Chapter 3

Expectation

3.1 Expectation

Definition 3.1. Let X be a discrete random variable. The **expectation** of X, denoted by $E(X)$, is defined as:

$$E(X) = \sum_{\text{all } x} xP(X = x).$$

The expectation of X is also referred to as the **mean** of X or the **expected value** of X.

In other words, the expected value of X is a *weighted average* of the possible values that X can take on, weighted by their probabilities. If the values are of equal probability, expectation is the simple average of all x: $E(X) = \frac{1}{n} \sum x$.

The expected value of X is a *number* (if it exists), $E(X) \in \mathbb{R}$. It is not a random variable such as a function of X.

Sometimes, we would like to omit the parentheses for simplicity and write $EX \equiv E(X)$. We also like to denote expectation by the greek letter $\mu \equiv E(X)$.

Example 3.1. The expectation of a Bernoulli random variable $X \sim \text{Bern}(p)$:

$$E(X) = 1 \times P(X = 1) + 0 \times P(X = 0) = p.$$

Example 3.2. The expectation of a Binomial random variable $X \sim \text{Bin}(n, p)$:

$$\begin{aligned}
E(X) &= \sum_{k=0}^{n} kp(k) \\
&= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=1}^{n} n \cdot \binom{n-1}{k-1} p^k q^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k} \\
&= np \underbrace{\sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-1-j}}_{\text{another Binomial PMF}} \\
&= np.
\end{aligned}$$

Example 3.3. Life expectancy is the average number of years a person is expected to live. It is a crucial indicator of the quality of living and one of the three components of the Human Development Index (HDI) (the other two components are education and per capita GDP). Here is a toy example to compute life expectancy with hypothetical data.[1]

(1)	(2)	(3)	(4)		(5)	(6)
Age	Population	Mortality rates	# Survive		# Died at age	P(Age)
0	200	1%	1000		10	1%
20	300	2%	990	=1000(1-1%)	20	2%
40	250	10%	970	=990(1-2%)	97	10%
60	150	20%	873	=970(1-10%)	175	17%
80	100	100%	699	=873(1-20%)	699	70%
Total	1000					

Table 3.1: Hypothetical mortality rates and life table

To simplify our analysis, we will assume there are only five possible ages: 0, 20, 40, 60, and 80. A baby is born at age 0, and can either die at that age or

[1] This is an overly simplified example that only serves to clarify the definition of expectation. See this tutorial from MEASURE Evaluation for the actual computation of life expectancy.

survive to age 20. We intentionally exclude intermediate ages such as 5 and 10 for the sake of computational simplicity.

It's important to note that life expectancy is not the same as the average age of the population. For instance, based on the hypothetical data presented, the average age can be calculated as:

$$\overline{\text{Age}} = (0 \times 200 + 20 \times 300 + 40 \times 250 + 60 \times 150 + 80 \times 100)/1000 = 33.$$

However, the expected age, denoted as $E(\text{Age})$, is defined as:

$$E(\text{Age}) = \sum \text{Age} \times P(\text{Age}).$$

To compute this expected value, we need to determine $P(\text{Age})$, the probability of living to a specific age or dying at that age. This requires consideration of the mortality rate at each age, which is given in Column 3.

Assuming 1000 babies are born at age 0, with a mortality rate of 1% at that age, we find that 99% of the babies survive to age 20. Thus, the number of babies that survive to age 20 is: $1000 \times (1 - 1\%) = 990$. We can apply similar calculations to determine the number of survivors at each subsequent age.

The number of individuals who die at a specific age (Column 5) is the difference between the number of survivors at that age and the next (Column 4). To find the probability of living to a specific age, we compute: $P(\text{Age}) =$ Column 4/1000.

Finally, we compute the expected value of age (or life expectancy) as follows:

$$E(Age) = 0 \times 1\% + 20 \times 2\% + 40 \times 10\% + 60 \times 17\% + 80 \times 70\% = 70.6.$$

This figure differs from the average age. Since the mortality rate is low at younger ages, the probabilities $P(\text{Age})$ for these ages are also low, while they are higher for older ages. This example illustrates the distinction between average and expected values. In everyday conversation, we may use these terms interchangeably, but in certain contexts, expected values can significantly differ from averages.

3.2 Linearity of expectation

Theorem 3.1. *For any random variables X, Y and any constant c,*

$$\begin{aligned} E(X+Y) &= E(X)+E(Y), \\ E(cX) &= cE(X). \end{aligned}$$

This property holds regardless of the dependencies between the random variables.

Proof. The proof is not as straightforward as it seems. It is hard to combine the two random variables:

$$E(X)+E(Y) = \sum_x xP(X=x) + \sum_y yP(Y=y) \overset{?}{=} \sum (x+y)P(X+Y=x+y).$$

The problem becomes easier if the number of possible values of X and Y are the same and all values are equally likely,

$$E(X)+E(Y) = \frac{1}{n}\sum x + \frac{1}{n}\sum y = \frac{1}{n}\sum(x+y) = E(X+Y).$$

The original problem is equivalent to the simple case if realizing that the weighted average is jut a simple average with repetitive values. For example,

$$1 \times \frac{1}{4} + 2 \times \frac{2}{4} + 3 \times \frac{1}{4} = \frac{1}{4}(1+2+2+3).$$

Imagine the sample space as being composed of "atom" outcomes $\{\omega\}$, each with equal probability $P(\omega)$. All random variable are function of these atoms, $X(\omega)$, and $Y(\omega)$. Therefore, the expectation formula can be rewritten as

$$E(X)+E(Y) = \sum_\omega X(\omega)P(\omega) + \sum_\omega Y(\omega)P(\omega) = \sum_\omega (X+Y)(\omega)P(\omega) = E(X+Y).$$

Here is another way to prove linearity for discrete random variables:

$$
\begin{aligned}
E(X+Y) &= \sum_{z=x+y} zP(X+Y=z)\\
E(X+Y) &= \sum_x \sum_y (x+y)P(X=x,Y=y)\\
&= \sum_x \sum_y xP(X=x,Y=y) + \sum_x \sum_y yP(X=x,Y=y)\\
&= \sum_x x \sum_y P(X=x,Y=y) + \sum_y y \sum_x P(X=x,Y=y)\\
&= \sum_x xP((X=x)\cap \bigcup_{\text{all } y}(Y=y)) + \sum_y yP(\bigcup_{\text{all } x}(X=x)\cap(Y=y))\\
&= \sum_x xP(X=x) + \sum_y yP(Y=y)\\
&= E(X)+E(Y).
\end{aligned}
$$

□

Corollary 3.1. *Further properties on the linearity of expectations:*

- If $Y = aX + b$, then $E(Y) = aE(X) + b$.
- $E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$
- $E(a_1X_1 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b$

Example 3.4. Redo the expectation of $X \sim \text{Bin}(n, p)$ with properties of expectation:

$$E(X) = E(X_1 + \cdots + X_n) = nE(X_i) = np$$

where $X_i \sim \text{Bern}(p)$.

Example 3.5. Let $X \sim \text{HGeom}(w, b, n)$. Find $E(X)$ the expected number of white balls. Similarly, we can decompose X:

$$X = I_1 + \cdots + I_n$$

where I_j equals 1 if the jth ball is white and 0 otherwise. We have said that $\{I_j\}$ are not independent, but the property of linearity still holds:

$$E(X) = E(I_1 + \cdots + I_n) = E(I_1) + \cdots + E(I_n).$$

Meanwhile we have

$$E(I_j) = P(j\text{-th ball is white}) = \frac{w}{w+b}$$

since unconditionally the jth ball is equally likely to be any of the balls. Thus, $E(X) = \frac{nw}{w+b}$.

Example 3.6. In a group of n people, what is the expected number of distinct birthdays among the n people (the expected number of days on which at least one of the people was born)? What is the expected number of people sharing a birthday (any day)?

Solution: Let X be the number of distinct birthdays, and write $X = I_1 + \cdots + I_{365}$, where

$$I_j = \begin{cases} 1 & \text{if someone was born on day } j \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$\begin{aligned} E(I_j) &= P(\text{someone was born on day } j) \\ &= 1 - P(\text{no one was born on day } j) \\ &= 1 - \left(\frac{364}{365}\right)^n. \end{aligned}$$

Then by linearity,

$$E(X) = 365\left(1 - \left(\frac{364}{365}\right)^n\right).$$

Let Y be the number of people sharing a birthday, and $Y = J_1 + \cdots + J_n$ where J_k is an indicator that the j-th person shares his birthday with somebody else.

$$\begin{aligned} E(J_k) &= P(\text{someone shares birthday with } k) \\ &= 1 - P(\text{no one shares birthday with } k) \\ &= 1 - \left(\frac{364}{365}\right)^{n-1}. \end{aligned}$$

Therefore,

$$E(Y) = \sum_{k=1}^{n} E(J_k) = n\left(1 - \left(\frac{364}{365}\right)^{n-1}\right).$$

For some numeric values, $E(Y) = 2.3$ if $n = 30$; $E(Y) = 6.3$ if $n = 50$.

Example 3.7. Suppose that there are n people sitting in a classroom with exactly n seats. At some point, everyone got up, ran around the room, and sat back down randomly (i.e., all seating arrangements are equally likely). What is the expected value of the number of people sitting in their original seat?

Solution: Number the people from 1 to n. Let X_i be the Bernoulli random variable with value 1 if person i returns to their original seat and value 0 otherwise. Since person i is equally likely to sit back down in any of the n seats, the probability that person i returns to their original seat is $1/n$. Therefore $E[X_i] = 1/n$. Now, let X be the number of people sitting in their original seat following the rearrangement. Then $X = X_1 + X_2 + \cdots + X_n$. By linearity of expected values, we have $E[X] = \sum E[X_i] = \sum 1/n = 1$.

Example 3.8. Let Π be a permutation over $\{1, 2, \ldots, n\}$. That is a reordering of the numbers. A fixed point of a permutation are the points not moved by the permutation. For example, in the permutation below

$$\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ \Pi & 2 & 4 & 3 & 1 \end{array}$$

The fixed point is 3. Find the expected number of fixed points of a random permutation.

Solution: Let X be the number of fixed points of a random permutation. Then $X = \sum_{k=1}^{n} \mathbf{1}_{\Pi(k)=k}$ where $\mathbf{1}_{\Pi(k)=k}$ indicates the k-th number stays the same after the permutation. By linearity,

$$E(X) = E\left(\sum_{k=1}^{n} \mathbf{1}_{\Pi(k)=k}\right) = \sum_{k=1}^{n} E\left(\mathbf{1}_{\Pi(k)=k}\right) = \sum_{k=1}^{n} \frac{1}{n} = 1.$$

Example 3.9 (Buffon's needle)**.** Rule a surface with parallel lines a distance d apart. What is the probability that a randomly dropped needle of length $l \leq d$ crosses a line?

Solution: Consider dropping any (continuous) curve of length l onto the surface. Imagine dividing up the curve into N straight line segments, each of length $\frac{l}{N}$. Let X_i be the indicator for the i-th segment crossing a line. Let X be the total number of times the curve crosses a line. Then,

$$E(X) = E(\sum X_i) = \sum E(X_i) = N \cdot E(X_i).$$

There could be infinitely many segments. It is hard to compute this expectation directly. But here we arrive an important Lemma: the expected number of crossings is proportional to the length of the curve, regardless of the shape of the curve. If we can compute $E(X)$ for some curve, the we can compute $E(X)$ for any length by scaling the value proportional to the length.

Consider a circle of diameter d. The circle always crosses the lines twice for sure. That is, $E(X_{\text{circle}}) = 2$. The length of the circle is πd. Therefore, the value of $E(X)$ for any curve of length l is given by

$$E(X) = \frac{2l}{\pi d}.$$

Now a needle can cross a line either 1 or 0 times. Thus, $E(X) = 1 \cdot P(X = 1) + 0 \cdot P(X = 0)$ is exactly the probability of a needle crossing a line.

Remark. This amazing example can be used to approximate the value of π. Let q be the probability of a needle crossing a line. q can be approximated by large number of simulations. Then $\pi \approx \frac{2l}{qd}$.

3.3 Multiplication and LOTUS

Theorem 3.2. If X and Y are independent, we have

$$E(XY) = E(X)E(Y).$$

In general, if $X_1, \ldots, X_n$ are independent, we have

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

Remark. The multiplication rule will not hold without independence.

Proof. For discrete and independent X, Y,

$$\begin{aligned}
E(XY) &= \sum_x \sum_y xyP(X = x, Y = y) \\
&= \sum_x \sum_y xyP(X = x)P(Y = y) \quad \text{if independent} \\
&= \sum_x xP(X = x) \sum_y yP(Y = y) \\
&= E(X)E(Y).
\end{aligned}$$

□

Remark. This is a sufficient but not necessary condition. $E(XY) = E(X)E(Y)$ does not imply independence. Consider a counter-example,

$$X = \begin{cases} 1 & \text{with prob. } 1/2 \\ 0 & \text{with prob. } 1/2 \end{cases}, \quad Z = \begin{cases} 1 & \text{with prob. } 1/2 \\ -1 & \text{with prob. } 1/2 \end{cases};$$

Then

$$Y = XZ = \begin{cases} -1 & \text{with prob. } 1/4 \\ 0 & \text{with prob. } 1/2 \\ 1 & \text{with prob. } 1/4 \end{cases}.$$

We have $E(X) = 1/2$, $E(Y) = 0$, $E(XY) = 0$. So $E(XY) = E(X)E(Y)$. But clearly X, Y are not independent.

Theorem 3.3 (Law of the unconscious statistician (LOTUS))**.** *Let X be a random variable, and g be a real-valued function of a real variable. If X has a discrete distribution, then*

$$E[g(X)] = \sum_{\textit{all } x} g(x)P(X = x).$$

LOTUS says we can compute the expectation of $g(X)$ without knowing the PMF of $g(X)$.

Proof. The idea is similar to the one we use to prove linearity. Imagine our sample space is composed of "atoms", and $X(\omega)$ maps some atoms to numbers.

The expectation of $g(X)$ can be rewritten as

$$E[g(X)] = \sum_{\omega} g(X(\omega))P(\omega).$$

If we group the atoms that compose the event $X = x$ together,

$$\begin{aligned} E[g(X)] &= \sum_{x} \sum_{\omega : X(\omega)=x} g(X(\omega))P(\omega) \\ &= \sum_{x} g(x) \sum_{\omega : X(\omega)=x} P(\omega) \\ &= \sum_{x} g(x)P(X = x). \end{aligned}$$

□

Example 3.10. Compute $E(X)$ and $E(X^2)$ given the following distribution.

X	0	1	2
X^2	0	1	4
P	1/4	1/2	1/4

Solution: According to the distribution table, we compute the expectations as

$$\begin{aligned} E(X) &= 0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4 = 1; \\ E(X^2) &= 0 \times 1/4 + 1 \times 1/2 + 4 \times 1/4 = 3/2. \end{aligned}$$

Note that $E(X^2) \neq [E(X)]^2$.

Remark. In general, $E[g(X)] \neq g(E(X))$. Linearity implies that if g is a linear function of X, then $E[g(X)] = g(E(X))$. For a nonlinear function g, the relationship between $E[g(X)]$ and $g(E(X))$ is determined case by case. We will get back to this point when we learn Jensen's inequality.

Example 3.11 (St. Petersburg Paradox)**.** Flip a fair coin over and over again until the head lands the first time. You will win 2^k dollars if the head lands in the k-th trial (including the successful trial). What is the expected payoff of this game?

Solution: Let $X = 2^k$. We want to find $E(X)$. The probability of the first head showing up in the k-th trial is $\frac{1}{2^k}$. Therefore,

$$E(X) = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \sum_{k=1}^{\infty} 1 = \infty$$

The expected payoff is infinitely high! This is against most people's intuition. This is because we intuitively think that $E(X) = E(2^k) = 2^{E(k)}$, which is a finite number.

3.4 Median and mode

The mean is called a measure of *central tendency* because it tells us something about the center of a distribution, specifically its center of mass. Other measures of central tendency that are commonly used in statistics are the median and the mode, which we now define.

Definition 3.2. We say that c is a **median** of a random variable X if $P(X \leq c) \geq 1/2$ and $P(X \geq c) \geq 1/2$.

Definition 3.3. For a discrete random variable X, we say that c is a **mode** of X if it maximizes the PMF: $P(X = c) \geq P(X = x)$ for all x. For a continuous random variable X with PDF f, we say that c is a **mode** if it maximizes the PDF: $f(c) \geq f(x)$ for all x.

Intuitively, the median is a value c such that half the mass of the distribution falls on either side of c (or as close to half as possible, for discrete random variables), and the mode is a value that has the greatest mass or density out of all values in the support of X. If the CDF F is a continuous, strictly increasing function, then $F^{-1}(1/2)$ is the median (and is unique).

Remark. A distribution can have multiple medians and multiple modes. Medians have to occur side by side; modes can occur all over the distribution.

Example 3.12. The main reason why the median is sometimes preferred over the mean is that the median is more robust to extreme values. Think about an income distribution. Higher incomes are rare, but their absolute values are high. Thus, the mean income tends be higher than what the mass of the population would earn. But the median is more robust to extreme values and is closer to the earnings of an "average" person. For example, the mean of China's income is ¥$2,561$ monthly in 2019; the median is only ¥$2,210$.

Income (monthly, yuan)	<1k	1-2k	2-5k	5-10k	10-20k	>20k
Population (million)	550	420	360	63	7.8	0.7

Table 3.2: China monthly income per capita. Source: NBS 2019.

Theorem 3.4. *Let X be an random variable with mean μ , and let m be a median of X.*

- *The value of c that minimizes the mean squared error $E\left(X-c\right)^2$ is $c=\mu$.*
- *A value of c that minimizes the mean absolute error $E\left|X-c\right|$ is $c=m$.*

3.5 Variance and covariance

Variance and standard deviation

Expectation is the most commonly used summary of a distribution, as it indicates where values are likely centered. However, it provides limited insight into the distribution's overall shape. For example, two random variables might have the same mean, yet one could have values spread far from the mean while the other has values tightly clustered around it. Variance, on the other hand, describes how far values in a distribution typically deviate from the mean, offering a measure of the distribution's dispersion.

Definition 3.4. The **variance** of a random variable X is defined as

$$Var(X) = E\left(X - EX\right)^2 .$$

The **standard deviation** of X is defined as

$$SD(X) = \sqrt{Var(X)}.$$

We often denote standard deviation by the greek letter $\sigma \equiv SD(X)$, and variance by σ^2.

Variance measures how far X typically deviates from its mean, but instead of averaging the differences, we average the squared differences to ensure both positive and negative deviations contribute. The expected deviation, $\mathbb{E}(X - \mathbb{E}(X))$, is always zero, so squaring avoids this cancellation. Since variance is in squared units, we take the square root to get the standard deviation, restoring the original units.

Why take squares? Sometimes we also use $\mathbb{E}(|X - \mathbb{E}(X)|)$ instead. But it is less common because the absolute value function isn't differentiable. Besides, squaring connects to geometric concepts like the distance formula and Pythagorean theorem, which have useful statistical meanings.

Theorem 3.5. *For any random variable X,*

$$Var(X) = E(X^2) - (EX)^2.$$

Proof. Let $\mu = E(X)$. By definition,

$$\begin{aligned} Var(X) &= E(X-\mu)^2 = E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2. \end{aligned}$$

□

Example 3.13. Find the variance for $X \sim \text{Bern}(p)$.

$$Var(X) = E(X^2) - E^2(X) = p - p^2 = p(1-p).$$

Theorem 3.6. *Variance has the following properties:*

- $Var(X) \geq 0$
- $Var(X + c) = Var(X)$
- $Var(cX) = c^2 Var(X)$
- *If X, Y are independent,* $Var(X + Y) = Var(X) + Var(Y)$.
- *If $X_1, X_2, \ldots, X_n$ are independent,* $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$.

Example 3.14. Find the variance for $X \sim \text{Bin}(n, p)$. $X = X_1 + \cdots + X_n$ where X_i are *i.i.d* Bernoulli distributions

$$Var(X) \overset{iid}{=} \sum_{i=1}^{n} Var(X_i) = np(1-p).$$

Covariance and correlation

For more than one random variable, it is also of interest to know the relationship between them. Are they dependent? How strong is the dependence? Covariance and correlation are intended to measure that dependence. But they only capture a particular type of dependence, namely linear dependence.

Definition 3.5. The **covariance** between random variables X and Y is defined as

$$Cov(X, Y) = E[(X - EX)(Y - EY)].$$

The covariance between X and Y reflects how much X and Y simultaneously deviate from their respective means. If $X > EX, Y > EY$ or $X < EX, Y < EY$ simultanenously, then $Cov(X, Y)$ tends be positive. Conversely, if $X > EX$ is pair with $Y < EY$ (or $X < EX$ paired with $Y > EY$), then $Cov(X, Y)$ tends to be negative.

Theorem 3.7. *For any random variables X and Y,*

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

Proof. Let $\mu_X = E(X)$ and $\mu_Y = E(Y)$. By definition,

$$\begin{aligned} Cov(X, Y) &= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\ &= E(XY) - E(X)E(Y). \end{aligned}$$

□

Theorem 3.8. *If X, Y are independent, they are uncorrelated. But the converse is false.*

Proof. $Cov(X,Y) = E(XY) - E(X)E(Y)$. Independence implies $E(XY) = E(X)E(Y)$. Thus, $Cov(X,Y) = 0$. But $Cov(X,Y) = 0$ does not necessarily imply independence. Consider the following counter example. Let X be a random variable that takes three values -1, 0, 1 with equal probability. And $Y = X^2$. X and Y are clearly dependent. But they their correlation is 0. Since $E(X) = 0$, $E(Y) = 2/3$, $E(XY) = E(X^3) = 0$, $Cov(X,Y) = 0$. □

Remark. Covariances and correlations provide measures of the extend to which two random variables are linearly related. If we plot the values of X and Y in the xy-plane, if the points form a straight line, that would signal a strong positive (if positive slope) or negative (if negative slope) correlation. It is possible that the correlation is 0 if X and Y are dependent but the relationship is nonlinear.

Theorem 3.9. *Covariance has the following properties:*

- $Cov(X,X) = Var(X)$
- $Cov(X,Y) = Cov(Y,X)$
- $Cov(cX,Y) = Cov(X,cY) = c\,[Cov(X,Y)]$
- $Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)$
- $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
- $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i,X_j)$

Proof. We only prove the variance-covariance property:

$$\begin{aligned}
Var(X+Y) &= E[(X+Y-\mu_X-\mu_Y)^2] \\
&= E[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)] \\
&= Var(X) + Var(Y) + 2Cov(X,Y).
\end{aligned}$$

□

Exercise 3.1. Find $Cov(X+Y,Z+W)$ and $Var(X-Y)$.

While $\mathrm{Cov}(X,Y)$ quantifies how X and Y vary together, its magnitude also depends on the absolute scales of X and Y (multiply X by a constant c, the covariance will be different). To establish a measure of association between X

and Y that is unaffected by arbitrary changes in the scales of either variable, we introduce a "standardized covariance" called correlation.

Definition 3.6. The **correlation** between random variables X and Y is defined as

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

We also denote correlation by $\rho \equiv Corr(X,Y)$.

Unlike covariance, scaling X or Y has no effect on the correlation. We can verify this:

$$Corr(cX,Y) = \frac{Cov(cX,Y)}{\sqrt{Var(cX)Var(Y)}} = \frac{cCov(X,Y)}{c\sqrt{Var(X)Var(Y)}} = Corr(X,Y).$$

Theorem 3.10. *For any random variable X and Y,*

$$-1 \leq Corr(X,Y) \leq 1.$$

Proof. Without loss of generality, assume X,Y both have variance 1, since scaling does not change the correlation. Let $\rho = Corr(X,Y) = Cov(X,Y)$. Then

$$\begin{aligned} Var(X+Y) &= Var(X) + Var(Y) + 2Cov(X,Y) = 2 + 2\rho \geq 0, \\ Var(X-Y) &= Var(X) + Var(Y) - 2Cov(X,Y) = 2 - 2\rho \geq 0. \end{aligned}$$

Thus $-1 \leq \rho \leq 1$. □

It is said that X and Y are **positively correlated** if $Corr(X,Y) > 0$, that X and Y are **negatively correlated** if $Corr(X,Y) < 0$, and that X and Y are **uncorrelated** if $Corr(X,Y) = 0$.

Theorem 3.11. *Suppose that X is a random variable and $Y = aX + b$ for some constants a, b, where $a \neq 0$. If $a > 0$, then $\rho_{XY} = 1$. If $a < 0$, then $\rho_{XY} = -1$.*

Proof. If $Y = aX+b$, then $E(Y) = aE(X)+b$. Thus, $Y - E(Y) = a(X - E(X))$. Therefore,

$$Cov(X,Y) = aE[(X - EX)^2] = aVar(X).$$

Since $Var(Y) = a^2Var(X)$, $\rho_{XY} = \frac{a}{|a|}$. The theorem thus follows. □

Example 3.15. Toss two coins. Let X be the number of Heads, and Y be the number of Tails. Find the covariance and correlation between X and Y.

Solution: Note that X and Y are counterparts to each other, $Y = 2 - X$. So we expect the correlation be negative. The expectation of X and Y are the same: $EX = EY = 1$. So we have $X - EX = -1, 0, 1$ and $Y - EY = 1, 0, -1$. The corresponding probabilities are 1/4,1/2,1/4 respectively. Therefore,

$$Cov(X, Y) = (-1) \times 1 \times 1/4 + 1 \times (-1) \times 1/4 = -1/2.$$

Since $Var(X) = Var(Y) = 1/2$, the correlation is

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{-1/2}{\sqrt{1/2 \times 1/2}} = -1.$$

Example 3.16. Let $X \sim HGeom(w, b, n)$. Find $Var(X)$.

Solution: Interpret X as the number of white balls in a sample of size n from an box with w white and b black balls. We can represent X as the sum of indicator variables, $X = I_1 + \cdots + I_n$, where I_j is the indicator of the j-th ball in the sample being white. Each I_j has mean $p = w/(w+b)$ and variance $p(1-p)$, but because the I_j are dependent, we cannot simply add their variances. Instead,

$$\begin{aligned} Var(X) &= Var\left(\sum_{j=1}^{n} I_j\right) \\ &= Var(I_1) + \cdots + Var(I_n) + 2\sum_{i<j} Cov(I_i, I_j) \\ &= np(1-p) + 2\binom{n}{2} Cov(I_i, I_j) \end{aligned}$$

In the last step, because of symmetry, for every pair i and j, $Cov(I_i, I_j)$ are the same.

$$\begin{aligned}
Cov(I_i, I_j) &= E(I_i I_j) - E(I_i)E(I_j) \\
&= P(i \text{ and } j \text{ both white}) - P(i \text{ is white})P(j \text{ is white}) \\
&= \frac{w}{w+b} \cdot \frac{w-1}{w+b-1} - p^2 \\
&= p\frac{Np-1}{N-1} - p^2 \\
&= \frac{p(p-1)}{N-1}
\end{aligned}$$

where $N = w + b$. Plugging this into the above formula and simplifying, we eventually obtain

$$Var(X) = np(1-p) + n(n-1)\frac{p(p-1)}{N-1} = \frac{N-n}{N-1}np(1-p).$$

This differs from the Binomial variance of $np(1-p)$ by a factor of $\frac{N-n}{N-1}$. This discrepancy arises because the Hypergeometric story involves sampling without replacement. As $N \to \infty$, it becomes extremely unlikely that we would draw the same ball more than once, so sampling with or without replacement essentially become the same.

Example 3.17 (PG exam)**.** Put k balls into n boxes. Let X be the number of empty boxes. Find $E(X)$ and $Var(X)$.

Solution: Define an indicator variable

$$I_j = \begin{cases} 1 & j\text{-th box is empty} \\ 0 & \text{otherwise} \end{cases}$$

Then $X = \sum_{j=1}^{n} I_j$. Unconditionally, the probability of one box being empty is $\left(\frac{n-1}{n}\right)^k$. Therefore,

$$E(I_j) = P(j\text{-th box is empty}) = \left(\frac{n-1}{n}\right)^k$$

for $j = 1, 2, \ldots, n$. It follows that

$$E(X) = \sum_{j=1}^{n} I_j = nE(I_j) = n\left(\frac{n-1}{n}\right)^k.$$

To compute the variance,

$$\begin{aligned}Var(X) = Var(I_1 + \cdots + I_n) &= \sum_{j=1}^{n} Var(I_j) + 2\sum_{i<j} Cov(I_i, I_j)\\ &= nVar(I_j) + 2\binom{n}{2} Cov(I_i, I_j),\end{aligned}$$

since by symmetry, $Var(I_j)$ is the same for all j and $Cov(I_i, I_j)$ is the same for all $i \neq j$. It suffices to compute $Var(I_j)$ and $Cov(I_i, I_j)$ for any j and $i \neq j$. Since I_j only takes number 0 and 1,

$$E(I_j^2) = \left(\frac{n-1}{n}\right)^k,$$

$$Var(I_j) = E(I_j^2) - (E(I_j))^2 = \left(\frac{n-1}{n}\right)^k - \left(\frac{n-1}{n}\right)^{2k}.$$

For the covariance term,

$$E(I_i I_j) = P(i, j \text{ are both empty}) = \left(\frac{n-2}{n}\right)^k,$$

$$Cov(I_i, I_j) = E(I_i I_j) - E(I_i)E(I_j) = \left(\frac{n-2}{n}\right)^k - \left(\frac{n-1}{n}\right)^{2k}.$$

Therefore,

$$Var(X) = n\left[\left(\frac{n-1}{n}\right)^k - \left(\frac{n-1}{n}\right)^{2k}\right] + 2\binom{n}{2}\left[\left(\frac{n-2}{n}\right)^k - \left(\frac{n-1}{n}\right)^{2k}\right].$$

3.6 Moments and MGF

Definition 3.7. Let X be a random variable with mean μ and variance σ^2 . For any positive integer n, the n-th **moment** of X is $E(X^n)$, the n-th **central moment** is $E(X-\mu)^n$, and the n-th **standardized moment** is $E\left(\frac{X-\mu}{\sigma}\right)^n$.

In accordance with this terminology, $E(X)$ is the first moment of X, $Var(X)$ is the second central moment of X. It is natural to ask if there are higher order moments. The answer is yes.

Definition 3.8. Let X be a random variable with mean μ, standard deviation σ, and finite third moment. The **skewness** of X is defined as

$$\text{Skew}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right].$$

Definition 3.9. The **Kurtosis** of X is defined as

$$\text{Kurt}(X) = \left[\left(\frac{X-\mu}{\sigma}\right)^4\right].$$

Skewness is the measure of the lopsidedness of the distribution; any symmetric distribution will have a third central moment, if defined, of zero. A distribution that is skewed to the left (the tail of the distribution is longer on the left) will have a negative skewness. A distribution that is skewed to the right (the tail of the distribution is longer on the right), will have a positive skewness.

Kurtosis is a measure of the heaviness of the tail of the distribution. If a distribution has heavy tails, the kurtosis will be high; conversely, light-tailed distributions have low kurtosis.

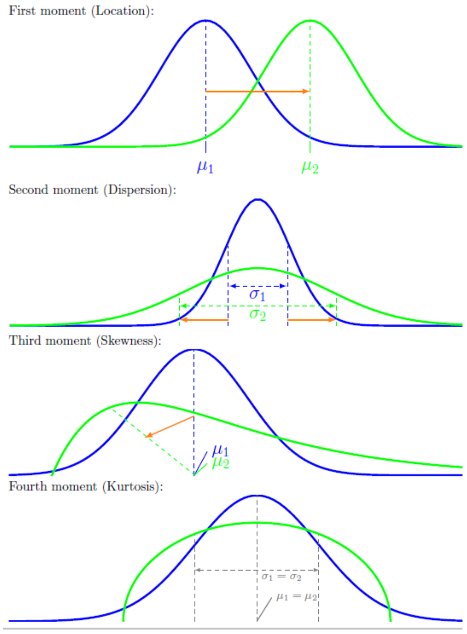

Figure 3.1: Moments and the shape of a distribution

We see that moments give information about the shape of a distribution. Different orders of moments captures different aspects of the distribution. In fact, if we know all the moments (moments of infinitely high order), we can exactly pin down the distribution.

Theorem 3.12. *For a distribution of mass or probability on a bounded interval, the collection of all the moments (of all orders, from* 0 *to* ∞*) uniquely determines the distribution.*

So there are two ways of fully characterize a distribution:

1. Listing all the possible values along with their associated probabilities;
2. Giving all the moments of the distribution.

It is somewhat like the analogous Taylor theorem in the probability theory. We can represent any distribution by a sequence of higher order "polynomials": $E(X), E(X^2), E(X^3), \ldots$

Definition 3.10. Let X be a random variable. For each real number t, define the **moment generating function** (MGF) as

$$M_X(t) = E\left(e^{tX}\right).$$

To see why it is "generating" moments, take the Taylor expansion of the exponential function:

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots$$

Hence,

$$M_X(t) = E\left(e^{tX}\right) = 1 + E(X)t + E(X^2)\frac{t^2}{2!} + \cdots$$

A natural question at this point is: What is the interpretation of t? The answer is that t has no interpretation in particular; it's just a bookkeeping device that we introduce in order to *encode* the sequence of moments in a differentiable function.

Theorem 3.13. *Let* $M_X(t)$ *be the MGF of* X*. Then the* n*-th moment of* X *is given by* $E(X^n) = M_X^{(n)}(0)$*, where* $M_X^{(n)}$ *denotes the* n*-th derivative of the MGF.*

Theorem 3.14. *The MGF of a random variable determines its distribution: if two random variables have the same MGF, they must have the same distribution.*

Theorem 3.15. *If X and Y are independent, then the MGF of $X + Y$ is the product of the individual MGFs:*

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Example 3.18. For $X \sim Bern(p)$, e^{tX} takes on the value e^t with probability p and the value 1 with probability q, so $M(t) = E\left(e^{tX}\right) = pe^t + q$. Since this is finite for all values of t, the MGF is defined on the entire real line.

Example 3.19. The MGF of a $Bin(n, p)$ random variable is $M(t) = (pe^t + q)^n$, since it is the product of n independent Bernoulli MGFs.

3.7 Inequalities*

This section introduces some of the most popular inequality in statistics and general mathematics. Interestingly, our probability theories can shed light on these inequalities that are otherwise hard to explain. We don't show formal proofs here, but just point out how these inequalities can be useful in statistics.

Theorem 3.16 (Cauchy-Schwarz inequality)**.**

$$\left|\sum x_i y_i\right| \leq \sqrt{\sum x_i^2}\sqrt{\sum y_i^2}$$

Proof. If X, Y have zero means, their correlation can be written as

$$\rho_{XY} = \frac{E(XY)}{\sqrt{E(X^2)E(Y^2)}}$$

Since $|\rho_{XY}| \leq 1$, we always have

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}.$$

Consider $\{x_i\}$ and $\{y_i\}$ as realizations of X and Y with equal probabilities, such that $E(X) = \frac{1}{n}\sum x_i$. The original inequality is thus proved. □

Theorem 3.17 (Jensen's inequality)**.** *For a convex function f, we have*

$$\frac{1}{n}\sum f(x_i) \geq f\left(\frac{1}{n}\sum x_i\right);$$

If f is concave, then

$$\frac{1}{n}\sum f(x_i) \leq f\left(\frac{1}{n}\sum x_i\right).$$

Proof. This is not a proof, but a special case that helps to understand Jensen's inequality. Since

$$Var(X) = E(X^2) - (E(X))^2 \geq 0$$

We have

$$E(X^2) \geq (E(X))^2.$$

Note that $f(X) = X^2$ is a convex function, and $E(*) = \frac{1}{n}\sum *$, we have shown the first inequality. The concave case is the opposite.

In general, if g is a convex function, then $E(g(X)) \geq g(E(X))$. If g is a concave function, then $E(g(X)) \leq g(E(X))$. In both cases, the only way that equality can hold is if there are constants a and b such that $g(X) = a + bX$ with probability 1. □

Theorem 3.18 (Markov inequality)**.** *Let X be a random variable, then*

$$P(|X| \geq a) \leq \frac{E|X|}{a}$$

That is, the probability of $|X|$ deviating from its mean by a multiple of a must be less than $1/a$.

Proof. Define a random variable

$$I_{|X|\geq a} = \begin{cases} 1 & \text{if } |X| \geq a \\ 0 & \text{if } |X| < a \end{cases}$$

Note that $P(|X| \geq a) = E(I_{|X|\geq a})$. It always holds that

$$a \cdot I_{|X|\geq a} \leq |X|$$

Therefore,

$$E\left[a \cdot I_{|X|\geq a}\right] \leq E|X|$$

Hence,

$$P(|X| \geq a) \leq \frac{E|X|}{a}.$$

□

For an intuitive interpretation, let X be the income of a randomly selected individual from a population. Taking $a = 2E(X)$, Markov's inequality says that $P(X \geq 2E(X)) \leq 1/2$, i.e., it is impossible for more than half the population to make at least twice the average income. This is clearly true, since if over half the population were earning at least twice the average income, the average income would be higher. Similarly, $P(X \geq 3E(X)) \leq 1/3$: you can't have more than 1/3 of the population making at least three times the average income, since those people would already drive the average above what it is.

Theorem 3.19 (Chebyshev inequality)**.** *Let X be a random variable with mean μ and standard deviation σ, then*

$$P\left(|X-\mu| > c\sigma\right) \leq \frac{1}{c^2}$$

That is, the probability of X deviating from its mean by a times the standard deviation must be less than $1/a^2$.

Proof. We first show

$$P(|X-\mu| > a) \leq \frac{\sigma^2}{a^2}$$

This is true by taking squares and applying the Markov inequality,

$$P(|X-\mu| > a) = P((X-\mu)^2 > a^2) \leq \frac{E(X-\mu)^2}{a^2} = \frac{\sigma^2}{a^2}.$$

Substitute $c\sigma$ for a, we have the original inequality. □

This gives us an upper bound on the probability of a random variable being more than c standard deviations away from its mean, e.g., there can't be more than a 25% chance of being 2 or more standard deviations from the mean. Given the mean and standard deviation of a random variable X, we know that $\mu \pm 2\sigma$ captures 75% of its possible values; $\mu \pm 3\sigma$ captures 90% of the possible values.