Analysis of Engineering and Scientific Data

Semester 1 – 2019

Sabrina Streipert s.streipert@uq.edu.au

Covariance and Correlation

Definition:

The **covariance** of X and Y is

$$\operatorname{cov}(X,Y) := \mathbb{E}\left[(X-\mathbb{E}[X])\,(Y-\mathbb{E}[Y])\right]$$

Basically, it is a measure for the amount of linear dependency between the variables.

The **correlation** (**correlation coefficient**) of X, Y is

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} \in [-1,1]$$

Properties of Variance and Covariance

- $\operatorname{cov}(X,Y) = \mathbb{E}\,[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

- $\text{cov}(X, Y) = \text{cov}(Y, X)$.

- $\text{cov}(aX + bY, Z) = a\text{cov}(X, Z) + b\text{cov}(Y, Z)$

- $\text{cov}(X, X) = \text{Var}(X)$

- **Marginal Variance:** $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{cov}(X, Y)$.

Example - revisited:

Recall the joint pmf for unfair dice example from last time.

1. $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[x])^2 = \sum_{i=1}^{3} i^2 \frac{1}{3} - \left(\sum_{i=1}^{3} i \frac{1}{3}\right)^2$

   $= \frac{1}{3} \sum_{i=1}^{3} i^2 - \frac{1}{9} \left(\sum_{i=1}^{3} i\right)^2 = \frac{1}{3}\left(\frac{3 \cdot 4 \cdot 7}{6}\right) - \frac{1}{9}\left(\frac{3 \cdot 4}{2}\right)^2 = \frac{14}{3} - 4 = \frac{2}{3}$

2. Covariance:

$$\operatorname{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \sum_{j=1}^{6}\sum_{i=1}^{3} ijp(X=i, Y=j) - \overbrace{2}^{\mathbb{E}[X]}\overbrace{\frac{7}{2}}^{\mathbb{E}[Y]}$$

$$= \sum_{j=1}^{4}\sum_{i=1}^{3} ij\frac{1}{18} + 1\cdot 5\cdot\frac{1}{18} + 1\cdot 6\cdot\frac{1}{18} + 2\cdot 5\cdot\frac{1}{9} + 2\cdot 6\cdot 0 + 3\cdot 5\cdot 0 + 3\cdot 6\cdot\frac{1}{9} - 7$$

$$= \frac{1}{18}\sum_{j=1}^{4} j\sum_{i=1}^{3} i + \frac{5}{18} + \frac{6}{18} + \frac{10}{9} + \frac{18}{9} - 7$$

$$= \frac{1}{18}\left(\frac{4\cdot 5}{2}\right)\left(\frac{3\cdot 4}{2}\right) + \frac{5}{18} + \frac{6}{18} + \frac{10}{9} + \frac{18}{9} - 7 = \frac{1}{18}$$

3. Correlation: $\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\cdot\sqrt{\operatorname{Var}(Y)}} = \frac{\frac{1}{18}}{\sqrt{\frac{2}{3}}\cdot\sqrt{\frac{35}{12}}} = \frac{1}{3\cdot\sqrt{70}}$

Since

$$\mathbb{E}[Y^2] = \sum_{j=1}^{6} j^2\frac{1}{6} = \frac{1}{6}\sum_{j=1}^{6} j^2 = \frac{1}{6}\frac{6\cdot 7\cdot 13}{6} = \frac{91}{6}$$

and therefore

$$\operatorname{Var}(X) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{182-147}{12} = \frac{35}{12}.$$

Conditional Probability Mass Function

Definition:

If X, Y are **discrete** R.V. and $P(X=x) > 0$, then the **conditional probability mass function** of Y given $X = x$ is:

$$P(Y=y \mid X=x) = \frac{P(Y=y, X=x)}{P(X=x)}$$

If X, Y are **continuous** R.V. and $f_X(x) > 0$, then the **conditional probability density function** of Y given $X = x$ is:

$$f_Y(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

Conditional Cumulative Distribution Function

conditional cdf:

$$F_Y(Y = y \mid X = x) = P(Y \leq y \mid X = x)$$

- If X, Y are discrete R.V. and $P(X = x) > 0$, then

$$F_Y(Y = y \mid X = x) = P(Y \leq y \mid X = x) = \frac{P(Y \leq y, X = x)}{P(X = x)}$$

- If X, Y are continuous R.V., then

$$F_Y(Y = y \mid X = x) = P(Y \leq y \mid X = x) = \int_{-\infty}^{y} f_Y(y \mid x)\,\mathrm{d}y$$

Conditional Expectation

- If X, Y are discrete R.V., then the **conditional expectation** of Y given $X = x$ is:

$$\mathbb{E}[Y \mid X] = \sum_y yP(Y = y \mid X = x))$$

  and the conditional expectation of X given $Y = y$ is:

$$\mathbb{E}[X \mid Y] = \sum_x xP(X = x \mid Y = y))$$

- If X, Y are continuous R.V., then the **conditional expectation** of Y given $X = x$ is:

$$\mathbb{E}[Y \mid X] = \int_{-\infty}^{\infty} y F_Y(y \mid x)\,\mathrm{d}y$$

and the conditional expectation of X given $Y = y$ is:

$$\mathbb{E}[X \mid Y] = \int_{-\infty}^{\infty} x F_X(x \mid y)\,\mathrm{d}x$$

Example:

We draw at random a point (X, Y) from the 10 points on the triangle D, see Figure 1.

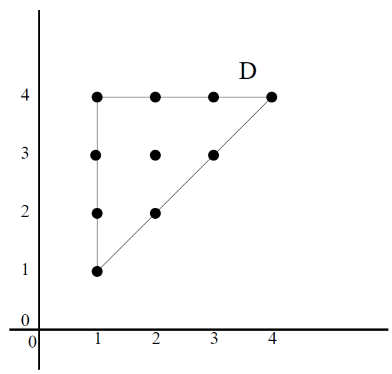

Figure 1: Drawing a point in D.

- Joint pmf: $P(X = i, Y = j) = \frac{1}{10} \quad (i, j) \in D$.

- Marginal pmf of X: $P(X = i) = \frac{5-i}{10}, \qquad i = 1, 2, 3, 4$

- Marginal pmf of Y:

  $P(Y = j) = \frac{j}{10}, \qquad j = 1, 2, 3, 4$

- Conditional pmf:

$$P(Y = j \mid X = i) = \frac{P(Y = j, X = i)}{P(X = i)} = \frac{\frac{1}{10}}{\frac{5-i}{10}} = \frac{1}{5-i}.$$

- Conditional Expectation:

  $E[Y|X = i] = \sum_{j=1}^{4} jP(Y = j \mid X = i) = \sum_{j=1}^{4} j\frac{1}{5-i} = \frac{1}{5-i}\sum_{j=1}^{4} j = \frac{1}{5-i}\frac{4\cdot 5}{2} = \frac{10}{5-i}$

Independence of two Random Variables

Definition:

X, Y are **independent R.V.** if any event defined by X is independent of every event defined by Y, i.e.,

- $$P((X \in A) \cap (Y \in B)) = P(X \in A)P(Y \in B)$$

  for any A and B,

- i.e.,

$$F(x, y) = F_X(x)F_Y(y)$$

- i.e., (if X, Y are discrete R.V.):

$$P(X = x, Y = y) = P_X(x)P_Y(y)$$

- i.e., (if X, Y are continuous R.V.):

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

Independence - Properties

- If X, Y are independent $\longrightarrow \text{cov}(X,Y) = 0$
- If X, Y are independent $\overset{?}{\longleftarrow} \text{cov}(X,Y) = 0$

  NO! For example, let $X \sim U(-1,1)$ then $\mathbb{E}[X] = 0$. Take $Y = g(X) = X^2$, then $\mathbb{E}[XY] = \mathbb{E}[X^3] = 0$ so $\text{cov}(X,Y) = 0$ but clearly the variables are dependent.

- If X, Y are independent $\longrightarrow \rho(X,Y) = 0$

- If X, Y are independent (recall: $\text{cov}(X,Y) = 0$)

$$\longrightarrow Var(aX+bY) = Var(ax)+Var(bY)+2\overbrace{\text{cov}(aX,bY)}^{=0} = a^2\text{Var}(X)+b^2\text{Var}(Y)$$

Example - revisited:

Recalling previous example, see Figure 1.

We note that

$$P(X = 2, Y = 2) = \frac{1}{10} \neq P(X = 2)\, P(Y = 2) = \frac{5-2}{10} \cdot \frac{2}{10} = \frac{6}{100}$$

$\longrightarrow X$ and Y are dependent

Consider now, that we draw at random a point (X,Y) from the 16 points on the square E, see Figure 2.

Then,

$$P(X = 2, Y = 2) = \frac{1}{16} = P(X = 2)\, P(Y = 2) = \frac{4}{16} \cdot \frac{4}{16} = \frac{1}{16}$$

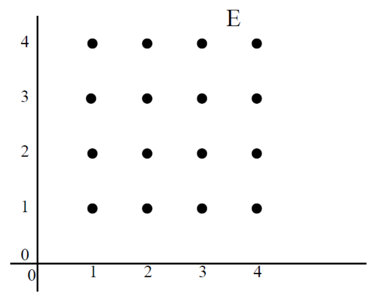

Figure 2: 16 points on the square E.

$\longrightarrow$ X and Y are independent

That does not yet imply independence, since equality has to hold for all values of x and y. To show that in fact X and Y are independent, one has to show: $P(X = x, Y = y) = P(X = x)P(Y = y)$.

Note that in fact X and Y are independent, since

$$\frac{1}{16} = P(X = i, Y = j) = P(X = i)P(Y = j) = \frac{4}{16} \cdot \frac{4}{16} = \frac{1}{16}$$

for any $(i, j) \in E$.

Generalization to multiple random variables

Let $X_1, X_2, \ldots, X_n$ be random variables (random vector):

- If X_i's are discrete, there exists a joint pmf p:

$$p(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n).$$

- If X_i's are continuous, there exists a joint pdf f:

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}.$$

- Joint cdf F:

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n).$$

- If $X_1, X_2, \ldots, X_n$ are discrete R.V., then they are **independent** if:

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_n = x_n),$$

  for all $x_1, x_2, \ldots$.

- If $X_1, X_2, \ldots, X_n$ are continuous R.V., then they are **independent** if:

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

- An infinite sequence $X_1, X_2, \ldots$ of R.V. is called independent if for any finite choice of parameters $i_1, i_2, \ldots, i_n$ (none of them the same), $X_{i_1}, \ldots, X_{i_n}$ are independent.

- Let $X_1, \ldots, X_n$ be discrete R.V.s, with means $\mu_1, \ldots, \mu_n$.

- Let $Y = a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$ where $a, b_1, \ldots, b_n$ are constants. Then

$$\mathbb{E}[Y] = \mathbb{E}[a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n] =$$
$$= a + b_1 \mathbb{E}[X_1] + \cdots b_n \mathbb{E}[X_n] = a + \mu_1 + b_1 \cdots + b_n \mu_n.$$

- If $X_1, \ldots, X_n$ are independent, then

$$\mathbb{E}[X_1 X_2 \cdots X_n] = \mathbb{E}[X_1]\mathbb{E}[X_2] \cdots \mathbb{E}[X_n].$$

Jointly Gaussian RVs

- The n-dimensional density of the random vector

$$\mathbf{X} = (X_1, \ldots, X_n)^\top$$

(column vector), with $X_1, \ldots, X_n$ independent and standard normal, is

$$f_X(x) = (2\pi)^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2}\mathbf{x}^\top \mathbf{x}}.$$

- We consider now the function (transformation) $Z = \boldsymbol{\mu} + B\mathbf{X}$. The pdf of Z is

$$f_{\mathbf{Z}}(z) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})},$$

where $\Sigma = BB^\top$.

- Z is said to have a multi-variate Gaussian (or normal) distribution with expectation vector $\boldsymbol{\mu}$ and covariance matrix Σ.

A very important property of the normal distribution is for independent

$$X_i \sim \mathsf{N}(\mu_i, \sigma_i^2), \quad i = 1, \ldots, n.$$

Specifically, the random variable

$$Y = a + \sum_{i=1}^{n} b_i\, X_i,$$

is distributed

$$\mathsf{N}\left(a + \sum_{i=1}^{n} b_i\, \mu_i, \sum_{i=1}^{n} b_i^2\, \sigma_i^2\right).$$

Consider the 2-dimensional case with $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$, and

$$B = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_1\sigma_2 & \sigma_2 \end{pmatrix}.$$

The covariance matrix is now

$$\Sigma = BB^\top = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Therefore, the density is

Correction

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{\boldsymbol{Z}}(z_1, z_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$

$$\times \exp\left\{ \frac{-1}{2(1-\rho^2)} \left(\frac{(z_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(z_1-\mu_1)(z_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(z_2-\mu_2)^2}{\sigma_2^2} \right) \right\}$$

This is the pdf of the bi-variate Gaussian distribution, which we encountered earlier.

Example A machine produces ball bearings with a $\mathsf{N}(1, 0.01)$ diameter (cm). The balls are placed on a sieve with a $\mathsf{N}(1.1, 0.04)$ diameter. The diameter of the balls and the sieve are assumed to be independent of each other. What is the probability that a ball will fall through?

Solution

- Let $X \sim \mathsf{N}(1, 0.01)$ and $Y \sim \mathsf{N}(1.1, 0.04)$.
- We need to calculate $P(Y > X) = P(Y - X > 0)$.

- But, $T := Y - X \sim \mathsf{N}(0.1, 0.05)$. Hence,

$$\begin{aligned} P(T > 0) &= P\left(\frac{T - 0.1}{\sqrt{0.05}} > -\frac{0.1}{\sqrt{0.05}}\right) \\ &= P\left(Z > -\frac{0.1}{\sqrt{0.05}}\right) = 1 - \Phi(-0.447) \approx 0.67. \end{aligned}$$

Transformations of R.V. - Motivation

1. Let X_1 is the amount of daily sugar intake of Australians and X_2 the sugar intake of Europeans, and X_3 of Asians. Suppose we are interested in the mean of the daily sugar intake across countries, that is

$$\frac{1}{3}(X_1 + X_2 + X_3)$$

2. Let $X_1, \ldots, X_n$ be the lifetimes of n components in a series system. Then, the lifetime of the system is

$$\min\{X_1, X_2, \ldots, X_n\}$$

3. Let $X_1, \ldots, X_n$ be the risk of a portfolio with n financial assets X_i. A risk averse person will look at

$$\max\{X_1, X_2, \ldots, X_n\}$$

to analyse the risk of the portfolio.

Transformations of R.V. - Properties

- Let X_i be discrete R.V. for $i = 1, \ldots, n$ and $Z = g(X_1, \ldots, X_n)$, then

$$\mathbb{E}[Z] = \sum_{x_1} \ldots \sum_{x_n} g(x_1, \ldots, x_n) P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

- Let X_i be continuous R.V. for $i = 1, \ldots, n$ and $Z = g(X_1, \ldots, X_n)$, then

$$\mathbb{E}[Z] = \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} g(x_1, \ldots, x_n) f(x_1, x_2, \ldots, x_n) \, \mathrm{d}x_1 \ldots \mathrm{d}x_n$$

Descriptive Statistics

- Visualisation of the data.

- Analysis and presentation of characteristics of the data.

Data types

Possible data types:

1. *Continuous quantitative* data $\longrightarrow$ values in continuous range (height, width, length, temperature, humidity, volume, area, and price)

2. *Discrete qualitative* data (*factor / categorical variable*) $\longrightarrow$ values in discrete range (number of family members, gender (male or female), count of objects).

   Discrete Sub-types:

   - *Nominal* factors = variables without order, such as males and females.
   - *Ordinal* factors = variable with a certain order, such as *age group.*

Data configurations

- Many possible data configurations
- Each configuration will consist of continuous and discrete (ordinal and nominal) variables.
- **Major configuration types:**
  - A single sample configuration consists of m scalars:
    $\mathcal{D} = \{x_1, x_2, \dots, x_m\}$.
    Nr of fisherman per day; $m = 365$ and $x_i = 0, 1, \dots$.
  - Two (or more) sets of samples:
    $$\mathcal{D} = \left\{ \left\{x_1^1, \dots, x_{m_1}^1\right\}, \left\{x_1^2, \dots, x_{m_2}^2\right\}, \dots, \left\{x_1^k, \dots, x_{m_k}^k\right\} \right\}.$$
    Nr of fisherman per day in k different regions.
  - Data tuples: $\mathcal{D} = \{(x_{1,1}, x_{1,2}), (x_{2,1}, x_{2,2}), \dots, (x_{m,1}, x_{m,2})\}$.
    $x_{i,1}$ = Nr of fisherman at ith day, $x_{i,2}$ is the number of fishing nets used at day i.
  - Generalization of tuples to vectors:
    $$\mathcal{D} = \{(x_{1,1}, \dots, x_{1,n}), \dots, (x_{m,1}, \dots, x_{m,n})\}$$
    $x_{i,1}$ = Nr of fisherman at ith day, $x_{i,2}$ is the number of fishing nets used at day i, $x_{i,3}$ = Sea-Surface temperature at day i, etc.

1. Data tables

The table **rows** represent observed measurements for *independent* variables (**columns**).

Observ.	variable 1	variable 2	$\cdots$	variable i	$\cdots$	variable n
1	.	.	.	.	.	.
2	.	.	.	.	.	.
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	.	.	.	.	.	.

```r
library(carData)
D <- Arrests
tail(D)

#or alterantive:
library(data.table)
print(data.table(D))
```

```
     released colour year age    sex employed citizen checks
5221      Yes  White 2002  22   Male      Yes     Yes      0
5222      Yes  White 2000  17   Male      Yes     Yes      0
5223      Yes  White 2000  21 Female      Yes     Yes      0
5224      Yes  Black 1999  21 Female      Yes     Yes      1
5225       No  Black 1998  24   Male      Yes     Yes      4
5226      Yes  White 1999  16   Male      Yes     Yes      3
```

Figure: Data on police arrests in Toronto for possession of marijuana.