NEWS-BASED LEARNING OF STATISTICS

Wing K. Fung and Philip L. H. Yu
The University of Hong Kong, Hong Kong
plhyu@hku.hk

Almost every day we come across statistics in our newspapers. Understanding these figures correctly not only gives us a better understanding of our social environment but also helps to prevent us from being taken in by misleading advertisements. This project investigates the teaching and learning of statistics through the use of statistical figures commonly found in newspapers and other mass media. These statistics have been used in specially designed courses such as "How to Read Figures in the Newspapers," a general education course offered to undergraduate students with or without a statistics background. Students find the topics interesting and appreciate the wide-ranging applications of statistics in different areas.

INTRODUCTION

Hong Kong, a small place of about 1,000 square kilometers with nearly 7 million people, is one the most densely populated areas in the world. It is a free society, and many international press companies have established their offices here. Competition in the mass media is keen in Hong Kong, and many news stories are often reported in detail.

People in Hong Kong come across statistics nearly every day in newspapers and on television and radio. Sometimes these figures are accompanied with explanations but in most cases they are not. These figures are often misinterpreted, misused or even abused, either unintentionally or intentionally. Understanding them correctly not only gives us a better understanding of our social environment but also prevents us from being misled. Since statistics are often found in local newspapers, a few years ago our Department introduced a new course entitled *How to Read Figures in the Newspapers* which covered the following topics: keys to index numbers, stock price movement-any trend?, modern Sherlock Holmes-forensic DNA, miracle or randomness, and validity of opinion poll. Since then, a few other courses such as *Crime and Punishment in Hong Kong* and *Understanding Financial Markets* which are largely news- and mass media-based have been set up. Moreover, in other courses we also use statistics found in the media such as SARS statistics, cancer probabilities, random sampling, and attempts to estimate attendance at rallies and marches. Students find this kind of news-based learning of statistics interesting and appreciate the wide-ranging applications of the subject.

In this paper, we illustrate our ideas with a number of examples of news-based learning of statistics. These examples are non-standard and look interesting; more standard examples, e.g., opinion poll and random sampling, are not considered. Students with an elementary statistics background are able to understand these examples.

EXAMPLE 1: A SERIES OF SEXUAL ASSAULTS

In the mid 1990s there was a series of sexual assaults in Hong Kong. This news was widely reported in local newspapers. The DNA fingerprinting/profiling technique, which was introduced in Hong Kong in the early 1990s, proved very useful to the local police since it had indicated that ten rapes and three murders in Tuen Mun had been committed by a single man, the so-called Tuen Mun Raptist. Since then, the powerfulness of DNA profiling for serious offence investigations has been widely reported from time to time in the mass media. Many students are interested in the technique and would like to know why and how statistics can be applied to crime investigation (Fung, 2001).

Consider a case where a crime has been committed. A blood stain (from a perpetrator) is found in the crime scene and a suspect has been identified. Suppose that the DNA found at the crime scene is typed with alleles (distinct types or lengths of DNA) {8, 9} (a person inherits two alleles from his or her parents) at the DNA locus TPOX and that the DNA of the suspect has the same alleles. Under the current legal system, the frequency or the random match probability of the alleles has to be calculated and reported in the courtroom. This can be obtained as $2p_8p_9$, where p_8 and p_9 are the frequencies of alleles 8 and 9 respectively. In the Hong Kong Chinese population, this frequency or match probability is evaluated as $2 \times 0.545 \times 0.1 = 0.109$; see Table 1 (Wong *et al*., 2001).

Table 1: Allele frequencies for locus TPOX in Hong Kong Chinese

Allele	8	9	10	11	12
Frequency	0.545	0.100	0.022	0.313	0.020

Many countries use 9 or more loci (of different chromosomes) for forensic DNA analysis. The overall match probability can be obtained by the product rule by multiplying individual probabilities over all loci. Therefore, it is not uncommon to find a figure as small as 1 in 1 billion. This shows the powerfulness of the product rule. Students find it amusing that such a small probability can be obtained based on a small number of loci.

Consider two possible explanations/hypotheses for the blood stain evidence, H_p: the perpetrator left the blood stain, and H_d: some other person left the blood stain. Based on the product rule, we may have arrived that "the chance/probability of observing this blood type if the blood came from someone other than the suspect is 1 in 1 billion." In the presentation of DNA evidence or forensic evidence in general, it was not uncommon in the early court cases which admitted DNA evidence to find a (incorrect) statement in the courtroom in such a form as "the chance that the blood came from someone else is 1 in 1 billion." People often mix up the meanings of these two quoted statements. They are explained as follows.

Let G_c be the DNA typing results for the crime sample. The first quoted sentence means

$$P(G_c \mid H_d) = 10^{-9} \text{ (1 in 1 billion),}$$

while the second says

$$P(H_d \mid G_c) = 10^{-9}.$$

The second sentence is equivalent to saying that "the chance that the blood came from the suspect is 99.9999999%." This is a very strong statement, suggesting a high chance that the suspect is the contributor of the blood stain. In fact, from the probability statement given above, it is clear that the order of H_d and G_c is mixed up in the second sentence in the assessment of the conditional probability. This is often called the error of the transposed conditional, and has been termed the "prosecutor's fallacy" (Thompson and Schumann, 1987; Evett and Weir, 1998), though defence lawyers may also make such a mistake.

Unfortunately, lawyers have often fallen into this kind of error. The following lists some of the examples (Evett and Weir, 1998): "*You testify that there is a 1 in 71 chance that a pair of contributors at random could have left the stain*" (people v. Simpson); "*Q: So the likelihood of this being any other man than Andrew Deen is one in three million? A: In three million, yes*" (R. v. Deen, U.K., 1994); "*Q: What is the combination, taking all those [alleles] into account? A: Taking them all into account, I calculated the chance of finding all those bands [alleles] and the conventional blood groups to be about one in 40 million. Q: The likelihood of it being anybody other than Alan Doheny? A: Is about one in 40 million.*" (*R. v Doheny, R. v G. Adams,* UK, 1997).

Students learn the importance on the correct interpretation of the conditional probability, and the serious consequences that can ensue from making the error of the transposed conditional.

EXAMPLE 2: MISFORTUNE OR COINCIDENCE

In February 2001, the mass media reported that, over a period of two years, 5 out of 48 professors in the Department of Applied Social Studies, Hong Kong Polytechnic University, had some form of cancer. The Department shared a building with the Department of Radiotherapy, and both the University itself and the public in general wondered if that was the cause. Was it a misfortune or coincidence?

Students are encouraged to discuss the above problem and explore different ways of formulating a problem. One possible solution is given below. Let p be the probability for an individual to have some form of cancer in a period of two years. According to Hong Kong

Cancer Registry of Hospital Authority, there were about 43,000 new cancer cases registered in Hong Kong during that period, and in a population with 6.8 million people, the value of *p* may be taken as 43,000/(6.8×10^{6}) = 6.32×10^{-3}. In a group of $n = 48$ randomly selected persons, the mean number of new cancer cases is equal to $\lambda = np = 0.30$. The probability that 5 or more people in such a group would get some forms of cancers could be evaluated as $\sum_{x\geq5} e^{-\lambda}\lambda^{x} \big/ x! = 0.000016$, which is very small. However if we group people in Hong Kong randomly into groups of 48, we expect on average $0.000016\times6.8\times10^{6}/48 = 2.3$ such groups to have 5 or more new cancer cases. It is still possible to observe such an incidence in a large population, but the expected number of occurrences is in fact rather small. Students can comment on the assumptions used in this calculation method. For example, is the sample of 48 professors random? Is the probability small enough? How small is small?

To address the small probability problem, one can compare the group of 48 professors who were situated next to a Department of Radiotherapy with a control group of 48 professors who were not next to a Department of Radiotherapy. Then a two-by-two table between case/control and cancer/ non-cancer can be constructed and a chi-square test of independence can be used to test for the problem. Of course, students may comment on how to select the control group and more than one answer for different control groups.

EXAMPLE 3: ESTIMATES OF THE NUMBER OF MAINLANDERS WITH RIGHT OF ABODE IN HONG KONG

In late 1998/early 1999, the Court of Final Appeal in Hong Kong made a ruling on Mainlanders with Right of Abode in Hong Kong. The ruling would have made Hong Kong permanent residents’ children born in the Mainland and children “born out of wedlock” of Hong Kong permanent residents eligible for the Right of Abode in Hong Kong. The Census and Statistics Department (C&SD) in Hong Kong was asked to conduct a survey to estimate the number of Mainlanders with Right of Abode.

The C&SD used a randomized response technique (Warner, 1965) for estimating the figure, and in mid 1999 an estimate of more than 1 million people (a sum over different categories including the first and second generations) having Right of Abode was reported. The issue was given headline coverage in the mass media for a few weeks and the validity of the statistical technique used was widely discussed.

The students were very interested in the way of obtaining the estimates when the newspaper cutting was shown to them. To simplify the problem for the students, it was slightly modified and posed in the following form.

The survey interviewee was told to select randomly a film holder from a bag that contained 10 film holders; 4 of which had a cover (C) and 6 that did not. Only the interviewee knew if the selected film holder had a cover. If it did, the interviewee was informed to answer question 1: Do you have children in Mainland China? Otherwise, the interview would answer question 2: Did you take any taxi trip in the last 7 days? So the interviewee would only answer Yes or No. But other people did not know which question he really answered. Suppose 24% of the respondents answered Yes. Let’s assume that 30% of the people indicated that they had taken taxi trips in an independent survey conducted in the same time period. Based on the total probability formula, we have

$$0.4p + 0.6\times0.3 = 0.24,$$

giving

$$p = 0.15$$

which is an estimate of the proportion of people who have children in Mainland. (In the real situation, the questions asked about the number of children the interviewees had in Mainland, and the number of times they had taken taxi trips in the last 7 days.)

EXAMPLE 4: CASE FATALITY RATE OF SARS

Three years ago the onset of severe acute respiratory syndrome (SARS), a lethal new infectious diseases, took Hong Kong and many other cities completely by surprise. According to a report published by the World Health Organisation (2003), out of a cumulative total of 7761 probable cases reported from 28 countries, 623 people died of SARS. The estimation of the case fatality rate of SARS has often been discussed in the community. According to WHO, the case fatality rate measures the proportion of all people with a disease who will die from the disease. During the SARS outbreak, the cumulative number (*T*) of reported SARS cases consisted of three components, namely, the cumulative number of fatalities (*F*), the number of recovered or discharged patients (*R*), and the number of patients remaining in wards (*W*). At the earlier stage of the outbreak, the WHO estimated the case fatality rate by $F / (F + R + W)$, and this calculation was used in many other affected areas such as Canada and Hong Kong. For example in a media session on 16 April 2003, the Secretary for Health, Welfare and Food of Hong Kong, Dr E.K. Yeoh, reported that the total number of SARS cases by that date was about 1200 and the death rate was 5%. However, at that time more than 70% of cases were still hospitalized. This estimate implicitly assumes that all patients currently hospitalized will recover. The WHO admitted on May 23, 2003 that this method underestimated the true rate unless the outbreak was over.

In fact, the estimation of case fatality rate can be posed as a problem of compositional data analysis. Note that there are three compositional proportions $f = F / T$, $r = R / T$ and $w = W / T$, where $f + r + w = 1$. The technique of compositional data analysis developed by Aitchison (1986) together with the ternary diagram is relevant to analyse the three compositions. Figure 1 shows the simplex diagram of the 3 compositional proportions (*r*, *w*, *f*) of the SARS cases in Hong Kong for the period April 1 – May 22, 2003 (running from left to right). We illustrate the diagram as follows: For example, as of April 30, which corresponds to the large solid circle point, the 3 proportions (r, w, f) measure about 0.5, 0.4 and 0.1. That is, 50% of the SARS patients had recovered, 40% still remained in wards and 10% had died. Using 10% as the case fatality rate was definitely an underestimate since it implicitly assumed the remaining 40% in wards would all recover.

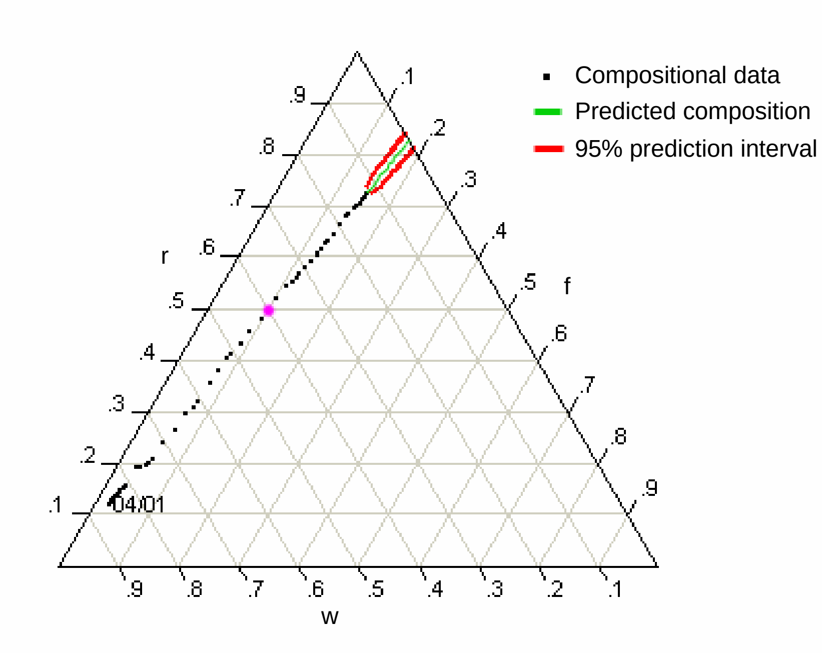

Figure 1: A simplex diagram of the compositional proportions of the SARS cases in Hong Kong for the period April 1 – May 22, 2003. A coordinate (r, w, f) represents the proportion of SARS patients who have recovered or discharged (r), still remain in wards (w), and have died (f).

In the simplex diagram, all the observations lie very close to a straight line. This linearity phenomenon indicates that the case fatality rate did not fluctuate over the period. We fitted a vector auto-regressive time series model to the log-ratios, $\log(F/R)$ and $\log(W/R)$, of the compositional data. Note that the model is invariant to the choice of the denominator. The predicted compositions were plotted in Figure 1, together with the confidence bands. Based on the model, we predicted that less than 1% of the SARS patients would remain in wards after late June, 2003. The final case fatality rate was predicted to be 17.0%, with a 95% prediction interval of (15.5%, 18.5%). This prediction was very close to the actual fatality figure of 17.0% (299 deaths out of 1755 cases).

When presenting various ways of estimating case fatality rates to students, this would help students appreciate the importance of a proper definition of a statistical measure. More discussions on statistical exploration from SARS can be found in Yu, Chan and Fung (2006) and Chan, Yu, Lam and Ho (2006).

EXAMPLE 5: COUNTING ATTENDANCE AT A MARCH

On July 1, 2004, hundreds of thousands of Hong Kong people marched in the streets to demand a more democratic system of government. How many participants marched that day? Many organizations presented different estimates. The march’s organizer, the Civil Human Rights Front, estimated that 530,000 people took part, while estimates made by the Hong Kong Police and Mingpao (a Hong Kong Chinese newspaper) put the figure at around 200,000. Which estimate is more reliable? Although they all adopted a density estimation approach, the accuracy of the count estimate largely depends on the accuracy of the density estimate. The Civil Human Rights Front later admitted that they took the largest possible density, i.e., the density obtained when the streets were fully crowded (around 2.5 persons per square feet). However, Mingpao found that the density was only around 0.85 to 1 persons per square feet).

An alterative approach is to use systematic sampling procedure by counting the number of marchers passing though a checkpoint during a time period (say 1 minute) on a regular basis (say for every 5 minutes). This method was adopted by several research teams from the University of Hong Kong, and their estimates ranged from 165,000 to 192,000.

No matter which estimate method is used, several issues must be taken into account: 1) Who should be counted? In other words, what is the definition of the target population? 2) How to take into account the fact that density is not a constant but time varying? and 3) How to estimate those marchers who left before reaching the point and those who joined midway?

Even once these issues have been appropriately taken into account, we still need to understand how large the sampling error is. Most of the time, a single point estimate is reported but because of sampling variation, it is possible that the actual number would vary somewhere around the point estimate. One crucial issue is on the size of the sampling variation. The estimate 200,000 ± 20,000 is clearly more reliable than the estimate 200,000 ± 50,000, given that both estimation methods are valid.

Students taking survey sampling methods can read different calculation methods from newspapers and comment on their methodologies and validity of their assumptions made. By showing some video clips of the march in the lecture, students can realise the practical difficulty of data collection in conducting a survey.

CONCLUDING REMARKS

The examples that we used were taken from real-life situations widely reported in the mass media, and students were very interested in discovering how those results were obtained. They were also curious about some of the figures, e.g., why over 1 million people have the right of abode, and why such a very high percentage of people (5 out of 48) had some form of cancer over a two-year period. Many of the statistical methods that we used can be grasped by students with a very limited statistics background. As many students find statistics boring and difficult to understand, we hope this news-based learning can make students find statistics interesting and appreciate the wide-ranging applications of the subject.

ACKNOWLEDGEMENTS

The authors thank the referees and Dr. David Wilmshurst for helpful comments that improved the presentation of the paper. This work is partly supported by a CRCG research grant of the University of Hong Kong.

REFERENCES

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Chan, J. S. K., Yu, P. L. H., Lam, Y. and Ho, A. P. K. (2006). Modeling SARS data using threshold Geometric Process. To appear in *Statistics in Medicine*.

Evett, I. W. and Weir, B. S. (1998). *Interpreting DNA Evidence*. Sunderland, MA: Sinauer Associates, Inc.

Fung, W. K. (2001). Teaching statistics using forensic examples (Invited Paper). *Bulletin of the 53rd Session of the International Statistical Institute*, 183-186.

Thompson, W. C. and Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trails. The prosecutor's fallacy and the defence attorney's fallacy. *Law and Human Behavior*, 11, 167-187.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

Wong, D. M., Law, M. Y., Fung, W. K., Chan, K. L., Li, C., Lun, T. S., Lai, K. M., Cheung, K. Y. and Chiu, C. T. (2001). Population data for 12 STR loci in Hong Kong Chinese. *International Journal of Legal Medicine*, 114, 281-284.

World Health Organisation. (2003). Severe acute respiratory syndrome (SARS): Status of the outbreak and lessons from the immediate future. *Communicable Disease Surveillance and Response Report* at Geneva on 20 May 2003.

Yu, P. L. H., Chan, J. S. K. and Fung, W. K. (2006). Statistical exploration from SARS. To appear in *The American Statistician*.