NMAI059 – Probability and Statistics 1

Mykhaylo Tyomkyn

Lecture 5 - Variance, joint distribution, independence of discrete rv's.

Two random variables can have the same mean, but very different deviations from it. To quantify how much X deviates from $\mu = \mathbb{E}(X)$ 'on average' we introduce the *variance*.

Definition 1 (Variance) *Let X be a discrete random variable with $\mathbb{E}(X) = \mu$. The* variance *of X is*

$$\text{Var}(X) := \mathbb{E}((X - \mu)^2) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

Remark 1 *The variance is defined only for variables with finite mean. However, even if $\mathbb{E}(X)$ is finite, the variance may not exist (be finite). On the other hand, since $(X - \mu)^2 \geq 0$ a.s., if the variance does not exist, we can informally write* $\text{Var}(X) = +\infty$.

Remark 2 *Just like the expectation, the variance is a function of the distribution ('doesn't care where the variable is coming from').*

Example 1 *Let $X \sim Bern(p)$, so that $\mathbb{E}(X) = p$. We have*

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}((X - p)^2) = (1 - p)^2 \cdot p + p^2 \cdot (1 - p) = p(1 - p)(p + (1 - p)) \\ &= p(1 - p).\end{aligned}$$

Example 2 *Let X be the outcome of a fair die throw, i.e., $X \sim Unif\{1, 2, 3, 4, 5, 6\}$. Then $\mathbb{E}(X) = 3.5$ and*

$$\text{Var}(X) = \frac{2.5^2 + 1.5^2 + 0.5^2 + 0.5^2 + 1.5^2 + 2.5^2}{6} = \frac{35}{12}.$$

Lemma 1 $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

Proof With $\mu = \mathbb{E}(X)$, applying linearity of expectation, we obtain

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}((X - \mu)^2) = \mathbb{E}(X^2 - 2\mu X + \mu^2) = \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 \\ &= \mathbb{E}(X^2) - 2\mu \cdot \mu + \mu^2 = \mathbb{E}(X^2) - \mu^2.\end{aligned}$$

□

Corollary 1 *For any discrete random variable X with finite mean and variance, we have*

$$\mathbb{E}(X^2) \geq \mathbb{E}(X)^2.$$

Theorem 1 *For any discrete random variable X with finite variance and $a, b \in \mathbb{R}$ we have*

$$\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X).$$

Proof Let $\mu = \mathbb{E}(X)$ and $Y := aX + b$. We know that $\mathbb{E}(Y) = a\mu + b$. Therefore,

$$\begin{aligned}\operatorname{Var}(Y) &= \mathbb{E}((Y - \mathbb{E}(Y))^2) = \mathbb{E}((aX + b - (a\mu + b))^2) = \mathbb{E}((a(X - \mu))^2) \\ &= a^2\mathbb{E}((X - \mu)^2) = a^2 \operatorname{Var}(X).\end{aligned}$$

□

Definition 2 (Standard deviation) *Let X be a discrete random variable with finite variance. The* standard deviation *of X is $\sigma = \sigma(X) := \sqrt{\operatorname{Var}(X)}$.*

The quantities $\mu(X) = \mathbb{E}(X)$ and $\sigma(X) = \sqrt{\operatorname{Var}(X)}$ are the most important characteristics of a random variable. But why don't we measure the deviation from the mean 'linearly', i.e. consider $\mathbb{E}(|X - \mu|)$ instead of $\sigma = \sqrt{\mathbb{E}((X - \mu)^2)}$? The reason is that the former is more difficult to handle computationally, but also that Var and σ are of fundamental importance in the theorems to come (Laws of large numbers, Central limit theorem).

Next, we would like to study interactions between random variables. So, suppose $X, Y\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ are discrete. Note that (X, Y) takes countably many values.

Definition 3 (Joint distribution) *For discrete random variables $X, Y\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$, define their* joint probability mass function $p_{X,Y}\colon \mathbb{R}^2 \to [0, 1]$,

$$p_{X,Y}(x, y) = \mathbb{P}(X = x \wedge Y = y).$$

[1] *Similarly, for n discrete random variables $X_1, \dots, X_n\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ their joint probability mass function is $p_{X_1,\dots,X_n}\colon \mathbb{R}^n \to [0, 1]$,*

$$p_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1 \wedge \cdots \wedge X_n = x_n).$$

When $\operatorname{Im} X$ and $\operatorname{Im} Y$ are finite, their joint distribution can be visualized by a two-dimensional table/matrix.

Example 3 *We toss a fair coin twice, counting 1 and 0 for 'heads' and 'tails', respectively. Let X be the sum of the two outcomes, and Y be the product. Then the values of $p_{X,Y}$ are as follows.*

Y \ X	*0*	*1*	*2*
0	*1/4*	*1/2*	*0*
1	*0*	*0*	*1/4*

[1] $\mathbb{P}(X = x \wedge Y = y)$ is shorthand for $\mathbb{P}(\{\omega \in \Omega\colon X(\omega) = x, Y(\omega) = y\})$. We shall also write $\mathbb{P}(X = x, Y = y)$ to denote the same.

Note that the row-sums and column-sums in the table above are precisely the values of p_Y and p_X, respectively. We say that p_X and p_Y are the *marginals* of $p_{X,Y}$. If, on the other hand, we fix a column and rescale the values in it to make their sum 1, we obtain the *conditionals*

$$p_{Y|X}(y \mid x) := \mathbb{P}(Y = y \mid X = x),$$

and symmetrically for rows and $p_{X|Y}$. For instance, in the above example we have $p_{X|Y}(0 \mid 0) = 1/3$. A particularly important case of a joint distribution is when X and Y are independent.

Definition 4 (Independence) *Two discrete random variables $X, Y\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ are* independent *if for every $(x, y) \in \operatorname{Im} X \times \operatorname{Im} Y$ the events $\{X = x\}$ and $\{Y = y\}$ are independent. In other words, when for all x and y we*[2] *have*

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

Similarly, $X_1, \ldots, X_n\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ are (mutually) independent if for all $(x_1, \ldots, x_n) \in \operatorname{Im} X_1 \times \cdots \times \operatorname{Im} X_n$ the events $\{X_1 = x_1\}, \ldots, \{X_n = x_n\}$ are independent.

Exercise 1 *Show that if X and Y are independent then for any $I, J \subseteq \mathbb{R}$ we have*

$$\mathbb{P}(X \in I, Y \in J) = \mathbb{P}(X \in I)\mathbb{P}(Y \in J).$$

Theorem 2 *For any two independent discrete random variables $X, Y\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ with finite means we have*

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

Proof

$$\begin{aligned}
\mathbb{E}(X)\mathbb{E}(Y) &= \sum_{x \in \operatorname{Im} X} x p_X(x) \sum_{y \in \operatorname{Im} Y} y p_Y(y) = \sum_{(x,y) \in \operatorname{Im} X \times \operatorname{Im} Y} xy p_X(x) p_Y(y) \\
&\overset{Indep}{=} \sum_{(x,y) \in \operatorname{Im} X \times \operatorname{Im} Y} xy p_{X,Y}(x, y) = \sum_{t \in \operatorname{Im} XY} t \sum_{\substack{(x,y) \in \operatorname{Im} X \times \operatorname{Im} Y \\ xy = t}} p_{X,Y}(x, y) \\
&= \sum_{t \in \operatorname{Im} XY} t\mathbb{P}(XY = t) = \mathbb{E}(XY).
\end{aligned}$$

□

Theorem 3 *For any two independent discrete random variables $X, Y\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ with finite variances we have*

$$\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

Proof Using Lemma 1, linearity of expectation and Theorem 2, we obtain

$$\begin{aligned}
\operatorname{Var}(X + Y) &= \mathbb{E}((X + Y)^2) - \mathbb{E}(X + Y)^2 = \mathbb{E}(X^2 + 2XY + Y^2) - (\mathbb{E}(X) + E(Y))^2 \\
&= \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - \mathbb{E}(X)^2 + 2\mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)^2 \\
&= \mathbb{E}(X^2) - \mathbb{E}(X)^2 + \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \operatorname{Var}(X) + \operatorname{Var}(Y).
\end{aligned}$$

[2] In expanded form: $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$.

□

By extension, for n independent random variables $X_1, \dots, X_n$ with finite means/variances we have $\mathbb{E}(X_1 \cdots X_n) = \mathbb{E}(X_1) \cdots \mathbb{E}(X_n)$ and $\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$.

Example 4 *Let $X \sim Bin(n,p)$. Then $X = X_1 + \cdots + X_n$, where $X_1, \dots, X_n$ are independent $Bern(p)$ random variables. Therefore,*

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = np(1-p).$$

And so, $\sigma(X) = \sqrt{np(1-p)}$.

Theorem 4 (LOTUS for joint distributions) *Let $X, Y\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ be discrete random variables and $g\colon \mathbb{R} \to \mathbb{R}$ be a function. Then $Z = g(X,Y)$ satisfies (assuming convergence)*

$$\mathbb{E}(Z) = \sum_{(x,y)\in \operatorname{Im} X \times \operatorname{Im} Y} g(x,y) p_{X,Y}(x,y).$$

Proof The proof essentially mirrors that of Theorem 2:

$$\sum_{(x,y)\in \operatorname{Im} X \times \operatorname{Im} Y} g(x,y) p_{X,Y}(x,y) = \sum_{z \in \operatorname{Im} Z} z \sum_{\substack{(x,y)\in \operatorname{Im} X \times \operatorname{Im} Y \\ g(x,y)=z}} \mathbb{P}(X = x, Y = y) = \sum_{z \in \operatorname{Im} Z} z \mathbb{P}(Z = z) = \mathbb{E}(Z).$$

□

Theorem 5 (Convolution formula for discrete rv's) *Let $X, Y\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ be discrete. Then their sum $Z = X + Y$ satisfies for all $z \in \operatorname{Im} Z$*

$$p_Z(z) = \sum_{x \in \operatorname{Im} X} p_{X,Y}(x, z-x) = \sum_{y \in \operatorname{Im} Y} p_{X,Y}(z-y, y).$$

If X and Y are independent then for all $z \in \operatorname{Im} Z$

$$p_Z(z) = \sum_{x \in \operatorname{Im} X} p_X(x) p_Y(z-x) = \sum_{y \in \operatorname{Im} Y} p_X(z-y) p_Y(y). \tag{1}$$

The proof is a direct application of the law of total probability and the definition of independence. The convolution formula provides another tool for constructing new probability measures. In notation, with p_Z as in (1) (so Z corresponds to the sum of independent X and Y) we write $p_Z = p_X * p_Y$ and call p_Z the *convolution* of p_X and p_Y (note that $p_X * p_Y = p_Y * p_X$). A remarkable feature of the convolution operator is that it preserves many 'typical' distributions.

Example 5 *Let $X \sim Bin(n,p)$ and $Y \sim Bin(m,p)$ be independent. Then $Z \sim Bin(n+m,p)$ (think of independent coin tosses). Hence,*

$$Bin(n+m,p) = Bin(n,p) * Bin(m,p).$$

Exercise 2 *Verify the above by a calculation, using the convolution formula.*