Arun Agarwal Applied Statistics & Data Science STAT 3503 HW 1

Problem 1 We have a simple random sample of articles (from FZ), in which we count the number of appearances of the word 'however'. We let X_i be the number of times the word 'however' appeared in the i^{th} article. Let $X_1, X_2, \ldots X_n$ be independently and identically distributed (iid) random variables. Let N be the population of interest and n be the sample population. Let $x_1, x_2, \ldots, x_n$ be our random sample of size 'n' taken from Poisson Distribution having parameter 'λ'. As given,

$$p(X_i = x_i|\lambda) = \text{Poisson}(x_i|\lambda) \text{ for } i = 1, \ldots, n.$$

Then, its probability mass function (pmf) is:

$$p(X_i = x_i|\lambda) = \begin{cases} e^{-\lambda}\lambda^x/x!, & \lambda \geq 0, \; p+e=1, \; x>0 \\ 0, & \text{else} \end{cases}$$

Now:

Q.1.1 By definition, the population of interest is the entire unit of people/things you consider in a study (labeled here as 'N'). Here, the population of interest is the articles of the popular Time Magazine contributor, FZ. As the name suggests, the population quantity of interest denotes the size of the population of interest, which for the sample population is called 'n'. A sampling unit refers to a singular value within a sample database that is being researched/modelled. Thus, the sampling units are the individual articles from FZ sampled, for which X_i represents the number of times the word 'however' appeared for that sampling unit (the i^{th} article).

Q.1.2 Estimands are quantities that are to be estimated in a statistical analysis (they are the target quantities). As is given in Q.1.3, the estimand is labeled as $\hat{\lambda}$, the average occurences of the word 'however' per document the unknown parameter we want to find to estimate our objective. However, this is the estimand given a Poisson distribution which does not need to be the case. Thus, one potentially useful estimand for studying writing style could be the probability of the word 'however' appearing (or maybe 'like', 'actually', 'lol', 'thus', etc.). Another potentially useful estimand for studying writing style is the average sentence length. One can, generally, examine word choice, punctuation, sentence structure, or other things as estimands for examining

Arun Agarwal Applied Statistics & Data Science HW1 Continued

writing style.

(Q1.3) As has been done in class and in notes, we can declare the technical nature of the quantities involved in the set-up/model above ($X_1, \ldots, X_n$; $x_1, \ldots, x_n$; λ; and n) using the following 2-by-2 table:

	Observed	Unobserved
Variable	Observed Random variables: $X_1, \ldots, X_n$	Latent Random variables
Constant	Known Constants: $x_1, \ldots, x_n, n$	Unknown Constants: λ

(what these quantities are can be seen at the beginning of my answer to this problem overall)

(Q1.4) The data generating process is a process in the real world that generates the data one is interested in. It describes the probabilistic process by which the data was generated and informs one on which quantities are variables and constants. As explicitly stated by the TA, pseudocode is requested to show the DGP. Thus,

for $i = 1, \ldots, n$ do:

| $X_i \overset{iid}{\sim} \text{Poisson}(x_i \mid \lambda)$

end

(Q1.5) A likelihood function is a function of the observed random variables, whatever they may be, given the constants that a researcher finds from the problem and model statements, or equivalently from the 2-by-2 table. Thus, likelihood = p(observed random variables | unknown constants);

As stated earlier, for a Poisson random variable X_i, the probability mass function is given by $P(X_i = x_i \mid \lambda) = f(x) = e^{-\lambda} \lambda^x / x!$, $x \in \{0, 1, \ldots, \infty\}$, $\lambda > 0$

Now, for the likelihood function $L(\lambda) = p(\cdot \mid \cdot)$, considering an IID sample from a Poisson variable: $L(\lambda) = p(X_i = x_i \mid \lambda)$; $L(\lambda) = \prod_{i=1}^{n} p(X_i = x_i \mid \lambda)$; $L(\lambda) = \left(\prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^x}{x!} \right)$, $\lambda > 0$, $x > 0$.

(Q1.6) Likelihood for a generic sample of n articles $(x_1, \ldots, x_n)$:

$$L(x_1, x_2, \ldots, x_n \mid \lambda_n) = p(X_1 = x_1 \mid \lambda)\, p(X_2 = x_2 \mid \lambda) \cdots p(X_n = x_n \mid \lambda)$$

$$L(x_1, x_2, \ldots, x_n \mid \lambda) = \prod_{i=1}^{n} p(X_i = x_i \mid \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^x}{x!}, \quad \lambda > 0, \; x > 0$$

$$L(x_1, x_2, \ldots, x_n \mid \lambda) = e^{-\lambda} \frac{\lambda^{x_1}}{x_1!} \cdots e^{-\lambda} \frac{\lambda^{x_n}}{x_n!} = e^{-n\lambda} \frac{\lambda^{(x_1 + x_2 + \cdots + x_n)}}{x_1!\, x_2! \cdots x_n!}$$

Arun Agarwal Applied Statistics & Data Science HW 1 Continued

$$L(\lambda|x_1,x_2,\dots,x_n) = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \quad \left(= e^{-n\lambda}\lambda^{\left(\sum_{i=1}^n x_i\right)}\prod_{i=1}^n \frac{1}{x_i!}\right)$$

Q1.7 Now, taking log-liklihood for a generic sample of 'n' articles $(x_1, x_2, \dots, x_n)$:

By applying the natural-log to the final expression in Q1.6, we get:

$$\ln L(\lambda|x_1,x_2,\dots,x_n) = (-n\lambda)\ln e + \left(\sum_{i=1}^n x_i\right)\ln\lambda - \ln\left(\prod_{i=1}^n x_i!\right)$$

$$\ln L(\lambda|x_1,x_2,\dots,x_n) = \ln(\lambda)\sum_{i=1}^n x_i - \ln\left(\prod_{i=1}^n x_i!\right) - n\lambda$$

We can also derive this answer by starting at our liklihood fun from 1.6:

$$L(\lambda|x_1,x_2,\dots,x_n) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^x}{x!}$$

$$\ln L(\lambda|x_1,x_2,\dots,x_n) = \sum_{i=1}^n \ln\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right)$$

$$\ln L(\lambda|x_1,x_2,\dots,x_n) = \sum_{i=1}^n \left[(-\lambda)\ln e + (x_i)\ln\lambda - \ln x_i!\right]$$

$$\boxed{\ln L(\lambda|x_1,x_2,\dots,x_n) = -n\lambda + \ln\lambda\sum_{i=1}^n x_i - \sum_{i=1}^n \ln(x_i!)}$$

NOTE: if we were to have taken the log of both sides instead of the natural log, we would not be able to simplify the expression as much and result with:

$$\log L(\lambda|x_1,x_2,\dots,x_n) = -n\lambda\log e + \log\lambda\sum_{i=1}^n x_i - \sum_{i=1}^n \log(x_i!)$$

NOTE: the TA said I could use the natural log version here instead of the log version ↓ If I had used the log, the answer would come out to 28.3005

Q.1.8 For the specific sample of 7 articles (12, 4, 5, 3, 7, 5, 6):

$n = 7$, $x_1 = 12$, $x_2 = 4$, $x_3 = 5$, $x_4 = 3$, $x_5 = 7$, $x_6 = 5$, $x_7 = 6$

$$\ln L\left(\lambda = \frac{\sum x_i}{n} \,\middle|\, x_1=12, x_2=4, x_3=5, x_4=3, x_5=7, x_6=5, x_7=6\right) =$$

$$\hookrightarrow -n\lambda + \ln\lambda\sum_{i=1}^n x_i - \sum_{i=1}^n \ln(x_i!)$$

We know that the mean and variance of poisson distribution are equal, by definition. In other words, $E(x_i) = \lambda$ and $V(x_i) = \lambda$. Then:

$$\sum_{i=1}^n x_i/n = \frac{12+4+5+3+7+5+6}{7} = \frac{42}{7} = 6 \quad \therefore \hat{\lambda} = 6, \text{ where } \hat{\lambda}$$

is an estimated value. Now:

$$\ln L(\hat{\lambda}|x_1=12, x_2=4, x_3=5, x_4=3, x_5=7, x_6=5, x_7=6) = -(7)(6) +$$

$$\hookrightarrow \ln(6)\sum_{i=1}^7 x_i - \sum_{i=1}^7 \ln(x_i!)$$

$$= -42 + (12+4+5+3+7+5+6)\ln 6 - [\ln 12! + \ln 4! + \ln 5! + \ln 3! + \ln 7! + \ln 5! + \ln 6!]$$

$$\approx -16.3825$$

If we were not to say $\hat{\lambda} = 6$, we would have

$$l(\lambda) = 42\ln\lambda - 7\lambda - 49.636$$

Q1.9 The maximum value of λ from the log-liklihood $l(\lambda)$ must follow the first order condition for a maximum ($\frac{d}{d\lambda} l(\lambda|x_1,x_2,\dots,x_n) = 0$). Thus:

$$\frac{d}{d\lambda} l(\lambda|x_1,x_2,\dots,x_n) = \frac{d}{d\lambda}\left(-n\lambda + \ln\lambda\sum_{i=1}^n x_i - \sum_{i=1}^n \ln(x_i!)\right) = -n + \frac{1}{\lambda}\sum_{i=1}^n x_i = 0 \text{ (to find the max)}$$

$$n = \frac{1}{\lambda}\sum_{i=1}^{n} x_i$$

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Thus, the maximum value of λ is just the sample mean of the n observations in the sample.

It should be noted that to verify that this is the maximum (and not the minimum), we should take the second derivative $\frac{d^2}{d^2\lambda}\ell(\lambda | x_1, x_2, \ldots, x_n)$ and verify it is less than 0. Thus:

$$\frac{d^2}{d^2\lambda}\ell(\lambda | x_1, x_2, \ldots, x_n) = \frac{d}{d\lambda}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \lambda\right)$$

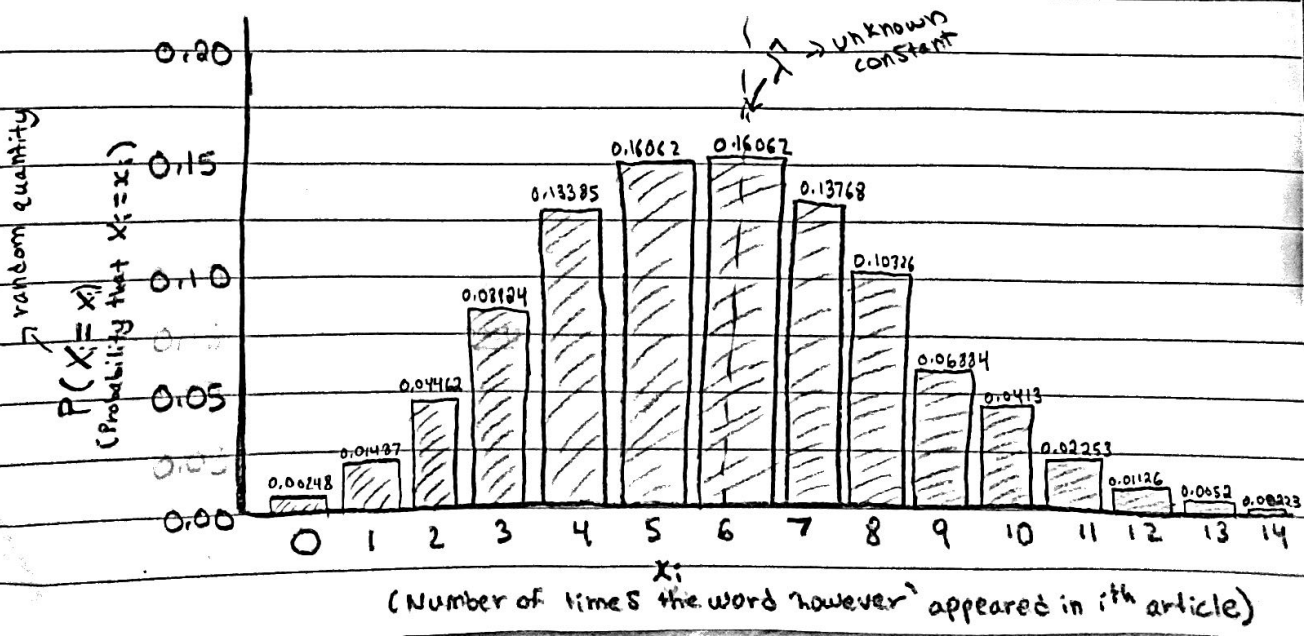
$$= -1 < 0 \quad \therefore \text{This is a maximum (not a minimum)}$$

$\therefore$ Likelihood function is maximized when $\hat{\lambda} = \sum_{i=1}^{n} x_i / n = \bar{x} = \text{sample mean}$

For our given sample of (1,2,4,5,3,7,5,6), we then expect the maximum value of λ to be $42/7 = 6$, the sample mean.

Next, the question asks to plot the log-likelihood $\ell(\lambda)$ in R for the given sample; however, I am an undergraduate non-STAT major with no experience coding in R. Thus, for this case, the Professor said this part of the question would not count towards my grade and would be extra credit. However, my coding attempt is attached in a separate file submission.

(1.10) As the TA stated, we can use a histogram as a graphical representation of this model. Thus, for $X \sim \text{Poisson}(\lambda)$, we have:



Arun Agarwal Applied Statistics & Data Science STAT 3503 HW1 cont.

Problem 1 Extra Credit We need to implicitly assume that the articles have the same length for a fair distribution and accurate calculation of λ. For example, with the given sample, (12, 9, 5, 3, 7, 5, 6), if $x_1 = 12$ has 2000 words but $x_4 = 3$ has 6000 words, we already get a completely new understanding of the data. There is no longer an independent and identically distributed (IID) Poisson distribution as each article of FZ can have different lengths, which directly affect the counts of words 'however'. With the given example, we now realize and must take into account the fact that x_1 has a higher count than x_4 but less words. If we were to truncate some of the articles to make it such that they all have the same length (as was implicitly assumed anyways), we would technically have a more accurate Poisson distribution and calculated lambda value (as the x_i values would change), but this would also cause a loss of ^data.^ In general though, we implicitly assume that the articles have the same length as this is necessary to model this situation as an IID Poisson Distribution. valuable Instead, we could use a ratio such as done in Problem 2

While not provided, I believe the implied common length to be 100 words.

Arun Agarwal Applied Statistics & Data Science STAT 3503 HW 1 Cont.

Problem 2) Similar to Problem 1, we have a simple random sample of articles (from FZ), in which we count the number of appearances of the word 'however'. We let X_i be the number of times the word 'however' appeared in the i^{th} article. Then, we let $X_1, X_2, \dots, X_n$ be independtly and identically distributed (IID) random variables. Let N be the population of interest and n be the sample population. Let $x_1, x_2, \dots, x_n$ be our random sample of size 'n' taken from Poisson Distribution having unknown parameter $v \cdot \frac{y_i}{1000}$. Here, $y_1, y_2, \dots, y_n$ represent the length of each article' (the number of words total). As given,

$$p(X_i = x_i | y_i, v, 1000) = \text{Poisson}(x_i | v \cdot \tfrac{y_i}{1000}) \text{ for } i = 1, 2, \dots, n$$

As stated by the TA, we can interpret the unknown parameter, $v \cdot \frac{y_i}{1000}$, as exactly equivalent to λ from problem 1. In other words, $\lambda = v \cdot \frac{y_i}{1000}$

Now, the probability mass function (pmf) is:

$$p(X_i = x_i | v \cdot \tfrac{y_i}{1000}) = \begin{cases} e^{(-v \cdot \frac{y_i}{1000})} (v \cdot \frac{y_i}{1000})^x / x!, & v \cdot \frac{y_i}{1000} \geq 0, \; p+q=1, \; x > 0 \\ 0, \text{ else} & \end{cases} \quad y_i > 0$$

Q2.1) As has been done in class and in notes, we can declare the technical nature of the quantities involved in the set-up/model above, $(X_1, \dots, X_n; x_1, \dots, x_n; y_1, \dots, y_n; v; n)$ using the following 2-by-2 table:

	Observed	Unobserved
Variable	Observed Random Variables: $X_1, \dots, X_n$	Latent Random Variables
Constant	Known Constants: $x_1, \dots, x_n$, $y_1, \dots, y_n$, n	Unknown Constants: v

(what these quantities are can be seen in the paragraph above)

Q2.2) As mentioned in the extra credit of problem 1, we had implicitly assumed that the articles have the same length for a fair distribution and accurate calculation of λ. However, as was noted, this was not actually a correct assumption and the articles FZ writes have different lengths. Then, there was not an independtly and identically distributed (IID) Poisson Distribution as each article of FZ had different lengths, directly affecting the word count of 'however.' Thus, we know have data for the length of each article, $y_1, \dots, y_n$, which we can and will use to have a more true IID Distribution. To take

the lengths into account, the unknown parameter, λ, was changed to be $v \cdot \frac{y_i}{1000}$, where $\frac{y_i}{1000}$ is the length of the article divided by 1000. This can be interpreted as a unit length (how many times the word 'however' appeared in a 1000 words of that article, i), which allows for an IID Poisson Distribution, as was implicitly assumed in Problem 1. Thus, it is the unit length for the appearance count.

(Q2.3) As stated by the TA, v can be interpreted exactly like the λ from Problem 1 except it is focusing on every 1000 words (to achieve the IID Poisson Distribution desired). If $\lambda = v \cdot \frac{y_i}{1000}$, then $v = \frac{\lambda \cdot 1000}{y_i}$, which again is interpreted as a density/mean of word count of 'however' on every 1000 words.

(Q2.4) The data generating process is a process in the real world that generates the data one is interested in. It describes the probabilistic process by which the data was generated, and informs one on which quantities are variables and constants. As explicitly stated by the TA, pseudocode is requested to show the DGP. Thus,

for $i = 1, \ldots, n$ do: ↷ λ
  $X_i \overset{iid}{\sim} \text{Poisson}(x_i \mid v \cdot \frac{y_i}{1000})$
end

(Q2.5) likelihood = p(observed random variables | unknown constants):
As stated earlier, the pmf here is $P(X_i = x_i \mid \lambda) = f(x) = e^{(-v \cdot \frac{y_i}{1000})} (v \cdot \frac{y_i}{1000})^{x_i} / x_i!$, $x \in \{0, 1, \ldots, \infty\}$, $\lambda > 0$, $y_i > 0$. Now, for the likelihood function $L(\lambda) = p(\cdot \mid \cdot)$, considering an IID sample from a Poisson variable,

$$L(\lambda) = p\left(X_i = x_i \mid v \cdot \frac{y_i}{1000}\right); \quad L(\lambda) = \prod_{i=1}^{n} P\left(X_i = x_i \mid v \cdot \frac{y_i}{1000}\right);$$

$$L(\lambda) = \prod_{i=1} e^{(-v \cdot \frac{y_i}{1000})} (v \cdot \frac{y_i}{1000})^{x} / x!, \quad v \cdot \frac{y_i}{1000} > 0, \; x > 0, \; y_i > 0$$

(Q2.6) Likelihood for a generic sample of n articles $(x_1, \ldots, x_n)$ and n article lengths $(y_1, \ldots, y_n)$:

$$L\left(x_1, x_2, \ldots, x_n \mid v \cdot \frac{y_i}{1000}\right) = p\left(X_1 = x_1 \mid v \cdot \frac{y_1}{1000}\right) \cdots p\left(X_n = x_n \mid v \cdot \frac{y_n}{1000}\right)$$

$$L(x_1, x_2, \ldots, x_n \mid v \cdot \frac{y_i}{1000}) = \prod_{i=1}^{n} p\left(X_i = x_i \mid v \cdot \frac{y_i}{1000}\right) = \prod_{i=1} \left(e^{(-v \cdot y_i/1000)} (v \cdot \frac{y_i}{1000})^{x} / x!\right)$$

$$L(x_1, x_2, \ldots, x_n \mid v \cdot \tfrac{y_i}{1000}) = e^{(-nv \cdot \prod_{i=1}^{n} \frac{y_i}{1000})} (v \cdot \prod_{i=1}^{n} \tfrac{y_i}{1000})^{x_1 + x_2 + \cdots + x_n} / x_1! x_2! \cdots x_n!$$

$$L(x_1, x_2, \ldots, x_n \mid v \cdot \tfrac{y_i}{1000}) = e^{(-nv \cdot \prod_{i=1}^{n} \frac{y_i}{1000})} \left((v \cdot \prod_{i=1}^{n} \tfrac{y_i}{1000})^{\sum_{i=1}^{n} x_i} / \prod_{i=1}^{n} x_i! \right)$$

$$L(x_1, x_2, \ldots, x_n \mid v \cdot \tfrac{y_i}{1000}) = e^{(\frac{-v}{1000} \cdot \sum_{i=1}^{n} y_i)} (v/1000)^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} y_i / \prod_{i=1}^{n} x_i!$$

Q2.7 Now, taking the log-likelihood for a generic sample of n articles, $(x_1, x_2, \ldots, x_n)$ we have:

$$\ln L(x_1, x_2, \ldots, x_n \mid v \cdot \tfrac{y_i}{1000}) = (-nv \cdot \prod_{i=1}^{n} \tfrac{y_i}{1000}) \ln e + \sum_{i=1}^{n} x_i \ln(v \cdot \prod_{i=1}^{n} \tfrac{y_i}{1000}) - \ln(\prod_{i=1}^{n} x_i!)$$

$$l(x_1, x_2, \ldots, x_n \mid v \cdot \tfrac{y_i}{1000}) = (-nv \cdot \prod_{i=1}^{n} \tfrac{y_i}{1000}) + \sum_{i=1}^{n} x_i \ln(v \cdot \prod_{i=1}^{n} \tfrac{y_i}{1000}) - \ln(\prod_{i=1}^{n} x_i!)$$

We can also derive this answer by starting at our likelihood function from Q2.6:

$$L(x_1, x_2, \ldots, x_n \mid v \cdot \tfrac{y_i}{1000}) = \prod_{i=1}^{n} e^{(-vy_i/1000)} (v \cdot \tfrac{y_i}{1000})^{x_i} / x_i!$$

$$\ln L(x_1, x_2, \ldots, x_n \mid v \cdot \tfrac{y_i}{1000}) = \sum_{i=1}^{n} \left[(\tfrac{-vy_i}{1000}) \ln e + (x_i) \ln(v \cdot \tfrac{y_i}{1000}) - \ln(x_i!) \right]$$

$$l(x_1, x_2, \ldots, x_n \mid v \cdot \tfrac{y_i}{1000}) = -nv \sum_{i=1}^{n} \tfrac{y_i}{1000} + \sum_{i=1}^{n} \left(x_i \ln(v \cdot \tfrac{y_i}{1000}) \right) - \sum_{i=1}^{n} \ln(x_i!)$$

$$l(x_1, x_2, \ldots, x_n \mid v \cdot \tfrac{y_i}{1000}) = \sum_{i=1}^{n} x_i \ln(\tfrac{v}{1000}) + \sum_{i=1}^{n} x_i \ln(y_i) - \tfrac{v}{1000} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(x_i!)$$

Q2.8 Given: $v = 10$, $y_1 = 1730$, $y_2 = 947$, $y_3 = 1830$, $y_4 = 1219$, $y_5 = 1100$

I will find the number of occurences of 'however' in each article:

Note: The TA said this was an optional problem because it requires a simulation in R. Since I could not figure out how to solve this question, I will not be doing it for extra credit.

Note: I think the numbers are $x_1 = 17$, $x_2 = 9$, $x_3 = 18$, $x_4 = 12$, $x_5 = 11$

y

$x_i = v \cdot \frac{y_i}{1000}$

Q2.9 Since I was unable to figure out how to generate the specific sample of occurences in the previous question, the majority of this question is unsolvable for me. However, substituting in what is known ($n = 5$, $v = 10$, $y_1 = 1730$, $y_2 = 947$, $y_3 = 1830$, $y_4 = 1210$, $y_5 = 1100$):

$$l(x_1, x_2, \ldots, x_n \mid 10 \cdot \tfrac{\sum y_i}{1000}) = -(5)(10) \sum_{i=1}^{5} \tfrac{y_i}{1000} + \sum_{i=1}^{n} x_i \ln(v \cdot \tfrac{y_i}{1000}) - \sum_{i=1}^{n} \ln(x!)$$

$$l(x_1, x_2, \ldots, x_n \mid 10 \cdot \tfrac{6817}{1000}) = -50 \left(\tfrac{6817}{1000} \right) + \sum_{i=1}^{n} x_i \, (81.3175) - \sum_{i=1}^{n} \ln(x!)$$

Note, If I substitute the x_i values written above, we get an answer of approximately -12

NOTE: This question is also considered extra credit since it relies on the last one (2.8) that needed the use of R.

Q2.10 The question asks to plot the log-likelihood $l(\lambda)$ in R for the given sample; while this is then an extra credit question, I have my coding attempt written in the separate file submission.

2.8 Continued

$$P(X_1 = x_1 \mid y_1, v, 1000) = e^{-17.3}(17.3)^{x_1}/x_1!$$

$$P(X_2 = x_2 \mid y_2, v, 1000) = e^{-9.47}(9.47)^{x_2}/x_2!$$

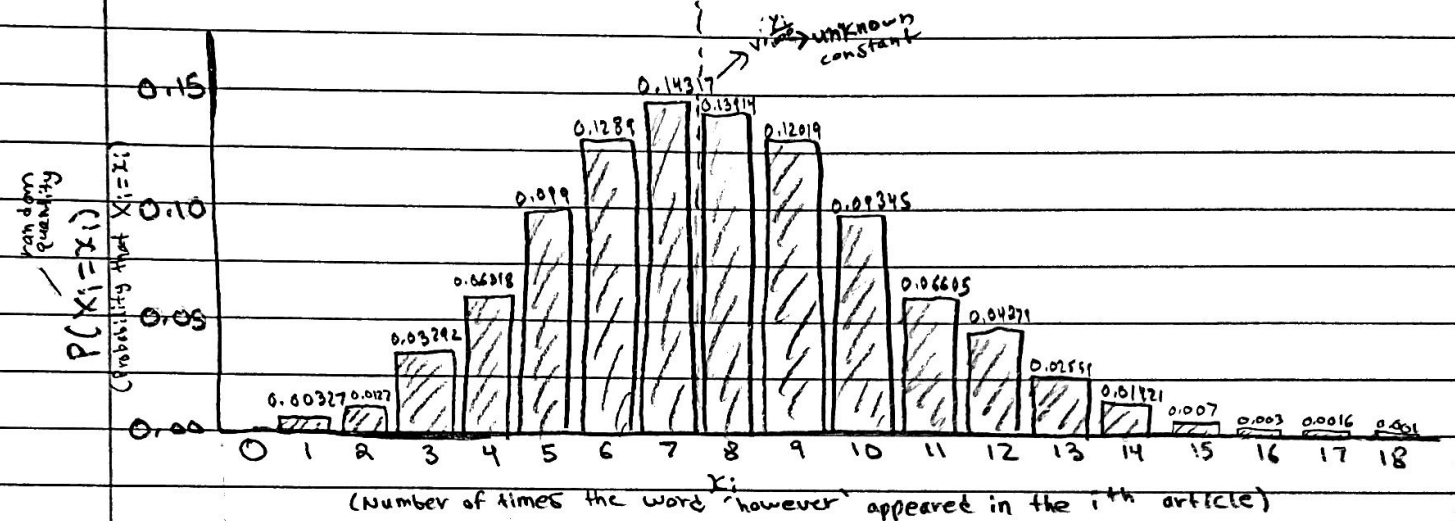
$$P(X_3 = x_3 \mid y_3, v, 1000) = e^{-18.3}(18.3)^{x_3}/x_3!$$

$$P(X_4 = x_4 \mid y_4, v, 1000) = e^{-12.1}(12.1)^{x_4}/x_4!$$

$$P(X_5 = x_5 \mid y_5, v, 1000) = e^{-11}(11)^{x_5}/x_5!$$

Arun Agarwal Applied Statistics & Data Science HW 1 Continued

(2.11) As the TA stated, we can use a histogram as a graphical representation of this model. Thus, for $X \sim \text{Poisson}(\lambda)$, we have:



(Problem 3) We have a simple random sample of articles (from FZ), in which we count the number of appearances of the word 'I'. We let X_i be the number of times the word 'I' appeared in the i^{th} article. Let N be the population of interest and n be the sample population. Let $x_1, x_2, \ldots, x_n$ be our random sample of size 'n'. We let Z_i indicate whether the i^{th} article is about politics, denoted by $Z_i = 1$, or not, denoted by $Z_i = 0$. We assume $X_1, \ldots, X_n$ are independent of one another conditionally on $Z_1, \ldots, Z_n$. Let $Z_1, \ldots, Z_n$ be independently and identically distributed (iid) random variables according to a Bernoulli Distribution with parameter π. As given,

$$p(Z_i \mid \pi) = \text{Bernoulli}(z_i \mid \pi) \quad \text{for } i = 1, \ldots, n$$

We further assume that the number of occurences of the word 'I' in an article follows a Poisson Distribution with unknown parameter $\lambda_{\text{Politics}}$. As given,

$$p(X_i = x_i \mid Z_i = 1, \lambda_{\text{Politics}}) = \text{Poisson}(x_i \mid \lambda_{\text{Politics}}) \quad \text{for } i = 1, \ldots, n$$

Finally, we assume that the number of occurences of the word 'I' in an article about any other topic follows a Binomial Distribution w/ size 1000 and unknown parameter θ_{other}. As given,

$$p(X_i = x_i \mid Z_i = 0, 1000, \theta_{\text{other}}) = \text{Binomial}(x_i \mid 1000, \theta_{\text{other}}) \quad \text{for } i = 1, \ldots, n$$

Arun Agarwal Applied Stat & Data Science HW 1 Continued

Q3.1 As has been done in class and in notes, we can declare the technical nature of the quantities involved in the set-up/model above ($X_1, \ldots, X_n, x_1, \ldots, x_n, Z_1, \ldots, Z_n, z_1, \ldots z_n, \pi, \lambda_{Politics}, \theta_{other}, n$):

	Observed	Unobserved
Variable	Observed Random Variables $Z_1, \ldots, Z_n$; $X_1, \ldots, X_n$	Latent Random Variables
Constant	Known Constants $z_1, \ldots, z_n$, n, $x_1, \ldots, x_n$,	Unknown Constants $\lambda_{Politics}$ θ_{other} π

(what these quantities are can be seen at the beginning of my answer to this problem overall)

Q3.2 As explicitly stated by the TA, pseudocode is requested to show the Data Generating Process. Thus:

```
for i = 1, ....., n do:
    if z_i = 1 : do:
        | X_i ~ Poisson(x_i | λ_Politics)
    end
    else:
        | X_i ~ Binomial(x_i | 1000, θ_other)
    end
```

Q3.3 This R simulation, while extra credit for me, was attempted, and my solution exists in the other file submission.

Q3.4 Likelihood for 1 article: $L_i(\lambda_{Politics}, \theta_{other}) = p(X_i = x_i | \lambda_{Politics}, \theta_{other})$

For 1 article, call it i, we have:

$L_i(\lambda_{Politics}, \theta_{other}) = P(X_i = x_i | Z_i = 0, 1000, \theta_{other}) P(Z_i = 0) + P(X_i = x_i | Z_i = 1, \lambda_{Politics}) P(Z_i = 1)$

$$L_i(\lambda_{Politics}, \theta_{other}) = \binom{1000}{x_i} \theta^{x_i} (1-\theta)^{1000-x_i} (1-\pi) + \frac{e^{-\lambda_{Politics}} \lambda_{Politics}^{(x_i)}}{x_i!} (\pi)$$

Q3.5 Likelihood for generic sample of n articles, $(x_1, \ldots, x_n)$:

$$L(x_1, x_2, \ldots, x_n; \lambda_{Politics}, \theta_{other}) = \prod_{i=1}^{n} \left[\binom{1000}{x_i} \theta^{x_i} (1-\theta)^{1000-x_i} (1-\pi) + \frac{e^{-\lambda_{Politics}} \lambda_{Politics}^{(x_i)}}{x_i!} (\pi) \right]$$

Arun Agarwal Applied Stat & Data Science HW1 Continued

Q3.6 log-likelihood for generic sample of n articles $(x_1, x_2, \ldots x_n)$:

Taking the natural log of the expression obtained in Q3.5:

$$\ln \mathcal{L}(x_1, x_2, \ldots, x_n; \lambda_{Politics}, \theta_{other}) = \ln \prod_{i=1}^{n} \left[\binom{1000}{x_i} \theta^{x_i} (1-\theta)^{1000-x_i} (1-\pi) + \frac{e^{-\lambda} \lambda^{(x_i)}}{x_i!} (\pi) \right]$$

$$\boxed{\ell(x_1, x_2, \ldots, x_n; \lambda_{Politics}, \theta_{other}) = \sum_{i=1}^{n} \ln \left[\binom{1000}{x_i} \theta^{x_i} (1-\theta)^{1000-x_i} (1-\pi) + \frac{e^{-\lambda} \lambda^{(x_i)}}{x_i!} (\pi) \right]}$$

NOTE: $\lambda_{Politics} = \lambda$, $\theta_{other} = \theta$

Q3.7 writing the log-likelihood for the specific sample of 8 articles (12, 4, 8, 3, 3, 10, 1, 9):

EXTRA CREDIT

$n = 8$, $x_1 = 12$, $x_2 = 4$, $x_3 = 8$, $x_4 = 3$, $x_5 = 3$, $x_6 = 10$, $x_7 = 1$, $x_8 = 9$

We know that the mean and variance of poisson distribution are equal, by definition. In other words, $E(x_i) = \lambda$ and $V(x_i) = \lambda$. Then:

$$\sum_{i=1}^{8} x_i / n = 12+4+8+3+3+10+1+9/8 = 6.25 \quad \therefore \hat{\lambda} = 6.25$$

Next, for a Binomial Distribution, it is known that $E(x_i) = n\theta$. Thus,

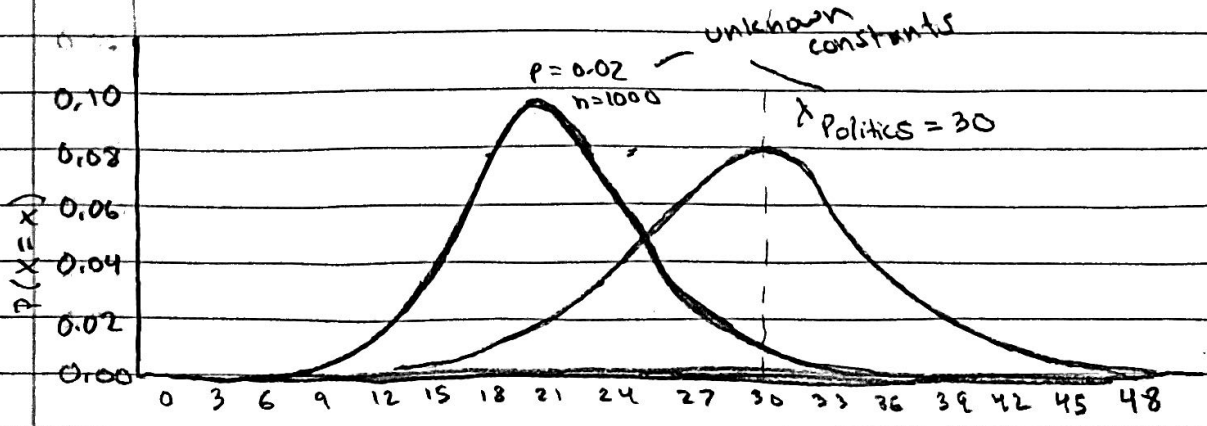
$$\sum_{i=1}^{8} x_i / n = 6.25 = 8\theta \Rightarrow \hat{\theta} = 0.78125$$

Now, we have $\hat{\theta}$ and $\hat{\lambda}$ and find the log-likelihood for the specific sample. Unfortunately, to find the log-likelihood, we also need π, and while $E(z_i) = \pi$, we have no information on the individual articles/samples (whether they are politics [1] or other [0]). Thus, I will choose a value of $p_i = 0.3$ to match what was given earlier:

$\ell(x_1, x_2, \ldots, x_n; \lambda_{Politics}, \theta_{other}) \simeq (-5.45) + (-3.301) + (-3.39792) + 2(-3.74798) + (-4.23257) + (-5.62139) + (-3.76256) = \boxed{-33.2614}$ → for $\pi = 0.3$

Also an extra credit question!

Q3.8 The TA said to draw two density curves, one for Poisson and one for Binomial:



Arun Agarwal Applied Stat & Data Science HW 1 Continued

(3 extra Credit) We always assume λ is an unknown constant, but it could be a variable, as is part of Problem 4. If we assume λ is not an unknown constant, then we must assume assume a distribution to it (it will have stochastic variation to it). Thus, it is not reasonable to always assume the rate λ is an unknown constant in all of our models.

Problem 4 We again have a random sample of articles (from F2) and count the occurrences of the word 'and'. We let x_i be the number of times the word 'end' appeared in the i^{th} article. Let $X_1, X_2, \ldots, X_n$ be IID random variables. Let N be the population of interest and n the sample population. Let $x_1, \ldots, x_n$ be our random sample of size 'n' taken from a Poisson Distribution having unknown parameter Λ. As given,

$$p(X_i = x_i \mid \Lambda = \lambda_i) = \text{Poisson}(x_i \mid \lambda_i) \text{ for } i = 1, \ldots, n$$

Assuming the rate Λ is distributed by Gamma distribution w/ parameters α and θ, we have:

$$f(\Lambda = \lambda_i \mid \alpha, \theta) = \text{Gamma}(\lambda_i \mid \alpha, \theta).$$

Q4.1 2-by-2 table:

	Observed	Unobserved
Variable	Observed Random variables $x_1, \ldots, x_n$	Latent Random variables $\lambda_1, \ldots, \lambda_n$
Constant	Known constants $x_1, \ldots, x_n$, n	Unknown constants α, θ

(what these quantities are can be seen above)

Q4.2 Pseudocode of Data Generating Process:

for $i = 1, \ldots, n$ do:

$\lambda_i \sim \text{Gamma}(\lambda_i \mid \alpha, \theta)$

$X_i \sim \text{Poisson}(x_i \mid \lambda_i)$

end

Arun Agarwal Applied Stat & Data science HW1 Cont.

Q4.3 This is an extra credit question that I decided not to do.

Q4.4 This is an extra credit question that I decided not to do

Q4.5 Writing the likelihood for 1 article, $L_i(\alpha,\Theta) = p(X_i = x_i | \alpha, \Theta)$: Using the total probability concept: $= \sum_j P(X_i = x_i | \lambda_j, \alpha, \Theta) P(\lambda_j)$. Now, using Poisson Distribution given by $P(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$, and since λ is distributed according to Gamma given by $f(x) = \frac{1}{\Gamma(\alpha)} \frac{1}{\theta^\alpha} x^{\alpha-1} e^{(-\frac{x}{\theta})}$, $x>0$, so $P(\lambda_j) = \frac{1}{\Gamma(\alpha)} \frac{1}{\theta^\alpha} \lambda_j^{\alpha-1} e^{(-\lambda_j/\theta)}$ → Now we have:

$$L_i(\alpha,\Theta) = p(X_i = x_i|\alpha,\Theta) = \sum_j \frac{\lambda_j^{x_i} e^{(-\lambda_j)}}{x_i!} \cdot \frac{1}{\Gamma(\alpha)} \cdot \frac{1}{\theta^\alpha} \lambda_j^{\alpha-1} e^{(-\lambda_j/\theta)}$$

Q4.6 Now, for a generic sample of n articles, $(x_1, x_2, \dots, x_n)$, we have:

$$L_i(\alpha|\Theta) = p(X_1 = x_1|\alpha,\Theta)\, p(X_2 = x_2|\alpha,\Theta) \cdots p(X_n = x_n|\alpha,\Theta)$$

$$L_i(\alpha|\Theta) = \prod_{i=1}^{n} p(X_i = x_i|\alpha,\Theta)$$

$$L_i(\alpha|\Theta) = \prod_{i=1}^{n} \sum_j \frac{\lambda_j^{x_i} e^{(-\lambda_j)}}{x_i!} \cdot \frac{1}{\Gamma(\alpha)} \cdot \frac{1}{\theta^\alpha} \lambda_j^{\alpha-1} e^{(-\lambda_j/\theta)}, \quad \lambda \in (0,\infty)$$

Then, the log-likelihood, $\ell(\alpha,\Theta)$ is:

$$\ln L = \ln \prod_{i=1}^{n} p(X_i = x_i|\alpha,\Theta) = \sum_{i=1}^{n} \ln p(X_i = x_i|\alpha,\Theta)$$

$$\ell(\alpha,\Theta) = \sum_{i=1}^{n} \ln\left[\sum_j \frac{\lambda_j^{x_i} e^{(-\lambda_j)}}{x_i!} \cdot \frac{1}{\Gamma(\alpha)} \cdot \frac{1}{\theta^\alpha} \lambda_j^{\alpha-1} e^{(-\lambda_j/\theta)}\right]$$

$$\ell(\alpha,\Theta) = \sum_{i=1}^{n}\left[-\ln(\Gamma(\alpha)) - \alpha\ln(\Theta) + \alpha - 1(\ln(\lambda_j)) - \frac{\lambda_j}{\theta}\left[\prod_{i=1}^{n} \ln\left(\sum_j \frac{\lambda_j^{x_i} e^{-\lambda_j}}{x_i!}\right)\right]\right]$$

Q4.7 Writing the log-likelihood, $\ell(\alpha,\Theta)$ for the specific sample of n=8 articles: (64, 61, 89, 55, 57, 76, 47, 55):

$n=8$, $x_1=64$, $x_2=61$, $x_3=89$, $x_4=55$, $x_5=57$, $x_6=76$, $x_7=47$, $x_8=55$:

$$\ln L = \sum_{i=1}^{n} \ln\left[\sum_j \frac{\lambda_j^{x_i} e^{(-\lambda_j)}}{x_i!} \cdot \frac{1}{\Gamma(\alpha)} \cdot \frac{1}{\theta^\alpha} \lambda_j^{\alpha-1} e^{(-\lambda_j/\theta)}\right] =$$

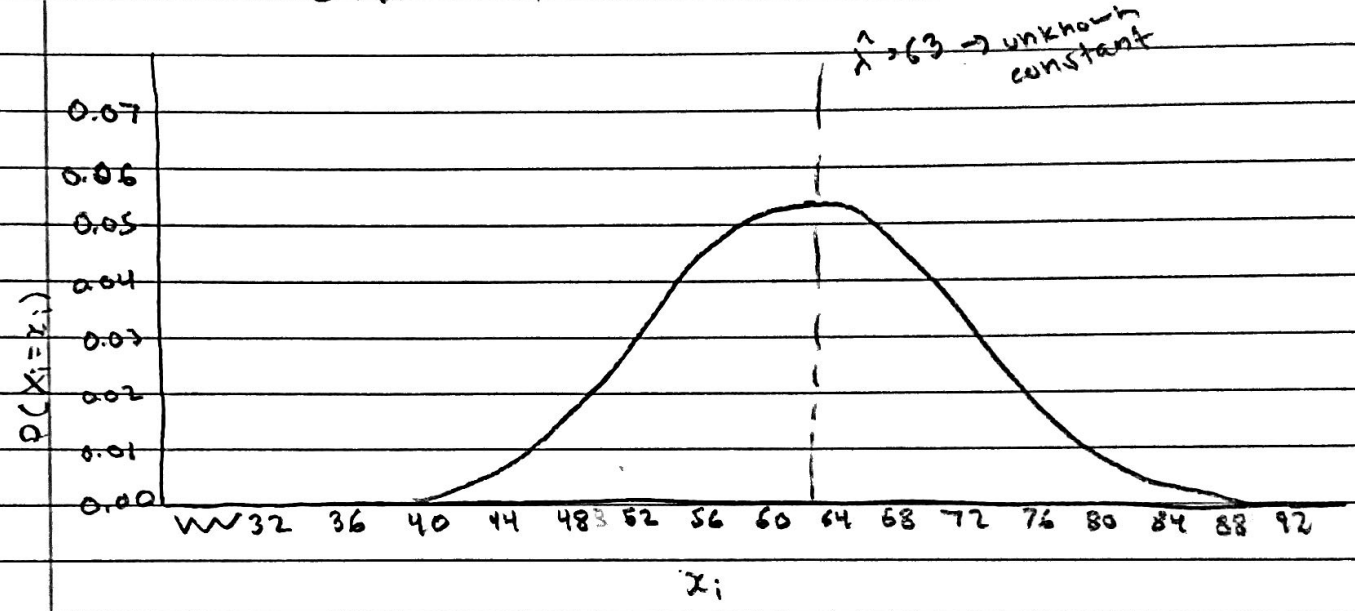
$$\ln\left[\sum_j \frac{\lambda_j^{64} e^{(-\lambda_j)}}{64!} \cdot \frac{1}{\Gamma(\alpha)} \cdot \frac{1}{\theta^\alpha} \lambda_j^{\alpha-1} e^{(-\lambda_j/\theta)}\right] + \dots + \ln\left[\sum_j \frac{\lambda_j^{55} e^{-\lambda_j}}{55!} \cdot \frac{1}{\Gamma(\alpha)} \cdot \frac{1}{\theta^\alpha} \lambda_j^{\alpha-1} e^{\left(\frac{-\lambda_j}{\theta}\right)}\right]$$

$$= 4\ln(\Gamma(\alpha)) - \ln(\theta)\alpha + \ln(\lambda_j)(\alpha-1) - \frac{\lambda_j}{\theta}\left[\prod_{i=1}^{n} \ln\left(\sum_j \frac{\lambda_j^{x_i} e^{-\lambda_j}}{x_i!}\right)\right]$$

64, 61, 89, 55, 57, 76, 47, 55

Arun Agarwal Applied Statistics & Data Science HW1 Continued

Q4.8 This is an **extra credit** question because it requires R to determine λ. If $\hat{\lambda}$ is 63, then we have:



Problem 4 Extra Credit (edo says) The very special probability mass function obtained for $p(X_i = x_i | \alpha, \Omega) = L_i(\alpha, \Omega)$ is Gaussian. Gamma mixture of Poisson is a Gaussian Distribution.
I will not be doing the second (final) extra credit given.