1 Marjorie is working on a software project, and wants to figure out whether her ability to fix bugs is improving. She knows that a few months ago she had a probability of 0.05 of fixing any bug that came her way without having to ask for help. She decides to perform a quick experiment to see whether she should believe that she has improved. So she counts the number of bugs that she has to ask for help with before she solves one on her own.

It turns out that the first bug she manages to fix without asking for help is the 3rd bug that comes her way. Use this observation to help Marjorie conduct a hypothesis test.

(a) Set up the Hypothesis test for this situation. That is, define an appropriate random variable X, state the distribution of X and state the Null Hypothesis as well as the alternative hypothesis. Use a 2-sided test.

Solution: We can think of what Marjorie is counting as the number of failures she has before her first success, thus it is obvious that X is geometrically distributedand the probability of success in each attempt is what she wants to know about. In particular she wants to know whether she has reason to believe whether she has reason to believe that her probability of success, p, has changed from the previous value of 0.05. SO:

$$X \sim \text{Geo}(p)$$
$$\text{H}_0 : p = 0.05$$
$$\text{H}_1 : p \neq 0.05$$

[3 marks]

(b) State the observation x that Marjorie has made of the Random variable X.

Solution: Since the first bug she manages to fix without asking for help is the third one she fails to fix the first two on her own. So her observation in $x = 2$.

[1 marks]

(c) Calculate the p-value for the observation that Marjorie has made of the Random Variable X.

Solution: The $p-value$ gives the probability to observe something as weird as what we observed, or weirder. In this case it is clear (for example by considering the expected value of the random variable) that the observation is surprisingly small. Therefore we need to sum up all the probabilities that are less than or equal to the observation. We also need to multiply this number by 2 to account for the equal probability in the upper tail (since we are doing a two-sided test). So:

$$\begin{aligned} p-\text{value} &= 2 \times \mathbb{P}(X \leq 2) \\ &= 2 \times (\mathbb{P}(X=0) + \mathbb{P}(X=1) + \mathbb{P}(X=2)) \\ &= 2 \times ((1-0.05)^0 0.05 + (1-0.05)^1 0.05 + (1-0.05)^2 0.05) \\ &= 2 \times (0.05 + 0.950.05 + 0.95^2 0.05) \\ &= 2 \times (0.142625...) \\ &\approx 0.29 \end{aligned}$$

[3 marks]

(d) State whether Marjorie has reason to believe that her programming skills have improved based on the p-value you have calculated.

Solution: A p-value of 0.29 is large, so despite how far the observation falls from the expected value of the function, it actually doesn't provide any evidence at all against H_0.

So, Marjorie has no reason to believe that her programming skills have improved.

Comment: This is a slightly counterintuitive property of the geometric distribution which is worth giving some attention; for a geometrically distributed random variable X the probability of $X = 0$ is always larger than any other outcome, regardless of the size of how small the probability of success is. In fact, in this case, we could have concluded that the p-value would be greater than 0.05 even before we started doing any calculations.

[1 mark]

2 Bill, an aspiring social media influencer, aims to explore the relationship between the number of posts he creates in an hour and the engagement those posts receive during that same hour. To achieve this, he plans to use a linear regression model to estimate the parameter β

such that $\mathbb{E}[Y_i] = \beta x_i$, where Y_i represents the number of unique engagements on the posts he makes during hour i, and x_i is the number of posts he makes in that hour.

Assume that the number of engagement that occurs during each hour is independent, conditional on the number of posts that Bill makes.

Clearly explain and justify all your calculations.

(a) Identify the suitable probability distribution to model this scenario. *Hint:* Use one of the models covered in the course. Keep in mind that the random variable Y_i counts events with no predefined upper limit.

Solution: We see that of the named distributions the one that best fits this situaotion is the POisson Random variable, so

$$Y_i \sim \text{Poisson}(\beta x_i).$$

[1 marks]

(b) Find the likelihood function and show that it can be expressed as:

$$L(\beta, y_1, y_2, \ldots, y_n) = K\beta^{n\overline{y}}e^{-n\overline{x}\beta}$$

where K is a constant with respect to the parameter β, and:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ and } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Solution: We Start by writing down the definition of the likelihood function as the probability of the observation that we have made. We then we put in the pmf for the Poisson random variable from part (a) and simplify the expression.

$$\begin{aligned}
L(\beta\,; y_1, \ldots, y_n) &= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | x_i\,; \beta) \\
&= \prod_{i=1}^{n} \frac{(\beta x_i)^{y_i}}{y_i\,!} e^{-\beta x_i} \\
&= \left(\prod_{i=1}^{n} \frac{x_i^{y_i}}{y_i!}\right) \beta^{(y_1+\ldots+y_n)}\, e^{-\beta(x_1+\ldots+x_n)} \\
&= K\, \beta^{(y_1+\ldots+y_n)}\, e^{-\beta(x_1+\ldots+x_n)} \\
&= K\, \beta^{n\overline{y}}\, e^{-n\overline{x}\beta}.
\end{aligned}$$

As required.

[4 marks]

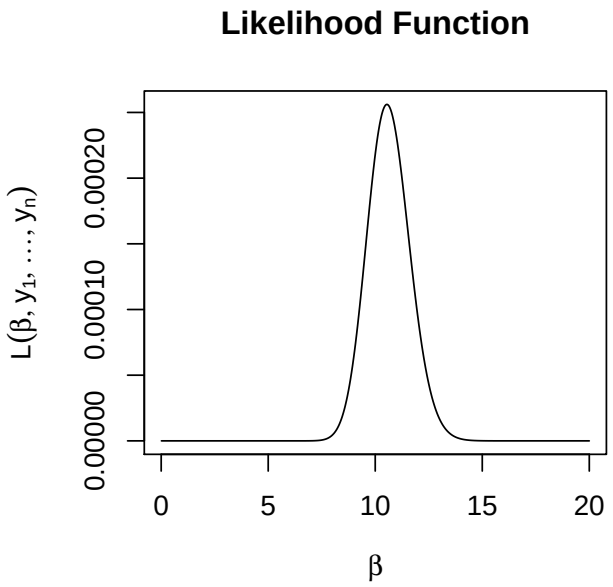

Figure 1: Likelihood Function for the Observed Number of Posts and Engagements

(c) Differentiate the likelihood function to find the maximum likelihood estimator of β, $\hat{\beta}$. You may refer to Figure 1 to justify your calculations. [4 marks]

Solution: We start by differentiating the likelihood function we found in part (b). We need to utilize both the product and chain rules of differentiation. We then simplify the expression as far as we can by breaking out everything we can outside of the brackets containing the sum that comes out of applying the product rule.

$$\begin{aligned}\frac{d}{d\beta}L(\beta\,;y_1,\ldots,y_n) &= \frac{d}{d\beta}\left(K\beta^{n\overline{y}}e^{-n\overline{x}\beta}\right)\\ &= K\left(n\overline{y}\beta^{(n\overline{y}-1)}\,e^{-n\overline{x}\beta} - \beta^{n\overline{y}}n\overline{x}\,e^{-n\overline{x}\beta}\right)\\ &= K\beta^{(n\overline{y}-1)}e^{-n\overline{x}\beta}\,(n\overline{y} - \beta n\overline{x})\end{aligned}$$

We then set the differential to 0 in order to find the value of β which maximizes the Likelihood function, and this value is $\hat{\beta}$, the

maximum likelihood estimate of β:

$$\frac{d}{d\beta} L(\beta; y_1, \ldots, y_n)\Big|_{\beta=\hat{\beta}} = 0$$
$$\Rightarrow K\hat{\beta}^{(n\overline{y}-1)} e^{-n\overline{x}\hat{\beta}} \left(n\overline{y} - \beta n\overline{x}\right) = 0$$

This implies Three possible values of $\hat{\beta}$. Either $\hat{\beta} = 0$, $\hat{\beta} = \infty$ or $(n\overline{y} - \beta n\overline{x}) = 0$. We can see from Figure 1 that the first two solutions don't maximize the likelihood function, so we proceed with the third possibility:

$$\begin{aligned} (n\overline{y} - \beta n\overline{x}) &= 0 \\ n\overline{y} &= \beta n\overline{x} \\ \frac{n\overline{y}}{n\overline{x}} &= \hat{\beta} \end{aligned}$$

where leaving the n in makes it more obvious that this can be rewritten as

$$\hat{\beta} = \frac{y_1 + \ldots + y_n}{x_1 + \ldots + x_n}.$$

So we have found the maximum likelihood estimate, and all that remains to turn it into a maximum likelihood estimator is to replace the observations y_i with the random variables Y_i:

$$\hat{\beta} = \frac{Y_1 + \ldots + Y_n}{x_1 + \ldots + x_n}.$$

(d) Bill tests his model by recording the number of posts he made and the corresponding engagements received during three different hours. The data is as follows:

Hour (i)	Number of posts (x_i)	Engagements received (y_i)
1	2	23
2	5	52
3	4	41

Using the maximum likelihood estimator found in part (c), find the maximum likelihood estimate $\hat{\beta}$ based on the provided data.

Solution: Since we have the maximum likelihood estimator from part (c), we find the maximum likelihood estimate by just putting in the values of the observations of the random variable as well as the predictor variables, so:

$$\hat{\beta} = \frac{23 + 52 + 41}{2 + 5 + 4} = 10.\overline{54}$$

[1 marks]

3 Consider the random variable X with the following probability density function:

$$f(x) = \begin{cases} \frac{6e^{-2x}}{c} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) Determine the value of the constant c

Solution: To find the value of the constant we use the fact that we know that the pdf has to have a total area of 1, so we simply integrate the pdf, and set the value to 1, and solve the resulting equation:

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)dx &= 1 \\ \int_{-\infty}^{0} 0dx + \int_{0}^{\infty} \frac{6e^{-2x}}{c}dx &= 1 \\ 0 + \left[-\frac{6e^{-2x}}{2c}\right]_0^{\infty} &= 1 \\ \frac{3}{c}\left[-e^{-2x}\right]_0^{\infty} &= 1 \\ \frac{3}{c}(-e^{-\infty} + e^{0}) &= 1 \\ \frac{3}{c}(0+1) &= 1 \\ c &= 3 \end{aligned}$$

[2 marks]

(b) Find the cumulative distribution function of X, $F_X(x)$.

Solution: First, it is obvious that for $x < 0$ we have $F_X(x) = 0$ since $f_X(x) = 0$ on this whole range. For $x \geq 0$, we find $F_X(x)$ by integrating the pdf from $-\infty$ (in practice 0) to the value x, remembering to use a dummy variable, u. So, for $x \geq 0$, we have:

$$\begin{aligned} F_X(x) &= 0 + \int_0^x 2e^{-2u}du \\ &= \left[-e^{-2u}\right]_0^x \\ &= -e^{-2x} + 1 \\ &= 1 - e^{-2x} \end{aligned}$$

Combining both ranges, the CDF is:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-2x} & \text{for } x \geq 0 \end{cases}$$

[3 marks]

(c) Find the probability that X takes a value between 0 and 1, $\mathbb{P}(0.5 < X < 1)$.

Solution: Since we have already found the cdf, $F_X(x)$ in part (b) the easiest way to find this probability is just to plug the values into the formula $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$, so we have:

$$\begin{aligned} \mathbb{P}(0.5 < X \leq 1) &= F_X(1) - F_X(0.5) \\ &= 1 - e^{-2} - (1 - e^{-2\times 0.5}) \\ &= e^{-1} - e^{-2} \\ &\approx 0.23 \end{aligned}$$

so

$$\mathbb{P}(0.5 < X \leq 1) \approx 0.23.$$

[2 marks]