Engsci 213 Discrete Probability Models - Problems (v. 1)

Oliver Maclaren
oliver.maclaren@auckland.ac.nz

Discrete probability models

Let's consider some simple problems involving discrete probability models.

The first part will be problems to do by hand, the second part is about using R to do probability calculations and simulation. (You can use whatever language you want but you will use R in the next (Data Analysis) module, so may as well start getting used to it!)

Note:
There are lots of extra worked problems and resources on canvas. E.g. in the past course book. I also put up a free (GNU) 'Introduction to Probability and Statistics Using R' by Kerns, too. It has lots of R code as well as example hand calculations.

There are also plenty of R tutorials etc on the internet. See the notebook I put up from lectures for some starting points (or just Google).

Problems

Corrupted discrete distribution

As part of a programming competition you are given a discrete probability distribution to process. It comes in a text file but the information is corrupted. You can only read in the following table:

x	0	1	2	3
$P(X = x)$	0.10	0.41	?	?
$P(X \leq x)$	0.10	?	0.78	?

where ? means the value has been lost. You know that the maximum possible value of x is 3 (no full columns are lost).

- What are $P(X = x)$ and $P(X \leq x)$ called? What other symbols do we use for them?
- Fill in the missing details of the distribution.
- Calculate the expected value of X, $E(X)$.

- Calculate the variance of X, $Var(X)$.
- Calculate the standard deviation

Binomial distrbution

You're working for a company that loves 'test-driven' software development. The piece of software you are working on has to pass 10 tests this week, one for each 'module'. Each test of a given module consists of evaluting a potentially erratically-behaving function (within that module) 5 times. Each of these functions is capable of returning one of 4 possible values for each function call.

You know the desired (unique) ouput for each function call during a test. Your company wants the software to behave well 'on average', i.e. return the correct result some percentage of the time.

To get a starting benchmark for the reliability of the functions you decide to compare the results to what you would get just by choosing values randomly. Consider one test consisting of 5 function calls, with 4 possible return values to each function call.

You model this 'random baseline' by picking an output for each function call at random and independently of each other. Let X be the number of the randomly generated guesses that are correct for the 5 function calls.

- You model this with the Binomial distribution. What are the parameters?

The Binomial table for this case is:

x	0	1	2	3	4	5
$P(X = x)$	0.2373	0.3955	0.2637	0.0879	0.0146	0.0010

- Verify it using a calculator/computer and the formula for Binomial probability function
- What is the probability of getting at least 3 function call results correct? Write out your reasoning explicitly

Remembering that you need to carry out 10 such tests, you define the new variable

$$T = X_1 + X_2 + ... + X_{10}$$

- What does this variable represent?
- What distribution do you expect it to have?
- What is $P(T = 10)$?
- If the company requires an overall success rate of at least 70% correct results from function calls, describe how you might calculate this.

Poisson processes and the Poisson distribution

You are tracking email arrivals to you email account. Emails arrive in your inbox one-by-one according to a Poisson process with rate $\lambda = 20$ emails per hour.

- What is the probability of receiving exactly 13 emails in a given hour? Evaluate it if you have a calculator, otherwise write out the expression to evaluate.
- What is the probability of receiving at least 3 emails in a given hour?
- What is the distribution modelling the number of emails in a (8-hour) working day period?
- What is the probability of receiving no more than 3 emails in one working day?

Using R

Interfacing with R

Open the PDF version of the R Markdown notebook I put on Canvas. See the links at the beginning for getting started with R. Alternatively, see the book 'Introduction to Probability and Statistics Using R' by Kerns, also on Canvas.

If you get stuck I recommending Goolging it - 'Stack Exchange' links are usually good! If you program often you will inevitably end up using this as your default strategy! E.g. try 'plot binomial distribution in R'. There will be lots of good links. Links starting with https://stat.ethz.ch give the official code documentation.

Basics

By referring to e.g. http://www.statmethods.net/advgraphs/probability.html, the lecture notebook or any other source of documentation

- plot the binomial probability function for a series of n, p combinations
- plot the cumulative distribution
- do the same for the Poisson distribution

Now, choose suitable n, p combinations so that the Poisson becomes a good approximation to the Binomial distribution. (See the lecture notebook if you get stuck).

- Plot a series of comparisons of the Binomial distribution given n, p values and the Poisson distribution for $\lambda = np$. Plot the Binomial distribution as a bar plot/probability histogram and the Poisson distribution as a line plot on the same figure.

Note that to do this last part with in-built functions is slightly subtle as the bar plots and line plots have different axes. See my notebook and/or http://www.r-bloggers.com/adding-lines-or-points-to-an-existing-barplot/.

Simulation

Consider the R command:

```
rbinom(10, 4, .7)
```

- Run it in R. What do you think it does? Search online to find out what.
- Plot a histogram of the output. Calculate the mean, variance and standard deviation.
- Do the same for different values.

Now, compare the simulated data to the theoretical distribution

- At the level of theoretical distribution/simulated histograms
- At the level of mean, variance and standard deviation summaries of the above full 'distribution information' (the summaries are called population parameters for theoretical quantities and sample statistics for simulated/sampled quantities)