STAT/MTHE 353: 4 - More on Expectations and Variances

T. Linder

Queen's University

Winter 2017

Expectations of Sums of Random Variables

Recall that if $X_1, \ldots, X_n$ are random variables with finite expectations, then

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

The X_i can be continuous or discrete or of any other type.

- The expectation on the left-hand-side is with with respect to the joint distribution of $X_1, \ldots, X_n$.
- The ith expectation on the right-hand-side is with with respect to the marginal distribution of X_i, $i = 1, \ldots, n$.

Often we can write a r.v. X as a sum of simpler random variables. Then $E(X)$ is the sum of the expectation of these simpler random variables.

Example: Consider $(X_1, \ldots, X_r)$ having multinomial distribution with parameters n and $(p_1, \ldots, p_r)$. Compute $E(X_i)$, $i = 1, \ldots, r$

Solution: ...

Example: Let $(X_1, \ldots, X_r)$ the multivariate hypergeometric distribution with parameters N and $n_1, \ldots, n_r$. Compute $E(X_i)$, $i = 1, \ldots, r$

Solution: ...

Example: (Matching problem) If the integers $1, 2, \ldots, n$ are randomly permuted, what is the probability that integer i is in the ith position? What is the expected number of integers in the correct position?

Solution: ...

Example: We have two urns. Initially Urn 1 contains n red balls and Urn 2 contains n blue balls. At each stage of the experiment we pick a ball from Urn 1 at random, also pick a ball from Urn 2 at random, and then swap the balls. Let $X = \#$ of red balls in Urn 1 after k stages. Compute $E(X)$ for even k.

Solution: ...

Conditional Expectation

- Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^T$ are two vector random variables defined on the same probability space.
- The distributions (joint marginals) of $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be described the pdfs $f_{\boldsymbol{X}}(\boldsymbol{x})$ and $f_{\boldsymbol{Y}}(\boldsymbol{y})$ (if both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are continuous) or by the pmfs $p_{\boldsymbol{X}}(\boldsymbol{x})$ and $p_{\boldsymbol{Y}}(\boldsymbol{y})$ (if both are discrete).
- The joint distribution of the pair $(\boldsymbol{X}, \boldsymbol{Y})$ can be described by their joint pdf $f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})$ or joint pmf $p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})$.
- The *conditional distribution* of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is described by either the conditional pdf

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

or the conditional pmf

$$p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})}{p_{\boldsymbol{Y}}(\boldsymbol{y})}$$

Remarks:

(1) In general, $\boldsymbol{X}$ and $\boldsymbol{Y}$ can have different types of distribution (e.g., one is discrete, the other is continuous).
Example: Let $n = m = 1$ and $X = Y + Z$, where Y is a Bernoulli(p) r.v. and $Z \sim N(0, \sigma^2)$, and Y and Z are independent. Determine the conditional pdf of X given $Y = 0$ and $Y = 1$. Also, determine the pdf of X.
Solution: ...

(2) Not all random variables are either discrete or continuous. Mixed discrete-continuous and even more general distributions are possible, but they are mostly out of the scope of this course.

Definitions

(1) The *conditional expectation* of X given $Y = y$ is the mean (expectation) of the distribution of X given $Y = y$ and is denoted by $E(X|Y = y)$.

(2) The *conditional variance* of X given $Y = y$ is the the variance of the distribution of X given $Y = y$ and is denoted by $\mathrm{Var}(X|Y = y)$.

- If both X and Y are *discrete*,

$$E(X|Y = y) = \sum_x x p_{X|Y}(x|y)$$

and $$\mathrm{Var}(X|Y = y) = \sum_x \big(x - E(X|Y = y)\big)^2 p_{X|Y}(x|y)$$

- In case both X and Y are *continuous*, we have

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx$$

and

$$\mathrm{Var}(X|Y = y) = \int_{-\infty}^{\infty} \big(x - E(X|Y = y)\big)^2 f_{X|Y}(x|y)\, dx$$

Special case: Assume X and Y are *independent*. Then (considering the discrete case)

$$p_{X|Y}(x|y) = p_X(x)$$

so that for all y,

$$E(X|Y = y) = \sum_x x p_{X|Y}(x|y) = \sum_x x p_X(x) = E(X)$$

A similar argument shows $E(X|Y = y) = E(X)$ if X and Y are independent continuous random variables.

Notation: Let $g(y) = E(X|Y = y)$. We define the *random variable* $E(X|Y)$ by setting

$$E(X|Y) = g(Y)$$

Similarly, letting $h(y) = \mathrm{Var}(X|Y = y)$, the random variable $\mathrm{Var}(X|Y)$ is defined by

$$\mathrm{Var}(X|Y) = h(Y)$$

For example, if X and Y are independent, then $E(X|Y = y) = E(X)$ (constant function), so

$$E(X|Y) = E(X)$$

The following are important properties of conditional expectation. We don't prove them formally, but they should be intuitively clear.

Properties

(i) (*Linearity of conditional expectation*) If X_1 and X_2 are random variables with finite expectations, then for all $a, b \in \mathbb{R}$,

$$\boxed{E(aX_1 + bX_2|Y) = aE(X_1|Y) + bE(X_2|Y)}$$

(ii) If $g : \mathbb{R} \to \mathbb{R}$ is a function such that $E[g(Y)]$ is finite, then

$$\boxed{E\big[g(Y)|Y\big] = g(Y)}$$

and if $E\big[g(Y)X\big]$ is finite, then

$$\boxed{E\big[g(Y)X|Y\big] = g(Y)E(X|Y)}$$

Theorem 1 (Law of total expectation)

$$E(X) = E\big[E(X|Y)\big]$$

Proof: Assume both X and Y are discrete. Then

$$\begin{aligned} E\big[E(X|Y)\big] &= \sum_y E(X|Y=y)p_Y(y) = \sum_y \Big(\sum_x x p_{X|Y}(x|y)\Big)p_Y(y) \\ &= \sum_y \Big(\sum_x x \frac{p_{X,Y}(x,y)}{p_Y(y)}\Big)p_Y(y) = \sum_y \sum_x x p_{X,Y}(x,y) \\ &= \sum_x x p_X(x) = E(X) \qquad \square \end{aligned}$$

Example: Expected value of geometric distribution...

Lemma 2 (Variance formula)

$$\mathrm{Var}(X) = E\big[\mathrm{Var}(X|Y)\big] + \mathrm{Var}\big[E(X|Y)\big]$$

Proof: Since $\mathrm{Var}(X|Y = y)$ is the variance of the conditional distribution of X given $Y = y$,

$$\mathrm{Var}(X|Y) = E[X^2|Y] - \big(E[X|Y]\big)^2$$

Taking expectation (with respect to Y),

$$E\big[\mathrm{Var}(X|Y)\big] = E(E[X^2|Y]) - E\big[\big(E[X|Y]\big)^2\big] = E(X^2) - E\big[\big(E[X|Y]\big)^2\big]$$

On the other hand,

$$\mathrm{Var}\big(E[X|Y]\big) = E\big[\big(E[X|Y]\big)^2\big] - \big(E\big[E(X|Y)\big]\big)^2 = E\big[\big(E[X|Y]\big)^2\big] - \big(E(X)\big)^2$$

so

$$\mathrm{Var}(X) = E(X^2) - \big(E(X)\big)^2 = E\big[\mathrm{Var}(X|Y)\big] + \mathrm{Var}\big[E(X|Y)\big] \qquad \square$$

Remarks:

(1) Let A be an event and X the indicator of A:

$$X = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A^c \text{ occurs} \end{cases}$$

Then $E(X) = P(A)$. Assuming Y is a discrete r.v., we have $E(X|Y = y) = P(A|Y = y)$ and the law of total expectation states

$$P(A) = E(X) = \sum_y E(X|Y = y)p_Y(y) = \sum_y P(A|Y = y)p_Y(y)$$

which is the *law of total probability*.
For continuous Y we have

$$P(A) = \int_{-\infty}^{\infty} E(X|Y = y)f_Y(y)\,dy = \int_{-\infty}^{\infty} P(A|Y = y)f_Y(y)\,dy$$

(2) The law of total expectation says that we can compute the mean of a distribution by conditioning on another random variable. This distribution can be a conditional distribution. For example, for r.v.'s X, Y, and Z,

$$E(X|Y = y) = E\big[E(X|Y = y, Z)|Y = y\big]$$

so that

$$\boxed{E(X|Y) = E\big[E(X|Y, Z)|Y\big]}$$

For example, if Z is discrete,

$$\begin{aligned} E(X|Y = y) &= \sum_z E(X|Y = y, Z = z)p_{Z|Y}(z|y) \\ &= \sum_z E(X|Y = y, Z = z)P(Z = z|Y = y) \end{aligned}$$

Exercise: Prove the above statement if X, Y, and Z are discrete.

Example: Repeatedly flip a biased coin which comes up heads with probability p. Let X denote the number of flips until 2 consecutive heads occur. Find $E(X)$.

Solution:

Example: (Simplex algorithm) There are n vertices (points) that are ranked from best to worst. Start from point j and at each step, jump to one of the better points at random (with equal probability). What is the expected number of steps to reach the best point?

Solution:

Minimum mean square error (MMSE) estimation

Suppose a r.v. Y is observed and based on its value we want to "guess" the value of another r.v. X. Formally, we want to use a function $g(Y)$ of Y to estimate the unobserved X in the sense of minimizing the *mean square error*

$$E\big[(X - g(Y))^2\big]$$

It turns out that $g^*(Y) = E(X|Y)$ is the optimal choice.

Theorem 3

Suppose X has finite variance. Then for $g^*(Y) = E(X|Y)$ and any function g

$$E\big[(X - g(Y))^2\big] \geq E\big[(X - g^*(Y))^2\big]$$

Proof: Use the properties of conditional expectation:

$$\begin{aligned}
&E\big[(X-g(Y))^2|Y\big] \\
&= E\big[(X-g^*(Y)+g^*(Y)-g(Y))^2|Y\big] \\
&= E\big[(X-g^*(Y))^2+(g^*(Y)-g(Y))^2+2(X-g^*(Y))(g^*(Y)-g(Y))|Y\big] \\
&= E\big[(X-g^*(Y))^2|Y\big]+E\big[(g^*(Y)-g(Y))^2|Y\big] \\
&\quad +2E\big[(X-g^*(Y))(g^*(Y)-g(Y))|Y\big] \\
&= E\big[(X-g^*(Y))^2|Y\big]+(g^*(Y)-g(Y))^2 \\
&\quad +2(g^*(Y)-g(Y))E\big[X-g^*(Y)|Y\big] \\
&= E\big[(X-g^*(Y))^2|Y\big]+(g^*(Y)-g(Y))^2 \\
&\quad +2(g^*(Y)-g(Y))\underbrace{\big[E(X|Y)-g^*(Y)\big]}_{=0} \\
&= E\big[(X-g^*(Y))^2|Y\big]+(g^*(Y)-g(Y))^2
\end{aligned}$$

Proof cont'd

Thus

$$E\big[(X-g(Y))^2|Y\big]=E\big[(X-g^*(Y))^2|Y\big]+(g^*(Y)-g(Y))^2$$

Take expectations on both sides and use the law of total expectation to obtain

$$E\big[(X-g(Y))^2\big]=E\big[(X-g^*(Y))^2\big]+E\big[g^*(Y)-g(Y))^2\big]$$

Since $\big(g^*(Y)-g(Y)\big)^2\ge 0$, this implies

$$E\big[(X-g(Y))^2\big]\ge E\big[(X-g^*(Y))^2\big] \qquad \square$$

Remark: Note that since $g^*(y)=E(X|Y=y)$, we have

$$E\big[\mathrm{Var}(X|Y)\big]=E\big[(X-g^*(Y))^2\big]$$

i.e., $E\big[\mathrm{Var}(X|Y)\big]$ is the mean square error of the MMSE estimate of X given Y.

Example: Suppose $X\sim N(0,\sigma_X^2)$ and $Z\sim N(0,\sigma_Z^2)$, where X and Z are independent. Here X represents a signal sent from a remote location which is corrupted by noise Z so that the received signal is $Y=X+Z$. What is the MMSE estimate of X given $Y=y$?

Random Sums

Theorem 4 (Wald's equation)

Let $X_1,X_2\ldots$ be i.i.d. random variables with mean μ. Let N be r.v. with values in $\{1,2,\ldots\}$ that is independent of the X_i's and has finite mean $E(N)$. Define $X=\sum_{i=1}^N X_i$. Then

$$E(X)=E(N)\mu$$

Proof:

$$\begin{aligned}
E(X|N=n) &= E(X_1+\cdots+X_N|N=n) \\
&= E(X_1+\cdots+X_n|N=n) \\
&= E(X_1|N=n)+\cdots+E(X_n|N=n) \\
&\qquad\qquad \text{(linearity of expectation)} \\
&= E(X_1)+\cdots+E(X_n) \qquad (N \text{ and } X_i \text{ are independent}) \\
&= n\mu
\end{aligned}$$

Proof cont'd: We obtained $E(X|N=n) = n\mu$ for all $n = 1, 2, \ldots$, i.e, $E(X|N) = N\mu$. By the law of total expectation

$$E(X) = E\big[E(X|N)\big] = E(N\mu) = E(N)\mu \qquad \square$$

Example: (Branching Process) Suppose a population evolves in generations starting from a single individual (generation 0). Each individual of the ith generation produces a random number of offsprings; the collection of all offsprings by generation i individuals forms generation $i+1$. The number of offsprings born to distinct individuals are independent random variables with mean μ. Let X_n be the number of individuals in the nth generation. Find $E(X_n)$.

Covariance and Correlation

Covariance

Definition Let X and Y be two random variables with finite variance. Their *covariance* is defined by

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

Properties:

(1)

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E(XY) - E\big[E(X)Y\big] - E\big[XE(Y)\big] + E\big[E(X)E(Y)\big] \\ &= E(XY) - 2E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

The formula $\boxed{\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)}$ is often useful in computations.

(2) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

(3) If $X = Y$ we obtain

$$\mathrm{Cov}(X,Y) = E\big[(X - E(X))^2\big] = \mathrm{Var}(X)$$

(4) For any constants a, b, c and d,

$$\begin{aligned}&\mathrm{Cov}(aX+b, cY+d) \\ &\quad= E\big[(aX+b-E(aX+b))(cY+d-E(cY+d))\big] \\ &\quad= E\big[a(X-E(X))c(Y-E(Y))\big] \\ &\quad= acE\big[(X-E(X))(Y-E(Y))\big] \\ &\quad= ac\,\mathrm{Cov}(X,Y)\end{aligned}$$

(5) If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$.

Proof: By independence, $E(XY) = E(X)E(Y)$, so

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = 0$$

Definition Let $X_1, \ldots, X_n$ be random variables with finite variances. The *covariance matrix* of the vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ is the $n \times n$ matrix $\mathrm{Cov}(\boldsymbol{X})$ whose (i,j)th entry is $\mathrm{Cov}(X_i, X_j)$.

Remarks:

- The ith diagonal entry of $\mathrm{Cov}(\boldsymbol{X})$ is $\mathrm{Var}(X_i)$, $i = 1, \ldots, n$
- $\mathrm{Cov}(\boldsymbol{X})$ is a symmetric matrix since $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$ for all i and j.

Some properties of covariance are easier to derive using a matrix formalism.

- Let $\boldsymbol{V} = \{Y_{ij}; i = 1, \dots, m,\ j = 1, \dots, n\}$ be an $m \times n$ matrix of random variables having finite expectations. We define $E(\boldsymbol{V})$ by taking expectations componentwise:

$$E(\boldsymbol{V}) = E\begin{bmatrix} Y_{11} & \dots & Y_{1n} \\ Y_{21} & \dots & Y_{2n} \\ \vdots & \ddots & \vdots \\ Y_{m1} & \dots & Y_{mn} \end{bmatrix} = \begin{bmatrix} E(Y_{11}) & \dots & E(Y_{1n}) \\ E(Y_{21}) & \dots & E(Y_{2n}) \\ \vdots & \ddots & \vdots \\ E(Y_{m1}) & \dots & E(Y_{mn}) \end{bmatrix}$$

- Now notice that the $n \times n$ matrix $(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T$ has $(X_i - E(X_i))(X_j - E(X_j))$ in its (i,j)th entry. Thus

$$\mathrm{Cov}(\boldsymbol{X}) = E\big[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T\big]$$

Lemma 5

Let $\boldsymbol{A}$ be an $m \times n$ real matrix and define $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$ (an m-dimensional random vector). Then

$$\boxed{\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}^T}$$

Proof: First note that by the linearity of expectation,

$$E(\boldsymbol{Y}) = E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}).$$

Thus

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{Y}) &= E\big[(\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T\big] \\ &= E\big[(\boldsymbol{A}\boldsymbol{X} - \boldsymbol{A}E(\boldsymbol{X}))(\boldsymbol{A}\boldsymbol{X} - \boldsymbol{A}E(\boldsymbol{X}))^T\big] \\ &= E\big[(\boldsymbol{A}(\boldsymbol{X} - E(\boldsymbol{X})))(\boldsymbol{A}(\boldsymbol{X} - E(\boldsymbol{X})))^T\big] \\ &= E\big[\boldsymbol{A}(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X})))^T\boldsymbol{A}^T\big] \\ &= \boldsymbol{A}E\big[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X})))^T\big]\boldsymbol{A}^T \\ &= \boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}^T \qquad \square \end{aligned}$$

For any m-vector $\boldsymbol{c} = (c_1, \dots, c_m)^T$ we also have

$$\mathrm{Cov}(\boldsymbol{Y} + \boldsymbol{c}) = \mathrm{Cov}(\boldsymbol{Y})$$

since $\mathrm{Cov}(Y_i + c_i, Y_j + c_j) = \mathrm{Cov}(Y_i, Y_j)$.

Thus

$$\boxed{\mathrm{Cov}(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{c}) = \boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}^T}$$

Let $m = 1$ so that $\boldsymbol{c} = c$ is a scalar and $\boldsymbol{A}$ is and $1 \times n$ matrix, i.e., $\boldsymbol{A}$ is a row vector $\boldsymbol{A} = \boldsymbol{a}^T = (a_1, \dots, a_n)$. Then

$$\mathrm{Cov}(\boldsymbol{a}^T\boldsymbol{X} + c) = \mathrm{Cov}\left(\sum_{i=1}^n a_i X_i + c\right) = \mathrm{Var}\left(\sum_{i=1}^n a_i X_i + c\right)$$

On the other hand,

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{a}^T\boldsymbol{X} + c) &= \boldsymbol{a}^T\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{a} = \sum_{i=1}^n \sum_{j=1}^n a_i a_j\,\mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n a_i^2\,\mathrm{Cov}(X_i, X_i) + 2\sum_{i<j} a_i a_j\,\mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n a_i^2\,\mathrm{Var}(X_i) + 2\sum_{i<j} a_i a_j\,\mathrm{Cov}(X_i, X_j) \end{aligned}$$

Hence

$$\boxed{\mathrm{Var}\left(\sum_{i=1}^n a_i X_i + c\right) = \sum_{i=1}^n a_i^2\,\mathrm{Var}(X_i) + 2\sum_{i<j} a_i a_j\,\mathrm{Cov}(X_i, X_j)}$$

Note that if $X_1, \dots, X_n$ are *independent*, then $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$, and we obtain

$$\mathrm{Var}\left(\sum_{i=1}^n a_i X_i + c\right) = \sum_{i=1}^n a_i^2\,\mathrm{Var}(X_i)$$

More generally, let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^T$ and let $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})$ be the $n \times m$ matrix with (i,j)th entry $\mathrm{Cov}(X_i, Y_j)$. Note that

$$\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = E\big[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T\big]$$

If $\boldsymbol{A}$ is a $k \times n$ matrix, $\boldsymbol{B}$ is an $l \times m$ matrix, $\boldsymbol{c}$ is a k-vector and $\boldsymbol{d}$ is an l-vector, then

$$\begin{aligned}
&\mathrm{Cov}(\boldsymbol{AX} + \boldsymbol{c}, \boldsymbol{BY} + \boldsymbol{d}) \\
&\quad = E\big[(\boldsymbol{AX} + \boldsymbol{c} - E(\boldsymbol{AX} + \boldsymbol{c}))(\boldsymbol{BY} + \boldsymbol{d} - E(\boldsymbol{BY} + \boldsymbol{d}))^T\big] \\
&\quad = \boldsymbol{A} E\big[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T\big] \boldsymbol{B}^T
\end{aligned}$$

We obtain

$$\boxed{\mathrm{Cov}(\boldsymbol{AX} + \boldsymbol{c}, \boldsymbol{BY} + \boldsymbol{d}) = \boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}^T}$$

We can now prove the following important property of covariance:

Lemma 6

For any constants $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$,

$$\mathrm{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \,\mathrm{Cov}(X_i, Y_j)$$

i.e., $\mathrm{Cov}(X, Y)$ is bilinear.

Proof: Let $k = l = 1$ and $\boldsymbol{A} = \boldsymbol{a}^T = (a_1, \ldots, a_n)$ and $\boldsymbol{B} = \boldsymbol{b}^T = (b_1, \ldots, b_m)$. Then we have

$$\begin{aligned}
\mathrm{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) &= \mathrm{Cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{b}^T\boldsymbol{Y}) = \mathrm{Cov}(\boldsymbol{AX}, \boldsymbol{BY}) \\
&= \boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}^T = \boldsymbol{a}^T\,\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{b} \\
&= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \,\mathrm{Cov}(X_i, Y_j) \qquad \square
\end{aligned}$$

The following property of covariance is of fundamental importance:

Lemma 7

$$|\mathrm{Cov}(X, Y)| \le \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$$

Proof: First we prove the *Cauchy-Schwarz inequality* for random variables U and V with finite variances. Let $\lambda \in \mathbb{R}$, then

$$\begin{aligned}
0 &\le E\big[(U - \lambda V)^2\big] = E(U^2 - 2\lambda UV + \lambda^2 V^2) \\
&= E(U^2) - 2\lambda E(UV) + \lambda^2 E(V^2)
\end{aligned}$$

This is a quadratic polynomial in λ which cannot have two *distinct real roots.*

Proof cont'd: Thus its discriminant cannot be positive:

$$4\big[E(UV)\big]^2 - 4E(U^2)E(V^2) \le 0$$

so we obtain

$$\boxed{[E(UV)]^2 \le E(U^2)E(V^2)}$$

Use this with $U = X - E(X)$ and $V = Y - E(Y)$ to get

$$\begin{aligned}
|\mathrm{Cov}(X, Y)| &= \big|E\big[(X - E(X))(Y - E(Y))\big]\big| \\
&\le \sqrt{E\big[(X - E(X))^2\big]E\big[(Y - E(Y))^2\big]} \\
&= \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)} \qquad \square
\end{aligned}$$

Correlation

Recall that $\operatorname{Cov}(aX, bY) = ab \operatorname{Cov}(X, Y)$. This is an undesirable property if we want to use $\operatorname{Cov}(X, Y)$ as a measure of association between X and Y. A proper normalization will solve this problem:

Definition The *correlation coefficient* between X and Y having nonzero variances is defined by

$$\rho(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X) \operatorname{Var}(Y)}}$$

Remarks:

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

$$\rho(aX + b, aY + d) = \rho(X, Y)$$

- Letting $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X^2 = \operatorname{Var}(X)$, $\sigma_Y^2 = \operatorname{Var}(Y)$, we have

$$\begin{aligned} \rho(X, Y) &= \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\operatorname{Cov}(X - \mu_X, Y - \mu_Y)}{\sigma_X \sigma_Y} \\ &= \operatorname{Cov}\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y} \right) \end{aligned}$$

  Thus $\rho(X, Y)$ is the covariance between the *standardized* versions of X and Y.

- If X and Y are independent, then $\operatorname{Cov}(X, Y) = 0$, so $\rho(X, Y) = 0$. On the other hand, $\rho(X, Y) = 0$ does not imply that X and Y are independent.

  Remark: If $\rho(X, Y) = 0$ we say that X and Y are *uncorrelated*.

  Example: Find random variables X and Y that are uncorrelated but not independent.

Example: Covariance and correlation for multinomial random variables...

Theorem 8

The correlation always satisfies

$$|\rho(X, Y)| \le 1$$

Moreover, $|\rho(X, Y)| = 1$ if and only if $Y = aX + b$ for some constants a and b ($a \neq 0$), i.e., Y is an affine function of X.

Proof: We know that $|\operatorname{Cov}(X, Y)| \le \sqrt{\operatorname{Var}(X) \operatorname{Var}(Y)}$, so $|\rho(X, Y)| \le 1$ always holds.

Let's assume now that $Y = aX + b$, where $a \neq 0$. Then

$$\operatorname{Cov}(X, Y) = \operatorname{Cov}(X, aX + b) = \operatorname{Cov}(X, aX) = a \operatorname{Cov}(X, X) = a \operatorname{Var}(X)$$

so

$$\rho(X, Y) = \frac{a \operatorname{Var}(X)}{\sqrt{\operatorname{Var}(X) a^2 \operatorname{Var}(X)}} = \frac{a}{\sqrt{a^2}} = \pm 1$$

Proof cont'd:

Conversely, suppose that $\rho(X, Y) = 1$. Then

$$\begin{aligned} \operatorname{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} \right) &= \operatorname{Cov}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}, \frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} \right) \\ &= \operatorname{Var}\left(\frac{X}{\sigma_X} \right) + \operatorname{Var}\left(\frac{Y}{\sigma_Y} \right) - 2 \operatorname{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y} \right) \\ &= \frac{\operatorname{Var}(X)}{\sigma_X^2} + \frac{\operatorname{Var}(Y)}{\sigma_Y^2} - 2 \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y} \\ &= 1 + 1 - 2 = 0 \end{aligned}$$

This means that $\dfrac{X}{\sigma_X} - \dfrac{Y}{\sigma_Y} = c$ for some constant c, so

$$Y = \frac{\sigma_Y}{\sigma_X} X - \sigma_Y c$$

If $\rho(X, Y) = -1$, consider $\operatorname{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} \right)$ and use the same proof $\square$

Remark: The previous theorem implies that correlation can be thought of as a measure of *linear association* (linear dependence) between X and Y. Recall the multinomial example...

Example: (Linear MMSE estimation) Let X and Y be random variables with zero means and finite variances $\sigma_X^2 > 0$ and $\sigma_Y^2 > 0$. Suppose we want to estimate X in the MMSE sense using a *linear* function of Y; i.e., we are looking for $a \in \mathbb{R}$ minimizing

$$E\big[(X - aY)^2\big]$$

Find the minimizing a and determine the resulting minimum mean square error. Relate the results to $\rho(X, Y)$.

Solution: ...