# Introduction to Probability
# Expectation and variance in a multivariate setting

# Sums of RVs

If $X_1$, $X_2$, ..., $X_n$ are $n$ random variables, then their sum

$$\sum_{j=1}^{n} X_j = X_1 + X_2 + \cdots + X_n$$

is also a random variable.

# Sums of RVs

These kinds of sums are easy to deal with if $X_1$, $X_2$, ..., $X_n$ are

**independent and identically distributed (IID)**:

- independent: $X_i \perp X_j$ for every $i \neq j$
- identically distributed: $X_i \sim X_j$ for every $i \neq j$,
  i.e. they have the same distribution

# Expectation of sums of RVs is a linear operation

The expected value of a sum of random variables can be calculated term-by-term.

$$E\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} E\left(X_j\right).$$

# Expectation is a linear operation

The expected value of a sum of scaled random variables can be calculated term-by-term, with constant multiples moving outside the sum. (Expectation is called a **linear** operation.)

$$E\left(\sum_{j=1}^{n} c_j X_j\right) = c_j \sum_{j=1}^{n} E\left(X_j\right).$$

# Expectation is a linear operation

This works for discrete and continuous random variables:

- If the $X_j$ are discrete, linearity is a property of summing.
- If the $X_j$ are continuous, linearity is a property of integrating.

Only the *finiteness* of each term's expectation is required; independence is *not* required here.

# Variance of sums of independent RVs

The variance of a sum of random variables is easy to calculate *if all the random variables are independent*; if $X_i \perp X_j$ for $i \neq j$, then

$$Var\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} Var\left(X_j\right).$$

If the $X_j$ are *not* independent, this fails - there are extra terms.

(Think about the square $\left(\sum_{j=1}^{n} X_j\right)^2$ - what terms cancel?)

# Sample Mean

The **sample mean** of $n$ IID samples,

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

is the average of $n$ sample points.

# Sample Mean

If $E(X_1) = \mu < \infty$ is the expectation of one sample,
then the sample mean, by the linearity of expectation, is

$$E(\overline{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}(n\mu) = \mu.$$

In other words, you expect the same from the sample average as you do from each sample.

# Statistics

The sample mean is an example of a **statistic**.

A **statistic** is simply a function of (IID) random variables,

$$Y = f(X_1, X_2, ..., X_n)$$

which is itself a random variable.

# Statistics; Unbiased Estimator

A statistic $Y$ is called an **unbiased estimator** of a parameter $a$ of

$$X_1, X_2, ..., X_n$$

if $E(Y) = a$.

Thus, the sample mean $\overline{X}_n$ is an unbiased estimator of $\mu$.

# Variance of Sample Mean

If $Var(X_1) = \sigma^2$, then the variance of the sample mean $\overline{X}_n$ is

$$Var(\overline{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right).$$

# Variance of Sample Mean

Since the $X_i$ are independent, this simplifies to

$$Var(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Thus, the variance of the sample mean shrinks as the number of IID samples $n$ increases.

**Note:** As $n \to \infty$, since $Var(\overline{X}_n) \to 0$, $\overline{X}_n$ converges to $\mu$.

# Sample Variance

The **sample variance*** of the IID samples $X_1, ..., X_n$ is defined by

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

$s_n^2$ is an unbiased estimator of $\sigma^2$.

*The use of $n-1$ instead of $n$ in the definition of $s_n^2$ is called **Bessel's correction**, after Friedrich Bessel (1784-1846).

# Sample Variance

**Proof** First, compute

$$E((X_i - \overline{X}_n)^2) = \frac{n+1}{n}\sigma^2 - 2E((\overline{X}_n - \mu)(X_i - \mu))$$

and sum from $i = 1$ to $n$ to get

$$\begin{aligned}(n-1)E(s_n^2) &= \sum_{i=1}^{n}\left[\frac{n+1}{n}\sigma^2 - 2E((\overline{X}_n - \mu)(X_i - \mu))\right] \\ &= (n+1)\sigma^2 - \frac{2n\sigma^2}{n} = (n-1)\sigma^2.\end{aligned}$$

Thus,

$$E(s_n^2) = \sigma^2. \blacksquare$$

# Coupon Collector's Problem: Collect 'em all!

There are $n$ collectables in a series.

You get one per package, packaged uniformly at random.

How many must you buy before you collect all $n$?

# Coupon Collector's Problem: Sum of Random Times

Let $T_n$ be the number of purchases we make before collecting all $n$.

We will compute $E(T_n)$ and $Var(T_n)$, but not its full PMF.

Let $T_j$ be the first time we get the $j$th new item in our set (order does not matter, just novelty).

Clearly, $T_1 = 1$ since the first purchase is always new.

# Coupon Collector's Problem: Sum of Geometrics

Let $W_1$ be the amount of time after $T_1$ that it takes to reach $T_2$.

That is, let $W_1 = T_2 - T_1$.

Each new purchase between times $T_1$ and $T_2$, with the goal of "getting a new collectable", is independent, and

- a "failure" with probability $\frac{1}{n}$ if you repeat the item you have;
- a "success" with probability $p_1 = \frac{n-1}{n}$ if you get a new one.

Thus, $W_1 \sim Geom(p_1 = \frac{n-1}{n})$.

# Coupon Collector's Problem: Sum of Geometrics

Likewise, let $W_k$ be the amount of time after $T_k$ until $T_{k+1}$.

That is, let $W_k = T_{k+1} - T_k$.

Then each purchase between times $T_k$ and $T_{k+1}$ is

- a "failure" with probability $\frac{k}{n}$;
- a "success" with probability $p_k = \frac{n-k}{n}$.

Hence, $W_k \sim Geom(p_k = \frac{n-k}{n})$, and we can conclude that

$$\begin{aligned} T_n &= T_1 + (T_2 - T_1) + \cdots + (T_n - T_{n-1}) \\ &= 1 + W_1 + \cdots + W_{n-1}. \end{aligned}$$

# Coupon Collector's Problem: Sum of Geometrics

The $W_k$ are independent (but not identically distributed), and

$$\begin{aligned} E(T_n) &= 1 + E(W_1) + \cdots + E(W_{n-1}) \\ &= 1 + \frac{1}{p_1} + \cdots + \frac{1}{p_{n-1}} \\ &= 1 + n\sum_{k=1}^{n-1} \frac{1}{n-k} = 1 + n\sum_{j=1}^{n} \frac{1}{j} \end{aligned}$$

$$Var(T_n) = n^2 \sum_{j=1}^{n-1} \frac{1}{j^2} - n\sum_{j=1}^{n-1} \frac{1}{j}.$$

# Coupon Collector's Problem: Logarithmic Expectation

As $n$ increases, these values both increase as well: asymptotics are

$$E(T_n) = 1 + n \sum_{j=1}^{n} \frac{1}{j} \sim n \ln(n)$$

$$Var(T_n) = n^2 \sum_{j=1}^{n-1} \frac{1}{j^2} - n \sum_{j=1}^{n-1} \frac{1}{j} \sim \frac{\pi^2}{6} n^2$$

# Moment Generating Function of a sum of RVs

If $X \perp Y$, then the MGF of $X + Y$ factors:

$$\begin{aligned} M_{X+Y}(t) &= E(e^{t(X+Y)}) \\ &= E(e^{tX} e^{tY}) \\ &= E(e^{tX}) E(e^{tY}) \quad (X \perp Y) \\ &= M_X(t) M_Y(t). \end{aligned}$$

# Sums of Poissons, Normals via MGFs

This is clear in, for example, the Poisson and normal distributions.

**Example**

$X \sim Poisson(\lambda)$, $Y \sim Poisson(\mu)$, and $X \perp Y$ implies

$$M_X(t)M_Y(t) = e^{\lambda(e^t-1)}e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)} = M_{X+Y}(t),$$

i.e. $X + Y \sim Poisson(\lambda + \mu)$.

# Sums of Poissons, Normals via MGFs

Example

$X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, and $X \perp Y$ implies

$$\begin{aligned} M_X(t)M_Y(t) &= e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2} e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2} \\ &= e^{(\mu_1+\mu_2)t + \frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2} = M_{X+Y}(t), \end{aligned}$$

i.e. $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

# Variance of sums of RVs

Assume $\mu_1 = E(X_1)$, $\mu_2 = E(X_2)$. Then, expanding,

$$\begin{aligned}
Var(X_1 + X_2) &= E((X_1 + X_2)^2) - [E(X_1 + X_2)]^2 \\
&= E(X_1^2 + 2X_1X_2 + X_2^2) - [\mu_1 + \mu_2]^2 \\
&= E(X_1^2 + 2X_1X_2 + X_2^2) - [\mu_1^2 + 2\mu_1\mu_2 + \mu_2^2] \\
&= E(X_1^2) + 2E(X_1X_2) + E(X_2^2) - \mu_1^2 - 2\mu_1\mu_2 - \mu_2^2 \\
&= [E(X_1^2) - \mu_1^2] + 2[E(X_1X_2) - \mu_1\mu_2] + [E(X_2^2) - \mu_2^2] \\
&= Var(X_1) + 2[E(X_1X_2) - \mu_1\mu_2] + Var(X_2).
\end{aligned}$$

# Covariance

The variance of a sum of two random variables,

$$Var(X_1 + X_2) = Var(X_1) + 2[E(X_1X_2) - \mu_1\mu_2] + Var(X_2),$$

has a middle term which needs a name, since it's not always zero.

The **covariance** of two random variables $X_1$ and $X_2$, with means $\mu_1$ and $\mu_2$, respectively is denoted $Cov(X_1, X_2)$, and defined by

$$Cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)].$$

The covariance is a measure of the *linear relationship* between the two random variables $X_1$ and $X_2$.

# Covariance

We have a computational formula for $Cov(X_1, X_2)$:

$$\begin{aligned} Cov(X_1, X_2) &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E(X_1 X_2 - \mu_1 X_2 - \mu_2 X_1 + \mu_1 \mu_2) \\ &= E(X_1 X_2) - \mu_1 \mu_2. \end{aligned}$$

# Covariance of Indicators

Let $X = 1_A$ and $Y = 1_B$ be indicator functions on $\Omega$.

The covariance of $X$ and $Y$ is

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(X)E(Y) \\ &= E(1_A 1_B) - E(1_A)E(1_B) \\ &= P(AB) - P(A)P(B). \end{aligned}$$

Thus, for indicator functions, the covariance is a measure of dependence of the events $A$ and $B$.

# Example: roll two fair D10s

Roll two fair 10-sided dice (each uniform on $\{0, 1, ..., 8, 9\}$). Let

- $X = \#$ of evens rolled ($X \sim Bin(2, \frac{5}{10})$, $E(X) = 1$),
- $Y = \#$ of sixes rolled ($Y \sim Bin(2, \frac{1}{10})$, $E(Y) = \frac{2}{10}$).

There are $10^2 = 100$ possible rolls on 2D10 (two 10-sided dice).

# Example: roll two fair D10s

You can easily see that $X \geq Y$

(six $\implies$ even; not even $\implies$ not six).

The joint PMF is

$$p_{X,Y}(x,y) = \begin{cases} (\frac{5}{10})^2 = 0.25 & x = 0, y = 0 \text{ (both odds)} \\ 2(\frac{4}{10})(\frac{5}{10}) = 0.40 & x = 1, y = 0 \text{ (1 not-six even, 1 odd)} \\ (\frac{4}{10})^2 = 0.16 & x = 2, y = 0 \text{ (both not-six even)} \\ 2(\frac{5}{10})(\frac{1}{10}) = 0.10 & x = 1, y = 1 \text{ (1 six, 1 odd)} \\ 2(\frac{1}{10})(\frac{4}{10}) = 0.08 & x = 2, y = 1 \text{ (1 six, 1 not-six even)} \\ (\frac{1}{10})^2 = 0.01 & x = 2, y = 2 \text{ (both six).} \end{cases}$$

# Example: roll two fair D10s

The cross-expectation $E(XY)$ is

$$E(XY) = 0(0.25+0.40+0.16)+1(0.10)+2(0.08)+4(0.01) = 0.30$$

and so their covariance is

$$Cov(X,Y) = E(XY) - E(X)E(Y) = 0.30 - 0.20 = 0.10.$$

These two RVs are slightly positively correlated.

# Properties of Covariance

- If $X_1 = X_2$, covariance is variance: if $E(X_1) = \mu_1$,

$$Cov(X_1, X_1) = E(X_1 X_1) - \mu_1 \mu_1 = E(X_1^2) - \mu_1^2 = Var(X_1).$$

- Scaling: if $a, b \in \mathbb{R}$, then $E(aX_1) = a\mu_1$ and $E(bX_2) = b\mu_2$,

$$Cov(aX_1, bX_2) = E(abX_1 X_2) - ab\mu_1\mu_2 = abCov(X_1, X_2).$$

- Sums: If $X, Y, W$ are three random variables,

$$Cov(X + Y, W) = Cov(X, W) + Cov(Y, W).$$

- Variance of a sum of $n$ RVs:

$$Var(X_1 + X_2 + ... + X_n) = \sum_{j=1}^{n} Var(X_j) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} Cov(X_i, X_j).$$

# Correlation

Covariance gives a measure of the linear relationship between two random variables, but it's not easy to understand.

The **correlation** between two random variables $X_1, X_2$, denoted

$$\rho(X_1, X_2),$$

is the *normalized* covariance in the following sense: let

$$E(X_1) = \mu_1,\ E(X_2) = \mu_2,\ Var(X_1) = \sigma_1^2,\ Var(X_2) = \sigma_2^2,$$

and $Cov(X_1, X_2) = \sigma_{12}$.

# Correlation

Define the *normalized* versions of $X_1$ and $X_2$ by

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1},\ Z_2 = \frac{X_2 - \mu_2}{\sigma_2}.$$

Then

$$E(Z_1) = 0,\ E(Z_2) = 0,\ Var(Z_1) = 1,\ \text{and}\ Var(Z_2) = 1,$$

and

$$\rho(X_1, X_2) = Cov(Z_1, Z_2) = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}.$$

# Example: roll two fair D10s

Roll two fair 10-sided dice. Let

- $X = \#$ of evens rolled ($X \sim Bin(2, \frac{5}{10})$, $E(X) = 1$),
- $Y = \#$ of sixes rolled ($Y \sim Bin(2, \frac{1}{10})$, $E(Y) = \frac{2}{10}$).

Their covariance and variances are

$$Cov(X, Y) = E(XY) - E(X)E(Y) = 0.30 - 0.20 = 0.10 = \frac{1}{10}$$
$$Var(X) = E(X^2) - E(X)^2 = 2(1/2)(1/2) = 0.5 = \frac{1}{2}$$
$$Var(Y) = E(Y^2) - E(Y)^2 = 2(1/10)(9/10) = 0.18 = \frac{9}{50}$$

and so their correlation is

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\frac{1}{10}}{\sqrt{\frac{1}{2} \cdot \frac{9}{50}}} = \frac{\frac{1}{10}}{\frac{3}{10}} = \frac{1}{3}.$$

# Properties of Correlation

- Correlation is always between -1 and 1: $-1 \leq \rho(X_1, X_2) \leq 1$.
- If $\rho(X_1, X_2) > 0$, we say $X_1$ and $X_2$ are **positively correlated**.
- If $\rho(X_1, X_2) < 0$, we say $X_1$ and $X_2$ are **negatively correlated**.
- If $\rho(X_1, X_2) = 0$, we call $X_1$ and $X_2$ **uncorrelated**.
- $X_2 = aX_1 + b \iff \rho(X_1, X_2) = \pm 1$
  (the function $sign(a) = 1_{\{a>0\}} - 1_{\{a<0\}}$ for $a \neq 0$):

$$\rho(X_1, aX_1 + b) = \frac{Cov(X_1, aX_1 + b)}{\sqrt{Var(X_1)Var(aX_1 + b)}} = \frac{aVar(X_1)}{|a|Var(X_1)} = sign(a).$$

# Relationship of RV independence to correlation

It is **VERY** important to remember that

$$X_1 \perp X_2 \implies \rho(X_1, X_2) = Cov(X_1, X_2) = 0,$$

but the converse is NOT true in general!

There are plenty of pairs of random variables $X_1$ and $X_2$ that are uncorrelated, but share a higher-order relationship, and so can have $\rho(X_1, X_2) = 0$ but are NOT independent.

# Relationship of RV independence to correlation

We can see this in a very simple way:

$$Cov(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2) = 0$$

$$\implies E(X_1 X_2) = E(X_1)E(X_2)$$

but

$$E(X_1 X_2) = E(X_1)E(X_2)$$

does NOT alone imply $X_1 \perp X_2$.

# Relationship of RV independence to correlation

Let $(X, Y)$ be a pair of random variables with joint PMF

$$p_{X,Y}(x, y) = \begin{cases} \frac{1}{4} & (-1, 1) \\ \frac{1}{2} & (0, -1) \\ \frac{1}{4} & (1, 1). \end{cases}$$

# Relationship of RV independence to correlation

Then $X$ and $Y$ are uncorrelated but not independent:

$$E(XY) = -1(1/4) + 0(1/2) + 1(1/4) = 0$$

$$E(X) = -1(1/4) + 0(1/2) + 1(1/4) = 0$$

$$E(Y) = 1(1/2) + -1(1/2) = 0$$

$$\implies Cov(X,Y) = E(XY) - E(X)E(Y) = 0,$$

$$\text{but } P(X=0, Y=1) = 0 \neq P(X=0)P(Y=1) = \left(\frac{1}{4}\right)\left(\frac{1}{2}\right) = \frac{1}{8}.$$

# Covariance, Correlation (Continuous)

The **covariance** and **correlation** of two continuous RVs, $X$ and $Y$, with means $\mu_X$, $\mu_Y$, come from the previous definitions and computational formulae:

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)E(Y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)dydx - \mu_X\mu_Y \end{aligned}$$

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

All general properties of covariance and correlation carry over.

# Bivariate Normal Random Variables

The **bivariate normal distribution** is a pair $(X, Y)$ of normal random variables that are correlated with correlation $\rho$.

The pair with marginals $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ can be generated by a linear transformation of a pair of independent standard normals $(Z, W)$ by

$$
\begin{aligned}
X &= \sigma_X Z + \mu_X, \\
Y &= \sigma_Y \rho Z + \sigma_Y \sqrt{1-\rho^2} W + \mu_Y,
\end{aligned}
$$

and can be represented by the joint PDF

$$f_{(X,Y)}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}.$$