Teaching Sampling and Hypothesis Testing

Koh Liang Cheung Daniel

Abstract

Sampling and Hypothesis Testing is a unit in the Statistics syllabus within Advanced "A" level H2 Mathematics. In this article, some approaches to teaching the topics are introduced and discussed.

1 Introduction and Background

Statistics is introduced in Secondary School in the form of Descriptive Statistics where students learn to summarize and analyze one variable data. In Junior College (JC), students learn about sampling methods and are also introduced to Statistical inference through hypothesis testing of a population mean.

Sampling Methods and Hypothesis Testing are tested in 2 separate questions at the A-levels. In a typical examination question on Sampling Method, students would be asked to describe in context how a *simple* random sample, *stratified* random sample or *systematic* random sample can be carried out in the given word problem scenario. In a hypothesis testing examination question, students would usually be provided with a set of observed data and asked firstly to find an *unbiased estimate* for the population mean (and population variance) and secondly to perform a 1 (or 2) tailed test of significance.

An astute student might wonder,

> "How does simple random sampling in a finite population context, lead to a *hypothesis test* of a population mean when in the test procedure, observed data is assumed to be a realization of independent and identically distributed random variables?"

The connection between data collection and decision making under uncertainty could be strengthened and made much more explicit. The learning of Statistics as a methodological discipline may be enhanced when JC students have the opportunity to view the topic of Statistics as an investigatory process. The teacher could lead students to consider the following:

1. Begin with a question about a population feature of interest.
2. Consider how to collect data and what information about the population can be derived from the data.
3. Draw a conclusion about the population using the data and consider the possible error in this decision.

In the author's school the topic of Sampling and Hypothesis Testing is introduced after the Probability Topics. We assume readers have familiarity with Permutations and

Combinations, Binomial and Normal Distributions, and the statement of the Central Limit Theorem.

2 A (Finite) Population Proportion

Sample Survey

In the Social Sciences, a survey is a common way to quickly obtain information about a population. For example, to obtain information about the proportion of the population who possess a certain characteristic or hold a certain view.

Example
All JC students take Project Work as an A-level subject in JC1 and surveys usually form a part of their report.

Suppose that a group of students working on their A-level Project Work (PW) wish to determine whether their school-mates hold a particular view on an issue they are researching on. They come up with 2 solutions and wish to obtain student feedback.

They obtain an email listing of all N (=1000) students, and conduct a **simple random sample** of say n (=28) students. In the survey, students are asked to choose between the 2 proposed solutions: "solution A" and "solution B". In particular, the PW group wishes to investigate the proportion of students who choose "A" as they believe it is the more workable one.

Denote X to be the number of students in a sample of 28 who choose A. For the survey sample obtained, it is revealed that $X = 21$ which is 75% (a more than half) of those surveyed.

Suppose the student population is indifferent to the 2 proposals, then how likely is it on the basis of pure chance that 21 out of 28 would choose "A" ? How small is this likelihood and hence how surprising?

To determine whether the observed result is surprising (or not) under the assumption that the student body has no real preference (null hypothesis), we can imagine that we could replicate this process of finding 28 students to survey and determine how many times we obtain 21 (or more) choices of "A". If this proportion is small then the actual observed survey results would provide evidence that students really do favour "A".

We can find this proportion systematically by considering a simple combinatorial problem encountered by JC students. We can compute the probability $P(X \geq 21)$ which is also known as the p-value. We proceed more generally.

Suppose an urn contains N balls numbered 1 to N of which $R\,(\le N)$ are red balls and the rest are white balls. A collection of $n\,(\le N)$ balls are selected at random without replacement. How many of these collections contain exactly $k\,(\le n)$ red balls?

There are $\binom{N}{n}$ possible collections of which $\binom{R}{k}\binom{N-R}{n-k}$ contain exactly k red balls.

Denote X to be the number of red balls obtained if we were to choose a collection of $n\,(\le N)$ balls at random without replacement. Then, $P(X=k)=\dfrac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$.

It should be clear that in the context of **Example** that if there are an equal number who choose "A" and "B", we would have $P(X=21)=\dfrac{\binom{500}{21}\binom{500}{7}}{\binom{1000}{28}}$=0.00405 (3sf) using a Graphics Calculator (GC). We can also compute $P(X\ge 21)=\sum_{k=21}^{28}P(X=k)=0.00568\,(3\text{sf})$.

This probability is very small so we would believe that the population is in favour of "solution A". Of course, there is always a possibility that we could be wrong and obtained a sample with extreme data simply by chance. However, based on this decision making procedure we would be wrong less than 1% of the time.

Discussion

We say that X follows a hyper-geometric distribution if $P(X=k)=\dfrac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$.

We observe that when the sample size is small relative to the population, then

$$\frac{P_k^M}{P_k^N}=\frac{M}{N}\left(\frac{M-1}{N-1}\right)\cdots\left(\frac{M-k+1}{N-k+1}\right)\approx\left(\frac{M}{N}\right)^k \text{ where } P_r^n=n(n-1)\cdots n(n-r+1).$$

That is, the probability of drawing k consecutive red balls ***with*** and ***without*** replacement makes little difference. We also have $\dfrac{P_{n-k}^{N-M}}{P_{n-k}^{N-k}}\approx\left(\dfrac{N-M}{N}\right)^{n-k}$ similarly.

Altogether, $$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}=\binom{n}{k}\left(\frac{P_k^M}{P_k^N}\right)\left(\frac{P_{n-k}^{N-M}}{P_{n-k}^{N-k}}\right)\approx\binom{n}{k}\left(\frac{M}{N}\right)^k\left(\frac{N-M}{N}\right)^{n-k}.$$

We recognize $\binom{n}{k}\left(\frac{M}{N}\right)^{k}\left(\frac{N-M}{N}\right)^{n-k}$ as a Binomial probability (mass function) and say that X has an "approximate" Binomial distribution if the sample size is small relative to the population size. We observe that when balls are drawn at random ***with replacement***, then the outcomes may be regarded as outcomes of independent and identically distributed random variables.

For the above example, using the Binomial model, $B(28,0.5)$, we can compute an approximate p-value, using the Graphics Calculator

$$P(X \geq 21) = 1 - P(X \leq 20) = 0.00627 \text{ (3sf)}.$$

Further Discussion
Technology in the use of GC has made computation of Hyper-geometric and Binomial probabilities a simple matter. Historically, Normal approximations could be utilized to approximate probabilities and find p-values. That is, we can match the

mean ($np = 28(0.5) = 14$) and variance ($np(1-p) = 28(0.5)(0.5) = 7$) from the Binomial Model $B(n = 28, p = 0.5)$, to obtain an approximate Normal Model, $N(\mu = 14, \sigma^2 = 7)$.

This leads us to a ***z-test*** which is at the heart of the Hypothesis Testing unit in the A-level syllabus! From the GC's ***Normal cumulative distribution function*** built-in function, we have

$$P(X \geq 21) \approx 0.00701 \text{ (3sf)}$$

where we can also use a continuity correction to improve the accuracy. The approximation is good because of the sample size which is justified by the ***Central Limit Theorem***.

3 Conclusion

Solving A-level examination hypothesis testing questions is often a mundane Graphics Calculator exercise and the focus on teaching and learning is often in completing a Hypothesis Testing template correctly, paying special attention to appropriate phrases used in the conclusion. Teachers can make the teaching of Statistics more meaningful by drawing on a careful selection of examples, and to use Mathematics for explanatory purposes.

The author is a teacher at Raffles Institution (JC) and an occasional A-level Project Work Oral Presentation assessor!