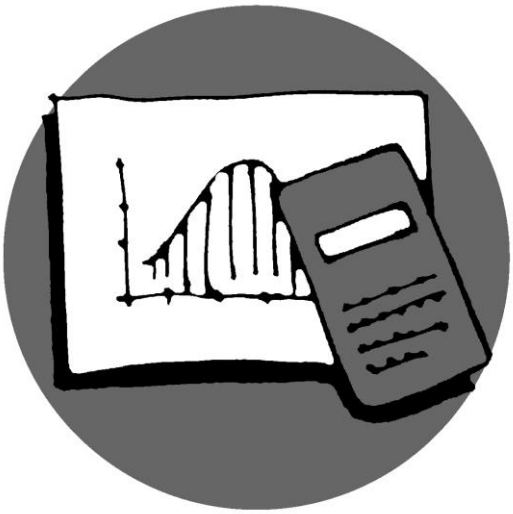
Normal and Binomial Distributions



Library, Teaching and Learning
2014

Lincoln University
Te Whare Wānaka o Aoraki
CHRISTCHURCH•NEW ZEALAND
New Zealand's specialist land-based university

- By now, you know about averages – means in particular – and are familiar with words like *data, standard deviation, variance, probability, sample, population*

- You **must** know how to use your calculator to enter data, and from this, access the mean, standard deviation and variance

- You need to become familiar with the various symbols used and their meanings – be able to "speak" the language:

  Sample statistics are estimates of *population parameters*:

	symbol used for the population *parameter*	symbol used for the sample *statistic*
mean	μ	$\bar{x}$
standard deviation	σ	s
variance	σ^2	s^2
standard error	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$

- You should appreciate that the analysis and interpretation of data is the basis of decision making. For example
  - Should the company put more money into advertising?
  - Should more fertilizer or water be applied to the crop?
  - Is it better to use Brand A or Brand B? etc

- There are many analytical processes. Which process you use depends on
  - What type of data you have – discrete or continuous
  - How many variables - one, two, many
  - What you want to know

Tests and Examination preparation

Practise on a regular basis – set aside, say, half an hour each night or every second night, and/or 3 times during the weekend – rather than a whole day or several hours just before a test.

Make sure your formula sheet is with you as you work, so that you become familiar with the information that is on it.

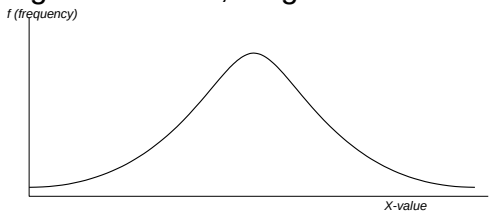
The following sections show summaries and examples of problems from the Normal distribution and the Binomial distribution.

Best practice

For each, study the overall explanation, learn the parameters and statistics used – both the words and the symbols, be able to use the formulae and follow the process.

Normal Distribution

- Applied to single variable continuous data that is distributed approximately normally. e.g. heights of plants, weights of lambs, lengths of time
  This distribution is graphed like:



- Used to calculate the probability of occurrences *less than, more than, between* given values
  e.g. "the probability that a randomly selected plant will be less than 70mm"
  "the probability that a randomly selected lamb will be heavier than 70kg"
  "the probability that the time taken will be between 10 and 12 minutes"

- Standard Normal tables give probabilities - you will need to be familiar with the Normal table and know how to use it.
  First need to calculate how many standard deviations above (or below) the mean a particular value is, i.e., calculate the value of the "standard score" or "Z-score".
  Use the following formula to convert a raw data value, X , to a standard score, Z :

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

eg. Suppose a particular population has μ= 4 and σ = 2. Find the probability of a randomly selected value being greater than 6.

The Z score corresponding to X = 6 is $Z = \frac{(6-4)}{2} = 1.$

(Z=1 means that the value X = 6 is 1 standard deviation above the mean.)

Now use standard normal tables to find P(Z>1) = 0.6587 *(more about this later)*.

Process:

L
E
A
R
N

- Draw a diagram and label with given values i.e.

$$\mu, (\textbf{population mean}) \sigma, (\textbf{pop s.d.}) \; and \; X \; (raw \; score)$$

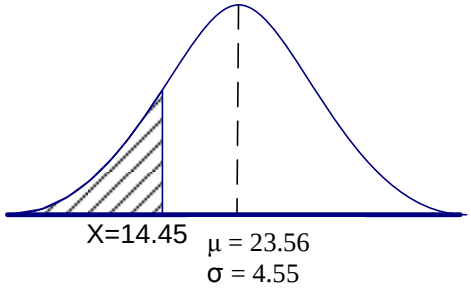
- Shade area required as per question
- Convert raw score (X) to standard score (Z) using formula
- Use tables to find probability, eg $p(Z < z)$.
- Adjust this result to required probability

That is, find p (an event) given a population mean and s.d. and a raw score.

Example 1

Wool fibre breaking strengths are normally distributed with mean μ = 23.56 Newtons and standard deviation, σ = 4.55.
What proportion of fibres would have a breaking strength of 14.45 or less?

- **Draw a diagram, label and shade area required:**



- **Convert** raw score (X) to standard score (Z) : $Z = \dfrac{14.45 - 23.56}{4.55} = -2.0$

  That is, the raw score of 14.45 is equivalent to a standard score of -2.0.
  It is negative because it is on the left side of the curve.

- **Use tables** to find probability and adjust this result to required probability:

$$P(X < 14.45) = P(Z < -2.0) = 0.0228$$

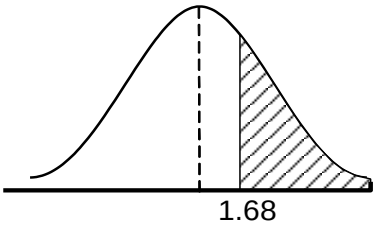
In this case there is no adjustment needed, since the shading in the diagram is similar to that shown in the standard normal table.
That is the proportion of fibres with a breaking strength of 14.46 or less is 2.28%.

***Note: Standard normal tables come in various forms. The ones used for these exercises show the probability of Z being less than z, i.e., P(Z<z). Some forms of standard normal tables give P(0<Z<z). Make sure you can use your table appropriately*.**

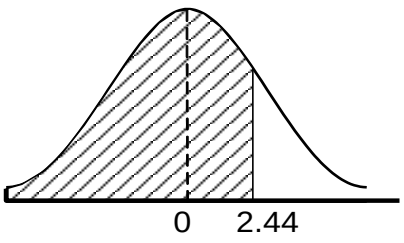
Examples:

Use the following examples to check your understanding of how to use the tables and make adjustments:

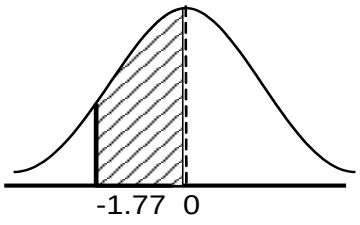
Example A
P(Z>1.68) = 1 – P(Z<1.68)
= 1 - 0.9535
= 0.0465



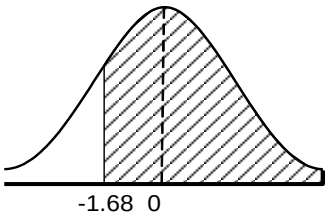
Example B
P(Z<2.44) =0.9927



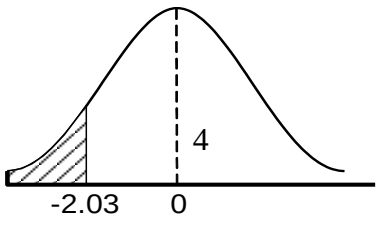
Example C
P(-1.77<Z<0) = 0.5 -P(Z<-1.77)
= 0.5 – 0.0384
= 0.4616



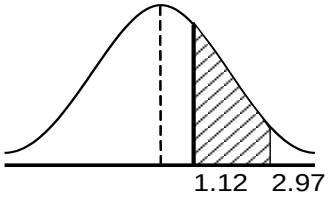
Example D
P(Z>-1.68) = 1 – P(Z< -1.68)
= 1 – .0465
= 0.9535



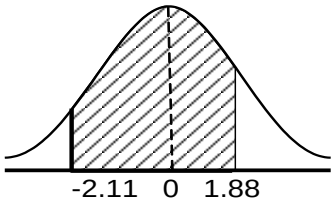
Example E
P(Z<-2.03) = 0.0212



Example F
P(1.12<Z<2.97) =P(Z<2.97)-P(Z<1.12)
= 0.9985 - 0.8686
= 0.1299



Example G

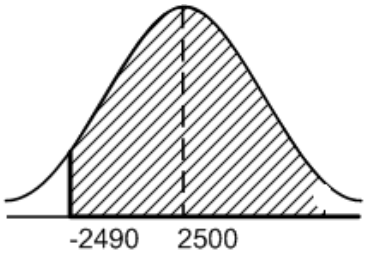


P(-2.11<Z<1.88)
= P(Z<1.88) - P(Z<-2.11)
= 0.9699 - 0.0174
= 0.9525

Example 2

A machine produces ball bearings, the weights of which are normally distributed with a mean of 2500 mg and a standard deviation of 8 mg.
(a) Find the percentage of ball bearings produced which weigh more than 2490 mg.
(b) If 675 ball bearings are chosen at random, how many could be expected to weigh more than 2490 mg?

Solution: (a) Let X be the weight of a ball bearing.



$$P(X > 2490) = P\left(Z > \frac{2490 - 2500}{8}\right) = P(Z > -1.25)$$

[*Change raw score to standard score*]
= 1 - P(Z<-1.25)
= 1 – 0.1056
= 0.8944 ie 89.44%

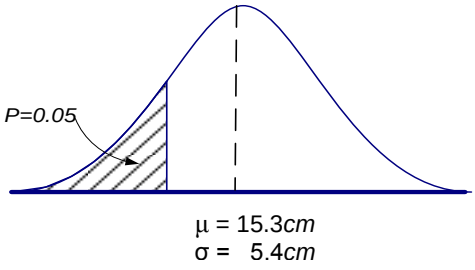
(b) 89.44% of 675 = 603.72 or 604

Inverse process: (to find a value for X, corresponding to a given probability)

- Draw a diagram and label
- Shade area given as per question
- Use probability tables to find Z -score
- Convert standard score (Z) to raw score (X) using inverse formula

Example:
Carrots entering a processing factory have an average length of 15.3 cm, and standard deviation of 5.4 cm. If the lengths are approximately normally distributed, what is the maximum length of the lowest 5% of the load?
(I.e., what value cuts off the lowest 5 %?)

- **Draw a diagram, label and shade area given as in question:**



- **Use standard Normal tables** to find the Z -score corresponding to this area of probability. Convert the standard score (Z) to a raw score (X) using the inverse formula: $X = Z \times \sigma + \mu$

For $P(Z < z) = 0.05$, the Normal tables give the corresponding z-score = -1.645. (Negative because it is below the mean.)

Hence the raw score is:

$$X = Z \times \sigma + \mu$$
$$= -1.645 \times 5.4 + 15.3$$
$$= 6.4$$

i.e., the lowest maximum length is 6.4cm

Exercises

1. For a particular type of wool the number of 'crimps per 10cm' follows a normal distribution with mean 15.1 and standard deviation 4.79.

   (a) What proportion of wool would have a 'crimp per 10 cm' measurement of 6 or less?
   (b) If more than 7% of the wool has a 'crimp per 10 cm' measurement of 6 or less, then the wool is unsatisfactory for a particular processing. Is the wool satisfactory for this processing?

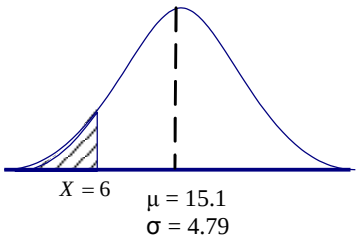
2. The finish times for marathon runners during a race are normally distributed with a mean of 195 minutes and a standard deviation of 25 minutes.
   a) What is the probability that a runner will complete the marathon within 3 hours?
   b) Calculate to the nearest minute, the time by which the first 8% runners have completed the marathon.
   c) What proportion of the runners will complete the marathon between 3 hours and 4 hours?

3. The download time of a resource web page is normally distributed with a mean of 6.5 seconds and a standard deviation of 2.3 seconds.
   a) What proportion of page downloads take less than 5 seconds?
   b) What is the probability that the download time will be between 4 and 10 seconds?
   c) How many seconds will it take for 35% of the downloads to be completed?

4. Potassium blood levels in healthy humans are normally distributed with a mean of 17.0 mg/100 ml, and standard deviation of 1.0 mg/100 ml. Elevated levels of potassium indicate an electrolyte balance problem, such as may be caused by Addison's disease. However, a test for potassium level should not cause too many "false positives".
   What level of potassium should we use so that only 2.5 % of healthy individuals are classified as "abnormally high"?

5. Household consumption of water in a particular suburb is found to be approximately normal with a mean of 279 litres per day and a standard deviation of 90 litres per day.
   a) What is the probability that a randomly selected household will use more than 360 litres per day?
   b) What is the probability that a randomly selected household will use between 200 and 350 litres per day?
   c) The council supplying the water want to charge the top 10% of water users a different rate from the rest. What is the minimum consumption (litres per day) that would put a household into the top 10%?

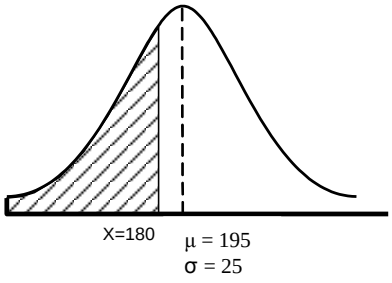
Worked solutions (Normal Distribution)

1. 

$$Z = \frac{(6 - 15.1)}{4.79} = -1.9$$

P(X < 6) = P(Z < -1.9) = <u>0.0287</u>

(a) That is the proportion of wool with a crimp of 6 or less is 2.87%.
(b) Yes, this is satisfactory since 2.87% is less than the stated 7% of the wool.

2. (a) 

$$x = 180 \Rightarrow z = \frac{180 - 195}{25} = -0.6$$

$$P(Z < -0.6) = 0.2743$$

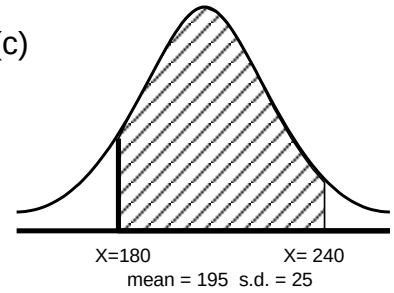
i.e. probability of a runner taking less than 3 hours (180 mins) is 0.2743

(b) $P = 0.08 \Rightarrow Z = -1.41$

Hence $-1.41 = \frac{X - 195}{25}$

$\Rightarrow X = -1.41 \times 25 + 195$

$= 159.75$ **or 160 mins**

(c) 

$$X = 180 \Rightarrow Z = \frac{180 - 195}{25} = -0.6$$

$$P(Z < -0.6) = 0.2743$$

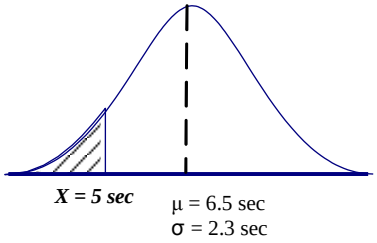
$$X = 240 \Rightarrow Z = \frac{240 - 195}{25} = 1.8$$

$$P(Z < 1.8) = 0.9641$$

$$\Rightarrow P(-0.6 < Z < 1.8) = 0.9641 - 0.2743 = 0.6898$$

i.e. proportion of runners taking between 3 and 4 hours (180 and 240 minutes) is approximately 70%.

3.

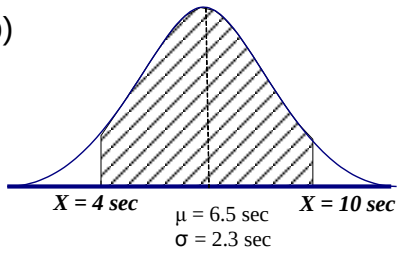


$$X = 5 \Rightarrow Z = \frac{(5 - 6.5)}{2.3} = -0.65$$

$$P(X < 5) = P(Z < -0.65) = 0.2578$$

(a) That is, approximately 26% of the downloads take less than 5 seconds.

(b)



$$X = 4 \Rightarrow Z = \frac{(4 - 6.5)}{2.3} = -1.09$$

$$X = 10 \Rightarrow Z = \frac{(10 - 6.5)}{2.3} = 1.52$$

$$P(Z < -1.09) = 0.1379$$

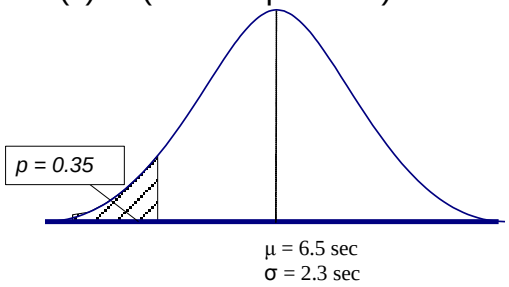
$$P(Z < 1.52) = 0.9357$$

$$\Rightarrow P(-1.9 < Z < 1.52) = 0.9357 - 0.1379$$

$$= 0.7978$$

That is, about 80% of downloads take between 4 and 10 seconds.

(c) (Inverse problem)



$$For\ P = 0.35, Z = -0.38$$

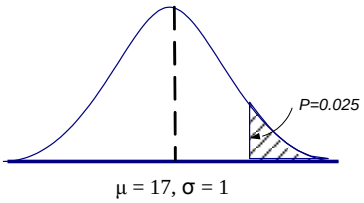
$$\Rightarrow X = -0.38 \times 2.3 + 6.5$$

$$= 5.63$$

That is, 35% of downloads are completed in 5.6 seconds or less.

4. (This is an inverse problem)

$$\mu = 17, \sigma = 1$$

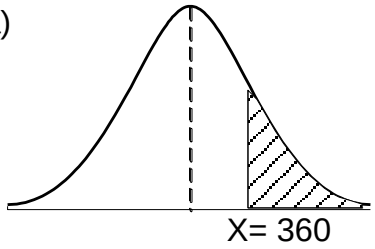


The z value (standard normal score) that cuts off the upper 2.5% is 1.96, so we find the value for potassium which is 1.96 SDs above the mean for a healthy person.

Start with $Z = \frac{x - \mu}{\sigma}$ *ie* $1.96 = \frac{x - 17}{1.0} \Rightarrow x = 1.96 \times 1.0 + 17 = 18.96$

That is X = 18.96 mg/100 ml.

5. (a)



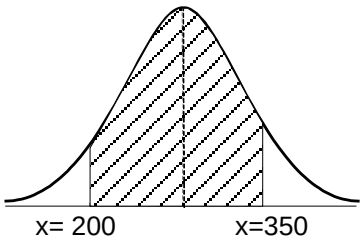
mean = 279

s.d. = 90

$$X = 360 \Rightarrow Z = \frac{(360 - 279)}{90} = 0.9$$

$$P(Z > 0.9) = 1 - 0.8159 = 0.1841$$

(b)



mean = 279

s.d. = 90

$$X = 200 \Rightarrow Z = \frac{(200 - 279)}{90} = -0.88$$

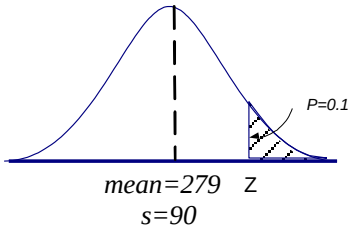
$$P(Z < -0.88) = 0.1894$$

$$X = 350 \Rightarrow Z = \frac{350 - 279}{90} = 0.79$$

$$P(Z < 0.79) = 0.7852$$

$$Hence\ P(-0.88 < Z < 0.79) = 0.7852 - 0.1894 = 0.5958$$

(c)



$$p = 0.1 \Rightarrow Z = 1.28$$

$$Hence, 1.28 = \frac{X - 279}{90}$$

$$\Rightarrow X = 1.28 \times 90 + 279 = 394.2$$

ie 10% of households use over 394 litres.

Binomial Distribution

- Applied to single variable discrete data where results are the numbers of "successful outcomes" in a given scenario.
  e.g.: no. of times the lights are red in 20 sets of traffic lights,
  no. of students with green eyes in a class of 40,
  no. of plants with diseased leaves from a sample of 50 plants

- Used to calculate the probability of occurrences *exactly, less than, more than, between* given values
  e.g. the "probability that the number of red lights will be exactly 5"
  "probability that the number of green eyed students will be less than 7"
  "probability that the no. of diseased plants will be more than 10"

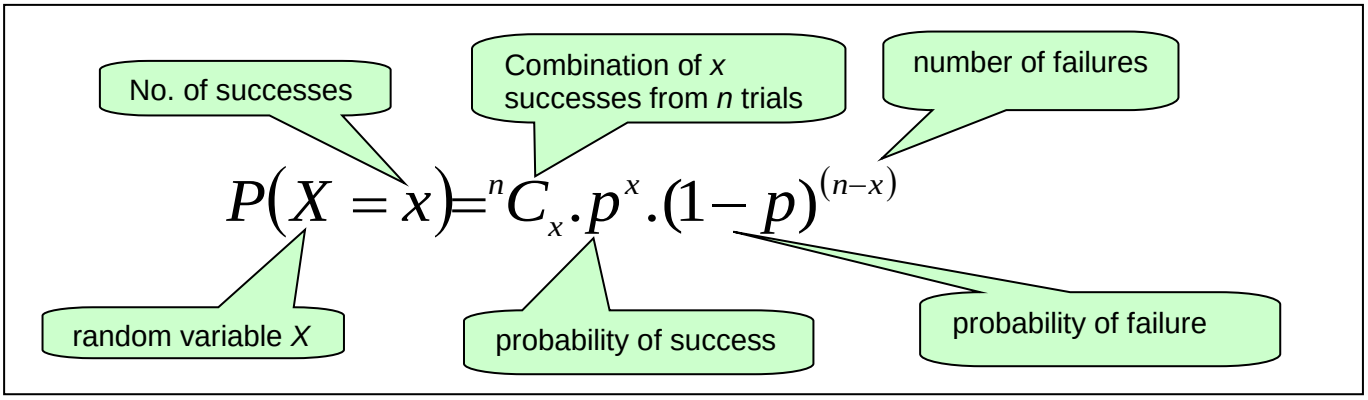
- Parameters, statistics and symbols involved are:

	population parameter symbol	sample statistic symbol
probability of success	π	p
sample size	N	n

- Other symbols:
  X , the number of successful outcomes wanted

  ${}^nC_x$, *or* ${}^nC_r$: the number of ways in which x "successes" can be chosen from a sample of size n. The ${}^nC_r$ key on your calculator can be used directly in the formula.

- Formula used:



Read as "the probability of getting 'x' successes is equal to the number of ways of choosing 'x' successes from *n* trials *times* the probability of success to the power of the number of successes required *times* the probability of failure to the power of the number of resulting failures."

Example

An automatic camera records the number of cars running a red light at an intersection (that is, the cars were going through when the red light was against the car). Analysis of the data shows that on average 15% of light changes record a car running a red light. Assume that the data has a binomial distribution. What is the probability that in 20 light changes there will be exactly three (3) cars running a red light?

Write out the key statistics from the information given:

$p = 0.15,\ n = 20,\ X = 3$

Apply the formula, substituting these values: $\boldsymbol{P}(\boldsymbol{X} = 3) = {}^{20}\boldsymbol{C}_3 \times 0.15^3 \times 0.85^{17} = 0.243$

That is, the probability that in 20 light changes there will be exactly three (3) cars running a red light is 0.24 (24%).

Practice (Binomial Distribution)

1 *Executives in the New Zealand Forestry Industry claim that only 5% of all old sawmills sites contain soil residuals of dioxin (an additive previously used for anti-sap-stain treatment in wood) higher than the recommended level. If Environment Canterbury randomly selects 20 old saw mill sites for inspection, assuming that the executive claim is correct*:

a) Calculate the probability that less than 1 site exceeds the recommended level of dioxin.

b) Calculate the probability that less than or equal to 1 site exceed the recommended level of dioxin.

c) Calculate the probability that at most (i.e., maximum of) 2 sites exceed the recommended level of dioxin.

2 *Inland Revenue audits 5% of all companies every year. The companies selected for auditing in any one year are independent of the previous year's selection*.

a) What is the probability that the company 'Ross Waste Disposal' will be selected for auditing exactly twice in the next 5 years?

b) What is the probability that the company will be audited exactly twice in the next 2 years?

c) What is the exact probability that this company will be audited at least once in the next 4 years?

3 *The probability that a driver must stop at any one traffic light coming to Lincoln University is 0.2. There are 15 sets of traffic lights on the journey.*

a) What is the probability that a student must stop at exactly 2 of the 15 sets of traffic lights?

b) What is the probability that a student will be stopped at 1 or more of the 15 sets of traffic lights?

Worked solutions (Binomial Distribution)

1. (a) $n = 20,\ p = 0.05,\ X < 1 \Rightarrow X = 0$

$P(X = 0) = {}^{20}C_0\, 0.05^0 \times 0.95^{20} = 0.3585$

(b) $n = 20,\ p = 0.05,\ X \le 1 \Rightarrow X = 0,1$

$P(X = 0) = 0.3584$

$P(X = 1) = {}^{20}C_1\, 0.05^1 \times 0.95^{19} = 0.3774$

$\Rightarrow P(X = 0,1) = 0.3585 + 0.3774 = 0.7359$

(c) $n = 20,\ p = 0.05,\ X \le 2 \Rightarrow X = 0,1,2$

$P(X = 0) = 0.3584$

$P(X = 1) = 0.0.3774$

$P(X = 2) = {}^{20}C_2\, 0.05^2 \times 0.95^{18} = 0.1887$

$\Rightarrow P(X = 0,1,2) = 0.3585 + 0.3774 + 0.1887 = 0.9246$

2. (a) $p = 0.05, n = 5, X = 2$

$\Rightarrow P(X = 2) = {}^{5}C_2 \times 0.05^2 \times 0.95^3 = 0.0214$ i.e. ≈ 2%

(b) $P(2\ in\ two\ years) = {}^{2}C_2 \times 0.05^2 \times 0.95^0 = 0.0025$ i.e. ≈ 0.25%

(c) $P(at\ least\ once) = P(X \ge 1)$

$= 1 - P(X = 0) = 1 - {}^{4}C_0 0.05^0 0.95^4$

$= 0.1855$

3. (a) $p = 0.2, n = 15, X = 2$

$\Rightarrow P(X = 2) = {}^{15}C_2 \times 0.2^2 \times 0.8^{13} = 0.2309$

i.e. ≈ 23%

(b) $P(X \ge 1) = 1 - P(X < 1) = 1 - P(X = 0)$

$= 1 - {}^{15}C_0 \times 0.2^0 \times 0.8^{15} = 0.9648$ i.e.≈ 97%