Introduction to Machine Learning

A quick refresher course in probability theory

Third lecture, 19.01.2022

Phuc Loi Luu, PhD
p.luu@garvan.org.au
luu.p.loi@googlemail.com

Roadmap for today

- *Expectation, variance, covariance, correlation, quantiles*
- *Independence, conditional probability, conditional independence*

Binomial distribution

The Binomial distribution describes the outcome from a fixed number, n, of Bernoulli trials. For example:

- X is the number of boys in a 3-child family: $n = 3$ *trials (children);* $p = \mathbb{P}(\textit{Boy}) = 0.5$ *for each child.*
- X is the number of 6's obtained in 10 rolls of a die: $n = 10$ *trials (die rolls);* $p = \mathbb{P}(\textit{Get a 6}) = 1/6$ *for each roll.*

Definition: Let X be the number of successes obtained in n independent Bernoulli trials, each of which has probability of success p.

Then X has the **Binomial distribution with parameters n and p.**

We write $X \sim \textit{Binomial}(n, p)$, *or* $X \sim \textit{Bin}(n, p)$.

The Binomial distribution counts the number of **successes** in a **fixed number** of Bernoulli trials.

Binomial distribution

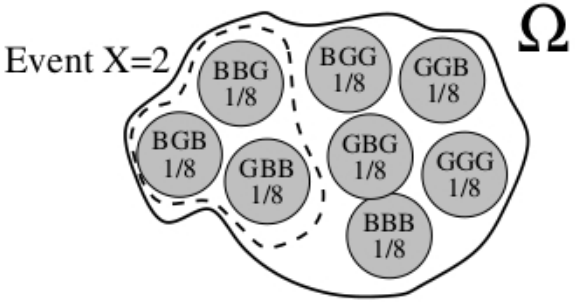
If $X \sim \text{Binomial}(n, p)$, then $X = x$ if there are x successes in the n trials. We don't care what order the successes occur in — in other words, we don't care *which* of the trials are successes and which are failures. However, we do have to bear in mind all the different orderings when we calculate the probabilities of the distribution.

Take the example of $X =$ number of boys in a 3-child family, so $X \sim \textit{Binomial}(n = 3, p = 0.5)$.

If we want to calculate $\mathbb{P}(X = 2)$, we have to take account of all the different ways that we can achieve $X = 2$:

$$\mathbb{P}(X = 2) = \mathbb{P}(BBG) + \mathbb{P}(BGB) + \mathbb{P}(GBB)\,.$$

In this case, there are 3 ways of getting the outcome we are interested in: 2 boys and 1 girl. How would we calculate the number of ways in general?



There are 3 trials (children), and we need to choose 2 of them to be boys. The number of ways of choosing 2 trials from 3 is:

$$^3C_2 = \binom{3}{2} = \frac{3!}{(3-2)!\,2!} = \frac{3 \times 2 \times 1}{1 \times (2 \times 1)} = 3.$$

Binomial distribution, examples

Question: How many ways are there of achieving 6 boys in a 10-child family?

Answer:

$$^{10}C_6 = \binom{10}{6} = \frac{10!}{(10-6)!\,6!} = 210 \qquad \text{— use calculator button } {}^{n}C_r\,.$$

Question: How many ways are there of achieving x successes in n trials?

Answer:

$$^{n}C_x = \binom{n}{x} = \frac{n!}{(n-x)!\,x!}\,.$$

Question: If each trial has probability p of being a success, what is the probability of getting the precise outcome $SFFSF$ from $n = 5$ trials?

Answer: $p \times (1-p) \times (1-p) \times p \times (1-p) = p^2(1-p)^3$. *This will be the same whatever order the successes and failures are in. But it only gives the probability for* <u>one</u> *ordering.*

Question: What is the probability of <u>one</u> ordering that contains x successes and $n - x$ failures?

Answer: $p^x(1-p)^{n-x}$.

Question: So what is the overall probability of achieving x successes in n trials: $\mathbb{P}(X = x)$ when $X \sim \text{Binomial}(n, p)$?

Answer: *(Number of orderings)* $\times$ *(probability of each ordering)* $= \binom{n}{x} p^x (1-p)^{n-x}$.

The probability function for the Binomial distribution

Let $X \sim \text{Binomial}(n, p)$. The probability function for X is:

$$f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for} \quad x = 0, 1, \ldots, n\,.$$

Note:

1. Importantly, $f_X(x) = 0$ *if* x *is not one of the values* $0, 1, \ldots, n$.
   The correct way to write the range of values is $x = 0, \ldots, n$.
   - — Writing $x \in [0, n]$ is **wrong,** because this includes decimals like 0.4.
   - — Writing $x = 0, 1, \ldots$ is **wrong,** because the range of values must stop at n: you can't have more than n successes in n trials.
2. $f_X(x)$ means, *'the probability function belonging to the r.v. I've named* X*'.*
   Use a capital X in the subscript and a lower-case x as the argument.

Geometric distribution

Like the Binomial distribution, the Geometric distribution is defined in terms of a sequence of Bernoulli trials.

- The Binomial distribution counts the *number of successes out of a fixed number of trials.*
- The Geometric distribution counts the *number of trials before the first success occurs.*

This means that the Geometric distribution counts the *number of failures before the first success.*

If every trial has probability p of success, we write: $X \sim$ *Geometric*(p).

Examples: 1) X =number of boys before the first girl in a family:
$X \sim$ *Geometric*$(p = 0.5)$.

2) Fish jumping up a waterfall. On every jump the fish has probability p of reaching the top.
Let X be *the number of failed jumps before the fish succeeds.*
Then $X \sim$ *Geometric*(p).

Properties of the Geometric distribution

i) Description

$X \sim \text{Geometric}(p)$ if X is the *number of failures before the first success in a series of Bernoulli trials with* $\mathbb{P}(\textit{success}) = p$.

ii) Probability function

For $X \sim \text{Geometric}(p)$,

$$f_X(x) = \mathbb{P}(X = x) = (1-p)^x p \quad \textit{for} \quad x = 0, 1, 2, \ldots$$

Explanation: $\mathbb{P}(X = x) = \underbrace{(1-p)^x}_{\text{need } x \text{ failures}} \times \underbrace{p}_{\text{final trial must be a success}}$

Difference between Geometric and Binomial: For the Geometric distribution, the trials must always occur in the order $\underbrace{FF\ldots F}_{x \text{ failures}}S$.

For the Binomial distribution, failures and successes can occur in any order: e.g. $FF\ldots FS$, $FSF\ldots F$, $SF\ldots F$, etc.

This is why the Geometric distribution has probability function

$$\mathbb{P}(x \text{ failures, 1 success}) = (1-p)^x p,$$

while the Binomial distribution has probability function

$$\mathbb{P}(x \text{ failures, 1 success}) = \binom{x+1}{x}(1-p)^x p.$$

iii) Mean and variance

For $X \sim \text{Geometric}(p)$,

$$\mathbb{E}(X) = \frac{1-p}{p} = \frac{q}{p}$$

$$\textit{Var}(X) = \frac{1-p}{p^2} = \frac{q}{p^2}$$

Negative Binomial distribution

The Negative Binomial distribution is a generalised form of the Geometric distribution:

- the Geometric distribution counts the number of *failures before the first success;*
- the Negative Binomial distribution counts the number of *failures before the k'th success.*

If every trial has probability p of success, we write: $X \sim \textbf{\textit{NegBin}}(k, p)$.

Examples: 1) X =number of boys before the second girl in a family:
$X \sim \textbf{\textit{NegBin}}(k = 2,\ p = 0.5)$.

2) Tom needs to pass 24 papers to complete his degree. He passes each paper with probability p, independently of all other papers. Let X be *the number of papers Tom fails in his degree.*

Then $X \sim \textbf{\textit{NegBin}}(24, p)$.

Properties of the Negative Binomial distribution

i) Description

$X \sim \text{NegBin}(k,\ p)$ if X is the *number of failures before the k'th success in a series of Bernoulli trials with $\mathbb{P}(\text{success}) = p$.*

ii) Probability function

For $X \sim \text{NegBin}(k,\ p)$,

$$f_X(x) = \mathbb{P}(X = x) = \binom{k+x-1}{x} p^k (1-p)^x \quad \text{for} \quad x = 0, 1, 2, \ldots$$

Properties of the Negative Binomial distribution

Explanation:

- For $X = x$, we need x failures and k successes.
- The trials stop when we reach the k'th success, so the last trial must be *a success.*
- This leaves x failures and $k-1$ successes to occur in *any order:* a total of $k-1+x$ trials.

For example, if $x = 3$ failures and $k = 2$ successes, we could have:

$$FFFS\underline{S} \quad FFSF\underline{S} \quad FSFF\underline{S} \quad SFFF\underline{S}$$

So:

$$\mathbb{P}(X = x) = \underbrace{\binom{k+x-1}{x}}_{\substack{(k-1)\ \textit{successes and}\ x\ \textit{failures} \\ \textit{out of}\ (k-1+x)\ \textit{trials.}}} \times \overbrace{p^k}^{k\ \textit{successes}} \times \underbrace{(1-p)^x}_{x\ \textit{failures}}$$

iii) Mean and variance

For $X \sim \text{NegBin}(k,\, p)$,

$$\mathbb{E}(X) = \frac{k(1-p)}{p} = \frac{kq}{p}$$

$$\textit{Var}(X) = \frac{k(1-p)}{p^2} = \frac{kq}{p^2}$$

These results can be proved from the fact that the Negative Binomial distribution is obtained as the sum of k independent Geometric random variables:

$$\begin{aligned} X &= Y_1 + \ldots + Y_k, \quad \text{where each } Y_i \sim \text{Geometric}(p), \quad Y_i \text{ indept,} \\ \Rightarrow \quad \mathbb{E}(X) &= k\mathbb{E}(Y_i) = \frac{kq}{p}, \\ \text{Var}(X) &= k\text{Var}(Y_i) = \frac{kq}{p^2}. \end{aligned}$$

iv) Sum of independent Negative Binomial random variables

If X and Y are *independent,*
and $X \sim \text{NegBin}(k,\, p)$, $Y \sim \text{NegBin}(m,\, p)$, with the same value of p, then

$$X + Y \sim \textit{NegBin}(k+m, p).$$

Expectation and variance of a random variable

- The ***expectation*** of a random variable is the value it takes ***on average.***
- The ***variance*** of a random variable measures how much the random variable ***varies about its average.***

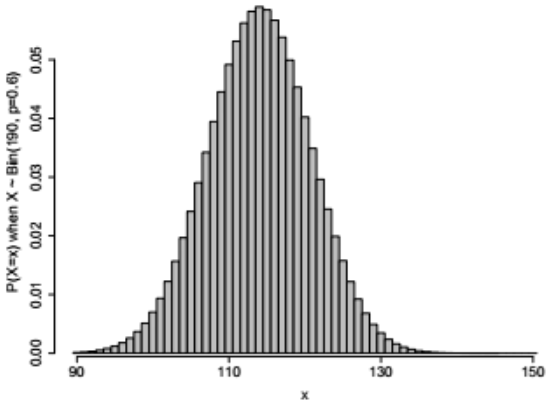
Expectation

Given a random variable X that measures something, we often want to know *what is the* **average** *value of* X*?*

For example, here are 30 random observations taken from the distribution $X \sim \text{Binomial}(n = 190, p = 0.6)$:

***R* command:** `rbinom(30, 190, 0.6)`

```
116 116 117 122 111 112 114 120 112 102
125 116  97 105 108 117 118 111 116 121
107 113 120 114 114 124 116 118 119 120
```



The average, or *mean*, of the ***first ten*** values is:

$$\frac{116 + 116 + \ldots + 112 + 102}{10} = 114.2.$$

Expectation of a random variable

The mean of the ***first twenty*** values is:

$$\frac{116 + 116 + \ldots + 116 + 121}{20} = 113.8.$$

The mean of the ***first thirty*** values is:

$$\frac{116 + 116 + \ldots + 119 + 120}{30} = 114.7.$$

The answers all seem to be close to *114*. What would happen if we took the average of hundreds of values?

100 values from Binomial(190, 0.6):

R command: `mean(rbinom(100, 190, 0.6))`
Result: `114.86`

Note: You will get a different result every time you run this command.

Expectation of a random variable

1000 values from Binomial(190, 0.6):

R command: `mean(rbinom(1000, 190, 0.6))`
Result: `114.02`

1 million values from Binomial(190, 0.6):

R command: `mean(rbinom(1000000, 190, 0.6))`
Result: `114.0001`

The average seems to be *converging to the value 114.*

The larger the sample size, *the closer the average seems to get to 114.*

If we kept going for larger and larger sample sizes, we would keep getting answers closer and closer to 114. This is because *114 is the DISTRIBUTION MEAN: the mean value that we would get if we were able to draw an infinite sample from the Binomial(190, 0.6) distribution.*

This distribution mean is called the *expectation, or expected value, of the Binomial(190, 0.6) distribution.*

It is a *FIXED property of the Binomial(190, 0.6) distribution.* This means it is a *fixed constant: there is nothing random about it.*

Expectation of a random variable

Definition: The **expected value**, also called the **expectation** or **mean**, of a discrete random variable X, *can be written as either* $\mathbb{E}(X)$, *or E(X), or* μ_X, *and is given by*

$$\mu_X = \mathbb{E}(X) = \sum_x x f_X(x) = \sum_x x \mathbb{P}(X = x)\,.$$

The expected value is a measure of the centre, or average, of the set of values that X can take, weighted according to the probability of each value.

If we took a very large sample of random numbers from the distribution of X, *their average would be approximately equal to* μ_X.

Expectation of a random variable

Ex1: Let X be the value that comes up with you roll a fair 6-sided die with the uniform distribution. Calculate expectation of X.

x	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6

$$E[X] = \sum_{X=1}^{X=6} xP(X=x) = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = \frac{7}{2}$$

Ex2: Let X be the value that comes up a 1 child family with 1 boy. Calculate expectation of X.

x	0	1
P(X=x)	1/2	1/2

Expectation of a random variable

Example: Let $X \sim \text{Binomial}(n = 190, p = 0.6)$. What is $\mathbb{E}(X)$?

$$\begin{aligned}\mathbb{E}(X) &= \sum_x x\mathbb{P}(X = x) \\ &= \sum_{x=0}^{190} x\binom{190}{x}(0.6)^x(0.4)^{190-x}.\end{aligned}$$

Although it is not obvious, the answer to this sum is $n \times p = 190 \times 0.6 = 114$. We will see why in Section 2.10.

Explanation of the formula for expectation

We will move away from the Binomial distribution for a moment, and use a simpler example.

Let the random variable X be defined as $X = \begin{cases} 1 & \text{with probability } 0.9, \\ -1 & \text{with probability } 0.1. \end{cases}$

X takes only the values 1 and -1. What is the 'average' value of X?

Expectation of a random variable

Using $\frac{1+(-1)}{2} = 0$ *would not be useful, because it ignores the fact that usually* $X = 1$, *and only occasionally is* $X = -1$.

Instead, think of observing X many times, say 100 times.

Roughly *90* of these 100 times will have $X = 1$.
Roughly *10* of these 100 times will have $X = -1$

The average of the 100 values will be roughly

$$\frac{90 \times 1 + 10 \times (-1)}{100},$$

$$= 0.9 \times 1 + 0.1 \times (-1)$$

$$(= 0.8.)$$

We could repeat this for any sample size.

Expectation of a random variable

As the sample gets large, the average of the sample will get ever closer to

$$0.9 \times 1 + 0.1 \times (-1).$$

This is why the distribution mean is given by

$$\mathbb{E}(X) = \mathbb{P}(X = 1) \times 1 + \mathbb{P}(X = -1) \times (-1),$$

or in general,

$$\mathbb{E}(X) = \sum_x \mathbb{P}(X = x) \times x.$$

> $\mathbb{E}(X)$ *is a fixed constant giving the average value we would get from a large sample of* X.

Linear property of expectation

Expectation is a ***linear*** operator:

Theorem 2.7: *Let a and b be constants. Then*

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

Proof:

Immediate from the definition of expectation.

$$\begin{aligned}
\mathbb{E}(aX + b) &= \sum_x (ax + b) f_X(x) \\
&= a \sum_x x f_X(x) + b \sum_x f_X(x) \\
&= a\,\mathbb{E}(X) + b \times 1. \qquad \square
\end{aligned}$$

Example 1: *finding expectation from the probability function*

Example 1: Let $X \sim \text{Binomial}(3, 0.2)$. Write down the probability function of X and find $\mathbb{E}(X)$.

We have:

$$\mathbb{P}(X = x) = \binom{3}{x}(0.2)^x(0.8)^{3-x} \text{ for } x = 0, 1, 2, 3.$$

x	*0*	*1*	*2*	*3*
$f_X(x) = \mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

Then

$$\begin{aligned}\mathbb{E}(X) = \sum_{x=0}^{3} x f_X(x) &= 0 \times 0.512 + 1 \times 0.384 + 2 \times 0.096 + 3 \times 0.008 \\ &= 0.6.\end{aligned}$$

Note: We have: $\mathbb{E}(X) = 0.6 = 3 \times 0.2$ *for* $X \sim$ *Binomial*$(3, 0.2)$.
We will prove in Section 2.10 that whenever $X \sim \text{Binomial}(n, p)$, then $\mathbb{E}(X) = np$.

Example 2: finding expectation from the probability function

Example 2: Let Y be Bernoulli(p) (Section 1.2). That is,

$$Y = \begin{cases} 1 & \textit{with probability } p, \\ 0 & \textit{with probability } 1-p. \end{cases}$$

Find $\mathbb{E}(Y)$.

y	*0*	*1*
$\mathbb{P}(Y=y)$	$1-p$	p

$$\mathbb{E}(Y) = 0 \times (1-p) + 1 \times p = p.$$

Expectation of a sum of random variables: E(X + Y)

For ANY random variables $X_1, X_2, \ldots, X_n$*,*

$$\mathbb{E}\,(X_1 + X_2 + \ldots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n).$$

In particular, $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ *for ANY* X *and* Y*.*

This result holds for ***any*** random variables $X_1, \ldots, X_n$. *It does NOT require* $X_1, \ldots, X_n$ *to be independent.*

We can summarize this important result by saying:

The expectation of a sum
is the sum of the expectations – ALWAYS.

The proof requires multivariate methods, to be studied in later courses.

Note: We can combine the result above with the linear property of expectation. For any constants $a_1, \ldots, a_n$, we have:

$$\mathbb{E}\,(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \ldots + a_n\mathbb{E}(X_n).$$

Expectation of a product of random variables: $E(XY)$

There are two cases when finding the expectation of a product:

1. **General case:**

For general X and Y, $\mathbb{E}(XY)$ *is NOT equal to* $\mathbb{E}(X)\mathbb{E}(Y)$.

   We have to find $\mathbb{E}(XY)$ either using their joint probability function (see later), or using their covariance (see later).

2. **Special case:** when X and Y are *INDEPENDENT:*

When X and Y are INDEPENDENT, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Variable transformations

We often wish to *transform* random variables through a function. For example, given the random variable X, possible transformations of X include:

$$X^2, \qquad \sqrt{X}, \qquad 4X^3, \qquad \ldots$$

We often summarize all possible variable transformations by referring to $Y = g(X)$ *for some function* g.

For discrete random variables, it is very easy to find the probability function for $Y = g(X)$, given that the probability function for X is known. Simply ***change all the values and keep the probabilities the same.***

Example 1: Let $X \sim \text{Binomial}(3, 0.2)$, and let $Y = X^2$. Find the probability function of Y.

The probability function for X is:

x	0	1	2	3
$\mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

Thus *the probability function for* $Y = X^2$ *is:*

y	0^2	1^2	2^2	3^2
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

This is because Y *takes the value* 0^2 *whenever* X *takes the value* 0, *and so on.*

Variable transformations

Thus the probability that $Y = 0^2$ is *the same as the probability that* $X = 0$.

Overall, we would write the probability function of $Y = X^2$ as:

y	0	1	4	9
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

To transform a discrete random variable, *transform the values and leave the probabilities alone.*

Variable transformations

Example 2: Mr Chance hires out giant helium balloons for advertising. His balloons come in three sizes: heights 2m, 3m, and 4m. 50% of Mr Chance's customers choose to hire the cheapest 2m balloon, while 30% hire the 3m balloon and 20% hire the 4m balloon.

The amount of helium gas in cubic metres required to fill the balloons is $h^3/2$, where h is the height of the balloon. Find the probability function of Y, the amount of helium gas required for a randomly chosen customer.

Let X be the height of balloon ordered by a random customer. The probability function of X is:

height, x (m)	*2*	*3*	*4*
$\mathbb{P}(X = x)$	*0.5*	*0.3*	*0.2*

Let Y be the amount of gas required: $Y = X^3/2$.
The probability function of Y is:

gas, y (m^3)	*4*	*13.5*	*32*
$\mathbb{P}(Y = y)$	*0.5*	*0.3*	*0.2*

Expected value of a transformed random variable

We can find the expectation of a transformed random variable just like any other random variable. For example, in Example 1 we had $X \sim \text{Binomial}(3, 0.2)$, and $Y = X^2$.

The probability function for X is:

x	0	1	2	3
$\mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

and for $Y = X^2$:

y	0	1	4	9
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

Thus the expectation of $Y = X^2$ is:

$$\begin{aligned} \mathbb{E}(Y) = \mathbb{E}(X^2) &= 0 \times 0.512 + 1 \times 0.384 + 4 \times 0.096 + 9 \times 0.008 \\ &= 0.84. \end{aligned}$$

Note: $\mathbb{E}(X^2)$ *is NOT the same as* $\{\mathbb{E}(X)\}^2$. *Check that* $\{\mathbb{E}(X)\}^2 = 0.36$.

Expected value of a transformed random variable

To make the calculation quicker, we could cut out the middle step of writing down the probability function of Y. Because we transform the values and keep the probabilities the same, we have:

$$\mathbb{E}(X^2) = 0^2 \times 0.512 + 1^2 \times 0.384 + 2^2 \times 0.096 + 3^2 \times 0.008.$$

If we write $g(X) = X^2$, this becomes:

$$\mathbb{E}\{g(X)\} = \mathbb{E}(X^2) = g(0) \times 0.512 + g(1) \times 0.384 + g(2) \times 0.096 + g(3) \times 0.008.$$

Clearly the same arguments can be extended to any function $g(X)$ and any discrete random variable X:

$$\boxed{\mathbb{E}\{g(X)\} = \sum_x g(x)\mathbb{P}(X = x).}$$

Expected value of a transformed random variable

Definition: For any function g and discrete random variable X, the expected value of $g(X)$ is given by

$$\mathbb{E}\{g(X)\} = \sum_x g(x)\mathbb{P}(X = x) = \sum_x g(x) f_X(x).$$

Example: Recall Mr Chance and his balloon-hire business from page 74. Let X be the height of balloon selected by a randomly chosen customer. The probability function of X is:

height, x (m)	2	3	4
$\mathbb{P}(X = x)$	0.5	0.3	0.2

(a) What is the average amount of gas required per customer?

Expected value of a transformed random variable

Gas required was $X^3/2$ from page 74.
Average gas per customer is $\mathbb{E}(X^3/2)$.

$$\begin{aligned}
\mathbb{E}\left(\frac{X^3}{2}\right) &= \sum_x \frac{x^3}{2} \times \mathbb{P}(X = x) \\
&= \frac{2^3}{2} \times 0.5 + \frac{3^3}{2} \times 0.3 + \frac{4^3}{2} \times 0.2 \\
&= 12.45 \ \textit{m}^3 \textit{ gas.}
\end{aligned}$$

(b) Mr Chance charges \$400$\times h$ to hire a balloon of height h. What is his expected earning per customer?

Expected value of a transformed random variable

Expected earning is $\mathbb{E}(400X)$.

$$\begin{aligned}
\mathbb{E}(400X) &= 400 \times \mathbb{E}(X) \qquad \textit{(expectation is linear)} \\
&= 400 \times (2 \times 0.5 + 3 \times 0.3 + 4 \times 0.2) \\
&= 400 \times 2.7 \\
&= \$1080 \textit{ per customer.}
\end{aligned}$$

(c) How much does Mr Chance expect to earn in total from his next 5 customers?

Expected value of a transformed random variable

Let $Z_1, \ldots, Z_5$ *be the earnings from the next 5 customers. Each* Z_i *has* $\mathbb{E}(Z_i) = 1080$ *by part (b). The total expected earning is*

$$\begin{aligned} \mathbb{E}(Z_1 + Z_2 + \ldots + Z_5) &= \mathbb{E}(Z_1) + \mathbb{E}(Z_2) + \ldots + \mathbb{E}(Z_5) \\ &= 5 \times 1080 \\ &= \$5400. \end{aligned}$$

Expected value of a transformed random variable

Suppose $X = \begin{cases} 3 & \text{with probability } 3/4, \\ 8 & \text{with probability } 1/4. \end{cases}$

Then $3/4$ of the time, X takes value 3, and $1/4$ of the time, X takes value 8.

So $\mathbb{E}(X) = \frac{3}{4} \times 3 + \frac{1}{4} \times 8.$

ADD UP THE VALUES
TIMES HOW OFTEN THEY OCCUR

Common mistakes in calculate expected value of a transformed random variable

i) $\mathbb{E}(\sqrt{X}) = \sqrt{\mathbb{E}X} = \sqrt{\frac{3}{4} \times 3 \quad + \quad \frac{1}{4} \times 8}$

ii) $\mathbb{E}(\sqrt{X}) = \sqrt{\frac{3}{4}} \times 3 \quad + \quad \sqrt{\frac{1}{4}} \times 8$

iii) $\mathbb{E}(\sqrt{X}) = \sqrt{\frac{3}{4} \times 3} \quad + \quad \sqrt{\frac{1}{4} \times 8}$

$= \sqrt{\frac{3}{4}} \times \sqrt{3} \quad + \quad \sqrt{\frac{1}{4}} \times \sqrt{8}$

Common mistakes in calculate expected value of a transformed random variable

What about $\mathbb{E}(\sqrt{X})$?

$$\sqrt{X} = \begin{cases} \sqrt{3} & \text{with probability } 3/4, \\ \sqrt{8} & \text{with probability } 1/4. \end{cases}$$

ADD UP THE VALUES
TIMES HOW OFTEN THEY OCCUR

$$\mathbb{E}(\sqrt{X}) = \frac{3}{4} \times \sqrt{3} + \frac{1}{4} \times \sqrt{8}.$$

Properties of Expectation

i) Let g and h be functions, and let a and b be constants. For any random variable X (discrete or continuous),

$$\mathbb{E}\Big\{ag(X) + bh(X)\Big\} = a\mathbb{E}\Big\{g(X)\Big\} + b\mathbb{E}\Big\{h(X)\Big\}.$$

In particular,

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

ii) Let X and Y be ANY random variables (discrete, continuous, independent, or non-independent). Then

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

More generally, for ANY random variables $X_1, \ldots, X_n$,

$$\mathbb{E}(X_1 + \ldots + X_n) = \mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n).$$

iii) Let X and Y be **independent** random variables, and g, h be functions. Then

$$\begin{aligned} \mathbb{E}(XY) &= \mathbb{E}(X)\mathbb{E}(Y) \\ \mathbb{E}\Big(g(X)h(Y)\Big) &= \mathbb{E}\Big(g(X)\Big)\mathbb{E}\Big(h(Y)\Big). \end{aligned}$$

Notes: 1. $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ is ONLY generally true if X and Y are *INDEPENDENT.*

2. If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. However, the converse is not generally true: it is possible for $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ even though X and Y are dependent.

Variance

Example: Mrs Tractor runs the Rational Bank of Remuera. Every day she hopes to fill her cash machine with enough cash to see the well-heeled citizens of Remuera through the day. She knows that the expected amount of money withdrawn each day is \$50,000. How much money should she load in the machine? \$50,000?

No: \$50,000 is the average, near the centre of the distribution. About half the time, the money required will be GREATER than the average.

How much money should Mrs Tractor put in the machine if she wants to be 99% certain that there will be enough for the day's transactions?

Answer: it depends how much the amount withdrawn *varies above and below its mean.*

For questions like this, we need the study of *variance.*

Variance is the *average squared distance of a random variable from its own mean.*

Variance

Definition: The **variance** of a random variable X *is written as either* $\mathrm{Var}(X)$ *or* σ_X^2, *and is given by*

$$\sigma_X^2 = \mathrm{Var}(X) = \mathbb{E}\left[(X - \mu_X)^2\right] = \mathbb{E}\left[(X - \mathbb{E}X)^2\right].$$

Similarly, the variance of a function of X is

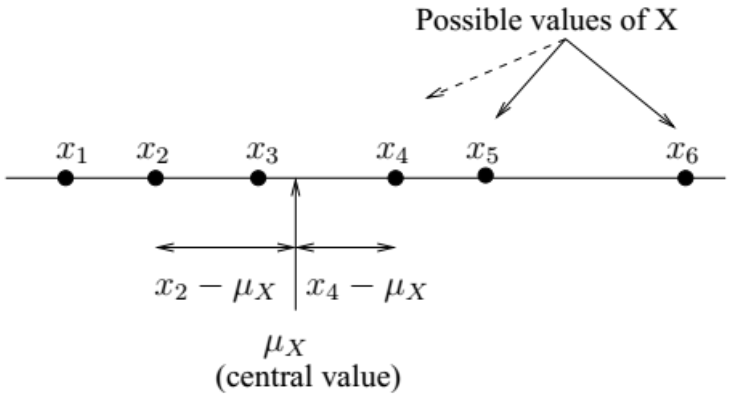
$$\mathrm{Var}(g(X)) = \mathbb{E}\left[\Big(g(X) - \mathbb{E}(g(X))\Big)^2\right].$$

Note: The variance *is the square of the standard deviation of X, so*

$$sd(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\sigma_X^2} = \sigma_X.$$

Variance as the average squared distance from the mean

The variance is a measure of how *spread out* are the values that X can take. It is the *average squared distance between a value of X and the central (mean) value, μ_X.*



$$\text{Var}(X) = \underbrace{\mathbb{E}}_{(2)}\ \underbrace{[(X-\mu_X)^2]}_{(1)}$$

(1) Take distance from observed values of X to the central point, μ_X. Square it to balance positive and negative distances.

(2) Then take the average over all values X can take: ie. if we observed X many times, find what would be the average squared distance between X and μ_X.

Note: The mean, μ_X, and the variance, σ_X^2, of X are just *numbers: there is nothing random or variable about them.*

Variance as the average squared distance from the mean

Example: Let $X = \begin{cases} 3 & \text{with probability } 3/4, \\ 8 & \text{with probability } 1/4. \end{cases}$

Then

$$\mathbb{E}(X) = \mu_X = 3 \times \frac{3}{4} + 8 \times \frac{1}{4} = 4.25$$

$$\begin{aligned} \mathit{Var}(X) = \sigma_X^2 &= \frac{3}{4} \times (3 - 4.25)^2 + \frac{1}{4} \times (8 - 4.25)^2 \\ &= 4.6875. \end{aligned}$$

When we observe X, we get either *3 or 8: this is random.*
But μ_X *is fixed at 4.25, and* σ_X^2 *is fixed at 4.6875, regardless of the outcome of X.*

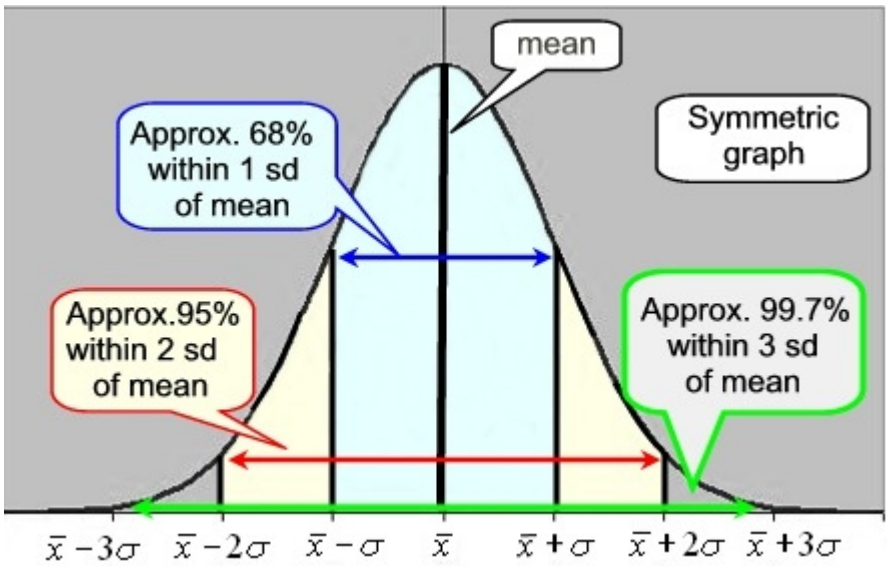
Variance as the average squared distance from the mean

For a discrete random variable,

$$\mathbf{Var}(X) = \mathbb{E}\left[(X-\mu_X)^2\right] = \sum_x (x-\mu_X)^2 f_X(x) = \sum_x (x-\mu_X)^2 \mathbb{P}(X=x).$$

This uses the definition of the expected value of a function of X:

$$\mathbf{Var}(X) = \mathbb{E}(g(X)) \quad \textit{where} \quad g(X) = (X-\mu_X)^2.$$



Variance as the average squared distance from the mean

$$\boxed{\textit{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2) - \mu_X^2}$$

Proof:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\left[(X-\mu_X)^2\right] \quad \text{by definition} \\
&= \mathbb{E}[\underbrace{X^2}_{\text{r.v.}} - 2\underbrace{X}_{\text{r.v.}}\underbrace{\mu_X}_{\text{constant}} + \underbrace{\mu_X^2}_{\text{constant}}] \\
&= \mathbb{E}(X^2) - 2\mu_X\mathbb{E}(X) + \mu_X^2 \quad \text{by Thm 2.7} \\
&= \mathbb{E}(X^2) - 2\mu_X^2 + \mu_X^2 \\
&= \mathbb{E}(X^2) - \mu_X^2 \,. \qquad \square
\end{aligned}$$

Note: $\mathbb{E}(X^2) = \sum_x x^2 f_X(x) = \sum_x x^2\,\mathbb{P}(X=x)$. This is not the same as $(\mathbb{E}X)^2$:

e.g.
$$X = \begin{cases} 3 & \text{with probability } 0.75, \\ 8 & \text{with probability } 0.25. \end{cases}$$

Then $\mu_X = \mathbb{E}X = 4.25$, so $\mu_X^2 = (\mathbb{E}X)^2 = (4.25)^2 = 18.0625$.
But
$$\mathbb{E}(X^2) = \left(3^2 \times \frac{3}{4} + 8^2 \times \frac{1}{4}\right) = 22.75.$$

Thus $\boxed{\mathbb{E}(X^2) \neq (\mathbb{E}X)^2 \textit{ in general.}}$

Properties of Variance

If a and b are constants and $g(x)$ is a function, then

i) $\mathsf{Var}(aX+b) = a^2\,\mathsf{Var}(X)$.

ii) $\mathsf{Var}(a\,g(X)+b) = a^2\,\mathsf{Var}\{g(X)\}$.

Proof:

(part (i))

$$\begin{aligned}
\mathsf{Var}(aX+b) &= \mathbb{E}\Big[\{(aX+b)-\mathbb{E}(aX+b)\}^2\Big] \\
&= \mathbb{E}\Big[\{aX+b-a\mathbb{E}(X)-b\}^2\Big] \quad \textit{by Thm 2.7} \\
&= \mathbb{E}\Big[\{aX-a\mathbb{E}(X)\}^2\Big] \\
&= \mathbb{E}\Big[a^2\{X-\mathbb{E}(X)\}^2\Big] \\
&= a^2\mathbb{E}\Big[\{X-\mathbb{E}(X)\}^2\Big] \quad \textit{by Thm 2.7} \\
&= a^2\,\mathsf{Var}(X)\,.
\end{aligned}$$

Part (ii) follows similarly.

Note: These are very different from the corresponding expressions for expectations (Theorem 2.7). Variances are more difficult to manipulate than expectations.

Example: finding expectation and variance from the probability function

Example: Recall Mr Chance and his balloon-hire business from page 74. Let X be the height of balloon selected by a randomly chosen customer. The probability function of X is:

height, x (m)	2	3	4
$\mathbb{P}(X = x)$	0.5	0.3	0.2

(a) What is the average amount of gas required per customer?

Gas required was $X^3/2$ from page 74.
Average gas per customer is $\mathbb{E}(X^3/2)$.

$$\mathbb{E}\left(\frac{X^3}{2}\right) = \sum_x \frac{x^3}{2} \times \mathbb{P}(X = x)$$

$$= \frac{2^3}{2} \times 0.5 + \frac{3^3}{2} \times 0.3 + \frac{4^3}{2} \times 0.2$$

$$= 12.45 \ m^3 \text{ gas.}$$

Find $\text{Var}(Y)$.

Example: finding expectation and variance from the probability function

Recall Mr Chance's balloons from page 74. The random variable Y is the amount of gas required by a randomly chosen customer. The probability function of Y is:

gas, y (m^3)	4	13.5	32
$\mathbb{P}(Y = y)$	0.5	0.3	0.2

Find $\text{Var}(Y)$.

First method: *use* $\text{Var}(Y) = \mathbb{E}[(Y - \mu_Y)^2]$*:*

$$\begin{aligned} \text{Var}(Y) &= (4 - 12.45)^2 \times 0.5 + (13.5 - 12.45)^2 \times 0.3 + (32 - 12.45)^2 \times 0.2 \\ &= 112.47. \end{aligned}$$

Second method: *use* $\mathbb{E}(Y^2) - \mu_Y^2$*: (usually easier)*

$$\begin{aligned} \mathbb{E}(Y^2) &= 4^2 \times 0.5 + 13.5^2 \times 0.3 + 32^2 \times 0.2 \\ &= 267.475. \end{aligned}$$

So $\text{Var}(Y) = 267.475 - (12.45)^2 = 112.47$ *as before.*

Variance of a sum of random variables: Var(X + Y)

There are two cases when finding the variance of a sum:

1. **General case:**

> *For general X and Y,*
> *$\mathrm{Var}(X+Y)$ is NOT equal to $\mathrm{Var}(X)+\mathrm{Var}(Y)$.*

We have to find $\mathrm{Var}(X+Y)$ using their covariance (see later courses).

2. **Special case:** when X and Y are *INDEPENDENT:*

> *When X and Y are INDEPENDENT,*
> *$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$.*

Class task for Expectation and Variance

A common problem for fire stations is the number of hoax calls received. Each hoax call must be answered as if it were a real fire, so it uses up fire station resources and personnel. A particular problem is that hoaxes are often more common at times when real fires are common, because the real fires put the idea into the hoaxers' minds. This means that fire stations can be most plagued by hoax calls at the times that they are busiest with real fires.

This question suggests how a fire station might consider modelling the occurrences of real fires and hoax calls, so that it can plan how to allocate its resources.

Let Y be the number of **real fires** that will occur over the next month.

Let X be the number of **hoax calls** that will occur over the next month.

Suppose that

$$\begin{aligned} Y &\sim \text{Poisson}(\lambda)\,, \\ X \,|\, Y &\sim \text{Poisson}(\alpha + \beta Y)\,. \end{aligned}$$

(a) State $E(Y)$ and $\text{var}(Y)$. Hence state $E(Y^2)$. [Hint: $\text{var}(Y) = E(Y^2) - (EY)^2$.]

(b) State $E(X \,|\, Y)$ and $\text{var}(X \,|\, Y)$. (Note that you should leave your result in terms of Y.)

(c) Using the formula for conditional expectation, show that $E(X) = \alpha + \beta\lambda$.

(d) The fire station needs to budget for the **total number of calls**, $X + Y$, because each hoax call must be answered as if it were real. Show that

$$E(X + Y) = \alpha + \beta\lambda + \lambda.$$

Class task for Expectation and Variance

(e) **Note:** the procedure used in this question is required

To find $\text{var}(X+Y)$, we will need to find $\text{cov}(X,Y) = E(XY) - E(X)E(Y)$. Using the formula for conditional expectation, we have

$$E(XY) = E_Y\left\{ E(XY \,|\, Y) \right\}.$$

Conditional on Y, we can take the Y outside the inner expectation as a constant:

$$E(XY) = E_Y\left\{ Y \times E(X \,|\, Y) \right\}.$$

Complete the working to show that

$$E(XY) = \alpha\lambda + \beta(\lambda + \lambda^2).$$

[Hint: you will need to use the result for $E(Y^2)$ from part (a).]

Covariance

Covariance is a measure of the association or dependence between two random variables X and Y. Covariance can be either positive or negative. (*Variance* is always positive.)

Definition: Let X and Y be any random variables. The **covariance** between X and Y is given by

$$\mathbf{cov}(X,Y) = \mathbb{E}\Big\{(X-\mu_X)(Y-\mu_Y)\Big\} = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

where $\mu_X = \mathbb{E}(X)$, $\mu_Y = \mathbb{E}(Y)$.

1. $\mathrm{cov}(X,Y)$ will be ***positive*** if large values of X tend to occur with large values of Y, and small values of X tend to occur with small values of Y. For example, if X is height and Y is weight of a randomly selected person, we would expect $\mathrm{cov}(X,Y)$ to be positive.

2. $\mathrm{cov}(X,Y)$ will be ***negative*** if large values of X tend to occur with small values of Y, and small values of X tend to occur with large values of Y. For example, if X is age of a randomly selected person, and Y is heart rate, we would expect X and Y to be negatively correlated (older people have slower heart rates).

3. If X and Y are independent, then there is no pattern between large values of X and large values of Y, so $\mathrm{cov}(X,Y) = 0$. However, $\mathrm{cov}(X,Y) = 0$ does NOT imply that X and Y are independent, unless X and Y are Normally distributed.

Covariance, example

Calculate covariance for the following data set:

x: 2.1, 2.5, 3.6, 4.0 (mean = 3.1)

y: 8, 10, 12, 14 (mean = 11)

Substitute the values into the formula and solve:

$Cov(X,Y) = \Sigma E((X-\mu)(Y-v)) / n-1$

$= (2.1-3.1)(8-11)+(2.5-3.1)(10-11)+(3.6-3.1)(12-11)+(4.0-3.1)(14-11) /(4-1)$

$= (-1)(-3) + (-0.6)(-1)+(.5)(1)+(0.9)(3) / 3$

$= 3 + 0.6 + .5 + 2.7 / 3$

$= 6.8/3$

$= 2.267$

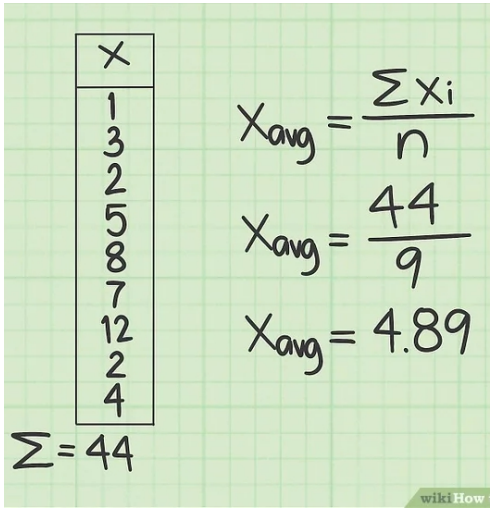
	S&P 500	ABC Corp.
2013	1,692	68
2014	1,978	102
2015	1,884	110
2016	2,151	112
2017	2,519	154

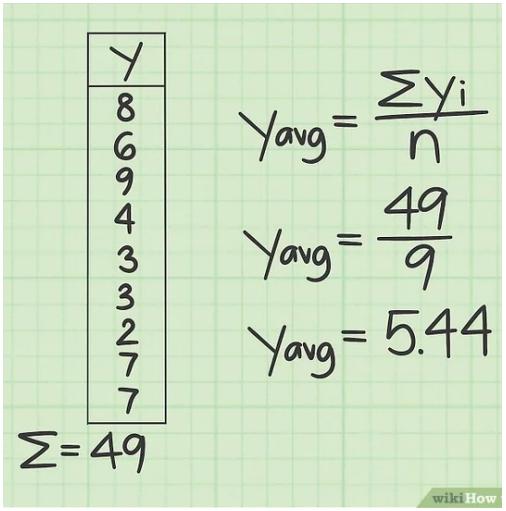
Daily Return for Two Stocks Using the Closing Prices

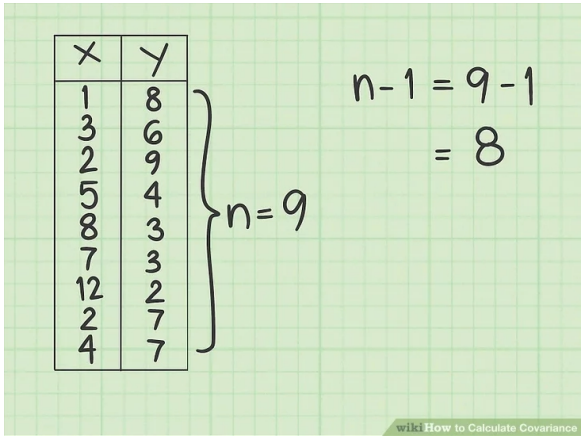
Day	ABC Returns	XYZ Returns
1	1.1%	3.0%
2	1.7%	4.2%
3	2.1%	4.9%
4	1.4%	4.1%
5	0.2%	2.5%

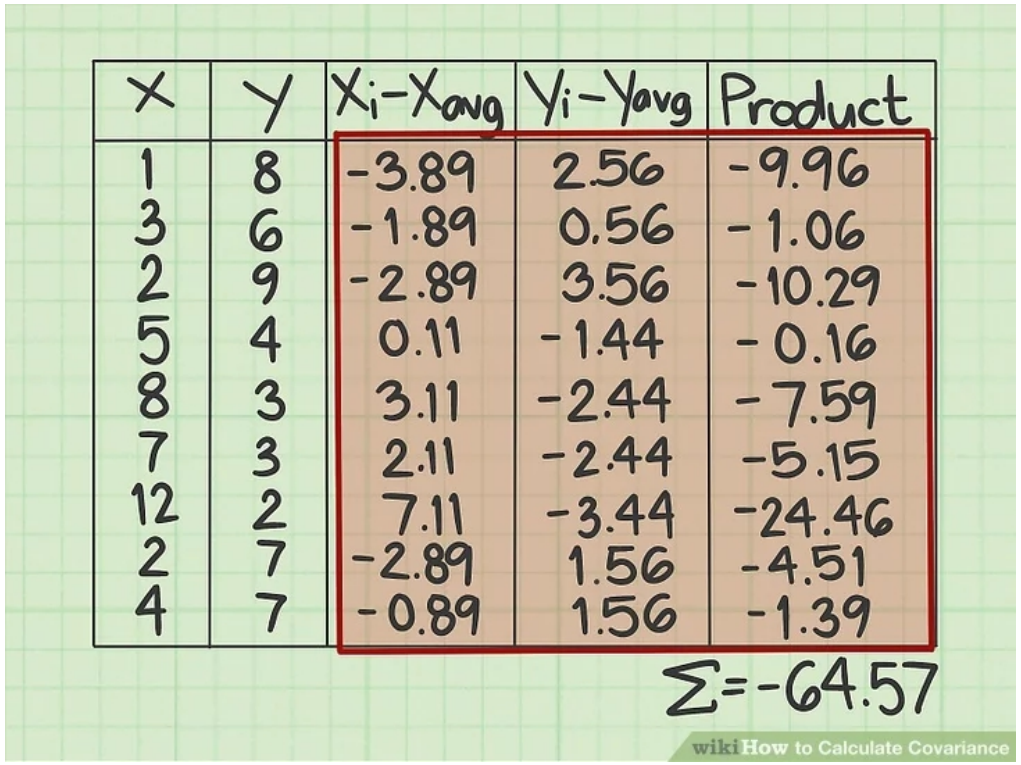
Covariance, example

Compute covariance of X and Y









https://www.wikihow.com/Calculate-Covariance

Covariance

Theorem 2.19. $\mathrm{Cov}(X_1, X_2) = \mathrm{E}(X_1X_2) - \mu_1\mu_2$.

Proof.

$$\begin{aligned}\mathrm{Cov}(X_1, X_2) &= \mathrm{E}\{(X_1 - \mu_1)(X_2 - \mu_2)\} \\ &= \mathrm{E}(X_1X_2 - X_1\mu_2 - \mu_1X_2 + \mu_1\mu_2) \\ &= \mathrm{E}(X_1X_2) - \mu_1\mu_2 - \mu_1\mu_2 + \mu_1\mu_2 \\ &= \mathrm{E}(X_1X_2) - \mu_1\mu_2.\end{aligned}$$

□

We have shown in the previous subsection how to obtain $\mathrm{E}(X_1X_2)$, and we can get the individual mean μ_i in the usual way from the marginal pdf of X_i.

A special case is when X_1 and X_2 are independent, for which $\mathrm{E}(X_1X_2) = \mathrm{E}(X_1)\mathrm{E}(X_2)$. As a result, if X_1 and X_2 are **independent**, then

$$\mathrm{Cov}(X_1, X_2) = 0. \tag{2.4}$$

Covariances arise naturally as soon as we try to find variances of linear combinations of r.v.'s that are not independent.

Covariance

Theorem 2.20.

(a) $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2)$;

(b) $\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) - 2\mathrm{Cov}(X_1, X_2)$;

(c) For constants a_1 and a_2,

$$\mathrm{Var}(a_1X_1 + a_2X_2) = a_1^2\mathrm{Var}(X_1) + a_2^2\mathrm{Var}(X_2) + 2a_1a_2\mathrm{Cov}(X_1, X_2);$$

(d) For any constants a_i,

$$\mathrm{Var}\left(\sum_i a_iX_i\right) = \sum a_i^2\mathrm{Var}(X_i) + 2\sum_i\sum_{j>i} a_ia_j\mathrm{Cov}(X_i, X_j).$$

Note 1. A useful way of remembering (a), (b) and (c) above is to compare the results with (a) $(X_1+X_2)^2$, (b) $(X_1-X_2)^2$, (c) $(a_1X_1+a_2X_2)^2$. Squares become variances and cross-products become covariances. This actually works for (d) as well.

Covariance

Note 2. The familiar results for the variances of independent r.v.'s, e.g., $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$, follow by noting that the covariances are zero for independent r.v.'s.

Proof. Items (a) and (b) are special cases of (c) ($a_1 = a_2 = 1$, and $a_1 = 1, a_2 = -1$, respectively), and item (d) follows easily from (c) by mathematical induction. We will prove (c).

$$
\begin{aligned}
&\mathrm{Var}(aX_1 + a_2X_2) \\
&= \mathrm{E}\left[\{a_1X_1 + a_2X_2 - \mathrm{E}(a_1X_1 + a_2X_2)\}^2\right] \\
&= \mathrm{E}\left[\{a_1(X_1 - \mu_1) + a_2(X_2 - \mu_2)\}^2\right] \\
&= \mathrm{E}\left\{a_1^2(X_1 - \mu_1)^2 + 2a_1a_2(X_1 - \mu_1)(X_2 - \mu_2) + a_2^2(X_2 - \mu_2)^2\right\} \\
&= a_1^2\mathrm{E}\left\{(X_1 - \mu_1)^2\right\} + 2a_1a_2\mathrm{E}\left\{(X_1 - \mu_1)(X_2 - \mu_2)\right\} + a_2^2\mathrm{E}\left\{(X_2 - \mu_2)^2\right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad [\text{by Theorem 2.16}] \\
&= a_1^2\mathrm{Var}(X_1) + 2a_1a_2\mathrm{Cov}(X_1, X_2) + a_2^2\mathrm{Var}(X_2).
\end{aligned}
$$

□

Covariance

Theorem 2.22. Let a, a_1, a_2 and a_3 be arbitrary constants.

(a) $\mathrm{Cov}(aX_1, X_2) = a\mathrm{Cov}(X_1, X_2)$.

(b) $\mathrm{Cov}(X_1 + X_2, X_3) = \mathrm{Cov}(X_1, X_3) + \mathrm{Cov}(X_2, X_3)$.

(c) $\mathrm{Cov}(a_1X_1 + a_2X_2, a_3X_3) = a_1a_3\mathrm{Cov}(X_1, X_3) + a_2a_3\mathrm{Cov}(X_2, X_3)$.

Proof. (a)

$$\begin{aligned}\mathrm{Cov}(aX_1, X_2) &= \mathrm{E}\{(aX_1 - a\mu_1)(X_2 - \mu_2)\} \\ &= a\mathrm{E}\{(X_1 - \mu_1)(X_2 - \mu_2)\} \\ &= a\mathrm{Cov}(X_1, X_2).\end{aligned}$$

(b)

$$\begin{aligned}\mathrm{Cov}(X_1 + X_2, X_3) &= \mathrm{E}\{(X_1 + X_2 - \mu_1 - \mu_2)(X_3 - \mu_3)\} \\ &= \mathrm{E}\{(X_1 - \mu_1)(X_3 - \mu_3)\} + \mathrm{E}\{(X_2 - \mu_2)(X_3 - \mu_3)\} \\ &= \mathrm{Cov}(X_1, X_3) + \mathrm{Cov}(X_2, X_3).\end{aligned}$$

(c) It follows easily from (a) and (b). □

Class task

5. Let X and Y be random variables such that $Var(X) = \sigma^2$, $Var(Y) = \tau^2$, and $Cov(X, Y) = \gamma$. Let a, b, c, d be constants. Find $Cov(aX + bY, cX + dY)$ in terms of the seven constants in this problem.

5. $Cov(aX + bY, cX + dY) = acCov(X, X) + adCov(X, Y) + bcCov(Y, X) + bdCov(Y, Y) = ac\sigma^2 + (ad + bc)\gamma + bd\tau^2$.

Correlation

The correlation coefficient of X and Y is a measure of the linear association between X and Y. It is given by the covariance, scaled by the overall variability in X and Y. As a result, the correlation coefficient is always between -1 and $+1$, so it is easily compared for different quantities.

Definition: The **correlation** between X and Y, also called the **correlation coefficient**, is given by

$$\boxed{\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.}$$

The correlation measures linear association between X and Y. It takes values only between -1 and $+1$, and has the same sign as the covariance.

The correlation is ± 1 if and only if there is a perfect linear relationship between X and Y, i.e. $\text{corr}(X,Y) = 1 \iff Y = aX + b$ for some constants a and b.

The correlation is 0 if X and Y are independent, but a correlation of 0 does not *imply* that X and Y are independent.

Correlation, example 1

Calculate Correlation of x and y

x	y
1	1
2	3
4	5
5	7

$$\rho = \left(\frac{1}{n-1}\right)\Sigma\left(\frac{x-\mu_x}{\sigma_x}\right)*\left(\frac{y-\mu_y}{\sigma_y}\right)$$

$$\mu_x = (1+2+4+5)\,/\,4$$

$$\mu_x = 12\,/\,4$$

$$\mu_x = 3$$

$$\sigma_x = \sqrt{\frac{1}{n-1}\Sigma(x-\mu_x)^2}$$

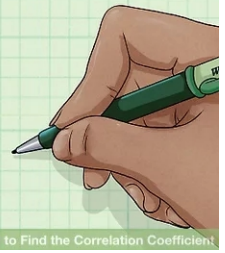
$$\sigma_x = \sqrt{\frac{1}{4-1}*((1-3)^2+(2-3)^2+(4-3)^2+(5-3)^2)}$$

$$\sigma_x = \sqrt{\frac{1}{3}*(4+1+1+4)}$$

$$\sigma_x = \sqrt{\frac{1}{3}*(10)}$$

$$\sigma_x = \sqrt{\frac{10}{3}}$$

$$\sigma_x = 1.83$$



$$\mu_y = (1+3+5+7)/4 \;=\; 4$$

$$\sigma_y = \sqrt{\frac{1}{4-1}*((1-4)^2+(3-4)^2+(5-4)^2+(7-4)^2)}$$

$$\sigma_y = \sqrt{\frac{1}{3}*(9+1+1+9)}$$

$$\sigma_y = \sqrt{\frac{1}{3}*(20)}$$

$$\sigma_y = \sqrt{\frac{20}{3}}$$

$$\sigma_y = 2.58$$

https://www.wikihow.com/Find-the-Correlation-Coefficient

Correlation, example 1

$$\rho = \left(\frac{1}{n-1}\right)\Sigma\left(\frac{x-\mu_x}{\sigma_x}\right)*\left(\frac{y-\mu_y}{\sigma_y}\right)$$

$$\rho = \left(\frac{1}{3}\right)*\left[\left(\frac{1\text{-}3}{1.83}\right)*\left(\frac{1\text{-}4}{2.58}\right)+\left(\frac{2\text{-}3}{1.83}\right)*\left(\frac{3\text{-}4}{2.58}\right)\right.$$
$$\left.+\left(\frac{4\text{-}3}{1.83}\right)*\left(\frac{5\text{-}4}{2.58}\right)+\left(\frac{5\text{-}3}{1.83}\right)*\left(\frac{7\text{-}4}{2.58}\right)\right]$$

$$\rho = \left(\frac{1}{3}\right)*\left(\frac{6+1+1+6}{4.721}\right)$$

$$\rho = \left(\frac{1}{3}\right)* 2.965$$

$$\rho = \left(\frac{2.965}{3}\right)$$

$$\rho = 0.988$$

wikiHow to Find the Correlation Coefficient

https://www.wikihow.com/Find-the-Correlation-Coefficient

Correlation, example 2

	Exercise (hour/day)	Weight (kg)
W1	0.25	78.0
W2	0.50	77.6
W3	0.75	77.1
W4	1.00	75
W5	1.50	70
W6	2.00	65

Correlation

TABLE 1.1 Table with four groups of numbers: What do they tell you?

Group A		Group B		Group C		Group D	
x	y	x	y	x	y	x	y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

https://www.kaggle.com/carlmcbrideellis/data-anscombes-quartet

Explore Data

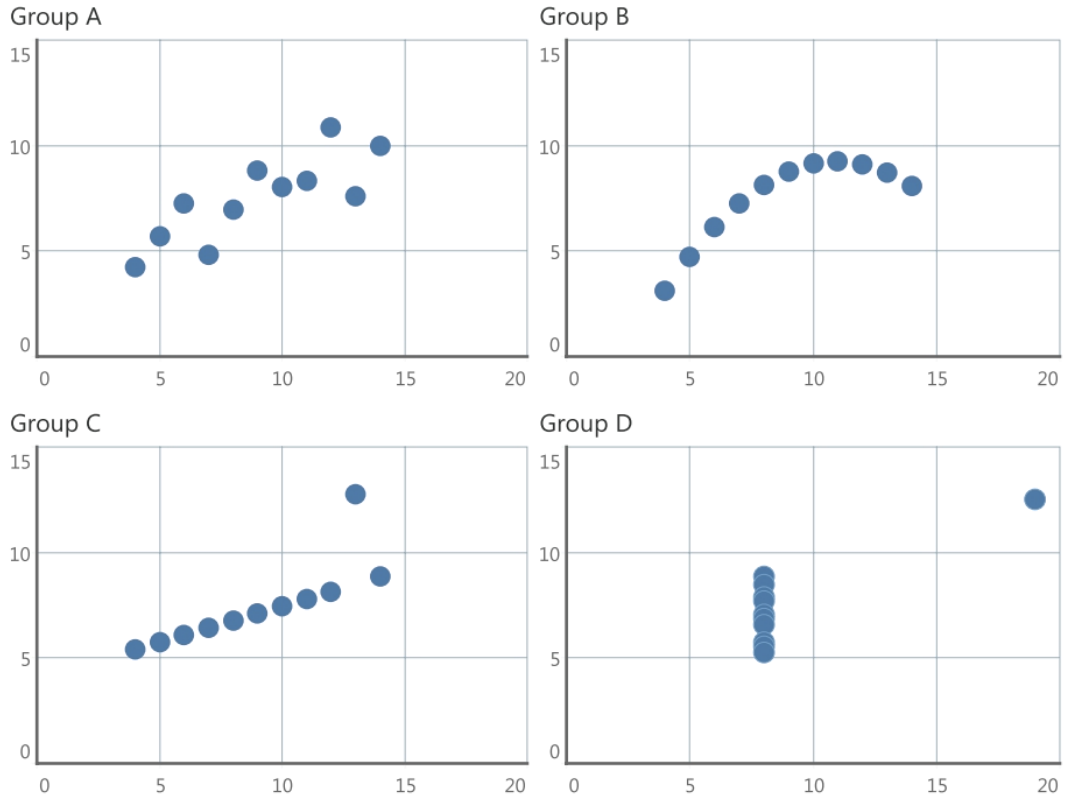


FIGURE 1.1 Now can you see a difference in the four groups?

“Anscombe’s Quartet”—in the paper “Graphs in Statistical Analysis” in 1973.

Theorem 2.23 (Properties of the correlation)**.**

(a) $-1 \leq \rho \leq 1$.

(b) $\rho^2 = 1 \iff X_2 = aX_1 + b$ for some constants a and b. ($\rho = 1$ if $a > 0$; $\rho = -1$ if $a < 0$.)

(c) $\rho = 0$ if X_1 and X_2 are independent, but the converse is not necessarily true.

Proof. (a) The result follows from setting up a r.v. $Z = aX_1 - X_2$ and arguing that its variance must be non-negative regardless of the value of a.

$$\begin{aligned} 0 &\leq \mathrm{Var}(Z) \\ &= \mathrm{Var}(aX_1 - X_2) \\ &= a^2\mathrm{Var}(X_1) - 2a\mathrm{Cov}(X_1, X_2) + \mathrm{Var}(X_2). \end{aligned}$$

This is a quadratic in a and since the quadratic must always be non-negative, it follows that the **discriminant** of the quadratic must be non-positive, i.e.,

$$4\left\{\mathrm{Cov}(X_1, X_2)\right\}^2 - 4\mathrm{Var}(X_1)\mathrm{Var}(X_2) \leq 0,$$

so

$$\frac{\{\mathrm{Cov}(X_1, X_2)\}^2}{\sigma_1^2 \sigma_2^2} \leq 1,$$

i.e., $\rho^2 \leq 1$.

(b) If $\rho^2 = 1$, then the discriminant is zero, so there is a value of a such that $\mathrm{Var}(aX_1 - X_2) = 0$. Since the variance is zero for this value of a, $aX_1 - X_2$ is a constant and is equal to $-b$, say, i.e., $X_2 = aX_1 + b$. When the discriminant is zero, the value of a which is the root of the quadratic is $\mathrm{Cov}(X_1, X_2)/\mathrm{Var}(X_1)$. Thus, a has the same sign as ρ. Conversely, if $X_2 = aX_1 + b$, direct calculation from the definition gives $\rho^2 = 1$.

(c) Suppose X_1 and X_2 are independent. We have proved in Theorem 2.18 that in this case $\mathrm{E}(X_1 X_2) = \mathrm{E}(X_1)\mathrm{E}(X_2)$, so $\mathrm{Cov}(X_1, X_2) = \mathrm{E}(X_1)\mathrm{E}(X_2) - \mu_1\mu_2 = 0$. Hence $\rho = 0$.

However, the converse is not necessarily true. As a counter-example, take the random vector having values $(1, 0)$, $(0, 1)$, $(-1, 0)$ and $(0, -1)$ with equal probability. The correlation is 0 but X_1 and X_2 are not independent because $f_{X_1,X_2}(0, 0) = 0$ and $f_{X_1}(0) \times f_{X_2}(0) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. □

Probability as an Expectation

Let A be any event. We can write $\mathbb{P}(A)$ as an expectation, as follows. Define the ***indicator function:***

$$I_A = \begin{cases} 1 & \textit{if event } A \textit{ occurs,} \\ 0 & \textit{otherwise.} \end{cases}$$

Then I_A is a ***random variable***, and

$$\begin{aligned} \mathbb{E}(I_A) &= \sum_{r=0}^{1} r\mathbb{P}(I_A = r) \\ &= 0 \times \mathbb{P}(I_A = 0) + 1 \times \mathbb{P}(I_A = 1) \\ &= \mathbb{P}(I_A = 1) \\ &= \mathbb{P}(A). \end{aligned}$$

Thus

$$\boxed{\mathbb{P}(A) = \mathbb{E}(I_A) \textit{ for any event } A.}$$

Conditional Expectation and Conditional Variance

Throughout this section, we will assume for simplicity that X and Y are discrete random variables. However, exactly the same results hold for continuous random variables too.

Suppose that X and Y are discrete random variables, possibly dependent on each other. Suppose that we fix Y at the value y. This gives us a set of conditional probabilities $\mathbb{P}(X = x \,|\, Y = y)$ for all possible values x of X. This is called the *conditional distribution of X, given that $Y = y$.*

Definition: Let X and Y be discrete random variables. The **conditional probability function** of X, given that $Y = y$, is:

$$\mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(X = x \textit{ AND } Y = y)}{\mathbb{P}(Y = y)}.$$

We write the conditional probability function as:

$$f_{X\,|\,Y}(x \,|\, y) = \mathbb{P}(X = x \,|\, Y = y).$$

Note: The conditional probabilities $f_{X\,|\,Y}(x \,|\, y)$ sum to one, just like any other probability function:

$$\sum_x \mathbb{P}(X = x \,|\, Y = y) = \sum_x \mathbb{P}_{\{Y=y\}}(X = x) = 1,$$

using the subscript notation $\mathbb{P}_{\{Y=y\}}$ of Section 2.3.

Conditional Expectation

We can also find the expectation and variance of X with respect to this conditional distribution. That is, if we know that the value of Y is fixed at y, then we can find the mean value of X *given that* Y takes the value y, and also the variance of X given that $Y = y$.

Definition: Let X and Y be discrete random variables. The **conditional expectation of X, given that $Y = y$**, is

$$\mu_{X\,|\,Y=y} = \mathbb{E}(X \,|\, Y = y) = \sum_x x f_{X\,|\,Y}(x \,|\, y).$$

$\mathbb{E}(X \,|\, Y = y)$ is the *mean value of X, when Y is fixed at y.*

Conditional expectation as a random variable

The unconditional expectation of X, $\mathbb{E}(X)$, is just *a number:*
e.g. $\mathbb{E}X = 2$ *or* $\mathbb{E}X = 5.8$.

The conditional expectation, $\mathbb{E}(X \mid Y = y)$, is *a number depending on* y.

If Y has an influence on the value of X, then Y will have an influence on the *average* value of X. So, for example, we would expect $\mathbb{E}(X \mid Y = 2)$ to be different from $\mathbb{E}(X \mid Y = 3)$.

We can therefore view $\mathbb{E}(X \mid Y = y)$ as a *function of* y, *say* $\mathbb{E}(X \mid Y{=}y) = h(y)$.

To evaluate this function, $h(y) = \mathbb{E}(X \mid Y = y)$, we:

i) *fix* Y *at the chosen value* y*;*

ii) *find the expectation of* X *when* Y *is fixed at this value.*

Conditional expectation as a random variable

However, we could also evaluate the function at a *random value* of Y:

i) *observe a random value of* Y*;*

ii) *fix* Y *at that observed random value;*

iii) *evaluate* $\mathbb{E}(X \mid Y =$ *observed random value*$)$.

We obtain a *random variable*: $\mathbb{E}(X \mid Y) = h(Y)$.

The randomness comes from the randomness in Y*, not in* X.

> *Conditional expectation,* $\mathbb{E}(X \mid Y)$*, is a random variable with randomness inherited from* Y*, not* X.

Conditional expectation of X given $Y = y$ is a number depending on y

Example: Suppose $Y = \begin{cases} 1 & \text{with probability } 1/8\,, \\ 2 & \text{with probability } 7/8\,, \end{cases}$

and $X \,|\, Y = \begin{cases} 2Y & \text{with probability } 3/4\,, \\ 3Y & \text{with probability } 1/4\,. \end{cases}$

If $Y = 1$, *then:* $X \,|\, (Y = 1) = \begin{cases} 2 & \textit{with probability } 3/4 \\ 3 & \textit{with probability } 1/4 \end{cases}$

$$\textit{so} \quad \mathbb{E}(X \,|\, Y = 1) = 2 \times \tfrac{3}{4} + 3 \times \tfrac{1}{4} = \tfrac{9}{4}.$$

If $Y = 2$, *then:* $X \,|\, (Y = 2) = \begin{cases} 4 & \textit{with probability } 3/4 \\ 6 & \textit{with probability } 1/4 \end{cases}$

$$\textit{so} \quad \mathbb{E}(X \,|\, Y = 2) = 4 \times \tfrac{3}{4} + 6 \times \tfrac{1}{4} = \tfrac{18}{4}.$$

Thus $\mathbb{E}(X \,|\, Y = y) = \begin{cases} 9/4 & \textit{if } y = 1 \\ 18/4 & \textit{if } y = 2. \end{cases}$

So $\mathbb{E}(X \,|\, Y = y)$ *is a number depending on* y, *or a function of* y.

Conditional expectation of X given random Y is a random variable

Example: Suppose $Y = \begin{cases} 1 & \text{with probability } 1/8\,, \\ 2 & \text{with probability } 7/8\,, \end{cases}$

and $X \,|\, Y = \begin{cases} 2Y & \text{with probability } 3/4\,, \\ 3Y & \text{with probability } 1/4\,. \end{cases}$

From above, $\mathbb{E}(X \,|\, Y) = \begin{cases} 9/4 & \textit{if } Y = 1 \text{ (}\textit{probability } 1/8\text{)}, \\ 18/4 & \textit{if } Y = 2 \text{ (}\textit{probability } 7/8\text{)}. \end{cases}$

So $\mathbb{E}(X \,|\, Y) = \begin{cases} 9/4 & \textit{with probability } 1/8, \\ 18/4 & \textit{with probability } 7/8. \end{cases}$

Thus $\mathbb{E}(X \,|\, Y)$ *is a random variable.*

The randomness in $\mathbb{E}(X \,|\, Y)$ *is inherited from* Y*, not from* X*.*

Conditional expectation is a very useful tool for finding the ***unconditional*** expectation of X (see below). Just like the Partition Theorem, it is useful because it is often easier to specify conditional probabilities than to specify overall probabilities.

Conditional variance

The conditional variance is similar to the conditional expectation.

- $\mathrm{Var}(X \mid Y = y)$ is the variance of X, when Y is fixed at the value $Y = y$.
- $\mathrm{Var}(X \mid Y)$ is a random variable, giving the variance of X when Y is fixed at a value to be selected randomly.

Definition: Let X and Y be random variables. The **conditional variance of X, given Y,** is given by

$$\mathrm{Var}(X \mid Y) = \mathbb{E}(X^2 \mid Y) - \Big\{\mathbb{E}(X \mid Y)\Big\}^2 = \mathbb{E}\Big\{(X - \mu_{X \mid Y})^2 \mid Y\Big\}$$

Like expectation, $\mathrm{Var}(X \mid Y = y)$ is a number depending on y (a function of y), while $\mathrm{Var}(X \mid Y)$ is a random variable with randomness inherited from Y.

Laws of Total Expectation and Variance

If all the expectations below are finite, then for ANY random variables X and Y, we have:

i) $\boxed{\mathbb{E}(X) = \mathbb{E}_Y\Big(\mathbb{E}(X \mid Y)\Big)}$ *Law of Total Expectation.*

Note that we can pick any r.v. Y, to make the expectation as easy as we can.

ii) $\mathbb{E}(g(X)) = \mathbb{E}_Y\Big(\mathbb{E}(g(X) \mid Y)\Big)$ *for any function g.*

iii) $\boxed{\mathit{Var}(X) = \mathbb{E}_Y\Big(\mathit{Var}(X \mid Y)\Big) + \mathit{Var}_Y\Big(\mathbb{E}(X \mid Y)\Big)}$

Law of Total Variance.

Note: $\mathbb{E}_Y$ and Var_Y denote expectation over Y and variance over Y,

i.e. the expectation or variance is computed over the distribution of the random variable Y.

The Law of Total Expectation says that *the total average is the average of case-by-case averages.*

- The total average is $\mathbb{E}(X)$;
- The case-by-case averages are $\mathbb{E}(X \mid Y)$ *for the different values of Y;*
- The average of case-by-case averages is *the average over Y of the Y-case averages:* $\mathbb{E}_Y\Big(\mathbb{E}(X \mid Y)\Big)$.

Laws of Total Expectation and Variance, example

Example: In the example above, we had: $\mathbb{E}(X \mid Y) = \begin{cases} 9/4 & \text{with probability } 1/8, \\ 18/4 & \text{with probability } 7/8. \end{cases}$

The total average is:

$$\mathbb{E}(X) = \mathbb{E}_Y\Big\{\mathbb{E}(X \mid Y)\Big\} = \frac{9}{4} \times \frac{1}{8} + \frac{18}{4} \times \frac{7}{8} = 4.22.$$

Laws of Total Expectation and Variance, proof

(i) is a special case of (ii), so we just need to prove (ii). Begin at RHS:

$$
\begin{aligned}
\text{RHS} \;=\; \mathbb{E}_Y\Big[\mathbb{E}(g(X)\,|\,Y)\Big] \;&=\; \mathbb{E}_Y\left[\sum_x g(x)\mathbb{P}(X=x\,|\,Y)\right] \\
&= \sum_y \left[\sum_x g(x)\mathbb{P}(X=x\,|\,Y=y)\right]\mathbb{P}(Y=y) \\
&= \sum_y\sum_x g(x)\mathbb{P}(X=x\,|\,Y=y)\mathbb{P}(Y=y) \\
&= \sum_x g(x)\sum_y \mathbb{P}(X=x\,|\,Y=y)\mathbb{P}(Y=y) \\
&= \sum_x g(x)\mathbb{P}(X=x) \qquad \text{(partition rule)} \\
&= \mathbb{E}(g(X)) \;=\; \text{LHS.}
\end{aligned}
$$

Laws of Total Expectation and Variance, proof

(iii) Wish to prove $\mathrm{Var}(X) = \mathbb{E}_Y[\mathrm{Var}(X \mid Y)] + \mathrm{Var}_Y[\mathbb{E}(X \mid Y)]$. Begin at RHS:

$$\mathbb{E}_Y[\mathrm{Var}(X \mid Y)] + \mathrm{Var}_Y[\mathbb{E}(X \mid Y)]$$

$$= \mathbb{E}_Y\left\{\mathbb{E}(X^2 \mid Y) - (\mathbb{E}(X \mid Y))^2\right\} + \left\{\mathbb{E}_Y\left\{[\mathbb{E}(X \mid Y)]^2\right\} - \left[\underbrace{\mathbb{E}_Y(\mathbb{E}(X \mid Y))}_{\mathbb{E}(X) \text{ by part (i)}}\right]^2\right\}$$

$$= \underbrace{\mathbb{E}_Y\{\mathbb{E}(X^2 \mid Y)\}}_{\mathbb{E}(X^2) \text{ by part (i)}} - \mathbb{E}_Y\{[\mathbb{E}(X \mid Y)]^2\} + \mathbb{E}_Y\{[\mathbb{E}(X \mid Y)]^2\} - (\mathbb{E}X)^2$$

$$= \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

$$= \mathrm{Var}(X) = \mathrm{LHS}. \qquad \square$$

Some important discrete distribution

1. Binomial distribution

Notation: $X \sim \text{Binomial}(n, p)$.

Description: number of successes in n independent trials, each with probability p of success.

Probability function:

$$f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n.$$

Mean: $\mathbb{E}(X) = np$.

Variance: $\text{Var}(X) = np(1-p) = npq$, where $q = 1 - p$.

Sum: If $X \sim \text{Binomial}(n, p)$, $Y \sim \text{Binomial}(m, p)$, and X and Y are **independent**, then

$$X + Y \sim \text{Bin}(n + m,\ p).$$

---

2. Poisson distribution

Notation: $X \sim \text{Poisson}(\lambda)$.

Description: arises out of the Poisson process as the number of events in a fixed time or space, when events occur at a constant average rate. Also used in many other situations.

Probability function: $f_X(x) = \mathbb{P}(X = x) = \dfrac{\lambda^x}{x!} e^{-\lambda} \quad \text{for} \quad x = 0, 1, 2, \ldots$

Mean: $\mathbb{E}(X) = \lambda$.

Variance: $\text{Var}(X) = \lambda$.

Sum: If $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$, and X and Y are **independent**, then

$$X + Y \sim \text{Poisson}(\lambda + \mu).$$

---

3. Geometric distribution

Notation: $X \sim \text{Geometric}(p)$.

Description: number of failures before the **first** success in a sequence of independent trials, each with $\mathbb{P}(\text{success}) = p$.

Probability function: $f_X(x) = \mathbb{P}(X = x) = (1-p)^x p \quad \text{for} \quad x = 0, 1, 2, \ldots$

Mean: $\mathbb{E}(X) = \dfrac{1-p}{p} = \dfrac{q}{p}$, where $q = 1 - p$.

Variance: $\text{Var}(X) = \dfrac{1-p}{p^2} = \dfrac{q}{p^2}$, where $q = 1 - p$.

Sum: if $X_1, \ldots, X_k$ are **independent**, and each $X_i \sim \text{Geometric}(p)$, then

$$X_1 + \ldots + X_k \sim \text{Negative Binomial}(k, p).$$

---

4. Negative Binomial distribution

Notation: $X \sim \text{NegBin}(k, p)$.

Description: number of failures before the ***k*th** success in a sequence of independent trials, each with $\mathbb{P}(\text{success}) = p$.

Probability function:

$$f_X(x) = \mathbb{P}(X = x) = \binom{k + x - 1}{x} p^k (1-p)^x \quad \text{for} \quad x = 0, 1, 2, \ldots$$

Mean: $\mathbb{E}(X) = \dfrac{k(1-p)}{p} = \dfrac{kq}{p}$, where $q = 1 - p$.

Variance: $\text{Var}(X) = \dfrac{k(1-p)}{p^2} = \dfrac{kq}{p^2}$, where $q = 1 - p$.

Sum: If $X \sim \text{NegBin}(k,\ p)$, $Y \sim \text{NegBin}(m,\ p)$, and X and Y are **independent**, then

$$X + Y \sim \text{NegBin}(k + m,\ p).$$

Examples of Conditional Expectation and Variance

1. Swimming with dolphins

Fraser runs a dolphin-watch business. Every day, he is unable to run the trip due to bad weather with probability p, independently of all other days. Fraser works every day except the bad-weather days, which he takes as holiday.

Let Y be the number of consecutive days Fraser has to work between bad-weather days. Let X be the total number of customers who go on Fraser's trip in this period of Y days. Conditional on Y, the distribution of X is

$$(X \,|\, Y) \sim \text{Poisson}(\mu Y).$$

(a) Name the distribution of Y, and state $\mathbb{E}(Y)$ and $\text{Var}(Y)$.

(b) Find the expectation and the variance of the number of customers Fraser sees between bad-weather days, $\mathbb{E}(X)$ and $\text{Var}(X)$.

Examples of Conditional Expectation and Variance

(a) *Let 'success' be 'bad-weather day' and 'failure' be 'work-day'.*
Then $\mathbb{P}(\textit{success}) = \mathbb{P}(\textit{bad-weather}) = p$.
Y *is the number of failures before the first success.*
So

$$Y \sim \textit{Geometric}(p).$$

Thus

$$\mathbb{E}(Y) = \frac{1-p}{p},$$

$$\textit{Var}(Y) = \frac{1-p}{p^2}.$$

(b) *We know* $(X \mid Y) \sim \textit{Poisson}(\mu Y)$*: so*

$$\mathbb{E}(X \mid Y) = \textit{Var}(X \mid Y) = \mu Y.$$

Examples of Conditional Expectation and Variance

By the Law of Total Expectation:

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}_Y\Big\{\mathbb{E}(X \mid Y)\Big\} \\ &= \mathbb{E}_Y(\mu Y) \\ &= \mu\mathbb{E}_Y(Y)\end{aligned}$$

$$\therefore \quad \mathbb{E}(X) = \frac{\mu(1-p)}{p}\,.$$

By the Law of Total Variance:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}_Y\Big(\mathrm{Var}(X \mid Y)\Big) + \mathrm{Var}_Y\Big(\mathbb{E}(X \mid Y)\Big) \\ &= \mathbb{E}_Y\Big(\mu Y\Big) + \mathrm{Var}_Y\Big(\mu Y\Big) \\ &= \mu\mathbb{E}_Y(Y) + \mu^2\,\mathrm{Var}_Y(Y) \\ &= \mu\left(\frac{1-p}{p}\right) + \mu^2\left(\frac{1-p}{p^2}\right) \\ &= \frac{\mu(1-p)(p+\mu)}{p^2}\,.\end{aligned}$$

Examples of Conditional Expectation and Variance

Checking your answer in R:

If you know how to use a statistical package like R, you can check your answer to the question above as follows.

```
> # Pick a value for p, e.g. p = 0.2.
> # Pick a value for mu, e.g. mu = 25
>
> # Generate 10,000 random values of Y ~ Geometric(p = 0.2):
> y <- rgeom(10000, prob=0.2)
>
> # Generate 10,000 random values of X conditional on Y:
> # use (X | Y) ~ Poisson(mu * Y) ~ Poisson(25 * Y)
> x <- rpois(10000, lambda = 25*y)

> # Find the sample mean of X (should be close to E(X)):
> mean(x)
[1] 100.6606
>
> # Find the sample variance of X (should be close to var(X)):
> var(x)
[1] 12624.47
>
> # Check the formula for E(X):
> 25 * (1 - 0.2) / 0.2
[1] 100
>
> # Check the formula for var(X):
> 25 * (1 - 0.2) * (0.2 + 25) / 0.2^2
[1] 12600
```

The formulas we obtained by working give $\mathbb{E}(X) = 100$ and $\text{Var}(X) = 12600$. The sample mean was $\overline{x} = 100.6606$ (close to 100), and the sample variance was 12624.47 (close to 12600). Thus our working seems to have been correct.

Reference

https://bookdown.org/probability/beta/conditional-probability.html