Exercise 6

Applied Statistics, IT University of Copenhagen

T = Theoretical Exercise, R = R Exercise

Preparation

- Read pages 92, 96–99, 102–108, 111–113, 132–139 from Verzani (2014).

Problems

1. Melencolia distribution (T)

Let X and Y be two random variables with the joint probability mass function given in Table 1. What is

(a) $P(X = Y)$?

(b) $P(X + Y = 5)$?

(c) $P(1 < X \leq 3, 1 < Y \leq 3)$?

(d) $P((X, Y) \in \{1, 4\} \times \{1, 4\})$?

2. Joint distribution (T)

Let X and Y be continuous random variables with the joint probability density function

$$f(x,y) = \begin{cases} cx + 1 & \text{if } x, y \geq 0, x + y < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

(a) Find the constant c.

Table 1: Probability mass function of the Melencolia distribution.

	a=1	a=2	a=3	a=4
b=1	0.1176471	0.0220588	0.0147059	0.0955882
b=2	0.0367647	0.0735294	0.0808824	0.0588235
b=3	0.0661765	0.0441176	0.0514706	0.0882353
b=4	0.0294118	0.1102941	0.1029412	0.0073529

(b) Compute the marginal distribution $f_X(x)$.

(c) Compute $P(Y < 2X^2)$.

3. Covariance (T)

Again, let X and Y be two random variables with the joint probability mass function given in Table 1. Compute $\text{Cov}(X, Y)$.

4. Correlation Coefficient (R)

(a) The data set `normtemp (UsingR)` contains body measurements for 130 healthy, randomly selected individuals. The variable `temperature` measures normal body temperature, and the variable `hr` measures resting heart rate. Make a scatter plot of the two variables. What does the plot show you? Find the Pearson correlation coefficient. How does the estimate relate to the scatter plot?

(b) The data set `nym.2002 (UsingR)` contains information about the 2002 New York city marathon. What do you expect the correlation between age and finishing time to be? Make a scatter plot and compute the correlation coefficient. Does the result match your expectation?

(c) The `batting` set (`UsingR`) data set contains baseball statistics for the 2002 Major League Baseball season. What is the correlation between the number of strikeouts (`SO`) and the number of home runs (`HR`)? Make a scatter plot to see whether there is any trend. Does the data suggest that in order to hit a lot of home runs one should strike out a lot?

5. Covariance, Correlatedness and Independence (R)

Load the data set by copying the data file from course webpage to your working directory and typing `load("mypnts.Rdata")`.

(a) Make a scatter plot of the points. Are the x and y coordinates correlated?

(b) Estimate the *covariance matrix* of the data set. It is a 2×2 matrix containing the all the pairwise covariances, i.e.,

$$\mathbf{C} = \begin{pmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Cov}(Y, Y) \end{pmatrix} \tag{2}$$

What can you see from the covariance matrix estimate?

(c) Apply the mapping

$$x' = ax + by \quad \text{and} \quad y' = cx + dy, \tag{3}$$

to the points where $a = 0.07$, $b = 0$, $c = 1$,and $d = 0.42$. This process is called *whitening*.

(d) Plot the mapped points (use the option 'asp=1' that gives unity aspect ratio), and their marginal distributions (`densityplot`) on both x' and y' axis. Compute the covariance matrix estimate for the mapped points. Are the mapped points uncorrelated? How about independent? Can you see why the mapping is called whitening?

(e) Apply rotation to the mapped points

$$x'' = cos(\alpha)x' - sin(\alpha)y' \quad \text{and} \quad y'' = sin(\alpha)x' + cos(\alpha)y', \tag{4}$$

where $\alpha = -\pi/6$.

Plot the rotated, mapped points, and the marginals on the new axes. Are the rotated, mapped coordinates uncorrelated? How about independent?

(f) What did you learn from this exercise?

Verzani, John. 2014. *Using R for Introductory Statistics.* CRC Press.