Lecture 7. Conditional Distributions with Applications

Igor Rychlik

Chalmers
Department of Mathematical Sciences

Probability, Statistics and Risk, MVE300 • Chalmers • April 2014

Random variables:

- Joint distribution of $X; Y$.
- Dependent random variables:
  - correlated normal variables,
  - expectation of h(X,Y), covariance.
- Conditional pdf and cdf.
- Law of total probabilities.
- Bayes formula.

Joint probability distribution function of X, Y:

Example Experiment: select at random a person in the classroom and measure his (her) length x [m] and weight y [kg]. Such an experiment results in two r.v. X; Y .

- For independent X, Y any statement A about X is independent of a statement B about Y, i.e. $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$
- For example let $A = "X \leq x"$ and $B = "X \leq x"$ then

$$\mathrm{P}(X \leq x \text{ and } Y \leq y) = \mathrm{P}(X \leq x) \cdot \mathrm{P}(Y \leq y)$$

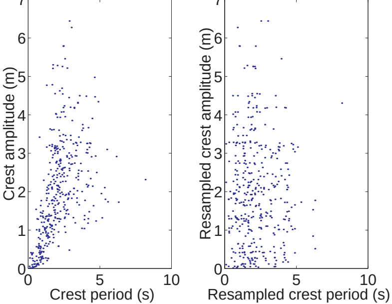
- Joint distribution of X; Y is a function

$$F_{XY}(x, y) = \mathrm{P}(X \leq x \text{ and } Y \leq y) = \mathrm{P}(X \leq x,\ Y \leq y)^{1}.$$

- X, Y are independent if and only if

$$F_{XY}(x, y) = F_X(x)F_Y(y) \tag{1}$$

[1] Similarly as for one dimensional case, the probability of any statement about the random variables X, Y is computable (at least in theory) when $F_{XY}(x, y)$ is known.



Wave data from North Sea. Scatter plot of crest period and crest amplitude (left); crest period T_c and crest amplitude A_c, *resampled* from original data (right).

Are T_c, A_c independent?

Very unlikely!

There were $n = 199$ waves measured. In order to get independent observations of T_c, A_c we choose 100 waves at random out of 199. Next we split the data in four groups defined by events $A = T_c \leq 1$, $B = A_c \leq 2$ and let $p = \mathrm{P}(A)$ and $q = \mathrm{P}(B)$. Data:

	B	B^c
A	16	2
A^c	49	33

[2]

[2] If T_c and A_c are independent then probabilities of four events AB, A^cB, AB^c and A^cB^c are defined by parameters p, q. The estimates are $p^* = 0.18$, $q^* = 0.65$. Now we can use χ^2 test to test hypothesis of independence, see blackboard.

$Q = 5.51$, $f = 4 - 2 - 1$.

n	α											
	0.9995	0.999	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001	0.0005
1	—	—	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$	$< 10^{-2}$	3.841	5.024	6.635	7.879	10.83	12.12
2	$< 10^{-2}$	$< 10^{-2}$	0.0100	0.0201	0.0506	0.1026	5.991	7.378	9.210	10.60	13.82	15.20
3	0.0153	0.0240	0.0717	0.1148	0.2158	0.3518	7.815	9.348	11.34	12.84	16.27	17.73
4	0.0639	0.0908	0.2070	0.2971	0.4844	0.7107	9.488	11.14	13.28	14.86	18.47	20.00
5	0.1581	0.2102	0.4117	0.5543	0.8312	1.145	11.07	12.83	15.09	16.75	20.52	22.11
6	0.2994	0.3811	0.6757	0.8721	1.237	1.635	12.59	14.45	16.81	18.55	22.46	24.10
7	0.4849	0.5985	0.9893	1.239	1.690	2.167	14.07	16.01	18.48	20.28	24.32	26.02
8	0.7104	0.8571	1.344	1.646	2.180	2.733	15.51	17.53	20.09	21.95	26.12	27.87
9	0.9717	1.152	1.735	2.088	2.700	3.325	16.92	19.02	21.67	23.59	27.88	29.67
10	1.265	1.479	2.156	2.558	3.247	3.940	18.31	20.48	23.21	25.19	29.59	31.42
11	1.587	1.834	2.603	3.053	3.816	4.575	19.68	21.92	24.72	26.76	31.26	33.14
12	1.934	2.214	3.074	3.571	4.404	5.226	21.03	23.34	26.22	28.30	32.91	34.82
13	2.305	2.617	3.565	4.107	5.009	5.892	22.36	24.74	27.69	29.82	34.53	36.48
14	2.697	3.041	4.075	4.660	5.629	6.571	23.68	26.12	29.14	31.32	36.12	38.11
15	3.108	3.483	4.601	5.229	6.262	7.261	25.00	27.49	30.58	32.80	37.70	39.72
16	3.536	3.942	5.142	5.812	6.908	7.962	26.30	28.85	32.00	34.27	39.25	41.31
17	3.980	4.416	5.697	6.408	7.564	8.672	27.59	30.19	33.41	35.72	40.79	42.88
18	4.439	4.905	6.265	7.015	8.231	9.390	28.87	31.53	34.81	37.16	42.31	44.43
19	4.912	5.407	6.844	7.633	8.907	10.12	30.14	32.85	36.19	38.58	43.82	45.97
20	5.398	5.921	7.434	8.260	9.591	10.85	31.41	34.17	37.57	40.00	45.31	47.50
21	5.896	6.447	8.034	8.897	10.28	11.59	32.67	35.48	38.93	41.40	46.80	49.01
22	6.404	6.983	8.643	9.542	10.98	12.34	33.92	36.78	40.29	42.80	48.27	50.51
23	6.924	7.529	9.260	10.20	11.69	13.09	35.17	38.08	41.64	44.18	49.73	52.00
24	7.453	8.085	9.886	10.86	12.40	13.85	36.42	39.36	42.98	45.56	51.18	53.48
25	7.991	8.649	10.52	11.52	13.12	14.61	37.65	40.65	44.31	46.93	52.62	54.95
26	8.538	9.222	11.16	12.20	13.84	15.38	38.89	41.92	45.64	48.29	54.05	56.41
27	9.093	9.803	11.81	12.88	14.57	16.15	40.11	43.19	46.96	49.64	55.48	57.86
28	9.656	10.39	12.46	13.56	15.31	16.93	41.34	44.46	48.28	50.99	56.89	59.30
29	10.23	10.99	13.12	14.26	16.05	17.71	42.56	45.72	49.59	52.34	58.30	60.73
30	10.80	11.59	13.79	14.95	16.79	18.49	43.77	46.98	50.89	53.67	59.70	62.16
40	16.91	17.92	20.71	22.16	24.43	26.51	55.76	59.34	63.69	66.77	73.40	76.09
50	23.46	24.67	27.99	29.71	32.36	34.76	67.50	71.42	76.15	79.49	86.66	89.56
60	30.34	31.74	35.53	37.48	40.48	43.19	79.08	83.30	88.38	91.95	99.61	102.7
70	37.47	39.04	43.28	45.44	48.76	51.74	90.53	95.02	100.4	104.2	112.3	115.6
80	44.79	46.52	51.17	53.54	57.15	60.39	101.9	106.6	112.3	116.3	124.8	128.3
90	52.28	54.16	59.20	61.75	65.65	69.13	113.1	118.1	124.1	128.3	137.2	140.8
100	59.90	61.92	67.33	70.06	74.22	77.93	124.3	129.6	135.8	140.2	149.4	153.2

CDF - some properties:

- $F_{XY}(x, y)$ is non-decreasing function of x, y. $F_{XY}(x, +\infty) = F_X(x)$ and $F_{XY}(+\infty, y) = F_Y(y)$
- A continuous cdf posses a **probability density** function $f_{XY}(x, y)$ such that
$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(\tilde{x}, \tilde{y})\, \tilde{x}\, \tilde{y}.$$
- Any positive function that integrates to one defines a cdf.
- For independent X, Y, $f_{XY}(x, y) = f_X(x)\, f_Y(y)$.
- If X, Y takes only finite (countable) number of values, for example $0, 1, 2, \ldots$. The function $p_{ij} = \mathrm{P}(X = i, Y = j)$ is called a **probability mass** function and
$$F_{XY}(x, y) = \sum_{i \le x} \sum_{j \le y} p_{ij}.$$

Example - Multinomial :

A probability-mass function p_{jk} often used in applications is the **multi-nomial distribution**. It is a generalization of the binomial distribution to higher dimensions:

$$P(X = j, Y = k) = \frac{n!}{j!\,k!\,(n-j-k)!}\, p_A^j p_B^k (1 - p_A - p_B)^{n-j-k}$$

for $0 \leq j + k \leq n$ and zero otherwise, p_A, and p_B are parameters.

X is $\text{Bin}(n, p_A)$ while Y is $\text{Bin}(n, p_B)$ but X, Y are in general dependent[3]:

$$P(X = 0, Y = 0) = (1 - p_A - p_B)^n \neq (1 - p_A)^n (1 - p_B)^n$$

Example 1 Solve **Problem 5.2:** Under assumption of independence what is probability that in five fires three are in family houses?

[3] In addition $Z = X + Y$ is $\text{Bin}(n, p_A + p_B)$ and take values $0, \dots, n$, and not $0, \dots, 2n$ what would be the case for independent X and Y.

Example: Normal pdf- and cdf-function:

The cdf of standard normal r.v. Z say is defined through its pdf-function:

$$P(X \le x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \, \mathrm{e}^{-\xi^2/2} \, \mathrm{d}\xi.$$

Let X, Y, be independent $N(0,1)$ variables then

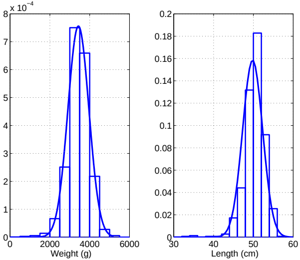
$$f_{XY}(x,y) = f_X(x)\, f_Y(x) = \frac{1}{2\pi} \, \mathrm{e}^{-(x^2+y^2)/2}$$

More generally if Z_1, Z_2 are independent standard normal then $X = m_X + \sigma_X Z_1$, $Y = m_Y + \sigma_Y Z_2$ are independent $N(m_X, \sigma_X^2)$ and $N(m_Y, \sigma_Y^2)$ having joint pdf

$$f_{XY}(x,y) = f_X(x)\, f_Y(x) = \frac{1}{2\pi\sigma_X\sigma_Y} \, \mathrm{e}^{-\frac{1}{2}\left(\frac{(x-m_X)^2}{\sigma_X^2} + \frac{(y-m_Y)^2}{\sigma_Y^2}\right)}.$$

As before $m_X = \mathrm{E}[X]$, $m_Y = \mathrm{E}[Y]$ while $\sigma_X^2 = \mathrm{V}[X]$, $\sigma_Y^2 = \mathrm{V}[Y]$.

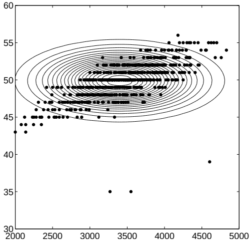
Example:

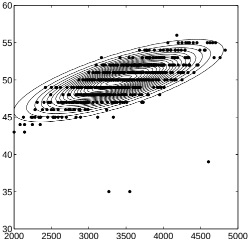


Normalized histogram of weights X (left) and length Y (right) of 750 newborn children in Malmö.

Solid line the normal pdf with $m_X = 3400$ g, $\sigma_X = 570$ g, $m_Y = 49.9$ cm, $\sigma_Y = 2.24$ cm[a].

[a] Three outliers has been removed.





Two dimensional Normal cdf:

The r.v. X, Y are jointly normal $(X, Y) \in N(m_X, m_Y, \sigma_X^2, \sigma_Y^2, \rho)$ and have pdf given by

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

$-1 \leq \rho \leq 1$. If $\rho = 0$ then X, Y are independent. If $\rho = 1$ or -1 Y is a linear function of X.[4]

For any constants a, b, c **if** $(X, Y) \in N(m_X, m_Y, \sigma_X^2, \sigma_Y^2, \rho)$ **then**

$$a + bX + cY \in N(m, \sigma^2), \quad m = a + b\, m_X + c\, m_Y, \quad \sigma^2 = ?.$$

[4] In the previous slight in right-bottom plot $\rho = 0.75$.

Expected value of $Z = h(X, Y)$:

Z is a random variable hence if one knows pdf or pmf then

$$\mathrm{E}[Z] = \int_{-\infty}^{+\infty} z\, f_Z(z)\,\mathrm{d}z \quad \text{or} \quad \mathrm{E}[Z] = \sum_z z\, p_z.$$

If the joint cdf $F_{XY}(x, y)$ is known then $F_Z(z) = \mathrm{P}(h(X, Y) \le z)$ can be computed. However this is not needed since

$$\mathrm{E}[Z] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x, y)\, f_{XY}(x, y)\,\mathrm{d}x\,\mathrm{d}y \quad \text{or} \quad \mathrm{E}[Z] = \sum_{x,y} h(x, y)\, p_{xy}.$$

Example 2 Derive fromula for expectation of $Z = aX + bY$.

Example 3 Derive fromula for expectation of $Z = X \cdot Y$ if X and Y are independent.

Covariance - correlation:

For any two independent r.v. X and Y, $\mathrm{E}[X \cdot Y] = \mathrm{E}[X]\mathrm{E}[Y]$ thus the difference

$$\mathrm{Cov}(X, Y) = \mathrm{E}[X \cdot Y] - E[X]E[Y] \tag{3}$$

is a measure of dependence between X and Y and is called covariance.

From (3) we see that $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y)$ and hence by changing the units of X and Y the covariance can have value close to zero and can be misinterpreted as being only weakly dependent. Consequently, one is also defining scaled covariance called correlation[5]

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{V}[X]\mathrm{V}[Y]}}, \qquad -1 \le \rho \le 1.$$

Example 4 Solve **Problem 5.3** on blackboard.

---

[5] If for X and Y correlation $|\rho| = 1$ then there are constants a; b (both not equal zero) such that $aX + bY = 0$ with probability one.

Covariance - variance of a sum:

When one has two random variables, their variances and covariances are often represented in the form of a symmetric matrix Σ, say,

$$\Sigma = \begin{bmatrix} \mathrm{V}[X] & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{V}[Y] \end{bmatrix}.$$

The variance of a sum of correlated variables will be needed for computation of variance in the following chapters. Starting from the definition of variance and covariance, the following general formula can be derived (do it as an exercise):

$$\mathrm{V}[aX + bY + c] = a^2\mathrm{V}[X] + b^2\mathrm{V}[Y] + 2ab\,\mathrm{Cov}(X,Y).$$

$$\Sigma = \mathrm{Cov}[\mathcal{E}_1, \mathcal{E}_2;\ \mathcal{E}_1, \mathcal{E}_2] \approx -\begin{bmatrix} \dfrac{\partial^2 l}{\partial\theta_1^2} & \dfrac{\partial^2 l}{\partial\theta_1\partial\theta_2} \\ \dfrac{\partial^2 l}{\partial\theta_2\partial\theta_1} & \dfrac{\partial^2 l}{\partial\theta_2^2} \end{bmatrix}^{-1} = -\left[\ddot{l}(\theta_1^*, \theta_2^*)\right]^{-1}$$

Conditional probability mass function

Suppose we are told that the event A, such that $\mathrm{P}(A) > 0$, has occurred, then probability that B occurs (is true), given that A has occurred, is

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}.$$

For discrete random variables X, Y with probability-mass function $p_{jk} = \mathrm{P}(X = j, Y = k)$ the conditional probabilities

$$\mathrm{P}(X = j|Y = k) = \frac{\mathrm{P}(X = j, Y = k)}{\mathrm{P}(Y = k)} = \frac{p_{jk}}{p_k} = p(j|k), \quad j = 0, 1, \ldots$$

It is easy to show that that $p(j|k)$, as a function of j, is a probability-mass function.

Example 5 **Problem 5.11:** Application of Law of Total Probability]

$$\mathrm{P}(B) = \mathrm{P}(B|A_1)\mathrm{P}(A_1) + \mathrm{P}(B|A_2)\mathrm{P}(A_2) + \cdots$$

(blackboard).

The conditional cdf $P(X \leq x|Y = y)$. and pdf

Suppose that we observed the value of Y, e.g. we know that $Y = y$, but X is not observed yet. An important question is if the uncertainty about X is affected by our knowledge that $Y = y$, i.e. if

$$F(x|y) = \mathrm{P}(X \leq x|Y = y)$$

depends on y[6].

For continuous r.v. X, Y it is not obvious how to define conditional probabilities given that "$Y = y$", since $\mathrm{P}(Y = y) = 0$ for all y.It is done using th conditional probability density

$$f(x|y) = \frac{f(x,y)}{f(y)}, \quad F(x|y) = \int_{-\infty}^{x} f(\widetilde{x}|y)\,\mathrm{d}\widetilde{x}$$

is the conditional distribution.

[6]If X and Y are independent then obviously $F(x|y) = F_X(x)$ and Y gives us no knowledge about X.

Law of Total Probability - continuous case

If X and Y have joint density $f(x, y)$ and B is a statement about X, then

$$P(B) = \int_{-\infty}^{+\infty} P(B|Y = y) f_Y(y)\, \mathrm{d}y. \quad P(B \mid Y = y) = \int_B f(x|y)\, \mathrm{d}x,$$

$\boxed{\textit{Example 6}}$ In remote located scientific station there is a supply of food for $Y \in Exp(a_Y)$ days. Waiting time for the new delivery is $X \in Exp(a_x)$. Compute probability that food will not finish, i.e. $P(X < Y)$ assuming that X, Y are independent.

Bayes Formula

In many examples the new piece of information is formulated in form of a statement that is true. For example let Y be strength of a wire and let C ="the wire passed preloading test of 1000kg", i.e. $C =$"$Y > 1000$" is true. If the likelihood $L(y) = \mathrm{P}(C|Y = y)$ is known then the density $f(y|C)$ is computed using Bayes formula

$$f_Y^{pos}(y) = f(y|C) = c\mathrm{P}(C|Y = y)f(y), \qquad c = 1/\mathrm{P}(C).$$

Law of total probability gives $\mathrm{P}(C) = \int_0^\infty \mathrm{P}(C|Y = y)f(y)\,dy$.

Typical problem in safety of existing structure:

Suppose a wire has strength Y and that loads during years i, X_i, are independent. We already can compute

$$\mathrm{P}(\text{"wire survives first years load"}) = \mathrm{P}(X_1 < Y) = \int_0^\infty F_{X_1}(y) f_Y(y)\, dy.$$

Suppose $B =$"wire survives first years load" is true. What is probability

$$\mathrm{P}(\text{"wire survives second year load"}) = \int_0^\infty F_{X_2}(y) f_Y^{post}(y)\, dy.$$

We need the posteriori strength $f_Y^{pos}(y)$, i.e. derivative of $F_Y(y|B)$.

Example 6 Do computations for unrealistic case that X_i and Y are exponentially distributed.