Applied Statistics

for Computer Science BSc, Exam

Probability Theory and Mathematical Statistics

for Computer Science Engineering BSc, Term grade

István Fazekas
University of Debrecen

2020/21 fall

This work was supported by the construction EFOP-3.4.3-16-2016-00021. The project was supported by the European Union, co-financed by the European Social Fund.

Main topics

1. Probability theory
2. Statistics

Mathematical tools: combinatorics, calculus
Computer tool: Matlab
Book:
Yates, Goodman:
Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers

Lecture 5

Joint distribution, correlation

Joint distribution of two discrete random variables

Let X and Y be discrete random variables.
Let the range of X be $x_1, x_2, \ldots,$
the range of Y be $y_1, y_2, \ldots.$
Then the joint distribution of X and Y is

$$p_{ij} = P(X = x_i, Y = y_j), \quad i, j = 1, 2, \ldots.$$

We see that these numbers are non-negative and their sum is equal to 1, that is

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} p_{ij} = 1$$

Marginal distributions

The distribution of X is

$$p_{i\cdot} = P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

and the distribution of Y is

$$p_{\cdot j} = P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

These are called the two marginal distributions.
These numbers are non-negative and
$\sum_{i=1}^{\infty} p_{i\cdot} = 1$,
$\sum_{j=1}^{\infty} p_{\cdot j} = 1$,

The joint distribution table (contingency table)

$$
\begin{array}{c|c|c|c|c}
X \backslash Y & y_1 & y_2 & \dots & \sum \\
\hline
x_1 & p_{11} & p_{12} & \dots & p_{1\cdot} \\
\hline
x_2 & p_{21} & p_{22} & \dots & p_{2\cdot} \\
\hline
\vdots & \vdots & \vdots & & \vdots \\
\hline
\sum & p_{\cdot 1} & p_{\cdot 2} & \dots & 1
\end{array}
\tag{1}
$$

On the margins of the table we can find the row and the column sums.

They are the marginal distributions.

Example 1

Roll two dice. Let X be the number shown by the first die, and Y be the number shown by the second die.
Their joint distribution is

$X \backslash Y$	1	2	$\dots$	$\sum$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\dots$	$\frac{1}{6}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\dots$	$\frac{1}{6}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\sum$	$\frac{1}{6}$	$\frac{1}{6}$	$\dots$	1

Example 2

Roll two dice. Let X be the number shown by the first die, and Y be again the number shown by the first die.
Their joint distribution is

$X \backslash Y$	1	2	$\ldots$	$\sum$
1	$\frac{1}{6}$	0	$\ldots$	$\frac{1}{6}$
2	0	$\frac{1}{6}$	$\ldots$	$\frac{1}{6}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\sum$	$\frac{1}{6}$	$\frac{1}{6}$	$\ldots$	1

Remark. We see that the joint distribution determine the marginal distributions, but NOT vice versa.

Exercise 1

Roll two dice. Let X be the number shown by the first die, and Y be the number shown by the second die. Find the distribution of

$$\max\{X, Y\}$$

Hint. Use an appropriate modification of the joint distribution table.
Homework.
Find the distribution of

$$\min\{X, Y\}$$

Solution of exercise 1.

$X \backslash Y$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

Let $Z = \max\{X, Y\}$. Then

$$P(Z=1) = \frac{1}{36},\ P(Z=2) = \frac{3}{36},\ P(Z=3) = \frac{5}{36},$$

$$P(Z=4) = \frac{7}{36},\ P(Z=5) = \frac{9}{36},\ P(Z=6) = \frac{11}{36}$$

Exercise 2.

Find the expectation of $Z = \max\{X, Y\}$.
Solution.

$$EZ = 1\frac{1}{36} + 2\frac{3}{36} + 3\frac{5}{36} + 4\frac{7}{36} + 5\frac{9}{36} + 6\frac{11}{36} =$$

$$= \frac{161}{36}$$

Homework. Find the expectation of $Z = \min\{X, Y\}$.

Exercise 3

Roll two dice. Let X be the number shown by the first die, and Y be the number shown by the second die. Find the distribution of

$$X + Y$$

Hint. Use an appropriate modification of the joint distribution table.

Homework.

Roll two dice. Let X be the number shown by the first die, and Y be the number shown by the second die. Find the distribution of

$$X \cdot Y$$

Solution of exercise 3. Let $Z = X + Y$.

$X \backslash Z$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$$P(Z=2) = \frac{1}{36},\ P(Z=3) = \frac{2}{36},\ P(Z=4) = \frac{3}{36},$$
$$P(Z=5) = \frac{4}{36},\ P(Z=6) = \frac{5}{36},\ P(Z=7) = \frac{6}{36},$$
$$P(Z=8) = \frac{5}{36},\ P(Z=9) = \frac{4}{36},\ P(Z=10) = \frac{3}{36},$$
$$P(Z=11) = \frac{2}{36},\ P(Z=12) = \frac{1}{36}$$

Homework

Roll two dice. Let X be the number shown by the first die, and Y be the number shown by the second die.

1. Find the expectation of $X + Y$.

1. Find the expectation of $X \cdot Y$.

Independence of discrete random variables

X and Y are called independent if

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j), \quad i, j = 1, 2, \dots .$$

That is

$$p_{ij} = p_{i\cdot} p_{\cdot j}, \quad \forall i, j.$$

Independence of discrete random variables

$X_1, X_2, \ldots, X_n$ are called pairwise independent if any two of them are independent.
$X_1, X_2, \ldots, X_n$ are called (totally) independent if

$$P(X_1 = x_{k_1}, X_2 = x_{k_2}, \ldots, X_n = x_{k_n}) =$$
$$= P(X_1 = x_{k_1})P(X_2 = x_{k_2}) \cdots P(X_n = x_{k_n})$$

is satisfied for any $x_{k_1}, \ldots, x_{k_n}$.
Remark. Total independence implies pairwise independence, but not vice versa.

A theorem for the product

If X and Y are independent random variables with $E|X| < \infty$ and $E|Y| < \infty$, then

$$E(X \cdot Y) = EX \cdot EY$$

Proof.

$$E(XY) = \sum_k \sum_l x_k y_l \; P(X = x_k, Y = y_l).$$

Because of independence

$$E(XY) = \sum_k \sum_l x_k y_l \; P(X = x_k) \; P(Y = y_l) =$$

$$= \sum_k x_k P(X = x_k) \sum_l y_l P(Y = y_l) = EX \cdot EY.$$

Convolution

Let X and Y be independent integer valued random variables.
Let $P(X = n) = p_n$, $P(Y = m) = q_m$, be their distributions, where $n, m = 0, \pm 1, \pm 2, \ldots$.
Let $Z = X + Y$. Then

$$s_k = P(Z = k) = \sum_{j=-\infty}^{\infty} p_j q_{k-j}, \quad k = 0, \pm 1, \pm 2, \ldots .$$

If X and Y have only non-negative integer values, then

$$s_k = P(Z = k) = \sum_{j=0}^{k} p_j q_{k-j}, \quad k = 0, 1, 2, \ldots .$$

Convolution of binomial random variables

Let X and Y be independent binomial random variables with parameters n_1, p, resp. n_2, p, that is

$$P(X = j) = \binom{n_1}{j} p^j (1-p)^{n_1 - j}, \quad j = 0, 1, \ldots, n_1,$$

$$P(Y = l) = \binom{n_2}{l} p^l (1-p)^{n_2 - l}, \quad l = 0, 1, \ldots, n_2.$$

Then for $Z = X + Y$

$$P(Z = k) = \sum_j P(X = j) P(Y = k - j) =$$

$$= \sum_j \binom{n_1}{j} \binom{n_2}{k-j} p^k (1-p)^{n_1 + n_2 - k} = \binom{n}{k} p^k (1-p)^{n-k},$$

where $n = n_1 + n_2$.

So the convolution is again a binomial distribution.

Convolution of binomial random variables

Above we used that

$$\sum_j \binom{n_1}{j}\binom{n_2}{k-j} = \binom{n_1+n_2}{k}$$

where the summation is applied for those values of j, for which $0 \le k-j \le n_2$ és $0 \le j \le n_1$.

Homework. Show that the sum of n independent p-parameter Bernoulli random variables has binomial distribution with parameters n, p.

Convolution

Homework. Show that the convolution of a Poisson distribution with parameter λ and a Poisson distribution with parameter μ is again a Poisson distribution with parameter $(\lambda + \mu)$.

Covariance

Definition. Let X and Y be random variables, $\mathrm{Var} X < \infty$, $\mathrm{Var} Y < \infty$.
Notation: $EX = m_X$, $EY = m_Y$.
The covariance of X and Y is

$$\boxed{\mathrm{cov}(X, Y) = E[(X - m_X)(Y - m_Y)]}$$

Remark. $\mathrm{Var} X = \mathrm{cov}(X, X)$
Remark.

$$\mathrm{cov}(X, Y) = E(XY) - m_X m_Y$$

Calculation of the covariance

Let

$$p_{ij} = P(X = x_i, Y = y_j), \quad i, j = 1, 2, \ldots$$

be the joint distribution of X and Y. Then

$$\mathrm{cov}(X, Y) = \sum_i \sum_j (x_i - m_X)(y_j - m_Y) p_{ij},$$

and

$$\mathrm{cov}(X, Y) = \sum_i \sum_j x_i y_j p_{ij} - m_X \cdot m_Y$$

Independence and covariance

Theorem. Let $\mathrm{Var}(X) < \infty$, $\mathrm{Var}(Y) < \infty$.
If X and Y are independent, then $\mathrm{cov}(\mathrm{X}, \mathrm{Y}) = 0$,
but not vice versa.
Proof. By independence
$E(XY) = EX \cdot EY$.
So
$\mathrm{cov}(X, Y) = E(XY) - EX \cdot EY = 0$.

Next example shows, that $\mathrm{cov}(\mathrm{X}, \mathrm{Y}) = 0$ does not imply independence.

Independence and covariance

Example. Let the range of X and Y be $-1, 0, +1$. Let their joint distribution be

$$P(X=0, Y=-1) = P(X=0, Y=+1) =$$

$$= P(X=-1, Y=0) = P(X=+1, Y=0) = 1/4.$$

Their joint distribution table is

$X \backslash Y$	-1	0	1	$\sum$
-1	0	1/4	0	1/4
0	1/4	0	1/4	1/2
1	0	1/4	0	1/4
$\sum$	1/4	1/2	1/4	1

Independence and covariance

Example (cont.).
Then $EX = EY = (-1)\cdot 1/4 + 0\cdot 1/2 + 1\cdot 1/4 = 0.$

$$E(XY) = (-1)\cdot(-1)\cdot 0 + (-1)\cdot 0\cdot 1/4 + (-1)\cdot 1\cdot 0 +$$

$$+0\cdot(-1)\cdot 1/4 + 0\cdot 0\cdot 0 + 0\cdot 1\cdot 1/4 + 1\cdot(-1)\cdot 0 + 1\cdot 0\cdot 1/4 + 1\cdot 1\cdot 0 = 0$$

So $\mathrm{cov}(X, Y) = 0$.
But

$$P(X = 0, Y = 0) = 0 \neq \frac{1}{2}\cdot\frac{1}{2} = P(X = 0)\cdot P(Y = 0),$$

so they are not independent.

Independence and covariance

Example. We win 1 EUR if a coin shows H, and pay 1 EUR if it shows T. Toss two coins. Let X be our win on the first coin and let Y our win on the second coin. Calculate the covariance of $X+Y$ and $X-Y$.

Solution

$$EX = EY = \frac{1}{2}1 + \frac{1}{2}(-1) = 0.$$

So

$$E(X+Y) = 0, \quad E(X-Y) = 0.$$

Moreover

$$E(X+Y)(X-Y) = EX^2 - EY^2 = 1 - 1 = 0.$$

Then

$$\operatorname{cov}[(X+Y)(X-Y)] = E(X+Y)(X-Y) - E(X+Y)E(X-Y) = 0 - 0 = 0.$$

Calculation of the covariance. Exercise

Roll two dice. Let X be the number shown by the first die, and Y be the number shown by the second die. Calculate the covariance of X and $Z = \max\{X, Y\}$.

Solution.

$X \backslash Z$	1	2	3	4	5	6	$\sum$
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
2	0	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
3	0	0	$\frac{3}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
4	0	0	0	$\frac{4}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
5	0	0	0	0	$\frac{5}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
6	0	0	0	0	0	$\frac{6}{36}$	$\frac{1}{6}$
$\sum$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	1

Calculation of the covariance. Exercise (cont.)

$$E(XZ) = 1\left[1+2+3+4+5+6\right]\frac{1}{36}+$$

$$+2\left[2\cdot 2+(3+4+5+6)\right]\frac{1}{36}+3\left[3\cdot 3+(4+5+6)\right]\frac{1}{36}+$$

$$+4\left[4\cdot 4+(5+6)\right]\frac{1}{36}+5\left[5\cdot 5+6\right]\frac{1}{36}+6\cdot 6\cdot 6\frac{1}{36}=$$

$$=\frac{616}{36}$$

Using Exercise 1,

$$\operatorname{cov}(X,Z) = E(XZ) - EX\cdot EZ = \frac{616}{36} - \frac{7}{2}\frac{161}{36} = \frac{105}{72}.$$

Properties of the covariance

The covariance is similar to the inner product.
Theorem. The covariance is symmetric, that is

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

The covariance is bilinear, that is

$$\operatorname{cov}(a_1 X_1 + a_2 X_2, Y) = a_1 \operatorname{cov}(X_1, Y) + a_2 \operatorname{cov}(X_2, Y).$$

Proof. Use the definition of the covariance.

The variance of a sum

Theorem.

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

If X and Y are independent, then

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

Proof. Use the previous theorems.

Correlation coefficient

The correlation coefficient is similar to the cosine of an angle.

Definition. Let $0 < \mathrm{Var}(X) < \infty$, $0 < \mathrm{Var}(Y) < \infty$.
The correlation coefficient of X and Y is

$$\boxed{\mathsf{corr}(X, Y) = \frac{\mathsf{cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}}$$

If $\mathsf{corr}(X, Y) = 0$, then we say that X and Y are uncorrelated.

If X and Y are independent, then X and Y are uncorrelated but not vice versa.

Correlation coefficient

Theorem. a) The value of $\mathrm{corr}(X, Y)$ always lies between -1 and $+1$.
b) $\mathrm{corr}(X, Y) = 1$ if and only if, when

$$Y = aX + b$$

for some numbers a and b with $a > 0$.
c) $\mathrm{corr}(X, Y) = -1$ if and only if, when

$$Y = aX + b$$

for some numbers a and b with $a < 0$.

Calculation of the covariance

Roll two dice. Let X be the number shown by the first die, and Y be the number shown by the second die.
Calculate $\operatorname{corr}(X, X+Y)$.
Solution.

$$EX = \frac{7}{2}, \;\; EX^2 = \frac{91}{6}, \;\; \operatorname{Var}(X) = \frac{35}{12}$$

Using independence

$$E(X+Y) = 2\frac{7}{2} = 7, \quad \operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) = \frac{70}{12}$$

and

$$\operatorname{cov}(X, X+Y) = \operatorname{cov}(X, X) + \operatorname{cov}(X, Y) = \operatorname{Var}(X) + 0 = \frac{35}{12}$$

Therefore

$$\operatorname{corr}(X, X+Y) = \frac{\operatorname{cov}(X, X+Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(X+Y)}} = \frac{\frac{35}{12}}{\sqrt{\frac{35}{12}}\sqrt{\frac{70}{12}}} = \frac{1}{\sqrt{2}}$$

Calculation of the covariance

Let X and Y be independent and identically distributed random variables. Assume that $0 < \mathrm{Var}(X) < \infty$.
Show that

$$\mathrm{corr}(X, X + Y) = \frac{1}{\sqrt{2}}$$