Assignment 03

Bayesian models & data analysis

2025-01-28

Part 1: Continuous random variables

1.1 A shooter tries to hit a target 5 times. Suppose the shooter misses the target by following margins (in centimetres) in 5 trials:

$5.76, 4.56, 2.33, 6.32, 5.4$

What is the sample space of the experiment if you want to anaylze the accuracy of the shooter in terms of distance from the target?

Consider a random variable X. The random variable X maps the above sample space to a discrete space Ω_x such that the target missed by a margin of grater than 5 centimetres is a "failure" otherwise it is a "success". And, an outcome in the sample space Ω_x represents the number of successes out of 5 trials.

(a) Describe the random variable X as a function of ω, where ω is member of the sample space Ω and it is the set of all five margins (in cm) by which the shooter misses the target in a trial.

(b) Define a probability distribution for the random variable X?
(Hint: The probability distribution of a random variable X is the probability-assigner function that assigns probability values to all possible values of the random variable X.)

1.2 Suppose a random variable X is normally distributed. The probability density function of the normal distribution is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

(a) Find the probability density of obtaining $x = 0$, given that $\mu = 1, \sigma = 1$.

(b) Find the probability density of obtaining $x = 1$, given that $\mu = 0, \sigma = 1$.

(c) You are given
The probability that the outcome occur between x_1 and x_2: $P(x_1 \leq X \leq x_2) = 0.3$
The probability that the outcome occur between x_1 and x_3: $P(x_1 \leq X \leq x_3) = 0.45$
Find the probability that the outcome occur between x_2 and x_3.

1.3 A continuos random variable X has the following probability density function
$f(x) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1}$
The domain of the function is $[0, 1]$ i.e., the random variable X can take any real number value between 0 and 1.

(a) Find the mode of the variable X given that $\alpha > 1$ and $\beta > 1$.
(Hint: The mode is the value of x at which the first derivative of the probability density function becomes zero, $\frac{d}{dx}f(x) = 0$.)

1.4 Suppose a random variable X is normally distributed. The probability density function (PDF) $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ assigns probability densities to the values of the random variable X.

(a) What are the mean and the variance of the distribution?
(Hint: Compare the above PDF with the PDF of a normal distribution with the mean μ and the variance σ^2: $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$)

(b) What is the probability density of obtaining X=0?

(c) What is the probability of obtaining a score of less than 0 i.e., $X < 0$?

(d) What is the probability of obtaining X=0?

Part 2: The distribution of distances between related words in natural languages

You are given corpora of three natural languages annotated with structural relationships among words.

For example, in a sentence shown below

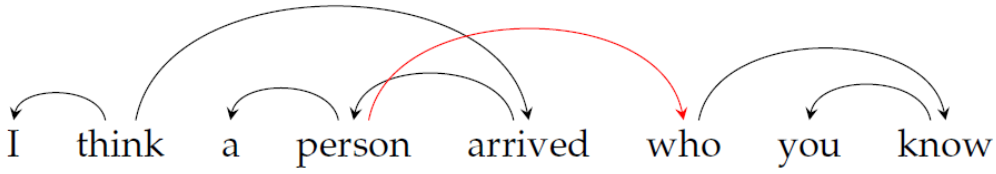


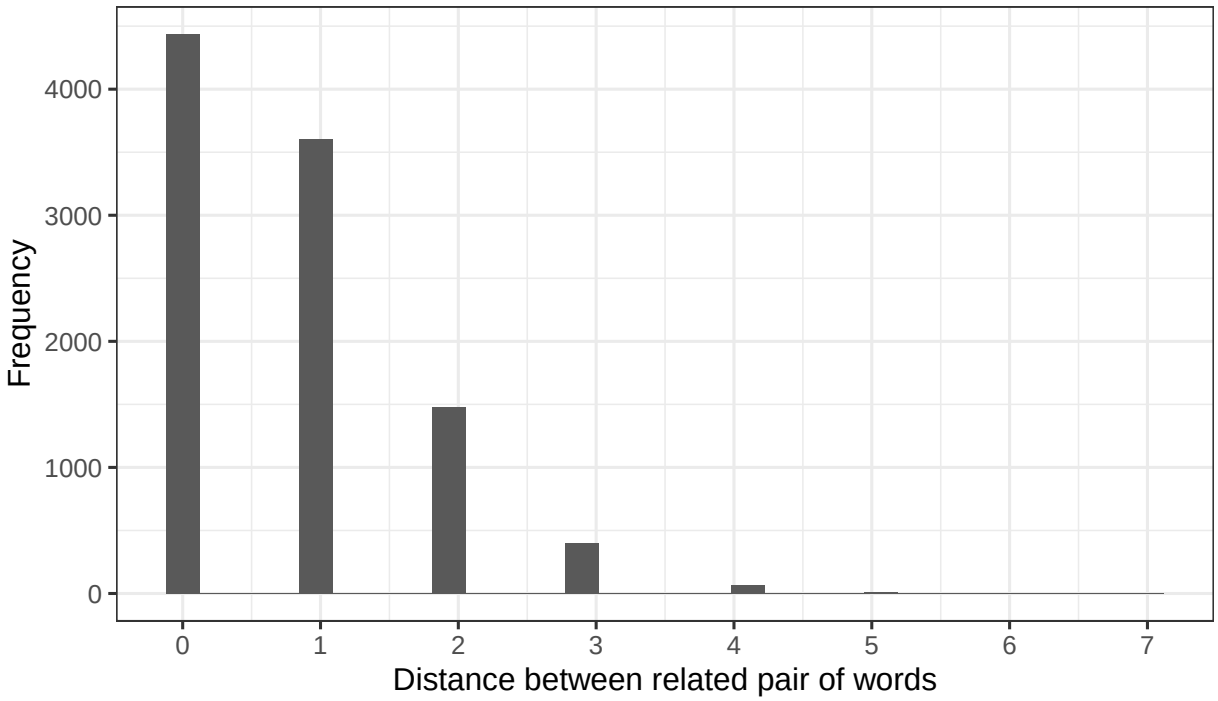
Figure 1: A graph representing the syntactic relationships between words in a sentence

The related word are connected via directed edges, going from one word to the other. (The above network-like representation is called a directed acyclic graph; it is used to encode hierarchical relationships among words in a sentence).

You can calculate the linear distances between pair of words connected via directed edges. For example, one word occur between the related word pair *person* and *who*; so, the linear distance between *person* and *who* is 1. Similarly, the distance between *I* and *think* is 0, the distance between *think* and *arrived* is 2.

In a natural language, the distances between related words follows a typical distribution as shown below:

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

The above distribution shows that the related word-pairs with linear distance of zero are much more frequent than word pairs with linear distance of say 3. And, the word pairs with linear distance of more than 5 rarely occur in natural languages.

The following probability mass function can provide a good approximation to the distribution of distances between related word-pairs.

$$f(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{1}$$

where k is the value of linear distance between word-pairs and can be 0, 1, 2, 3, etc, and λ represent the average distance between related words over an entire corpus. The function $f(k, \lambda)$ assigns probability to occurence of word-pairs with distance k in a corpus.

Suppose the distribution of distances between words in three languages, English, German and Czech is given by

English: $f(k) = \frac{\lambda_1^k e^{-\lambda_1}}{k!}$

German: $f(k) = \frac{\lambda_2^k e^{-\lambda_2}}{k!}$

Czech: $f(k) = \frac{\lambda_3^k e^{-\lambda_3}}{k!}$

where $\lambda_1, \lambda_2, \lambda_3 < 1$

2.1 You are given $\lambda_3 < \lambda_1 < \lambda_2$. Which language has highest number of word pair with linear distance 0?

2.2 You are given $\lambda_1 = 0.7$, $\lambda_2 = 0.8$, and $\lambda_3 = 0.6$. Find the number of related word-pairs

with linear distance of 1 in each language given that the total number related words in English, German and, Czech corpora are 10^5, 10^7 and 10^8 respectively.

2.3 Suppose you artificially generate directed acyclic graphs and calculate the linear distances between the nodes connected via directed edges. Let us call this random structure corpus.

The distribution of linear distances between nodes in the random structure corpus is given by

$$f(k) = \frac{2^k e^{-2}}{k!} \tag{2}$$

How is the distribution of linear distances in natural languages different from that in the random structure corpus?

Part 3: The likelihood function

Suppose a random variable X has the probability density function $f(x, \theta)$ where θ is a parameter of the probability density function and x is a value of the random variable X. You can write:

$X \sim f(x, \theta)$

The probability density function (PDF) $f(x, \theta)$ tells you the probability density of generating an outcome x when the value of θ is known/assumed. For example, if you know/assume $\theta = 2$, you can calculate the probability density for different values of the random variable such as $X = 5$, $X = 3$, $X = 100$. Basically, the PDF is viewed as a function of x when θ is fixed.

However, you can also view the PDF in a different manner. You can calculate the probability density of obtaining a given, fixed outcome x for different values of θ. That is, the PDF can be viewed as a function θ when x is fixed. This alternative characterization of the PDF is called **the likelihood function**.

The likelihood function is a function that maps the values of the parameter θ to probability densities, when the sample x is taken as a fixed, observed quantity.

In sum, the PDF is a function of x where θ is assumed to be fixed; the likelihood function is a function of parameter θ when the sample x is fixed/known.

The likelihood function is often represented by $\mathcal{L}(\theta|x)$:

$\mathcal{L}(\theta|x) = f(x, \theta)$ when x is fixed

Use the above information to do the following exercise.

3.1 In a visual word recognition experiment, a participant has to recognize whether a string shown on screen is a meaningful word (e.g., "book") or a non-word (e.g., "bktr"). The participant is asked to answer "yes" if the shown string is a meaningful word, and "no" if it is a meaningless non-word. Suppose a participant is shown 5 strings on the screen one by one. The time taken by the participants to recognize each string is shown below (in milliseconds):

Recognition time for 5 strings: $303.25, 443, 220, 560, 880$

Suppose the random variable X represents the string recognition times.

A researcher proposes a hypothesis that the string recognition times are generated by the probability density functions $f(x, \mu)$:

$X \sim f(x, \mu)$

such that,

$f(x, \mu) = \frac{1}{x\sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2}}$

(a) Plot the graph of the likelihood function with respect to values of μ, assuming that x is fixed to 220.

(Hint: Choose a range of values for μ and plug those values in the function $f(x, \mu)$ along with $x = 220$ to get probability densities; plot a graph with μ on x-axis and the corresponding values of $f(x, \mu)$ on y-axis.)

(b) Graph the likelihood function when x is (fixed as) the observed sample of recognition times i.e., $303.25, 443, 220, 560, 880$.

(Hint: The probability density for $\mu = \mu_1$ and $x = [x_1, x_2, x_3, ..., x_n]$ will be $f(x_{1:n}, \mu_1) = \frac{1}{\left(\prod_{i=1}^{n} x_i\right)\left(\sqrt{2\pi}\right)^n} e^{-\frac{\sum_{i=1}^{n} (\log x_i - \mu_1)^2}{2}}$.)

(c) For what value of μ, the likelihood (probability density) of obtaining the observed sample $303.25, 443, 220, 560, 880$ is maximum? You do not need to be precise, you can tell the approximate value.

Note: You do not need the following information to do the above problem, this is only for your understanding.

Why do we need a likelihood function?
Suppose you are planning to collect data for an experiment. Before you collect any data, you can only speculate which outcomes are more/less likely using your assumptions about the underlying process. For example, if you assume that the underlying process has the PDF $f(x; \theta)$, you would know which values of x are more likely to be generated in the experiment.
But after you have collected a sample of data, say x_{obs}. You can now calculate: how likely it is that a parameter value θ_i would have generated the observed data x_{obs}. That is, you can compute likelihood for different values of θ after seeing the data x_{obs}.
For Bayesian inference, it is critical to write the PDF as a function of parameter values θ when the data x is known. Thus, the PDF f(x | theta) represents our assumption about the underlying process and the likelihood function, $\mathcal{L}(\theta|x)$ tells which values of θ are more likely when the data x is known and given.

Part 4: Distributions in R

4.1 Practice using the **pnorm** function

The R function **pnorm(x,mean,sd)** calculates the probability of obtaining a value less than x from a normal distribution with mean=mean and standard deviation=sd. (Check ?pnorm).

Given a normal distribution with mean 500 and standard deviation 100, use the **pnorm** function to calculate the following:

(a) The probability of obtaining a value of 700 or less

(b) The probability of obtaining a value of 900 or more

(c) The probability of obtaining values between 200 and 800 from this distribution

4.2 Practice using the **dnorm** function

Given a normal distribution with mean 650 and standard deviation 125. Calculate the probability density of obtaining a value of 550 from this distribution.

(Hint: Use dnorm(x=550,mean=650,sd=125).)

4.3 Practice using the function **rnorm**

The following line of code draws n values from a normal distribution with mean 300 and standard deviation 200.

```
xsim <- rnorm(n, 300, 200)
```

Draw 10000 values from the above normal distribution (set $n = 10000$ in the above code). Calculate the mean, standard deviation, minimum, and maximum of the values contained in the vector *xsim*.