

# Problem Set 3

C. Durso

February 06, 2019

These questions were rendered in R markdown through RStudio (<https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>, <http://rmarkdown.rstudio.com> ).

Please complete the following tasks regarding the data in R. Please generate a solution document in R markdown and upload the .Rmd document and a rendered .doc, .docx, or .pdf document. Your work should be based on the data's being in the same folder as the .Rmd file. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

1. This question revisits the problem of whether two null hypotheses that we applied to the polio trials are equivalent. Though they both resulted in very low probabilities of the data under the respective null hypotheses, you will show that they aren't exactly the same.

Consider two non-empty multisets  $C = \{c_1, c_2, \dots, c_n\}$  and  $D = \{d_1, d_2, \dots, d_m\}$  of binary outcomes, say “success” and “failure”, with  $k$  successes in  $C$  and  $l$  successes in  $D$ .

Define a random variable  $X$  as the number of successes in a sample of size  $n$  drawn without replacement from  $\{c_1, c_2, \dots, c_n, d_1, d_2, \dots, d_m\}$  in such a way that all samples are equally likely. Recall that the number of samples is  $\binom{n+m}{n}$ .

Define another random variable  $Y$  to be *Binomial*  $\left(k+l, \frac{n}{n+m}\right)$ .

(These are abstract versions of the probability models for a couple of the null hypotheses considered in `inference_example_polio.Rmd` if you take  $C$  to be the Vaccinated population,  $D$  to be the Placebo population,  $k$  to be the number of polio cases in the Vaccinated population, and  $l$  to be the number of polio cases in the Placebo population.

Then  $X$  corresponds to the null model that the number of polio cases in the two populations is consistent with pooling the two populations, randomly selecting  $n$  cases to be designated as Vaccinated and counting the number of polio cases in this sample.

The random variable  $Y$  corresponds to the null hypothesis that the  $k+l$  polio cases were assigned to the Vaccinated group with probability  $\frac{n}{m+n}$  and to the Placebo group with probability  $1 - \frac{n}{m+n} = \frac{m}{m+n}$ .)

Please calculate the probability that  $X = 0$  and the probability that  $Y = 0$ . Use these to demonstrate that distribution of  $X$  is not equal to the distribution of  $Y$  when  $k, l, m$ , and  $n$  are all non-zero. You may use the fact that  $0 < r \leq p < q$  implies  $\frac{p}{q} > \frac{p-r}{q-r}$ . Suggestion: simplify the factorials in the probability that  $X = 0$  in such a way that you have  $k+l$  terms in the numerator and denominator. (10 points)

2. The parts of this problem lead to a proof that that if  $X$  is a Poisson random variable with  $\lambda = \lambda_x$  and  $Y$  is a Poisson random variable with  $\lambda = \lambda_y$  and  $X$  and  $Y$  are independent, then the random variable  $W = X + Y$  is a Poisson random variable with  $\lambda = \lambda_x + \lambda_y$ . Please complete each part rather than providing a proof of the result directly. Recall that the probability that  $X = k$  is  $\frac{(\lambda_x)^k \exp(-\lambda_x)}{k!}$ , and likewise for  $Y$ . (2 points each)
  - a. Consider the probability space  $Q$  with outcomes in  $\mathbb{N} \times \mathbb{N}$  such that the first component has the same distribution as  $X$ , the second component has the same distribution  $Y$ , and the two components are independent. What is the probability of the outcome  $(1, 2)$ ?
  - b. The probability that  $W = n$  is equal to the sum of the probabilities of the outcomes  $(k, l)$  in  $Q$  for which  $k+l = n$ :  $P(W = n) = \sum_{(k,l) \text{ such that } k+l=n} P(X = k)P(Y = l)$ . Please rewrite this using a summation of the form  $\sum_{k=0}^n \dots$

- c. Rewrite the summation from 2.b using the Poisson formulas for the probabilities.
  - d. Recall the binomial formula:  $(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$ . Multiply your summation from 2.c. by  $\frac{n!}{n!}$ . Explain how rearranging terms verifies the claim that  $W$  is a Poisson random variable with  $\lambda = \lambda_x + \lambda_y$ .
3. Normal Distributions shift and scale while remaining Normal. This greatly simplifies working with Normal distributions. This property is not generally true of parametrized families of distributions. (5 points each)
- a. Prove that if  $X$  is a random variable with the  $Normal(\mu, \sigma^2)$  distribution, then  $X + a$  is a random variable with the  $Normal(\mu + a, \sigma^2)$  distribution. Please do this by showing that the cumulative distribution for  $X + a$  defined in terms of the cumulative distribution for  $X$  equals the cumulative distribution for a random variable with the  $Normal(\mu + a, \sigma^2)$  distribution.
  - b. Prove that if  $X$  is a random variable with the  $Normal(\mu, \sigma^2)$  distribution, then  $cX$  is a random variable with the  $Normal(c\mu, c^2\sigma^2)$  distribution. Please do this by showing that the cumulative distribution for  $cX$  equals the cumulative distribution for a random variable with the  $Normal(c\mu, c^2\sigma^2)$  distribution.
4. The code below draws samples of size 100,000 from the Poisson distributions with  $\lambda = 5, 25, 50, 100$  and puts them in a matrix. For each sample, calculate the maximum likelihood values of  $\mu$  and  $\sigma^2$  to approximate the sample as a sample from a Normal distribution. For each  $\lambda$ , plot the density histogram of the sample and the density curve of the Normal distribution with the maximum likelihood values of  $\mu$  and  $\sigma^2$  for the sample on the same graph. (7 points)

```
set.seed(34567)
mat<-matrix(c(rpois(100000,5),rpois(100000,25),rpois(100000,50),rpois(100000,100)),
            ncol=4)
```

5. Consider approximating the probability  $P(X = k)$  for a Poisson random variable  $X$  with parameter  $\lambda$  by the probability that a Normal random variable with parameters  $\mu = \sigma^2 = \lambda$  lies in  $(k - 0.5, k + 0.5)$ . These should be close to the values you found by simulation. (Remember the parameter “sd” is  $\sigma$ , not  $\sigma^2$ .) Using R, calculate the sum of the absolute values of the differences between the probabilities for  $k = 0, 1, \dots, 1000$  for the values  $\lambda = 5, 25, 50, 100$ . Please base the Poisson probabilities on the probability density function rather than on simulation.

This is a measure of how well the Normal distribution approximates the Poisson distribution: it’s the sum over all reasonably probable  $k$ s of the absolute value of errors from using the approximation.

What do you observe about this difference as  $\lambda$  increases? Note this is a computation based on the theoretical probabilities, not on simulations. (8 points)

6. Repeat exercise 5 using the Binomial distributions with probability of success=0.7 and the size parameter equal to 5, 25, 50, 100. This time take  $\mu = 0.7n$  and  $\sigma^2 = n(0.7)(1 - 0.7)$ . These are close to the values you would find by simulation. For each  $n$ , you can limit the summation to  $k = 0, 1, \dots, n$ . (4 points)
7. View the results of question 5 as giving measures of the quality of a Normal approximation for sums of 5, 25, 50, and 100 independent Poisson random variables with  $\lambda = 1$  and the results of question 6 as giving measures of the quality of a Normal approximation for sums of independent random variables that equal 1 with probability 0.7 and that equal 0 with probability 0.3. What do you observe about the quality of the approximation as the number of terms in the sum increases? (3 points)