MAST20004 Probability

Tutorial Set 4

1. The data below are taken from a study by Geissler and are given in Fisher's book "Statistical methods for research workers". A large number (53,680) of German families with 8 children were contacted, and for each family, the number X of male children was noted. The results are given below.

x	0	1	2	3	4	5	6	7	8
obs freq	215	1485	5331	10649	14959	11929	6678	2092	342

   (a) Calculate the relative frequencies.

   (b) Compare them with the values of the probability mass function of the binomial distribution with parameters $n = 8$ and $p = 0.51$ (the value $p = 0.51$ is considered to be the probability of a male birth for Caucasian parents). Comment on your findings.

   NB. In the subject of statistics, we are often expected to use data such as this to *estimate* the value of a parameter. For example, imagine that we have good reason to think that these data come from a binomial model with some value of p, but we don't know this value. How can we decide on the best value of p that is consistent with the data?

   Many of you will study questions like this in MAST20005 Statistics next semester.

Solution: If $X \stackrel{d}{=} \text{Bi}(8, 0.51)$ then, for $x = 1, 2, \ldots, 8$, $p_X(x) = \binom{8}{x}(0.51)^x(0.49)^{8-x}$.
The answers are given in the following table.

x	0	1	2	3	4	5	6	7	8
obs freq	215	1485	5331	10649	14959	11929	6678	2092	342
rel freq	0.0040	0.0277	0.0993	0.1984	0.2787	0.2222	0.1244	0.0390	0.0064
bin pmf	0.0033	0.0277	0.1008	0.2098	0.2730	0.2273	0.1183	0.0352	0.0046

The relative frequency and the binomial probability mass function look 'quite similar'. Thus, on an intuitive level, we might be tempted to conclude that the binomial model with $p = 0.51$ is indeed a good model for children of the same parents. This would indicate, for example, that the sexes of different children in the same family are independent.

However, to make statements such as this, we really need a more rigorous analysis. This analysis is the realm of statistics, which many of you will study in MAST20005 Statistics. Two immediate questions are

1. If we take the binomial model as given, what is the value of p that gives the best fit? This is *parameter estimation*.
2. How close is the distribution given by the relative frequency to the theoretical pmf. This is a question of *goodness of fit*.

2. A game is played in which a coin is tossed until a tail turns up. A payout is given depending on the number of heads that turn up before the first tail. What is the expected payout if

   (a) the payout for n heads is \$$n$?

   (b) the payout for n heads is \$$2^n$?

   Solution: The number N of heads before the first tail is a geometric random variable with probability mass function $P_N(n) = (1/2)^{n+1}$. Therefore, when the payout for n heads is \$$n$, the expected payout is

$$\sum_{n=0}^{\infty} n \left(\frac{1}{2}\right)^{n+1} = \left(\frac{1}{2}\right)^2 \frac{1}{(1-1/2)^2} = 1.$$

   When the payout for n heads is \$$2^n$, the expected payout is

$$\sum_{n=0}^{\infty} 2^n \left(\frac{1}{2}\right)^{n+1} = \frac{1}{2} \sum_{n=0}^{\infty} 1,$$

   which does not exist. This is an example of a random variable for which the expectation is not defined. Some of you may have heard of the gambling strategy where a player doubles his/her bet every time he/she loses, so that when he/she eventually wins he/she will be ahead. The expected loss for such a gambler is like the expected payout in the 2^n case above. It does not exist, because the series diverges to infinity. This is the St. Petersburg Paradox that was discussed in lectures – see Slides 7 and 162 - 164.

3. Consider a sequence of independent trials where three outcomes are possible, success (S), failure (F), and neither success nor failure (N), with $\mathbb{P}(S) = p$, $\mathbb{P}(F) = q$, and $\mathbb{P}(N) = 1-p-q$. Let X be the number of successes and Y the number of failures in n trials. Write down an expression for $\mathbb{P}(X = x, Y = y)$ (this is known as the *joint* probability mass function of X and Y).

   Solution: The probability of a sequence of n trials with x successes and y failures is $p^x q^y (1-p-q)^{n-x-y}$. The number of ways we can get x successes and y failures in n trials is $\dfrac{n!}{x!y!(n-x-y)!}$. Therefore, for $0 \le x, y \le n$ with $0 \le x + y \le n$,

$$\mathbb{P}(X = x, Y = y) \;=\; \frac{n!}{x!y!(n-x-y)!}\, p^x q^y (1-p-q)^{n-x-y}.$$

   This is a special case of the *multinomial* distribution. Here we have a sequence of n independent trials with k possible outcomes, say, $1, 2, \ldots, k$, with respective probabilities of occurring, $p_1, p_2, \ldots, p_k$. If, for $i = 1, 2, \ldots, k$, X_i is the number of times outcome i occurs, then, for $0 \le x_1, x_2, \ldots, x_k \le n$ with $x_1 + x_2 + \ldots + x_k = n$,

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) \;=\; \frac{n!}{x_1! x_2! \ldots x_k!}\, p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}.$$

4. Let $X \stackrel{d}{=} \mathrm{Nb}(r, p)$ and $Y \stackrel{d}{=} \mathrm{Nb}(s, p)$, where r and s are positive integers. What is the distribution of $X + Y$?

   Solution: In a sequence of Bernoulli trials with probability of success p, X can be interpreted as the number of failures before the rth success. Similarly, Y can be interpreted as the number of failures before the sth success. Thus, $X + Y$ can be interpreted as the number of failures before the $(r+s)$th success. Therefore, $X + Y \stackrel{d}{=} \mathrm{Nb}(r+s, p)$.

5. Suppose a jar has n chips numbered $1, 2 \ldots, n$. A person draws a chip, replaces it, draws another and returns it and so on until he gets a chip which has been drawn before. He then stops. Let X be the number of drawings required to accomplish this objective. Find
   (a) The distribution function $F(k)$ of X. (Hint: It is easiest first to compute $\mathbb{P}(X > k)$.)
   (b) The probability mass function $p(k)$ of X.
   (c) Find an expression (a sum) for $\mathbb{E}(X)$.

   Think about how this question relates to the 'birthday problem' that we mentioned in Lecture 1.

   Solution: Note that $S_X = \{1, 2, \ldots, n+1\}$.

   (a) For $2 \leq k \leq n$, the event $\{X > k\}$ occurs only when the first k selected chips are all different. So

$$\mathbb{P}(X > k) = \frac{n-1}{n}\frac{n-2}{n}\cdots\frac{n-k+1}{n} = \frac{(n-1)!}{(n-k)!n^{k-1}}.$$

   Therefore, for $2 \leq k \leq n$,

$$F(k) = 1 - \frac{(n-1)!}{(n-k)!n^{k-1}}.$$

   $F(k) = 0$ for $k < 2$ and $F(k) = 1$ for $k \geq n+1$.

   (b) For $2 \leq k \leq n+1$,

$$\begin{aligned} p(k) &= F(k) - F(k-1) \\ &= \frac{(n-1)!}{(n-k+1)!n^{k-2}} - \frac{(n-1)!}{(n-k)!n^{k-1}} \\ &= \frac{(n-1)!}{(n-k+1)!n^{k-2}}\left[1 - \frac{n-k+1}{n}\right] \\ &= \frac{(k-1)(n-1)!}{(n-k+1)!n^{k-1}}. \end{aligned}$$

   (c)

$$\mathbb{E}(X) = \sum_{k=2}^{n+1} k\frac{(k-1)(n-1)!}{(n-k+1)!n^{k-1}}.$$

   This can be calculated easily for any fixed n, but a neat analytic expression for this sum in terms of n is not easy to derive.

   For the birthday problem think of the 'chips' as dates and the selection process as corresponding to randomly selected people choosing the chip corresponding to their birthday. Thus $n = 365$.

MAST20004 Probability

Computer Lab 4

In this lab you

- simulate a two stage random experiment involving coins with 'random bias' i.e. different types of coins have differing probabilities of coming up heads.

MATLAB programming

In this lab you will need to study the operation of some parts of the supplied programs. To help you with this remember to

- use the help system (you can get help on any specific command simply by highlighting it in the program, right clicking and choosing 'Help on Selection' from the contextual menu. The lookfor command is also useful when searching for keywords.);
- follow the hints built into the comment lines in the programs themselves;
- continue to work in groups of 2 or 3.

Exercise A - Coins with random bias

1. Consider the following two stage random experiment.
   (a) In the first stage of the random experiment we choose a coin at random and with replacement from an urn containing a mixture of different types of biased coins. The different coin types have different probabilities of coming up heads when you toss them. We will be able to vary the probabilities of getting a head across the different types and also the proportion of the different types of coins in the urn.
   (b) Initially there will be just three types of coins present in equal numbers - one third with $\mathbb{P}(H) = 0.05$, one third with $\mathbb{P}(H) = 1/2$, and one third with $\mathbb{P}(H) = 0.95$. Therefore you can think of the random variable $P = \mathbb{P}(H)$ for a randomly chosen coin as having a discrete uniform distribution over those three values.
   (c) In the second stage of the random experiment we toss the chosen coin a fixed number of times and count X, the number of heads. Initially we do 100 tosses.
   (d) We can think of this experiment as involving the 'randomisation' of the p parameter in a Binomial distribution, effectively replacing it by a discrete random variable P. So we can write $X \stackrel{d}{=} \text{Bi}(n, P)$ as an extension of the fixed parameter notation $X \stackrel{d}{=} \text{Bi}(n, p)$.
   (e) The program **Lab4ExA.m** which simulates this two stage random experiment is available on the Maths and Stats server, together with its supporting function.
2. For the initial set up of coin types and proportions in the urn write down the probability mass function for P and calculate $\mathbb{E}(P)$ and $V(P)$.

3. Use the m-file editor to open up **Lab4ExA.m** and the function **rand_discrete.m** that it calls and study these programs carefully. The 'random-prob-of-head' variable calculated in stage 1 uses a call to the function 'rand_discrete'. Study this function and try to figure out for yourself how it works using the hints in the program itself. It uses a standard simulation technique which may be familiar to some of you. Also using help for various commands try to understand how the variable 'numheads' is calculated in stage 2. Your tutors will give you a short time to think about these tasks and then explain the answers to the whole group so you can move on.

4. Run **Lab4ExA.m** and consider the empirical probability mass function that it generates. Can you explain its shape? Test your understanding by changing the set up for the urn - both the possible and number of values for the probabilities of a head and the relative proportions of coins of different types.

5. Using your insights into the form of the probability mass function for X, can you calculate $\mathbb{P}(X = 2)$ theoretically for the case where 'ntosses' is set to 10 and the initial set up for the urn is used? [**Hint:** use the law of total probability and the probability mass function for the binomial distribution. Use your calculator (or program Matlab to do the calculations if you wish!)].

6. Run the simulation for this set up and check your theoretical calculations against the simulation output. You may wish to repeat the simulation with a larger number of repetitions.

7. Now set the initial urn set up to have equal proportions of coins with probabilities of tossing a head equal to all tenths from 0.1 to 0.9 (use **possvalues=[0.1:0.1:0.9]** and **probmasses=(1/9)*ones(1,9)**) and set 'ntosses' to 100. Run the simulation and think about the outcome. Conjecture what output would be produced by having an urn with equal proportions of coins with $\mathbb{P}(H) = 1/4$, or $\mathbb{P}(H) = 1/2$, or $\mathbb{P}(H) = 3/4$. Test your conjecture by running the simulation for that set up.

8. (Extension task) Extend your conjecture by considering what the distribution of X might be if $P = \mathbb{P}(H)$ was uniformly and **continuously** distributed on [0,1]? Test your conjecture by modifying **Lab4ExA.m**. (You only need to change slightly one line of the program).