5.1 – Mean, Median, & Mode

definitions

Mean: Mean is <u>average</u>. It is calculated by adding all values and dividing by the total number of values.

$$\underset{\overset{\nearrow}{\text{mean}}}{\mu} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Median: the middle value in a data set. If there are an even number of values, the median is the mean of the two middle values The best way to find the median is to arrange values from smallest to largest.

ex $2, 3, \underset{\text{median}}{4}, 5, 6$ ex $2, 3, 4, 5, 6, 7$

$\frac{4+5}{2} = 4.5$ median

Mode: the value that occurs most often in a set of values. There can be more than one mode.

ex $2, 3, 3, 3, 3, 5$

Mode = 3

ex $1, 1, 2, 3, 4, 4, 5$

Modes: 1 and 4

Example 1 – The Blue Jays score these amounts of runs in their last 9 games:

4, 7, 2, 4, 10, 5, 6, 7, 7

Find the **mean, median,** and **mode:**

$$\text{Mean} = \mu = \frac{4+7+2+4+10+5+6+7+7}{9} = 5.78 \text{ runs per game}$$

Median: $2, 4, 4, 5, 6, \underbrace{7, 7, 7}_{\text{mode}}, 10$ Median = 6 runs

Mode: 7 runs

Example 2 – Cheesy Burgers Restaurant gets the following reviews out of 10:

8, 6, 9, 7, 8, 4, 10, 7

Find the **mean, median,** and **mode:**

$$\mu = \frac{8+6+9+7+8+4+10+7}{8} = 7.375$$

median: $4, 6, 7, 7, 8, 8, 9, 10$ $\frac{7+8}{2} = \frac{15}{2} = 7.5$

mode: 7 and 8

Example 3 – For 30 randomly selected high school students, the following IQ frequency distribution was obtained. Determine the **mean, median,** and **mode.**

Class Limits	Frequency
$80 \le x < 90$	2
$90 \le x < 100$	9
$100 \le x < 110$	11
$110 \le x < 120$	5
$120 \le x < 130$	2
$130 \le x < 140$	1

for each range of class limits, use the middle value

ex for 80-90, use 85

raw data: 85, 85, 95, 95, 95, 95, 95, 95, 95,
95, 95, 105, 105, 105, 105, 105, 105,
105, 105, 105, 105, 105, 115, 115,
115, 115, 115, 125, 125, 135

$$\mu = \frac{2(85) + 9(95) + 11(105) + 5(115) + 2(125) + 135}{30} = 104.67$$

Median: 105

Mode: 105

Example 4 – 10 numbers have a mean of 37. If one number is removed, the mean is 38. What was the number that was removed?

Sum of 10 numbers $= 10 \times 37 = 370$

Sum of 9 numbers $= 9 \times 38 = 342$

Number removed was: $370 - 342 = \boxed{28}$

Example 5 – The mean age of 4 people is 39.25. The ages of three of the people are 20, 35, and 60. What is the age of the fourth person?

$$\mu = \frac{x_1 + x_2 + x_3 + x_4}{4}$$

$$39.25 = \frac{20 + 35 + 60 + x_4}{4}$$

$$39.25 = \frac{115 + x_4}{4}$$

$$4(39.25) = \frac{\cancel{4}(115 + x_4)}{\cancel{4}}$$

$$157 = \cancel{115} + x_4$$
$$-115 \quad \cancel{-115}$$

$$\boxed{x_4 = 42}$$

Check: $\frac{20 + 35 + 60 + 42}{4} = 39.25$ ✓

5.2 – Standard Deviation

definition

What is **standard deviation?** It's a measure of the dispersion of data about the mean

ex 1, 7, 13, 19, 25 AND 11, 12, 13, 14, 15

$\mu = 13$ data more widely dispersed so larger standard deviation

$\mu = 13$ data not widely dispersed so small standard deviation

How do you calculate **standard deviation?**

σ = sigma = standard deviation

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_n-\mu)^2}{n}}$$

Why is **standard deviation** useful in **statistics?**

- it helps to explain the nature of your data (is it generally grouped together, or spread apart?)
- it helps to make predictions about new/future data points

Example 1 – Find the mean for each set of data.

a) 5, 6, 7, 8, 9

$$\mu = \frac{5+6+7+8+9}{5}$$

$$\mu = 7$$

b) 3, 5, 7, 9, 11

$$\mu = \frac{3+5+7+9+11}{5}$$

$$\mu = 7$$

Analyzing the data, can you predict which set will have a larger standard deviation? How can you tell? (b) will have a larger σ as the data is more widely dispersed.

Calculate the **standard deviation** for each:

* Notice the data is (b) is dispersed by 'double' so σ for (b) is double σ for (a)

a)
$$\sigma = \sqrt{\frac{(5-7)^2+(6-7)^2+(7-7)^2+(8-7)^2+(9-7)^2}{5}}$$

$$\sigma = \sqrt{\frac{(-2)^2+(-1)^2+(0)^2+(1)^2+(2)^2}{5}}$$

$$\sigma = \sqrt{\frac{4+1+0+1+4}{5}} = \sqrt{\frac{10}{5}} = \sqrt{2}$$

$$\sigma = 1.414$$

(b)
$$\sigma = \sqrt{\frac{(3-7)^2+(5-7)^2+(7-7)^2+(9-7)^2+(11-7)^2}{5}}$$

$$\sigma = \sqrt{\frac{(-4)^2+(-2)^2+(0)^2+(2)^2+(4)^2}{5}}$$

$$\sigma = \sqrt{\frac{16+4+0+4+16}{5}} = \sqrt{\frac{40}{5}} = \sqrt{8}$$

$$\sigma = 2.83$$

Example 2 – Calculate the **standard deviation** for the following sets of data:

Daily Commute Time (mins)	Number of Employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2
Total	25

Daily Commute Time (mins)	Number of Employees
0 to less than 10	2
10 to less than 20	10
20 to less than 30	9
30 to less than 40	3
40 to less than 50	1
Total	25

First, predict which table will have the higher standard deviation. Why?

First table as data is more dispersed among different daily commute times

1ST Table

data : 5,5,5,5, 15,15,15,15,15,15,15,
15,15, 25,25,25,25,25,25,35,
35,35,35,45,45

$$\mu = \frac{535}{25} = 21.4 \text{ minutes}$$

same mean !

$$\sigma = \sqrt{\frac{4(5-21.4)^2 + 9(15-21.4)^2 + 6(25-21.4)^2 + 4(35-21.4)^2 + 2(45-21.4)^2}{25}}$$

$$\sigma = \sqrt{\frac{1075.84 + 368.64 + 77.76 + 739.84 + 1113.92}{25}}$$

$$\boxed{\sigma = 11.62}$$

↖ first table has higher σ

2nd Table

data: 5,5,15,15,15,15,15,15,15,15,15,15,
25,25,25,25,25,25,25,25,25,35,35,
35,45

$$\mu = \frac{535}{25} = 21.4 \text{ mins}$$

$$\sigma = \sqrt{\frac{2(5-21.4)^2 + 10(15-21.4)^2 + 9(25-21.4)^2 + 3(35-21.4)^2 + (45-21.4)^2}{25}}$$

$$\sigma = \sqrt{\frac{537.92 + 409.6 + 116.64 + 554.88 + 556.96}{25}}$$

$$\boxed{\sigma = 9.33}$$

What would a standard deviation of zero signify?

Signifies that every data point is the same.

ex How many times did you brush your teeth each day? 2, 2, 2, 2, 2, 2 $\mu = 2$ $\sigma = 0$

Example 3 - Without doing any calculations, what is the relationship between the standard deviation of 1, 2, 3, 4, 5 and the standard deviation of:

a) 17, 19, 21, 23, 25

double the σ of 1,2,3,4,5

b) 100, 105, 110, 115, 120

5 times the σ of 1,2,3,4,5

5.3A – The Normal Distribution Part 1

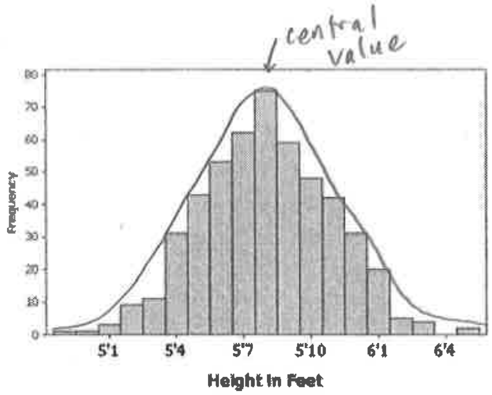
Data can be **distributed** in many ways. It can have many more 'smaller' values than 'larger' values, or *visa versa*. Or, it can be very jumbled up, some smaller values that are common, as well as larger values that are common:

more smaller values

more larger values

jumbled

However, there are many cases where data is symmetrical (or almost) around a central value, and this is called a normal distribution.

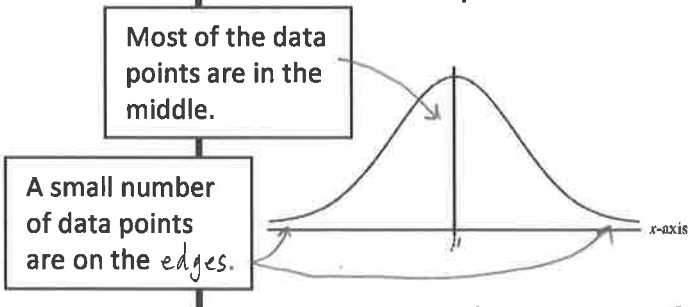


Normal Distribution

definition

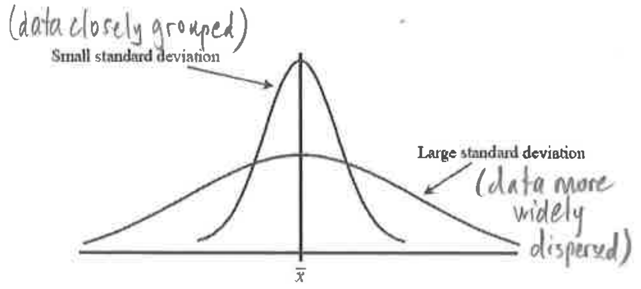
The **Normal Distribution** is a bell shaped curve used in statistical analysis to model the distribution of values in a data set. Examples of data that follow a normal distribution are: heights or weights of people or other species, sizes or weights of goods manufactured at a factory, marks on a test, length of time a battery lasts, milk produced by a cow in a day, etc. If data follows a normal distribution, it is easier to make predictions using the data.

The **Normal Distribution** has the mean (μ) in the middle of the curve, & the shape of the curve is dependent on the standard deviation (σ) of the data set.



Most of the data points are in the middle.

A small number of data points are on the edges.



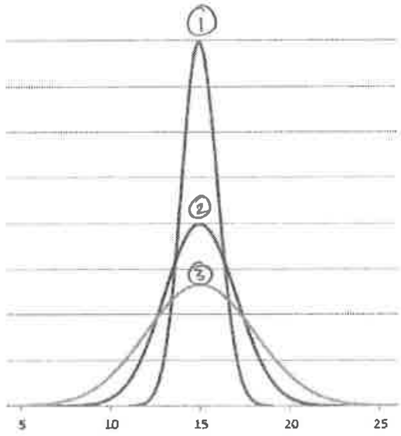
Here are some characteristics of the **Normal Distribution**:
- it is bell shaped and symmetric about the mean
- the shape (tall & skinny OR short & wide) is due to the standard deviation
- the enclosed area under is always equal to 1, which signifies the probability (100%) of a score falling within the bell curve.
- the curve will never touch the *x*-axis; it extends to infinity in both directions

Example 1 – What can you say about the mean (μ) for each normal distribution? What about their standard deviations (σ)?

μ is the same for each curve

Curve 1 has the smallest σ

Curve 3 has the largest σ

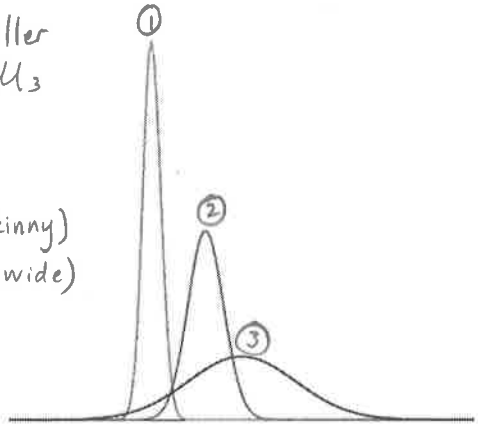


Example 2 – What can you say about the mean for each normal distribution? What about their standard deviations?

μ_1 is smaller than μ_2 which is smaller than μ_3

OR: $\mu_1 < \mu_2 < \mu_3$

Curve 1 has the smallest σ (tall+skinny)
Curve 3 has the largest σ (short + wide)



There are many different normal curves with different μ and σ. By transforming each raw score into a *z*-score, which is a measure of how many standard deviations the value is from the mean, you can get a sense of how far that raw score is from the mean compared to other raw scores.

How to Calculate Z-Scores:

$$Z = \frac{\text{difference between } x \text{ and } \mu}{\text{standard deviation}} = \frac{x-\mu}{\sigma}$$

Z: number of standard deviations that x is away from the mean μ
μ: mean
x : a particular score
σ : standard deviation

Note that there are just as many negative z-scores (when $x < \mu$) as there are positive z-scores.

Example 3 – The test score mean was 76% with a standard deviation of 6%. If somebody scored 67% on their test, what is their **z-score**? What about 80%?

$$Z = \frac{x-\mu}{\sigma} = \frac{67-76}{6} = \boxed{-1.5}$$

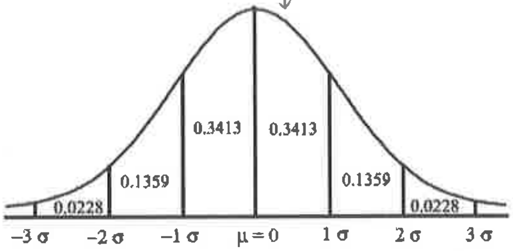
67% is 1 and a half standard deviations below the mean

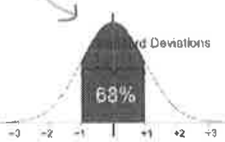
$$Z = \frac{80-76}{6} = 0.\bar{6}$$

80% is $0.\bar{6}$ standard deviations $(\frac{2}{3})$ above the mean

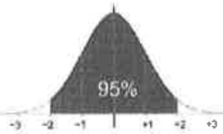
The **Standard Normal Distribution**: The value under each area in the curve represents the decimal value (multiply by 100 for percent) of the number of raw scores in that vicinity.

34% + 34% = 68%

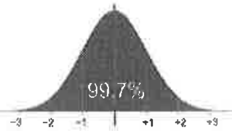




⊛ **68%** of values are within **1 standard deviation** of the mean



⊛ **95%** of values are within **2 standard deviations** of the mean



⊛ **99.7%** of values are within **3 standard deviations** of the mean

Notice that the *z*-score is negative if smaller than the mean.

It is good to know the standard deviation because we can say that any value is: **likely** to be within 1 standard deviation (68% chance), **very likely** to be within 2 standard deviations (95% chance), and **almost certainly** within 3 standard deviations (99.7% chance). Anything outside of this can be deemed an **outlier**.

Example 4 – Professor Hardmarker is marking a test and the following scores out of 60 result: 20, 15, 26, 32, 18, 28, 35, 14, 26, 22, 17.

① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪

a) Any observations?

low set of scores for a test out of 60

b) Calculate the **mean**: $\mu = \frac{253}{11} = 23$

mean of test $= \frac{23}{60}$

c) Calculate the **standard deviation**:

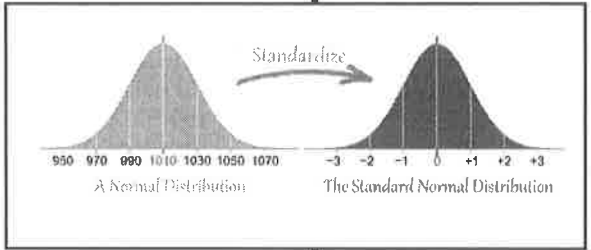
$$\sigma = \sqrt{\frac{(20-23)^2+(15-23)^2+(26-23)^2+(32-23)^2+(18-23)^2+(28-23)^2+(35-23)^2+(14-23)^2+(26-23)^2+(22-23)^2+(17-23)^2}{11}} = 6.63$$

d) The test must have been really hard, so the Prof decides to only fail those who are more than one standard deviation below the average (he **standardizes** the exam). Use **z-scores** to find out how many people will fail:

$Z = \frac{x-\mu}{\sigma}$ ① $Z = \frac{20-23}{6.63} = -0.45$ safe, ② $Z = \frac{15-23}{6.63} = -1.21$ fails

③ $\frac{26-23}{6.63} = 0.45$ safe ④ 1.36 safe ⑤ −0.75 safe ⑥ 0.75 safe ⑦ 1.81 safe

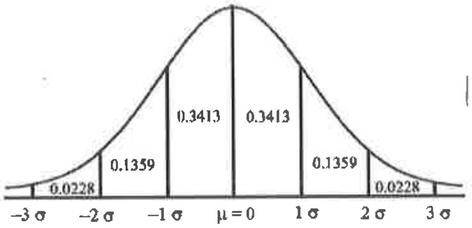
⑧ $\frac{14-23}{6.63} = -1.36$ fails ⑨ 0.45 safe ⑩ −0.15 safe ⑪ −0.90 safe



Instead of 9 students failing the test, after standardizing, only 2 students failed.

5.3B – The Normal Distribution Part 2

A standard normal distribution curve can also be used to estimate probabilities of many different possible results to a statistical study. We already know percentages for z-scores of 1, 2, and 3, as seen below:



What is the percentage that a raw score has a *z*-score below 1?

$100 \times (0.0228 + 0.1359 + 0.3413 + 0.3413)$

$= 84.13\%$

However, what, for instance, is the percentage chance of a raw score that has a *z*-score of below 1.5? We cannot tell using the graphic above. We can use a **'Standard Normal Distribution Table'** to find the answer. The table value always indicates the area or probability to the LEFT of the *z*-score.

$Z = 1.50$ on table $= 0.9332$

$0.9332 \times 100 = 93.32\%$

Example 1 – Using last day's worksheet, let's answer #2e. The math exam had a mean of 62% with a standard deviation of 12. What percentage of students originally earned an A grade (86%)?

% of students below 86% $= 0.0228 + 0.1359 + 0.3413 + 0.3413 + 0.1359$

$= 0.9772 = 97.72\%$

$62 + 12 + 12 = 86\%$

so $86\% = \mu + 2\sigma$

% above 86% $= 100 - 97.72 = \boxed{2.28\%}$

Let's say the university wanted about 10% of the students to get an A in each course. Do you think the professor should standardize the scores (scale the test results)? If so, can you explain how the professor may go about this process (think about the area under the normal distribution and how this can help)?

Yes, he should standardize!

86 is $\mu + 2\sigma$ so currently 97.72% getting below 86

We want 86% to be 0.90

On table 0.90 has a Z score of 1.28

so any score above $\mu + 1.28\sigma$ should be an A

$62 + 1.28(12) = 62 + 15.36 = 77.36\%$

anything above 77.36% is an A

Example 2 – If IQ scores are normally distributed with a mean of 100 and standard deviation of 15, determine:

a) the *z*-score for 128

$Z = \frac{128 - 100}{15} = 1.87$

b) the probability that a randomly selected person has an IQ less than 128

look up $Z = 1.87$ on table : 0.9693

96.93% chance

c) the probability that a randomly selected person has an IQ more than 105

$Z = \frac{105 - 100}{15} = 0.33$

look up $Z = 0.33$ on table : 0.6293

62.93% have an IQ less than 105

so $100 - 62.93 = 37.07\%$ have an IQ more than 105.

Example 3 – Using the **standard normal distribution table**, find:

a) the % that a z-score is below -2.57

-2.57 on table : 0.0051 = 5.1% chance

b) the % that a z-score is above -1.34

-1.34 on table : 0.0901

9.01% below -1.34

so 100 - 9.01 = 90.99% chance above -1.34

c) the % that a z-score is below 1.75

1.75 on table : 0.9599

95.99 % chance

d) the % that a z-score is above 3.31

3.31 on table : 0.9995

99.95% chance below 3.31

so 0.05% chance above 3.31

(1 in every 2000)

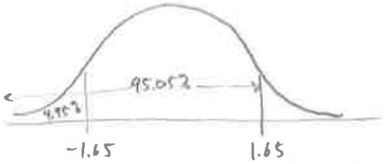
e) the % that a z-score is between -1.65 and 1.65

-1.65 on table : 0.0495

4.95% below

1.65 on table : 0.9505

95.05% below



so between -1.65 and 1.65

= 95.05 - 4.95

= 90.1% chance

Example 4 – The grade point average at Capital City High is 2.6, with a standard deviation of 0.5. If the top 15% of all students are eligible to attend Uvic, what is the minimum GPA needed to attend?

top 15% parallel to bottom 85%

so look up 0.85 on table to get z score of 1.04

$$z = \frac{x - \mu}{\sigma}$$

$$1.04 = \frac{x - 2.6}{0.5}$$

$x = 2.6 + 1.04(0.5)$

$x = 2.6 + 0.52$

$x = 3.12$ is the minimum GPA.

Example 5 – A manufacturer of cell phones indicated a mean of 26 months before there is a need of repairs, with a standard deviation of 6 months. What length of time for the warranty should the manufacturer set such that less than 10% of all cell phones will need repairs during the warranty period?

$\mu = 26$ mo.

$\sigma = 6$ mo.

find z score for less than 10%

0.10 = z-score of -1.28

$$z = \frac{x - \mu}{\sigma}$$

$$-1.28 = \frac{x - 26}{6}$$

$x = 26 + (-1.28)(6)$

$x = 18.32$

The warranty should be for 18 months

5.4 – Confidence Interval for Means

sample vs population

Suppose you wanted to find out the average time that track athletes in Canada can run a 5K. How might you go about collecting the information necessary for this?

ask as many Canadian track athletes as possible their 5k time, using random sampling

Population: A particular group, section, or type of people

ex Population for above is all track athletes in Canada

Sample: a portion of a population

Therefore, in most cases where data is being collected, a sample is used rather than surveying the entire population (due to time, money, and convenience issues).

If you survey a sample of the population to find out the average 5K time, how accurate will the sample mean ($\bar{x}$) be, meaning how close is it to the population mean (μ), the actual average time 5K time of all Canadian track athletes? What would you think the accuracy would depend on?

- how many people are asked (sample size)
- method used for sample (random sampling vs, say, cluster sampling)

Statisticians have developed a way to assess the accuracy of extrapolating a survey mean ($\bar{x}$) to a population mean (μ), by a method called **Confidence Interval for Means**. This is based on the premise that the sample data collection was random (to minimize any bias).

For our example above, finding the average 5K time, let's say 200 athletes were randomly surveyed, and the sample mean $\bar{x}$ = 20 mins. Let's say a **Confidence Interval for Mean** analysis was done. The results would read something like this: *'The mean 5K time for track athletes in Canada is 20 mins, with results accurate to within 1.2 points, 19 times out of 20.'*

So what does this mean?

1.2 points: It means that the mean time (μ) could be anything between 20 ± 1.2 min, so 18.8 min - 21.2 min

19 times out of 20: the mean will be between 18.8-21.2 min 19 times out of 20 that you conduct the identical survey

Overall Result:

The mean 5k time is 20 ± 1.2 min with 95% confidence

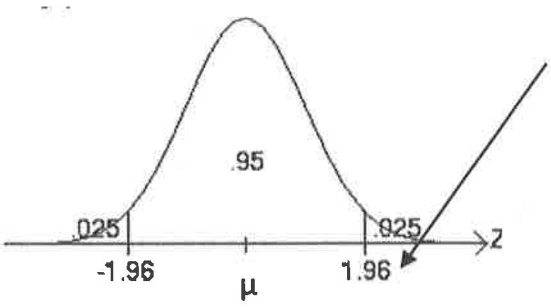
Example 1 – A survey was conducted to find the average height of teenagers in Victoria. 500 teens were sampled and the sample mean $\bar{x}$ = 165cm. A **Confidence Interval for Mean** analysis was run and the final results read like this: *'The average height of teenagers in Victoria is 165cm, with results accurate to 5 points, 19 times out of 20.'* Describe what this means:

The average height of teenagers in Victoria is 165cm ± 5cm
(160 - 170cm)

with 95% confidence

How is **Confidence Interval for Means** calculated?

Most often, 19 times out of 20 is the standard for a survey, meaning surveyors want to state their results with 95% confidence.



z-score for 0.975 (0.025 + 0.95): use the table to look up and you get 1.96 (and -1.96 on the left end).

The Confidence Interval should depend on the info above, but it should also depend on the standard deviation of the data (if the data is more spread out, there's a greater chance that the sample mean ($\bar{x}$) will differ from the population mean (μ).

The standard deviation of the **sample** is symbolized by **s**. The standard deviation for the population (which is unknown) is symbolized by **σ**. It is widely accepted that as long as the sample size exceeds 30, **s** is close enough to **σ** to use in the calculation.

The Confidence Interval also depends on the sample size. Asking 1 000 000 Canadians should yield a more accurate $\bar{x}$ than asking 20 Canadians.

Here is the formula for calculating a **Confidence Interval for a Mean:**

For 95% confidence, $Z_{\frac{\alpha}{2}} = 1.96$

$$\bar{x} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

where

μ = population mean
$\bar{x}$ = sample mean
$Z_{\frac{\alpha}{2}}$ = standard deviation z–score separated into 2 tails
n = sample size

s = standard deviation of sample
σ = standard deviation of population
(same as s if n > 30)

Example 2 – A random sample of 64 teenagers in Victoria have a mean height, $\bar{x}$ = 160cm, with a standard deviation of 15. Find a 95% confidence interval for the mean of the population (μ).

find: $\bar{x} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < x + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

n = sample size = 64

$\bar{x}$ = 160cm

$s = \sigma = 15$ $(n > 30)$

95% confidence

$\frac{5}{2} = 2.5$

so look up z score for 97.5% or 0.975 = 1.96

$160 - 1.96\left(\frac{15}{\sqrt{64}}\right)$ $160 + 1.96\left(\frac{15}{\sqrt{64}}\right)$

$160 - 3.675$ $160 + 3.675$

$156.325 < \mu < 163.675$

The mean is 160cm with results accurate to 3.675 points, 19 times out of 20.

Example 3a – A random sample of 1000 adult male Canadians were asked their weight, and $\bar{x}$ = 78kg, with a standard deviation of 15. Find a 90% confidence interval for μ.

$\bar{x}$ = 78 kg

$n = 1000$

$s = \sigma = 15$

90%

$\frac{10}{2} = 5$

so look up z score for 95, or 0.95 : 1.64

$78 - 1.64\left(\frac{15}{\sqrt{1000}}\right)$ $78 + 1.64\left(\frac{15}{\sqrt{1000}}\right)$

$78 - 0.78$ $78 + 0.78$

$77.22 < \mu < 78.78$

The mean is 78 kg with results accurate to 0.78 points, 9 times out of 10.

Example 3b – Find a 95% confidence interval for μ.

$78 - 1.96\left(\frac{15}{\sqrt{1000}}\right)$ $78 + 1.96\left(\frac{15}{\sqrt{1000}}\right)$

$78 - 0.93$ $78 + 0.93$

$77.07 < \mu < 78.93$

The mean is 78 kg with results accurate to 0.93 points, 19 times out of 20

If correct 95% of time instead of 90% of time, range will have to be a bit larger!