STATISTICS

Module 2 *of*
MATHEMATICS AND STATISTICS

Marina Santacroce

marina.santacroce@unibocconi.it
Department of Decision Sciences, Bocconi University

L11. Expectation of functions of two (or more) random variables: covariance, correlation, expectation and variance of the sum.

Outline

Expected value of a function of two random variables

We generalize the formula used to compute the expected value of a function of one random variable by considering a function of two random variables.

Prop. X and Y are two discrete random variables with joint mass probability function p, then

$$\mathbb{E}(g(X,Y)) = \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}} g(x,y)p(x,y)$$

If X and Y are continuous with joint density f, then

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)\, f(x,y)\, dxdy$$

Expected value of the sum of random variables

We study the particular case $\mathbb{E}(g(X, Y)) = \mathbb{E}(X + Y)$.
Applying the formula with $g(x, y) = x + y$

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Indeed, if X and Y are continuous:

$$\begin{aligned}
\mathbb{E}(X + Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x + y)\, f(x, y)\, dxdy \\
&= \int_{-\infty}^{+\infty} x \int_{-\infty}^{+\infty} f(x, y)\, dy\, dx + \int_{-\infty}^{+\infty} y \int_{-\infty}^{+\infty} f(x, y)\, dx\, dy \\
&= \int_{-\infty}^{+\infty} x\, f_X(x)\, dx + \int_{-\infty}^{+\infty} y\, f_Y(y)\, dy = \mathbb{E}(X) + \mathbb{E}(Y)
\end{aligned}$$

Important Applications

We generalize the previous formula to consider n random variables

$$\mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n).$$

- If $X \sim \text{Bin}(n, p)$, $X = X_1 + \cdots + X_n$, with $X_i \sim \text{Be}(p)$ (independent) $\Rightarrow$
$$\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = np$$
- If $X \sim \Gamma(n, \lambda)$, $X = X_1 + \cdots + X_n$, $X_i \sim$Exp(λ)(independent), $\Rightarrow$
$$\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = \frac{n}{\lambda}$$

Consider $X_1, \ldots, X_n$ (independent) identically distributed random variables, i.i.d. for short, with common expected value μ.

$X_1, \ldots, X_n$ is said a random sample.

$\boxed{\overline{X} = \frac{1}{n}\sum_{i:1}^{n} X_i}$ represents the sample mean and

$$\mathbb{E}(\overline{X}) = \frac{1}{n}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)) = \mu$$

Independence and expected values

Prop. If the variables X and Y are *independent*, and g, h are two real function, then

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

- $$\begin{aligned}\mathbb{E}(g(X)h(Y)) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x)h(y)\, f(x,y)\, dxdy \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x)h(y)\, f_X(x)f_Y(y)\, dxdy \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x)f_X(x)\, h(y)f_Y(y)\, dxdy \\ &= \int_{-\infty}^{+\infty} g(x)f_X(x)\, dx \int_{-\infty}^{+\infty} h(y)f_Y(y)\, dy \\ &= \mathbb{E}(g(X))\mathbb{E}(h(Y))\end{aligned}$$
- Check the discrete case.

Independence and moment generating functions

If X and Y are two independent random variables the moment generating function of $X + Y$ is

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t), \quad t \in \mathbb{R}$$

In particular we find some notable results:

- $X \sim \text{Binom}(n, p)$ and $Y \sim \text{Binom}(m, p) \Rightarrow X + Y \sim \text{Binom}(n + m, p)$

$$\phi_{X+Y}(t) = \left(1 - p(1 - e^t)\right)^{n+m}$$

- $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2) \Rightarrow X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$

$$\phi_{X+Y}(t) = e^{-(\lambda_1+\lambda_2)(1-e^t)}$$

Independence and moment generating functions: notable results for continuous random variables

- $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\lambda) \;\Rightarrow\; X + Y \sim \text{Gamma}(2, \lambda)$

$$\phi_{X+Y}(t) = \left(\frac{\lambda}{\lambda - t}\right)^2$$

- $X \sim \text{N}(\mu_1, \sigma_1^2)$ and $Y \sim \text{N}(\mu_2, \sigma_2^2) \;\Rightarrow\; X + Y \sim \text{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

$$\phi_{X+Y}(t) = e^{t(\mu_1+\mu_2)+(\sigma_1^2+\sigma_2^2)\frac{t^2}{2}}$$

- $X_1, \ldots, X_n$ i.i.d. and $X_i \sim \text{N}(\mu, \sigma^2) \Rightarrow \bar{X} \sim \text{N}(\mu, \frac{\sigma^2}{n})$

$$\phi_{\bar{X}}(t) = \phi_{X_1+\cdots+X_n}\left(\frac{t}{n}\right) = e^{t\mu+\frac{\sigma^2}{n}\frac{t^2}{2}}$$

Covariance

Def. X and Y are two real random variables.

$$\mathrm{COV}(X,Y) = \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

is said the *covariance* between X and Y.

Moreover, if the $\mathrm{COV}(X,Y)=0$ the random variables are said *uncorrelated*.

Example: $X \sim \mathrm{U}([-1,1])$ and denote by $Y = X^2 \;\Rightarrow\; \mathrm{COV}(X,Y)=0$.

Proposition: If X and Y are independent, then the random variables are uncorrelated, i.e. $\mathrm{COV}(X,Y)=0$.

Indeed,

$$\mathrm{COV}(X,Y) = \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))] = \mathbb{E}[(X-\mathbb{E}(X))]\mathbb{E}[(Y-\mathbb{E}(Y))] = 0$$

Rem. *As the previous example shows, the viceversa is not true in general*!!!

Thus in general uncorrelated random variables are *far from* being independent!

Exceptions are:

- X and Y bivariate normal;
- X and Y two (uncorrelated) Bernoulli.

In those two cases if the variables are uncorrelated, they are independent.

Bivariate normal

The random vector (X, Y) is *bivariate normal* or *bivariate Gaussian*, with parameters $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma_X^2 > 0, \sigma_Y^2 > 0$, $-1 < \rho < +1$ if the joint density, defined for $(x, y) \in \mathbb{R}^2$, is

$$f(x,y) = (2\pi)^{-1} \frac{1}{\sigma_X \sigma_Y \sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)}$$

The parameters represent respectively the expectations and the variances of the marginals X and Y, and the correlation coefficient $\rho(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$.

If X and Y are *uncorrelated*, i.e. $\rho = 0 \Rightarrow X$ and Y *independent*, since

$$f(x,y) = \underbrace{\frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}}}_{f_X(x)} \underbrace{\frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}}}_{f_Y(y)}, \quad (x,y) \in \mathbb{R}^2$$

Properties of the Covariance and Variance of the sum

Verify the following equalities hold true

a) $\mathrm{COV}(X, Y) = \mathrm{COV}(Y, X)$;

b) $\mathrm{COV}(X, X) = Var(X)$;

c) $\mathrm{COV}(X, a) = 0$, for a constant a;

d) $\mathrm{COV}(aX, Y) = a\mathrm{COV}(X, Y)$, $\quad \mathrm{COV}(X, bY) = b\mathrm{COV}(X, Y)$, for two constants a and b;

e) $\mathrm{COV}(aX + bZ + c, Y) = a\mathrm{COV}(X, Y) + b\mathrm{COV}(Z, Y)$ for three constants a, b and c.

The covariance is *bilinear*, i.e. for constants a_i, b_j we have

$$\mathrm{COV}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \mathrm{COV}(X_i, Y_j)$$

From this formula we immediately find

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{COV}(X_i, X_j)$$

Important cases are:

- If $X_1, \ldots X_n$ are (pairwise) *uncorrelated* $\Rightarrow \text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$
- If $X_1, \ldots X_n$ are *independent* $\Rightarrow \text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$;
- $X_1, \ldots, X_n$ are i.i.d. random variables with variance σ^2. Then

$$\text{Var}(\sum_{i:1}^{n} X_i) = n\sigma^2$$

and the variance of the sample mean $\overline{X} = \frac{1}{n}\sum_{i:1}^{n} X_i$ is

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Rem: If $X_1, \ldots X_n$ are i.i.d. and $X_i \sim \text{N}(\mu, \sigma^2) \Rightarrow \overline{X} \sim \text{N}(\mu, \frac{\sigma^2}{n})$

Correlation

Def. X and Y are real random variables and $Var(X)Var(Y) \neq 0$.

$$\rho(X, Y) = \frac{\mathrm{COV}(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

is said the *correlation* of X and Y.

Note that $\rho(X, Y) = 0 \iff X$ and Y are *uncorrelated*.

Prop. $-1 \leq \rho(X, Y) \leq +1$.

- Let $\sigma_X = \sqrt{Var(X)}$ and $\sigma_Y = \sqrt{Var(Y)}$.

$$0 \leq Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = Var\left(\frac{X}{\sigma_X}\right) + Var\left(\frac{Y}{\sigma_Y}\right) + 2\mathrm{COV}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) =$$

$$= 1 + 1 + 2\frac{\mathrm{COV}(X, Y)}{\sigma_X \sigma_Y}) = 2(1 + \rho(X, Y)) \quad \Rightarrow \quad \rho(X, Y) \geq -1$$

Similarly, $Var(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}) \geq 0 \Longrightarrow \rho(X, Y) \leq 1$.

$\rho^2 = 1$

Def. If $\rho(X, Y) = \pm 1$, the random variables are said *perfectly correlated*.
Two random variables are perfectly correlated $\Longleftrightarrow$ Y is a *linear function* of X.

- $\rho(X, Y) = \pm 1$ corresponds respectively to $Var(\frac{X}{\sigma_X} \mp \frac{Y}{\sigma_Y}) = 0$, i.e. to $\mathbb{P}\left(\frac{X}{\sigma_X} \mp \frac{Y}{\sigma_Y} = c\right) = 1 \quad \Rightarrow \quad Y = a + bX$ con $b = \pm\frac{\sigma_Y}{\sigma_X}$ and $a = \mp\sigma_Y c$.

ρ^2 is a measure of the *linear dependence* of X and Y.