### 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 respect to the joint distribution of $X_1, \ldots, X_n$.
- The $i$th expectation on the right-hand-side is 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 $i$th position? What is the expected number of integers in the correct position?

**Solution:** ...

**Example:** (HW problem in 2010) 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 \( \mathbf{X} = (X_1, \ldots, X_n)^T \) and \( \mathbf{Y} = (Y_1, \ldots, Y_m)^T \) are two vector random variables defined on the same probability space.
- The distributions (joint marginals) of \( \mathbf{X} \) and \( \mathbf{Y} \) can be described by their joint pdfs \( f_{X,Y}(x,y) \) (if both \( \mathbf{X} \) and \( \mathbf{Y} \) are continuous) or by the joint pmfs \( p_{X,Y}(x,y) \) (if both are discrete).
- The joint distribution of the pair \( (\mathbf{X}, \mathbf{Y}) \) can be described by their joint pdf \( f_{X,Y}(x,y) \) or joint pmf \( p_{X,Y}(x,y) \).
- The conditional distribution of \( \mathbf{X} \) given \( \mathbf{Y} = y \) is described by either the conditional pdf
  \[
  f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}
  \]
  or the conditional pmf
  \[
  p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}
  \]

Definitions

1. The conditional expectation of \( \mathbf{X} \) given \( \mathbf{Y} = y \) is the mean (expectation) of the distribution of \( \mathbf{X} \) given \( \mathbf{Y} = y \) and is denoted by \( E(\mathbf{X}|\mathbf{Y} = y) \).

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

- If both \( \mathbf{X} \) and \( \mathbf{Y} \) are discrete,
  \[
  E(\mathbf{X}|\mathbf{Y} = y) = \sum_x xp_{X|Y}(x|y)
  \]
  and
  \[
  \text{Var}(\mathbf{X}|\mathbf{Y} = y) = \sum_x (x - E(\mathbf{X}|\mathbf{Y} = y))^2 p_{X|Y}(x|y)
  \]
- In case both \( \mathbf{X} \) and \( \mathbf{Y} \) are continuous, we have
  \[
  E(\mathbf{X}|\mathbf{Y} = y) = \int_{-\infty}^{\infty} xf_{X|Y}(x|y) \, dx
  \]
  and
  \[
  \text{Var}(\mathbf{X}|\mathbf{Y} = y) = \int_{-\infty}^{\infty} (x - E(\mathbf{X}|\mathbf{Y} = y))^2 f_{X|Y}(x|y) \, dx
  \]

Remarks:

1. In general, \( \mathbf{X} \) and \( \mathbf{Y} \) can have different types of distribution (e.g., one is discrete, the other is continuous).

   Example: Let \( n = m = 1 \) and \( \mathbf{X} = Y + Z \), where \( Y \) is a Bernoulli(\( \mu \)) r.v. and \( Z \sim N(0, \sigma^2) \). and \( \mathbf{Y} \) and \( \mathbf{Z} \) are independent. Determine the conditional pdf of \( \mathbf{X} \) given \( \mathbf{Y} = 0 \) and \( \mathbf{Y} = 1 \). Also, determine the pdf of \( \mathbf{X} \).

   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.

Special case: Assume \( \mathbf{X} \) and \( \mathbf{Y} \) are independent. Then (considering the discrete case)

\[
 p_{X|Y}(x|y) = p_X(x)
\]

so that for all \( y \),

\[
 E(\mathbf{X}|\mathbf{Y} = y) = \sum_x xp_{X|Y}(x|y) = \sum_x xp_X(x) = E(\mathbf{X})
\]

A similar argument shows \( E(\mathbf{X}|\mathbf{Y} = y) = E(\mathbf{X}) \) if \( \mathbf{X} \) and \( \mathbf{Y} \) 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 \( h(y) = \text{Var}(X|Y = y) \), the random variable \( \text{Var}(X|Y) \) is defined by

\[
\text{Var}(X|Y) = h(Y)
\]

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

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

**Theorem 1 (Law of total expectation)**

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

**Proof:** Assume both \( X \) and \( Y \) are discrete. Then

\[
E[E(X|Y)] = \sum_y E(X|Y = y)p_Y(y) = \sum_y \left( \sum_x x p_{X|Y}(x|y) \right) p_Y(y) = \sum_y \left( \sum_x x p_{X,Y}(x,y) \right) p_Y(y) = \sum_y \sum_x x p_{X,Y}(x,y) = \sum_x x p_X(x) = E(X)
\]

**Example:** Expected value of geometric distribution...

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 \( a, b \in \mathbb{R} \),

\[
E(aX_1 + bX_2|Y) = aE(X_1|Y) + bE(X_2|Y)
\]

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

\[
E[g(Y)|Y] = g(Y)
\]

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

\[
E[g(Y)X|Y] = g(Y)E(X|Y)
\]

**Lemma 2 (Variance formula)**

\[
\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}[E(X|Y)]
\]

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

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

Taking expectation (with respect to \( Y \)),

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

On the other hand,

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

so

\[
\text{Var}(X) = E(X^2) - (E(X))^2 = E[\text{Var}(X|Y)] + \text{Var}[E(X|Y)]
\]
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

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

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:

(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 $Z$,

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

so that

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

For example, if $Z$ is discrete,

$$E(X|Y = y) = \sum_z E(X|Y = y, Z = z)p_{Z|Y}(z|y)$$

$$= \sum_z E(X|Y = y, Z = z)p(Z = z|Y = y)$$

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

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

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

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 $g$,

$$E[(X - g(Y))^2] \geq E[(X - g^*(Y))^2]$$
**Proof:** Use the properties of conditional expectation:

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

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

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

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

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

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

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

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

i.e., \( E[\text{Var}(X|Y)] \) 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 \( Y = y \)?

**Proof cont'd**

Thus

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

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

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

Since \( (g^*(Y) - g(Y))^2 \geq 0 \), this implies

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

\[
\square
\]

**Random Sums**

**Theorem 4 (Wald’s equation)**

Let \( X_1, X_2, \ldots \) be i.i.d. random variables with mean \( \mu \). 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:**

\[
E(X|N = n) = E(X_1 + \cdots + X_n|N = n)
\]

\[
= E(X_1 + \cdots + X_n|N = n)
\]

\[
= E(X_1|N = n) + \cdots + E(X_n|N = n)
\]

(linearity of expectation)

\[
= E(X_1) + \cdots + E(X_n)
\]

\[
= n\mu
\]

\[
\text{Work in progress.}
\]
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

$$E(X) = E[E(X|N)] = E(N\mu) = E(N)\mu \quad \Box$$

Example: (Branching Process) Suppose a population evolves in generations starting from a single individual (generation 0). Each individual of the $i$th 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 $\mu$. Let $X_n$ be the number of individuals in the $n$th generation. Find $E(X_n)$.

2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
3. If $X = Y$ we obtain

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

4. For any constants $a$, $b$, $c$ and $d$,

$$\text{Cov}(aX + b, cY + d)$$

$$= E[(aX + b - E(aX + b))(cY + d - E(cY + d))]$$

$$= E[a(X - E(X))c(Y - E(Y))]$$

$$= acE[(X - E(X))(Y - E(Y))]$$

$$= ac\text{Cov}(X, Y)$$

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

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

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

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

Remarks:

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

Covariance and Correlation

**Covariance**

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

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

**Properties:**

1. $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$

2. $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$

3. If $X$ and $Y$ are independent, then $\text{Cov}(X, Y) = 0$.
Some properties of covariance are easier to derive using a matrix formalism.

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

\[
E(V) = E\left[\begin{bmatrix} Y_{11} & \ldots & Y_{1n} \\ Y_{21} & \ldots & Y_{2n} \\ \vdots & \ddots & \vdots \\ Y_{m1} & \ldots & Y_{mn} \end{bmatrix}\right] = \left[\begin{bmatrix} E(Y_{11}) & \ldots & E(Y_{1n}) \\ E(Y_{21}) & \ldots & E(Y_{2n}) \\ \vdots & \ddots & \vdots \\ E(Y_{m1}) & \ldots & E(Y_{mn}) \end{bmatrix}\right]
\]

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

\[
\text{Cov}(X) = E[(X - E(X))(X - E(X))^T]
\]

For any \( m \)-vector \( c = (c_1, \ldots, c_m)^T \) we also have

\[
\text{Cov}(Y + c) = \text{Cov}(Y)
\]

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

Thus

\[
\text{Cov}(AX + c) = A\text{Cov}(X)A^T
\]

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

\[
\text{Cov}(a^TX + c) = \text{Cov}\left(\sum_{i=1}^n a_iX_i + c\right) = \text{Var}\left(\sum_{i=1}^n a_iX_i + c\right)
\]

Lemma 5

Let \( A \) be an \( m \times n \) real matrix and define \( Y = AX \) (an \( m \)-dimensional random vector). Then

\[
\text{Cov}(Y) = A\text{Cov}(X)A^T
\]

Proof: First note that by the linearity of expectation,

\[
E(Y) = E(AX) = AE(X).
\]

Thus

\[
\text{Cov}(Y) = E[(Y - E(Y))(Y - E(Y))^T]
\]

\[
= E[(AX - AE(X))(AX - AE(X))^T]
\]

\[
= E[(A(X - E(X)))(A(X - E(X)))^T]
\]

\[
= AE[(X - E(X))(X - E(X))^T]A^T
\]

\[
= A\text{Cov}(X)A^T \quad \square
\]

On the other hand,

\[
\text{Cov}(a^TX + c) = a^T\text{Cov}(X)a = \sum_{i=1}^n \sum_{j=1}^n a_ia_j\text{Cov}(X_i, X_j)
\]

\[
= \sum_{i=1}^n a_i^2\text{Var}(X_i) + 2\sum_{i<j} a_ia_j\text{Cov}(X_i, X_j)
\]

Hence

\[
\text{Var}\left(\sum_{i=1}^n a_iX_i + c\right) = \sum_{i=1}^n a_i^2\text{Var}(X_i) + 2\sum_{i<j} a_ia_j\text{Cov}(X_i, X_j)
\]

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

\[
\text{Var}\left(\sum_{i=1}^n a_iX_i + c\right) = \sum_{i=1}^n a_i^2\text{Var}(X_i)
\]
More generally, let $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)^T$ and let $\text{Cov}(\mathbf{X}, \mathbf{Y})$ be the $n \times m$ matrix with $(i, j)$th entry $\text{Cov}(X_i, Y_j)$. Note that

\[
\text{Cov}(\mathbf{X}, \mathbf{Y}) = E[(\mathbf{X} - E(\mathbf{X}))(\mathbf{Y} - E(\mathbf{Y}))^T]
\]

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

\[
\text{Cov}(\mathbf{AX} + \mathbf{c}, \mathbf{BY} + \mathbf{d}) = \mathbf{A} \text{Cov}(\mathbf{X}, \mathbf{Y}) \mathbf{B}^T
\]

The following property of covariance is of fundamental importance:

**Lemma 7**

\[
|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}
\]

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

\[
0 \leq E[(U - \lambda V)^2] = E(U^2 - 2\lambda UV + \lambda^2 V^2) = E(U^2) - 2\lambda E(UV) + \lambda^2 E(V^2)
\]

This is a quadratic polynomial in $\lambda$ which cannot have two distinct real roots.

We can now prove the following important property of covariance:

**Lemma 6**

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

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

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

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

\[
\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \text{Cov}(a^T \mathbf{X}, b^T \mathbf{Y}) = \text{Cov}(\mathbf{AX}, \mathbf{BY})
\]

\[
= A \text{Cov}(\mathbf{X}, \mathbf{Y}) B^T = a^T \text{Cov}(\mathbf{X}, \mathbf{Y}) b
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, Y_j)
\]

**Proof cont’d:** Thus its discriminant cannot be positive:

\[
4[E(UV)]^2 - 4E(U^2)E(V^2) \leq 0
\]

so we obtain

\[
[E(UV)]^2 \leq E(U^2)E(V^2)
\]

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

\[
|\text{Cov}(X, Y)| = |E[(X - E(X))(Y - E(Y))]| \leq \sqrt{E[(X - E(X))^2]E[(Y - E(Y))^2]}
\]

\[
= \sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}
\]

\[\square\]
Correlation

Recall that $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$. This is an undesirable property if we want to use $\text{Cov}(X, Y)$ 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

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

**Remarks:**

- Since $\text{Var}(aX + b) = a^2 \text{Var}(X)$,
  $$\rho(aX + b, aY + d) = \rho(X, Y)$$

**Theorem 8**

The correlation always satisfies

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

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 $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}$, so $|\rho(X, Y)| \leq 1$ always holds.

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

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

so

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

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

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

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

**Proof cont’d:**

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

$$\text{Var}(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}) = \text{Cov}(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}, \frac{X}{\sigma_X} - \frac{Y}{\sigma_Y})$$

$$= \text{Var}(\frac{X}{\sigma_X}) + \text{Var}(\frac{Y}{\sigma_Y}) - 2 \text{Cov}(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y})$$

$$= \frac{\text{Var}(X)}{\sigma_X^2} + \frac{\text{Var}(Y)}{\sigma_Y^2} - 2 \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$= 1 + 1 - 2 = 0$$

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

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

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

**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 \( Y \) are 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 \textit{linear} function of \( Y \); i.e., we are looking for \( a \in \mathbb{R} \) minimizing \( E[(X - aY)^2] \).

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

\textit{Solution:} ...