Individual Project
CHAN King Yeung (1155119394)
Dr. HO Kwok Wah
STAT2001A Basic Concepts in Statistics and Probability I
18 December 2019

Part A

A group of data can be divided into several parts of the component. While comparing 2 groups of data, says group 1 and group 2, it happens a case that each component in group 1 has some properties which is better than that of group 2, but these properties cannot be reflected in overall data. The outcome of data gives people an illusion. This kind of misleading results is known as Simpson's Paradox.

People treat this phenomenon as a paradox because of its misleading results. People may assume there is a connection between the whole group and its parts. In other words, it is a fallacy. The following will illustrate how the fallacy works. On one hand, it is arguing the whole has a particular property by appealing to its parts have the same property. "Each of my thesis statement in my essay is perfect, so my essay is perfect." We know it is not true because there is a case that all the thesis are contradictive. On the other hand, it is arguing the parts have a particular property by appealing to the whole has the same property. "An orchestra has outstanding performance, so each player must be playing well in the team." However, if you have heard the story of "lan yu chong shu" (濫竽充數), you know it is not true. Simpson's Paradox is very similar in about examples of fallacy, but it describes in a quantitative representation. Thus, people would say it is a paradox.

Let's consider the following case (data are hypothetical) to summarise Simpson's Paradox. Last year, 2 groups of students, students in group 1 attend tutorials and students in group 2 do not attend tutorials. Given that some students have taken elementary courses about statistics, says STAT1011, and some have not taken any course related to statistics. The following tables describe the distribution of their final result in STAT2001.

Attend tutorial	Taken STAT1011	Have not taken STAT1011
Pass	44	26
Fail	4	26

Do not attend tutorial	Taken STAT1011	Have not taken STAT1011
Pass	173	1
Fail	21	23

The above tables illustrate that students who attend tutorials have a higher passing rate in STAT2001 whatever they have taken STAT1011 beforehand or not (comparing students have taken STAT1011: 91.67% versus 89.18%; and comparing students have not taken STAT1011: 50% versus 4.17%). However, the overall passing rate of STAT2001 if students do not attend tutorials have a higher rate than that of attending tutorials (70% versus 79.82%).

As a student, it is too confused to decide whether to attend tutorials or not since the above statistics seem to contradictive. In decision making, a case involving Simpson's Paradox would mislead one's will as we do not know which option is representative about the fact. In the above example, the lecturer would use the comparison of passing rate for with and without taken STAT1011 to encourage students to attend tutorials. Students may think tutorials are useful to their studies and tend to attend tutorials. Meanwhile, classmates would use the comparison of overall passing rate to persuade not to attend tutorials. Students may think tutorials make thing worse and tend to not going to attend tutorials. Each of their claims only contains things that favour their argument, which is complicated to make a correct decision.

Overall speaking, Simpson's Paradox is a kind of fallacy demonstrated with data. It leads a confusion in decision making because of the wrong impression about the result if we do not fully understand the situation.

In statistics, we are often interested in finding the moment, such as mean and variance, to measure the central tendency of a set of values. The moment generating function, as the name suggested, is a function used to generate the moment of random variables. It exists on all real number t, and defines on a finite (for discrete random variables) or open internal (for continuous random variables) about $t = 0$.

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum e^{tx} f(x) & for\ discrete\ random\ variable \\ \int_{-\infty}^{\infty} e^{tx} f(x) & for\ continuous\ random\ variable \end{cases}, t \in \mathbb{R}$$

To obtain the r-th moment of a distribution, we need to apply differentiation r times and evaluating the result at $t = 0$. (Remarks: r is a non-negative integer, at $r = 0$ it return $M_X(0) = 1$)

First moment, $r = 1$

$$M'(t) = \frac{d}{dt} E(e^{tX}) = \begin{cases} \frac{d}{dt} \sum e^{tx} f(x) = \sum x e^{tx} f(x) \\ \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f(x)\,dx = \int_{-\infty}^{\infty} x e^{tx} f(x)\,dx \end{cases} = E(Xe^{tX})$$

$$M'(0) = \begin{cases} \sum x f(x) \\ \int_{-\infty}^{\infty} x f(x)\,dx \end{cases} = E(X)$$

Second moment, $r = 2$

$$M''(t) = \frac{d}{dt} M'(t) = \begin{cases} \frac{d}{dt} \sum x e^{tx} f(x) = \sum x^2 e^{tx} f(x) \\ \frac{d}{dt} \int_{-\infty}^{\infty} x e^{tx} f(x)\,dx = \int_{-\infty}^{\infty} x^2 e^{tx} f(x)\,dx \end{cases} = E(X^2 e^{tX})$$

$$M''(0) = \begin{cases} \sum x^2 f(x) \\ \int_{-\infty}^{\infty} x^2 f(x)\,dx \end{cases} = E(X^2)$$

r-th moment

$$M^{(r)}(t) = E(X^r e^{tX})$$
$$M^{(r)}(0) = E(X^r)$$

Another property of the moment generating function is that is can uniquely characterise the distribution of random variables. In other words, it is another way to represent the probability distribution of random variables, similar to the probability mass function and probability density function. With this useful property, we can distinguish the probability distribution of random variables quickly and easily.

Suppose there are 2 random variables X and Y, where $M_X(t)$ and $M_Y(t)$ exist respectively. If $M_X(t) = M_Y(t)$ for all values of t, we can say that X and Y follow the same kind of probability distribution. Given that $M_X(t) = (0.8e^t + 0.2)^n$, we can easily state that the random variable X is Binomial distributed with parameter 0.8 because the $M_X(t)$ equals to the moment generating function of Binomial distribution for all values of t.

Furthermore, the moment generating function aids to find the distribution of random variables in a trivial way. What if we want to find the distribution of $Z = X + Y$, where X and Y are Exponential random variables with parameter θ, given they are independent. Without using the moment generating function, we need to spend so much time calculating the cumulative distribution function and the probability density function of Z.

$$F_Z(z) = P(Z \leq z) = P(X + Y \leq z) = \int_0^z \int_0^{z-x} \frac{1}{\theta} e^{\frac{-x}{\theta}} \frac{1}{\theta} e^{\frac{-y}{\theta}}\, dy\, dx = 1 - e^{-\frac{z}{\theta}} - \frac{1}{\theta} z e^{-\frac{z}{\theta}}$$

$$f_Z(z) = \frac{d}{dz}\left(1 - e^{-\frac{z}{\theta}} - \frac{1}{\theta} z e^{-\frac{z}{\theta}}\right) = \frac{z^{2-1} e^{-\frac{z}{\theta}}}{\theta^2 (2-1)!}$$

However, with the use of the moment generating function, we can easily to notice that Z is distributed as Gramma with parameter θ and $\alpha = 2$.

$$M_{X+Y}(t) = E\left[e^{t(X+Y)}\right] = E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t) = \frac{1}{1-\theta t} \cdot \frac{1}{1-\theta t} = \frac{1}{(1-\theta t)^2}, t < \frac{1}{\theta}$$

Although the moment generating function seems to be very useful for time-saving, there are also some limitations in using the moment generation. Let's recall the definition of the moment generating function is $M_X(t)$ exists for all real number t. The distribution of random variables needed to bound within a range. Yet, some random variables do not fulfil the above requirement, just like lognormal random variable does not define for all values of t. We should consider whether the random variable exists for the moment generating function. As said previously, the moment generating function is useful for finding the moment of a set of values. However, things may get worse if the formula of the moment generating function is difficult to differentiate it. Consider the moment generating function of Hypergeometric distribution.

$$M_{Hypergeometric}(t) = \frac{\binom{N-k}{n}\ {}_2F_1(-n, -K; N-K-n+1; e^t)}{\binom{N}{n}}$$

It is too complicated to find the moment in such a case. The moment generating function is useful only when we can easily to perform differentiation and evaluate the function, such as Poisson distribution.

$$M_{Poisson}(t) = e^{\lambda(e^t - 1)}$$
$$M'_{Poisson}(t) = \lambda e^{t + \lambda(e^t - 1)}$$

All in all, the moment generating function is a technic for finding the moment of a set of values in an efficient way. Also, it is an alternative to describe the probability distribution of random variables.

In many industries, especially in manufacturing, people treat quality control as critical since they want to free from defects to improve their productivity. Mistakes can bring someone or a company into trouble. Even if the error is not very significant, it may bring some negative noise about the company, just like my laugh when the glass of Tesla's Cybertruck is broken. Statistical process control is often adapted to monitor and control the quality of a process. It measures conforming products to discover potential problems with the process, which is important in decision making.

In the article "SPC of a Near Zero-Defect Process Subject to Random Shocks, Quality and Reliability Engineering International", Xie and Goh suggest a 2 stage model deal with the issue in long terms of sampling with conformities and non-conformities. In quality control, we often use u-chart and Shewhart control chart to measure the quality. However, there are many points fall outside the upper control limit and result from a large number of false alarms when unidentified and uncontrollable mechanism take place to cause non-conformities. Simply speaking, let's imagine there is a product line which produces either a perfect unit or infected one. Each unit contains many items, tends to infinity here. The reason for defective products is that there is an unexpected mechanism, which named as "shocks", and it has a probability p about the occurrence. Additionally, within the infected unit, the number of defected items is random variable following Binomial distribution with parameters n. Since each product is independent, the probability of having k non-conformities within a non-conforming unit does not change throughout, it is Binomial distributed but not Hyprtgeometric. As mentioned previously, faulty may put some stakeholders at risk. Thus, Xie and Goh propose the 2 stage model to test if the product line is going worse, known as the situation of "out of control".

Xie and Goh demonstrate the 2 stage model with 2 detailed descriptions about each product unit. The first one is that there is a probability p that "shocks" happen. When "shocks" occurs, the current product will defect. In general, the probability of "shocks" is supposed to be small. However, "shocks" can make any number of infected items, known as non-conformities. Those infected items follow Binomial distribution with parameters n, which is the second assumption made by authors. It is true that the probability mass function of Binomial distribution is similar to that of Poisson distribution when the size n tends to infinity, given small p.

$$\lim_{\mathrm{n}\to\infty} \binom{n}{x} p^x (1-p)^{n-x} \approx \frac{\lambda^x e^{-\lambda}}{x!}$$

Thus, to calculate friendly, the authors make use of the Binomial approximation to Poisson distribution with parameter $\lambda = np$. Having the above descriptions, we can easily understand that the probability of k non-conformities within a non-conforming unit.

$$P(k\ defective\ items\ in\ a\ defective\ unit) = \frac{\lambda^k e^{\lambda}}{k!}$$

The 2 stage model, as the name implies, has 2 stages. The model aims to determine the probability of "shocks" p and the average number of non-conformities within a non-conforming unit λ. Therefore, the primary step of the model is to find out the general equations of the interested parameters. On the one hand, it is noticeable that perfect items in the sample are all items within the perfect unit and perfect items within the defected unit. In such a way, it can be affirmed mathematically by the law of total probability.

$$\begin{aligned} P(perfect\ items\ in\ the\ sample) &= P(prefect\ unit) + P(defective\ unit)P(prefect\ items\ in\ defective\ unit) \\ &= (1-p) + pe^{-\lambda} \end{aligned}$$

On the other hand, it establishes another statement which is infected items in the sample happen only when the unit is defective, such that

$$\begin{aligned} P(k\ defective\ items\ in\ the\ sample) &= P(defective\ unit)P(k\ defective\ items\ in\ defective\ unit) \\ &= p\frac{\lambda^k e^{\lambda}}{k!} \end{aligned}$$

In the first stage, the conditions of parameters p and λ are obtained. Consequently, in the second stage, the model estimates parameters p and λ using maximum likelihood estimation (details of estimation is omitted as it is out of syllabus in STAT2001).

Having the estimated parameters p and λ, how can we make use of these statistics? Xie and Goh illustrate the production exists “out of control” situation when the estimated probability p is significantly larger than the proposed one. Apart from that, there is a signal that is “out of control” situation when the probability p remains unchanged but the average of non-conformities increases. It implies a higher chance that a defected unit contains a huge amount of infected items. With the above information, the one, who in charge decision making, needs to make some modification about the process of production.

The 2 stage model is realistic since it makes use of the frequency view about the probability. In such a situation can repeat many times under the same condition to discover the parameter by estimating the relative frequency of the event in the collection of experiment results. Other than the read-write error example given the authors, here is another application about the 2 stage model. There are many flights every day. We treat each flight as a product unit in this situation. There are random “shocks” happening in the engine on the aeroplane and “shocks” cause an aeroplane accident. The number of casualties is known as the number of defective items. It has a sense that although there is an accident on the aeroplane, it is still possible that the number of injuries and deaths is zero, which would be very nice. However, when a crash on the aeroplane happens, the number of casualties is extremely unimaginable (here we assume n tends to infinity). Thus, such a situation applicable to the 2 stage model. For the department of engineering, engineers can make use of 2 stage model to deal with such issue to guarantee the safety of passengers (the maximum likelihood estimation is out of scope in STAT2001, thus we skip the parts of estimation and decision making by last year data about aeroplane accidents.).

However, just as every coin has 2 sides, the 2 stage model also has both advantages and disadvantages. The 2 stage model ignores defected items caused by other reason than random “shocks”. The model focus on the defection due to random “shocks” which is not precise for the whole process. Illustrate by the above example, there may be a passenger pass his or her lives by food poisoning and it is obviously not related to the “shocks” in the engine. We can only consider the 2 stage model as a small part in decision making since we know the probability of random “shocks” is very small. Besides, the 2 stage mode only make use of the empirical data which may not reflect the true situation. The data are biased for the instability of the process. When there is a consecutive large or small non-conformities in the sample, the model can only take account of the inputted data. As the model has so many bugs in considering the situation, the results of the model are not accurate for professional decision making, especially some circumstances involving human life similar to the above example.

Spite of everything, defection caused by random “shocks” is identifiable using the 2 stage model in a simple and practical way. The model is convenient for decision making in a random “shock”. If a production involves more than one random “shocks”, it is possible to use a bivariate method to improve the 2 stages in a multiple dimensions way. Although the proposed model is not perfect, it is very user-friendly to someone who new to probability model and quality control management.