Homework #01

SMDS-2023-2024

Statstical Methods in Data Science II a.y. 2022-2023

M.Sc. in Data Science

deadline: April 26th, 2024

A. Simulation

1. Consider the following joint discrete distribution of a random vector (Y, Z) taking values over the bi-variate space:

$$\begin{aligned} \mathcal{S} = \mathcal{Y} \times \mathcal{Z} &= \{(1,1);(1,2);(1,3);\\ &\quad (2,1);(2,2);(2,3);\\ &\quad (3,1);(3,2);(3,3)\} \end{aligned}$$

The joint probability distribution is provided as a matrix `J` whose generic entry `J[y,z]` $= Pr\{Y = y, Z = z\}$

```
J
```

```
     1    2    3
1 0.06 0.17 0.10
2 0.10 0.12 0.11
3 0.14 0.02 0.18
```

```
S
```

```
      row col
(1,1)   1   1
(1,2)   1   2
(1,3)   1   3
(2,1)   2   1
(2,2)   2   2
(2,3)   2   3
(3,1)   3   1
(3,2)   3   2
(3,3)   3   3
```

You can load the matrix `S` of all the couples of the states in $\mathcal{S}$ and the matrix `J` containing the corresponding bivariate probability masses from the file "Hmwk.RData". How can you check that J is a probability distribution?

---

1. Answer:

We can check whether J is a *valid probability distribution* by verifying if it satisfies the two properties of *non-negativity* and *summation to 1*. Since J is discrete we will refer to the discrete versions of the properties.

The first states that all elements of the probability distribution must be non-negative:

$Pr\{Y = y, Z = z\} \geq 0 \; \forall y, z \in \{1, 2, 3\}$

We can write a simple conditional statement to check if this condition is satisfied:

```r
# function to verify if a probability
# distribution is non-negative
verify_non_negativity <- function(A) {
    return(sum(A >= 0) == length(A))
}

verify_non_negativity(J)
```

```
[1] TRUE
```

As we can see and was evident from the previous code block that printed the matrix the first property is satisfied. The property of summation to 1 instead states that the elements of the probability distribution must sum up to 1:

$\sum_{y,z \in \{1,2,3\}} Pr\{Y = y, Z = z\} = 1$

This also can be checked with a simple conditional statement:

```r
# function to verify is a probability
# distribution has sum equal to 1
verify_sum_to_one <- function(A) {
    return(all.equal(sum(A), 1))
}
```

It's important to notice that, to account for *machine precision*, we used the method `all.equal` instead of the `==` operator.

```r
verify_sum_to_one(J)
```

```
[1] TRUE
```

Also the second condition is satisfied, therefore we can say that J is indeed a valid probability distribution.

---

2. How many *conditional distributions* can be derived from the joint distribution `J`? Please list and derive them.

2. Answer:

We can derive 6 conditional probability distributions by fixing on either Y or Z a value in $\{1, 2, 3\}$; we have 2 random variables with 3 possible values each, therefore $3 * 2 = 6$ conditional distributions. To derive them we have just to implement the definition of conditional probability:

$Pr(Y = y|Z = z) = \frac{Pr(Y=y,Z=z)}{Pr(Z=z)}$

Where at the numerator we have the joint distribution, provided by the matrix J, and at the denominator the *marginal distribution* of the value we condition on, defined as:

$Pr(Z = z) = \sum_y Pr(Y = y, Z = z)$

Base elements of this second formula are elements of the joint distribution, therefore we have derived all ingredients we need for the computation. In the previous definitions Y and Z can be swapped to get the formulas for Z conditioned on Y. Below we implement the functions to get the marginal and conditional distributions:

```r
# support of y
Y = 1:3
# support of z
Z = 1:3

# marginal distribution of z
get_marginal_distro_Z <- function() {
    marginal = c()
    for (z in 1:3) {
        marginal = c(marginal, sum(J[, z]))
    }
    return(marginal)
}

# marginal distribution of y
get_marginal_distro_Y <- function() {
    marginal = c()
    for (y in 1:3) {
        marginal = c(marginal, sum(J[y, ]))
    }
    return(marginal)
}

# marginal distribution of y given z
get_conditional_distro_Y <- function(y, z) {
    return(J[y, z]/get_marginal_distro_Z()[z])
}
```

```
# marginal distribution of z given y
get_conditional_distro_Z <- function(y, z) {
    return(J[y, z]/get_marginal_distro_Y()[y])
}
```

Here we call the functions to get for each possible conditioning the conditional probability distribution:

Pr(Y | Z = 1):

```
get_conditional_distro_Y(Y, 1)
```

```
        1         2         3 
0.2000000 0.3333333 0.4666667 
```

Pr(Y | Z = 2):

```
get_conditional_distro_Y(Y, 2)
```

```
         1          2          3 
0.54838710 0.38709677 0.06451613 
```

Pr(Y | Z = 3):

```
get_conditional_distro_Y(Y, 3)
```

```
        1         2         3 
0.2564103 0.2820513 0.4615385 
```

Pr(Z | Y = 1):

```
get_conditional_distro_Z(1, Z)
```

```
        1         2         3 
0.1818182 0.5151515 0.3030303 
```

Pr(Z | Y = 2):

```
get_conditional_distro_Z(2, Z)
```

```
        1         2         3 
0.3030303 0.3636364 0.3333333 
```

Pr(Z | Y = 3):

```
get_conditional_distro_Z(3, Z)
```

```
         1          2          3 
0.41176471 0.05882353 0.52941176 
```

---

3. Make sure they are probability distributions.

---

3. Answer:

To make sure every conditional distribution we created is a probability distribution we can simply use the functions defined over point $A.1$ and check whether they give us a positive response. To reduce the amount of code required we create a new function that checks them both:

```
# function that verifies if a probability
# distribution satisfies both properties
is_valid_distro <- function(A) {
    return(verify_non_negativity(A) & verify_sum_to_one(A))
}
```

Pr(Y | Z = 1):

```
is_valid_distro(get_conditional_distro_Y(Y, 1))
```

```
[1] TRUE
```

Pr(Y | Z = 2):

```
is_valid_distro(get_conditional_distro_Y(Y, 2))
```

```
[1] TRUE
```

Pr(Y | Z = 1):

```
is_valid_distro(get_conditional_distro_Y(Y, 3))
```

```
[1] TRUE
```

Pr(Z | Y = 1):

```
is_valid_distro(get_conditional_distro_Z(1, Z))
```

```
[1] TRUE
```

Pr(Z | Y = 2):

```
is_valid_distro(get_conditional_distro_Z(2, Z))
```

```
[1] TRUE
```

Pr(Z | Y = 3):

```
is_valid_distro(get_conditional_distro_Z(3, Z))
```

```
[1] TRUE
```

---

4. Can you simulate from this `J` distribution? Please write down a working procedure with few lines of R code as an example. Can you conceive an alternative approach? In case write down an alternative working procedure with few lines of R

---

4. Answer:

We can simulate from the `J` by using the `sample` method. We will have to treat the matrix as a vector and turn back the indices of the vector into tuples:

```r
index_to_tuple <- function(n) {
    row = ((n - 1)%%3) + 1
    col = ((n - 1)%/%3) + 1

    return(c(row, col))
}

sample_from_J <- function() {
    indices = sample(length(J), size = 1, prob = J)
    tuples = index_to_tuple(indices)
    return(tuples)
}
```

We can think of an alternative method that exploits the previously computed marginal and conditional distributions. We first sample Z from its marginal distribution and then sample Y from the distribution conditioned on the result of the first sampling.

```r
# sampling from marginal Z
sample_from_marginal_Z <- function() {
    distro = get_marginal_distro_Z()
    sample(length(distro), size = 1, prob = distro)
}

# sampling from Y conditioned on Z
sample_from_Y_conditioned <- function(z) {
    distro = get_conditional_distro_y(Y, z)
    sample(length(distro), size = 1, prob = distro)
}

sample_marginal_conditional_1 <- function() {
    z = sample_from_marginal_Z()
    y = sample_from_Y_conditioned(z)
    return(c(y, z))
}
```

Of curse we can also do the opposite, first sampling from the marginal distribution of Y and then from the conditional distribution of Z conditioned on the result of the first sampling.

```r
# sampling from marginal Y
sample_from_marginal_Y <- function() {
    distro = get_marginal_distro_Y()
    sample(length(distro), size = 1, prob = distro)
}

# sampling from Z conditioned on Y
sample_from_Z_conditioned <- function(y) {
    distro = get_conditional_distro_z(y, Z)
    sample(length(distro), size = 1, prob = distro)
}

sample_marginal_conditional_2 <- function() {
    y = sample_from_marginal_Z()
    z = sample_from_Y_conditioned(y)
    return(c(y, z))
}
```

B. Bulb lifetime: a conjugate Bayesian analysis of exponential data

You work for Light Bulbs International. You have developed an innovative bulb, and you are interested in characterizing it statistically. You test 20 innovative bulbs to determine their lifetimes, and you observe the following data (in hours), which have been sorted from smallest to largest.

1, 13, 27, 43, 73, 75, 154, 196, 220, 297,
344, 610, 734, 783, 796, 845, 859, 992, 1066, 1471

Based on your experience with light bulbs, you believe that their lifetimes Y_i can be modeled using an exponential distribution conditionally on θ where $\psi = 1/\theta$ is the average bulb lifetime.

1. Write the main ingredients of the Bayesian model.

1. Answer:

The main ingredients of the Bayesian model are

- **Prior distribution**: the prior distribution $\pi(\theta)$ is a probability distribution that represents the knowledge we have about the parameter of interest θ before any evidence gathered from observing the data is taken into account.
- **Likelihood function**: the likelihood function $f(\theta|y)$

2. Choose a conjugate prior distribution $\pi(\theta)$ with mean equal to 0.003 and standard deviation 0.00173.

Since our knowledge about the problem suggests us that the problem can be modeled using an exponential distribution the choice for a conjugate prior ends on a Gamma distribution. We indicate the shape and rate parameters of the Gamma respectively with α and β, while we use μ and σ^2 for mean and variance.

Since $\mu = \frac{\alpha}{\beta}$ and $\sigma^2 = \frac{\alpha}{\beta^2}$ we can easily derive:

$\alpha = \frac{\mu^2}{\sigma^2} = \frac{0.003^2}{0.00173^2} = 3.007$

$\beta = \frac{\mu}{\sigma^2} = \frac{0.003}{0.00173^2} = 1002.372$

Therefore:

$\pi(\theta) \sim Gamma(\alpha = 3.007, \beta = 1002.372)$

3. Argue why with this choice you are providing only a vague prior opinion on the average lifetime of the bulb.

4. Show that this setup fits into the framework of the conjugate Bayesian analysis.

5. Based on the information gathered on the 20 bulbs, what can you say about the main characteristics of the lifetime of your innovative bulb? Argue that we have learnt some relevant information about the θ parameter and this can be converted into relevant information about the unknown average lifetime of the innovative bulb $\psi = 1/\theta$.

6. However, your boss would be interested in the probability that the average bulb lifetime $1/\theta$ exceeds 550 hours. What can you say about that after observing the data? Provide her with a meaningful Bayesian answer.

C. Exchangeability

Let us consider an infinitely exchangeable sequence of binary random variables $X_1, ..., X_n, ...$

1. Provide the definition of the distributional properties characterizing an infinitely echangeable binary sequence of random variables $X_1, ..., X_n, ...$. Consider the De Finetti representation theorem relying on a suitable distribution $\pi(\theta)$ on $[0, 1]$ and show that

$$\begin{aligned} E[X_i] &= E_\pi[\theta] \\ E[X_iX_j] &= E_\pi[\theta^2] \\ Cov[X_iX_j] &= Var_\pi[\theta] \end{aligned}$$

1. Answer:

TODO: Provide the definition of the distributional properties characterizing an infinitely echangeable binary sequence

Showing $\mathbf{E}[\mathbf{X_i}] = \mathbf{E}_\pi[\theta]$:

From the definition of the expected value defined over a binary value we have that:

$$E[X_i] = \sum_{x_i=0}^{1} x_i Pr(X_i = x_i)$$

As starting hypothesis our sequence $X_1, ..., X_n, ...$ is infinitely exchangeable, therefore De Finetti's Theorem holds. We write the formula from De Finetti's theorem as was presented in the course's slides, since we are working with binary random variables:

$$Pr(X_1 = x_1, ..., X_n = x_n) = \int_{[0,1]} \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\pi(\theta)d\theta$$

This holds for any n, therefore we can fix $n = 1$ from which we get permutations consisting of a single element that we will call X_i, that is the same as the one presented in the first formula:

$$Pr(X_i = x_i) = \int_{[0,1]} \theta^{x_i}(1-\theta)^{1-x_i}\pi(\theta)d\theta$$

We can therefore rewrite the latter as:

$$E[X_i] = \sum_{x_i=0}^{1} x_i \int_{[0,1]} \theta^{x_i}(1-\theta)^{1-x_i}\pi(\theta)d\theta$$

When $x_i = 1$ the element of the sum is equal to zero has it is multiplied by zero. We then rewrite the formula substituting x_i with 1:

$$E[X_i] = 1 \int_{[0,1]} \theta^1 (1-\theta)^{1-1} \pi(\theta) d\theta = \int_{[0,1]} \theta \pi(\theta) d\theta$$

The element on the left of the equation follows exactly the definition of expected value for continuous random variables, from which we conclude that:

$$E[X_i] = E_\pi[\theta]$$

Showing $\mathbf{E}[\mathbf{X_i X_j}] = \mathbf{E}_\pi[\theta^\mathbf{2}]$:

We start in a similarly to the previous proof, with the formula of the expected value, this time of the product of two binary random variables:

$$E[X_i X_j] = \sum_{x_i=0}^{1} \sum_{x_j=0}^{1} x_i x_j Pr(X_i = x_i, X_j = x_j)$$

Still as before we write the formula from De Finetti's Theorem, this time fixing $n = 2$:

$$Pr(X_i = x_i, X_j = x_j) = \int_{[0,1]} \theta^{x_i} (1-\theta)^{1-x_i} \theta^{x_j} (1-\theta)^{1-x_j} \pi(\theta) d\theta = \int_{[0,1]} \theta^{x_i + x_j} (1-\theta)^{2-x_i-x_j} \pi(\theta) d\theta$$

Substituting this into the expected value formula get us:

$$E[X_i X_j] = \sum_{x_i=0}^{1} \sum_{x_j=0}^{1} x_i x_j \int_{[0,1]} \theta^{x_i + x_j} (1-\theta)^{2-x_i-x_j} \pi(\theta) d\theta$$

However the only element of those summations that is not zero is the one where both x_i and x_j are equal to 1:

$$E[X_i X_j] = \int_{[0,1]} \theta^2 (1-\theta)^0 \pi(\theta) d\theta = \int_{[0,1]} \theta^2 \pi(\theta) d\theta$$

Which is the definition of the second moment of θ, therefore:

$$E[X_i X_j] = E_\pi[\theta^2]$$

Showing $\mathbf{Cov}[\mathbf{X_i}, \mathbf{X_j}] = \mathbf{Var}_\pi[\theta]$:

Let's recall the definition of covariance between two random variables X_i and X_j:

$$Cov[X_i, X_j] = E[X_i X_j] - E[X_i] E[X_j]$$

Let's also recall the definition of variance of a random variable A:

$$Var(A) = E[A^2] - E[A]^2$$

We have already found the results of the two components of the covariance, namely:

$$E[X_i X_j] = E_\pi[\theta^2], \qquad E[X_i] = E[X_j] = E_\pi[\theta]$$

And by substituting them we understand they follow the definition of variance of θ:

$$Cov[X_i, X_j] = E_\pi[\theta^2] - E_\pi[\theta]E_\pi[\theta] = E_\pi[\theta^2] - E_\pi[\theta]^2 = Var_\pi[\theta]$$

2. Prove that any couple of random variabes in that sequence must be non-negatively correlated.

2. Answer:

Let's recall the definition of correlation between two random variables X_i and X_j, that is the Pearson Correlation Coefficient:

$$Cor[X_i, X_j] = \rho_{X_i, X_j} = \frac{Cov[X_i, X_j]}{\sigma_{X_i}\sigma_{X_j}}$$

Where σ is the standard deviation, that is the square root of the variance:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{Var[X]} \geq 0$$

The denominator of the correlation formula is always non-negative as it is the product of two standard deviations and they are always non-negative. We have already studied the numerator in the previous points of this exercise and we concluded that:

$$Cov[X_i, X_j] = Var_\pi[\theta] \geq 0$$

We know that the variance is always non-negative, therefore also the numerator is non-negative. Since both numerator and denominator of the correlation formula are non-negative we can conclude that any couple of random variables of the sequence must be non-negatively correlated.

3. Find what are the conditions on the distribution $\pi(\cdot)$ so that $Cor[X_i X_j] = 1$.

3. Answer:

Recalling again the definition of correlation stated above, the get a value of correlation equal to 1 we need the numerator and the denominator of the formula to be the same:

$$Cor[X_i, X_j] = 1 \iff Cov[X_i, X_j] = \sigma_{X_i}\sigma_{X_j}$$

Since X_i and X_j are Bernoulli random variables with the same parameter θ they have the same variance $Var(X_i) = Var(X_j) = \theta(1-\theta)$:

$$\sigma_{X_i}\sigma_{X_j} = \sqrt{Var[X_i]}\sqrt{Var[X_j]} = \sqrt{\theta(1-\theta)}\sqrt{\theta(1-\theta)} = \theta(1-\theta)$$

Then we must impose:

$$Cov[X_i, X_j] = \theta(1-\theta)$$

Recalling one of the results of the first part of this exercise, that is $Cov[X_i, X_j] = Var_\pi[\theta]$ what we are imposing is:

$$Var_\pi[\theta] = \theta(1-\theta)$$

4. What do these conditions imply on the type and shape of $\pi(\cdot)$? (make an example).

4. Answer:

© 2023-2024 - Statistical Methods in Data Science and Laboratory II - 2023-2024

```
Last update by LT: Sun Apr 21 12:35:46 2024
```