Introductory exercises

Eawag Summer School in Environmental Systems Analysis

Note, for the first three exercises you need only pencil and paper.

1. Joint, marginal and conditional distributions I ★

The joint discrete probability table of $P_{A,B}(a,b)$ is given below:

	B.1	B.2	B.3
A.1	0.2	0.1	0.3
A.2	0.1	0.1	0.2

Derive the following probabilities:

- $P_{A,B}(1,2)$
- $P_B(2)$
- $P_{A|B}(1|2)$
- Are A and B independent?

2. Joint, marginal and conditional distributions II ★

Assume the probability densities $p(E \mid B)$, $p(B)$, $p(A, D \mid E)$, and $p(C \mid B, E)$ are known.

- Draw the corresponding directed acyclic graph of the conditional probabilities to visualize the independence structure.

- Derive $p(B, C, E)$
- Derive the joint distribution of A, B, C, D, and E.
- Derive $p(A, B \mid C, D, E)$
- Derive $p(A \mid D)$
- Derive $p(A \mid B, E)$

3. Compound distribution

Assume that:

$$\mu \sim f_\mu(\cancel{m}\,\mu) = \begin{cases} 0.1 \exp(-0.1\cancel{m}\,\mu) & \cancel{m}\,\mu \geq 0 \\ 0 & \text{else} \end{cases}$$

and

$$X \sim f_{X\mid\mu=\cancel{m}}(x \mid \cancel{m}\,\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\cancel{m}\,\mu)^2}{2}\right)$$

This means X is normal -ly distributed with mean μ and μ itself is exponentially distributed.

Derive and interpret:

- $f_{X,\mu}(x, m)$
- $f_X(x)$, a so called compound distribution.
- $P(\mu > 5)$
- $f_{X,\mu\mid\mu>5}(x, \cancel{m}\,\mu)$
- $f_{X\mid\mu>5}(x)$

It is not the aim to find closed forms for the integrals.

4. Sampling vs. evaluating random variables ★

~~Assume~~ Take two random variables X and Y with the following distributions:

$$X \sim \text{Uniform}(0, 1)$$
$$Y \sim \text{Normal}(2, 10)$$

1. Evaluate the probability density $f_X(0.8)$ and $f_Y(0.8)$.
2. Generate 10000 samples from both random variables. Visualize the distributions as histograms.

Another random variable is defined as a function of X as follows:

$$Z = \sin(2\pi X)\sqrt{X}$$

While it is difficult to derive the probability density of Z, sampling from it is easy. 3. Derive the density $f_Z(z)$

~~3.~~ 4. Generate 10000 samples from Z by first sampling from X and then transforming the samples. Visualize as histogram.

Hints

R **Julia**

Most important univariate probability distributions are already implemented in R. Type `?Distributions` to get an overview. For every distribution `__` four functions are defined with the following naming scheme:

```
d__(x, ...)   # evaluate pdf at x
p__(x, ...)   # evaluate cdf at x
q__(p, ...)   # evaluate the p-th quantile
r__(n, ...)   # sample n random numbers
```

For example, for the normal distribution the functions are called `dnorm()`, `pnorm()`, `qnorm()`, and `rnorm()`.

Histograms are generated with the function `hist`. You can adjust the number of bins with the argument `breaks`, e.g. `hist(rnorm(10000), breaks=100)`.

5. Generating data ★

Generate two samples of fictional observations (each of class `matrix`) denoted as $\boldsymbol{Y}_{\text{obs,indep}}$ and $\boldsymbol{Y}_{\text{obs,dep}}$. The former should contain 1000 realisations of two *independent* random variables (as two columns of the matrix) and the latter of two *dependent* ones. Use means of $\boldsymbol{\mu} = (3, 8)$ for both samples. Use the standard deviations $\boldsymbol{\sigma}_{\text{obs,indep}} = (2, 5)$ for the independent variables constituting $\boldsymbol{Y}_{\text{obs,indep}}$, and the covariance matrix.

$$\boldsymbol{\Sigma}_{\text{obs,dep}} = \begin{pmatrix} 4 & 8 \\ 8 & 25 \end{pmatrix}$$

for the dependent variables constituting $\boldsymbol{Y}_{\text{obs,dep}}$.

Hints

R **Julia**

In R, objects of a certain class can often be constructed by a function that matches the class name, such as `matrix()`. You can use `rnorm()` and `cbind()` to construct $\boldsymbol{Y}_{\text{obs,indep}}$ and `rmvnorm()` from the package `mvtnorm` to construct $\boldsymbol{Y}_{\text{obs,dep}}$.

6. Analyzing and visualizing data ★

Perform some preliminary analysis of the data generated in Task 5:

a. What are the interquartile and the 90%-interquantile ranges of your samples?
b. Plot and compare the histograms and the densities of all the marginals
c. Compare the scatterplots of $\boldsymbol{Y}_{\text{obs,indep}}$ and $\boldsymbol{Y}_{\text{obs,dep}}$

d. Compute the covariance and the correlation matrix of $Y_{\text{obs,indep}}$ and $Y_{\text{obs,dep}}$

Which of the above steps reveal a potential correlation structure in your data?

Hints

... produce scatter plots, density plots

R	Julia

Try to arrange multiple plots in the same window by setting `par(mfrow=c(<nrow>,<ncol>))` . Use `quantile()` to calculate the interquantile range; use `hist()` and `plot(density())` to visualize the data; use `cov()` , `cor()` for the covariance and the correlation, respectively.

7. Working with dataframes ★

Real data contain often columns of different data types (e.g. numbers and strings). Dataframes are designed to work with this kind of data conveniently.

a. Import the file `./data/model_growth.csv` as dataframe. Perform some analyses similar to the ones in Task 6.

what's the point? b. Add a new column (say `C_new`) to the growth data frame ~~to~~ it that contains some random values .

Hints

R	Julia

Read the data using `read.table("</path/to/somefile.txt>", header=TRUE)` to indicate that the first row are the column names (use file `../data/model_growth.csv`). To select the column `C_M` from a dataframe, say `data` , you can use `data$C_M` or `data[,"C_M"]` .

To add columns, you can use `cbind` to column-bind the new data to the available matrix and convert everything to a `data.frame` . Then, you can rename the columns by assigning the desired names with function `colnames()` applied to the newly created `data.frame` .

8. Error propagation and functions

~~It is generally known that,~~ I if $f()$ is non-linear, generally,

$$f(E[X]) \neq E[f(X)]$$

where E is the expected value and X is a random variable. Define a non-linear function in R, e.g. $f(x) = \sin \sqrt{x}$. In order to avoid negative values, generate some realizations of a log-normally distributed random variable X. Calculate $f(E[X])$, $E[f(X)]$, $\mathrm{Var}[X]$, $\mathrm{Var}[f(X)]$ and compare them.

Hints

R	Julia

Use `rlnorm` to sample from a log-normal distribution, which accepts the mean and the standard deviation on the log-scale, not on the original scale. Additionally, keep in mind that in R a general function can be defined as:

```
function.name <- function(arg1,arg2){
  result <- arg1 + arg2 # or any other operation
  return(result)
}
```

Most basic functions are already available, and those include both `sin` and `sqrt`. Try `?sin` in the R console to get access to the manual of the harmonic functions. These can be used anywhere in the code, including inside a custom function.