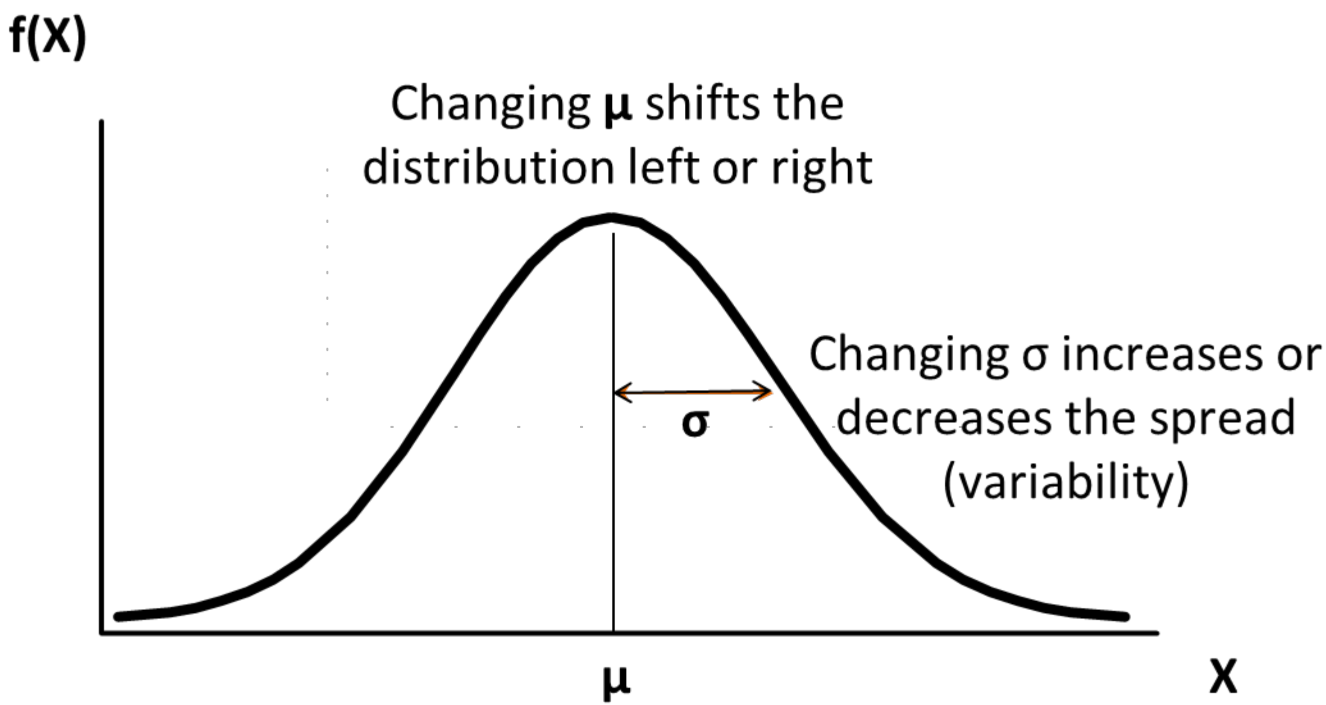
Bell-shaped or **Symmetrical**

- Mean, median and mode are equal
- Central location is determined by the mean, μ
- Spread is determined by the standard deviation, σ
- The random variable X has an infinite theoretical range: + ∞ to - ∞



Mean = Median = Mode

+The Normal Distribution (cont)



+The Normal Distribution (cont)

By varying the parameters μ and σ, we obtain different normal distributions

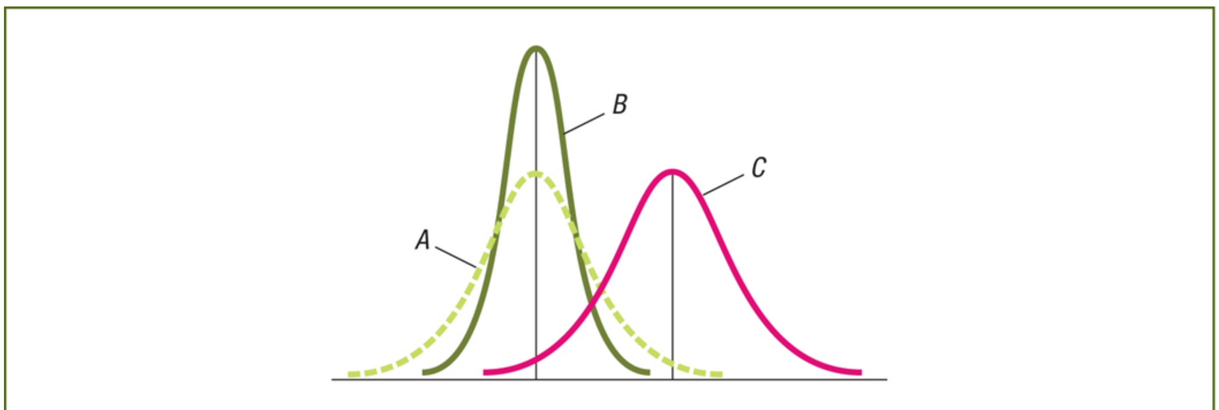


Figure 6.3
Three normal distributions

+The Normal Distribution (cont)

Any normal distribution (with any mean and standard deviation combination) can be transformed into the standardised normal distribution (Z)

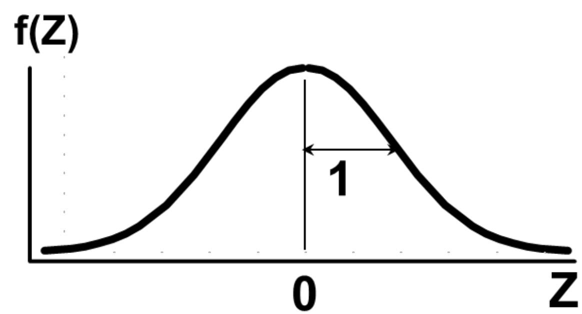
Translate any X to the Standardised Normal (the Z distribution) by subtracting from any particular X value the population mean and dividing by the population standard deviation

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

+The Normal Distribution (cont)

The Standardised Normal Distribution

- Also known as the ‘Z distribution’
- Mean is 0
- Standard deviation is 1



- Values above the mean have positive Z-values
- Values below the mean have negative Z-values

+The Normal Distribution (cont)

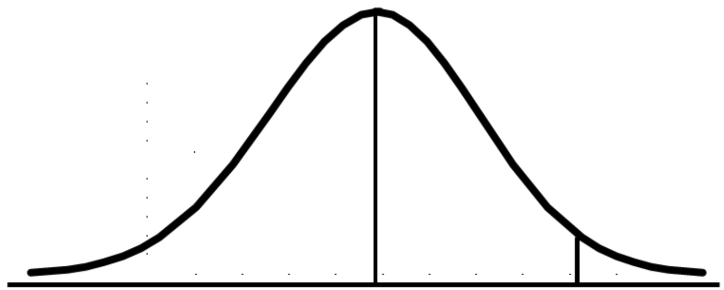
Example:

If X is distributed normally with mean of 100 and standard deviation of 50, the Z value for X = 200 is:

$$Z = \frac{X - \mu}{\sigma} = \frac{200 - 100}{50} = 2.0$$

This says that X = 200 is two standard deviations (2 increments of 50 units) above the mean of 100

+The Normal Distribution (cont)

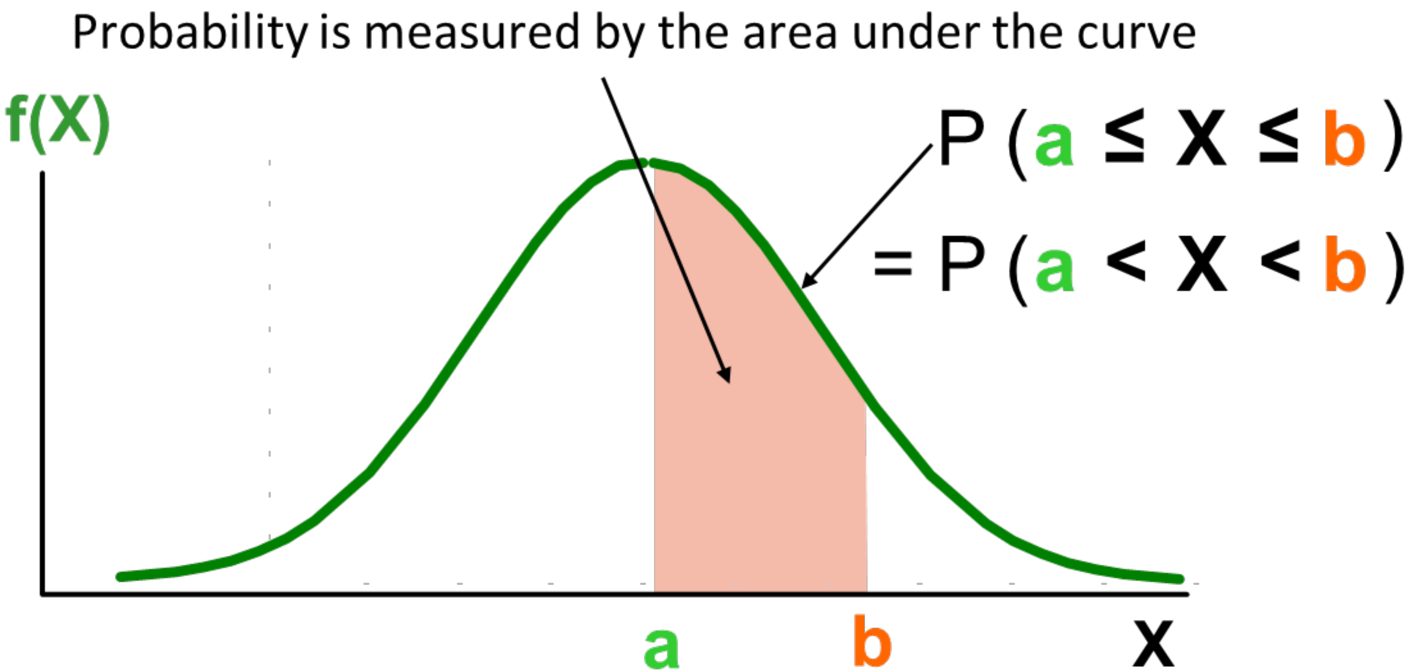


100	200	X	(μ = 100, σ = 50)
0	2.0	Z	(μ = 0, σ = 1)

Note: the distribution is the same - only the scale has changed. We can express the problem in original units (X) or in standardised units (Z)

+The Normal Distribution (cont)

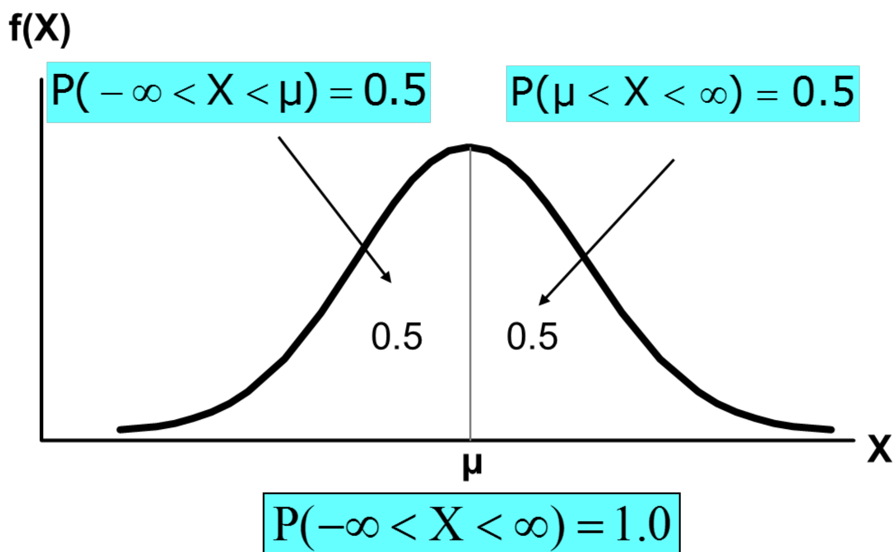
Finding Normal Probabilities



+The Normal Distribution (cont)

Probability as Area Under the Curve

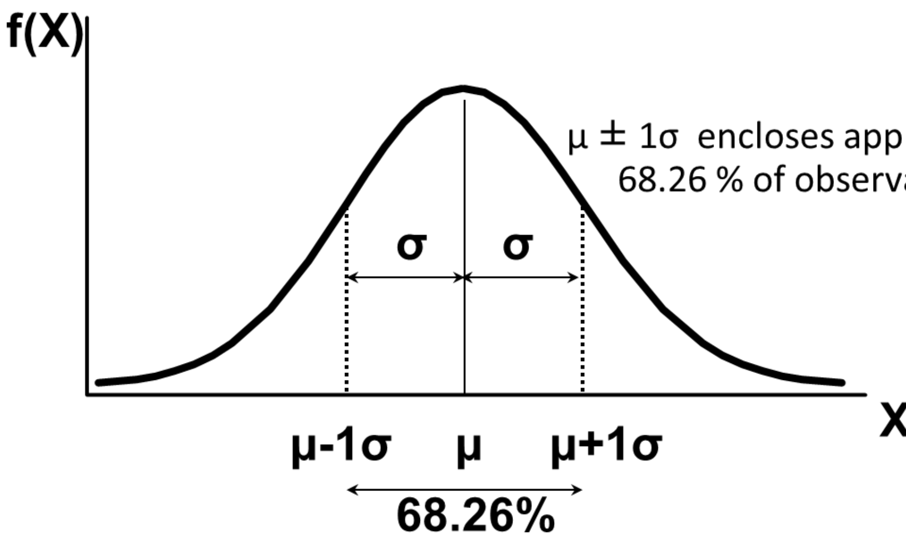
The total area under the curve is 1.0, and the curve is symmetric, so half is above the mean, half is below



+The Normal Distribution (cont)

Empirical Rules

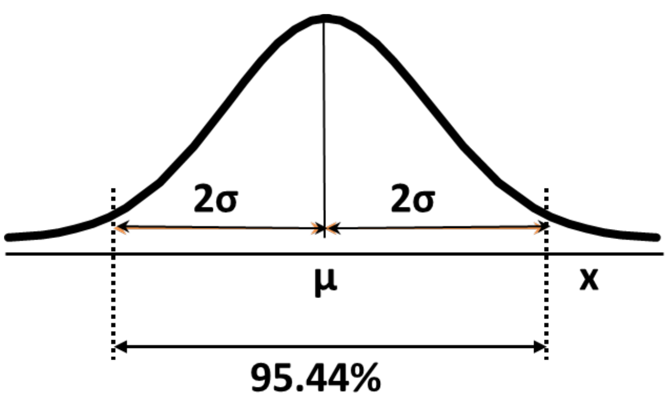
There are some general rules as to what we can say about the distribution of values around the mean

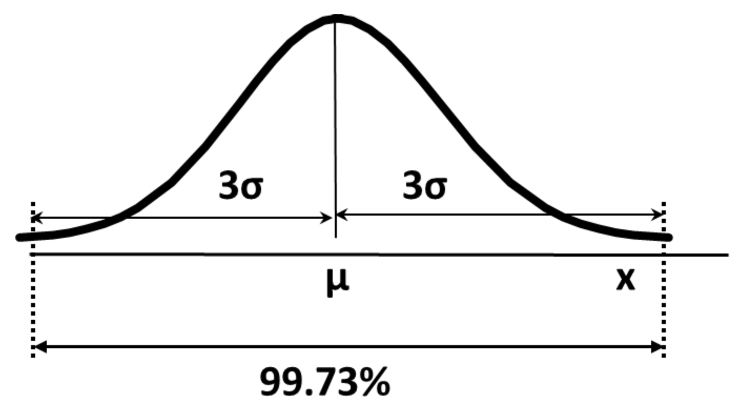


+The Normal Distribution (cont)

Empirical Rules (cont)

- $\mu \pm 2\sigma$ covers approximately 95.44% of observations
- $\mu \pm 3\sigma$ covers approximately 99.73% of observations





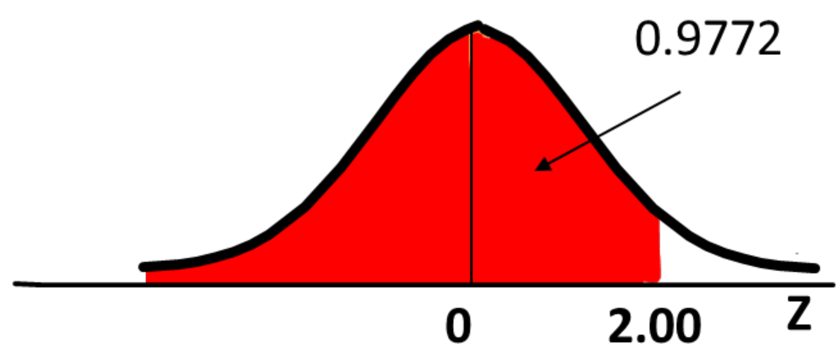
The Normal Distribution (cont)

The Standardised Normal Distribution Table

- The Cumulative Standardised Normal Distribution table in the textbook (Appendix Table E.2) gives the probability less than a desired value for Z
- Once Z<-6, the values listed become so small as to be effectively 0 in area

Example:

P(Z < 2.00) = 0.9772



+The Normal Distribution (cont)

The Standardised Normal Distribution Table

The *column* gives the value of Z to the second decimal point

Z	0.00	0.01	0.02
0.0			
0.1			
.			
.			
.			
2.0	.9772		

The *row* shows the value of Z to the first decimal point

The value within the table gives the probability from Z = +6 down to the desired Z value

P(Z < 2.00) = 0.9772

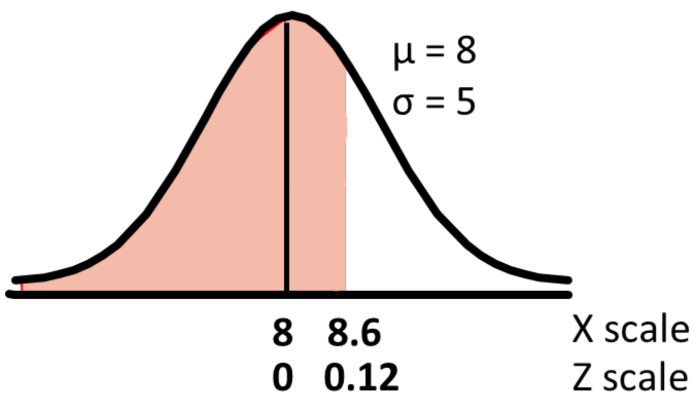
+The Normal Distribution (cont)

Example:

Suppose X is normally distributed with mean 8.0 and standard deviation 5.0. Find $P(X < 8.6)$

$$Z = \frac{X - \mu}{\sigma} = \frac{8.6 - 8.0}{5.0} = 0.12$$

- Therefore, $P(X < 8.6)$ is the same as $P(Z < 0.12)$



Z	.00	.01	.02
0.0	.5000	.5040	.5080
0.1	.5398	.5438	**.5478**
0.2	.5793	.5832	.5871
0.3	.6179	.6217	.6255

+The Normal Distribution (cont)

Finding the X Value for a Known Probability involves 4 steps:

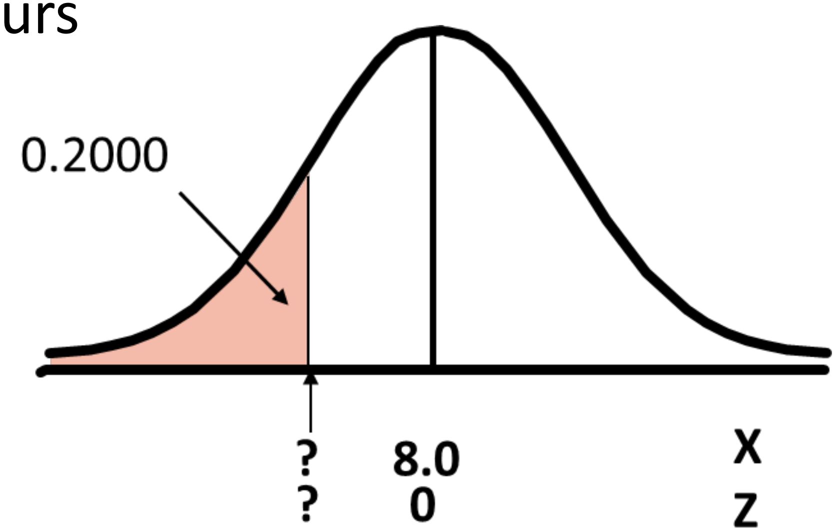
1. Draw a normal curve placing all known values on it such as mean of X and Z
2. Shade in area of interest and find cumulative probability
3. Find the Z value for the known probability
4. Convert to X units using the formula

+The Normal Distribution (cont)

Example:

Suppose X is normally distributed with a Mean of 8 hours and a Standard Deviation of 5 hours. Find the value of X for the lowest 20% of hours

Steps 1 and 2



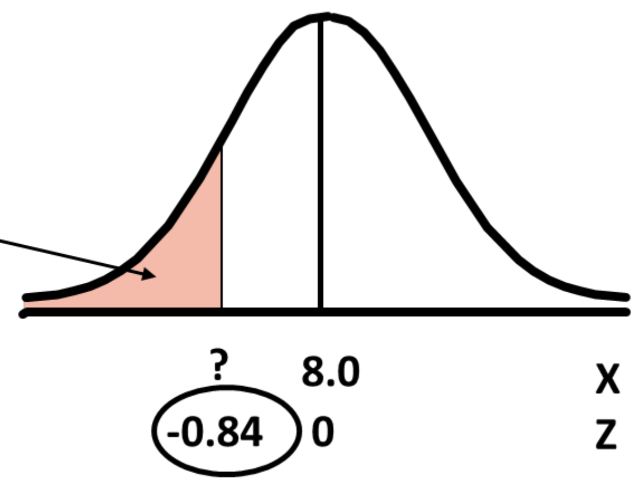
+The Normal Distribution (cont)

Step 3

Table E.2 (Portion)

Z	...	.03	**.04**	.05
-0.9	...	.1762	.1736	.1711
-0.8	...	.2033	**.2005**	.1977
-0.7	...	.2327	.2296	.2266

20% area in the lower tail is consistent with a Z value of **-0.84 (closest)**



The Normal Distribution (cont)

Step 4

The following formula is simply our Z formula rearranged in terms of X

$$\begin{aligned} X &= \mu + Z\sigma \\ &= 8.0 + (-0.84)5.0 \\ &= 3.80 \end{aligned}$$

Note: Z = -0.84 (not +0.84) since we are dealing with the left-hand side of the curve

Answer:

- 20% of the values are less than 3.8 hours

+The Normal Distribution (cont)

Figure 6.18
Microsoft Excel worksheet for calculating normal probabilities

	A	B	
1	**Normal probabilities**		
2			
3	**Common data**		
4	**Mean**	**7**	
5	**Standard deviation**	**2**	
6			
7	**Probability for *X* <=**		
8	***X* value**	**3.5**	
9	***Z* value**	**–1.75**	**=STANDARDIZE(B8, B4, B5)**
10	***P*(*X*<=3.5)**	**0.0401**	**=NORM.DIST(B8, B4, B5, TRUE)**
11			
12	**Find *X* and *Z* given cum. pctage.**		
13	**Cumulative percentage**	**10.00%**	
14	***Z* value**	**–1.2816**	**=NORM.S.INV(B13)**
15	***X* value**	**4.4369**	**=NORM.INV(B13, B4, B5)**

+Evaluating Normality

This section presents two approaches for evaluating whether a set of data can be approximated by the normal distribution:

- Compare the data set's characteristics with the properties of the normal distribution
- Construct a normal probability plot

+Evaluating the Properties

The normal distribution has several important theoretical properties:

- It is symmetrical, thus the mean and median are equal
- It is bell shaped, thus the empirical rule applies
- The interquartile range equals approximately 4/3 standard deviations
- The range is infinite

+Evaluating the Properties (cont)

To check for normality, compare the actual data characteristics with the corresponding properties from an underlying normal distribution, as follows:

Construct charts and observe their appearance

- e.g. a box-and-whisker plot or a frequency distribution and plot the histogram or polygon

+Evaluating the Properties (cont)

Calculate descriptive numerical measures and compare the characteristics of the data with the theoretical properties of the normal distribution

- e.g. Compare the mean and median. Is the interquartile range approx. 1.33 times the standard deviation? Is the range approx. six times the standard deviation?

Evaluate how the values in the data are distributed

- e.g. Do 2/3 of the values lie between the mean ± 1 standard deviation. Do approx. 4/5 of the values lie between the mean ± 1.28 standard deviations. Do approx. 19 of every 20 values lie between the mean ± 2 standard deviations

+Constructing a Normal Probability Plot

A normal probability plot is a graphical approach for evaluating whether data are normally distributed

One common approach is called the quantile-quantile plot. In this method, each ordered value is transformed to a Z score and plotted along with the ordered data values of the variable

- e.g. If you have a sample of, say, n = 19, the Z value for the smallest value corresponds to a cumulative area of $\frac{1}{n+1} = \frac{1}{19+1} = \frac{1}{20}$ = 0.05. The Z value for a cumulative area of 0.05 (from Table E.2) is -1.65

+ Constructing a Normal Probability Plot

Table 6.7
Ordered values and corresponding Z values for a sample of $n = 19$

Ordered value	Z value	Ordered value	Z value
1	−1.65	11	0.13
2	−1.28	12	0.25
3	−1.04	13	0.39
4	−0.84	14	0.52
5	−0.67	15	0.67
6	−0.52	16	0.84
7	−0.39	17	1.04
8	−0.25	18	1.28
9	−0.13	19	1.65
10	0.00		

+Constructing a Normal Probability Plot

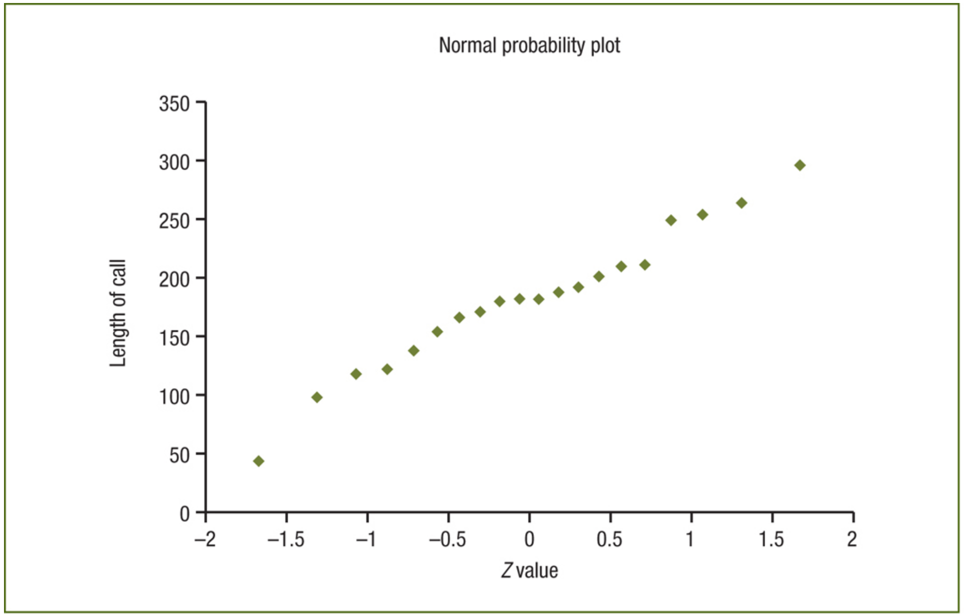


Figure 6.22
PHStat2 normal probability plot for length of call

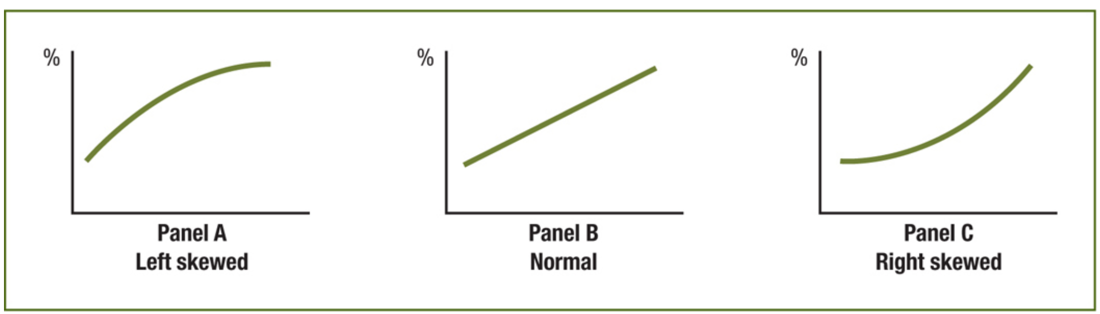


Figure 6.21
Normal probability plots for a left-skewed distribution, a normal distribution and a right-skewed distribution

+The Uniform Distribution

The uniform distribution is a probability distribution that has equal probabilities for all possible outcomes of the random variable

Sometimes also called a rectangular distribution

The Uniform Probability Density Function is:

$$\boxed{f(X) = \frac{1}{(b - a)}}$$ if $a \leq X \leq b$ and 0 elsewhere

where:

f(X) = value of the density function at any X value
a = minimum value of X
b = maximum value of X

+The Uniform Distribution (cont)

The mean of a uniform distribution is:

$$\boxed{\mu = \frac{a + b}{2}}$$

The standard deviation is:

$$\boxed{\sigma = \sqrt{\frac{(b-a)^2}{12}}}$$

+The Exponential Distribution

The exponential distribution is a continuous distribution that is right skewed and ranges from zero to positive infinity

It’s is widely used in waiting line (or queuing) theory to model the length of time between random and independent events, or the time to the first occurrence of an event. For example:

- time between arrivals of customers at a bank's ATM or a fast-food restaurant
- time between patients entering a hospital emergency room
- time between hits on a website
- time between outages to an Internet banking system
- time to failure of a certain item or component

+The Exponential Distribution (cont)

The Exponential and Poisson distributions are closely related

The Poisson distribution is used to count the number of times an event occurs in some interval, while the Exponential distribution is used to measure the interval between Poisson events or until the first event

The exponential distribution is defined by a single parameter, λ, the expected number of events per interval

Note: this is the mean of the corresponding Poisson distribution

+The Exponential Distribution (cont)

Probability that an Exponential Random Variable is less than A

If X is an exponential random variable, $0 \leq X \leq \infty$, then

$$\boxed{P(X < A) = 1 - e^{-\lambda A}}$$

where: λ = expected number of events in interval

e = 2.71828 ... is the base of natural logarithms

A is a given value of the exponential random variable X

+The Exponential Distribution (cont)

Example:

Customers arrive at the service counter at the rate of 15 per hour

What is the probability that the arrival time between consecutive customers is less than three minutes?

The mean number of arrivals per hour is 15, so λ = 15

Three minutes is 0.05 hours, so A = 0.05

$$P(X < .05) = 1 - e^{-\lambda A} = 1 - e^{-(15)(0.05)} = \boxed{0.5276}$$

So there is a 52.76% probability that the arrival time between successive customers is less than three minutes

+The Exponential Distribution (cont)

You can also use Microsoft Excel to calculate this probability The following figure shows a Microsoft Excel worksheet, using the Excel inbuilt exponential function EXPON.DIST(x,lambda,cummulative)

	A	B	
1	Exponential probability		
2			
3	Data		
4	λ	20	
5	*X* value	0.1	
6			
7	Results		
8	*P*(<=*X*)	0.8647	=EXPON.DIST(B5, B4, TRUE)

Figure 6.25 Microsoft Excel worksheet for finding exponential probabilities

+Sampling Distributions

When we take a **Sample**, the attributes of a variable are called Statistics and those for a **Population**, are called Parameters

This hold true for both Numeric (e.g a Mean) and Categorical (e.g a Proportion) variables

Our purpose in taking a sample is to make statistical inferences and draw conclusions about the population, **not** the sample

Hypothetically, to use the sample statistic to estimate the population parameter, we should examine *every* possible sample. A sampling distribution is the distribution of the results if we actually selected all possible samples

Sampling Distribution of the Mean

Summary Measures for the **Population** Distribution

$$\mu = \frac{\sum X_i}{N}$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

Summary Measures of **Sampling** Distribution

$$\mu_{\overline{X}} = \frac{\sum \overline{X}_i}{N}$$

$$\sigma_{\overline{X}} = \sqrt{\frac{\sum (\overline{X}_i - \mu_{\overline{X}})^2}{N}}$$

+Standard Error of the Mean

Different samples of the same size from the same population will yield different sample means

A measure of the variability in the sample mean from sample to sample is given by the Standard Error of the Mean

$$\boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$$

This assumes that sampling is done with replacement or sampling is done without replacement from a large or infinite population

Note: the standard error of the mean decreases as the sample size increases

+Sampling from Non-normally Distributed Populations – The Central Limit Theorem

If the Population is NOT Normal, we can apply the Central Limit Theorem

The CLT states that, as the sample size (i.e. the number of values in each sample) gets large enough, (generally $n \geq 30$), the sampling distribution of the mean is approximately normally distributed

This is true regardless of the shape of the distribution of the individual values in the population

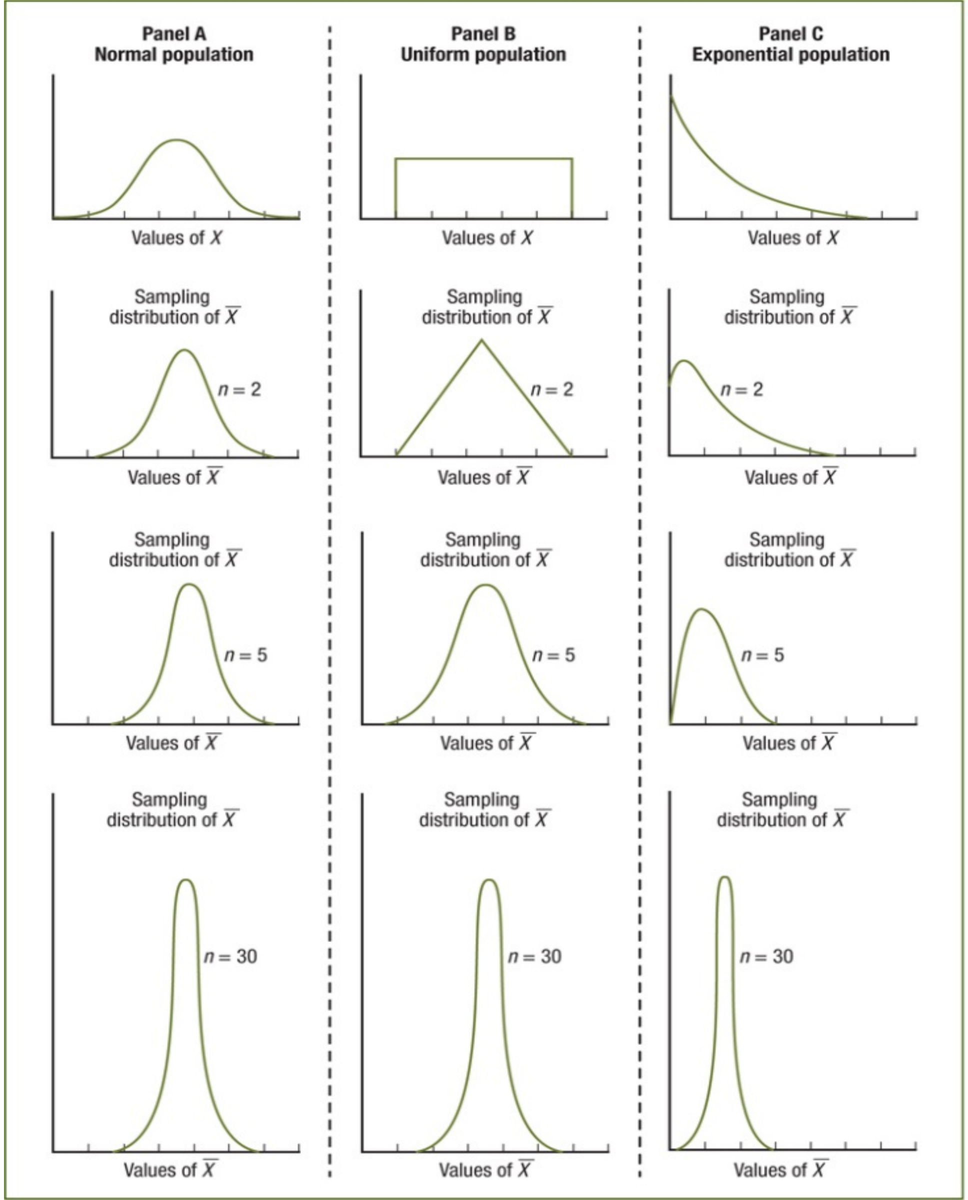


Figure 7.4
Sampling distribution of the mean for different populations for samples of $n = 2$, 5 and 30

+Z Formula for Sampling Distribution

If the population is normal OR the Central Limit Theorem is applicable, then we can use the normal distribution and the Z table to find probabilities for the sample mean

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Where:

$\bar{X}$ = sample mean

μ = population mean

σ = population standard deviation

n = sample size

+Sampling Distribution of the Proportion

π is the proportion of items in the population with a characteristic of interest

p is the **sample proportion** and provides an estimate of π

$$p = \frac{X}{n}$$

$$= \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

+Standard Error of the Proportion

The underlying distribution of the sample proportion is binomial

It can be approximated by a normal distribution if n $\pi \geq 5$ and n(1- π) ≥ 5 with the resulting mean equal to π and standard error equal to:

$$\boxed{\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}}$$

+Z Formula for Proportions

We standardise p to a Z value with the following formula:

$$Z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\dfrac{\pi(1 - \pi)}{n}}}$$