ASGR Assignment 3 Questions

Aaron Robotham

1) [10 marks total]

Remember the example of rare blood testing at Perth Bayes Health. Alas, in Sydney they only have access to Sydney Bayes Health. The nominal test is the same as at Perth Bayes Health, so the test is 99.9% accurate at correctly determining you are positive, and 99% accurate at correctly determining you are negative. Again 99.9% of people do not have the disease. However, their practices are much sloppier than Perth Bayes Health- 5% of samples are mixed up, and 2% of the time the computer will return a random (50/50 positive/negative) result. Alas, you once again test positive. Whilst annoyed at the sloppiness of the testing procedure (no doubt you will move to Perth some time soon), you still want to know how worried you should be. What is the answer? [As a warning, LLM get this question very wrong, so do not be guided by those.]

Re marking - perfect solution = 10; good effort but +/- 0.1% = 9; good effort but +/- 0.5% = 8; good effort but +/- 3%/2% = 7; good effort but +/- 5%/3% = 6; good effort 5 but worse; various degrees of wrong effort 4-1; no effort 0.

---

2) [10 marks total]

a) [2] With the seed set to 2025, the following processes happen:
   i) 10,000 random samples are made of a Poisson distribution with $\lambda = 10$
   ii) 10,000 random samples are made of a t distribution with df=3
   iii) Following this a further 10,000 random samples are made from a Normal distribution where the means (μ) are specified by our Poisson samples from i (i.e. you have 10,000 different means) and the standard-deviations (σ) from the square-root of the absolute value of our samples from ii (i.e. you have 10,000 different standard deviations).

What is the mean, median and standard-deviation of the resulting distribution? What sampling process dominates the standard deviation of the final distribution?

b) [2] Remember from the lectures that the 1-σ range of the Normal distribution contains 68.3% of the probability mass.

   i) What value of *df* is required for the Student t distribution to contain the same percentage of probability mass within $-1.5 < q < 1.5$?
   ii) What base **R** distribution with default arguments contains this percentage of probability mass up to $q = 1.148854$?

c) [2] Which event is more likely: sampling $x = 2$ from a Normal ($\mu = 1$, $\sigma = 3$) followed by $x = 3$ from a χ^2 ($\nu = 10$), or sampling $x = 1$ from an exponential ($\lambda = 3$) followed by $x = 5$ from an F ($\nu_1 = 5$, $\nu_2 = 5$)?

d) [2] There are 1000 cows in a 100 hectare field. They wander about aimlessly and randomly. What is the chance that at a given time there are exactly 7 cows within a given hectare of the field?

e) [2] There is typically 1 Milky Way mass galaxy per Mpc^3 in the Universe. Surveys have been done of a few Gpc^3 of the Universe to date, and individual 100 Mpc^3 volumes often contain as little as 0 Milky Way mass galaxies, and sometimes 300+. What does this suggest about the distribution of galaxies in the Universe?

---

3) [10 marks total]

A bike factory produces sequentially numbered bikes, so in principal if 5 bikes were randomly found with the numbers 20, 40, 60, 80 and 100 you might reasonably be able to estimate how many bikes have been made by the factory to date. With this situation in mind, imagine if one (and only one) bike was found, and it had the ID #1001.

a) [5] What would a Bayesian statistician estimate as the mean, median and mode of the total number of bikes produces (assume no priors on the bike production distribution)?

b) [5] In a similar scenario, we know that the factory definitely cannot have built more than 10,000 bikes, so what would we estimate the mean, median and mode to be now?

---

4) [10 marks total]

a) [6] Starting from the origin, we make 100 steps sampling from a Normal distribution ($\mu = 0$, $\sigma = 1$) on a 100 dimension Cartesian grid (i.e. one Normal sample in each dimension). What percentage of the time will our final distance be greater than 8/9/10 from the origin?
   i) Show this via Monte Carlo simulation for a reasonable number of simulations.
   ii) Show this analytically using the appropriate distribution in **R**.

b) [4] Two chefs grab 5 (chef A) and 10 (chef B) potatoes from one of 2 supply bins in a kitchen that come from very different farms with very different potatoes that vary hugely in the average weights. They weigh the potatoes, giving the following in kilograms:

```
potAi
```

```
## [1] 1.0070508 0.9769823 1.0129288 0.9439524 1.1558708
```

```
potBi
```

```
##  [1] 1.1280555 1.0238213 1.2528337 1.0503812 0.8272729 0.8951107 1.1690184
##  [8] 1.0825540 1.1294763 1.0549097
```

i) If the null hypothesis is both chefs took their sample from the same bin, can this be rejected at the 5% level?

This happens again with new batches of potatoes where the farms concerned now vary in their weight variances. This time one of the chefs does not directly zero the scales so there is a systematic offset in the measured weights.

```
potAii
```

```
## [1] 0.8972054 1.0873289 1.0014641 1.0685665 1.0449437
```

```
potBii
```

```
##  [1] 1.528735 1.606035 1.616779 1.571635 1.649665 1.580785 1.595488 1.605457
##  [9] 1.562277 1.620391
```

ii) Again, can the null that the potatoes were taken from the same bin be rejected at the 5% level?

---

5) [10 marks total]

The file quake.csv (in the data directory) contains all Earthquakes greater than or equal to magnitude 8 since 1900 (taken from www.isc.ac.uk).

a) [2] How many magnitude greater than or equal to 8.1 earthquakes per year do we expect?
b) [4] How consistent with exponential are the delay times between magnitude greater than or equal to 8.1 earthquakes?
c) [4] Make a plot of where all magnitude greater than or equal to 8 earthquakes occurred on a map of the Earth, scaling the point size by the magnitude of the earthquake, and colours by the date (make the result visually appealing and intuitive). What can you say about the distribution?

There are quite a few bits of data wrangling required to tackle this question before even getting to the statistics part, and some plotting functionality in **R** is needed but has not been explicitly discussed in the course. Remember, Google is your friend (on this occasion at least).