STATS 10X WORKSHOP

EXAM PREP 2: CHAPTERS 4, 6 & 7

SAT 2 & WED 27 OCTOBER 2010

Students **MUST REGISTER** for all workshops with
The Student Learning Centre, 3rd Floor, Information Commons

Statistical help available at the SLC

The Student Learning Centre (SLC) offers help for STATS 10x by offering:

- one-on-one tutoring help, and
- a number of workshops

One-on-one help over S2 2010 including exam period

One-on-one assistance for STATS 10x is available at the SLC. Check appointment availability and book at SLC reception in person (third floor, Information Commons building) or by calling 373-7599 ext. 88850.

Note: SLC tutors are not allowed to help students complete their assignments.

SLC STATS 10x Exam Prep Workshops

Any questions regarding STATS 10x workshops should be forwarded to:

Leila Boyle; SLC Statistics Co-ordinator: l.boyle@auckland.ac.nz

These twelve workshops (six different sessions, each repeated twice) are held prior to the exam, from Saturday 2 October until Monday 1 November 2010 (inclusive).

These workshops concentrate on questions reviewing the **basic concepts**, rather than questions on finer details. They are designed to assist students to achieve a pass and **don't cover all material**.

The timetable for these workshops is available at this workshop, at SLC Reception and on Leila's website. Please enrol in each of your preferred workshops by EITHER:

- ***Dropping by*** **the SLC Reception to enrol in person (Room 320, Level 3, Information Commons Building, 11 Symonds Street) OR**
- ***Emailing*** **slc@auckland.ac.nz with your name, ID number, and the name, date and time of the workshop/s you wish to attend OR**
- ***Calling*** **the SLC Reception on 373-7599 ext. 88850 and book over the phone.**

Useful Websites

- SLC webpage: **www.slc.auckland.ac.nz**
- Cecil: **https://cecil.auckland.ac.nz**
- **Leila's website for STATS 10x SLC workshop handouts & information: www.stat.auckland.ac.nz/~leila**

Revision Notes

Chapter 4 - Probability

Look at blue pages for extra test/exam questions for practice

- A **probability** is a number between 0 and 1 that quantifies uncertainty.
- There are two main sources of probabilities that we will deal with.
  1. Probabilities using a model – some models that may involve equally likely outcomes are *tossing a coin* and *rolling a die*
  2. Probabilities from data
- A **random experiment** is an experiment where the outcome cannot be predicted.
- A **sample space** is the collection of all possible outcomes.
- An **event** is a collection of outcomes. An event **occurs** if any outcome making up that event occurs.
- If all **outcomes** are **equally likely:** $\text{pr}(A) = \dfrac{\text{no. of outcomes in } A}{\text{total no. of outcomes}}$
- The **complement** of an event A, denoted $\overline{A}$, occurs if A does not occur. A, and $\overline{A}$ are **mutually exclusive** events, ie they CANNOT occur at the same time.
- General probability rules:
  1. $\text{pr}(S) = 1$
  2. $\text{pr}(\overline{A}) = 1 - \text{pr}(A)$
- **Statistical Independence** – two events (A & B) are statistically independent if knowing whether B has occurred gives no new information about the chances of A occurring.

  i.e. $\text{pr}(A|B) = \text{pr}(A)$ and $\text{pr}(A \text{ and } B) = \text{pr}(A) \times \text{pr}(B)$

- **Two Types of Test/Exam Questions**

  1. Given a table of numbers/proportions, find the probability:
  - ☞ Easier question/s (can be between 1 and 3 of this type).
  - ☞ *May* want to convert the table into table of probabilities first.

2. Given a short story with proportions, percentages and/or counts about two factors (qualitative variables), find the probability:

- ☞ Harder question/s (can be 1 or 2 of this type).
- ☞ Need to *interpret* the story first, and then construct a table.
- ☞ Use the table to find 1 or 2 probabilities.
- ☞ Steps to constructing a table:

  Step 1: highlight numbers

  Step 2: highlight factors

  Step 3: define factor levels

  Step 4: label table

  Step 5: enter appropriate table total

  Step 6: enter row/column totals from story

  Step 7: enter cell numbers from story

  Step 8: enter remaining numbers by +/-

- **Four Types of Probability Calculations**

1. Probability of <u>AN EVENT</u> (basic/simple)

- ☞ pr(A) → pr (an event)

2. Probability of an event <u>AND</u> another event:

- ☞ pr(A and B) → pr(one event and another event)
- ☞ *Finding* pr(A) <u>*and*</u> pr(B) (intersection)

3. Probability of an event <u>OR</u> another event:

- ☞ pr(A or B) → pr(one event or another event)
- ☞ <u>*Add*</u> pr(A) *to* pr(B), then <u>*subtract*</u> pr(A and B)

Use
GRAND TOTAL (TABLE TOTAL)

4. <u>CONDITIONAL</u> Probability:

☞ Harder to detect but will usually have one of the key words:

- "Given that…"
- "Of those…"
- "Among those…"

Use **_ROW TOTAL/S_** *OR* **_COLUMN TOTAL/S_**

- Look for language that <u>restricts you to part of the table</u> instead of the whole table.
- pr (one event | another event) = pr(A|B) = $\frac{\text{pr}(A \text{ and } B)}{\text{pr}(B)}$

Examples 1 to **4** are about the following information.

Market researchers and pollsters worry that having chosen a sample, the people that they fail to contact may differ in important ways from the ones that they do contact. Research was done where interviewers telephoned designated houses up to 3 times to make contact with the residents. Residents' income levels as well as the number of calls required to contact each resident are tabulated below:

Income	First call respondents	Second call respondents	Third call respondents	Total
Less than $10,000	28	4	2	34
$10,000-30,000	51	11	3	65
$30,000-50,000	41	17	4	62
$50,000-70,000	88	27	11	126
$70,000 or more	96	55	36	187
Total	304	114	56	474

Example 1: The proportion of residents who were contacted on the first call is:

Pr (   ) = ________________

= (4dp)

Example 2: The proportion of residents earning $50,000 or more who were contacted on the first call is:

Pr (   ) = ________________

= (4dp)

Example 3: The percentage of residents who were contacted on the first call and earning less than $10,000 is:

Pr (   ) = ________________

= (2dp)

Example 4: The probability a resident was earning less than $50,000 or was contacted on the second call is:

Pr (   ) = ________________

= (4dp)

Chapter 6 – Continuous Random Variables

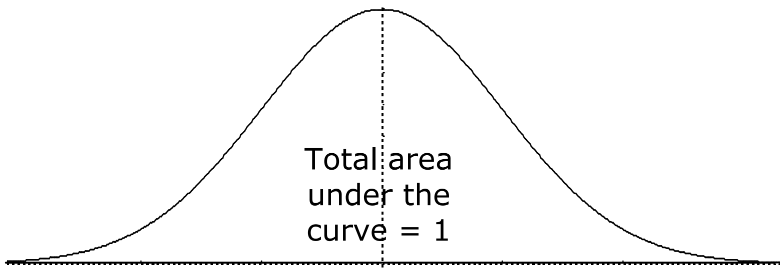
Look at blue pages for good notes and test/exam questions for practice

- A density curve is the probability distribution of a continuous random variable.

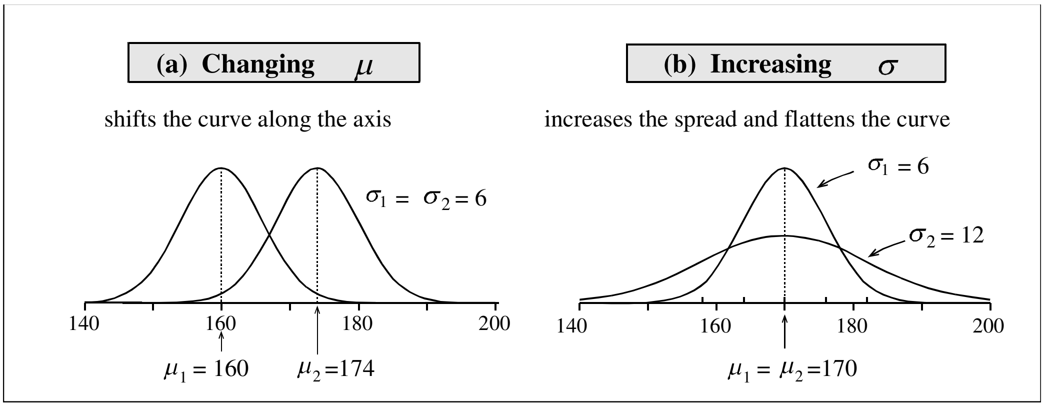
- There are no gaps between the values that a continuous random variable can take and therefore, when we calculate probabilities for a continuous random variable it does not matter whether **interval endpoints** are included or excluded

Normal Distribution

- The Normal Distribution has a probability density function curve, which is smooth, **bell-shaped**, and **symmetric**.



- The shape of the curve is solely determined by the parameters μ (mean) and σ (standard deviation).



- The Normal distribution is important because it:
  - fits a lot of data particularly well
  - can be used to approximate other distributions
  - is very important in statistical inference

- If X is a continuous random variable from a Normal distribution then:
  - $E(X) = \mu$ and $sd(X) = \sigma$
  - Probability distribution function of X is written: $X \sim \text{Normal}(\mu, \sigma)$

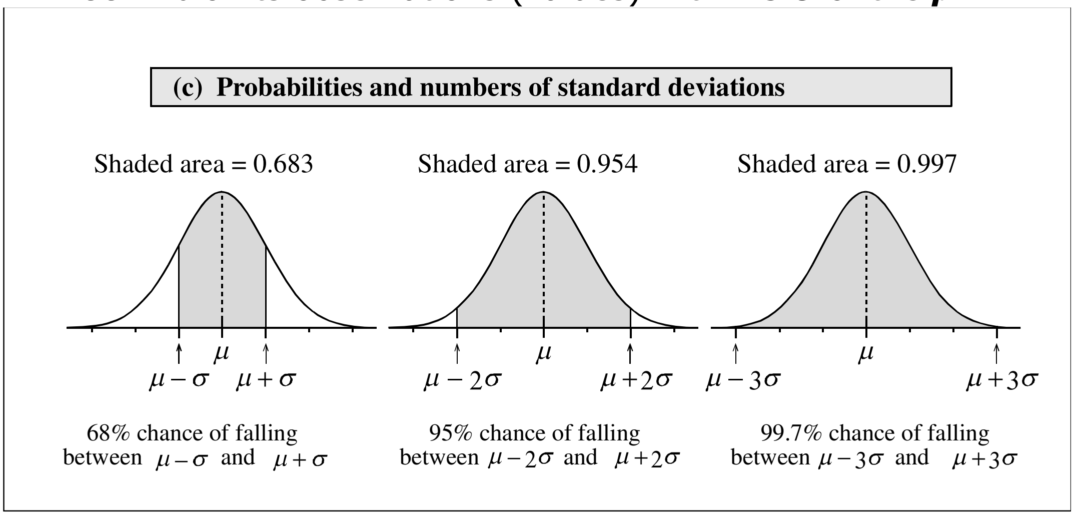
Chapter 6 test/exam questions

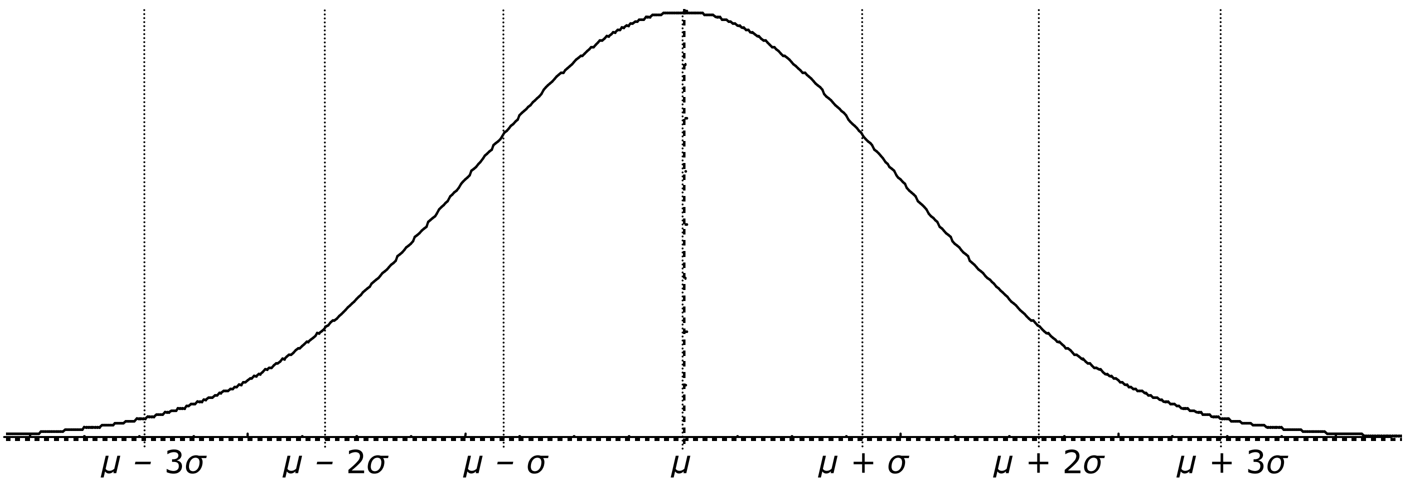
When doing Chapter 6 problems, it is sensible to draw a Normal curve and then mark on it what is known and what is unknown. There are **three (3)** types of Chapter 6 test/exam questions:

1. *True/False (Normal) Chapter 6 problem*

There will be five statements, each about one or the other or both of two different Normal distributions. Use the 68-95-99.7% rule or z-scores to determine whether four of the statements are true or false. The fifth statement will probably be comparing the means (centres/averages) and standard deviations (spread/variability) of the two distributions.

- **68-95-99.7% rule**: A population with a Normal distribution has:
  - ✓ 68% of its observations (values) within 1 σ of the μ
  - ✓ 95% of its observations (values) within 2 σ of the μ
  - ✓ 99.7% of its observations (values) within 3 σ of the μ



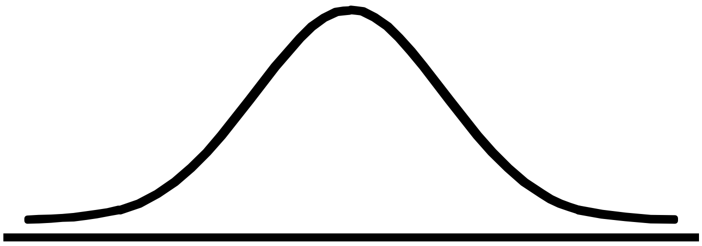


The **z-score**, $z = \dfrac{x - \mu}{\sigma}$, is a standardised number. It represents the number of standard deviations, σ, the value of x is away from the mean, μ. We can use z-scores to compare two or more different Normal distributions.

2. ***Normal probability problem, i.e. find a probability associated with a number***

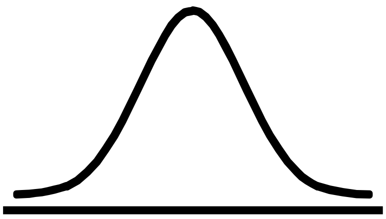
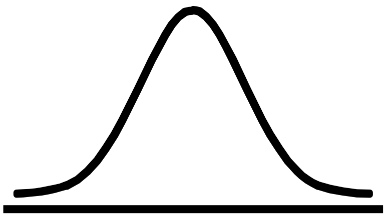
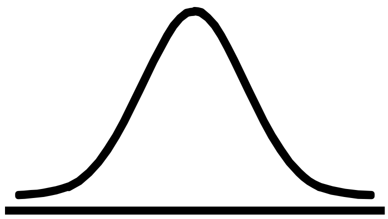
When finding a probability, shade the desired area under the curve and then devise a way to obtain it using lower tail probabilities which is all the computer can give. There are three types of Normal probability problems:

- *Find a lower tail probability (area)*
  The computer can find/give the answer directly.



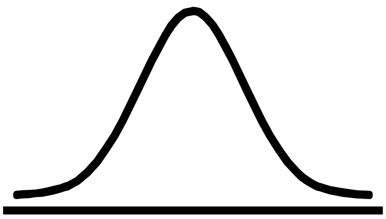
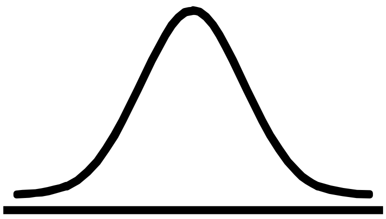
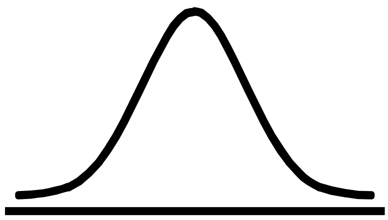
pr($X \leq x$)

- *Find an upper tail probability (area)*
  The computer cannot find/give the answer directly so subtract the lower tail from 1.

pr($X \geq x$) = 1 − pr($X \leq x$)

- *Find a probability (area) between two numbers*
  The computer cannot find/give the answer directly so subtract the lower tail beneath the smaller number from the lower tail beneath the larger number.

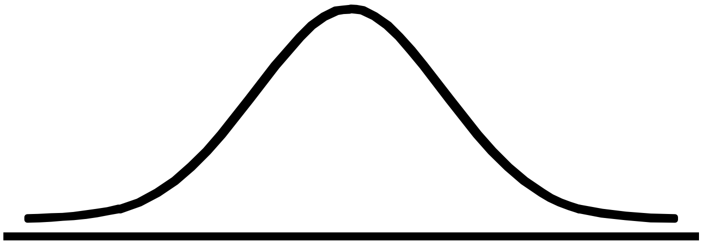
  

pr($a \leq X \leq b$) = pr($X \leq b$) − pr($X \leq a$)

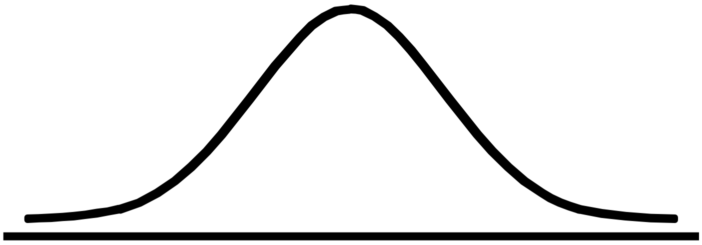
3. *Inverse Normal problem, i.e. find a number associated with a probability*

This type of problem occurs when we know the probability (e.g. the highest 10% in the class) and we need to find out the number associated with it, *x* (e.g. the mark). There are three types of inverse Normal problems:

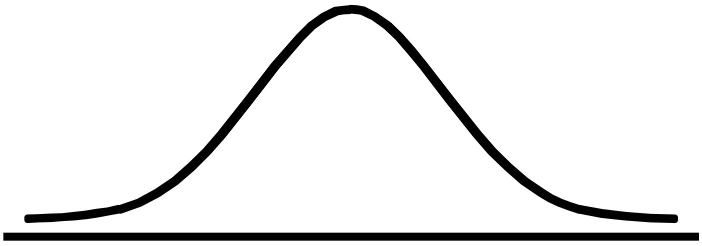
- *Given a lower tail probability, find the number associated with it*
  The computer can find/give the answer directly.



- *Given an upper tail probability, find the number associated with it*
  The computer cannot find/give the answer directly so subtract the upper tail probability from 1 & use the lower tail probability to find the answer.



- *Given a central area/probability, find the two numbers associated with it (the lower limit and the upper limit)*
  The computer can give the two limits as long as you use the lower tails/areas beneath each of them.



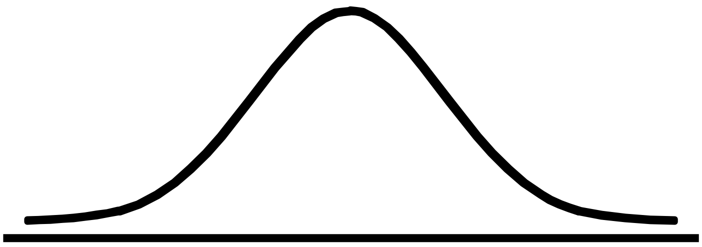
Examples 5 to **10** are about the following information.

Empirical studies have provided support for the belief that a common stock's annual rate of return is approximately Normally distributed. Suppose that you have invested in the stock of a company for which the annual return rate has an expected value of 30 (in percent) and a standard deviation of 10 (in percent).

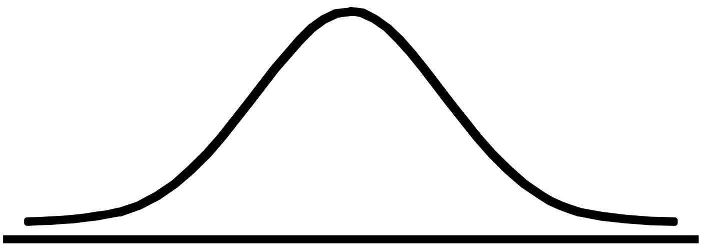
Normal with mean = 30.0000 and standard deviation = 10.0000

x	Pr(X<=x)	Pr(X<=x)	x
10	0.0228	0.20	21.58
25	0.3085	0.25	23.26
40	0.8413	0.40	27.47
50	0.9772	0.75	36.74

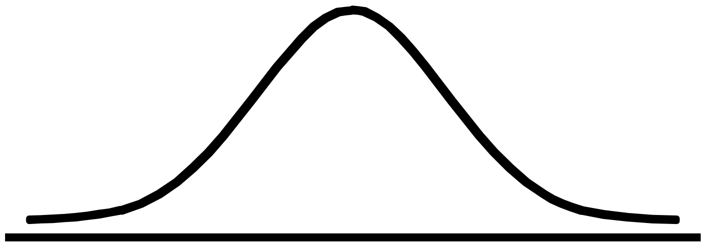
Example 5: The proportion of stock that will give an annual return rate of between 25% and 50% is:



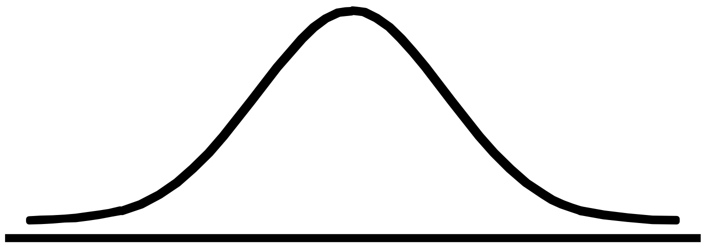
Example 6: There is a 20% probability that the annual rate of return will be less than:



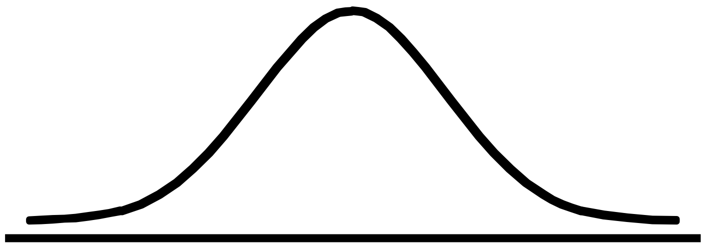
Example 7: The stock return rates of the central 50% are:



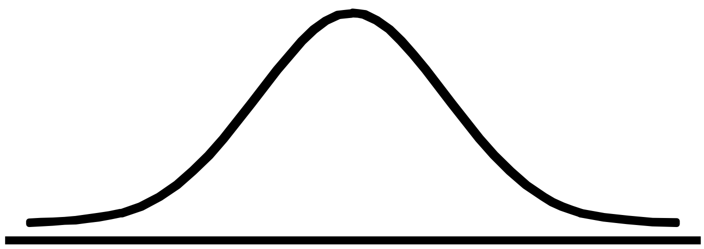
Example 8: The probability that the stock will give an annual return rate of at least 40% is:



Example 9: Sixty percent of the stock will give an annual return rate of at least:



Example 10: The probability that the stock will give an annual return rate of no more than 10% is:



Revision Notes

Chapter 7 – Sampling Distributions of Estimates

Look at blue pages for good notes, 26 statements (2 of which are false!) and test/exam questions for practice

- Statistics is concerned with finding out about the real world and aspects of it specific to our area of interest. Statistical tools allow us to deal with the **uncertainty** present in all samples due to **sampling variation** which occurs because we are unable to survey the entire population of interest.

- We are usually unable to survey the entire population (take a census) as it is too large and/or there are:
  - ✓ budget constraints
  - ✓ time limits
  - ✓ logistical barriers

- This means we are unable to establish the **parameters** of interest within our population, such as:
  - ✓ Population mean, μ
  - ✓ Population standard deviation, σ

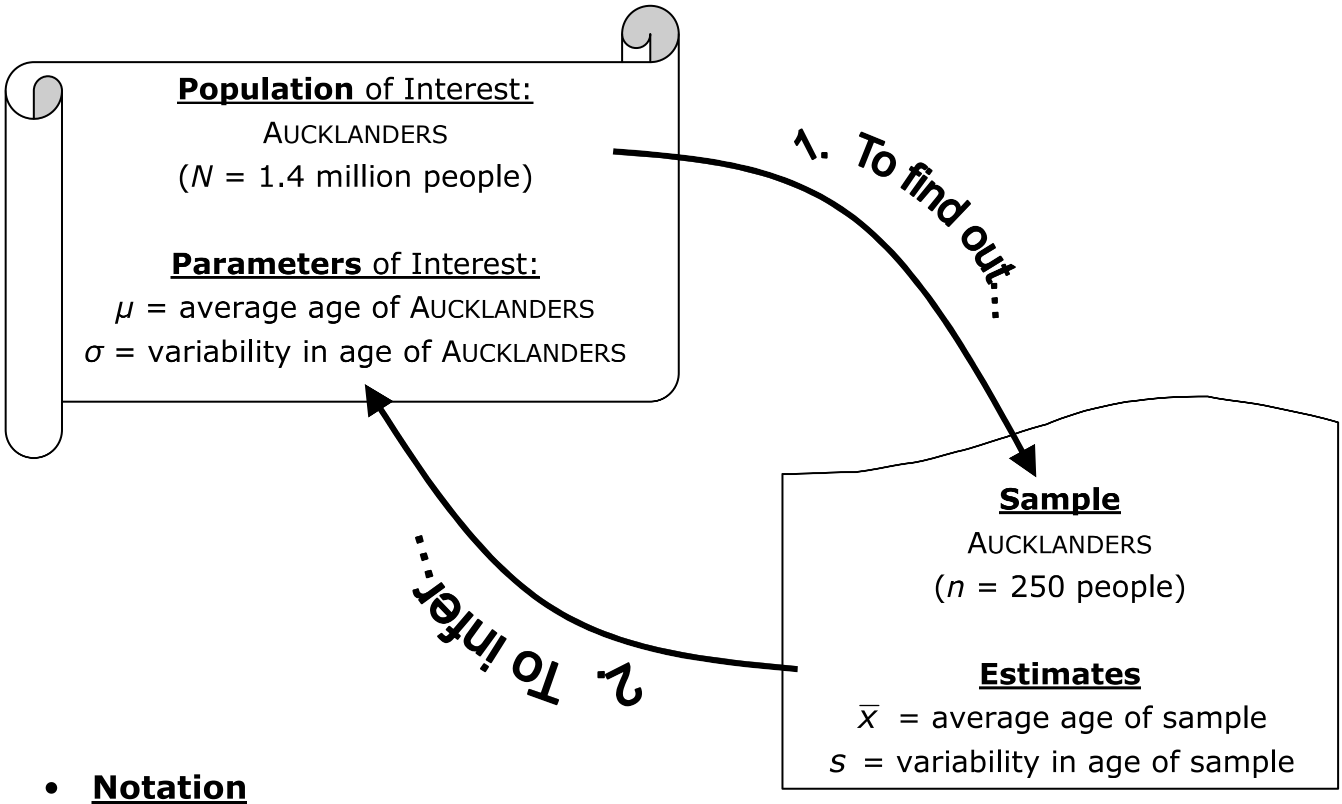
- This means that the **parameter** of interest is an ***<u>unknown</u> numerical characteristic*** for that particular population.

- To estimate an ***<u>unknown</u> numerical characteristic*** **(parameter)** for our population of interest, we take a sample and find a sample **estimate** from it (that is, we make a **statistical inference**). The **sample estimates** of the above **population parameters** are:
  - ✓ Sample mean, $\bar{x}$
  - ✓ Sample standard deviation, $sd(\bar{x})$ or σ_{n-1} or s

- Usually $^{\wedge}$HATS or $\overline{}$BARS are used to distinguish between **sample estimates** and **population parameters**.

- Random variables X_1, X_2, … X_n, form a random sample from a distribution if:
  - ✓ they all have the same distribution; and
  - ✓ they are independent of one another.

- The big question which we will answer in Chapter 7 is "But how can we trust the sample estimates ($\bar{x}$ and s) from our single sample of size n?"



- **Notation**

  In statistics we use **CAPITAL letters** to refer to the **variable of interest** for the population and **small letters** to specify the **actual "number" observed** for that variable in our particular sample.

Variable of Interest **CAPITAL** LETTER	Actual "number" **small** LETTER
X	x
$\overline{X}$	$\overline{x}$

- **The sample mean – $\overline{X}$**

If X_1, X_2, … X_n, form a random sample from a distribution where $E(X_i) = \mu$ and $sd(X_i) = \sigma$, then,

✓ The **expected value** of the sample mean, $E(\overline{X})$ is calculated by:

$$\mu_{\overline{X}} = E(\overline{X}) = \mu$$

✓ The **standard deviation** of the sample mean, $sd(\overline{X})$ is calculated by:

$$\sigma_{\overline{X}} = sd(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

- ## The Central Limit Theorem

  ✓ The Central Limit Theorem tells us that the larger the sample size, the closer the distribution of $\overline{X}$ comes to a Normal distribution. Even if the distribution of X is non-Normal, the distribution of $\overline{X}$ will be **approximately** Normal for a sufficiently large sample size n.

  ✓ If X is from a "well-behaved" distribution (i.e. symmetric, no outliers) the Central Limit Theorem works reasonably fast. n = 10 may be sufficient!

  ✓ In **general**, n = 30 works well for most distributions *except* distributions that are severely skewed or have large outliers.

  ✓ If the distribution is **severely skewed**, n = 50 should be sufficient.

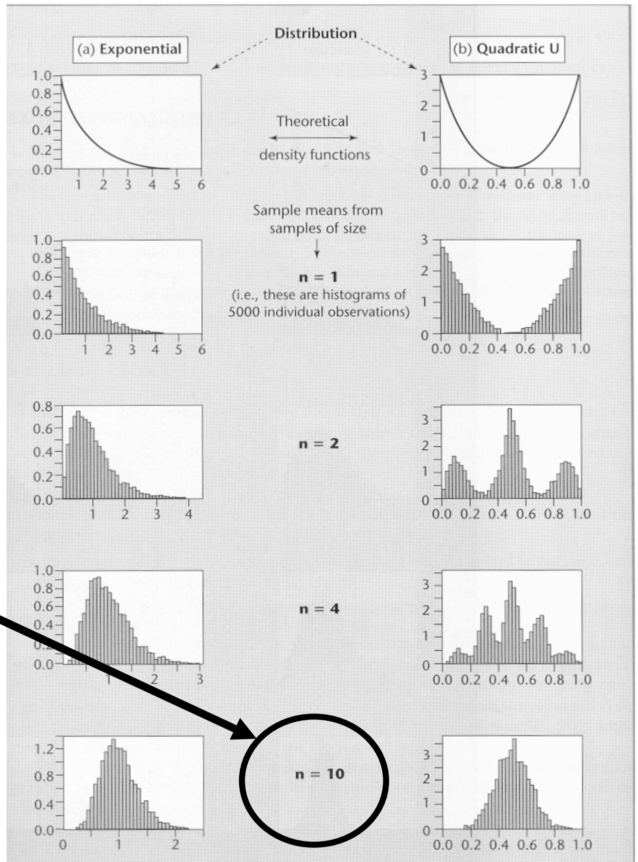


FIGURE 7.2.4 Histograms of means calculated from random samples from the Exponential and Quadratic U distributions.

Chance Encounters, C.J. Wild & G.A.F. Seber, p286

  ✓ If X is from a Normal distribution, then $\overline{X}$ is **exactly** Normally distributed.

- $\overline{X}$ is an **unbiased** estimator of μ because $E(\overline{X}) = \mu$.

- ## Standard errors

However, as $sd(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$, it is not a useful measure of the precision of $\overline{x}$, because we do not know the value of σ. Therefore, we have to use the ***standard error*** of the sample mean to estimate the precision of $\overline{x}$ as an estimate of μ:

$$\text{The standard error of } \overline{x} = \frac{\text{sample standard deviation}}{\sqrt{\text{sample size}}} = \text{se}(\overline{x}) = \frac{s}{\sqrt{n}}$$

- **Student's *t*-distribution**

  ✓ Parameter: Degrees of Freedom (*df*).

  ✓ Bell shaped and centred at 0 like the (Standard) Normal (0,1) distribution but it's more variable.

  ✓ As *df* becomes larger, the Student (*df*) distribution becomes more and more like the Standard Normal distribution.

  ✓ Student's *t*-distribution ($df = \infty$) and Normal (0,1) are the same distribution.

  ✓ The random sample from a Normal distribution : $T = \dfrac{\overline{X} - \mu}{se(\overline{X})}$ is exactly distributed as Student($df = n - 1$)

  ✓ Methods based on this distribution works very well even for small samples that are from very non-Normal distributions.

  ✓ By t_{df} (*prob*), we mean the number t such that when $T \sim$ Student(*df*), $\text{pr}(T \geq t) = prob$; that is, the tail area above t (that is to the right of t on the graph) is *prob*:

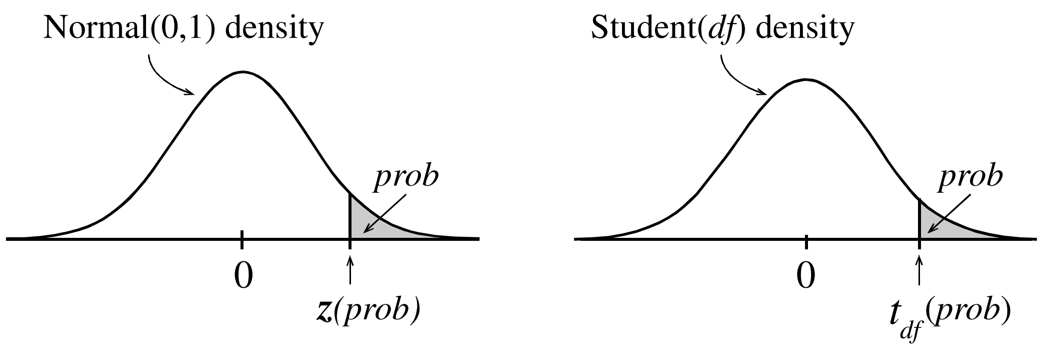


Figure 7.6.2 The *z(prob)* and *t(prob)* notations.

Chapters 4, 6 & 7 – Questions

Questions 1 to **3** refer to the following information.

At the beginning of the semester, stage one statistics students at the University of Auckland were surveyed. One question asked: 'How would you describe your Excel experience?'. A total of 918 students answered this question. Each answer was classified according to the response given by the student, and the stream the student attended. The results are given in the table below, where 107, 108 and 101 refer to the various streams.

	Stream			
Response	107	108	101	**Total**
None	15	36	102	153
Very Little	44	89	119	252
Some	74	150	200	424
Lots	9	29	51	89
Total	142	304	472	918

Table: Responses to question regarding Excel experience.

1. Of those students in this survey who had no Excel experience at the start of the course, the proportion who were in stream 101 is:

(1) 0.216
(2) 0.667
(3) 0.167
(4) 0.111
(5) 0.514

2. The proportion of students in this survey who had no Excel experience at the start of the course or were in stream 101 is:

(1) 0.514
(2) 0.667
(3) 0.167
(4) 0.111
(5) 0.570

3. The proportion of students in this survey who had no Excel experience at the start of the course is:

(1) 0.216
(2) 0.667
(3) 0.167
(4) 0.111
(5) 0.514

Questions 4 to **6** refer to the following information.

A study was performed to investigate liver damage in 21 patients with Hepatitis C. Two qualitative variables were measured for each patient, as listed below.

Gender: The gender of each patient.

Histology: A category indicating the amount of damage to the liver of each patient (mild, medium, severe).

Consider the two-way table of counts for Gender and Histology given below.

	Histology			
Gender	mild	medium	severe	**Total**
female	3	5	3	11
male	1	5	4	10
Total	4	10	7	21

Table: A two-way table of Gender and Histology

4. The proportion of patients with medium liver damage is:

(1) 10/21
(2) 5/10
(3) 5/21
(4) 5/11
(5) 11/21

5. The proportion of women with medium liver damage is:

(1) 10/21
(2) 5/10
(3) 5/21
(4) 5/11
(5) 11/21

6. The proportion of patients who female and have medium liver damage is:

(1) 10/21
(2) 5/10
(3) 5/21
(4) 5/11
(5) 11/21

Questions 7 to **9** are about the following information.

In 1989, 7013 New Zealanders died from coronary heart disease. The numbers of deaths classified by age and gender is:

Age	Males	Females	Total
<45	74	19	93
45-54	259	69	328
55-64	701	190	891
≥65	3017	2684	5701
Total	4051	2962	7013

7. The probability that a randomly chosen member of this population was female and aged 45 or over is:

(1) $\frac{69}{328}$

(2) $\frac{69}{2962}$

(3) $\frac{69+190+2684}{2962}$

(4) $\frac{69+190+2684}{7013}$

(5) $\frac{328+891+5701}{7013}$

8. Given that a randomly chosen member of this population was male, the probability that he was younger than 45 at the time of his death is:

(1) $\frac{74}{493}$

(2) $\frac{74}{4051} \times \frac{74}{93}$

(3) $\frac{74}{4051}$

(4) $\frac{74}{7013}$

(5) $\frac{93+4051-74}{7013}$

9. The percentage of this population that was aged 55 or over is:

(1) 13%
(2) 51%
(3) 9%
(4) 94%
(5) 6%

Questions 10 to **12** are about the following information.

As part of it's responsibilities to EEO (Equal Employment Opportunities), a New Zealand Government department sent out a survey to its employees. In one section of the survey the employees were asked to consider the statement "*Ethnic and cultural diversity is welcomed in this workplace*" and to make one selection from a given list of responses. The table below cross classifies ethnic origin of the respondent and their chosen response to the statement.

Ethnic Origin	1 Agree Strongly	2 Agree Somewhat	3 Neither Agree nor Disagree	4 Disagree Somewhat	5 Disagree Strongly	6 No Opinion	Total
Pakeha (NZ European)	32	35	11	10	10	6	104
Maori	1	7	2	2	0	0	12
Non-NZ European	1	1	0	0	0	0	2
Pacific Islander	3	4	4	1	0	0	12
Asian	4	5	2	0	0	0	11
Other	1	0	10	3	4	0	18
Total	42	52	29	16	14	6	159

10. The percentage of those surveyed who **agreed strongly** with the statement is:
    (1) 30.8%
    (2) 26.4%
    (3) 0.26%
    (4) 0.59%
    (5) 20.1%

11. The probability that an employee randomly selected from this group is a **Pacific Islander** who either **agreed somewhat** or **agreed strongly** with this statement is:
    (1) 0.074
    (2) 0.075
    (3) 0.044
    (4) 0.583
    (5) 0.591

12. Of those who **disagreed** strongly with the statement, the proportion who were Pakeha is:
    (1) 0.654
    (2) 0.714
    (3) 0.088
    (4) 0.096
    (5) 0.063

Questions 13 to **16** are about the following information.

An analysis of telephone calls made from a firm's office indicates that the length of a call is Normally distributed with a mean, mu, of 240 seconds, and a standard deviation, sigma, of 40 seconds.

Normal with mean = 240.0000 and standard deviation = 40.0000

x	P(X<=x)	P(X<=x)	x
3	0.0000	0.0010	116.3907
179	0.0636	0.0100	146.9461
180	0.0668	0.1000	188.7379
181	0.0701	0.2000	206.3352
300	0.9332	0.8000	273.6648
		0.9000	291.2621
		0.9900	333.0539

13. The probability of a call lasting less than 3 minutes is:
    (1) 0
    (2) 0.0636
    (3) 0.9332
    (4) 0.0668
    (5) 0.0701

14. We would expect only 1% of calls to be shorter than:
    (1) 116 seconds
    (2) 189 seconds
    (3) 147 seconds
    (4) 291 seconds
    (5) 333 seconds

15. The central 80% of calls would fall between:
    (1) 206 and 273 seconds
    (2) 189 and 291 seconds
    (3) 147 and 333 seconds
    (4) 189 and 273 seconds
    (5) 206 and 291 seconds

16. Let $X_1, X_2, \ldots, X_{15}$ be the times (in seconds) of 15 randomly selected phone calls made from the firm's offices. Let $\overline{X}$ be the mean of $X_1, X_2, \ldots, X_{15}$. The standard deviation of $\overline{X}$, sd($\overline{X}$), is approximately:
    (1) 2.7
    (2) 600
    (3) 240
    (4) 10.3
    (5) 40

Questions 17 to **19** refer to the following information.

A study was performed in which one of the measurements taken was the duration of radiotherapy sessions, D. 183 sessions were timed, and the durations were analysed and found to be approximately Normally distributed. The mean session duration, $\overline{x}_D$, was 12.7 minutes and the standard deviation of the session durations, s_D, was 3.4 minutes.

In the following questions assume $D \sim \text{Normal}(\mu_D = 12.7,\ \sigma_D = 3.4)$.

17. One session lasted 15.3 minutes. How many standard deviations is this away from the mean session duration?

(1) 1.41 (4) 2.6
(2) 0.76 (5) -1.41
(3) -0.76

Use the following computer output to answer **Question 18**:

```
Normal with mean = 12.7000 and standard deviation = 3.40000

            x   P( X <= x)
       0.1000   0.0001
       0.9000   0.0003
       8.3427   0.1000
      17.0573   0.9000
```

18. The greatest session duration that is exceeded by 90% of all sessions is:

(1) 0.0003 (4) 17.0573
(2) 0.0001 (5) 8.3427
(3) -7.3427

19. Let $X_1, X_2, \ldots, X_9$ be the durations (in minutes) of 9 randomly selected radiotherapy sessions from the study. Let $\overline{X}$ be the mean of $X_1, X_2, \ldots, X_9$. The standard deviation of $\overline{X}$, sd($\overline{X}$), is approximately:

(1) 1.411 (4) 1.133
(2) 3.4 (5) 0.378
(3) 12.7

Questions 20 to **22** refer to the information given below.

Hypholoma Capnoides is a pleasant tasting fungus (mushroom) which looks very much like the generally taller, poisonous fungus *Sulphur Tuft*. It has been suggested that the two fungi can be distinguished by use of the following guide: if the fungus is shorter than the threshold height of 8cm then classify it as the edible *Hypholoma Capnoides*, but if the fungus is taller than 8cm then classify it as the poisonous *Sulphur Tuft*.

Let X_H be the height of a *Hypholoma Capnoides* mushroom and X_S be the height of a *Sulphur Tuft* fungus. X_H is well modelled by a Normal distribution with a mean of 6.5cm and a standard deviation of 1.76cm (X_H ~ Normal(μ_H = 6.5cm, σ_H = 1.76cm)) whereas X_S is well modelled by a Normal distribution with mean 9.5cm and a standard deviation 1.25cm (X_S ~ Normal(μ_S = 9.5cm, σ_S = 1.25cm)). Assume X_H and X_S are independent random variables.

The table below shows a selection of probabilities from the distributions of X_H and X_S.

```
Normal with mean = 6.5                Normal with mean = 9.5
and standard deviation = 1.76         and standard deviation = 1.25
x          pr(X<=x)                   x          Pr(X<=x)
2.4        0.01                       6.6        0.01
4.2        0.10                       8.0        0.12
5.0        0.20                       8.5        0.20
5.3        0.25                       8.7        0.25
5.5        0.30                       8.8        0.30
6.1        0.40                       9.2        0.40
6.5        0.50                       9.5        0.50
6.9        0.60                       9.8        0.60
7.4        0.70                       10.2       0.70
7.7        0.75                       10.3       0.75
8.0        0.80                       10.6       0.80
8.8        0.90                       11.1       0.90
10.6       0.99                       12.4       0.99
```

Table: A selection of probabilities from Normal(6.5, 1.76) and Normal(9.5, 1.25) distributions.

20. Under the suggested classification plan, the probability of classifying a fungus as poisonous *Sulphur Tuft* when it really is an edible *Hypholoma Capnoides* is approximately:

(1) 0.80
(2) 0.25
(3) 0.88
(4) 0.20
(5) 0.12

21. The interquartile range for X_H is:
   (1) 1.6cm
   (2) 2.4cm
   (3) 3.0cm
   (4) 2.5cm
   (5) 2.1cm

22. To ensure that the probability of mistaking a poisonous *Sulphur Tuft* for an edible *Hypholoma Capnoides* is 1%, the threshold height should be changed from 8cm to approximately:
   (1) 6.5cm
   (2) 12.4cm
   (3) 2.4cm
   (4) 6.6cm
   (5) 10.6cm

Question 23 refers to the following information.

A recent study was designed to investigate the abundance and size of snapper in the Cape Rodney – Okakari Point Marine Reserve and in an adjacent non-reserve region. Fishing surveys were conducted in both regions.

The data on the lengths of the snapper caught have been explored. We have decided to model the distribution of the length of a **Reserve** snapper, X_R, and the distribution of the length of a **non-Reserve** snapper, X_{NR}, as follows:

$$X_R \sim \text{Normal}(\mu_R = 360.18\text{mm}, \sigma_R = 94.47\text{mm})$$

$$X_{NR} \sim \text{Normal}(\mu_{NR} = 257.09\text{mm}, \sigma_{NR} = 59.35\text{mm})$$

23. Use the following computer output in this question.

```
Normal with mean = 360.18000 and standard deviation = 94.470000

            X           P( X <= x)
            0.2000      0.0001
            0.8000      0.0001
            239.1118    0.1000
            280.6720    0.2000
            439.6880    0.8000
            481.2482    0.9000
```

80% of Reserve snapper are longer than:
   (1) 280.7mm
   (2) 239.1mm
   (3) 360.2mm
   (4) 481.2mm
   (5) 439.7mm

24. $X_1,\ldots,X_{35}$ form a random sample from a moderately skewed distribution. Let $\overline{X}$ denote the average of these 35 variables. The Central Limit Theorem tells us that $\overline{X}$ has, approximately:

(1) a Student's *t*-distribution
(2) a moderately skewed distribution.
(3) a Normal distribution.
(4) an *F*-distribution.
(5) a χ^2-distribution.

25. Let X_1, X_2, …, X_n be a random sample from a distribution with mean μ and standard deviation σ. Which **one** of the following statements is **false**?

(1) $E(X_1) = E(X_2) = \ldots = E(X_n)$.
(2) $X_1, X_2, \ldots, X_n$ all have the same distribution.
(3) $X_1, X_2, \ldots, X_n$ are independent of each other.
(4) $sd(X_1) = sd(X_2) = \ldots = sd(X_n)$.
(5) $nX_1 = X_1 + X_2 + \ldots + X_n$.

26. Which **one** of the following statements is **false**?

(1) If the distribution of a random variable, X, is exactly Normal, then, according to the Central Limit Theorem, the distribution of $\overline{X}$ will be approximately Normal.
(2) se($\overline{x}$) is an estimate of $sd(\overline{X})$.
(3) $\overline{x}$ is an unbiased estimate of μ since $E(\overline{X}) = \mu$.
(4) An estimate is a known quantity computed from data.
(5) A parameter is a numerical characteristic of a population or distribution.

Questions 27 and **28** refer to the following information.

Let X_1, X_2, …, X_{100} be a random sample from a skewed distribution with mean, $\mu = 6$ and standard deviation, $\sigma = 2.05$.

27. If $\overline{X}$ is the sample mean, then $\mu_{\overline{X}}$, and standard deviation, $\sigma_{\overline{X}}$, of $\overline{X}$ are:

(1) $\mu_{\overline{X}} = 6$, $\sigma_{\overline{X}} = 0.205$
(2) $\mu_{\overline{X}} = 6$, $\sigma_{\overline{X}} = 2.05$
(3) $\mu_{\overline{X}} = 60$, $\sigma_{\overline{X}} = 2.05$
(4) $\mu_{\overline{X}} = 60$, $\sigma_{\overline{X}} = 20.5$
(5) $\mu_{\overline{X}} = 6$, $\sigma_{\overline{X}} = 20.5$

28. The distribution of $\overline{X}$ is best described as:

(1) exactly Normal.
(2) approximately Normal.
(3) approximately Binomial.
(4) unknown.
(5) exactly Binomial.

29. Given a random sample from a population with a population mean μ, which **one** of the following statements is **false**?

(1) The distribution of the sample mean, $\overline{X}$, is exactly Normally distributed if the distribution of the population is exactly Normal.

(2) The sample mean is an unbiased estimate of the population mean since $E(\overline{X}) = \mu$.

(3) The standard error of the sample mean is an estimate of the standard deviation of the sample mean.

(4) The mean of the distribution of the sample mean is equal to the population mean. Ie, $\mu_{\overline{X}} = \mu$.

(5) For large samples, the distribution of the sample mean, $\overline{X}$, is exactly Normally distributed.

30. Given a simple random sample, which **one** of the following is **false**?

(1) The mean of the distribution of the sample mean, $\mu_{\overline{X}}$, is equal to the population mean, μ.

(2) The sample mean is an unbiased estimate of the population mean since E[$\overline{X}$] = μ.

(3) The distribution of the sample mean, $\overline{X}$, is approximately Normal for very large samples only if the distribution from which the sample has been drawn is not skewed.

(4) The standard error of the sample mean is an estimate of the standard deviation of the sample mean.

(5) Increasing the sample size by a factor of 4 will double the accuracy of the sample mean as an estimate of the population mean.

31. $X_1,\ldots,X_{25}$ form a random sample from a distribution with mean $\mu = 10$ and standard deviation $\sigma = 9$. The sample mean, $\overline{X}$, has mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$ given by:

(1) $\mu_{\overline{X}} = 25 \times 10$, $\sigma_{\overline{X}} = \sqrt{25} \times 10$

(2) $\mu_{\overline{X}} = \dfrac{10}{25}$, $\sigma_{\overline{X}} = \dfrac{9}{25}$

(3) $\mu_{\overline{X}} = 10$, $\sigma_{\overline{X}} = 9$

(4) $\mu_{\overline{X}} = 10$, $\sigma_{\overline{X}} = \sqrt{\dfrac{9}{25}}$

(5) $\mu_{\overline{X}} = 10$, $\sigma_{\overline{X}} = \dfrac{9}{\sqrt{25}}$

32. Let X_1, X_2, …, X_n be a random sample from a distribution with mean μ and standard deviation σ. Let $\overline{X}$ be the sample mean. Which **one** of the following statements is **false**?
(1) $\overline{X}$ may not be approximately Normally distributed for small samples.
(2) $\overline{X}$ is a random variable.
(3) sd($\overline{X}$) = sd(X).
(4) E($\overline{X}$) = E(X).
(5) $\overline{X}$ is at least approximately Normally distributed for large samples.

33. Which **one** of the following statements is **false**?
(1) A parameter is a numerical characteristic of a population or distribution.
(2) An estimate is a quantity computed from data.
(3) The standard deviation of $\overline{X}$ is usually greater than the standard deviation of X.
(4) The standard error of the sample mean, se($\overline{x}$), estimates the precision of $\overline{x}$ as an estimate of μ.
(5) $\overline{x}$ is an unbiased estimate of μ.

34. A random variable T has a Student's t-distribution with df degrees of freedom. Which **one** of the following statements is **false**?
(1) The t-distribution for small degrees is flatter with wider tails than the Standard Normal distribution.
(2) The mean of T is zero.
(3) The t-distribution gets closer to the Normal(0,1) distribution as df gets larger.
(4) The t-distribution gets closer to the Normal(0,1) distribution as df gets smaller.
(5) The distribution of T is symmetric.

35. Suppose $X_1, X_2, \ldots, X_{11}$ are the weights of 11 randomly selected packets of M & M's which come from a distribution with mean μ = 54 and standard deviation σ = 2. Then the distribution of the sample mean has mean $\mu_{\overline{X}}$ and standard deviation $\sigma_{\overline{X}}$ given by:
(1) $\mu_{\overline{X}} = 4.91$, $\sigma_{\overline{X}} = 0.182$
(2) $\mu_{\overline{X}} = 594$, $\sigma_{\overline{X}} = 2.000$
(3) $\mu_{\overline{X}} = 54$, $\sigma_{\overline{X}} = 0.603$
(4) $\mu_{\overline{X}} = 54$, $\sigma_{\overline{X}} = 2.000$
(5) $\mu_{\overline{X}} = 4.91$, $\sigma_{\overline{X}} = 0.603$

36. The Central Limit Theorem says that the distribution of the mean, $\overline{X}$, of a random sample:
    (1) is a Student's *t*-distribution with $n-1$ degrees of freedom.
    (2) has an *approximately* Chi-Square distribution when the mean is large.
    (3) is an *F*–distribution with degrees of freedom which can be calculated from the sample data.
    (4) is unknown.
    (5) is *approximately* Normal in large samples.

37. Let $X_1,\ldots,X_{10}$ be a random sample of size 10 from a distribution with $\mu = 12$ and $\sigma = 2$, then the distribution of $\overline{X}$ has expected value E[$\overline{X}$] and standard deviation *sd*[$\overline{X}$] where:
    (1) E[$\overline{X}$] = 1.2, *sd*[$\overline{X}$] = 2
    (2) E[$\overline{X}$] = 12, *sd*[$\overline{X}$] = 0.2
    (3) E[$\overline{X}$] = 1.2, *sd*[$\overline{X}$] = 0.63
    (4) E[$\overline{X}$] = 12, *sd*[$\overline{X}$] = 0.63
    (5) E[$\overline{X}$] = 120, *sd*[$\overline{X}$] = 6.32

38. Which one of the following statements is **false?**
    (1) For large random samples, the true value of μ lies inside the interval $\overline{x} \pm 2se(\overline{x})$ for a little more than 95% of all samples taken.
    (2) For random samples for a Normal distribution, $T = (\overline{X} - \mu)/se(\overline{X})$ is exactly distributed as Student's *t*-distribution ($df = n - 1$).
    (3) The precision of an estimate refers to its variability – one estimate is less precise than another if it has more variability.
    (4) The Student's *t*-distribution ($df = \infty$) distribution has 'fatter' or 'heavier' tails than the Normal($\mu = 0$, $\sigma = 1$) distribution.
    (5) For large random samples, $T = (\overline{X} - \mu)/se(\overline{X})$ is distributed as approximately Normal(μ_X=0, σ_X=1).

ANSWERS

1.	**(2)**	2.	**(5)**	3.	**(3)**	4.	**(1)**	5.	**(4)**	6.	**(3)**
7.	**(4)**	8.	**(3)**	9.	**(4)**	10.	**(2)**	11.	**(3)**	12.	**(2)**
13.	**(4)**	14.	**(3)**	15.	**(2)**	16.	**(4)**	17.	**(2)**	18.	**(5)**
19.	**(4)**	20.	**(4)**	21.	**(2)**	22.	**(4)**	23.	**(1)**	24.	**(3)**
25.	**(5)**	26.	**(1)**	27.	**(1)**	28.	**(2)**	29.	**(5)**	30.	**(3)**
31.	**(5)**	32.	**(3)**	33.	**(3)**	34.	**(4)**	35.	**(3)**	36.	**(5)**
37.	**(4)**	38.	**(4)**								