

Statistics Lesson #6:

Confidence Intervals and the Normal Approximation to the Binomial Distribution

Confidence Intervals

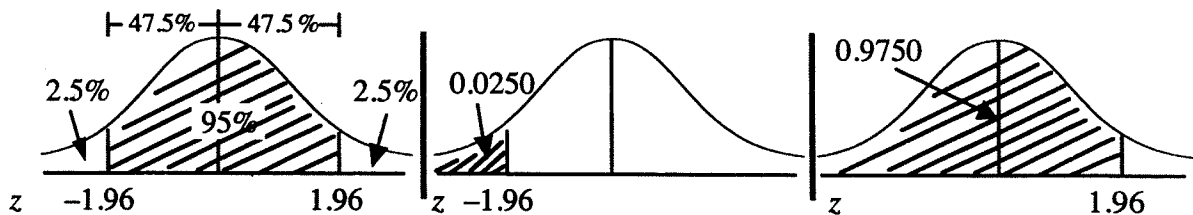
An important aspect of statistics is the level of confidence one has in making predictions, or inferences on a population based on the factual data from the sample.

When data is gathered from a sample of a population, measures of central tendency and dispersion (and other measures) are determined. From the sample, data which is known to be true is then used to make inferences, or predictions on the population.

The level of confidence one has in which they know what is true for the sample may also be true for a population is called a **confidence interval**. A confidence interval is symmetrical about the mean.

95% Confidence Intervals

A 95% confidence interval has 95% of the data contained symmetrical about the mean. This means there is 47.5% to the left of the mean and 47.5% to the right of the mean.



The z -scores for a 95% interval are -1.96 and 1.96 .

Therefore the data interval for a 95% confidence interval is represented as $\mu \pm 1.96\sigma$.

We will use the data and scenario from Lesson 3 without the histogram and the data.

“Glowbright is the name of a 60 watt light bulb manufactured by Duralong Bulbs Inc. The company tested 44 bulbs to determine the mean life of the bulbs and their standard deviation. The lifetime, in hours, of the 44 bulbs tested is shown.”

In lesson 2 we calculated the mean and standard deviation of the data to be approximately $\mu = 900$ and $\sigma = 50$.

a) Calculate the 95% confidence interval for this scenario.

$$\text{Lower: } \text{invNorm}(0.025, 900, 50) = 802$$

$$\text{Upper: } 1 - 0.025 = 0.975$$

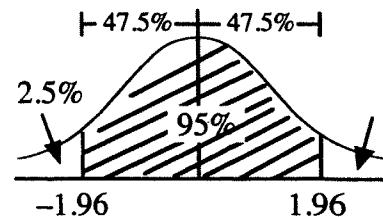
$$\text{invNorm}(0.975, 900, 50) = 998$$

b) Write in words what the confidence interval means for this scenario.

Glowbright can be 95% confident that their bulbs will last between 802 and 998 hours.

95% Confidence Interval Notes

1. The z -scores for a 95% confidence interval are ± 1.96 .
2. The data calculation for a 95% confidence interval is $\mu \pm 1.96\sigma$.
3. Always round the lower limit of the data interval down and the upper limit of the data interval up.
4. The confidence interval express the level of certainty that what is true for the sample mean is also true for a population mean.



An agriculturist recorded data on the output of an ancient grain, quinoa, on 100 sections out of 500 sections of land. She determined that the average production of quinoa is 35.2 kL per section with a standard deviation of 3.2 kL. She also determined that the data was normally distributed.

$$\mu = 35.2 \quad \sigma = 3.2$$

- a) Calculate the numerical 95% confidence interval for the average number of kL of quinoa per section of land in 500 sections of land.

$$\text{Lower: } \text{invNorm}(0.025, 35.2, 3.2) = 28.92 \\ = 28$$

$$\text{Upper: } \text{invNorm}(0.975, 35.2, 3.2) = 41.47 \\ = 42$$

- b) Write a concluding statement for a).

The agriculturist can be 95% confident that the output of quinoa on 500 sections of land will be between 28 kL and 42 kL.

Complete Assignment Questions #1 - #2

The Binomial Distribution (Review from Probability Lesson 1)

The **binomial distribution** is a type of probability distribution based on a binomial experiment. A **binomial experiment** is a probability experiment with the following conditions:

1. There are a fixed number, n , of identical trials in the experiment.
2. Only 2 outcomes are possible at each trial. The outcome of each trial is classified as:
 - “success” (the occurrence of a specified event),
 - or*
 - “failure” (the non-occurrence of a specified event).
3. The trials are independent.
4. In each trial, the probability of success, (denoted by p), and the probability of failure, (denoted by $1 - p$) remains the same from trial to trial.
5. The variable is the number of successes in n trials.

Calculations with a Binomial Distribution

Consider the following problem.

“ A professor gives a unit test consisting of 50 multiple choice questions.
Each question has four responses, only one of which is the correct answer.

What is the probability that a student passes the test by guessing each of the answers?”

a) Complete the following to review a binomial experiment:

- A trial for this experiment is a multiple choice question.
- Each trial is identical.
- The fixed number of trials for this experiment is 50.
- At each trial only two outcomes are possible. The “success” is the event of getting a multiple choice question correct. The “failure” is getting a multiple choice question wrong.
- The trials are independent of one another.
- In each trial the probability of success is 0.25.
- In each trial the probability of failure is 0.75.

b) In this problem we are required to calculate the probability of 25 or more successes. That is, we would have to determine the probability of the following:

$P(25 \text{ right and } 25 \text{ wrong}) + P(26 \text{ right and } 24 \text{ wrong}) + P(27 \text{ right and } 23 \text{ wrong}) + \dots + P(50 \text{ right and } 0 \text{ wrong})$

This calculation would require 26 separate independent probability questions added up!
Or we could use the binomial pdf feature and sum the probabilities (page 61).

It has also been discovered that instead of doing binomial calculations, we can use the normal distribution curve to approximate an answer for a binomial distribution.

This approximation can only be used when the data is large and if the conditions $np > 5$ and $n(1 - p) > 5$ are met. For this level of mathematics it will be assumed that all binomial data given fulfill these conditions and thus the normal distribution can be used to answer and approximate binomial distribution problems.

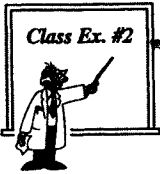
The Normal Approximation to the Binomial Distribution

In order to use the normal distribution curve for a binomial distribution, the normal distribution requires a mean and standard deviation.

The mean and standard deviation for binomial data ONLY which will be approximated by using the normal distribution curve are the same formulas from Lesson 2 of this unit.

Mean and Standard Deviation of the Binomial Probability Distribution ONLY

$$\mu = np \qquad \sigma = \sqrt{np(1 - p)}$$



Consider the problem discussed earlier

“A professor gives a unit test consisting of 50 multiple choice questions. Each question has four responses, only one of which is the correct answer.

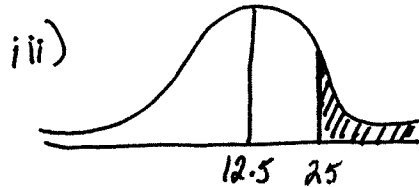
What is the probability that a student passes the test by guessing each of the answers?”

- a) Answer the problem, in percent, to four decimal places, using the normal approximation to the binomial distribution.

$$p = \frac{1}{4} = 0.25 \quad 1-p = 1 - \frac{1}{4} = \frac{3}{4} = 0.75$$

i) $\mu = np$
 $= 50(0.25)$
 $= 12.5$

ii) $\sigma = \sqrt{np(1-p)}$
 $= \sqrt{50(0.25)(1-0.25)}$
 $= 3.061862178...$
 $\hat{=} 3.06$



iv) $P(x > 25) = \text{normalcdf}(25, 10^9, 12.5, 3.06)$
 $= 0.00002205 = 0.0022\%$
 $= 0.0022\%$ probability that a student will pass the test by guessing

- b) Find a 95% confidence for the average mark on this test if all of the students guessed.

Lower $\rightarrow \text{invNorm}(0.0250, 12.5, 3.06) = 6.50... \text{ Round Lower DOWN} \rightarrow 7$

Upper $\rightarrow \text{invNorm}(0.9750, 12.5, 3.06) = 18.49... \text{ Round Upper UP} \rightarrow 19$

We will be 95% confident that between 7 to 19 students will pass the test if they all guess at the answers.



Statement:

The probability of rain in a region of Brazil is 0.15 on any given day.

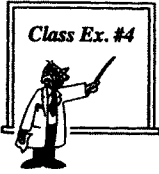
Consider the following question to the above statement.

“What is the average number of days a person who lives in this region could expect to experience weather other than rain in a year of 365 days?”

- a) Is there enough information in the statement and question to consider this a binomial experiment? *Yes, only 2 outcomes; rain or no rain*

- b) Answer the question in the quotation marks by writing a concluding statement about the rain in the region for the year.

$n = 365$ $p = 0.85$ $1-p = 0.15$	$\mu = np$ $= 365(0.85)$ $\mu = 310.25$	On average, a person can expect weather other than rain on approximately 310 days of the year in this region of Brazil
---	---	--



Reasearchers studied the length of whiskers in seals. They found that the gene for the length of whiskers in seals has two variations. One variation would be the dominant gene for long whiskers (code W) within a certain interval. The other variation would be for the recessive gene (code w) within a certain interval.

- a) A Punnet square is a type of grid used in genetics to indicate all the possible outcomes of a genetic cross of a dominant and recessive genes, also called the checkerboard. Complete the Punnet square to show the sample space for the offspring of two seal parents who both carry the dominant (W) gene for the long whiskers.

		Father Seal	
		W	w
Mother Seal	W	WW	wW
	w	Ww	ww

- b) Calculate the probability that one offspring from these parents will have long whiskers. Note: Having at least one dominant gene in the pairing means a baby seal will have long whiskers.

$$P(\text{long}) = \frac{3}{4} = 0.75$$

- c) The researchers determined in their original sample that approximately 35% of the seals have short whiskers. Calculate, to the nearest hundredth, the mean and standard deviation for the number of seals in a new sample of 1075 seals that will have short whiskers.

$$n = 1075$$

$$p = 0.35$$

$$1-p = 0.65$$

$$\mu = np$$

$$= 1075(0.35)$$

$$= 376.25$$

$$\sigma = \sqrt{np(1-p)}$$

$$= \sqrt{1075(0.35)(0.65)}$$

$$= 15.64$$

- d) Calculate and describe in words the following:

- i) The 95% confidence interval for the **number** of seals that will have short whiskers in the new sample.

$$\text{Lower: } \text{invNorm}(0.025, 376.25, 15.64) = 345.6$$

$$= 345$$

$$\text{Upper: } \text{invNorm}(0.975, 376.25, 15.64) = 406.9$$

$$= 407$$

We can be 95% confident that between 345 and 407 seals will have short whiskers.

- ii) The 95% confidence interval for the **percentage**, to the nearest tenth of a percent, of seals that will have short whiskers in the new sample.

$$\text{Lower: } \frac{345}{1075} = 32.1\%$$

$$\text{Upper: } \frac{407}{1075} = 37.9\%$$

We can be 95% confident that between 32.1% and 37.9% of the seals will have short whiskers.

Complete Assignment Questions #3 - #10

Assignment

1. A study done on Calgary Transit has found that for a sample of 35 buses, 175 people on average use the bus each hour with a standard deviation of 14 people. Determine a 95% confidence interval for the population average number of people of the entire fleet of busses.

$$\mu = 175, \sigma = 14$$

$$\text{Lower: invNorm}(0.025, 175, 14) = 147.56 \\ = 147$$

$$\text{Upper: invNorm}(0.975, 175, 14) = 202.44 \\ = 203$$

We can be 95% confident that between 147 and 203 people will ride the bus each hour.

2. The average level of contaminants found in a sample of hydroplurate, an new chemical additive, was 8.2% with a standard deviation of 0.5%. Determine a 95% confidence interval, to the nearest percent, for the population average of impurities in hydroplurate

$$\mu = 8.2, \sigma = 0.5$$

$$\text{Lower: invNorm}(0.025, 8.2, 0.5) = 7.22 \\ = 7$$

$$\text{Upper: invNorm}(0.975, 8.2, 0.5) = 9.18 \\ = 10$$

We can be 95% confident that the average level of impurities is from 7% to 10%.

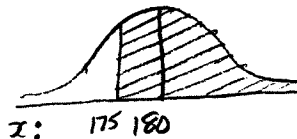
3. A web site manager has determined that she will complete 90% of internet orders to customers within one day of receiving the order. The web site manager expects 200 new customers in the next week. $n = 200, p = 0.9, 1 - p = 0.1$

- a) Calculate; i) the mean and ii) standard deviation to the nearest hundredth.

$$\mu = np \\ = 200(0.9) \\ = 180$$

$$\sigma = \sqrt{np(1-p)} \\ = \sqrt{200(0.9)(0.1)} \\ = 4.24$$

- b) Calculate the probability, to four decimal places, that sales are completed to 175 or more customers.



$$P(x > 175) = \text{normalcdf}(175, 1 \times 10^{99}, 180, 4.24) \\ = 0.8808$$

There is 0.8808 probability that sales are completed to 30 or fewer customers.

4. From previous records, 90% of the students who take "Easydrive" instruction, passed the provincial driving test at the first attempt. If 1600 students took the "Easydrive" instruction determine; $n=1600, p=0.9, 1-p=0.1$

- a) the mean and standard deviation of the number of students who passed the provincial driving test at the first attempt.

$$\begin{aligned} \mu &= np \\ &= 1600(0.9) \\ &= 1440 \end{aligned} \qquad \begin{aligned} \sigma &= \sqrt{np(1-p)} \\ &= \sqrt{1600(0.9)(0.1)} \\ &= 12 \end{aligned}$$

- b) a 95% confidence interval for the number of students who passed the provincial driving test at the first attempt.

Lower: $\text{inv Norm}(0.025, 1440, 12) = 1416.48$
 $= 1416$

Upper: $\text{inv Norm}(0.975, 1440, 12) = 1463.52$
 $= 1464$

- c) a 95% confidence interval for the percentage of students who passed the provincial driving test at the first attempt.


Lower $\frac{1416}{1600} = 88.5\%$ We can be 95% confident that between 88.5% and 91.5% of the students will pass on their first try.

Upper $\frac{1464}{1600} = 91.5\%$

5. Dennis surveyed a group of students for a statistics project and determined that 6 out of every 10 students ate their lunches at least once a week at a mall located across the street from his school. The school has a population of 1 920 students.


$n = 1920, p = 0.6, 1-p = 0.4$

- a) Find the probability, to the nearest hundredth of a percent, that more than 1 100 students eat their lunch at least once a week at the mall.

$$\begin{aligned} \mu &= np \\ &= 1920(0.6) \\ &= 1152 \end{aligned} \qquad \begin{aligned} \sigma &= \sqrt{np(1-p)} \\ &= \sqrt{1920(0.6)(0.4)} \\ &= 21.47 \end{aligned}$$


$x: 1100 \quad 1152$
 $P(x > 1100) = \text{normalcdf}(1100, 10^{99}, 1152, 21.47)$
 $= 0.9923$
 $= 99.23\%$

- b) Find the probability, to the nearest hundredth of a percent, that between 900 students and 1200 students eat their lunch at least once a week at the mall.



$x: 900 \quad 1152 \quad 1200$

$$\begin{aligned} P(900 < x < 1200) &= \text{normalcdf}(900, 1200, 1152, 21.47) \\ &= 0.9873 \\ &= 98.73\% \end{aligned}$$

- c) Determine a 95% confidence interval for the average number of students who eat their lunch at least once a week at the mall.

Lower: $\text{inv Norm}(0.025, 1152, 21.47) = 1109$

Upper: $\text{inv Norm}(0.975, 1152, 21.47) = 1195$

We can be 95% confident that the average number of students who eat lunch at the mall at least once a week is from 1109 to 1195.

6. 14 out of every 20 high school graduates in a large city take post secondary courses.

a) Determine the probability that, in a random selection of 90 high school graduates, at least 65 are taking post secondary courses. Answer to the nearest hundredth.

$$\begin{aligned}
 n &= 90 & \mu &= np & \sigma &= \sqrt{np(1-p)} \\
 p &= \frac{14}{20} = 0.7 & &= 90(0.7) & &= \sqrt{90(0.7)(0.3)} \\
 1-p &= 0.3 & &= 63 & &= 4.35
 \end{aligned}$$

$P(X > 65) = \text{normalcdf}(65, 10^{99}, 63, 4.35)$
 $= 0.32$

b) Determine the probability that, in a random selection of 150 high school graduates, less than 100 of them are taking post secondary courses. Answer to the nearest hundredth of a percent.

$$\begin{aligned}
 n &= 150 & \mu &= np & \sigma &= \sqrt{np(1-p)} \\
 p &= 0.7 & &= 150(0.7) & &= \sqrt{150(0.7)(0.3)} \\
 1-p &= 0.3 & &= 105 & &= 5.61
 \end{aligned}$$

$P(X < 100) = \text{normalcdf}(-1 \times 10^{99}, 100, 105, 5.61)$
 $= 0.1864 = 18.64\%$

c) Determine a 95% confidence interval for a school population of 1200 in the same city for the average number of students who take post secondary courses

$$\begin{aligned}
 n &= 1200 & \mu &= np & \sigma &= \sqrt{np(1-p)} \\
 p &= 0.7 & &= 1200(0.7) & &= \sqrt{1200(0.7)(0.3)} \\
 1-p &= 0.3 & &= 840 & &= 15.87
 \end{aligned}$$

Lower: $\text{invNorm}(0.025, 840, 15.87) = 808$
 Upper: $\text{invNorm}(0.975, 840, 15.87) = 872$
 We can be 95% confident that on average 808 to 872 students take post secondary courses.

7. Of the voters in a certain town 60% are in favour of building a ring road to by-pass the town. A random sample of 105 voters is interviewed. Calculate the probability, to the nearest hundredth that;

of a percent,

a) at least two-thirds of the sample are in favour of the ring road

$$\begin{aligned}
 \mu &= np & &= \frac{2}{3}(105) = 70 & P(X > 70) &= \text{normalcdf}(70, 10^{99}, 63, 5.02) \\
 &= 105(0.6) & & & &= 0.0816 \\
 &= 63 & & & &= 8.16\% \\
 \sigma &= \sqrt{np(1-p)} & & & & \\
 &= \sqrt{105(0.6)(0.4)} & & & & \\
 &= 5.02 & & & &
 \end{aligned}$$

The probability that two-thirds favour a ring road is 8.16%.

b) a majority of the sample are not in favour of the ring road

$$\begin{aligned}
 \mu &= np & \text{majority} &= \frac{105}{2} & P(X > 52) &= \text{normalcdf}(52, 10^{99}, 42, 5.02) \\
 &= 105(0.4) & &= 52.5 & &= 0.0142 \\
 &= 42 & &= 53 & &= 1.42\% \\
 \sigma &= \sqrt{np(1-p)} & & & & \\
 &= \sqrt{105(0.4)(0.6)} & & & & \\
 &= 5.02 & & & &
 \end{aligned}$$

The probability that a majority are not in favour is 1.42%

8. Researchers studied the colour of leaves on a new plant discovered in the last 10 years. The plant leaves have two colour variations. One variation would be the dominant gene for solid green (code G) and the other variation is for the recessive gene for striped green code (g) The other variation would be for the recessive gene (code w) within a certain interval.

- a) Complete the checkerboard to show the sample space for the offspring of two plants who both carry the dominant (G) gene for the solid green leaves.
- b) Calculate the probability that one plant from these parents will have striped leaves.

		Solid Green leaves	
		G	g
Striped Green Leaves	G	GG	gG
	g	Gg	gg

$$P(\text{striped}) = \frac{1}{4} = 0.25$$

- c) The researchers determined in their original sample that approximately 70% of the plants have solid green leaves. Calculate the mean and standard deviation (to two decimal places) for the number of leaves in a new sample of 750 plants that will have solid green leaves.

$$\begin{aligned} n &= 750 & \mu &= np & \sigma &= \sqrt{np(1-p)} \\ p &= 0.7 & &= 750(0.7) & &= \sqrt{750(0.7)(0.3)} \\ 1-p &= 0.3 & &= 525 & &= 12.55 \end{aligned}$$

- d) Calculate and describe in words the following:

- i) The 95% confidence interval for the **number** of plants that will have solid green leaves in the new sample.

$$\text{Lower: } \text{invNorm}(0.025, 525, 12.55) = 500$$

$$\text{Upper: } \text{invNorm}(0.975, 525, 12.55) = 550$$

The researchers can be 95% confident that between 500 and 550 plants will have solid green leaves.

- ii) The 95% confidence interval, to the nearest whole percent, for the **percentage** of leaves that will have solid green leaves in the new sample.

$$\text{Lower} = \frac{500}{750} = 66\%$$

$$\text{Upper} = \frac{550}{750} = 74\%$$

The researchers can be 95% confident that between 66% and 74% of the plants will have solid green leaves.

Use the following information to answer questions #9 and #10

10 % of the chocolates produced at Cardy's Chocolates are labeled as defective because of their shape (not their taste). Cardy's Chocolates takes these chocolates and sells them at a 50% discount under the label "Not so Perfect Chocolates". Cardy's Chocolates plans to produce 525 chocolates for their upcoming Valentine's day special.

Multiple Choice

9. The mean, to the nearest whole chocolate, for the number of **non defective** chocolates are

- A. 7 $n=525$ $\mu=np$
 B. 22 $p=0.9$ $=525(0.9)$
 C. 53 $1-p=0.1$ $=472.5$
 D. 473

Numerical Response

10. The standard deviation, to the nearest tenth for the number of **defective** chocolates is _____ .

(Record your answer in the numerical response box from left to right)

6	.	9	
---	---	---	--

$$\begin{aligned} \sigma &= \sqrt{np(1-p)} \\ &= \sqrt{525(0.1)(0.9)} \\ &= 6.9 \end{aligned}$$

Assignment Key

Note: Answers may vary slightly if the z-score tables are used instead of the normal distribution features of a graphing calculator.

1. 147 to 203 2. 7% to 10% 3. a) i) 180 ii) 4.24 b) 0.8808
 4. a) $\mu = 1440, \sigma = 12$ b) 1416 to 1464 c) 88.5% to 91.5%
 5. a) 99.23% b) 98.73% c) 1109 to 1195
 6. a) 0.32 b) 18.64% c) 808 to 872 7. a) 8.16% b) 1.42%
 8. a) see chart below b) $\frac{1}{4}$ c) $\mu = 525, \sigma = 12.55$ d) i) 500 to 550 ii) 66% to 74%

Solid Green leaves		
	G	g
Striped Green Leaves	G	g
	GG	Gg
	gG	gg

9. D

10.

6	.	9	
---	---	---	--