NMAI059 – Probability and Statistics 1

Mykhaylo Tyomkyn

Lecture 8 - Random variables: odds and ends.

Definition 1 (Covariance) *Let $X, Y\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ be either both discrete, or both (jointly) continuous. The* covariance *of X and Y, denoted $Cov(X, Y)$ is defined (subject to convergence) as*

$$Cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X)(Y - \mathbb{E}(Y)).$$

Note that $Cov(X, X) = \mathrm{Var}(X)$.

Lemma 1 *$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.*

Proof Using linearity of expectation,

$$\begin{aligned} Cov(X, Y) &= \mathbb{E}((X - \mathbb{E}(X)(Y - \mathbb{E}(Y)) = \mathbb{E}(XY - \mathbb{E}(X)Y - \mathbb{E}(Y)X + \mathbb{E}(X)\mathbb{E}(Y)) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

□

Note that the covariance operator is symmetric, i.e., $Cov(X, Y) = Cov(Y, X)$ and using Lemma 1 one can see that it is also bilinear: $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$.

If $Cov(X, Y) = 0$, that is when $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, we say that X and Y are *uncorrelated*. This is the case when X and Y are independent, but in general uncorrelated random variables need not be independent[1], as we saw in Lecture 7. We say that X and Y are *positively/negatively correlated* if $Cov(X, Y) > 0$ and if $Cov(X, Y) < 0$, respectively.

Theorem 1 (Probabilistic Cauchy-Schwarz inequality) *Let $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then $XY \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, and*

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2).$$

Proof Given X and Y, let $f\colon \mathbb{R} \to \mathbb{R}$ be the real function defined as

$$f(t) = \mathbb{E}((tX + Y)^2).$$

By linearity of expectation we can write it as

$$f(t) = \mathbb{E}(t^2X^2 + 2tXY + Y^2) = t^2\mathbb{E}(X^2) + 2t\mathbb{E}(XY) + \mathbb{E}(Y^2) =: at^2 + bt + c,$$

[1]Intuitively, the covariance indicates how likely both variables will be simultaneously above or below their respective expectations.

where we set $a = \mathbb{E}(X^2), b = 2\mathbb{E}(XY)$ and $c = \mathbb{E}(Y^2)$. Observe now that $(tX + Y)^2 \geq 0$ always, and therefore,

$$0 \leq \mathbb{E}((tX + Y)^2) = f(t),$$

for all values of t. So, the quadratic function $at^2 + bt + c$, takes only non-negative values, meaning its discriminant must be non-positive: $b^2 \leq 4ac$. This translates to

$$4\mathbb{E}(XY)^2 \leq 4\mathbb{E}(X^2)\mathbb{E}(Y^2).$$

$\square$

Definition 2 (Correlation) *Let $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ with $Var(X), Var(Y) \neq 0$. The* (Pearson) correlation *of X and Y is defined as*

$$\rho(X, Y) := \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}.$$

Theorem 2

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

Equivalently,

$$Cov(X, Y)^2 \leq Var(X)Var(Y).$$

Proof Let $X' = X - \mathbb{E}(X)$ and $Y' = Y - \mathbb{E}(Y)$. Then by the properties of $\mathbb{E}$, Var and Cov we have $\mathbb{E}(X') = \mathbb{E}(Y') = 0$, $Var(X') = Var(X)$, $Var(Y') = Var(Y)$ and $Cov(X', Y') = Cov(X, Y)$. So, to prove the statement, it suffices to show that

$$Cov(X', Y')^2 \leq Var(X')Var(Y').$$

Now observe that $Cov(X', Y') = \mathbb{E}(X'Y')$, $Var(X') = \mathbb{E}(X'^2)$ and $Var(Y') = \mathbb{E}(Y'^2)$. Hence, $Cov(X', Y')^2 \leq Var(X')Var(Y')$ is a direct consequence of Theorem 1. $\square$

Exercise 1 *Examine the above proofs to determine the relationship between X and Y, given $\rho(X, Y) = 1$ or $\rho(X, Y) = -1$.*

Theorem 3 (Variance of a sum) *Let $X_1, \dots, X_n \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ be all discrete or (mutually) jointly continuous, and let $X = X_1 + \cdots + X_n$. Then $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, and*

$$Var(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i \neq j} Cov(X_i, X_j).$$

Proof By bilinearity and symmetry of the covariance,

$$\begin{aligned} Var(X) &= Var(X_1 + \cdots + X_n) = Cov(X_1 + \cdots + X_n, X_1 + \cdots + X_n) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j) = \sum_{i=1}^{n} Cov(X_i, X_i) + 2\sum_{i \neq j} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Var(X_i) + 2\sum_{i \neq j} Cov(X_i, X_j). \end{aligned}$$

$\square$

Theorem 4 (Convolution formula for continuous rv's) *Let $X, Y \in (\Omega, \mathcal{F}, \mathbb{P})$ be jointly continuous. Then $Z = X + Y$ is continuous, with the probability density function*

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)\, dx = \int_{-\infty}^{\infty} f_{X,Y}(z-y, y)\, dy.$$

In particular, if X and Y are independent we have

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy.$$

In that case we write $f_Z = f_X * f_Y (= f_Y * f_X)$ and call f_Z the (continuous) convolution *of f_X and f_Y.*

We skip the proof, but notice that the above is analogous to the discrete convolution formula.

Example 1 (Convolution for the normal distribution) *Let $X, Y \sim \mathcal{N}(0,1)$ be independent. Then, for $Z = X + Y$ we have*

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-x)^2}{2}}\, dx \\ &= \frac{1}{2\pi} e^{-\frac{z^2}{4}} \int_{-\infty}^{\infty} e^{-x^2 + zx - \frac{z^2}{4}}\, dx = \frac{1}{2\pi} e^{-\frac{z^2}{4}} \int_{-\infty}^{\infty} e^{-(x-\frac{z}{2})^2}\, dx \\ &= \frac{1}{2\pi} e^{-\frac{z^2}{4}} \cdot \sqrt{\pi} = \frac{1}{2\sqrt{\pi}} e^{-\frac{z^2}{4}}, \end{aligned}$$

and thus $Z \sim \mathcal{N}(0,2)$. This holds for the normal distribution more generally:[2]

$$\mathcal{N}(\mu_1, \nu_1) * \mathcal{N}(\mu_2, \nu_2) = \mathcal{N}(\mu_1 + \mu_2, \nu_1 + \nu_2).$$

Exercise 2 (Convolution for the Gamma distribution) *Recall that the $Gamma(r, \alpha)$ distribution is defined on $[0, \infty)$ via the pdf*

$$\gamma_{r,\alpha} = \frac{1}{\Gamma(r)} \alpha^r x^{r-1} e^{-\alpha x}, \quad \textit{where} \;\; \Gamma(r) = \int_0^{\infty} x^{r-1} e^{-x}\, dx.$$

Prove that

$$\Gamma(r, \alpha) * \Gamma(s, \alpha) = \Gamma(r+s, \alpha).$$

Recall also that $Gamma(1, \alpha)$ is the exponential distribution $Exp(\alpha)$. Thus, for a positive integer r, the sum of r independent $Exp(\alpha)$-variables is $Gamma(r, \alpha)$-distributed.[3]

Lastly, let us introduce two fundamental probabilistic inequalities, which apply to discrete and continuous random variables alike.

[2] The behaviour of the parameters should not come as a surprise, given that μ and ν are the mean and variance of $\mathcal{N}(\mu, \nu)$, respectively.

[3] An interpretation: when the time until the next meteorite impact is $Exp(\alpha)$-distributed, the time for the r-th meteorite to arrive is distributed with $Gamma(r, \alpha)$, while the number of meteorites in a time interval is Poisson-distributed.

Theorem 5 (Markov's inequality) *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ satisfy $X \geq 0$ a.s. Then for every $t > 0$ we have*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

This has a very simple interpretation: "if the average age in a room is 20, at most 50% of the people in it can be above 40."

Proof Let $Y = t \cdot \mathbb{1}_{\{X \geq t\}}$. Then $X \geq Y$, so $Y \in L^1$ and

$$\mathbb{E}(X) \geq \mathbb{E}(Y) = t \cdot \mathbb{P}(X \geq t).$$

□

Theorem 6 (Chebyshev's inequality) *Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ and $c > 0$. Then*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq c) \leq \frac{Var(X)}{c^2}.$$

Proof Applying Markov's inequality (Theorem 5) with $Z = (X - \mathbb{E}(X))^2$ and $t = c^2$ gives

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq c) = \mathbb{P}(Z \geq c^2) \leq \frac{\mathbb{E}(Z)}{c^2} = \frac{Var(X)}{c^2}.$$

□

Example 2 *Consider 100 independent (fair) coin tosses. How likely are we to see at least 60 'heads'? That is, for $X \sim Bin(100, 1/2)$ estimate $\mathbb{P}(X \geq 60)$. By symmetry and Chebyshev's inequality (as $\mathbb{E}(X) = 50$ and $Var(X) = 25$) we have*

$$\mathbb{P}(X \geq 60) = \mathbb{P}(X - 50 \geq 10) = \frac{1}{2}\mathbb{P}(|X - 50| \geq 10) \leq \frac{1}{2}\frac{\mathrm{Var}(X)}{10^2} = \frac{1}{8} = 12.5\%.$$

This is not bad for a first estimate, however, we will shortly see that the actual probability to get 60 or more heads is much smaller, namely about 2.5%.