7 Joint distributions: conditioning, correlation, and transformations

7a Conditioning in terms of densities

DISCRETE PROBABILITY states that

(7a1) $$\mathbb{P}\left(Y = y \,\middle|\, X = x \right) = \frac{\mathbb{P}\left(Y = y, X = x \right)}{\mathbb{P}\left(X = x \right)},$$

(7a2) $$\mathbb{P}\left(X = x \right) = \sum_y \mathbb{P}\left(X = x, Y = y \right),$$

that is,

(7a3) $$p_{Y|X=x}(y) = \frac{p_{X,Y}(x,y)}{p_X(x)},$$

(7a4) $$p_X(x) = \sum_y p_{X,Y}(x,y).$$

CONTINUOUS PROBABILITY, assuming existence of $f_{X,Y}$ (2-dim density), states that $f_X(x) = \int f_{X,Y}(x,y)\,dy$ (recall 5c14), which is a continuous counterpart of (7a4). The following definition is a natural continuous counterpart of (7a3).

7a5 Definition. Let random variables X, Y have a joint density $f_{X,Y}$. The *conditional density* of Y given $X = x$ is

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

whenever $f_X(x) \neq 0$.[98]

7a6 Exercise. Let $f_{X,Y}$ be continuous at (x_0, y_0), and f_X be continuous at x_0, and $f_X(x_0) > 0$. Then

$$f_{Y|X=x_0}(y_0) = \lim_{\varepsilon\to0,\delta\to0} \frac{1}{2\delta} \mathbb{P}\left(y_0 - \delta < Y < y_0 + \delta \,\middle|\, x_0 - \varepsilon < X < x_0 + \varepsilon \right).$$

Prove it. (Hint: recall 5b11.)

7a7 Example. Let (X, Y) be distributed uniformly on the disk $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \le 1\}$. Then the conditional density of Y given $X = x$ is the density of the uniform distribution on $(-\sqrt{1-x^2}, +\sqrt{1-x^2})$ whenever $x \in (-1, +1)$.

The conditional density $f_{Y|X=x}$ satisfies the two conditions, $\forall y \; f_{Y|X=x}(y) \ge 0$ and $\int_{-\infty}^{+\infty} f_{Y|X=x}(y)\,dy = 1$; by 2e12, it is a density of a one-dimensional distribution; the latter is denoted by $P_{Y|X=x}$ and called the *conditional distribution* of Y given $X = x$. As

[98] Existence of f_X is ensured by 5c14; $f_X(x) = \int f_{X,Y}(x,y)\,dy$.

any other 1-dim distribution, it has a (cumulative) distribution function — the *conditional distribution function*

(7a8) $$F_{Y|X=x}(y) = \int_{-\infty}^{y} f_{Y|X=x}(y_1)\,dy_1\,,$$

a quantile function — the *conditional quantile function*

(7a9) $$Y^*\,(\,p\,|\,X=x\,)\,,$$

an expectation (if exists) — the *conditional expectation*

(7a10) $$\mathbb{E}\,\big(\,Y\,\big|\,X=x\,\big) = \int_{-\infty}^{+\infty} y f_{Y|X=x}(y)\,dy = \int_{-\infty}^{+\infty} y\,dF_{Y|X=x}(y) = \int_0^1 Y^*(p|X=x)\,dp\,.$$

Some of them are functions of y, others are not, but anyway, they all are functions of x. Substituting X for x, we get random variables (functions of X), namely, the conditional density

(7a11) $$f_{Y|X}(y) = \frac{f_{X,Y}(X,y)}{f_X(X)}$$

(note that the denominator is non-zero with probability 1), the conditional distribution function

(7a12) $$F_{Y|X}(y) = \int_{-\infty}^{y} f_{Y|X}(y_1)\,dy_1\,,$$

the conditional quantile function

(7a13) $$Y^*\,(\,p\,|\,X\,)\,,$$

the conditional expectation

(7a14) $$\mathbb{E}\,\big(\,Y\,\big|\,X\,\big) = \int_{-\infty}^{+\infty} y f_{Y|X}(y)\,dy = \int_{-\infty}^{+\infty} y\,dF_{Y|X}(y) = \int_0^1 Y^*(p|X)\,dp\,.$$

7a15 Exercise. For X, Y as in 7a7 show that

$$f_{Y|X}(0) = \frac{1}{2\sqrt{1-X^2}}\,; \qquad f_{Y|X}\Big(\frac{\sqrt3}{2}\Big) = \frac{1}{2\sqrt{1-X^2}}\mathbf{1}_{(-1/2,1/2)}(X)\,;$$

$$F_{Y|X}(0) = \frac12\,; \qquad F_{Y|X}\Big(\frac{\sqrt3}{2}\Big) = \begin{cases} \frac{\sqrt3/2+\sqrt{1-X^2}}{2\sqrt{1-X^2}} & \text{when } X\in(-1/2,1/2),\\ 1 & \text{otherwise;}\end{cases}$$

$$Y^*(\tfrac12|X) = 0\,; \qquad Y^*(\tfrac34|X) = \tfrac12\sqrt{1-X^2}\,;$$

$$\mathbb{E}\,\big(\,Y\,\big|\,X\,\big) = 0\,.$$

Find the support for each of these 7 random variables.

Bayes formula,

$$\mathbb{P}\left(A \,\middle|\, B \right) = \frac{\mathbb{P}\left(B \,\middle|\, A \right) \mathbb{P}\left(A \right)}{\mathbb{P}\left(B \right)},$$

well-known in discrete probability, has a continuous counterpart:

$$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y) f_X(x)}{f_Y(y)} \tag{7a16}$$

(follows immediately from 7a5).

All said (in Sect. 7a) about dimensions 1 and 2 holds also for other dimensions. You can easily formulate such generalizations. (However, Y^* works only for one-dimensional Y.)

7b Conditioning in general

For now, conditioning is defined for two cases separately: discrete two-dimensional distributions, and distributions having two-dimensional densities. It would be more satisfactory to treat these two cases as special cases of a general definition. Also, some important two-dimensional distributions are intractable for now. Namely, let X have a density but Y be discrete. Then their joint distribution cannot be discrete (since X is not discrete), and cannot have a two-dimensional density (since Y has no density, recall 5c14).

A general approach to conditioning is based on the following idea. Given two random variables X, Y, we try to find another random variable U such that X, U are independent and Y is a function of X, U. Then, given $X = x$, the conditional distribution of Y is obtained from the (conditional = unconditional) distribution of U by a (one-dimensional) transformation.

How to find the needed representation $Y = \varphi(X, U)$? It may be done via the conditional quantile function.

We try it first on a very simple discrete case,

X	0	0	1	1
Y	0	1	0	1
probability	p_{00}	p_{01}	p_{10}	p_{11}

with some positive probabilities p_{kl}. Discrete probability gives us the conditional distribution of Y given $X = 0$,

$$\begin{aligned} \mathbb{P}\left(Y = 0 \,\middle|\, X = 0 \right) &= \frac{p_{00}}{p_0}, \\ \mathbb{P}\left(Y = 1 \,\middle|\, X = 0 \right) &= \frac{p_{01}}{p_0}, \end{aligned} \qquad p_0 = p_{00} + p_{01} = \mathbb{P}\left(X = 0 \right).$$

The corresponding quantile function is

$$Y^*\left(p \,|\, X = 0 \right) = \begin{cases} 0 & \text{if } p < p_{00}/p_0, \\ 1 & \text{if } p > p_{00}/p_0. \end{cases}$$

Similarly,

$$Y^*\,(\,p\,|\,X=1\,) = \begin{cases} 0 & \text{if } p < p_{10}/p_1, \\ 1 & \text{if } p > p_{10}/p_1, \end{cases}$$

$p_1 = p_{10} + p_{11} = \mathbb{P}\big(\,X=1\,\big)$. We introduce $U \sim \mathrm{U}(0,1)$ independent of X and

$$\tilde{Y} = \varphi(X,U) = Y^*\,(\,U\,|\,X\,)\,,$$

that is, $\varphi(0,u) = Y^*\,(\,u\,|\,X=0\,)$ and $\varphi(1,u) = Y^*\,(\,u\,|\,X=1\,)$. The joint distribution of $X, \tilde{Y}$ is equal to the joint distribution of X, Y. For example, $\mathbb{P}\big(\,X=0,\,\tilde{Y}=0\,\big) = \mathbb{P}\big(\,X=0\,\big)\mathbb{P}\big(\,U < p_{00}/p_0\,\big) = p_0 \cdot p_{00}/p_0 = p_{00} = \mathbb{P}\big(\,X=0,\,Y=0\,\big)$. The (unconditional) distribution of $\varphi(0,U)$ is equal to the conditional distribution of Y given $X=0$. For example, $\mathbb{P}\big(\,\varphi(0,U)=0\,\big) = \mathbb{P}\big(\,U < p_{00}/p_0\,\big) = \mathbb{P}\big(\,Y=0\,\big|\,X=0\,\big)$. And the (unconditional) distribution of $\varphi(1,U)$ is equal to the conditional distribution of Y given $X=1$.

Similarly, the general approach conforms to the elementary (discrete) approach for every discrete two-dimensional distribution. You may say: no real progress here, we still used our old good discrete conditioning. Yes, we did, but it is also possible to do from scratch. Namely, we may seek two thresholds a, b such that, defining $\tilde{Y}$ by

$$\tilde{Y} = \varphi(X,U)\,, \qquad \begin{aligned} \varphi(0,u) &= \begin{cases} 0 & \text{if } u < a, \\ 1 & \text{if } u > a, \end{cases} \\ \varphi(1,u) &= \begin{cases} 0 & \text{if } u < b, \\ 1 & \text{if } u > b, \end{cases} \end{aligned}$$

we get the joint distribution of $(X,\tilde{Y})$ the same as of (X,Y).

7b1 Exercise. Prove that the two joint distributions are equal if and only if $a = p_{00}/p_0$, $b = p_{10}/p_1$.

The general approach will give us new results soon (in 7c). Before that, however, we try it on the second old case. Let X, Y have a two-dimensional density

$$f_{X,Y}(\cdot,\cdot)\,,$$

then we have the conditional density 7a5 and the conditional quantile function (7a9),

$$Y^*\,(\,p\,|\,X=x\,)\,.$$

As before, we introduce $U \sim \mathrm{U}(0,1)$ independent of X and

$$\tilde{Y} = \varphi(X,U) = Y^*\,(\,U\,|\,X\,)\,.$$

In order to prove that $(X,\tilde{Y})$ is distributed like (X,Y) it is sufficient to check that $F_{X,\tilde{Y}} = F_{X,Y}$. The probability

$$F_{X,\tilde{Y}}(x,y) = \mathbb{P}\big(\,X \le x, \tilde{Y} \le y\,\big) = \mathbb{P}\big(\,X \le x, Y^*(U|X) \le y\,\big)$$

may be calculated by means of the joint density of X and U,

$$f_{X,U}(x,u) = f_X(x) f_U(u)\,, \qquad f_U(u) = 1 \text{ for } u \in (0,1)\,.$$

We integrate $f_{X,U}$ on the set

$$B = \{(x_1,u) : x_1 \le x,\; Y^*(u|X=x_1) \le y\}\,.$$

First, we keep $x_1 \in (-\infty, x]$ fixed and integrate in u,

$$\int_{B_{x_1}} f_U(u)\,du = \mathbb{P}\bigl(\, Y^*(\cdot|X=x_1) \le y \,\bigr) = F_{Y|X=x_1}(y) = \frac{\int_{-\infty}^{y} f_{X,Y}(x_1,y_1)\,dy_1}{f_X(x_1)}\,.$$

Second, we integrate it in x_1,

$$F_{X,\tilde Y}(x,y) = \iint_B f_{X,U}(x_1,u)\,dx_1 du = \int_{-\infty}^{x} \biggl(\int_{B_{x_1}} f_U(u)\,du \biggr) f_X(x_1)\,dx_1 = \\ \int_{-\infty}^{x} \biggl(\int_{-\infty}^{y} f_{X,Y}(x_1,y_1)\,dy_1 \biggr) dx_1 = F_{X,Y}(x,y)\,.$$

We see that $(X,\tilde Y)$ is distributed like (X,Y). The (unconditional) distribution of $\varphi(x,U)$ has the density $f_{Y|X=x}$, it is the conditional distribution of Y given $X=x$. Thus, the general approach conforms to the approach of 7a whenever a two-dimensional density exists.

We see that the general approach is consistent with the two special cases. The following result shows self-consistency of the general approach: conditional distributions do not depend on the choice of the representation $\tilde Y = \varphi(X,U)$.

7b2 Proposition. Let $Y_1 = \varphi_1(X_1,U_1)$ and $Y_2 = \varphi_2(X_2,U_2)$, where $\varphi_1,\varphi_2 : \mathbb{R}^2 \to \mathbb{R}$ are Borel functions, $X_1,U_1 : \Omega_1 \to \mathbb{R}$ are independent random variables, and $X_2,U_2 : \Omega_2 \to \mathbb{R}$ are independent random variables. If the two pairs (X_1,Y_1) and (X_2,Y_2) are identically distributed, then

$$\varphi_1(x,U_1) \text{ and } \varphi_2(x,U_2) \text{ are identically distributed}$$

for almost all x w.r.t. the distribution $P_{X_1} = P_{X_2}$.

Universal applicability of the general approach is ensured by the following result.

7b3 Proposition. For every two-dimensional distribution P there exist independent random variables X,U and a Borel function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ such that the pair $\bigl(X,\varphi(X,U)\bigr)$ is distributed P.

Moreover, it is always possible to choose $U \sim \mathrm{U}(0,1)$ and $\varphi(x,u)$ increasing in u (for every x); then $\varphi(x,u) = Y^*\,(\,u\,|\,X=x\,)$.

Taking 7b2 and 7b3 into account we may define conditioning as follows.

7b4 Definition. (a) Let P be a two-dimensional distribution. Its conditional distribution P_x is the distribution of $\varphi(x,U)$ where a Borel function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ and independent random variables X,U are chosen such that the pair $\bigl(X,\varphi(X,U)\bigr)$ is distributed P.

(b) Let $X,Y : \Omega \to \mathbb{R}$ be random variables. The conditional distribution $P_{Y|X=x}$ is P_x where $P = P_{X,Y}$.

Similarly to densities, conditional distributions are defined up to an arbitrary change on a negligible set (of x).

7c Combining discrete and continuous

We consider a pair X, Y of random variables, X (continuous) and Y (discrete). For simplicity we assume that Y takes on two values $0, 1$ only, equiprobably,

$$\mathbb{P}\big(Y=0\big)=\frac{1}{2}=\mathbb{P}\big(Y=1\big).$$

Conditioning on Y is elementary,

$$F_{X|Y=0}(x)=\mathbb{P}\big(X\le x\,\big|\,Y=0\big)=\frac{\mathbb{P}\big(X\le x,\,Y=0\big)}{\mathbb{P}\big(Y=0\big)}=2\mathbb{P}\big(X\le x,\,Y=0\big);$$
$$F_{X|Y=1}(x)=2\mathbb{P}\big(X\le x,\,Y=1\big).$$

We assume that these two distributions have densities,[99]

$$F_{X|Y=0}(x)=\int_{-\infty}^{x} f_{X|Y=0}(x_1)\,dx_1\,,$$
$$F_{X|Y=1}(x)=\int_{-\infty}^{x} f_{X|Y=1}(x_1)\,dx_1\,.$$

Then X has (unconditional) density

$$f_X(x)=\frac{1}{2}f_{X|Y=0}(x)+\frac{1}{2}f_{X|Y=1}(x)\,;$$

indeed, $F_X(x)=\mathbb{P}\big(X\le x\big)=\mathbb{P}\big(Y=0\big)\mathbb{P}\big(X\le x\,\big|\,Y=0\big)+\mathbb{P}\big(Y=1\big)\mathbb{P}\big(X\le x\,\big|\,Y=1\big)=\int_{-\infty}^{x} f_X(x_1)\,dx_1$ for f_X as above.

What about conditional probabilities

$$\mathbb{P}\big(Y=0\,\big|\,X=x\big)\,,\qquad \mathbb{P}\big(Y=1\,\big|\,X=x\big)\,,$$

are they well-defined? How to calculate them? The elementary approach cannot answer, since $\mathbb{P}\big(X=x\big)=0$ for all x.

We try the general approach of 7b:

$$\tilde{Y}=\varphi(X,U)=\begin{cases}0 & \text{if } U<g(X),\\ 1 & \text{if } U>g(X);\end{cases}$$

can we find g such that $(X,\tilde{Y})$ is distributed like (X,Y)?

We have (similarly to the second part of 7b)

$$F_{X,Y}(x,0)=\mathbb{P}\big(X\le x,\,Y=0\big)=\mathbb{P}\big(Y=0\big)\mathbb{P}\big(X\le x\,\big|\,Y=0\big)=\frac{1}{2}\int_{-\infty}^{x} f_{X|Y=0}(x_1)\,dx_1\,;$$

[99] It is sufficient that X has a (unconditional) density.

$$F_{X,\tilde{Y}}(x,0) = \mathbb{P}\left(X \le x,\, \tilde{Y} = 0 \right) = \mathbb{P}\left(X \le x,\, U < g(X) \right) =$$
$$= \int_{-\infty}^{x} \biggl(\int_0^{g(x_1)} f_U(u)\,du \biggr) f_X(x_1)\,dx_1 = \int_{-\infty}^{x} g(x_1) f_X(x_1)\,dx_1\,;$$

they became equall (for all x) when

$$g(x) = \frac{1}{2} \frac{f_{X|Y=0}(x)}{f_X(x)}\,.$$

The (unconditional) distribution of $\varphi(x,U)$ gives us the conditional distribution of Y given $X = x$:

$$\mathbb{P}\left(Y = 0 \,\middle|\, X = x \right) = \mathbb{P}\left(\varphi(x,U) = 0 \right) = \mathbb{P}\left(U < g(x) \right) = g(x)\,.$$

So,

$$\mathbb{P}\left(Y = 0 \,\middle|\, X = x \right) = \frac{f_{X|Y=0}(x)\mathbb{P}\left(Y = 0 \right)}{f_X(x)}\,,$$

which is another case of Bayes formula; compare it with (7a16).

The same holds for any discrete Y (taking on a finite or countable set of values):

(7c1) $$f_X(x) = \sum_y \mathbb{P}\left(Y = y \right) f_{X|Y=y}(x)\,;$$

(7c2) $$\mathbb{P}\left(Y = y \,\middle|\, X = x \right) = \frac{f_{X|Y=y}(x)\mathbb{P}\left(Y = y \right)}{f_X(x)}\,.$$

All said (in Sect. 7c) about dimensions 1 and 2 holds also for other dimensions. You can easily formulate such generalizations. (However, Y^* works only for one-dimensional Y.)

7d Back to unconditional: total probability, expectation, density

Discrete probability states that

$$\mathbb{E}\left(\mathbb{E}\left(Y \,\middle|\, X \right)\right) = \mathbb{E}\left(Y\right),$$

that is, the expectation of Y may be calculated in two stages: first $\mathbb{E}\left(Y \,\middle|\, X \right)$, second, expectation of the first. Continuous probability states the same.

7d1 Theorem. Let X, Y be random variables, Y being integrable. Then the conditional expectation $\mathbb{E}\left(Y \,\middle|\, X \right)$ exists with probability 1, is an integrable random variable, and

$$\mathbb{E}\left(\mathbb{E}\left(Y \,\middle|\, X \right)\right) = \mathbb{E}\left(Y\right).$$

It follows immediately that

(7d2) $$\mathbb{E}\left(\mathbb{E}\left(\varphi(X,Y) \,\middle|\, X \right)\right) = \mathbb{E}\left(\varphi(X,Y)\right)$$

whenever $\varphi(X,Y)$ is integrable.

On the other hand, let $Y = \mathbf{1}_A$ be the indicator of an event $A \subset \Omega$; then $\mathbb{E}\,(Y) = \mathbb{P}\,\big(A\big)$, $\mathbb{E}\,\big(Y \,\big|\, X\big) = \mathbb{P}\,\big(A \,\big|\, X\big)$, and we get

$$\mathbb{P}\,\big(A\big) = \mathbb{E}\,\big(\mathbb{P}\,\big(A \,\big|\, X\big)\big)\,; \tag{7d3}$$

expectation of the conditional probability is the unconditional probability. That is a continuous counterpart of the *total probability formula* of discrete probability:

$$\mathbb{P}\,\big(A\big) = \mathbb{P}\,\big(A \,\big|\, B_1\big)\,\mathbb{P}\,\big(B_1\big) + \dots + \mathbb{P}\,\big(A \,\big|\, B_n\big)\,\mathbb{P}\,\big(B_n\big)\,;$$
$$\mathbb{P}\,\big(A\big) = \mathbb{P}\,\big(A \,\big|\, X = x_1\big)\,\mathbb{P}\,\big(X = x_1\big) + \dots + \mathbb{P}\,\big(A \,\big|\, X = x_n\big)\,\mathbb{P}\,\big(X = x_n\big)\,.$$

For proving Theorem 7d1 we need a fact close to the Fubini theorem.

7d4 Lemma. Let X, Y be independent random variables and $\varphi : \mathbb{R}^2 \to \mathbb{R}$ a Borel function such that $\varphi(X, Y)$ is integrable. Then

$$\mathbb{E}\,\varphi(X,Y) = \mathbb{E}\,\psi(X) \qquad \text{where } \psi(x) = \mathbb{E}\,\varphi(x,Y)\,.$$

Proof. Similarly to the proof of 5d12, we restrict ourselves to the model of (5d2),

$$(\Omega, \mathcal{F}, P) = \big((0,1)\times(0,1), \mathcal{B}_2|_{(0,1)\times(0,1)}, \operatorname{mes}_2|_{(0,1)\times(0,1)}\big)\,;$$
$$X(\omega) = X(\omega_1,\omega_2) = X^*(\omega_1)\,,$$
$$Y(\omega) = Y(\omega_1,\omega_2) = Y^*(\omega_2)\,,$$

and apply Fubini theorem 5c11:

$$\begin{aligned}\mathbb{E}\,\varphi(X,Y) &= \iint_{(0,1)\times(0,1)} \varphi\big(X(\omega_1,\omega_2), Y(\omega_1,\omega_2)\big)\,d\omega_1 d\omega_2 = \\ &= \iint_{(0,1)\times(0,1)} \varphi\big(X^*(\omega_1), Y^*(\omega_2)\big)\,d\omega_1 d\omega_2 = \int_0^1 \bigg(\int_0^1 \varphi\big(X^*(\omega_1), Y^*(\omega_2)\big)\,d\omega_2\bigg) d\omega_1 = \\ &= \int_0^1 \psi\big(X^*(\omega_1)\big) d\omega_1 = \mathbb{E}\,\psi(X)\,,\end{aligned}$$

since

$$\psi(x) = \mathbb{E}\,\varphi(x,Y) = \int_0^1 \varphi\big(x, Y^*(\omega_2)\big)\,d\omega_2\,.$$

□

Proof of Theorem 7d1. We may assume that $Y = \varphi(X,U)$ where X, U are independent. We apply 7d4 to X, U (rather than X, Y):

$$\psi(x) = \mathbb{E}\,\varphi(x,U) = \mathbb{E}\,\big(Y \,\big|\, X = x\big)\,;$$
$$\mathbb{E}\,Y = \mathbb{E}\,\varphi(X,U) = \mathbb{E}\,\psi(X) = \mathbb{E}\,\big(\mathbb{E}\,\big(Y \,\big|\, X\big)\big)\,.$$

□

7d5 Exercise. Give a more elementary proof (using 7a but not 7b) of Theorem 7d1 for the case of X, Y having a joint density.

Hint: use the conditional density, and apply Fubini theorem 5c11.

7d6 Example. Let $X \sim \mathrm{U}(0,1)$, and the conditional distribution of Y given $X = x$ be $\mathrm{U}(0,x)$. Then

$$\mathbb{E}\big(Y \,\big|\, X = x\big) = \frac{1}{2}x; \quad \mathbb{E}\big(Y \,\big|\, X\big) = \frac{1}{2}X; \quad \mathbb{E}(Y) = \mathbb{E}\Big(\frac{1}{2}X\Big) = \frac{1}{4}.$$

Also,

$$f_{X,Y}(x,y) = f_X(x) f_{Y|X=x}(y) = \mathbf{1}_{(0,1)}(x) \cdot \frac{1}{x}\mathbf{1}_{(0,x)}(y) = \begin{cases} \frac{1}{x} & \text{if } 0 < y < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

and we may calculate $\mathbb{E}(Y)$ without 7d1;

$$\mathbb{E}(Y) = \iint y f_{X,Y}(x,y)\, dxdy = \iint_{0<y<x<1} \frac{y}{x}\, dxdy = \begin{cases} = \int_0^1 \big(\int_0^x \frac{y}{x}\, dy\big)\, dx = \int_0^1 \frac{x}{2}\, dx = \frac{1}{4}; \\ = \int_0^1 \big(\int_y^1 \frac{y}{x}\, dx\big)\, dy = \int_0^1 y \cdot (-\ln y)\, dy = \big(-\frac{y^2}{2}\ln y + \frac{y^2}{4}\big)\big|_0^1 = \frac{1}{4}. \end{cases}$$

7d7 Exercise. Let random variables X, Y have a joint density $f_{X,Y}$. Then

$$f_Y(y) = \mathbb{E}\, f_{Y|X}(y).$$

Prove it. (Hint: recall (7a11) and 5c14.)

7d8 Example. Let X, Y be as in 7d6. Then

$$f_{Y|X}(y) = \frac{1}{X}\mathbf{1}_{(0,X)}(y) = \frac{1}{X}\mathbf{1}_{(y,\infty)}(X);$$

$$f_Y(y) = \mathbb{E}\, f_{Y|X}(y) = \mathbb{E}\Big(\frac{1}{X}\mathbf{1}_{(y,\infty)}(X)\Big) = \int_y^1 \frac{1}{x}\, dx = \ln x\big|_y^1 = -\ln y$$

for $y \in (0,1)$; otherwise $f_Y(y) = 0$. Another way to f_Y:

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\, dx = \int_y^1 \frac{1}{x}\, dx = -\ln y.$$

Having f_Y, we may use it for calculating $\mathbb{E}(Y)$ once again:

$$\mathbb{E}(Y) = \int_{-\infty}^{+\infty} y f_Y(y)\, dy = \int_0^1 y \cdot (-\ln y)\, dy = \frac{1}{4}.$$

(Do you understand, why the last integral here is the same as the last integral in 7d6?)

All said (in Sect. 7d) about dimensions 1 and 2 holds also for other dimensions. You can easily formulate such generalizations.

7e Correlation

Distributions are quite diverse; instead of describing them in detail, we may sometimes prefer a rough description via two parameters, the expectation $\mathbb{E}(X)$ and the variance $\operatorname{Var}(X) = \sigma_X^2$. It may be called the second-order (or quadratic) description, since first and second moments of X are used. The second-order description is insensitive to the distinction between discrete and continuous. For 1-dim case recall (4f1), (4f2).

A dependence between X and Y cannot influence $\operatorname{Var}(X)$, $\operatorname{Var}(Y)$ (these involve marginal distributions only), but influences $\operatorname{Var}(X+Y)$, $\operatorname{Var}(X-Y)$ and, more generally, $\operatorname{Var}(aX+bY)$.

7e1 Exercise. Let X, Y have second moments.[100] Then $aX + bY$ has second moment for any $a, b \in \mathbb{R}$. Prove it. (Hint: $(u+v)^2 \le (u+v)^2 + (u-v)^2 = 2(u^2+v^2)$.)

Treated as a function of the coefficients a, b, the variance $\operatorname{Var}(aX+bY)$ is a quadratic form, which is well-known from discrete probability:

$$\begin{aligned}\mathbb{E}\big((aX+bY)^2\big) - \big(\mathbb{E}(aX+bY)\big)^2 &= a^2\mathbb{E}(X^2) + 2ab\mathbb{E}(XY) + b^2\mathbb{E}(Y^2) - (a\mathbb{E}X + b\mathbb{E}Y)^2 = \\ &= a^2\big(\mathbb{E}(X^2) - (\mathbb{E}X)^2\big) + 2ab\big(\mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)\big) + b^2\big(\mathbb{E}(Y^2) - (\mathbb{E}Y)^2\big),\end{aligned}$$

that is,

$$\operatorname{Var}(aX+bY) = a^2\operatorname{Var}(X) + 2ab\operatorname{Cov}(X,Y) + b^2\operatorname{Var}(Y), \tag{7e2}$$

where the *covariance* is defined by

$$\operatorname{Cov}(X,Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) \tag{7e3}$$

whenever X, Y have second moments.

7e4 Exercise. $\operatorname{Cov}(X,Y) = \mathbb{E}\big((X-\mathbb{E}X)(Y-\mathbb{E}Y)\big)$. Prove it. (Hint: just open the brackets.)

7e5 Exercise. $\operatorname{Cov}(aX+bY, Z) = a\operatorname{Cov}(X,Z) + b\operatorname{Cov}(Y,Z)$. Prove it. What about $\operatorname{Cov}(X, aY+bZ)$? What about $\operatorname{Cov}(aX+bY, cU+dV)$?

Geometrically, a quadratic form corresponds to an ellipse. In order to simplify the situation, we turn to *standardized* random variables $\tilde X, \tilde Y$:

$$\begin{gathered}\tilde X = \frac{X - \mathbb{E}(X)}{\sigma_X}, \qquad X = \sigma_x \cdot \tilde X + \mathbb{E}X, \\ \mathbb{E}(\tilde X) = 0, \quad \operatorname{Var}(\tilde X) = 1,\end{gathered} \tag{7e6}$$

[100] That is, $\mathbb{E}(X^2) < \infty$, or equivalently $\operatorname{Var}(X) < \infty$ (and the same for Y). It follows that X, Y have first moments, that is, $\mathbb{E}|X| < \infty$ (and the same for Y).

and the same for Y.[101] Of course, we assume that $\sigma_X \ne 0$, $\sigma_Y \ne 0$.[102] The covariance between $\tilde X$ and $\tilde Y$ is called the *correlation coefficient:*

$$\rho(X,Y) = \operatorname{Cov}(\tilde X, \tilde Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y}\,. \tag{7e7}$$

Random variables X, Y are called *uncorrelated,* if $\rho(X,Y) = 0$.

7e8 Exercise. Let random variables X, Y have second moments. If X, Y are independent then they are uncorrelated. Prove it. The converse is wrong. Find a counterexample.

Remark. Being uncorrelated means a single equality $\operatorname{Cov}(X,Y) = 0$. Being independent means a continuum of equalities, $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all $x, y \in \mathbb{R}$. You see, the latter is much stronger than the former.

So,

$$\operatorname{Var}(aX + bY) = a^2\sigma_X^2 + 2ab\sigma_X\sigma_Y\,\rho(X,Y) + b^2\sigma_Y^2\,. \tag{7e9}$$

We'll investigate (in the second order) dependence between $\tilde X$, $\tilde Y$; afterwards we'll return to X, Y easily.

Note that

$$\operatorname{Var}(\tilde X + \tilde Y) = 2\bigl(1 + \rho(X,Y)\bigr)\,, \qquad \operatorname{Var}(\tilde X - \tilde Y) = 2\bigl(1 - \rho(X,Y)\bigr)\,, \tag{7e10}$$

therefore

$$-1 \le \rho(X,Y) \le 1\,. \tag{7e11}$$

Extremal cases $\rho(X,Y) = \pm 1$ are simple. If $\rho(X,Y) = 1$ then $\operatorname{Var}(\tilde X - \tilde Y) = 0$, therefore $\tilde X = \tilde Y$ (think, why), which means a linear functional dependence between X and Y;

$$\begin{aligned} Y &= \frac{\sigma_Y}{\sigma_X}(X - \mathbb{E}\,X) + \mathbb{E}\,Y \quad \text{if } \rho(X,Y) = +1\,; \\ Y &= -\frac{\sigma_Y}{\sigma_X}(X - \mathbb{E}\,X) + \mathbb{E}\,Y \quad \text{if } \rho(X,Y) = -1 \end{aligned} \tag{7e12}$$

(the latter case is similar to the former).

Now assume that $\rho(X,Y) \ne \pm 1$, that is, $-1 < \rho(X,Y) < 1$. Note that $\tilde X + \tilde Y$ and $\tilde X - \tilde Y$ are uncorrelated, that is, $\operatorname{Cov}(\tilde X + \tilde Y, \tilde X - \tilde Y) = 0$ (why?). Introduce

$$U = \widetilde{\tilde X + \tilde Y} = \frac{\tilde X + \tilde Y}{\sqrt{2(1+\rho)}}\,, \qquad V = \widetilde{\tilde X - \tilde Y} = \frac{\tilde X - \tilde Y}{\sqrt{2(1-\rho)}} \tag{7e13}$$

(of course, $\rho = \rho(X,Y)$); then

$$\begin{gathered} \tilde X = \sqrt{\frac{1+\rho}{2}}U + \sqrt{\frac{1-\rho}{2}}V\,, \qquad \tilde Y = \sqrt{\frac{1+\rho}{2}}U - \sqrt{\frac{1-\rho}{2}}V\,; \\ \operatorname{Var}(U) = 1\,, \quad \operatorname{Var}(V) = 1\,, \quad \operatorname{Cov}(U,V) = 0\,. \end{gathered} \tag{7e14}$$

[101] Often, the standardized random variable is denoted by $\hat X$ rather than $\tilde X$. However, $\hat X$ is widely used for denoting an estimator (predictor) of X.

[102] Is it possible that $\sigma_X = 0$? What does it mean?

Random variables of the form $aX + bY + c$ are the same as random variables of the form $aU + bV + c$ (with a, b, c changed appropriately).

First, consider the case $a = \cos\varphi$, $b = \sin\varphi$, $c = 0$. The random variable $Z = U\cos\varphi + V\sin\varphi$ satisfies $\mathbb{E}\,Z = 0$, $\sigma_Z = 1$. On the other hand, the function $(u, v) \mapsto u\cos\varphi + v\sin\varphi$ on the plane vanishes at the origin, and its maximal value on the unit disk $u^2 + v^2 \le 1$ is equal to 1, thus to σ_Z.

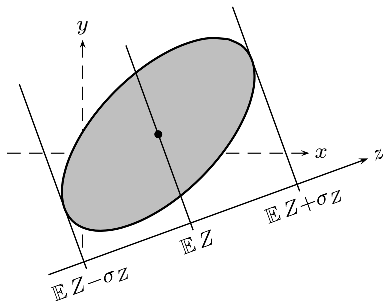
Second, consider the case $c = 0$; arbitrary a, b may be represented as $a = r\cos\varphi$, $b = r\sin\varphi$ where $r = \sqrt{a^2 + b^2}$. The random variable $Z = aU + bV = r(U\cos\varphi + V\sin\varphi)$ satisfies $\mathbb{E}\,Z = 0$, $\sigma_Z = r$. On the other hand, the function $(u, v) \mapsto au + bv = r(u\cos\varphi + v\sin\varphi)$ on the plane vanishes at the origin, and its maximal value on the unit disk $u^2 + v^2 \le 1$ is equal to r, thus to σ_Z.

Third, consider the general case, $Z = aU + bV + c$. Here $\mathbb{E}\,Z = c$ and $\mathbb{E}\,Z + \sigma_Z = c + \sqrt{a^2 + b^2}$. On the other hand, the function $(u, v) \mapsto au + bv + c$ is equal to $c = \mathbb{E}\,Z$ at the origin, and its maximal value on the unit disk $u^2 + v^2 \le 1$ is equal to $\mathbb{E}\,Z + \sigma_Z$. Also, its minimal value on the disk is equal to $\mathbb{E}\,Z - \sigma_Z$.

Now we return from U, V to X, Y. The disk $u^2 + v^2 \le 1$ on the u, v-plane turns into an ellipse on the x, y-plane, call it the *concentration ellipse* of X, Y. Its explicit form is rather frightening,

(7e15)
$$\frac{\Big(\frac{x - \mathbb{E}\,X}{\sigma_X} + \frac{y - \mathbb{E}\,Y}{\sigma_Y}\Big)^2}{2(1+\rho)} + \frac{\Big(\frac{x - \mathbb{E}\,X}{\sigma_X} - \frac{y - \mathbb{E}\,Y}{\sigma_Y}\Big)^2}{2(1-\rho)} \le 1\,,$$

but its property is easy to understand.



7e16. Consider the random variable $Z = aX + bY + c$ and the function $(x, y) \mapsto ax + by + c$ on the plane.

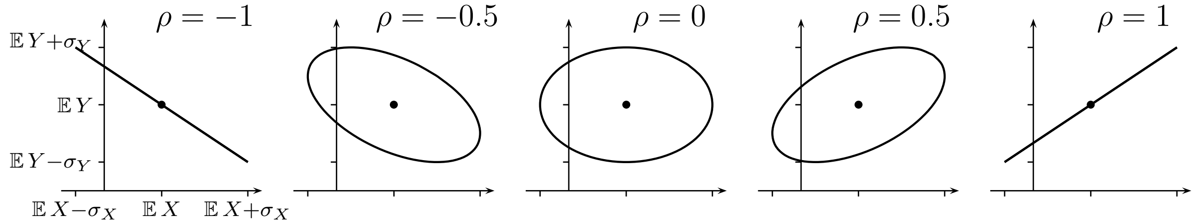
(a) $\mathbb{E}\,(Z)$ is equal to the value of the function at the center of the concentration ellipse.

(b) $\mathbb{E}\,(Z) + \sigma_Z$ is equal to the maximal value of the function on the concentration ellipse.

(c) $\mathbb{E}\,(Z) - \sigma_Z$ is equal to the minimal value of the function on the concentration ellipse.

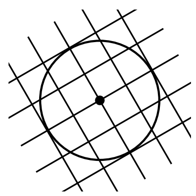
In fact, 7e16 holds also in the extremal cases $\rho = \pm 1$, however, the concentration ellipse degenerates into a straight segment connecting two points, $(\mathbb{E}\,X - \sigma_X, \mathbb{E}\,Y - \sigma_Y)$ and $(\mathbb{E}\,X +$

$\sigma_X, \mathbb{E}\,Y + \sigma_Y)$ if $\rho = +1$, or $(\mathbb{E}\,X - \sigma_X, \mathbb{E}\,Y + \sigma_Y)$ and $(\mathbb{E}\,X + \sigma_X, \mathbb{E}\,Y - \sigma_Y)$ if $\rho = -1$.

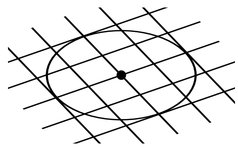


7e17 Exercise. The following three conditions are equivalent:
(a) Two random variables $Z_1 = a_1 U + b_1 V + c_1$, $Z_2 = a_2 U + b_2 V + c_2$ are uncorrelated.
(b) Two vectors (a_1, b_1), (a_2, b_2) are orthogonal.
(c) Lines $a_1 u + b_1 v = \text{const}$ are orthogonal to lines $a_2 u + b_2 v = \text{const}$.
Prove it.

On the x, y-plane the situation is more complicated (than on the u, v-plane), since the concentration ellipse is not just a disk. Instead of treating orthogonality as a relation between two directions, we may treat it as a relation between the two directions *and the disk,*



which may be generalized to an ellipse,



Such directions are called *conjugate* (w.r.t. the ellipse). That relation is invariant under linear transformations, and so, 7e17(a,c) may be transferred to the x, y-plane as follows.

7e18. Two random variables $Z_1 = a_1 X + b_1 Y + c_1$, $Z_2 = a_2 X + b_2 Y + c_2$ are uncorrelated if and only if two directions $a_1 x + b_1 y = \text{const}$, $a_2 x + b_2 y = \text{const}$ are conjugate w.r.t. the concentration ellipse.

The *optimal linear predictor* $\hat{Y}$ for Y from X is, by definition, a random variable of the form $\hat{Y} = aX + b$ that minimizes (over $a, b \in \mathbb{R}$) the mean square error $\mathbb{E}\,(\hat{Y} - Y)^2$.

First, we'll find the optimal linear predictor $\hat{\tilde{Y}}$ for $\tilde{Y}$ from $\tilde{X}$; afterwards we'll return to X, Y easily. We have

$$\mathbb{E}\,(a\tilde{X} + b - \tilde{Y})^2 = \operatorname{Var}(a\tilde{X} + b - \tilde{Y}) + \big(\mathbb{E}\,(a\tilde{X} + b - \tilde{Y})\big)^2 = \operatorname{Var}(a\tilde{X} - \tilde{Y}) + b^2\,;$$

the optimal value of b is evidently 0. Further,

$$\operatorname{Var}(a\tilde{X}-\tilde{Y}) = a^2\operatorname{Var}(\tilde{X}) - 2a\operatorname{Cov}(\tilde{X},\tilde{Y}) + \operatorname{Var}(\tilde{Y}) = a^2 - 2\rho a + 1 = (a-\rho)^2 + 1 - \rho^2\,;$$

the optimal value of a is evidently $\rho = \rho(X,Y)$. So,

$$\hat{\tilde{Y}} = \rho\tilde{X}\,; \qquad \mathbb{E}\,\big(\hat{\tilde{Y}} - \tilde{Y}\big)^2 = 1-\rho^2\,. \tag{7e19}$$

Note that the prediction error $\hat{\tilde{Y}} - \tilde{Y}$ is uncorrelated with the predictor $\hat{\tilde{Y}}$, as well as with $\tilde{X}$:

$$\operatorname{Cov}(\hat{\tilde{Y}} - \tilde{Y}, \tilde{X}) = \operatorname{Cov}(\rho\tilde{X} - \tilde{Y}, \tilde{X}) = \rho\operatorname{Var}(\tilde{X}) - \operatorname{Cov}(\tilde{Y},\tilde{X}) = 0\,.$$

Now we return from $\tilde{X},\tilde{Y}$ to X,Y:

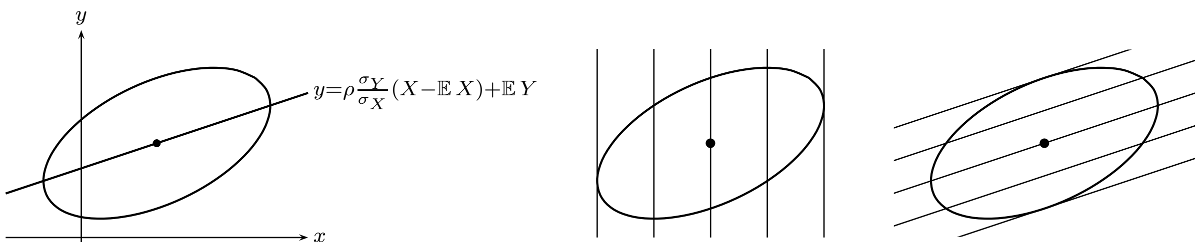
$$\hat{Y} = \sigma_Y\hat{\tilde{Y}} + \mathbb{E}\,Y = \sigma_Y\rho\tilde{X} + \mathbb{E}\,Y = \sigma_Y\rho\frac{X-\mathbb{E}\,X}{\sigma_X} + \mathbb{E}\,Y\,,$$

so, the optimal linear predictor is

$$\hat{Y} = \rho(X,Y)\frac{\sigma_Y}{\sigma_X}(X-\mathbb{E}\,X) + \mathbb{E}\,Y\,, \tag{7e20}$$

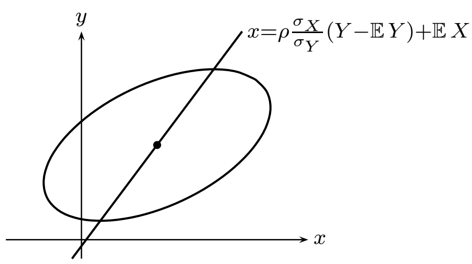
and the mean square error is

$$\mathbb{E}\,(\hat{Y}-Y)^2 = (1-\rho^2)\sigma_Y^2\,. \tag{7e21}$$



$x = \text{const}$ error = const

Note that the optimal linear predictor $\hat{X}$ for X from Y is another line (unless $\rho = \pm 1$).



Generalizations (of Sect. 7e) for higher dimensions are well-known, but involve multidimensional ellipsoids, matrices etc.

7f Transformations

Recall the simplest case, a *linear* 1-dim transformation $Y = aX + b$, studied in Sect. 3a. In the smooth case we have (see (3a6))

$$|a| f_Y(y) = f_X(x) \qquad (y = ax + b,\ a \ne 0)\,. \tag{7f1}$$

The same holds in full generality. Namely, Y has a density if and only if X has a density, in which case (7f1) holds almost everywhere.

The coefficient $|a|$ appears because Lebesgue measure is not preserved by the transformation:

$$\operatorname{mes}\big(T(B)\big) = |a| \operatorname{mes}(B) \tag{7f2}$$

for all Borel sets $B \subset \mathbb{R}$; here $T(x) = ax+b$ and, of course, $T(B) = \{T(x) : x \in B\}$. Formula (7f2) evidently holds for intervals. Its validity for Borel sets follows from 1f11.

Turn to a *linear* 2-dim transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$,

$$T(x, y) = \big(ax + by,\ cx + dy\big)\,. \tag{7f3}$$

If $B = (0, 1) \times (0, 1)$ is the unit square, then $T(B)$ is a parallelepiped of the area $|ad - bc|$. In general,

$$\operatorname{mes}_2\big(T(B)\big) = |J| \operatorname{mes}_2(B) \tag{7f4}$$

for all Borel sets $B \subset \mathbb{R}^2$; here T is given by (7f3), and

$$J = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \tag{7f5}$$

is the so-called Jacobian of T; we assume that $J \ne 0$ (which is equivalent to existence of the inverse transformation T^{-1}). Formula (7f4) is a simple geometric fact for polygons B; therefore it holds (again by 1f11) for all Borel sets. It can be deduced that[103]

$$|J| \cdot f_{U,V}(u, v) = f_{X,Y}(x, y) \qquad \text{when } (u, v) = T(x, y) \tag{7f6}$$

almost everywhere. Here (X, Y) is a 2-dim random variable, (U, V) is another 2-dim random variable, and $(U, V) = T(X, Y)$ is assumed (with probability 1). Existence of $f_{U,V}$ is equivalent to existence of $f_{X,Y}$.

For *nonlinear* smooth 1-dim transformations we have (recall (3b1))

$$\left|\frac{dy}{dx}\right| f_Y(y) = f_X(x) \qquad \text{when } y = \varphi(x) \tag{7f7}$$

provided that the transformation is one-one (continuity of f_X, f_Y, stipulated in Section 3, may be discarded now). Comparing (7f1), (7f6) and (7f7) it is easy to guess that

$$\left|\frac{\partial(u, v)}{\partial(x, y)}\right| f_{U,V}(u, v) = f_{X,Y}(x, y) \qquad \text{when } (u, v) = T(x, y) \tag{7f8}$$

[103] A sandwich argument, similar to the proof of 5b4, is used.

for any (nonlinear) 2-dim one-one transformation T; here

(7f9) $$\frac{\partial(u,v)}{\partial(x,y)} = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix}$$

is a convenient notation for the Jacobian of T. We'll deduce (7f8) from some results of Analysis, formulated below (7f11, 7f12).

Let $D \subset \mathbb{R}^2$ be an open set, and $T : D \to \mathbb{R}^2$ a map, $T(x,y) = \big(u(x,y), v(x,y)\big)$. Assume that functions u, v have continuous first-order partial derivatives on D. Introduce the Jacobian

(7f10) $$J_T(x,y) = \frac{\partial(u,v)}{\partial(x,y)} = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix} .$$

Call T *smooth,* if it is one-one (from D onto $T(D)$) and $J_T(x,y) \ne 0$ for all $(x,y) \in D$.

7f11 Lemma. If T is smooth then $T(D)$ is an open subset of $\mathbb{R}^2$, and the inverse map $T^{-1} : T(D) \to D$ is smooth, and

$$\frac{\partial(u,v)}{\partial(x,y)} \cdot \frac{\partial(x,y)}{\partial(u,v)} = 1 .$$

I give no proof.

7f12 Theorem. (Change of variables). Let $T : D \to \mathbb{R}^2$ be a smooth transformation, $D_1 = T(D)$, and $f : D_1 \to \mathbb{R}$ a Borel function. Then

$$\iint_D f\big(T(x,y)\big) \, |J_T(x,y)| \, dxdy = \iint_{D_1} f(u,v) \, dudv$$

provided that f is integrable on D_1 and the function $(x,y) \mapsto f\big(T(x,y)\big) \, |J_T(x,y)|$ is integrable on D. Otherwise, both functions are non-integrable.

I give no proof.

7f13 Exercise. Prove that $\operatorname{mes}_2(A) = 0 \iff \operatorname{mes}_2\big(T(A)\big) = 0$ for Borel sets $A \subset D$.

Hint. Apply 7f12 to $f = \mathbf{1}_{T(A)}$.

7f14 Theorem. Let $T : D \to \mathbb{R}^2$ be a smooth transformation, $D_1 = T(D)$. Let $X, Y, U, V : \Omega \to \mathbb{R}$ be random variables such that $(X,Y) \in D$ and $T(X,Y) = (U,V)$ almost sure.[104] Then X, Y have a joint density $f_{X,Y}$ if and only if U, V have a joint density $f_{U,V}$, and if they have, then[105]

$$\left| \frac{\partial(u,v)}{\partial(x,y)} \right| f_{U,V}(u,v) = f_{X,Y}(x,y) \qquad \text{when } (u,v) = T(x,y) .$$

[104]Therefore $(U,V) \in D_1$ almost sure.

[105]Of course, $f_{X,Y}$ vanishes outside D, and $f_{U,V}$ vanishes outside D_1.

Proof. Assume that X, Y have a joint density $f_{X,Y}$.[106] Define $f_{U,V}$ on D_1 by $f_{U,V}(u,v) = J_{T^{-1}}(u,v) f\big(T^{-1}(u,v)\big)$, then $f_{U,V}$ satisfies the equality $\big|\frac{\partial(u,v)}{\partial(x,y)}\big| f_{U,V}(u,v) = f_{X,Y}(x,y)$ when $(u,v) = T(x,y)$. We have to prove that $f_{U,V}$ is a density for (U,V), which means

$$\iint_A f_{U,V}(u,v)\,dudv = P_{U,V}(A)$$

for every Borel set $A \subset D_1$. Apply 7f12 to $\mathbf{1}_A f_{U,V}$:

$$\iint_D \mathbf{1}_A\big(T(x,y)\big) \underbrace{f_{U,V}\big(T(x,y)\big)|J_T(x,y)|}_{f_{X,Y}(x,y)}\,dxdy = \iint_{D_1} \mathbf{1}_A(u,v) f_{U,V}(u,v)\,dudv\,;$$

$$\iint_{T^{-1}(A)} f_{X,Y}(x,y)\,dxdy = \iint_A f_{U,V}(u,v)\,dudv\,;$$

the left-hand side is equal to $\mathbb{P}\big(\,(X,Y) \in T^{-1}(A)\,\big) = \mathbb{P}\big(\,(U,V) \in A\,\big) = P_{U,V}(A)$. □

Some examples:

$$\begin{matrix} U = X + Y \\ V = X \end{matrix} \quad\Longrightarrow\quad f_{U,V}(u,v) = f_{X,Y}(x,y)\,;$$

$$\begin{matrix} X = R\cos\Phi \\ Y = R\sin\Phi \end{matrix} \quad\Longrightarrow\quad r f_{X,Y}(x,y) = f_{R,\Phi}(r,\varphi)\,.$$

Some implications:

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_{X,Y}(x, u-x)\,dx\,;$$

$$f_{\sqrt{X^2+Y^2}}(r) = r\int_0^{2\pi} f_{X,Y}(r\cos\varphi, r\sin\varphi)\,d\varphi\,;$$

$$f_\Phi(\varphi) = \int_0^\infty r f_{X,Y}(r\cos\varphi, r\sin\varphi)\,dr\,.$$

Generalizations for higher dimensions d are straightforward; they involve determinants of $d \times d$ matrices.

7g Some paradoxes, remarks etc

7g1 Exercise. There exist random variables X, Y taking on values in $\{1, 2, 3, \dots\}$ such that

$$\mathbb{E}\big(\,Y\,\big|\,X\,\big) > X \quad\text{but also}\quad \mathbb{E}\big(\,X\,\big|\,Y\,\big) > Y$$

with probability 1 (the conditional distributions being integrable). Find an example. (Hint: $\mathbb{P}\big(\,X = 2^{k-1}, Y = 2^k\,\big) = \mathbb{P}\big(\,Y = 2^{k-1}, X = 2^k\,\big) = \frac{1}{2}(1-p)p^{k-1}$.)

However, for integrable X, Y it cannot happen. Prove it. What about continuous X, Y?

[106]The other case, assuming that U, V have a joint density, is similar (with T^{-1} instead of T).

7g2 Exercise. There exist random variables X, Y, Z taking on values in $\{\dots, -2, -1, 0, 1, 2, \dots\}$ such that

$$\mathbb{E}\left(X \,\middle|\, Y \right) > 0 \quad \text{however} \quad \mathbb{E}\left(X \,\middle|\, Z \right) < 0$$

with probability 1 (the conditional distributions being integrable). Find an example. (Hint: $\mathbb{P}\left(X = (-2)^{k-1} \right) = (1-p)p^{k-1}$; Y is the integral part ("floor") of $k/2$, Z — of $(k-1)/2$.)

However, for integrable X it cannot happen. Prove it.

Remark. Think, what would you prefer: winning 10^{10} dollars with probability 10^{-3}, or winning 10^{100} dollars with probability 10^{-4}?

7g3 Exercise. The conditional expectation $\mathbb{E}\left(Y \,\middle|\, X \right)$ is the *optimal predictor*[107] for Y from X, that is, a random variable of the form $\hat{Y} = \varphi(X)$ that minimizes (over all Borel functions φ) the mean square error $\mathbb{E}\,(\hat{Y} - Y)^2$. Prove it under some appropriate assumptions about X, Y. (Hint: apply Theorem 7d1 to $(\hat{Y} - Y)^2$.)

Remark. A statistician often knows (with a reasonable precision) the correlation coefficient, but not the joint density.

7g4 Exercise. Let X, Y have a joint density $f_{X,Y}$. Then

(a) The conditional density of X, given $Y = 0$, is equal to

$$\frac{f_{X,Y}(x,0)}{f_Y(0)} = \frac{f_{X,Y}(x,0)}{\int f_{X,Y}(x_1,0)\,dx_1} \,.$$

(b) The conditional density of X, given $Y/X = 0$, is equal to

$$\frac{|x| f_{X,Y}(x,0)}{\int |x_1| f_{X,Y}(x_1,0)\,dx_1} \,.$$

(c) Conditions $Y = 0$ and $Y/X = 0$ are equivalent.

Do you agree with (a), (b), (c)? Are they consistent?

Remark. If someone told you that he observed a zero value of a nonatomic random variable, do not believe.

7g5 Exercise. Consider the uniform distribution on the circle $\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$, and the Borel subset $B = \{\left(\cos(\pi a), \sin(\pi a)\right) : a \text{ is rational }\}$ of the circle.

(a) The distribution is invariant under rotations (around the origin).

(b) The set B is invariant under rotations by πa for all rational a.

(c) Thus, the conditional distribution on B[108] is invariant under rotations by πa for all rational a.

Do you agree with (a), (b), (c)? Can you use (c) for calculating the conditional distribution?

Remark. If someone told you that he observed a *rational* value of a nonatomic random variable, do not believe.

[107] Not just linear.

[108] That is, the conditional distribution of a random point of the circle, given that the point belongs to B.