Level 1 – AS91035 – 1.10 – 4 Credits – Internal

Investigate a given multivariate data set using the statistical enquiry cycle

Sample question: Bag weights

Below is some data on twelve students in New Zealand. They have been randomly chosen from those that took part in the CensusAtSchool survey. Half of the students are year 10 and half are year 11.

Based on this sample, make a guess about all year 10s and 11s in New Zealand; who has the heavier bags?

Gender	Age	Country	Bag weight	Year	Region
male	13	Zimbabwe	7.9	10	Canterbury Region
male	14	New Zealand	5.9	10	Wellington Region
female	14	New Zealand	3.7	10	Canterbury Region
female	14	England	5.2	10	Auckland Region
male	14	New Zealand	6	10	Canterbury Region
female	14	New Zealand	5.5	10	Waikato Region
female	15	New Zealand	4	11	Other North Island
male	15	New Zealand	3.3	11	Auckland Region
female	15	New Zealand	5	11	Waikato Region
female	15	New Zealand	2	11	Northland Region
male	15	New Zealand	5	11	Auckland Region
female	15	New Zealand	6	11	Auckland Region

Practice internal

As well as going through this workbook, you'll do a practice internal at the same time. You'll use the same context as the sample question: bag weights of year 10 and 11 students in New Zealand. But you'll use a proper sample of 60; 30 year 10s and 30 year 11s. The data is spread over the next few pages.

Gender	Age	Country	Travel	Time to travel	Bag weight	Year	Region
female	14	New Zealand	bus	30	4	10	Auckland Region
female	14	New Zealand	bike	15	3	10	Bay of Plenty Region
female	14	New Zealand	bus	15	5.5	10	Northland Region
female	14	New Zealand	walk	30	6.6	10	Canterbury Region
female	14	New Zealand	motor	10	4	10	Bay of Plenty Region
female	13	New Zealand	motor	10	4.4	10	Auckland Region
male	14	New Zealand	walk		2.8	10	Otago Region
male	14	New Zealand	walk	12	8	10	Auckland Region
male	14	New Zealand	motor	2	8.2	10	Wellington Region
female	14	New Zealand	bus	15	3.5	10	Auckland Region
male	14	New Zealand	walk	17	2.2	10	Auckland Region
female	14	Serbia	bus	15	1	10	Auckland Region
female	14	New Zealand	walk	10	0	10	Waikato Region
male	14	New Zealand	walk	1	2	10	Waikato Region
female	14	New Zealand	motor	8	5.9	10	Auckland Region
male	14	Papua New Guinea	motor	20	1	10	Canterbury Region
male	14	Scotland	walk	20	4.3	10	Canterbury Region
female	14	New Zealand		30	6.2	10	Waikato Region
female	14	New Zealand	motor	5	1.1	10	Bay of Plenty Region
female	14	New Zealand	bus	30		10	Auckland Region
female	14	New Zealand		20	3.1	10	Canterbury Region
female	14	New Zealand	walk	20		10	Auckland Region
male	14	New Zealand	walk	20	3.1	10	Waikato Region
female	15	New Zealand	walk	1	8	10	Waikato Region
female	14	New Zealand	bus	50	3.3	10	Auckland Region
female	14	New Zealand	walk	38	11	10	Auckland Region
male	14	India	motor	8	6	10	Bay of Plenty Region
male	14	New Zealand	bus	25	7.1	10	Other South Island
female	14	England	bus	45	4.5	10	Bay of Plenty Region
female	14	New Zealand	motor	5	2	10	Auckland Region
female	15	New Zealand	walk	15	5	11	Waikato Region
female	15	New Zealand	bus	15	5	11	Otago Region
male	15	New Zealand	bus	45	0	11	Waikato Region
female	15	New Zealand	walk	3	2.2	11	Waikato Region
female	15	New Zealand	motor	5	6.6	11	Auckland Region
male	15	New Zealand	motor	45	1.4	11	Auckland Region
female	15	New Zealand	walk	15	4.5	11	Waikato Region

female	15	Australia	motor	3	4.4	11	Otago Region
male	14	New Zealand	walk	30	2	11	Auckland Region
	15	New Zealand	walk	13		11	Auckland Region
male	15	Thailand	bike	28	2	11	Wellington Region
female	15	Scotland	bus	40		11	Other North Island
female	15	Pakistan	bus	30	4	11	Auckland Region
female	15	New Zealand	train	20	5.5	11	Auckland Region
female	15	New Zealand	motor	15	4	11	Auckland Region
female	15	New Zealand	motor	25	1.3	11	Auckland Region
female	15	New Zealand	motor	9	2.6	11	Wellington Region
female	15	New Zealand	train	45	6	11	Auckland Region
female	15	New Zealand	walk	2	6	11	Auckland Region
female	15	Nauru	train	25	1	11	Auckland Region
male	15	New Zealand	motor	15	0	11	Auckland Region
male	15	New Zealand	walk	45	1	11	Auckland Region
female	15	New Zealand	motor	5	3	11	Wellington Region
female	15	New Zealand	bus	30	4	11	Bay of Plenty Region
male	15	New Zealand	walk	12	9	11	Otago Region
female	15	New Zealand	motor	25	4.4	11	Waikato Region
female	15	New Zealand	walk	40	3.5	11	Canterbury Region
female	11	New Zealand	motor	10	6.3	11	Otago Region
female	15	England	motor	15	4.9	11	Auckland Region
female	15	New Zealand	bus	30	4.1	11	Waikato Region

Grade requirements

The requirements for each grade are below. Keep these in mind when doing the practice internal. These also apply for the actual internal at the end of the unit.

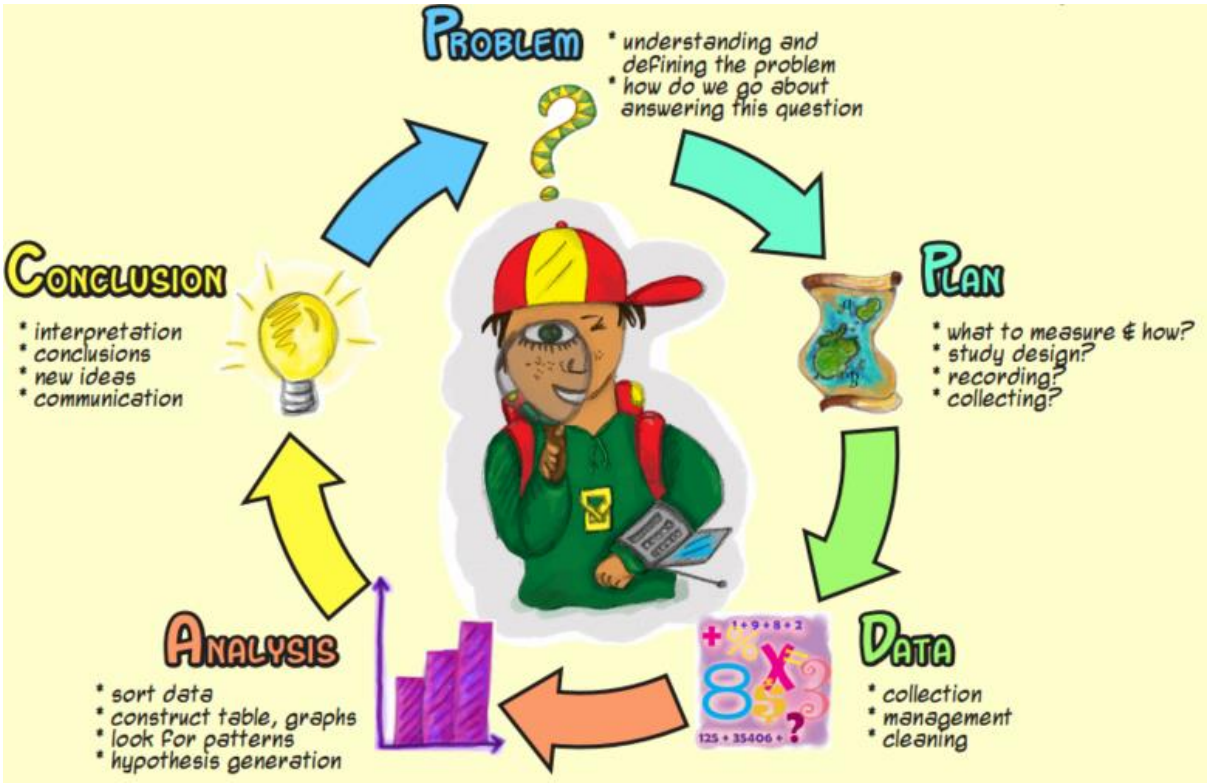
	Achieved	Merit	Excellence
Problem	Pose a question that is comparative. Include the variable you're investigating, the groups being compared, and the population.		
Plan	Nothing to do here.		
Data	Draw a graph (box and whisker plot) and give a summary statistic.	Draw a graph (box and whisker plot) and give **summary statistics**.	Draw **two graphs** (dot plot and box and whisker plot) and give summary statistics
Analysis	Compare two features of the distributions. This could involve comparing the middle 50%, shift and overlap, shape, spread, unusual or interesting features.	Compare two features of the distributions **in context with evidence.** This could involve comparing the middle 50%, shift and overlap, shape, spread, unusual or interesting features.	Compare **three features** of the distributions in context with evidence **and insight**. This could involve comparing the middle 50%, shift and overlap, shape, spread, unusual or interesting features.
Conclusion	An inference is made. *-or-* Use your analysis to answer the question.	An inference is made. Use your **inference** to answer the question **with evidence.** **Show understanding of the context or sampling variability.**	An inference is made. Use your inference to answer the question with **key** evidence **summarised in context**. **Refer to the population when answering the question.** **Show understanding of the context and sampling variability.** **Considers other explanations for the findings**

Lesson One: Inference and the statistical enquiry cycle

What is inference? Inference is a noun meaning "a conclusion reached on the basis of evidence and reasoning" AKA a guess. The verb form is 'infer' which has the synonyms deduce, work out, and conclude. In this standard, you will use a sample of a population to make an inference about that population.

The statistical enquiry cycle (PPDAC cycle) is used to remind us of everything we need to do to make good inferences. If you have another way to remember that works for you then use that but be warned that you'll have to work harder.

What does PPDAC stand for?



Problem

In this first stage, you state the problem as a question, e.g. Are bag weights of year 11 students heavier than year 10 students, in New Zealand?

Plan

In this standard, you are given the data to use so there is no need to make a plan for collecting it.

Data

In this third stage, you make the data useful. Useless data is, well… useless. To make the data useful you need to summarise the data with summary statistics and display the data with data displays.

Analysis

In this fourth stage, you describe the patterns in detail. You mention anything that might help you infer about the population. Use your sample to make an inference about the population.

Conclusion

In this fifth stage, you use your inference about the population to answer your question. This is where you show off your understanding of the context and of sampling variability as well as considering other explanations for your findings.

Lesson 1 questions

Q1. Explain the PPDAC cycle in student-speak.

Q2. Explain what a population is and what a sample is.

Lesson Two: Problem

It turns out that writing a good question to define a problem is rather complex mostly because you need to make clear that you're not looking for a definite answer, only an educated guess, an inference.

This is the beginning of your report so it is a good time to mention what your population is and what your sample is. This is where you get to show off that you know they're different things. For example, in the bag weights context, the population is all year 10s and 11s in New Zealand, and our sample is 12 students randomly chosen from those that took part in the CensusAtSchool survey.

It is also a good time to mention what you're comparing. In the bag weights context, it's obviously bag weights. But perhaps it's more complicated than that. What if a student has two bags, did they combine the weights or only choose one? Do some students have cubbies and others don't? Do sports bags count or only school bags? Were the bags measured over several days or only one? What if a student left their bag at home that day or was taking their books home at the end of the year? It may turn out that the real-world reality is more complicated than you originally thought.

You can mention what you initially think the sample will show. In science this is called a hypothesis. There's no point in mentioning this if you don't also explain why.

Lesson 2 questions

Suggest a possible question for each investigation below. Include the groups to compare and the variable to investigate.

Q1. How do different students get to school?

Q2. How do school days differ for students in the primary school and high school?

Q3. What do different students use their phones for?

Q4. How do different sports compare?

Q5. Who has the healthiest lunches at school?

Q6. How different are students in their gaming habits?

Practice internal: Write a question for the bag weights practice internal.

Lesson Three: Plan

In this standard, the data is already given to you so there is no need to make a plan to collect it. Move on to the next lesson.

Lesson Four: Data: Summary statistics

Data wants to be useful. To make it useful, we do two things: calculate summary statistics and display the data in data displays.

Summary statistics summarise the data so anyone can see roughly what the data looks like without having to look at *all* the data. It helps greatly with comparing two groups.

There are 10 summary statistics:

Measures of centre	Quartiles	Extremes	Measures of spread
Mean Median Mode	Lower quartile Upper quartile	Minimum Maximum	Range Inter-quartile range Standard deviation

Mean
The mean is the easiest to calculate but is not the most useful as it's affected by extreme values. To calculate the mean add all the values together and divide by the number of values.

Median
The median is harder to calculate because it requires the data to be in numerical order. The strength of the median is that it's unaffected by extreme values. This makes it a good summary statistic to use as evidence of a shifted centre. To calculate the median, order the data from smallest to largest and select the middle number. If there are two middle values, find halfway between them.

Mode
The mode is laughably easy to calculate but hardly useful as a measure of centre. To find the mode, find the value that occurs most often. There can be more than one mode.

Lower quartile
While the median gives us the middle of the data, the lower quartile gives us the lower quarter of the data. This is useful in seeing the shape of the data. To calculate the lower quartile, keep the data ordered from smallest to largest and select the number halfway between the minimum and the median.

Upper quartile
The upper quartile gives us the upper quarter of the data. When combined with the lower quartile and median, it gives us a strong picture of the shape of the data. To calculate the upper quartile, keep the data ordered from smallest to largest and select the number halfway between the median and the maximum.

Minimum
The minimum helps give us a better picture of the shape of the data. It is the smallest value in the data.

Maximum
The maximum also helps give us a better picture of the shape of the data. It is the largest value in the data.

Range
The range is easy to calculate but is easily influenced by extreme values so isn't very useful as a measure of spread. To calculate the range, take the maximum minus the minimum.

Inter-quartile range
The inter-quartile range is easy to calculate and isn't influenced by extreme values making it very useful. It gives the spread of the middle 50% of the data. To calculate the inter-quartile range, take the upper quartile minus the lower quartile.

Standard deviation
The standard deviation is hard to calculate and hard to interpret. But it is an excellent measure of spread so you can be sure that data with a higher standard deviation will have a larger spread. This makes it a good summary statistic to use as evidence of spread. Use a calculator to calculate the standard deviation. See the appendix for how to calculate standard deviation by hand.

Data that is 'normal' is within $\pm$ 1 standard deviation of the mean. Anything less than 1 s.d. from the mean is unusually small. And anything more than 1 s.d. is unusually large. Not extremely large, just larger than most.

Lesson 4 questions

Refer to the data set on the next page. Find the summary statistics for these comparisons:

Q1. Reaction times of females and males.
Q2. Heights of females and males.
Q3. Time taken to travel to school for females and males.
Q4. Number of texts sent on a cell phone for females and males.
Q5. Waking times for females and males.
Practice internal: Find the summary statistics for the bag weights practice internal.

gender	languages	height	travel time	reaction	celltxtsend	celltxtrec	bedtime	waketime	year	region
female	1	154	40	0.395	2	0	21:00:00	6:30:00	9	Auckland Region
female	3	144	4	0.667	4	14	22:15:00	7:30:00	7	Waikato Region
female	2	171	15	1.047	100	99	22:30:00	7:45:00	10	Other North Island
female	1	156	25	0.44	34	43	23:45:00	7:00:00	12	Wellington Region
female	2	164	10	0.525	14	13	21:30:00	6:30:00	9	Other North Island
female	1	165	6	0.448	30	30	22:00:00	6:45:00	9	Bay of Plenty Region
female	1	175	21	0.778	0	0	21:45:00	7:15:00	9	Auckland Region
female	1		6	0.872	200	200	24:00:00	7:30:00	9	Auckland Region
female	1	151	6	0.487	10	25	22:45:00	8:30:00	10	Wellington Region
female	1	182	15	4.493	20	25	22:00:00	6:15:00	11	Waikato Region
male	1	152	8	0.437	5	5	21:45:00	6:45:00	7	Auckland Region
male	4	160	31	0.377	0	0	22:30:00	6:30:00	9	Canterbury Region
male	1	180	2	0.394	0	0	20:30:00		8	Wellington Region
male	2	182		1.079	50	50	22:00:00	5:45:00	11	Waikato Region
male	1	183	80	0.47	200	200	23:15:00	6:15:00	13	Other South Island
male	3	140	15	0.434	4	4	18:45:00	6:00:00	9	Auckland Region
male	1	152	5	0.593	0	0	9:30:00	19:30:00	5	Auckland Region
male	1	174	8	1.804	20	25	12:45:00	8:00:00	9	Auckland Region
male	1	132	2		0	0	21:45:00	8:00:00	5	Auckland Region
male	1	130		45.699	100	100	21:00:00	7:00:00	8	Auckland Region

Lesson Five: Data: Data displays

The other way to make data easy to see is to show it in data displays. There are two data displays we'll use: dot plots and box and whisker plots.

Dot plots

Dot plots give us a picture of the shape of the data. They are the best display to do first because they give an overall picture. Display the reaction times of females and males with a dot plot (the data is from lesson 4).

Box and whisker plots

Box and whisker plots give a good comparison between the two groups. They show some of the summary statistics in a visual form so you'll need to have calculated them first. Display the reaction times of females and males with a box and whisker plot.

Lesson 5 questions

Refer to the data set on page 12. Draw both a dot plot and a box and whisker plot for these comparisons:

~~Q1. Reaction times of females and males.~~

Q2. Heights of females and males.

Q3. Time taken to travel to school for females and males.

Q4. Number of texts sent on a cell phone for females and males.

Q5. Waking times for females and males.

Practice internal: Draw a dot plot and box and whisker plot for the bag weights practice internal.

Lesson Six: Analysis

The most powerful analysis is done with your eyes. By literally looking at the data in your data displays and observing how it's distributed, we can make some powerful claims about the sample.

The summary statistics make really good evidence for your observations of the distribution.

To describe how the data is distributed, we refer to the features of the data distribution. You could choose to refer to all of these and be systematic or you could choose to refer to only the most relevant ones.

Middle 50%
The middle 50% of the data lies between the lower and upper quartile. Comparing the middle 50% of both groups focuses on typical values rather than extreme values. You could comment on the shift and/or spread of the middle 50%

Shift
Is the data shifted between on group and the other? You could focus on the middle 50% or on the entire sample.

Overlap
An alternative to shift. This describes how much the middle 50% of each group overlap one another: mostly, somewhat, or not at all.

Shape
Box and whisker plots are great for most things but they're rubbish at showing the shape of the data. Use your dot plot for this. We expect the shape of the data to be bell-shaped, if it's not there's usually an interesting reason which is worth thinking about. If it's not bell-shaped, is it bi-modal? Is it skewed? Are there clusters?

Spread
How spread out is the data? You could look at the middle 50% or all of the sample.

Unusual or interesting features
This is the "misc." category for all of the things you want to say but don't fit easily anywhere else.

Lesson 6 questions

Refer to the data set on page 12. Analyse the features of the data distribution for these comparisons:

Q1. Reaction times of females and males.

Q2. Heights of females and males.

Q3. Time taken to travel to school for females and males.

Q4. Number of texts sent on a cell phone for females and males.

Q5. Waking times for females and males.

Lesson Seven: Inference

You can analyse your sample as much as you want but remember your problem. In the bag weights context it is: "Are bag weights of year 11 students heavier than year 10 students, in New Zealand?" This is referring to the population of all year 11 students and all year 10 students not just the sample you've just analysed. Somehow we need to make a decision about something we don't know anything about.

Fortunately, we know one thing about the population: our sample came from it.

But what does it mean for year 11s to have heavier bags than year 10s? It doesn't mean that the year 11s have a heavier median than the year 10s in our sample; we want to describe all year 11s and 10s in New Zealand. It also doesn't mean that every year 11 in New Zealand has a heavier bag than every year 10 in New Zealand. Let's use the two medians of the population to decide the difference.

(You could write this into your problem if you would like. For example "Is the median bag weight for year 11 students heavier than the median bag weight for year 10 students, in New Zealand?")

If our groups are so separate that there is no overlap, we are confident that there is a significant difference between the two medians.

If there is almost total overlap, the two medians are very close together and our sample is insufficient information to tell us which median is bigger/taller/greater (or heavier in the bag weight context).

The sample you have will be somewhere in the middle of these two extreme cases. To decide if the population groups are separate enough for your sample to be sufficient information to make a decision, use this method:

$$\frac{distance\ between\ medians}{overall\ visual\ spread}$$

What…?

The distance between the medians is, well, the distance between the medians in your sample. The overall visual spread is the highest upper quartile minus the lowest lower quartile.

In the first scenario, the no overlap scenario, the $\frac{DBM}{OVS}$ is close to 1. So we can say that scores close to 1 mean there is a significant difference.

In the second scenario, the almost total overlap scenario, the $\frac{DBM}{OVS}$ is close to 0. So we can say that scores close to 0 mean there is no significant difference.

You will always get a score between 0 and 1. But what's the boundary between a bad score and a good score, no significant difference and a significant difference? It's not 0.5. It depends on your sample size. For samples with at least 30 data points per group, it's $\frac{1}{3}$ or 0.333333…. For samples with at least 100 data points per group, it's $\frac{1}{5}$ or 0.2

Lesson 7 questions

Refer to your box and whisker plots from lesson 5. Make an inference about the population for each comparison:

Q1. Reaction times of females and males.

Q2. Heights of females and males.

Q3. Time taken to travel to school for females and males.

Q4. Number of texts sent on a cell phone for females and males.

Q5. Waking times for females and males.

Practice internal: Make an inference for the bag weights practice internal.

Lesson Eight: Conclusion

This is where you use everything you've done so far to answer your question. Use your inference to answer your question for the population. Summarise your analysis with key evidence and mention contextual factors (in the bag weights scenario this includes cubbies, multiple bags, when the bags were measured, etc.). Show that you understand that if there is a difference in the population then even if you take another sample, you would expect the result to be the same though the samples will look different. Consider alternative explanations for your findings (this is more of those contextual factors).

A sample achieved conclusion:

A sample merit conclusion:

A sample excellence conclusion:

Lesson 8 questions

Use the work you've done so far to conclude each comparison:

Q1. Reaction times of females and males.

Q2. Heights of females and males.

Q3. Time taken to travel to school for females and males.

Q4. Number of texts sent on a cell phone for females and males.

Q5. Waking times for females and males.

Practice internal: Conclude the bag weights practice internal.

More practice internals

Done! You have just completed a practice internal comparing the bag weights of year 10 and 11 students in New Zealand. Do at least one practice internal by yourself before the class does another one together. Get a new sample from CensusAtSchool of at least 30 students in each group. You could change the groups you're comparing, the variable you're investigating, or both.

Appendix: Standard deviation

The standard deviation is the square root of the variance. But wait, what's the variance, I hear you ask. The variance is the average of the squared differences from the mean.

What…?

Let's take this step by step to calculate the standard deviation.

1. Let's take an imaginary data set of five values: {2, 5, 5, 8, 10}
2. Find the mean:
   $\frac{2+5+5+8+10}{5} = 6$
3. Find the difference for each value from the mean:
   $2-6=-4$
   $5-6=-1$
   $5-6=-1$
   $8-6=2$
   $10-6=4$
4. Square the differences to find the "squared differences from the mean":
   $(-4)^2 = 16$
   $(-1)^2 = 1$
   $(-1)^2 = 1$
   $2^2 = 4$
   $4^2 = 16$
5. Find the average (mean) of the squared distances from the mean. This is the variance:
   $\frac{16+1+1+4+16}{5} = 7.6$
6. Find the square root of the variance. This is the standard deviation:
   $\sqrt{7.6} = 2.756809750418044$

The standard deviation is 2.8 (1 d.p.). Check this on a calculator.

For more explanation google "maths is fun standard deviation".