problem set no. 1

learning objectives. compute likelihoods, both for a generic sample, i.e., $(x_1, ..., x_n)$, and for a specific sample, i.e., $(2, 3, 6, 4, 8, 5, 6, 2, 3, 6, 5)$; write some short programs to generate fake data sets from a given model and plot the corresponding likelihoods.

problem 1. set-up: you are interested in studying the writing style of a popular Time Magazine contributor, FZ. you collect a simple random sample of his articles and count how many times he uses the word `however` in each of the articles in your sample, $(x_1, ..., x_n)$. In this set-up, x_i is the number of times the word `however` appeared in the i-th article.

question 1.1. (10 points) define the population of interest, the population quantity of interest (the thing you're interested in the population), and the sampling units.

question 1.2. (10 points) what are potentially useful estimands for studying writing style? (hint: you are interested in comparing FZ writing style to that of other contributors.)

question 1.3. (10 points) model: let X_i denote the quantity that captures the number of times the word `however` appears in the i-th article. let's assume that the quantities $X_1, ...X_n$ are independent and identically distributed (IID) according to a Poisson distribution with unknown parameter λ.

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

using the 2-by-2 table of what's variable/constant versus what's observed/unknown, declare what's the technical nature (random variable, latent variable, known constant or unknown constant) of the quantities involved the set-up/model above: $X_1, ..X_n$, $x_1, ...x_n$, λ and n.

question 1.4. (10 points) write the data generating process for the model above.

question 1.5. (10 points) define the likelihood $L(\lambda) = p(\cdot \mid \cdot)$ for this model as a function of $p(\cdot \mid \cdot)$.

question 1.6. (10 points) write the likelihood $L(\lambda)$ for a generic sample of n articles, $(x_1, ..., x_n)$.

question 1.7. (10 points) write the log-likelihood $\ell(\lambda)$ for a generic sample of n articles, $(x_1, ..., x_n)$.

question 1.8. (10 points) write the log-likelihood $\ell(\lambda)$ for the following specific sample of 7 articles $(12, 4, 5, 3, 7, 5, 6)$ (you could use ... to abbreviate it).

question 1.9. (10 points) plot the log-likelihood $\ell(\lambda)$ (on a computer) for the same specific sample of 7 articles $(12, 4, 5, 3, 7, 5, 6)$. What is the maximum value of λ (approximately)?

question 1.10. (10 points) draw a graphical representation of this model, which explicitly shows the random quantities and the unknown constants only.

Extra credit mmmh ... something is amiss. the articles FZ writes have different lengths. if we model the word occurrences in each article as IID Poisson random variables with rate λ, we are implicitly assuming that the articles have the same length. why? *(10 points; extra credit)* and if that is true, what is the implied common length? *(10 points; extra credit)*

problem 2. set-up: you collect another random sample of articles penned by FZ and count how many times he uses the word `however` in each of the articles in your sample, $(x_1, ..., x_n)$. you also count the length of each article in your sample, $(y_1, ..., y_n)$. In this set-up, x_i is the number of times the word `however` appeared in the i-th article, as before, and y_i is the total number of words in the i-th article.

question 2.1. (10 points) model: let X_i denote the quantity that captures the number of times the word `however` appears in the i-th article. let's assume that the quantities $X_1, ...X_n$ are independent and identically distributed (IID) according to a Poisson distribution with unknown parameter $\nu \cdot \frac{y_i}{1000}$:

$$p(X_i = x_i \mid y_i, \nu, 1000) = \text{Poisson}(x_i \mid \nu \cdot \frac{y_i}{1000}) \quad \text{for} \quad i = 1, ..., n.$$

using the 2-by-2 table of what's variable/constant versus what's observed/unknown, declare what's the technical nature (random variable, latent variable, known constant or unknown constant) of the quantities involved the set-up/model above: $X_1, ..X_n$, $x_1, ...x_n$, $y_1, ...y_n$, ν and n.

question 2.2. (10 points) what is the interpretation of $\frac{y_i}{1000}$ in this model? explain.

question 2.3. (10 points) what is the interpretation of ν in this model? explain.

question 2.4. (10 points) write the data generating process for the model above.

question 2.5. (10 points) define the likelihood $L(\nu) = p(\cdot \mid \cdot)$ for this model as a function of $p(\cdot \mid \cdot)$.

question 2.6. (10 points) write the likelihood $L(\nu)$ for a generic sample of n articles, $(x_1, ..., x_n)$, and n article lengths, $(y_1, ..., y_n)$.

question 2.7. (10 points) write the log-likelihood $\ell(\nu)$ for a generic sample of n articles, $(x_1, ..., x_n)$, and n article lengths, $(y_1, ..., y_n)$.

question 2.8. (10 points) Simulate the number of occurrences of the word `however` for 5 articles using the data generating process. Assume $\nu = 10$ and corresponding article lengths $y = (1730, 947, 1830, 1210, 1100)$. Record the number of occurrences of `however` in each article.

question 2.9. (10 points) write the log-likelihood $\ell(\nu)$ for the following the specific sample of occurrences you generated in the previous question and their corresponding 5 article lengths $(1730, 947, 1830, 1210, 1100)$ (you can use ... to abbreviate).

question 2.10. (10 points) Plot the log-likelihood from the previous question (on a computer). Does the maximum occur near $\nu = 10$?

question 2.11. (10 points) draw a graphical representation of this model, which explicitly shows the random quantities and the unknown constants only.

OK, that was a more reasonable model. but FZ writes about different topics. our model is not capturing that. is FZ more prone to offering his own opinions when he writes about politics than when he writes about other topics? let's investigate.

problem 3. set-up. you collect a random sample of articles penned by FZ and count how many times he uses the certain word `I` in each of the articles in your sample, $(x_1, ..., x_n)$. In this set-up, x_i is the number of times the word `I` appeared in the i-th article.

question 3.1. (10 points) model: let X_i denote the quantity that captures the number of times the word `I` appears in the i-th article. let Z_i indicate whether the i-th article is about politics, denoted by $Z_i = 1$, or not, denoted by $Z_i = 0$. let's assume that the quantities $X_1, ..., X_n$ are independent of one another conditionally on the corresponding values of $Z_1, ..., Z_n$. let's assume that the quantities $Z_1, ..., Z_n$ are independent and identically distributed (IID) according to a Bernoulli distribution with parameter π,

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

let's further assume that the number of occurrences of the word `I` in an article about politics follows a Poisson distribution with unknown parameter $\lambda_{Politics}$,

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

and that the number of occurrences of the word `I` in an article about any other topic follows a Binomial distribution with size 1000 and unknown parameter θ_{Other},

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

using the 2-by-2 table of what's variable/constant versus what's observed/unknown, declare what's the technical nature (random variable, latent variable, known constant or unknown constant) of the quantities involved the set-up/model above: $X_1, ..X_n$, $x_1, ...x_n$, $Z_1, ..Z_n$, $z_1, ...z_n$, $\pi, \lambda_{Politics}$, θ_{Other} and n.

question 3.2. (10 points) write the data generating process for the model above.

question 3.3. (10 points) simulate 1000 values of X_i in R from the data generating process assuming $pi = 0.3$, $\lambda_{Politics} = 30$ and $\theta_{Other} = 0.02$. Plot the values of $X_i|Z_i = 1$ and $X_i|Z_i = 0$ as two histograms on the same plot. Color the histograms by the value of Z_i so the two populations can be distinguished.

question 3.4. (10 points) write the likelihood for 1 article, $L_i(\lambda_{Politics}, \theta_{Other}) = p(X_i = x_i \mid \lambda_{Politics}, \theta_{Other})$.

question 3.5. (10 points) write the likelihood $L(\lambda_{Politics}, \theta_{Other})$ for a generic sample of n articles, $(x_1, ..., x_n)$.

question 3.6. (10 points) write the log-likelihood $\ell(\lambda_{Politics}, \theta_{Other})$ for a generic sample of n articles, $(x_1, ..., x_n)$.

question 3.7. (10 points) write the log-likelihood $\ell(\lambda_{Politics}, \theta_{Other})$ for the following specific sample of 8 articles $(12, 4, 8, 3, 3, 10, 1, 9)$.

question 3.8. (10 points) draw a graphical representation of this model, which explicitly shows the random quantities and the unknown constants only.

Extra credit wait, but is it reasonable to assume that the rate λ is an unknown constant in all of our models? it seems like a stretch. *(10 points; if you agree)*

problem 4. This one is for 8109 ONLY!! set-up: let's go back to the simplest possible set-up for this exercise. you collect a random sample of articles penned by FZ and count how many times he uses the word `and` in each of the articles in your sample, $(x_1, ..., x_n)$. In this set-up, x_i is the number of times the word `and` appeared in the i-th article, as before.

question 4.1. (10 points) model: let X_i denote the quantity that captures the number of times the word `and` appears in the i-th article. let's assume that the quantities $X_1, ...X_n$ are independent and identically distributed (IID) according to a Poisson distribution with unknown parameter Λ,

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

in addition, let's assume that the rate Λ is distributed according to a Gamma distribution with unknown parameters α and θ,

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

using the 2-by-2 table of what's variable/constant versus what's observed/unknown, declare what's the technical nature (random variable, latent variable, known constant or unknown constant) of the quantities involved the set-up/model above: $X_1, ..X_n$, $x_1, ...x_n$, Λ, $\lambda_1, ...\lambda_n$, α, θ and n.

question 4.2. (10 points) write the data generating process for the model above.

question 4.3. (10 points) simulate 1000 values from the data generating process. Assume $\alpha = 10$ and $\theta = 1$. Compute the mean and variance of the X_i.

question 4.4. (10 points) simulate 1000 values assuming $\lambda_i = 10$ for all i (ignore the Gamma distribution). Compute the mean and variance of the X_i now. How do they compare to the mean and variance you calculated in question 4.3?

question 4.5. (10 points) write the likelihood for 1 article, $L_i(\alpha, \theta) = p(X_i = x_i \mid \alpha, \theta)$.

question 4.6. (10 points) write the log-likelihood $\ell(\alpha, \theta)$ for a generic sample of n articles, $(x_1, ..., x_n)$.

question 4.7. (10 points) write the log-likelihood $\ell(\alpha, \theta)$ for the following specific sample of 8 articles $(64, 61, 89, 55, 57, 76, 47, 55)$.

question 4.8. (10 points) draw a graphical representation of this model, which explicitly shows the random quantities and the unknown constants only.

Extra credit do you recognize the very special probability mass function you just obtained for $p(X_i = x_i \mid \alpha, \theta) = L_i(\alpha, \theta)$? *(10 points; extra credit)* excellent! you just proved a useful result: Gamma mixture of Poisson is a ... .

Generate samples from this distribution and verify graphically that you get the distribution looks the same as that in 4.3 (you must use appropriate parameters you identified above). *(10 points; extra credit)*