A guide to Post-Primary statistical inference

Strand 1 Section 1.7 lists learning outcomes related to ***statistical inference*** which deals with the principles involved in generalising observations from a ***sample*** to ***the whole population***. Such ***generalisations*** are valid only if the data are ***representative*** of that larger group.

> *A representative sample is one in which the relevant characteristics of the sample members are generally the same as those of the population.*
>
> *An improper or **biased** sample tends to systematically favour certain outcomes and can produce misleading results and erroneous conclusions.*

Randon sampling is a way to remove ***bias*** in sample selection, and tends to produce representative samples. At ***JC HL*** and ***all levels*** at ***LC***, students are required to

- recognise how sampling variability influences the use of sample information to make statements about the population

Whilst ***LC HL*** students are required to go beyond this and

- *use simulations to explore the variability of sample statistics from a known population, to construct sampling distributions and to draw conclusions about the sampling distribution of the mean*

At ***JC HL***, and ***LC FL*** and ***LC OL***, students should experience the consequences of non-random selection and develop a basic understanding of the principles involved in random selection procedures. At ***LC HL***, learners extend this understanding; they explore simulations that produce frequency distributions of sample means and conclude from these explorations that when we take a large number of random samples of the same size and get a frequency distribution of the sample means, this distribution – called ***the sampling distribution of the mean*** – tends to become a normal distribution and

- If the sample size is large ($n \geq 30$) then for any population, no matter what its distribution, the sampling distribution of the mean will be approximately normal
- This normal distribution will have a mean equal to the population mean with standard deviation $\frac{\sigma}{\sqrt{n}}$. This is calld the ***standard error of the mean***.

Suppose a group of students was investigating the sporting preferences of students in their school. At ***JC FL*** and ***JC OL***, students might survey the whole class; students at this level are ***not*** required to *look beyond the data* and no generalisation is required. At ***JC HL*** and at ***all levels*** at ***LC***, students begin to acknowledge that it is possible to *look beyond the data*. They would gather data from a ***sample*** and ***generalise*** to a larger group. In order to be able to ***generalise*** to all students at the school a ***representative sample*** of students from the school is needed. This can be done by selecting a ***simple random sample*** of students from the school.

At each of the levels ***JC HL***, ***LC FL, LC OL*** and ***LC HL***, students are required to deal with ***sampling variability*** in increasingly sophisticated ways.

Consider the data below gathered from a ***simple random sample*** of 50 students.

		Do You Like Soccer?		Row Total
		Yes	**No**	
Do You like Rugby?	**Yes**	25	4	29
	No	6	15	21
Column Total		31	19	**50**

Suppose, before the study began, a teacher ***hypothesised***: *I think that more than 50% of students in this school like Rugby*. Because 58% ($\frac{29}{50}$ = 58%) of the sample like rugby there is ***evidence*** to support the teachers claim. However, because we have only a sample of 50 students, it is ***possible*** that 50% of ***all*** the students like rugby but the variation due to random sampling might produce 58% or even more who like rugby.The statistical question, then, is whether the sample result of 58% is reasonable from the variation we expect to occur when selecting a random sample from a population with 50% successes? Or, in simple terms, ***What is a possible value for the true population proportion based on the sample evidence?***

At ***JCHL*** and ***LCFL*** it is sufficient for students to acknowledge sampling variability; a typical response at this level would be ….*although 58% of this sample reported that they like rugby, it is possible that a larger or smaller proportion would like rugby if a different sample was chosen. 58% is close to 50% and it is possible that 50% of all the students like rugby….* At this level, the acknowledgement of variability is more evident in the planning stage with students deciding to choose a large sample or perhaps several small samples and average the findings in order to reduce the sampling error. [If this cohort were deaing with numerical data and were looking for a set of posible values for the ***population mean*** the possible set of values could be determined by looking at the distribution of the data with respect to the ***sample mean*** and the ***range***.]

Building on this understanding, a more sophisticated approach to inference involves finding a set of possible values by using the **margin of error.**

The true population proportion = The sample proportion $\pm$ *Margin of Error*

The margin of error is estimated as $\frac{1}{\sqrt{n}}$ where ***n*** is the sample size and refers to the maximum value of the radius of the 95% confidence interval.

This is the level of inference required by ***OL*** students at Leaving Certificate. A ***LC OL*** student might therefore conclude

…there is evidence to support the teachers claim that more than 50% of students in the school like rugby because, based on the sample data, any values in the range 44% - 72% are possible values for the proportion of students in the school who like rugby…

[If this cohort were deaing with numerical data and were looking for a set of posible values for the ***population mean*** the possible set could be determined by engaging with the ***empirical rule.*** The empirical rule formalises the undertanding students get from examining the spread of the distribution with respect to the mean. Knowing the proportion of values that lie within approx 1,2 or 3 standard deviations from the mean allows students to determine what is a ***possible set of values for the population mean***.]

LC HL students are required to build further on these ideas and make more accurate estimates of the ***possible values*** of the ***true population proportion in the case of categorical data*** *or* ***the population mean in the case of numerical data.*** To do this they

- *construct 95% confidence intervals for the population mean from a large sample and for the population proportion, in both cases using z tables*

Constructing ***confidence intervals*** brings two ideas together:

- sampling variability and the idea of the ***standard error of the population proportion/mean***
- the ***empirical rule*** – 95% of the data lies within 1.96 standard deviations of the mean.

The set of possible values, or the ***confidence interval***, is

Sample mean/proportion ± 1.96 standard error

In the case being examined, the set of possible values for the ***true population proportion*** would be given by

$$\text{Sample proportion} \pm 1.96 \text{ standard error} = .58 \pm 1.96\sqrt{\frac{.58(1-.58)}{50}}$$

$$= .7168 \text{ or } .4432$$

So, the ***true population proportion*** lies beween 44.32% and 71.68%.

Compare this with the set of values obtained using the margin of error. ***LCHL*** students can examine the effect of increasing the sample size on the ***precision*** of the estimate.

LC OL students should understand a hypothesis as a ***theory*** or ***statement*** whose truth has yet to be proven. However, ***LC HL*** students must develop this idea and deal formally with ***hypothesis testing***. They

- perform univariate large sample tets of the population mean (two-tailed z-test only)
- use and interpret p-values

The ***p-value*** represents the chance of observing the result obtained in the sample, or a value more extreme, when the hypothesised value is in fact correct. A small p-value would support the teacher's claim because this would have indicated that, if the population proportion was 0.50 (50%), it would be very unlikely that an observation of 0.58 (58%) would be observed.

A large sample hypothesis test of the population ***mean*** has 4 components:

1. **A test statistic:** This is a standard normal ***z score*** that is the difference between the value we have ***observed for the sample*** and the ***hypothesised value for the poulation*** divided by the ***standard error of the mean.***

2. **A decision rule:** Reject the hypothesised value if ***z > 1.96*** or ***z < -1.96***

3. **A rejection zone:** ***z > 1.96 or z <-1.96***

4. **Critical values**: ***z = 1.96 , z = -1.96*** since we are using the 5% level of significance.