MAST90105 Lab and Workshop 4 Solutions

1 Lab

In this lab., Mathematica will be used because it provides very good facilities for examining the shape of distributions as the numbers which define them change.

A Mathematica notebook called Lab4.nb will be the base for you to explore the examples from Lectures this week further. Please start by opening this notebook and from the `Evaluation` menu select the item `Evaluate Notebook`. This will cause all of the commands to be evaluated - this is needed to activate the commands that you need for your work in this lab.

1. In the section of the notebook labelled Skewness and Kurtosis

   a. use the sliders, or open the windows next to them with plus, in the graph to find the skewness and kurtosis recording them in the following table for the **Binomial Distribution**:

   Solution: Each cell has the skewness followed by the kurtosis:

	p=0.1	*p=0.3*	*p=0.5*	*p=0.7*	*p=0.9*
n=1	*2.7, 8.1*	*0.9, 1.8*	*0.0, 1.0*	*-0.9, 1.8*	*-2.7, 8.1*
n=10	*0.8, 3.5*	*0.3, 2.9*	*0.0, 2.8*	*-0.3, 2.9*	*-0.8, 3.5*
n=50	*0.4, 3.1*	*0.1,3.0*	*0.0,3.0*	*-0.1,3.0*	*-0.4, 3.1*
n=100	*0.3, 3.1*	*0.1, 3.0*	*0.0,3.0*	*-0.1,3.0*	*-0.3, 3.1*
n=250	*0.2, 3.0*	*0.1, 3.0*	*0.0,3.0*	*-0.1,3.0*	*-0.2, 3.1*
n=500	*0.1, 3.0*	*0.0, 3.0*	*0.0, 3.0*	*-0.0, 3.0*	*-0.1,3.0*

   b. Comment on the trends you find in your table, in particular across the rows and down the columns

   Solution: Going down the columns, ie as the sample size increases, the skewness gest closer to 0. This is saying that the shape of the pmf gets more symmetrical. Going down the columns, the kurtosis approaches the value 3. Going accross the rows, the skewness entries are anti-symmetrical about $p = 0.5$ (ie the absolute values are the same for $p = 0.3, 0.7$ but the signs are different). The skewness is 0 for $p = 0.5$. The kurtosis entries are symmetrical about $p = 0.5$. The kurtosis gets closest to 3.0 quicker for p closer to 0.5.

   c. Copy the Mathematica `Manipulate` command into the cell below the Skewness and Kurtosis graph. Alter the `Manipulate` command to show the Negative Binomial distribution. Note that Mathematica - as does R - uses the number of failures before the nth success, so the values are always $0, 1, \cdots$ rather than $n, n+1, n+2\cdots$. Note also that Mathematica uses n rather than r for the success number. Record the results in the **Negative Binomial** table:

Solution: Each cell has the skewness followed by the kurtosis:

	p=0.1	*p=0.3*	*p=0.5*	*p=0.7*	*p=0.9*
n=1	*2.0, 9.0*	*2.0, 9.1*	*2.1, 9.5*	*2.4, 10.6*	*3.5, 17.1*
n=10	*0.6, 3.6*	*0.6, 3.6*	*0.7, 3.7*	*0.8, 3.8*	*1.1, 4.4*
n=50	*0.3, 3.1*	*0.3, 3.1*	*0.3, 3.1*	*0.3, 3.2*	*0.5, 3.3*
n=100	*0.2, 3.1*	*0.2, 3.1*	*0.2, 3.1*	*0.2, 3.1*	*0.3, 3.1*
n=250	*0.1, 3.0*	*0.1, 3.0*	*0.1, 3.0*	*0.2, 3.0*	*0.2, 3.1*
n=500	*0.1, 3.0*	*0.1, 3.0*	*0.1, 3.0*	*0.1, 3.0*	*0.2, 3.0*

d. Comment on the trends you find in the table, in particular across the rows and down the columns

Solution: Going down the columns, ie as the sample size increases, the skewness gest closer to 0. This is saying that the shape of the pmf gets more symmetrical. Going down the columns, the kurtosis approaches the value 3. Going accross the rows, the skewness entries increase. The kurtosis entries get further away from 3.0. The rate of departure from 0 and 3 decreases as the sample size increases.

2. In the section of the Notebook labelled Distance between distributions for Sampling with and Without Replacement, `n` stands for the sample size, `p` for the probability of success (whatever that might be in the sampling context) and `t` for the population size, so that, using the Mathematica notation, .

a. Use the sliders, to take the sample size `n` to be 500, 1000 and 3000. Record the three distances between the distributions for each of the combinations in the table:

Solution: n=500. Each cell has the distance recorded:

	p=0.1	*p=0.3*	*p=0.5*	*p=0.7*	*p=0.9*
t=n+100	*0.41*	*0.41*	*0.41*	*0.41*	*0.41*
t=n+500	*0.17*	*0.17*	*0.17*	*0.17*	*0.17*
t=n+1,500	*0.07*	*0.07*	*0.07*	*0.07*	*0.07*
t=n+2,000	*0.05*	*0.05*	*0.05*	*0.05*	*0.05*
t=n+10,000	*0.01*	*0.01*	*0.01*	*0.01*	*0.01*
t=16,000,000	7.56×10^{-6}	7.56×10^{-6}	7.56×10^{-6}	7.56×10^{-6}	7.56×10^{-6}

n=1000:

	p=0.1	*p=0.3*	*p=0.5*	*p=0.7*	*p=0.9*
t=n+100	*0.52*	*0.452*	*0.52*	*0.52*	*0.52*
t=n+500	*0.26*	*0.26*	*0.26*	*0.26*	*0.26*
t=n+1,500	*0.12*	*0.12*	*0.12*	*0.12*	*0.12*
t=n+2,000	*0.1*	*0.1*	*0.1*	*0.1*	*0.1*
t=n+10,000	*0.02*	*0.02*	*0.02*	*0.02*	*0.02*
t=16,000,000	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*

n=3000:

	p=0.1	*p=0.3*	*p=0.5*	*p=0.7*	*p=0.9*
t=n+100	*0.68*	*0.68*	*0.68*	*0.68*	*0.68*
t=n+500	*0.44*	*0.44*	*0.44*	*0.44*	*0.44*
t=n+1,500	*0.26*	*0.26*	*0.26*	*0.26*	*0.26*
t=n+2,000	*0.22*	*0.22*	*0.22*	*0.22*	*0.22*
t=n+10,000	*0.06*	*0.06*	*0.06*	*0.06*	*0.06*
t=16,000,000	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*

b. Comment on the trends in the table, in particular across the rows and columns

Solution: The values for the maximum probability difference are constant across rows. They decrease down columns towards 0.0 for the 16 million population size, although the difference is shown as greater than 0 for sample size 500. The differences are greater for smaller population sizes for the larger sample sizes. Looking at the graphs, the larger sample sizes have more fine-grained pmf's

c. Looking at the Mathematica code, what is the purpose of defining the variables `lower` and `upper`? In particular, what is the significance of $n * p$ and $(n * p * (1 - p))^{0.5}$? (Hint: look at the definitions of `HypergeometricDistribution` and `BinomialDistribution` and then work out which entries from these vectors are plotted and which are used to compute the distance)

Solution: The variables are the lower and upper ends of the plotting interval. The significance of np *and* $(n * p * (1 - p))^{0.5}$ *is that they are the mean and standard deviation of the Binomial distribution. So the probabilities are plotted from the mean minus 5 standard deviations to the mean plus 5 standard deviations. Since the Hypergeometric probabilities are more peaked around the same mean, plotting all of the non-zero probabilities (to the resolution of the screen) for the Binomial will ensure the same for the Hypogeometric distribution.*

d. Why is the variable `total` introduced?

Solution: The variable ***`total`*** *is introduced so that the population size for the Hypogeometric distribution is always at least the sample size in the plots and distance calculations.*

e. (*challenging*) Can you relate the formula for `mpd` to what was claimed in lectures, namely the maximum difference between probabilities for the Binomial and the Hypergeometric? (hint: use the fact that probabilities add to one, and think about the set of numbers for which the Hypergeomtric probability is greater than the Binomial probability and vice-versa.)

Solution: The variable D was defined as

$$D = max_A|P(X \in A) - P(Y \in A)|$$

where X, Y are random variables with the Binomial and Hypergeometric distributions. Suppose $B = \{i : P(X = i) \geq P(Y = i)\}$. Then for any A with $P(X \in A) \geq P(Y \in A)$, rule (c) for probability gives:

$$\begin{aligned}|P(X \in A) - P(Y \in A)| &= P(X \in A) - P(Y \in A) \\ &= P(X \in A \cap B) - P(Y \in A \cap B) + P(X \in A \cap B^c) - P(Y \in A \cap B^c \\ &\leq P(X \in A \cap B) - P(Y \in A \cap B)\end{aligned}$$

since $P(X \in A \cap B^c) - P(Y \in A \cap B^c) = \sum_{i \in A \cap B^c} P(X = i) - Y(Y = i) \leq 0$. But

$$1 = P(X \in B) + P(X \in B^c) = P(Y \in B) + P(Y \in B^c)$$

so

$$P(X \in B) - P(Y \in B) = P(Y \in B^c) - P(X \in B^c) \qquad (1)$$

Thus we can apply a similar argument to show that for A with $P(X \in A) < P(Y \in A)$, we have $|P(X \in A) - P(Y \in A)| \leq P(Y \in B^c) - P(X \in B^c) = P(X \in B) - P(Y \in B)$. Thus B is a set for which $D = P(X \in B) - P(Y \in B)$. Finally, equation 1 also shows that $D = \frac{1}{2}\sum_i |P(X = i) - P(Y = i)|$.

3. **The objective of this question is to use Mathematica to look at the actual distance between Binomial and Poisson distributions.**

   a. The bound in Lectures said that the maximum difference between any Binomial and the corresponding Poisson probability, denoted D in the Mathematica plots, was bounded by p, the probability of success in the Binomial.

   b. Choosing a variety of values for p between 0 and 1, record the values of the maximum probability difference, D, for different values of n. Comment, particularly on whether the relationship with n is monotone.

Solution: Each cell has the distance recorded.

	n=1	*n= 10*	*n= 50*	*n= 100*	*n = 500*
p=0.0001	$1. \times 10^{-8}$	9.99×10^{-8}	4.96×10^{-7}	9.85×10^{-7}	4.64×10^{-6}
p=0.001	$1. \times 10^{-6}$	9.86×10^{-6}	4.64×10^{-5}	8.6×10^{-5}	2.28×10^{-4}
p=0.01	9.95×10^{-5}	8.68×10^{-4}	2.29×10^{-3}	2.78×10^{-3}	2.46×10^{-3}
p=0.1	9.52×10^{-3}	2.93×10^{-2}	2.6×10^{-2}	2.58×10^{-2}	2.56×10^{-2}
p = 0.3	7.78×10^{-2}	8.64×10^{-2}	8.57×10^{-2}	8.65×10^{-2}	8.61×10^{-2}
p = 0.5	1.97×10^{-1}	1.72×10^{-1}	1.67×10^{-1}	1.67×10^{-1}	1.66×10^{-1}
p = 0.9	5.34×10^{-1}	5.48×10^{-1}	5.1×10^{-1}	5.05×10^{-1}	5.04×10^{-1}

For fixed n, the distance increases with p. For fixed p, the relationship with n is monotone increasing for $p = 0.0001, 0.001, 0.01$ *but not for* $p = 0.1, 0.3, 0.5, 0.9$. *For larger p, the distance is approximately constant across different n and well away from 0.*

c. Find the maximum difference that you can between D and p and write it down. What do the plots show about the guidance that is given in many textbooks that "the Poisson approximation to Binomial can be used when n is large, p is small and np is moderate"?

Solution: The maximum difference on the graph occurs with $n = 1, p = 0.99$ *when the difference is 0.68. This can easily be seen by changing the definition of mpd in the Mathematica code to be* $p - D$. *The advice is excessively conservative. There is no need for n to be large and indeed for p small the quality of the approximation decreases as n increases. The quality depends mainly on p and is good for small p.*

d. Try to find n and p that minimise the ratio between p and D. Report on your findings.

Solution: This can be done by changing the definition of mpd in the Mathematica code to be $\frac{p}{D}$. *The ratio is minimised by* $p = 0.99, n = 100$ *at 1.2. For small values of p the approximation is good, but the ratio tends to be large for small n and then decrease to become relatively stable.*

2 Workshop

1. A warranty is written on a product worth \$10,000 so that the buyer is given \$8000 if it fails in the first year, \$6000 if it fails in the second, \$4000 if it fails in the third, \$2000 if it fails in the fourth, and zero after that. Its probability of failing in a year is 0.1; failures are independent of those of other years. What is the expected value of the warranty?

   - *Let X be such that the product fails at the X-th year. Let $Y = u(X)$ be the amount of money (value of the warranty) the buyer is given.*
   - *Then*

   $$Y = u(X) = \begin{cases} 8000, & x = 1 \\ 6000, & x = 2 \\ 4000, & x = 3 \\ 2000, & x = 4 \\ 0, & x \geq 5. \end{cases}$$

   - *It can be seen that the pmf of Y is*

y	*8000*	*6000*	*4000*	*2000*	*0*
$P(Y = y)$	*0.1*	$0.9 \cdot 0.1$	$0.9^2 \cdot 0.1$	$0.9^3 \cdot 0.1$	$1 - \{0.1 + 0.9 \cdot 0.1 + 0.9^2 \cdot 0.1 + 0.9^3 \cdot 0.1\}$

   - *So* $E(u(X)) = E(Y) = 8000(0.1) + 6000(0.9 \cdot 0.1) + 4000(0.9^2 \cdot 0.1)$ $+2000(0.9^3 \cdot 0.1) + 0 \cdot P(Y = 0) = 1809.8$.

2. Define the pmf and give the values of μ and σ^2 when the moment-generating function (mgf) of X is defined by

   a. $M(t) = \left(\frac{0.6e^t}{1-0.4e^t}\right)^2, \quad t < -\ln(0.4).$

      - *X has a negative binomial distribution $NB(2, 0.6)$.*
      - *Pmf* $f(x) = \binom{x-1}{1}0.6^2 0.4^{x-2}$, $x = 2, 3, \cdots$.
      - $\mu = \frac{r}{p} = \frac{2}{0.6} = \frac{10}{3}$ *and* $\sigma^2 = \frac{rq}{p^2} = \frac{2 \times 0.4}{0.6^2} = \frac{20}{9}$.

3. Let X equal the number of people selected at random that you must ask in order to find someone with the same birthday as yours. Assuming each day of the year is equally likely (and ignoring February 29).

   a. What probability distribution does X have? Namely, what is the pmf of X?

      - *X has a geometric distribution Geo*$(p = 1/365)$.

b. Give the mean and variance of X.

- $\mu = \frac{1}{p} = 365$ *and* $\sigma^2 = \frac{q}{p^2} = \frac{364/365}{1/365^2} = 364 \times 365 = 132860.$

c. Find $P(X > 400)$ and $P(X < 300)$.

- $P(X > 400) = q^{400} = \left(\frac{364}{365}\right)^{400} = 0.3337.$
- $P(X < 300) = 1 - P(X > 299) = 1 - q^{299} = 1 - \left(\frac{364}{365}\right)^{299} = 0.5597.$

4. Let X equal the number of flips of a fair coin that are required to observe the same face on consecutive flips. Find the pmf of X.

- $X = k$ $(k \geq 2)$ *if and only if we have one of the two outcomes:*

$$\underbrace{\cdots THTH}_{k-2}TT, \quad \underbrace{\cdots HTHT}_{k-2}HH$$

- *Assuming flips are independent, the probability of each of the two outcomes is* $\underbrace{\frac{1}{2}\cdot\cdots\cdot\frac{1}{2}}_{k-2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2^k}$ *and hence* $\Pr(X = k) = 2\cdot\frac{1}{2^k} = \frac{1}{2^{k-1}}$. $k = 2, 3, \ldots.$

5. Suppose that a basketball player can make a free throw 60% of the time. Let X equal the minimum number of free throws that this player must attempt to make a total of 10 shots,

a. What probability distribution does X have? Namely, what is the pmf of X?

- *X has a negative binomial distribution $NB(10, 0.6)$.*
- *Pmf* $f(x) = \binom{x-1}{9}0.6^{10}0.4^{x-10}$, $x = 10, 11, \cdots.$

b. Give the mean and variance of X.

- $\mu = E(X) = \frac{r}{p} = \frac{10}{0.6} = \frac{50}{3}.$
- $\sigma^2 = \mathrm{Var}(X) = \frac{rq}{p^2} = \frac{100}{9}.$

c. Find $P(X = 16)$.

- $P(X = 16) = \binom{15}{9}0.6^{10}0.4^{6} = 0.1240.$

6. Let X have a Poisson distribution with a variance of 3. Find $P(X = 2)$.

- *For Poisson random variable,* $\mu = \sigma^2 = \lambda.$
  So $P(X = 2) = \frac{\lambda^2 e^{-\lambda}}{2!} = \frac{3^2 e^{-3}}{2} = 4.5e^{-3} = 0.224.$

7. Flaws in a certain type of drapery material appear on the average of one in 150 square feet. If we assume the Poisson distribution, find the probability of at most one flaw in 225 square feet.

- *Let X be the number of flaws on a piece of drapery of 225 square feet. Then* $X \overset{d}{=} Poi(\lambda = 225/150).$

- $P(X \leq 1) = P(X = 0) + P(X = 1) = e^{-225/150} + \frac{225}{150}e^{-225/150} = 0.5578.$

8. Suppose that the probability of suffering a side effect from a certain flu vaccine is 0.005. Also suppose 1000 persons are inoculated.

   a. Find the exact probability that at most 1 person suffers using a binomial distribution.

      - $P(X \leq 1) = 0.995^{1000} + 1000 \times 0.005 \times 0.995^{999} = 0.040091.$

   b. Find approximately the probability that at most 1 person suffers using a Poisson distribution.

      - $P(X \leq 1) \approx e^{-1000\times 0.005} + (1000 \times 0.005)^1 \times e^{-1000\times 0.005} = 6e^{-5} = 0.040428.$

9. A hospital obtains 40% of its flu vaccine from Company A, 50% from Company B, and 10% from Company C. From past experience it is known that 3% of the vials from A are ineffective, 2% from B are ineffective, and 5% from C are ineffective. The hospital test 5 vials from each shipment. If at least one of the five is ineffective, find the conditional probability of that shipment coming from C.

   - *Let X be the number of ineffective vials in the sample of 5. Then,*
     if the shipment is from A, X has a binomial distribution, i.e. $X|A \overset{d}{=} b(5, 0.03)$;
     if the shipment is from B, X has a binomial distribution, i.e. $X|B \overset{d}{=} b(5, 0.02)$;
     if the shipment is from C, X has a binomial distribution, i.e. $X|C \overset{d}{=} b(5, 0.05)$.
   - $P(X \geq 1|A) = 1 - P(X = 0|A) = 1 - (1 - 0.03)^5 = 1 - 0.97^5$
     $P(X \geq 1|B) = 1 - P(X = 0|B) = 1 - (1 - 0.02)^5 = 1 - 0.98^5$
     $P(X \geq 1|C) = 1 - P(X = 0|C) = 1 - (1 - 0.05)^5 = 1 - 0.95^5$
   - *By Bayes's Theorem,*

$$\begin{aligned} P(C|\{X \geq 1\}) &= \frac{P(C)P(X \geq 1|C)}{P(A)P(X \geq 1|A) + P(B)P(X \geq 1|B) + P(C)P(X \geq 1|C)} \\ &= \frac{0.1(1 - 0.95^5)}{0.4(1 - 0.97^5) + 0.5(1 - 0.98^5) + 0.1(1 - 0.95^5)} = 0.1779. \end{aligned}$$

10. If X has a Poisson distribution so that $3P(X = 1) = P(X = 2)$, find $P(X = 4)$.

    - $3P(X = 1) = P(X = 2)$ *implies* $3\lambda e^{-\lambda} = \frac{\lambda^2}{2!}e^{-\lambda}$. *So* $\lambda = 6$.
      Accordingly $P(X = 4) = \frac{6^4}{4!}e^{-6} = 54e^{-6} = 0.1339.$

11. One of four different prizes was randomly put into each box of a cereal. If a family decided to buy this cereal until it obtained at least one of each of the four different prizes, what is the expected number of boxes of cereal that must be purchased?

- *Let T_1 be the number of purchases - 1 until the first different prize appears. Then the event that $[T_1 = i], i = 1, 2, \cdots$ is the event that the second, third, ... up to the ith prizes are all identical to the first one and the $i + 1$th is different. The probability of this is $\frac{1}{4}^{i-1}\frac{3}{4}$ so that T_1 is Geometric with success probability $\frac{3}{4}$. Let T_2 be the additional number of purchases after the first different prize has appeared until the second different box appears. Whatever the value of T_1 and whatever the identity of the first two different prizes, T_2 has a Geometric distribution with success probability $\frac{1}{2}$. Let T_3 be the additional number of purchases after the first two different prizes to the first one appear until the final prize appears. Whatever the values of T_1, T_2 and whatever the identity of the first three different boxes, T_3 has a Geometric distribution with success probability $\frac{1}{4}$. The number of boxes of cereal that must be purchased to get all 4 prizes is $T_1 + 1 + T_2 + T_3$. Hence, the expected number of boxes of cereal that must be purchased is $E(T_1 + 1 + T_2 + T_3) = E(T_1) + 1 + E(T_2) + E(T_3) = \frac{4}{3} + 1 + \frac{2}{1} + \frac{4}{1} = 8\frac{1}{3}$ since the expected value of a Geometric random variable is the reciprocal of the success probability.*