DSCI 551 Quiz 2 Solutions

Instructions

rubric={mechanics:1}

- Complete the quiz by editing this R Markdown (`.Rmd` file).
- When you are finished, knit to `.pdf` and submit both files (by "submit" we mean `add`, `commit` and `push` the `.Rmd` and `.pdf` files).
- In this quiz, both the steps and the answer are being graded. Please show your work.
- Some questions require a correct answer to a previous question in order to get it right. For example, the second part of 2.1 requires the first part of 2.1 to be answered correctly. You will be graded as if prerequisite questions are correct.
- Feel free to look at the lecture slides and R documentation (there will be *very light* coding in R required)

Part 1

Let X_1 denote the height of a person selected at random from a certain population, and let X_2 denote the weight of the person. Suppose (X_1, X_2) have a bivariate normal distribution with with $\mu_1 = 5.8$ feet, $\mu_2 = 130$ lbs, $\sigma_1 = 0.2$ foot, $\sigma_2 = 10$ lbs and $\rho = 0.515$.

Hint: you may find the following R functions useful. Feel free to consult the documentation for these functions. You can always as us to check if you're using the function properly.

- `pnorm`: the cdf of a Normal distribution
- `dnorm`: the pdf/density of a Normal distribution
- `qnorm`: the quantile function of a Normal distribution

1.1

rubric={reasoning:3}

What is the marginal distribution of X_1?

$$X_1 \sim N(\mu_1 = 5.8, \sigma_1 = 0.2).$$

1.2

rubric={reasoning:3}

Find the probability that a randomly selected individual's height is between 5.5 feet and 6 feet.

```r
mu1 = 5.8 #ft
sigma1 = 0.2 #ft

pnorm(6,mu1,sigma1) - pnorm(5.5,mu1,sigma1)
```

```
## [1] 0.7745375
```

1.3

rubric={reasoning:3}

Find the 90th percentile (i.e., the 0.9-quantile) of X_1.

```r
qnorm(0.9,mu1,sigma1)
```

```
## [1] 6.05631
```

1.4

rubric={reasoning:3}

Given that the person's height is 6.3 feet, what is the probability that his/her weight is greater than 160 lbs?

$$\mathbb{P}(X_2 > 160 \mid X_1 = 6.3) \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(6.3 - \mu_1), \sqrt{\sigma_2^2(1-\rho^2)}\right).$$

Hence, $\mathbb{P}(X_2 > 160 \mid X_1 = 6.3)$ is given by:

```r
mu2 = 130 #lb
sigma2 = 10 #lb
rho = 0.515

M <-  mu2+rho*(sigma2/sigma1)*(6.3-mu1)
S <- sqrt(sigma2^2*(1-rho^2))
c(M,S)
```

```
## [1] 142.875000   8.571902
```

```r
1-pnorm(160,mean = M,sd = S)
```

```
## [1] 0.02286883
```

1.5

rubric={reasoning:3}

Define $Y = 2X_1 - X_2 + 1$. Find the expected value of Y.

$$\begin{array}{rcc} \mathbb{E}(Y) & = & 2\mathbb{E}(X_1) - \mathbb{E}(X_2) + \mathbb{E}(1) \\ \mathbb{E}(Y) & = & 2\mu_1 - \mu_2 + 1 \\ \mathbb{E}(Y) & = & -117.4. \end{array}$$

1.6

rubric={reasoning:2}

The variance of Y (defined in Question 1.5) is 96.04. What is the distribution of Y?

$$\begin{array}{rcc} Var(Y) & = & 4Var(X_1) + Var(X_2) - 4Cov(X, Y) \\ Var(Y) & = & 4\sigma_1^2 + \sigma_2^2 - 4\sigma_1\sigma_2\rho \\ Var(Y) & = & 96.04. \end{array}$$

Hence,

$$Y \sim N(\mu_Y = -117.4, \sigma_Y = \sqrt{96.04}).$$

Part 2

Eight travelers from Canada to the US are surveyed. It is found that the second, third, and seventh travelers have purchased travel insurance, whereas the others have not. Suppose that the travelers are independent of one another.

Let p be the proportion of all travelers from Canada to the US who purchase travel insurance.

2.1

rubric={reasoning:3}

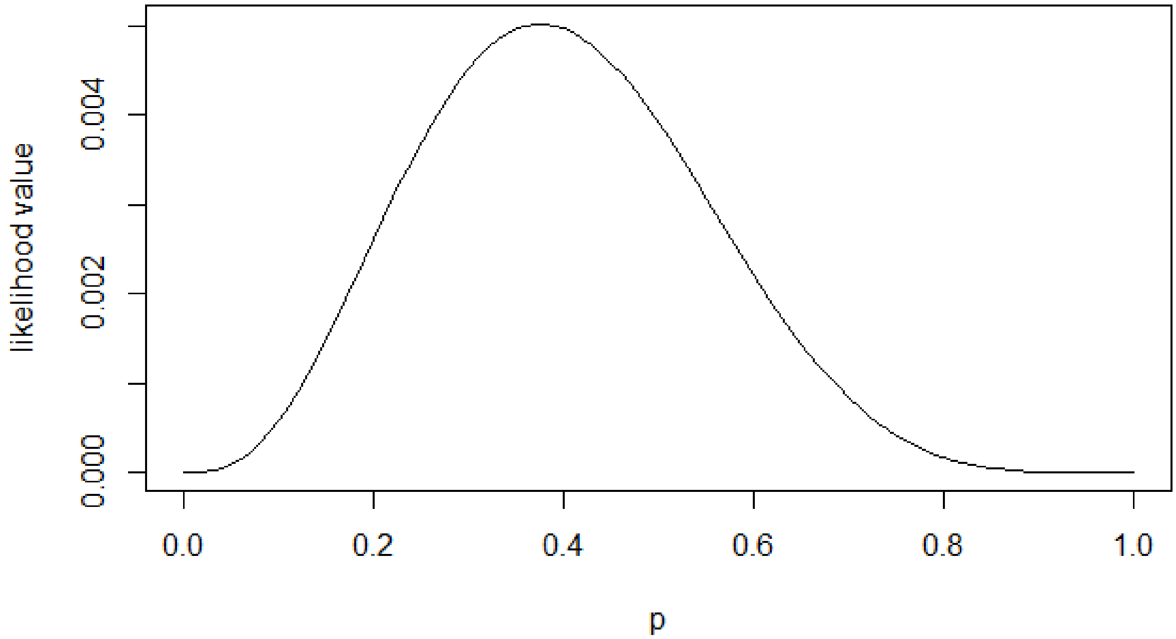


Figure 1: Likelihood function

Suppose I tell you that the true p is one of 0.2, 0.5, or 0.6. Which one will you choose as your estimate? Use the likelihood function above to justify your choice. **Briefly** explain your answer in 1 sentence.

Solutions

Of the possible true values for p, I will choose the value that has the greatest maximum likelihood (lowest negative log likelihood). That value is 0.5. In this particular problem, it is also acceptable to choose the value that is closest to the MLE ($p_{\text{MLE}} = 0.375$); however, this is not true in general (why?).

```
nll.2.1 <- function(p) -3*log(p) - 5*log(1-p)
p.true <- c(.2, .5, .6)
sapply(p.true, nll.2.1)
```

```
## [1] 5.944031 5.545177 6.113931
```

2.2

rubric={reasoning:3}

Based on the plot in Question 2.1 above, what is the maximum likelihood estimate of p (just give a rough approximation; “eyeball” the plot). **Briefly** explain your answer in 1 sentence.

Solutions

The MLE for p is 0.375, at which $\mathcal{L}(p \mid x_i)$ is greatest.

2.3

rubric={reasoning:3}

Define X as a Bernoulli rv corresponding to a randomly chosen individual from the specified population, so that

$$X = \begin{cases} 1 & \text{if a traveller purchased insurance;} \\ 0 & \text{if a traveller didn't purchase insurance.} \end{cases}$$

The pmf is

$$P(X = x) = \begin{cases} p, & \text{if } x = 1 \text{ (i.e., if a traveller purchased insurance);} \\ 1 - p, & \text{if } x = 0 \text{ (i.e., if a traveller didn't purchase insurance)} \end{cases}$$

Write a mathematical form for the likelihood function for the above data. Be sure to indicate what the *variable* of the function is.

Solutions

Three of eight individuals purchased insurance. From the definition of likelihood given a sample,

$$\mathcal{L}(p \mid \{x_i\}_{i=1}^{8}) = \prod_{i=1}^{8} \mathbb{P}(X = x_i) = p^3(1-p)^5.$$

Part 3

3.1

rubric={reasoning:3}

Suppose random variables X and Y have a correlation of 0.2. Can we say whether X and Y are independent/dependent? If we can say, then which one is it? **Briefly** explain your answer in one or two sentences.

Solutions

They are dependent. Non-zero correlation implies non-zero covariance, which implies that $\mathbb{E}(XY) \neq \mathbb{E}(X)\mathbb{E}(Y)$. In particular,

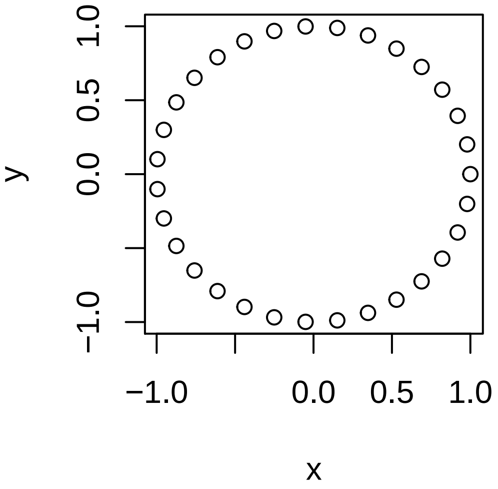
$$\rho_{X,Y} \neq 0 \quad \Longrightarrow \quad \operatorname{Cov}(X,Y) \neq 0 \quad \Longrightarrow \quad \mathbb{E}(XY) \neq \mathbb{E}(X)\mathbb{E}(Y).$$

3.2

rubric={reasoning:3}

Consider the following scatterplot.

```
t <- seq(0, 2*pi, length.out=32)[-1]
x <- cos(t)
y <- sin(t)
plot(x,y)
```



What is the correlation of these data (at least approximately)? Are these data dependent or independent? **Briefly** explain your answer in one or two sentences.

Solutions

These data are dependent: if I know a value `x[i]` for some `i`, then I also know the corresponding `y[i]` value, as one uniquely determines the other. Nevertheless, the Pearson correlation is 0.

It is simple to compute the correlation numerically using R's `cor` function (we round the result to machine precision for clarity).

```
cor(x,y) %>%
  round(1e-16)
```

```
## [1] 0
```

It can also be computed analytically, relying only on symmetry of the data and properties of sin and cos. Firstly, letting $x =$ `x` and $y =$ `y`,

$$\mathbb{E}(xy) = \frac{1}{31}\sum_{j=1}^{31}\cos(2j\pi/31)\sin(2j\pi/31) = \frac{1}{31}\sum_{j=-15}^{15}\cos(2j\pi/31)\sin(2j\pi/31) = 0$$

with the last equality following from the fact that $\cos\cdot\sin : \mathbb{R} \to \mathbb{R}$ is a π-periodic odd function. Secondly, $\mathbb{E}(x)\mathbb{E}(y) = 0$, because

$$\mathbb{E}(y) = \frac{1}{31}\sum_{j=1}^{31}\sin(2j\pi/31) = \frac{1}{31}\sum_{j=-15}^{15}\sin(2j\pi/31) = 0$$

again by symmetry of the data and the fact that sin is 2π-periodic and odd. Hence $\mathbb{E}(xy) = \mathbb{E}(x)\mathbb{E}(y) = 0$. (What happens if the points are no longer evenly distributed on a circle? Can you think of another pattern that is not a circle, but which gives rise to the same phenomenon?)