Advanced Statistics

3. Moments of Random Variables

Christian Aßmann

Chair of Survey Statistics and Data Analysis – Otto-Friedrich-Universität Bamberg

The expected value, or expectation, of a random variable represents its average value and can be thought of as a measure of the center of its pdf.

DEFINITION (EXPECTATION; DISCRETE CASE): The expected value of a discrete random variable exists, and is defined by

$$EX = \sum_{x \in R(X)} x \cdot f(x), \quad \text{iff} \quad \sum_{x \in R(X)} |x \cdot f(x)| = \sum_{x \in R(X)} |x| \cdot f(x) < \infty.$$

REMARK: The existence condition ensures that the sum $\sum_{x \in R(X)} xf(x)$ defining the expectation is absolutely convergent.

Furthermore, note that absolute convergence implies standard convergence, that is

$$\sum_{x \in R(X)} |x| \cdot f(x) < \infty \quad \Rightarrow \quad \Big| \sum_{x \in R(X)} x \cdot f(x) \Big| < \infty.$$

such that the sum defining the expectation is finite and exists.

(CONTINUES)

Remark (continued): Also note that, if

$$R(X) \text{ is finite and } |x| < \infty \; \forall \; x \in R(X) \qquad \Rightarrow \qquad \sum_{x \in R(X)} |x| \cdot f(x) < \infty,$$

such that the existence condition is automatically satisfied; but if

$R(X)$ is countably infinite there is no guarantee that $\sum_{x \in R(X)} |x| \cdot f(x) < \infty$.

Finally note that standard convergence does not ensure the uniqueness of the converged value in the countably infinite case. This means that a change of the ordering of the terms in an infinite sum can result in a change of the value of the sum. The absolute convergence ensures the uniqueness of the converged value (see Ex. below). $\diamondsuit$

EXAMPLE: Consider the experiment of rolling a die, and recall the pdf for the number of dots facing up given by

$$f(x) = \frac{1}{6}\mathbb{I}_{\{1,2,\ldots,6\}}(x) \quad \text{with} \quad R(X) = \{1, 2, \ldots, 6\}.$$

The expected value equals $\mathrm{E}X = \sum_{x=1}^{6} x \frac{1}{6}\mathbb{I}_{\{1,2,\ldots,6\}}(x) = 3.5.$ $\quad ||$

EXAMPLE: Consider a random variable with pdf

$$f(x_k) = \frac{1}{2^k} \quad \text{with} \quad R(X) = \{x_k = (-1)^k \frac{2^k}{k}, k = 1, 2, \ldots\}.$$

The sum defining the expectation is

$$\begin{aligned} \sum_{k=1}^{\infty} x_k f(x_k) &= \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \\ &= -\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} 1^k \\ &= -\ln(1+1). \qquad \left[\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} x^k \Big|_{x\in(-1,1]} = \ln(1+x) \right] \end{aligned}$$

Thus, the sum is convergent. (CONTINUES)

Example (continued): But the sum is not absolutely convergent since

$$\sum_{k=1}^{\infty} |x_k| f(x_k) = \sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>1/2} + \underbrace{\frac{1}{5} + \cdots + \frac{1}{8}}_{>1/2} + \cdots = \infty.$$

Thus, the uniqueness of the converged value (and hence of the expected value) is not ensured. $||$

The expectation of a continuous random variable is defined as follows:

DEFINITION (EXPECTATION; CONTINUOUS CASE): The expected value of a continuous random variable exists, and is defined by

$$EX = \int_{-\infty}^{\infty} x \cdot f(x)\,dx, \quad \text{iff} \quad \int_{-\infty}^{\infty} |x| \cdot f(x)\,dx < \infty.$$

REMARK: The existence condition is necessary to ensure that the improper Riemann integral $\int_{-\infty}^{\infty} x \cdot f(x)\,dx$ (and hence the expectation) exists. $\diamondsuit$

EXAMPLE: Consider the pdf $f(x) = 3x^2\mathbb{I}_{[0,1]}(x)$. The expected value equals

$$EX = \int_{-\infty}^{\infty} x \cdot 3x^2\mathbb{I}_{[0,1]}(x)dx = 3\int_0^1 x^3 dx = 0.75. \quad ||$$

EXAMPLE: Consider a random variable with pdf

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \qquad -\infty < x < \infty \qquad \text{(Cauchy distribution).}$$

This is the classical example of a random variable whose expected value does not exist. In order to see this, write

$$\int_{-\infty}^{\infty} |x| f(x) dx = \int_{-\infty}^{\infty} \frac{|x|}{\pi}\frac{1}{1+x^2} dx = \frac{2}{\pi}\int_0^{\infty} \frac{x}{1+x^2} dx,$$

For any positive number a we obtain

$$\int_0^a \frac{x}{1+x^2} dx = \left[\frac{\ln(1+x^2)}{2}\right]_{x=0}^{x=a} = \frac{\ln(1+a^2)}{2}.$$

Thus,

$$\int_{-\infty}^{\infty} |x| f(x) dx = \lim_{a\to\infty} \frac{2}{\pi}\int_0^a \frac{x}{1+x^2} dx = \frac{1}{\pi}\lim_{a\to\infty} \ln(1+a^2) = \infty. \qquad ||$$

In applications, the following result w.r.t. sufficient conditions for the existence of the expectation can be useful:

THEOREM 3.1 *If* $|x| < c \; \forall \; x \in R(X)$*, for some choice of* $c \in (0, \infty)$*. Then* $\mathrm{E}X$ *exists.*

PROOF: For a discrete random variable we obtain

$$\sum_{x \in R(X)} |x| \cdot f(x) \Big|_{|x|<c} < \sum_{x \in R(X)} c \cdot f(x) = c \cdot \sum_{x \in R(X)} f(x) = c < \infty;$$

The proof for the continuous case is analogous. $\square$

REMARK: The theorem indicates that the expectation exists if the outcomes of the random variables are bounded. $\diamondsuit$

In many situations we are interested in the expectation of a function of a random variable rather than the expectation of the random variable itself.

Consider, for example, the revenue of a company $Y = p \cdot X$, where p is the selling price, which is fixed, and X represents the (random) selling. How might $\mathrm{E}Y = \mathrm{E}(p \cdot X)$ be determined ?

The following theorem identifies a straightforward approach of obtaining the expectation of a function $Y = g(X)$ of a random variable X.

Theorem 3.2 *Let X be a random variable with pdf $f(x)$. Then the expectation of random variable $Y = g(X)$ is given by*

$$Eg(X) = \begin{cases} \displaystyle\sum_{x \in R(X)} g(x) f(x) & \text{(discrete)} \\ \displaystyle\int_{-\infty}^{\infty} g(x) f(x)\, dx & \text{(continuous).} \end{cases}$$

Proof: (Discrete Case) Since the outcome y is equivalent to the event $x \in \{x : g(x) = y\}$, the pdf of $Y = g(X)$, say h, can be represented as

$$h(y) = P_Y(y) = P_X(\{x : g(x) = y, x \in R(X)\}) = \sum_{x \in \{x : g(x) = y\}} f(x).$$

Thus,

$$\begin{aligned} \mathrm{E}g(X) = \mathrm{E}Y &= \sum_{y \in R(Y)} y \cdot h(y) = \sum_{y \in R(Y)} y \cdot \sum_{x \in \{x : g(x) = y\}} f(x) \\ &= \sum_{y \in R(Y)} \sum_{x \in \{x : g(x) = y\}} g(x) \cdot f(x) \qquad (y\text{: fixed value in the inner sum}) \\ &= \sum_{x \in R(X)} g(x) \cdot f(x) \qquad \left(\textstyle\sum_{y \in R(Y)} \sum_{x \in \{x : g(x) = y\}} \text{ is equivalent to summing over all } x \in R(X)\right). \end{aligned}$$

PROOF (CONTINUED): The proof for the continuous case is analogous. □

EXAMPLE: Consider the experiment of rolling a die and the number of dots facing up denoted by X. The expectation of the function $Y = g(X) = X^2$ is

$$\mathsf{E}X^2 = \sum_{x=1}^{6} x^2 \cdot \frac{1}{6} \mathbb{I}_{\{1,2,\ldots,6\}}(x) = \frac{91}{6}. \qquad ||$$

REMARK: An implication of Theorem 3.2 is that the expectation of an indicator function equals the probability of the set being indicated:

Let X be a variable with pdf f and define

$$\mathbb{I}_A(x) = \begin{cases} 1 & x \in A \\ 0 & \text{else.} \end{cases}$$

Then,

$$\mathrm{E}\mathbb{I}_A(X) = \begin{cases} \displaystyle\sum_{x \in R(X)} \mathbb{I}_A(x) \cdot f(x) = \sum_{x \in A} f(x) = P(x \in A) & \text{(discrete case)} \\ \displaystyle\int_{x \in R(X)} \mathbb{I}_A(x) \cdot f(x) dx = \int_{x \in A} f(x) dx = P(x \in A) & \text{(continuous case).} \end{cases}$$

It follows that probabilities can be represented as expectations. $\diamondsuit$

The following theorem indicates that, in general $\mathrm{E}g(X) \neq g(\mathrm{E}X)$:

THEOREM 3.3 (JENSEN'S INEQUALITY) *Let X be a non-degenerate random variable*[1] *with expectation EX, and let g be a continuous function on an open interval I containing $R(X)$ (that is $R(X) \subseteq I$).*

$$\begin{aligned} \text{If } g \text{ is convex on } I, &\qquad \text{then } Eg(X) \geq g(EX); \\ \text{if } g \text{ is strictly convex on } I, &\qquad \text{then } Eg(X) > g(EX). \end{aligned}$$

[1] A degenerate random variable has only one outcome that is assigned a probability of 1.

PROOF: Let $\ell(x)$ be a tangent to $g(x)$ at point $g(\mathrm{E}X)$, say $\ell(x) = a + bx$ (see Fig.10). Now, if g is convex on I ($g'' \geq 0$), we have

$$g(x) \;\geq\; \ell(x) = a + bx \quad \forall x \in I.$$

Thus, for a (discrete) X with pdf f, we obtain

$$\begin{aligned}
\mathrm{E}g(X) = \sum_{x \in R(X)} g(x) f(x) \quad &\geq \quad \sum_{x \in R(X)} (a + bx) f(x) = a + b\mathrm{E}X \\
&= \quad \ell(\mathrm{E}X) \qquad (\text{def. of } \ell(x)) \\
&= \quad g(\mathrm{E}X) \qquad (\ell \text{ is tangent at } \mathrm{E}X),
\end{aligned}$$

so that $\mathrm{E}g(X) \geq g(\mathrm{E}X)$, as was to be shown (for a continuous X analogous).

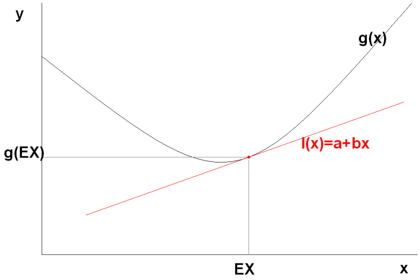


Fig. 10.

(CONTINUES)

PROOF (CONTINUED): Now, if g is strictly convex on I ($g'' > 0$), we have

$$g(x) > \ell(x) = a + bx \qquad \forall x \in I \quad \text{for which } x \neq \mathrm{E}X.$$

Then, assuming that no $x \in R(X)$ is assigned probability one (this means that X is non-degenerate), the previous inequality results become strict, implying

$$\mathrm{E}g(X) > g(\mathrm{E}X), \qquad \text{as was to be shown.} \qquad \square$$

REMARK: Jensen's Inequality also applies to concave functions. If g is concave, then $\mathrm{E}g(X) \leq g(\mathrm{E}X)$. $\diamondsuit$

EXAMPLE: One immediate application of Jensen's Inequality shows that

$$\mathrm{E}X^2 \geq (\mathrm{E}X)^2, \qquad \text{since } g(x) = x^2 \text{ is convex.}$$

Note that this implies that $\mathrm{var}(X) = \mathrm{E}X^2 - (\mathrm{E}X)^2 \geq 0$. $\|$

Some Properties of the Expectation Operation

THEOREM 3.4 *If c is a constant, then $E(c) = c$.*

PROOF: $\mathrm{E}(c) = \int_{-\infty}^{\infty} c \cdot f(x)dx = c\int_{-\infty}^{\infty} f(x)dx = c.$ $\square$

THEOREM 3.5 *If c is a constant, then $E(cX) = cEX$.*

PROOF: $\mathrm{E}(cX) = c\int_{-\infty}^{\infty} x \cdot f(x)dx = c\mathrm{E}X.$ $\square$

THEOREM 3.6 $E\sum_{i=1}^{k} g_i(X) = \sum_{i=1}^{k} Eg_i(X)$.

PROOF: Let $g(X) = \sum_{i=1}^{k} g_i(X)$. Then, by Theorem 3.2,

$$\mathrm{E}\sum_{i=1}^{k} g_i(X) = \underbrace{\int_{-\infty}^{\infty}\sum_{i=1}^{k} g_i(x)f(x)dx = \sum_{i=1}^{k}\int_{-\infty}^{\infty} g_i(x)f(x)dx}_{\text{(additivity property of Riemann integrals)}} = \sum_{i=1}^{k}\mathrm{E}g_i(X). \quad \square$$

Note that Theorem 3.6 indicates that the expectation of a sum is the sum of the expectations.

COROLLARY 3.1 $E(a + bX) = a + bEX$.

PROOF: This follows directly from Theorem 3.6, by defining $g_1(X) = a$ and $g_2(X) = bX$. $\square$

Multivariate Extensions

So far, we considered the expectation of a function of a univariate random variable. This concept can be generalized to a function of a multivariate random variable as indicated in the following theorem:

THEOREM 3.7 *Let $(X_1, ..., X_n)$ be a multivariate random variable with joint pdf $f(x_1, ..., x_n)$. Then the expectation of random variable $Y = g(X_1, ..., X_n)$ is given by*

$$EY = \begin{cases} \sum \cdots \sum_{(x_1,\ldots,x_n)\in R(X)} g(x_1, ..., x_n) f(x_1, ..., x_n) & \text{(discrete)} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, ..., x_n) f(x_1, ..., x_n)\, dx_1 \cdots dx_n & \text{(continuous).} \end{cases}$$

PROOF: This follows from a direct extension of the proof of Theorem 3.2 for the univariate case. $\square$

EXAMPLE: Consider a bivariate random variable (X_1, X_2) with joint pdf

$$f(x_1, x_2) = 6x_1 x_2^2 \mathbb{I}_{[0,1]}(x_1)\mathbb{I}_{[0,1]}(x_2),$$

and the function $g(x_1, x_2) = .5(x_1 + x_2)$. Then, by Theorem 3.7,

$$\mathrm{E}g(X_1, X_2) = \int_0^1 \int_0^1 \left[\frac{1}{2}(x_1 + x_2)\right] \cdot \left[6x_1 x_2^2\right] dx_1 dx_2 = .7083. \qquad ||$$

The expectation property in Theorem 3.6 concerning the sum of functions of univariate random variables can be extended to sums of functions of multivariate random variables. This is indicated in the following theorem:

THEOREM 3.8 $E\sum_{i=1}^{k} g_i(X_1,\ldots,X_n) = \sum_{i=1}^{k} Eg_i(X_1,\ldots,X_n)$.

PROOF: (Continuous case) Let $g(X_1,\ldots,X_n) = \sum_{i=1}^{k} g_i(X_1,\ldots,X_n)$. Then, by Theorem 3.7 and the additivity property of Riemann integrals,

$$\begin{aligned} \mathrm{E}\sum_{i=1}^{k} g_i(X_1,\ldots,X_n) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\sum_{i=1}^{k} g_i(x_1,\ldots,x_n)f(x_1,\ldots,x_n)dx_1\cdots dx_n \\ &= \sum_{i=1}^{k}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g_i(x_1,\ldots,x_n)f(x_1,\ldots,x_n)dx_1\cdots dx_n \\ &= \sum_{i=1}^{k}\mathrm{E}g_i(X_1,\ldots,X_n) \end{aligned}$$

(discrete case analogous). □

COROLLARY 3.2 *$E\sum_{i=1}^{k} X_i = \sum_{i=1}^{k} EX_i$ (Expectation of a sum is the sum of the expectations).*

PROOF: This follows by Theorem 3.8 with $g_i = X_i$. □

In the case that the random variables are independent, the expectation of a product is the product of the expectations as indicated in the following theorem:

THEOREM 3.9 *Let $X_1, \dots X_n$ be independent random variables. Then $E\prod_{i=1}^n X_i = \prod_{i=1}^n EX_i$.*

PROOF: (Continuous case) Let $g(X_1, \dots, X_n) = \prod_{i=1}^n X_i$. Then, by Theorem 3.7 we have

$$\begin{aligned}
\mathrm{E}\prod_{i=1}^n X_i &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\prod_{i=1}^n x_i\right] \underbrace{f(x_1, \dots, x_n)}_{=\prod_{i=1}^n f_i(x_i) \text{ by independence}} dx_1 \cdots dx_n \\
&= \prod_{i=1}^n \int_{-\infty}^{\infty} x_i f(x_i) dx_i \\
&= \prod_{i=1}^n \mathrm{E}X_i
\end{aligned}$$

(discrete case analogous). $\square$

3.3 Conditional Expectation

So far, we have considered unconditional expectations, this means the expectations of unconditional/marginal distributions. If we take the expectation w.r.t. a conditional distribution, we have the conditional expectation.

The conditional expectation is one of the most important concepts used in econometrics and empirical economics, and is the key element of regression analysis.

Regression analysis attempts to explain the expectation of a random variable as a function of related random variables, for example, the expected consumption of a household as a function of the realized household income.

DEFINITION (**CONDITIONAL EXPECTATION**): Let $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ be random vectors with joint pdf $f(x_1, \ldots, x_n, y_1, \ldots, y_m)$. Let $g(Y_1, \ldots, Y_m)$ be a real-valued function of $(Y_1, \ldots, Y_m)$.

Then the conditional expectation of $g(Y_1, \ldots, Y_m)$, given $(x_1, \ldots, x_n) \in B$, is defined as

(discrete) $\quad \mathrm{E}\left[g(Y_1, \ldots, Y_m) \mid (x_1, \ldots, x_n) \in B\right]$

$$= \sum_{(y_1, \ldots, y_m) \in R(Y)} \cdots \sum g(y_1, \ldots, y_m)\, f(y_1, \ldots, y_m \mid (x_1, \ldots, x_n) \in B)$$

(continuous) $\quad \mathrm{E}\left[g(Y_1, \ldots, Y_m) \mid (x_1, \ldots, x_n) \in B\right]$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, \ldots, y_m)\, f(y_1, \ldots, y_m \mid (x_1, \ldots, x_n) \in B)\, dy_1 \cdots dy_m.$$

REMARK: An important special case of the definition given above obtains by setting $g(Y_1, ..., Y_n) = Y$, where Y is a univariate random variable,

$$\mathrm{E}[Y \mid (x_1, \ldots, x_n) \in B] = \begin{cases} \displaystyle\sum_{y \in R(Y)} y \cdot f\left(y \mid (x_1, \ldots, x_n) \in B\right) & \text{(discrete)} \\ \displaystyle\int_{-\infty}^{\infty} y \cdot f\left(y \mid (x_1, \ldots, x_n) \in B\right) dy & \text{(continuous).} \end{cases} \quad \diamondsuit$$

EXAMPLE: Consider a bivariate random variable (X, Y) with joint pdf

$$f(x, y) = \frac{1}{96}(x^2 + 2xy + 2y^2)\mathbb{I}_{[0,4]}(x)\mathbb{I}_{[0,2]}(y).$$

What is the conditional expectation $E(Y|x = 1)$?

To answer this, we need to establish $f(y|x = 1) = f(x, y)/f_X(x)\Big|_{x=1}$.

The marginal pdf for X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy = \left[\frac{1}{48}x^2 + \frac{1}{24}x + \frac{1}{18}\right] \mathbb{I}_{[0,4]}(x),$$

such that $\qquad f(y|x = 1) = \left[.088235 + .176471(y + y^2)\right] \mathbb{I}_{[0,2]}(y).$

(CONTINUES)

EXAMPLE (CONTINUED): Thus, we find that

$$\mathrm{E}(Y|x=1) = \int_{-\infty}^{\infty} y \cdot f(y|x=1)dy = \int_0^2 y \cdot \left[.088235 + .176471(y+y^2)\right] dy = 1.3529. \quad ||$$

REMARK: *All* properties of expectations discussed above also apply analogously to conditional expectations. $\diamondsuit$

REMARK: The conditional expectation $\mathrm{E}(Y|(x_1,...,x_n) \in B)$ was introduced as being conditional on a *particular event* $(x_1,...,x_n) \in B$.

Rather than specifying a particular event, we might conceptualize leaving the event for $(X_1,...,X_n)$ *unspecified* and interpret the conditional expectation of Y as a function of $(X_1,...,X_n)$ denoted by $\mathrm{E}(Y|X_1,...,X_n)$.

Note that $\mathrm{E}(Y|X_1,...,X_n)$ is then a function of random variables and, therefore, itself a random variable.

$\mathrm{E}(Y|x_1,...,x_n)$ is referred to as the regression function of a regression of Y on the X_is. $\diamondsuit$

EXAMPLE: Recall the Ex. of the bivariate random variable with joint pdf

$$f(x,y) = \frac{1}{96}(x^2 + 2xy + 2y^2)\mathbb{I}_{[0,4]}(x)\mathbb{I}_{[0,2]}(y).$$

The regression function of a regression of Y on X is obtained as

$$\begin{aligned} \mathrm{E}(Y|x) &= \int_{-\infty}^{\infty} y \cdot \frac{f(x,y)}{f_X(x)}\,dy = \int_0^2 \frac{y \cdot (x^2 + 2xy + 2y^2)\mathbb{I}_{[0,4]}(x)}{(2x^2 + 4x + \frac{16}{3})\mathbb{I}_{[0,4]}(x)}\,dy \\ &= \frac{2x^2 + \frac{16}{3}x + 8}{2x^2 + 4x + \frac{16}{3}} \quad \text{for} \quad x \in [0,4]. \end{aligned}$$

For $x \notin [0,4]$, the regression function is not defined.

Note that the regression function is a nonlinear function in x. $||$

The following theorem (referred to as the law of iterated expectation) indicates how we obtain the unconditional expectation of $g(Y)$ from the conditional expectation of the random variable $g(Y)$ conditional on the random variable X.

THEOREM 3.11 $E[E(g(Y)|X)] = Eg(Y)$.

PROOF: (Continuous case) Let $f(x, y)$ be the joint pdf of X and Y, $f_X(x)$ the marginal pdf of X, and $f(y|x)$ the conditional pdf. Then

$$\begin{aligned}
\mathrm{E}[\, \underbrace{\mathrm{E}(g(Y)|X)}_{\text{random variable !}} \,] &= \int_{R(X)} [\mathrm{E}(g(Y)|x)] \cdot f_X(x)dx = \int_{R(X)} \left[\int_{R(Y)} g(y) \cdot f(y|x)dy \right] \cdot f_X(x)dx \\
&= \int_{R(X)} \int_{R(Y)} g(y) \cdot f(x, y)dydx \qquad (f(y|x)f_X(x) = f(x, y)) \\
&= \int_{R(Y)} g(y) \left[\int_{R(X)} f(x, y)dx \right] dy \qquad (\textstyle\int f(x, y)dx = f_Y(y)) \\
&= \mathrm{E}g(Y).
\end{aligned}$$

(Discrete case analogous). $\square$

REMARK: The law of iterated expectations straightforwardly extends to the case where the random variables Y and/or X are multivariate (being random vectors). In particular, we get

$$\mathsf{E}[\mathsf{E}(g(Y_1, ... Y_n)|X_1, ..., X_n)] = \mathsf{E}g(Y_1, ..., Y_n). \quad \diamondsuit$$

Moments of random variables are expectations of power functions of random variables. They can be used to measure certain characteristics of the pdf of the random variable, for example, the dispersion and skewness.

There are two types of moments, namely, non-central and central moments.

DEFINITION (*r*TH NON-CENTRAL MOMENT): Let X be a random variable with pdf $f(x)$. Then the rth non-central moment of X, denoted by μ_r', is defined as

$$\mu_r' = \mathrm{E}\left(X^r\right) = \begin{cases} \displaystyle\sum_{x \in R(X)} x^r f(x) & \text{(discrete)} \\ \displaystyle\int_{-\infty}^{\infty} x^r f(x)\, dx & \text{(continuous).} \end{cases}$$

REMARK: The first non-central moment is simply the expectation (also called the mean) of the random variable, that is $\mu_1' = \mathrm{E}(X)$, and will be denoted by the symbol μ .

Furthermore note that $\mu_0' = \mathrm{E}\left(X^0\right) = 1$. $\diamondsuit$

Definition (**rth central moment**): Let X be a random variable with pdf $f(x)$. Then the rth central moment of X, denoted by μ_r, is defined as

$$\mu_r = \mathrm{E}\,(X-\mu)^r = \begin{cases} \displaystyle\sum_{x \in R(X)} (x-\mu)^r f(x) & \text{(discrete)} \\ \displaystyle\int_{-\infty}^{\infty} (x-\mu)^r f(x)\,dx & \text{(continuous).} \end{cases}$$

Remark: Note that $\mu_0 = \mathrm{E}(X-\mu)^0 = 1$, and $\mu_1 = \mathrm{E}(X-\mu) = 0$. $\diamondsuit$

The second central moment is commonly known as the variance.

Definition (**Variance and standard deviation**): The **variance** of a random variable X is the 2nd central moment, $\mathrm{var}(X) = \mathrm{E}\,(X-\mu)^2$, and will be denoted by the symbol σ^2.

The non-negative square root of $\mathrm{var}(X)$ is the **standard deviation** of X and will be denoted by the symbol σ.

The variance and standard deviation are measures of the dispersion of a distribution around the mean. The larger the variance, the larger the dispersion.

The relationship between the variance and the dispersion can be examined by means of **Chebyshev's inequality** which is a special case of **Markov's inequality**.

THEOREM 3.12 (MARKOV'S INEQUALITY) *Let X be a random variable with pdf f, and let g be a nonnegative function of X. Then*

$$P(g(x) \geq a) \leq \frac{Eg(X)}{a} \qquad \text{for any } a > 0.$$

PROOF: (Discrete case) We can decompose $\mathrm{E}g(X)$ into

$$\begin{aligned}
\mathrm{E}g(X) &= \sum_{x\in R(X)} g(x)f(x) = \underbrace{\sum_{\{x:g(x)<a\}} \underset{(\geq 0)(\geq 0)}{g(x)f(x)}}_{\geq 0} + \sum_{\{x:g(x)\geq a\}} g(x)f(x) \\
&\geq \sum_{\{x:g(x)\geq a\}} g(x)f(x) \\
&\geq \sum_{\{x:g(x)\geq a\}} af(x) \qquad (\text{since } g(x)\geq a \ \forall\ x\in\{x: g(x)\geq a\}) \\
&= a\sum_{\{x:g(x)\geq a\}} f(x) \ = \ aP(g(x)>a),
\end{aligned}$$

and thus $\frac{\mathrm{E}g(X)}{a} \geq P(g(x)\geq a)$. (Continuous case analogous). $\square$

REMARK: Markov's Inequality states that the probability for $g(x) \geq a$ has **always** an upper bound (independent of the probability distribution) as long as $g(x)$ is non-negative valued (see Fig.11).

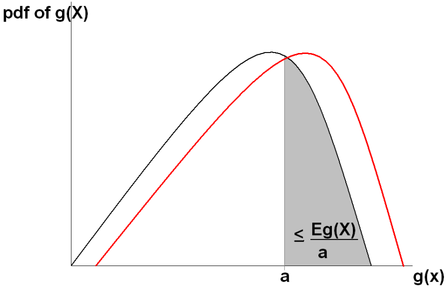


Fig. 11.

Note that the upper bound for the probability is increasing with the expectation $\mathrm{E}g(X)$. $\diamondsuit$

As a special case of Markov's Inequality we obtain Chebyshev's Inequality.

Corollary 3.3 (Chebyshev's inequality)

$$P(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \textit{for} \quad k > 0.$$

Proof: Let

$$g(x) = \frac{(x-\mu)^2}{\sigma^2} \geq 0, \qquad \text{where} \qquad \mathrm{E}g(X) = \frac{\mathrm{E}(X-\mu)^2}{\sigma^2} = 1,$$

and for convenience, set $a = k^2 > 0$. Then Markov's inequality implies

$$P\left(\frac{(x-\mu)^2}{\sigma^2} \geq k^2\right) \leq \frac{\mathrm{E}\,\frac{(X-\mu)^2}{\sigma^2}}{k^2} = \frac{1}{k^2}.$$

Doing some obvious algebra, we get the inequality

$$P(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \quad \square$$

Remark: From obvious algebra, we also get $P(|x-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}$. $\diamondsuit$

Chebyshev's Inequality allows us to examine the relationship between the variance σ^2 and the dispersion of a pdf.

For this purpose, set in Chebyshev's Inequality $k\sigma = c$, where $c > 0$. Then

$$P(|x-\mu| \geq c) \leq \left. \frac{\sigma^2}{c^2} \right|_{\sigma^2 \to 0} \to 0.$$

This implies that as $\sigma^2 \to 0$, the pdf concentrates over the interval $(\mu - c, \mu + c)$ for any arbitrarily small positive c (see Fig.12).

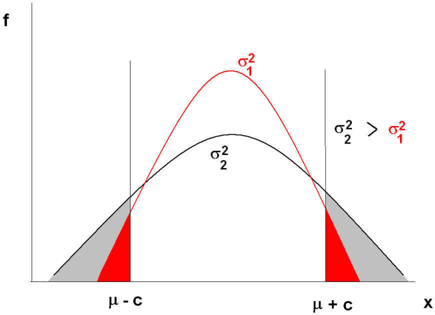


Fig. 12.

The third central moment $E(X-\mu)^3$ can be used as a measure of **skewness** of the pdf of X, that means the deviation from symmetry around μ.

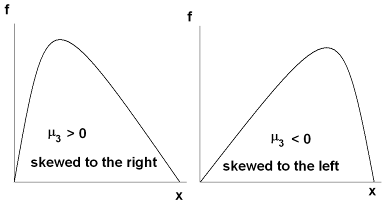
DEFINITION (SYMMETRY OF A PDF): The pdf f is said to be symmetric around μ iff

$$f(\mu+\delta)=f(\mu-\delta) \quad \text{for any} \quad \delta>0.$$

Otherwise f is said to be skewed.

- A symmetric pdf has $\mu_3 = E(X-\mu)^3 = 0$ [2].
- For $\mu_3 > 0$ ($\mu_3 < 0$) the pdf said to be skewed to the right (left)- see Fig.13.

Fig. 13.



[2]Though the condition $\mu_3 = 0$ is necessary for a symmetric pdf, it is not sufficient - see Mittelhammer (1996, p.136).

Existence of Moments

With respect to the existence of non-central moments, the following theorem is useful:

THEOREM 3.13 *If EX^r exists for an $r > 0$, then EX^s exists $\forall s \in [0, r]$.*

The theorem implies that if we can show the existence of the r-th order non-central moment, then all lower-order non-central moments are known to exist.

PROOF: (Continuous Case) We need to show that if $\int_{-\infty}^{\infty} |x|^r f(x)dx < \infty$, then $\int_{-\infty}^{\infty} |x|^s f(x)dx < \infty$ for $s \leq r$. For this purpose, define the sets

$$A_{<1} = \{x : |x|^s < 1\} \qquad \text{and} \qquad A_{\geq 1} = \{x : |x|^s \geq 1\},$$

$$\text{such that} \qquad \int_{-\infty}^{\infty} |x|^s f(x)dx = \int_{x \in A_{<1}} |x|^s f(x)dx + \int_{x \in A_{\geq 1}} |x|^s f(x)dx.$$

Since $f(x) \geq |x|^s f(x) \; \forall \; x \in A_{<1}$, we can write

$$P(|x|^s < 1) = \int_{x \in A_{<1}} f(x)dx \;\; \geq \;\; \int_{x \in A_{<1}} |x|^s f(x)dx.$$

Now note that, for $r > s$, we have $|x|^r \geq |x|^s \; \forall \; x \in A_{\geq 1}$. It follows that

$$\int_{x \in A_{\geq 1}} |x|^r f(x)dx \;\; \geq \;\; \int_{x \in A_{\geq 1}} |x|^s f(x)dx.$$

Finally, we have

$$\begin{aligned} \int_{-\infty}^{\infty} |x|^s f(x)dx \quad &\leq \quad P(|x|^s < 1) + \int_{x \in A_{\geq 1}} |x|^r f(x)dx \\ &\leq \quad P(|x|^s < 1) + \int_{-\infty}^{\infty} |x|^r f(x)dx \qquad (\text{since } \textstyle\int_{x \in A_{<1}} |x|^r f(x)dx \geq 0 \text{ }) \\ &< \quad \infty \qquad\qquad (\text{since } P(|x|^s < 1) \in [0,1] \text{ and } EX^r \text{ exists }). \end{aligned}$$

PROOF (CONTINUED): The proof for the discrete case is analogous. □

REMARK: Theorem 3.13 also implies, that if $\mathrm{E}X^r$ does not exist, then necessarily $\mathrm{E}X^k$ cannot exist for $k > r$. Otherwise, Theorem 3.13 would be contradicted. ♢

EXAMPLE: Consider the pdf

$$f(x) = \frac{2}{(x+1)^3}\mathbb{I}_{[0,\infty)}(x).$$

Examine $\mathrm{E}X^\alpha$, that is

$$\mathrm{E}X^\alpha = \int_0^\infty \frac{x^\alpha 2}{(x+1)^3}dx = 2\int_1^\infty (y-1)^\alpha y^{-3} dy,$$

(the 2nd Eq. obtains by substituting $y = x + 1$, so that $y - 1 = x$ and $dy = dx$). If $\alpha = 2$, we get

$$\mathrm{E}X^2 = 2\lim_{b\to\infty}\left[\ln(y) + 2y^{-1} - \frac{1}{2}y^{-2}\right]_{y=1}^{y=b} = \infty.$$

Thus, $\mathrm{E}X^2$ does not exist. By Theorem 3.13, moments of order greater than 2 also do not exist. ||

With respect to the existence of central moments, the following theorem is useful:

THEOREM 3.14 *If $E(Y-\mu)^r$ exists for an $r>0$, then $E(Y-\mu)^s$ exists $\forall s \in [0,r]$.*

PROOF: This follows from Theorem 3.13 upon defining $X = Y - \mu$. $\square$

In addition to the moments of a random variable, there exist further useful measures of pdf characteristics, including the median and quantiles.

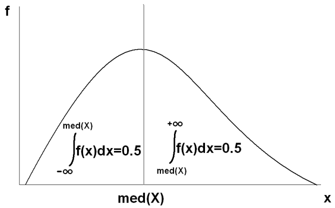
DEFINITION (MEDIAN): Any number, b, satisfying

$$P(x \leq b) \geq 1/2 \qquad \text{and} \qquad P(x \geq b) \geq 1/2$$

is called a **median** of X and is denoted by $\text{med}(X)$.

The median is a measure for the center of the distribution (see Fig.14).

Fig. 14.



The median is a special quantile of a distribution.

DEFINITION (QUANTILE): Any number, b_p, satisfying

$$P\left(x \leq b_p\right) \geq p \qquad \text{and} \qquad P\left(x \geq b_p\right) \geq 1 - p$$

is called a **quantile** of X of order p (or the $(100p)$th percentile of the distribution of X).

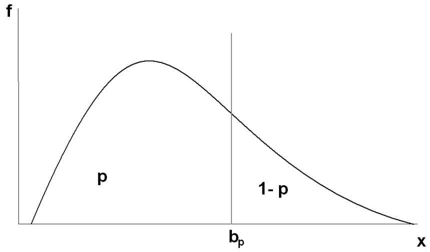


Fig. 15.

As its name suggests, the Moment-Generating Function (MGF) can be used to determine moments of a random variable. However, the main use of the MGF is not to generate moments, but to help in characterizing a distribution.

Definition (Moment-Generating Function): The MGF of a random variable X, denoted by $M_X(t)$, is

$$M_X(t) = \mathrm{E}e^{tX} = \begin{cases} \displaystyle\sum_{x \in R(X)} e^{tx} f(x) & \text{(discrete)} \\ \displaystyle\int_{-\infty}^{\infty} e^{tx} f(x)\, dx & \text{(continuous)}, \end{cases}$$

provided that the expectation exists for t in some neighborhood of 0. That is, there exists an $h > 0$ such that $\mathrm{E}e^{tX}$ exists $\forall\ t \in (-h, h)$.

Remark: The condition that $M_X(t)$ be defined $\forall\ t \in (-h, h)$ is a technical condition ensuring that $M_X(t)$ is differentiable at the point $t = 0$. This is a property which will become evident shortly. $\diamondsuit$

The following theorem indicates how the MGF generates non-central moments.

THEOREM 3.15 *Let X be a random variable for which the MGF $M_X(t)$ exists. Then*

$$\mu_r' = EX^r = \left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0}$$

PROOF: (Continuous Case) Assuming that we can differentiate under the integral sign (interchanging the order of integration and differentiation), we have[3]

$$\begin{aligned} \frac{d^r M_X(t)}{dt^r} &= \frac{d^r}{dt^r}\int_{-\infty}^{\infty} e^{tx} f(x)dx = \int_{-\infty}^{\infty} \frac{d^r}{dt^r}\left[e^{tx} f(x)\right] dx \\ &= \left.\int_{-\infty}^{\infty} x^r e^{tx} f(x)dx\right|_{t=0} = \mathrm{E}X^r. \end{aligned}$$

(Discrete case analogous). $\square$

[3] If $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$ exists for $t \in (-h, h)$, then $d^r M_X(t)/dt^r$ exists $\forall\ t \in (-h, h)$ and for all positive integers r, and the derivative of $M_X(t)$ can be found by differentiating under the integral sign (see Mittelhammer 1996, Lemma 3.3, p.142.). More details can be found in Casella and Berger (2002) in Chap. 2.4.

EXAMPLE: Consider the pdf

$$f(x) = e^{-x}\mathbb{I}_{(0,\infty)}(x) \qquad \text{(pdf of an exponential distribution).}$$

The MGF is given by

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx}e^{-x}\mathbb{I}_{(0,\infty)}(x)dx = \int_0^{\infty} e^{x(t-1)}dx \\ &= \left[\frac{e^{x(t-1)}}{t-1}\right]_{x=0}^{x=\infty}\Bigg|_{t<1} = 0 - \frac{1}{t-1} = \frac{1}{1-t}. \end{aligned}$$

The mean and the 2nd non-central moment are given by

$$\mu = \frac{dM_X(t)}{dt}\bigg|_{t=0} = \frac{1}{(1-t)^2}\bigg|_{t=0} = 1, \qquad \mu_2' = \frac{d^2M_X(t)}{dt^2}\bigg|_{t=0} = \frac{2}{(1-t)^3}\bigg|_{t=0} = 2. \quad ||$$

REMARK: The MGF $M_X(t) = \mathrm{E}e^{tX}$ can be written as a series expansion in terms of the moments of the pdf of X.

In particular, a Taylor-series expansion of $g(t) = e^{tX}$ around $t = 0$ yields

$$\begin{aligned} M_X(t) = \mathrm{E}e^{tX} &= \mathrm{E}\left(e^{0X} + \frac{1}{1!}[Xe^{0x}]t + \frac{1}{2!}[X^2e^{0x}]t^2 + \frac{1}{3!}[X^3e^{0x}]t^3 + \cdots\right) \\ &= 1 + \mu_1' t + \frac{1}{2!}\mu_2' t^2 + \frac{1}{3!}\mu_3' t^3 + \cdots \\ &= \sum_{i=0}^{\infty} \frac{1}{i!} t^i \mu_i'. \end{aligned}$$

This representation indicates that if the MGF exists, it characterizes an infinite set of moments. But note that the existence of all moments is not equivalent to the existence of the MGF[4]. $\diamondsuit$

[4] An example for a distribution that does not have an existing MGF but non-central moments that all exist and are finite can be found in Casella and Berger (2002), Example 2.3.10.

The following list summarizes useful elementary results for MGFs:

Let $X_1, ..., X_n$ be independent random variables having respective MGFs $M_{X_i}(t)$, $i = 1, \ldots, n$. Then we get

- for $Y = aX_i + b$ the MGF
$$M_Y(t) = \mathrm{E}e^{(aX_i+b)t} = e^{bt} M_{X_i}(at);$$
- for $Y = \sum_{i=1}^n X_i$ the MGF
$$M_Y(t) = \mathrm{E}e^{(\sum_{i=1}^n X_i)t} = \underbrace{\mathrm{E}\left(\prod_{i=1}^n e^{X_i t}\right) = \prod_{i=1}^n \mathrm{E}e^{X_i t}}_{\text{by independence}} = \prod_{i=1}^n M_{X_i}(t);$$
- for $Y = \sum_{i=1}^n a_i X_i + b$ the MGF
$$M_Y(t) = e^{bt} \prod_{i=1}^n M_{X_i}(a_i t).$$

The following theorem indicates that the MGF can be useful for identifying the pdf of a given random variable

Theorem 3.16 (MGF Uniqueness Theorem) *If an MGF exists for a random variable X having pdf $f(x)$, then*

- *the MGF is unique;*
- *and, conversely, the MGF determines the pdf of X uniquely, at least up to a set of points having probability 0.*

For the proof see, e.g., Widder (1961,p. 41), *Advanced Calculus*, Englewood Cliffs, Prentice-Hall.

Remark: The theorem says that there is essentially a one-to-one correspondence between pdfs and MGFs:

> A pdf has one and only one MGF associated with it, if an MGF exists at all.
>
> Furthermore, there is typically only one pdf associated with a given MGF. (If there is more than one pdf, then they differ only at a set of points having probability 0).

(continues)

REMARK (CONTINUED): This correspondence between pdfs and MGFs implies the following: If the MGF of a given random variable X is known, and if one knows a pdf that produces exactly this MGF, then the pdf can treated as the pdf of the random variable X. $\diamondsuit$

EXAMPLE: Suppose Z has an MGF defined by $M_Z(t) = \frac{1}{1-t}$ for $|t| < 1$.

Now, consider the pdf

$$f(x) = e^{-x}\mathbb{I}_{(0,\infty)}(x), \quad \text{which has an MGF} \quad M_X(t) = \frac{1}{1-t}$$

(see the previous Example). Then, by the uniqueness theorem, the pdf of Z can be specified as

$$Z \sim f(z) = e^{-z}\mathbb{I}_{(0,\infty)}(z). \quad ||$$

Multivariate Extensions

The MGF can be extended to the case of a multivariate random variable.

Definition (Moment-Generating Function; multivariate): The MGF of a multivariate random variable $X = (X_1, \ldots, X_N)'$ is

$$M_X(t) = \mathrm{E}e^{t'X} = \mathrm{E}e^{\sum_{i=1}^{n} t_i X_i}, \qquad \text{where} \qquad t = (t_1, ..., t_n)',$$

provided that the expectation exists for all t_i in some neighborhood of 0, $i = 1, ..., n$. That is, there exists an $h > 0$ such that $\mathrm{E}e^{t'X}$ exists $\forall\ t_i \in (-h, h)$, $i = 1, ..., n$.

Remark: The rth order non-central moment of X_i obtains from the rth order partial derivative w.r.t. t_i

$$\mu_r'(X_i) = \mathrm{E}X_i^r = \left.\frac{\partial^r M_X(t)}{\partial t_i^r}\right|_{t=0}.$$

(continues)

REMARK (CONTINUED): If we take cross partial derivatives of the multivariate MGF, we obtain *joint non-central moments* (which will be discussed in the next section)

$$\mathsf{E}X_i^r X_j^s = \left.\frac{\partial^{r+s} M_X(t)}{\partial t_i^r \partial t_j^s}\right|_{t=0}. \quad \diamondsuit$$

REMARK: An analog of the MGF uniqueness theorem establishing a correspondence between joint pdfs and multivariate MGFs applies to the multivariate MGF. $\diamondsuit$

In the case of multivariate random variables, *joint moments* characterize the relationship between the individual variables.

DEFINITION (JOINT NON-CENTRAL MOMENT): Let X and Y be two random variables with joint pdf $f(x, y)$. Then the joint non-central moment of (X, Y) of order (r, s) is defined as

$$\mu'_{r,s} = \mathrm{E}\left(X^r Y^s\right) = \begin{cases} \displaystyle\sum_{x \in R(X)} \sum_{y \in R(Y)} x^r y^s f(x, y) & \text{(discrete)} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r y^s f(x, y)\, dxdy & \text{(continuous).} \end{cases}$$

DEFINITION (JOINT CENTRAL MOMENT): Let X and Y be two random variables with joint pdf $f(x,y)$. Then the joint central moment of (X,Y) of order (r,s) is defined as

$$\mu_{r,s} = \begin{cases} \displaystyle\sum_{x \in R(X)} \sum_{y \in R(Y)} (x - \mathsf{E}X)^r (y - \mathsf{E}Y)^s f(x,y) & \text{(discrete)} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mathsf{E}X)^r (y - \mathsf{E}Y)^s f(x,y)\, dx dy & \text{(continuous).} \end{cases}$$

The joint moment of order $(1,1)$, namely $\mu_{1,1}$, is commonly known as the covariance, which measures the 'linear association' between X and Y.

DEFINITION (COVARIANCE): The **covariance between the random variables X and Y** is the joint central moment of the order $(1,1)$,

$$\operatorname{cov}(X,Y) = \mathsf{E}\,(X - \mathsf{E}X)\,(Y - \mathsf{E}Y)\,,$$

and will be denoted by the symbol σ_{XY}.

REMARK: The covariance can be represented in terms of non-central moments, namely

$$\begin{aligned}\sigma_{XY} &= \mathrm{E}\,(X-\mathrm{E}X)\,(Y-\mathrm{E}Y) = \mathrm{E}[XY-(\mathrm{E}X)Y-(\mathrm{E}Y)X+(\mathrm{E}X)(\mathrm{E}Y)] \\ &= \mathrm{E}XY-(\mathrm{E}X)(\mathrm{E}Y).\end{aligned}$$

From this relationship we obtain the result that

$$\mathrm{E}XY = (\mathrm{E}X)(\mathrm{E}Y) \quad \text{iff} \quad \sigma_{XY} = 0. \quad \diamondsuit$$

The covariance has an upper bound in absolute values, which depends on the variances of the corresponding random variables. This upper bound follows from the **Cauchy-Schwarz Inequality**.

Theorem 3.17 (Cauchy-Schwarz Inequality) $(EWZ)^2 \leq EW^2 EZ^2$.

Proof: Consider the random variable $(\lambda_1 W + \lambda_2 Z)^2$ which is non-negative $\forall\ (\lambda_1, \lambda_2)$;

$$\begin{aligned}
\Rightarrow \quad & \mathrm{E}(\lambda_1 W + \lambda_2 Z)^2 \geq 0, \qquad \forall\ (\lambda_1, \lambda_2) \\
\Rightarrow \quad & \lambda_1^2 \mathrm{E}W^2 + \lambda_2^2 \mathrm{E}Z^2 + 2\lambda_1\lambda_2 \mathrm{E}WZ \geq 0 \qquad \forall\ (\lambda_1, \lambda_2) \\
\Rightarrow \quad & [\lambda_1, \lambda_2] \cdot \underbrace{\begin{bmatrix} \mathrm{E}W^2 & \mathrm{E}WZ \\ \mathrm{E}WZ & \mathrm{E}Z^2 \end{bmatrix}}_{=A} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \geq 0 \qquad \forall\ (\lambda_1, \lambda_2).
\end{aligned}$$

Note that the last inequality is the defining property of positive semidefiniteness for the (2×2) matrix A. The positive semidefiniteness of A requires that

$$\mathrm{E}W^2 \geq 0, \qquad \mathrm{E}Z^2 \geq 0, \qquad |A| = \mathrm{E}W^2 \mathrm{E}Z^2 - (\mathrm{E}WZ)^2 \geq 0.$$

The last inequality implies that $(\mathrm{E}WZ)^2 \leq \mathrm{E}W^2 \mathrm{E}Z^2$. $\square$

The Cauchy-Schwarz Inequality allows us to establish an upper bound for the covariance indicated in the following theorem.

Theorem 3.18 (Covariance bound) $|\sigma_{XY}| \leq \sigma_X \sigma_Y$.

Proof: Let $W = (X - \mathrm{E}X)$ and $Z = (Y - \mathrm{E}Y)$ in the Cauchy-Schwarz Inequality. Then

$$[\mathrm{E}\,(X - \mathrm{E}X)\,(Y - \mathrm{E}Y)]^2 \leq \mathrm{E}\,(X - \mathrm{E}X)^2\,\mathrm{E}\,(Y - \mathrm{E}Y)^2\,,$$

such that $\sigma_{XY}^2 \leq \sigma_X^2 \sigma_Y^2$ or equivalently $|\sigma_{XY}| \leq \sigma_X \sigma_Y$. $\square$

Using this upper bound, we can define a normalized version of the covariance, the so-called **correlation**.

Definition (Correlation): The correlation between the random variables X and Y is defined by

$$\mathrm{corr}(X, Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

From the upper bound of the covariance, we obtain directly an upper bound for the correlation as indicated in the following theorem.

Theorem 3.19 (Correlation bound) $-1 \leq \rho_{XY} \leq 1$.

Proof: This follows directly from the upper bound for the covariance $|\sigma_{XY}| \leq \sigma_X \sigma_Y$.
$\square$

A fundamental relationship between the covariance and the stochastic (in)dependence is indicated in the next theorem.

Theorem 3.20 *If X and Y are independent, then $\sigma_{XY} = 0$ and $\rho_{XY} = 0$.*

Proof: (Discrete case) If X and Y are independent, then $f(x, y) = f_X(x) \cdot f_Y(y)$. It follows that

$$\begin{aligned}
\sigma_{XY} &= \sum_{x \in R(X)} \sum_{y \in R(Y)} (x - \mathrm{E}X)(y - \mathrm{E}Y) f_X(x) f_Y(y) && \text{(by def. of } \sigma_{XY}\text{)} \\
&= \sum_{x \in R(X)} (x - \mathrm{E}X) f_X(x) \sum_{y \in R(Y)} (y - \mathrm{E}Y) f_Y(y) \\
&= (\mathrm{E}X - \mathrm{E}X) \qquad (\mathrm{E}Y - \mathrm{E}Y) = 0.
\end{aligned}$$

(Continuous case analogous). $\square$

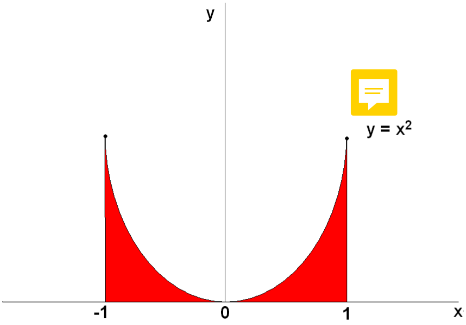
REMARK: The converse of Theorem 3.20 is not true: The fact that $\sigma_{XY} = 0$ *does not necessarily imply* that X and Y are independent. This is illustrated in the following example. $\diamondsuit$

EXAMPLE: Let X and Y have the joint pdf $f(x, y) = 1.5\mathbb{I}_{[-1,1]}(x)\mathbb{I}_{[0,x^2]}(y)$.

→ Note that this is a uniform pdf with support given by the points (x, y) on and below the parabola $y = x^2$ (see Fig.16).

→ The range of y depends on x, so that the support of $f(y|x)$ depends on x, and thus X and Y must be dependent!

Fig. 16.



(CONTINUES)

$\rightarrow$ Nonetheless, $\sigma_{XY} = 0$. To see this, note that

$$\begin{aligned}
\mathrm{E}XY &= 1.5 \int_{-1}^{1} \int_{0}^{x^2} xy dy dx = 1.5 \int_{-1}^{1} \left[x \frac{1}{2} y^2 \right]_{y=0}^{y=x^2} dx = 1.5 \int_{-1}^{1} \frac{1}{2} x^5 dx = 0, \\
\mathrm{E}X &= \int_{-1}^{1} x \underbrace{\left[\int_{0}^{x^2} 1.5 dy \right]}_{= f_X(x)} dx = 0, \qquad \mathrm{E}Y = 0.3.
\end{aligned}$$

Therefore, $\sigma_{XY} = \mathrm{E}XY - \mathrm{E}X\mathrm{E}Y = 0$. $\quad ||$

The next theorem indicates that if the correlation takes its maximum absolute value, that means $\rho_{XY} = 1$ or $\rho_{XY} = -1$, then there is a perfect linear relationship between X and Y.

Theorem 3.21 *If $\rho_{XY} = 1$ or -1, then $P(y = a + bx) = 1$, where $b \neq 1$.*

Proof: Define $Z = \lambda_1(X - \mathrm{E}X) + \lambda_2(Y - \mathrm{E}Y)$, where $\mathrm{E}Z = 0$ and

$$\begin{aligned} \mathrm{var}(Z) &= \mathrm{E}Z^2 = \lambda_1^2\mathrm{E}(X - \mathrm{E}X)^2 + \lambda_2^2\mathrm{E}(Y - \mathrm{E}Y)^2 + 2\lambda_1\lambda_2\mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y) \\ &= [\lambda_1, \lambda_2] \cdot \underbrace{\begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}}_{=A} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \geq 0 \qquad \forall\, (\lambda_1, \lambda_2). \end{aligned}$$

Now, if $\rho_{XY} = 1$ or -1, then $\sigma_{XY}^2 = \sigma_X^2\sigma_Y^2$, which implies that $|A| = 0$ such that A is singular. In this case, the columns of A are linearly dependent, so that there exist $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$ such that

$$\begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

A solution for λ_1 and λ_2 is given by $\lambda_1 = \sigma_{XY}/\sigma_X^2$ and $\lambda_2 = -1$.
Since $\mathrm{var}(Z) = 0$ at those values for λ_1, λ_2 , we have

$$P(z = \mathrm{E}Z) = P(z = 0) = 1.$$

Inserting the definition of Z together with $\lambda_1 = \sigma_{XY}/\sigma_X^2$ and $\lambda_2 = -1$ yields

$$P(y = [\mathrm{E}Y - (\sigma_{XY}/\sigma_X^2)\mathrm{E}Y] + [\sigma_{XY}/\sigma_X^2]\, x) = P(y = a + b\, x) = 1. \quad \square$$

REMARK: If $\rho_{XY} = 1$ or -1 such that $P(y = a + bx) = 1$, then the joint pdf $f(x, y)$ is **degenerate**. All the probability mass of $f(x, y)$ is concentrated above the line $y = a + bx$.

This generates a perfect linear relationship between X and Y (see Fig.17). $\diamondsuit$

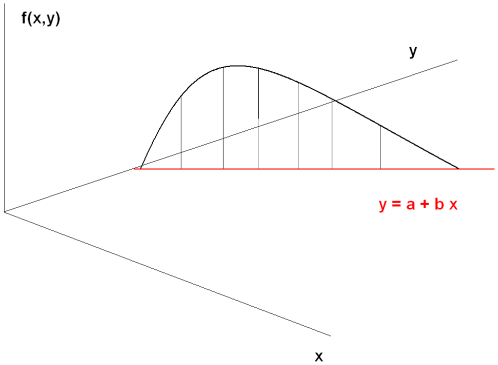


Fig. 17.

Our earlier results for the mean and variance of random variables can be extended to obtain the mean and variance of linear combinations of random variables. The first result concerns the mean.

THEOREM 3.22 *Let $Y = \sum_{i=1}^{n} a_i X_i$, where the a_is are real constants. Then $EY = \sum_{i=1}^{n} a_i EX_i$.*

PROOF: This follows directly from Theorem 3.8 indicating that the expectation for a sum of random variables is equal to the sum of their expectations. $\square$

REMARK: The matrix representation of this result obtains as follows. Let

$$\boldsymbol{a} = (a_1, \ldots, a_n)' \quad \text{and} \quad \boldsymbol{X} = (X_1, \ldots, X_n)'.$$

Then $\qquad Y = \boldsymbol{a}'\boldsymbol{X} \qquad$ such that $\qquad \mathrm{E}Y = \boldsymbol{a}'\mathrm{E}\boldsymbol{X}. \qquad \diamondsuit$

For the variance of a linear combination, we have the following result.

THEOREM 3.23 *Let $Y = \sum_{i=1}^{n} a_i X_i$, where the a_is are real constants. Then*

$$\sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_{X_i}^2 + 2 \sum_{i<j} a_i a_j \sigma_{X_i X_j}.$$

PROOF: We have

$$\begin{aligned}
\sigma_Y^2 &= \mathrm{E}(Y - \mathrm{E}Y)^2 = \mathrm{E}\left[\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} a_i \mathrm{E}X_i\right]^2 = \mathrm{E}\left[\sum_{i=1}^{n} a_i (X_i - \mathrm{E}X_i)\right]^2 \\
&= \mathrm{E}\left[\sum_{i=1}^{n} a_i^2 (X_i - \mathrm{E}X_i)^2 + 2 \sum_{i<j} a_i a_j (X_i - \mathrm{E}X_i)(X_j - \mathrm{E}X_j)\right] \\
&= \sum_{i=1}^{n} a_i^2 \sigma_{X_i}^2 + 2 \sum_{i<j} a_i a_j \sigma_{X_i X_j}. \quad \square
\end{aligned}$$

In order to rewrite this result in matrix notation we shall define the **covariance matrix** of a multivariate random variable.

Definition (Covariance matrix): The covariance matrix of the n-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is the $n \times n$ symmetric matrix

$$\mathbf{Cov}(\boldsymbol{X}) = \mathrm{E}(\boldsymbol{X} - \mathrm{E}\boldsymbol{X})(\boldsymbol{X} - \mathrm{E}\boldsymbol{X})' = \begin{pmatrix} \sigma^2_{X_1} & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_n} \\ \sigma_{X_2 X_1} & \sigma^2_{X_2} & \cdots & \sigma_{X_2 X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_n X_1} & \sigma_{X_n X_2} & \cdots & \sigma^2_{X_n} \end{pmatrix}.$$

Remark:

- The variance of the ith variable in $\boldsymbol{X}$ is given by the (i, i)th diagonal entry in the covariance matrix.
- The covariance of the ith and jth variable is displayed in the (i, j)th as well as in the (j, i)th off-diagonal entry in the covariance matrix.

  Note that this implies that a covariance matrix is symmetric, that is $\mathbf{Cov}(\boldsymbol{X}) = \mathbf{Cov}(\boldsymbol{X})'$. $\diamondsuit$

REMARK: Let $\boldsymbol{a} = (a_1, \ldots, a_n)'$ and $\boldsymbol{X} = (X_1, \ldots, X_n)'$. Then the variance of $Y = \boldsymbol{a}'\boldsymbol{X}$ given in Theorem 3.23 can obviously be represented as

$$\sigma_Y^2 = \boldsymbol{a}'\mathbf{Cov}(\boldsymbol{X})\boldsymbol{a}.$$

Note that since a variance is non-negative ($\sigma_Y^2 \geq 0$) the expression $\boldsymbol{a}'\mathbf{Cov}(\boldsymbol{X})\boldsymbol{a}$ is also non-negative for any a.

This implies that a covariance matrix is necessarily positive semidefinite ! $\diamondsuit$

The preceding results can be extended to the case where $\boldsymbol{Y}$ is a vector of linear combinations of a random vector $\boldsymbol{X}$.

Theorem 3.24 *Let* $\boldsymbol{Y} = \boldsymbol{AX}$*, where* $\boldsymbol{A} = (a_{hm})$ *is a* $k \times n$ *matrix of real constants, and* $\boldsymbol{X} = (X_i)$ *is an* $n \times 1$ *vector of random variables. Then* $\mathrm{E}\boldsymbol{Y} = \boldsymbol{A}\mathrm{E}\boldsymbol{X}$*.*

Proof: The linear combination $\boldsymbol{AX}$ can be written as

$$\boldsymbol{Y} = \boldsymbol{AX} = \begin{bmatrix} a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n \\ \vdots \\ a_{k1}X_1 + a_{k2}X_2 + \cdots + a_{kn}X_n \end{bmatrix},$$

such that $\mathrm{E}\boldsymbol{Y} = \boldsymbol{A}\mathrm{E}\boldsymbol{X}$ follows immediately from the application of Theorem 3.22 (expectation of one linear combination) to each of the k linear combinations. □

For the covariance matrix of k linear combinations, we have the following result.

Theorem 3.25 *Let $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{A} = (a_{hm})$ is a $k \times n$ matrix of real constants, and $\boldsymbol{X} = (X_i)$ is an $n \times 1$ vector of random variables. Then $\mathbf{Cov}(\boldsymbol{Y}) = \boldsymbol{A}\mathbf{Cov}(\boldsymbol{X})\boldsymbol{A}'$.*

Proof: We have by definition of a covariance matrix

$$\mathbf{Cov}(\boldsymbol{Y}) \quad = \quad \mathrm{E}(\boldsymbol{Y} - \mathrm{E}\boldsymbol{Y})(\boldsymbol{Y} - \mathrm{E}\boldsymbol{Y})',$$

where $\qquad \boldsymbol{Y} - \mathrm{E}\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} - \boldsymbol{A}\mathrm{E}\boldsymbol{X} = \boldsymbol{A}(\boldsymbol{X} - \mathrm{E}\boldsymbol{X}).$

Thus we get

$$\begin{aligned}
\mathbf{Cov}(\boldsymbol{Y}) &= \mathrm{E}\left[\boldsymbol{A}(\boldsymbol{X} - \mathrm{E}\boldsymbol{X})(\boldsymbol{X} - \mathrm{E}\boldsymbol{X})'\boldsymbol{A}'\right] \\
&= \boldsymbol{A}\left[\mathrm{E}(\boldsymbol{X} - \mathrm{E}\boldsymbol{X})(\boldsymbol{X} - \mathrm{E}\boldsymbol{X})'\right]\boldsymbol{A}' \\
&= \boldsymbol{A}\left[\mathbf{Cov}(\boldsymbol{X})\right]\boldsymbol{A}'. \quad \square
\end{aligned}$$