RANDOM VARIABLES

Documents prepared for use in courses C22.0103 and B01.1305,
New York University, Stern School of Business

edit date 2012.FEB.02

Cover photo:
Hawk, Washington Square Park,
New York, 2012.JAN.25.

Random variables are used to describe the outcomes of situations subject to randomness. Here “situations” could refer to games of chance, sales of a business in one week, biodata on a person, and many others. A random variable could be observable many times (as in recording sales for week after week after week) or might be observable only once (as in the outcome of a single election). Random variables are usually described with upper-case letters, and their possible values are usually described with lower-case letters.

There is clear interest in the outcomes of random variables. The descriptions $X = 14.6$ or $Y > 200$ or $-1.0 \le Z \le -0.4$ provide important facts. From the analyst’s perspective, there is a great need to describe the probabilities of outcomes. What, for example, is $P[\ X = 14.6\ ]$? Can we express $P[\ X = x\ ]$ as a simple function of x? In working from the probability descriptions, it will be possible to say many things about random variables, *even before data are collected*.

> In the expression $P[\ X = x\ ]$, the upper-case X is the random variable itself and represents the phenomenon, which might be the number of orders in one hour at Delaware Deli. The lower-case x is a stand-in for a possible value. If we have $x = 15$, we are asking $P[\ X = 15\ ]$. The symbol x is an algebra symbol in the conventional sense.

Random variables are divided into these two broad categories:

> Discrete random variables are obtained by counting and have values for which there are no in-between values. These values are typically the integers 0, 1, 2, …. These are described by their probability functions $P[\ X = x\ ]$. These functions are also called *probability mass functions*.
>
> Continuous random variables are obtained by measuring. Between any two possible values are other possible values. If Y is a height in inches, there are certainly values between 57 inches and 58 inches. This property holds even if we happen to round our data to the nearest inch, quarter inch, or even 0.001 inch.

Random variables are usually denoted by upper case (capital) letters. The possible values are denoted by the corresponding lower case letters, so that we talk about events of the form [X = x]. The random variables are described by their probabilities. For example, consider random variable X with probabilities

x	0	1	2	3	4	5
P[$X = x$]	0.05	0.10	0.20	0.40	0.15	0.10

You can observe that the probabilities sum to 1.

The notation $P(x)$ is often used for P[$X = x$]. The notation $f(x)$ is also used. In this example, $P(4) = 0.15$. The symbol P (or f) denotes the probability function, also called the probability mass function.

The *cumulative* probabilities are given as $F(x) = \sum_{i \le x} P(i)$. The interpretation is that $F(x)$ is the probability that X will take a value less than or equal to x. The function F is called the cumulative distribution function (CDF). This is the only notation that is commonly used. For our example,

$$F(3) \quad = \mathrm{P}[X \le 3] \quad = \mathrm{P}[X{=}0] + \mathrm{P}[X{=}1] + \mathrm{P}[X{=}2] + \mathrm{P}[X{=}3]$$

$$= \quad 0.05 \quad + \quad 0.10 \quad + \quad 0.20 \quad + \quad 0.40 \quad = \quad 0.75$$

One can of course list all the values of the CDF easily by taking cumulative sums:

x	0	1	2	3	4	5
P[$X = x$]	0.05	0.10	0.20	0.40	0.15	0.10
$F(x)$	0.05	0.15	0.35	0.75	0.90	1.00

The values of F increase.

The *expected value* of X is denoted either as E(X) or as μ. It's defined as $\mathrm{E}(X) = \sum_x x\ \mathrm{P}(x) = \sum_x x\ \mathrm{P}[X = x]$. The calculation for this example is

$$\mathrm{E}(X) \quad = \quad 0 \times 0.05 + 1 \times 0.10 + 2 \times 0.20 + 3 \times 0.40 + 4 \times 0.15 + 5 \times 0.10$$

$$= \quad 0.00 \quad + \quad 0.10 \quad + \quad 0.40 \quad + \quad 1.20 \quad + \quad 0.60 \quad + \quad 0.50 \quad = \quad 2.80$$

This is also said to be the mean of the probability distribution of X.

The probability distribution of X also has a standard deviation, but one usually first defines the variance. The variance of X, denoted as Var(X) or σ^2, or perhaps σ_X^2, is

$$\text{Var}(X) = \sum_x (x-\mu)^2 \ \text{P}(x) \ = \ \sum_x (x-\mu)^2 \ \text{P}(X = x)$$

This is the expected square of the difference between X and its expected value, μ. We can calculate this for our example:

x	x – 2.8	(x – 2.8)^2	P[X = x]	(x – 2.8)^{2}P[X = x]
0	-2.8	7.84	0.05	0.392
1	-1.8	3.24	0.10	0.324
2	-0.8	0.64	0.20	0.128
3	0.2	0.04	0.40	0.016
4	1.2	1.44	0.15	0.216
5	2.2	4.84	0.10	0.484

The variance is the sum of the final column. This value is 1.560.

This is *not* the way that one calculates the variance, but it does illustrate the meaning of the formula. There’s a simplified method, based on the result $\sum_x (x-\mu)^2 \ \text{P}[X = x] \ = \ \left\{ \sum_x x^2 \ \text{P}[X = x] \right\} - \mu^2$. This is easier because we’ve already found μ, and the sum $\sum_x x^2 \ \text{P}[X = x]$ is fairly easy to calculate because the x values here are small integers. For our example, this sum is

$$0^2 \times 0.05 + 1^2 \times 0.10 + 2^2 \times 0.20 + 3^2 \times 0.40 + 4^2 \times 0.15 + 5^2 \times 0.10 = 9.40$$

Then $\sum_x (x-\mu)^2 \ \text{P}[X = x] \ = \ 9.40 - 2.8^2 \ = \ 9.40 - 7.84 \ = \ 1.56$. This is the same number as before, although obtained with rather less effort.

The standard deviation of X is determined from the variance. Specifically, SD(X) = σ = $\sqrt{\text{Var}(X)}$. In this situation, we find simply σ = $\sqrt{1.56} \approx 1.2490$.

It should be noted that random variables also obey, at least approximately, a variant on the empirical rule used with data. Specifically, for a random variable X with mean μ and standard deviation σ, we have

$$\text{P}[\ \mu - \sigma \ \le \ X \ \le \ \mu + \sigma \ ] \ \approx \ \tfrac{2}{3}$$

$$\text{P}[\ \mu - 2\sigma \ \le \ X \ \le \ \mu + 2\sigma \ ] \ \approx \ 95\%$$

1. Suppose that you are rolling a die eight times. Find the probability that the face with two spots comes up exactly twice.

SOLUTION: Let X be the number of "successes," meaning the number of times that the face with two spots comes up. This is a binomial situation with $n = 8$ and $p = \frac{1}{6}$. The probability of exactly two successes is $\text{P}[\ X = 2\ ] = \binom{8}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^6 = 28 \times \frac{5^6}{6^8}$. This can be done with a calculator. There are various strategies to organize the arithmetic, but the answer certainly comes out as about 0.260476.

> Some calculators have keys like $\boxed{x^y}$, and these can be useful to calculate expressions of the form 6^8. Of course, 6^8 can always be calculated by careful repeated multiplication. The calculator in Microsoft Windows will find 6^8 through the keystrokes 6, *y*, 8, =.

2. The probability of winning at a certain game is 0.10. If you play the game 10 times, what is the probability that you win at most once?

SOLUTION: Let X be the number of winners. This is a binomial situation with $n = 10$ and $p = 0.10$. We interpret "win at most once" as meaning "$X \leq 1$." Then

$$\text{P}[\ X \leq 1\ ] = \text{P}[\ X = 0\ ] + \text{P}[\ X = 1\ ] = \binom{10}{0} 0.10^0 \times 0.90^{10} + \binom{10}{1} 0.10^1 \times 0.90^9$$

$$= 0.90^{10} + 10 \times 0.10^1 \times 0.90^9 = 0.90^{10} + 0.90^9 = 0.90^9 \,(\ 0.90 + 1\ )$$

$$= 0.90^9 \times 1.90 \approx 0.736099$$

3. If X is binomial with parameters n and p, find an expression for $\text{P}[\ X \leq 1\ ]$.

SOLUTION: This is the same as the previous problem, but it's in a generic form.

$$\text{P}[\ X \leq 1\ ] = \text{P}[\ X = 0\ ] + \text{P}[\ X = 1\ ] = \binom{n}{0} p^0 \left(1 - p\right)^n + \binom{n}{1} p^1 \left(1 - p\right)^{n-1}$$

$$= \left(1 - p\right)^n + np\left(1 - p\right)^{n-1} = \left(1 - p\right)^{n-1}\left(\left(1 - p\right) + np\right)$$

$$= \left(1 - p\right)^{n-1}\left(1 + \left(n - 1\right) p\right)$$

4. The probability is 0.038 that a person reached on a “cold call” by a telemarketer will make a purchase. If the telemarketer calls 40 people, what is the probability that at least one sale will result?

SOLUTION: Let X be the resulting number of sales. Certainly X is binomial with $n = 40$ and $p = 0.038$. This “at least one” problem can be done with this standard trick:

$$P[\ X \geq 1\ ] \ = \ 1 - P[\ X = 0\ ] \ = \ 1 - \binom{40}{0} 0.038^0 \times 0.962^{40} \ = \ 1 \ - \ 0.962^{40}$$

$$\approx \ 0.787674.$$

5. The probability is 0.316 that an audit of a retail business will turn up irregularities in the collection of state sales tax. If 16 retail businesses are audited, find the probability that

(a) exactly 5 will have irregularities in the collection of state sales tax.
(b) at least 5 will have irregularities in the collection of state sales tax.
(c) fewer than 5 will have irregularities in the collection of state sales tax.
(d) at most 5 will have irregularities in the collection of state sales tax.
(e) more than 5 will have irregularities in the collection of state sales tax.
(f) no more than 5 will have irregularities in the collection of state sales tax.
(g) no fewer than 5 will have irregularities in the collection of state sales tax.

SOLUTION: Let X be the number of businesses with irregularities of this form. Note that X is binomial with $n = 16$ and $p = 0.316$. The calculations requested here are far too ugly to permit hand calculation, so a program like Minitab should be used.

(a) asks for $P[\ X = 5\ ] \ = \ \binom{16}{5} 0.316^5 \times 0.684^{11} \ = \ 0.2110$. This was done by Minitab.

(b) asks for $P[\ X \geq 5\ ]$. Use Minitab to get $1 - P[\ X \leq 4\ ] \ = \ 1 - 0.3951 = 0.6049$.

(c) asks for $P[\ X \leq 4\ ] = 0.3951$.

(d) asks for $P[\ X \leq 5\ ] = 0.6062$.

(e) asks for $P[\ X > 5\ ]$. Use Minitab to get $1 - P[\ X \leq 5\ ] = 1 - 0.6062 = 0.3938$.

(f) asks for $P[\ X \leq 5\ ] = 0.6062$. This is the same as (d).

(g) asks for $P[\ X \geq 5\ ] = 0.6049$. This is the same as (b).

6. A certain assembly line produces defects at the rate 0.072. If you observe 100 items from this list, what is the smallest number of defects that would cause a 1% rare-event alert? Specifically, if X is the number of defects, find the smallest value of k for which $P[\ X \geq k\ ] \leq 0.01$.

SOLUTION: This will require an examination of the cumulative probabilities for X. Since Minitab computes cumulative probabilities in the ≤ form, rephrase the question as searching for the smallest k for which P[$X \leq k$-1] ≥ 0.99. Here is a set of cumulative probabilities calculated by Minitab:

x	P[$X \leq x$]
11	0.94417
12	0.97259
13	0.98751
14	0.99471
15	0.99791

The first time that the cumulative probabilities cross 0.99 occurs for x = 14. This corresponds to k-1, so we report that k = 15. The smallest number of defects which would cause a 1% rare-event alert is 15.

7. If you flip a fair coin 19 times, what is the probability that you will end up with an even number of heads?

SOLUTION: Let X be binomial with n = 19 and $p = \frac{1}{2}$. This seems to be asking for P[X = 0] + P[X = 2] + P[X = 4] + …. + P[X = 18], which is an annoying calculation. However, we've got a trick. Consider the first 18 flips. The cumulative number of heads will either be even or odd. If it's even, then the 19th flip will preserve the even total with probability $\frac{1}{2}$. If it's odd, then the 19th flip will convert it to even with probability $\frac{1}{2}$. At the end, our probability of having an even number of heads must be $\frac{1}{2}$. This trick *only* works when $p = \frac{1}{2}$.

8. Suppose that you are playing roulette and betting on a single number. Your probability of winning on a single turn is $\frac{1}{38} \approx 0.026316$. You would like to get at least three winners. Find the minimum number of turns for which the probability of three or more winners is at least 0.80.

SOLUTION: The problem asks for the smallest n for which P[$X \geq 3$] ≥ 0.80. Since Minitab computes cumulative probabilities in the ≤ form, we'll convert this question to finding the smallest n for which P[$X \leq 2$] ≤ 0.20.

This now requires a trial-and-error search. It helps to set up a diagram in which the first row is for an n that's likely to be too small and a last row for an n that is likely to be too big.

n	P[$X \leq 2$]	value of n is
20	0.9851	too small
200	0.1011	too large

Then intervening positions can be filled in. Let’s try n = 100. This would result in P[$X \leq 2$] = 0.5084, revealing that n = 100 is too small. The table gets modified to this:

n	P[$X \leq 2$]	value of n is
20	0.9851	too small
100	0.5084	too small
200	0.1011	too large

Now try n = 150, getting P[$X \leq 2$] = 0.2420. This says that n = 150 is too small, but not by much. Here’s what the table looks like:

n	P[$X \leq 2$]	value of n is
20	0.9851	too small
100	0.5084	too small
150	0.2420	too small
200	0.1011	too large

After a little more effort, we get to this spot:

n	P[$X \leq 2$]	value of n is
20	0.9851	too small
100	0.5084	too small
150	0.2420	too small
160	0.2050	too small
161	0.2016	too small
162	0.1982	just right!
200	0.1011	too large

For n = 162, the cumulative probability drops below 0.20 for the first time. This is the requested number of times to play the game.

Recall that the binomial coefficient $\binom{n}{r}$ is used to count the number of possible selections of *r* things out of *n*. Using *n* = 6 and *r* = 2 would provide the number of possible committees that could be obtained by selecting two people out of six.

The computational formula is $\binom{n}{r} = \frac{n!}{r!\,(n-r)!}$. For *n* = 6 and *r* = 2, this would give $\binom{6}{2} = \frac{6!}{2! \times 4!} = \frac{6\times5\times4\times3\times2\times1}{(2\times1)\ \times\ (4\times3\times2\times1)} = \frac{6\times5}{2\times1}$ = 15. If the six people are named *A*, *B*, *C*, *D*, *E*, and *F*, these would be the 15 possible committees:

A B
A C
A D
A E
A F
B C
B D
B E
B F
C D
C E
C F
D E
D F
E F

There are a few useful manipulations:

0! = 1 (by agreement)

$$\binom{n}{r} = \binom{n}{n-r}$$

In the committee example, $\binom{6}{2} = \binom{6}{4}$, so that the number of selections of two people to be on the committee is exactly equal to the number of selections of four people to *leave off* the committee.

$$\binom{n}{0} = \binom{n}{n} = 1$$

$$\binom{n}{1} = n$$

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

The hypergeometric distribution applies to the situation in which a random selection is to be made from a finite set. This is most easily illustrated with drawings from a deck of cards. Suppose that you select five cards at random from a standard deck of 52 cards.

> This description certainly suggests a card game in which five cards are dealt to you from a standard deck. The game of poker generally begins this way. The fact that cards will also be dealt to other players does not influence the probability calculations, as long as the identities of those cards are not known to you.

You would like to know the probability that your five cards (your "hand") will include exactly two aces. The computational logic proceeds along these steps:

* There are $\binom{52}{5}$ possible selections of five cards out of 52. These selections are equally likely.

* There are four aces in the deck, and there are $\binom{4}{2}$ ways in which you can identify two out of the four.

* There are 48 non-aces in the deck, and there are $\binom{48}{3}$ ways in which you can identify three of these non-aces.

* The number of possible ways that your hand can have exactly two aces and exactly three non-aces is $\binom{4}{2} \times \binom{48}{3}$. This happens because every selection of the aces can be matched with every selection of the non-aces.

* The probability that your hand will have exactly two aces is $\dfrac{\binom{4}{2} \times \binom{48}{3}}{\binom{52}{5}}$.

The computation is not trivial. If you are deeply concerned with playing poker, the number $\binom{52}{5}$ will come up often. It's $\dfrac{52!}{5! \times 47!} = \dfrac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1} = 2{,}598{,}960$.

Now note that $\binom{4}{2} = 6$, $\binom{48}{3} = \dfrac{48 \times 47 \times 46}{3 \times 2 \times 1} = 17{,}296$. Then you can find the desired probability $\dfrac{6 \times 17{,}296}{2{,}598{,}960} \approx 0.039930$. This is about 4%.

This technology can be generalized in a useful random variable notation. We will let random X be the number of special items in a sample of n taken at random from a set of N.

> You will sometimes see a distinction between "sampling with replacement" and "sampling without replacement." The issue comes down to whether or not each sampled object is returned to the set of N before the next selection is made. (Returning an item to the set of N would make it possible for that item to appear in the sample more than once.) In virtually all applications, the sampling is *without* replacement.
>
> The sampling is usually done sequentially, but it does not have to be. In a card game, the same probabilities would apply even if you were given all your cards in a single clump from the top of a well-shuffled deck. Of course, dealing out cards in clumps violates the etiquette of the game.

The process is sometimes described as "taking a sample of n from a finite population of N." We then use these symbols:

N	population size
n	sample size
M	number of special items in the population
$N - M$	number of non-special items in the population
X	(random) number of special items in the sample

This table lays out the notation:

	General notation	Card example
Population size	N	52 (cards in deck)
Sample size	n	5 (cards you will be dealt)
Special items in the population	M	4 (aces in the deck)
Non-special items in the population	$N - M$	48 (non-aces in the deck)
Random number of special items in the sample	X	Number of special items in your hand (we asked for this to be 2 in the example)

The probability structure of X is given by the hypergeometric formula

$$\text{P}[\ X = x\ ] \;=\; \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

For our example, this was $\dfrac{\binom{4}{2}\times\binom{48}{3}}{\binom{52}{5}}$.

In a well-formed hypergeometric calculation

the numerator upper numbers add to the denominator upper number
(as 4 + 48 = 52)

the numerator lower numbers add to the denominator lower number
(as 2 + 3 = 5)

As an interesting curiousity, it happens that we can write the hypergeometric probability in the alternate form

$$P[\, X = x \,] \;=\; \frac{\binom{n}{x}\binom{N-n}{M-x}}{\binom{N}{M}}$$

In our example, this would be $\dfrac{\binom{4}{2}\times\binom{48}{3}}{\binom{52}{5}} = \dfrac{\binom{5}{2}\times\binom{47}{2}}{\binom{52}{4}}$.

The program Minitab, since release 13, can compute hypergeometric probabilities. Suppose that you would like to see the probabilities associated with the number of spades that you get in a hand of 13 cards. The game of bridge starts out by dealing 13 cards to each player, so this question is sometimes of interest to bridge players.

In column 1 of Minitab, lay out the integers 0 through 13. You can enter these manually, or you can use this little trick:

Calc ⇒ Make Patterned Data ⇒ Simple Set of Numbers

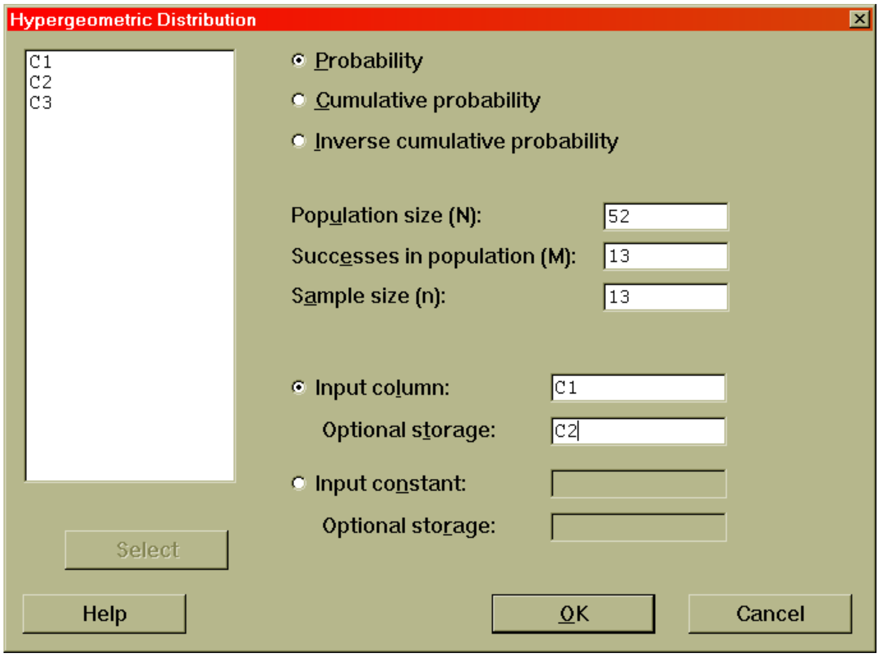
On the resulting panel, enter the information indicated. . .

Store patterned data in:	(enter C1)
From first value:	(enter 0)
To last value:	(enter 13)

Minitab can then easily find all the probabilities at once:

Calc ⇒ Probability Distributions ⇒ Hypergeometric

Fill out the resulting panel to look like this:



Hypergeometric Distribution

C1
C2
C3

● Probability
○ Cumulative probability
○ Inverse cumulative probability

Population size (N): 52
Successes in population (M): 13
Sample size (n): 13

● Input column: C1
Optional storage: C2
○ Input constant:
Optional storage:

The first 13, successes in population, corresponds to the symbol *M*. The second 13, sample size, corresponds to our symbol *n*.

When you click **OK**, the entire column of probabilities appears in column C2.

Here are some typical problems.

Example 1: What is the most likely number of spades that you will get in a hand of 13 cards?

Solution: If you examine the output that Minitab produced, you'll see

P[X = 2] = 0.205873
P[X = 3] = 0.286330
P[X = 4] = 0.238608

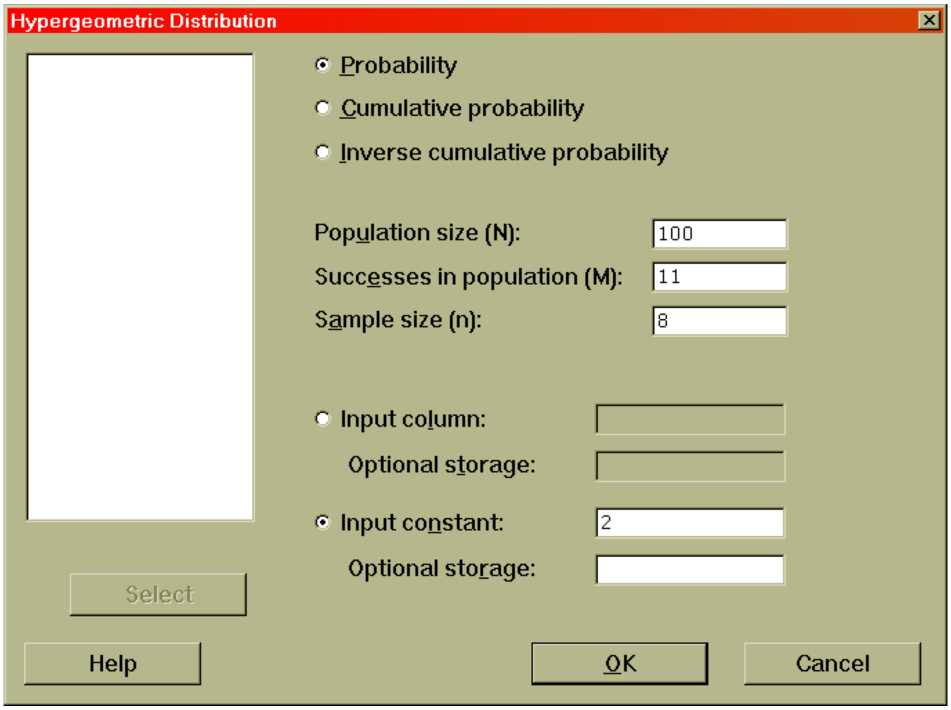
All the other probabilities are much smaller. Thus, you're most likely to get three spades.

Example 2: Suppose that a shipment of 100 fruit crates has 11 crates in which the fruit shows signs of spoilage. A quality control inspection selects 8 crates at random, opens these selected crates, and then counts the number (out of 8) in which the fruit shows signs of spoilage. What is the probability that exactly two crates in the sample show signs of spoilage?

Solution: Let X be the number of bad crates in the sample. This is a hypergeometric random variable with N = 100, M = 11, n = 8, and we ask P[X = 2]. This probability is

$$\frac{\binom{11}{2}\binom{89}{6}}{\binom{100}{8}}$$

The arithmetic is possible, but it's annoying. Let's use Minitab for this. The detail panel should be this:


Hypergeometric Distribution

● Probability
○ Cumulative probability
○ Inverse cumulative probability

Population size (N): 100
Successes in population (M): 11
Sample size (n): 8

○ Input column:
Optional storage:

● Input constant: 2
Optional storage:

Minitab will produce this information in its session window:

Probability Density Function

```
Hypergeometric with N = 100, M = 11, and n = 8

x  P( X = x )
2    0.171752
```

The requested probability is 0.171752.

💻💻💻💻💻💻💻💻💻 THE POISSON RANDOM VARIABLE 💻💻💻💻💻💻💻💻💻

The Poisson random variable is obtained by counting outcomes. The situation is not governed by a pre-set sample size, but rather we observe over a specified length of time or a specified spatial area. There is no conceptual upper limit to the number of counts that we might get. The Poisson would be used for

The number of industrial accidents in a month
The number of earthquakes to strike Turkey in a year
The number of maple seedlings to sprout in a 10 m × 10 m patch of meadow
The number of phone calls arriving at your help desk in a two-hour period

The Poisson has some similarities to the binomial and hypergeometric, so we'll lay out the essential differences in this table:

	Binomial	Hypergeometric	Poisson
Number of trials	n	n	no concept of sample size
Population size	Infinite (trials could go on indefinitely)	N	no concept of population size
Event probability	p	$\frac{M}{N}$	no concept of event probability
Event rate	no concept of event rate	no concept of event rate	λ

The Poisson probability law is governed by a rate parameter λ. For example, if we are dealing with the number of industrial accidents in a month, λ will represent the expected rate. If X is this random variable, then the probability law is

$$\text{P}[\ X = x\ ] \ = \ e^{-\lambda} \frac{\lambda^x}{x!}$$

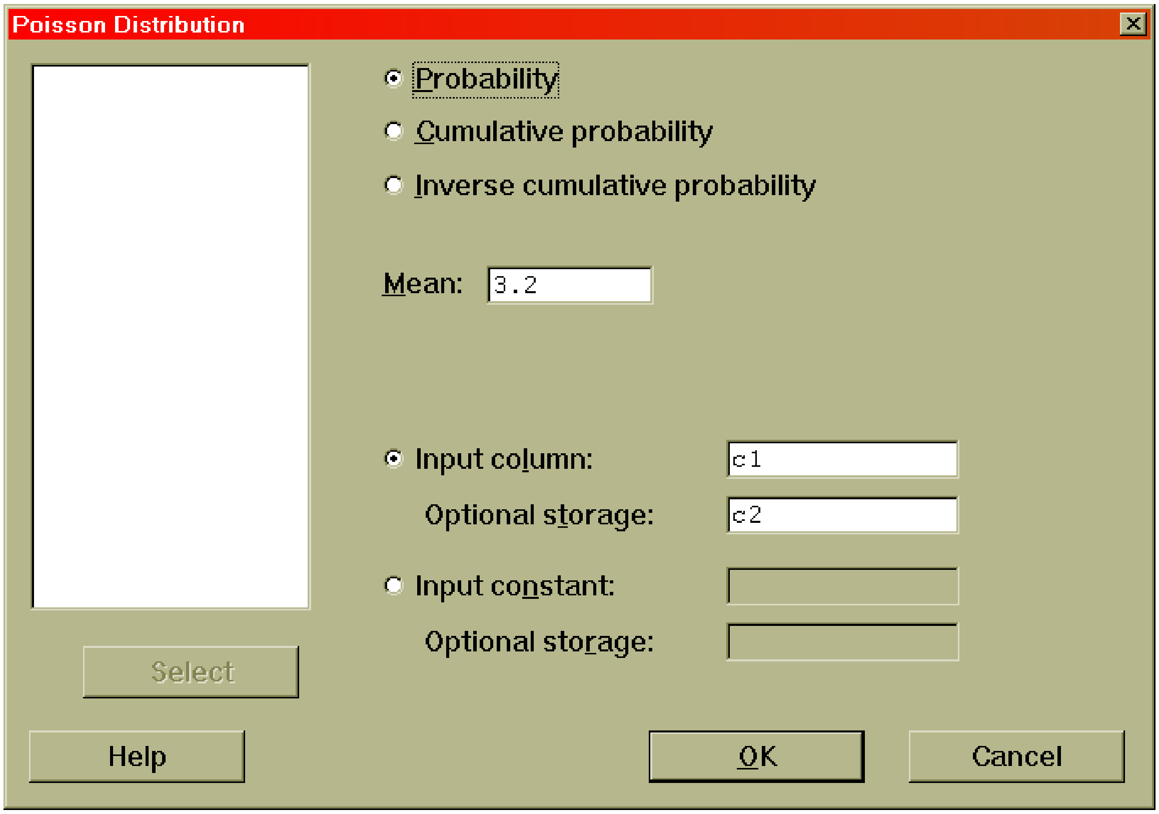
This calculations can be done for x = 0, 1, 2, 3, 4, … There is no upper limit.

If the rate is 3.2 accidents/month, then the probability that there will be exactly two accidents in any month is

$$\text{P}[\ X = 2\ ] \ = \ e^{-3.2} \frac{3.2^2}{2!} \ \approx \ 0.040762 \ \frac{10.24}{2} \ \approx \ 0.2087$$

Minitab can organize these calculations easily. In a column of the data sheet, say C1, enter the integers 0, 1, 2, 3, 4, …., 10. It's easy to enter these directly, but you could also use **Calc ⇒ Make Patterned Data**. Then do **Calc ⇒ Probability Distributions ⇒ Poisson**.

The information panel should then be filled as indicated:



Poisson Distribution

● Probability
○ Cumulative probability
○ Inverse cumulative probability

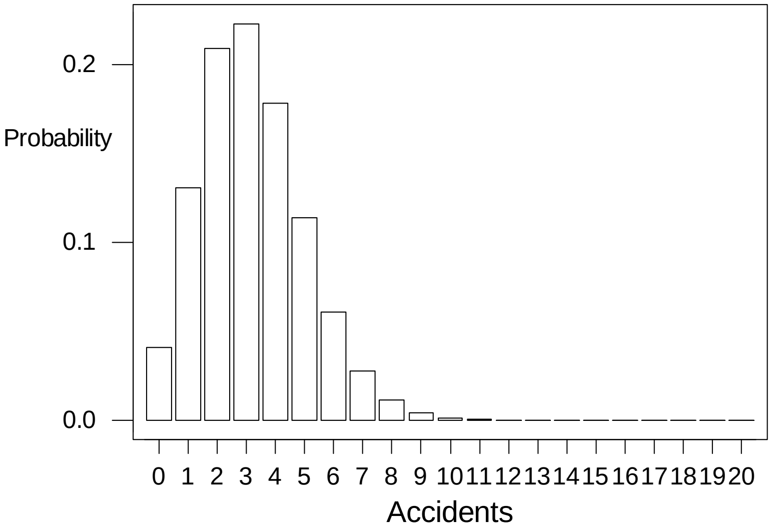
Mean: 3.2

● Input column: c1
Optional storage: c2
○ Input constant:
Optional storage:

The probabilities that result from this operation are these:

Accidents	Probability
0	0.040762
1	0.130439
2	0.208702
3	0.222616
4	0.178093
5	0.113979
6	0.060789
7	0.027789
8	0.011116
9	0.003952
10	0.001265

Here is a graph of the probability function for this Poisson random variable:



This is drawn all the way out to 20 accidents, but it's clear that nearly all the probability action is below 12.

EXAMPLE: The number of calls arriving at the Swampside Police Station follows a Poisson distribution with rate 4.6/hour. What is the probability that exactly six calls will come between 8:00 p.m. and 9:00 p.m.?

SOLUTION: Let X be the random number arriving in this one-hour time period. We'll use $\lambda = 4.6$ and then find P[$X = 6$] = $e^{-4.6} \frac{4.6^6}{6!} \approx 0.1323$.

EXAMPLE: In the situation above, find the probability that exactly 7 calls will come between 9:00 p.m. and 10:30 p.m.

SOLUTION: Let Y be the random number arriving during this 90-minute period. The Poisson rate parameter expands and contracts appropriately, so the relevant value of λ is $1.5 \times 4.6 = 6.9$. We find P[$Y = 7$] = $e^{-6.9} \frac{6.9^7}{7!} \approx 0.1489$.

The Poisson random variable has an expected value that is exactly λ. The standard deviation is $\sqrt{\lambda}$.

EXAMPLE: If X is a Poisson random variable with $\lambda = 225$, would it be unusual to get a value of X which is less than 190?

SOLUTION: If we were asked for the exact number, we'd use Minitab to find $P[\ X \le 189\ ] \approx 0.0077$. This suggests that indeed it would be unusual to get an X value below 190. However, we can get a quick approximate answer by noting that E X = mean of $X = \lambda = 225$, and SD(X) = $\sqrt{\lambda} = \sqrt{225} = 15$. The value 190 is $\frac{225-190}{15} \approx 2.33$ standard deviations below the mean; yes, it would be unusual to get a value that small.

This chart summarizes some relevant facts for the three useful discrete random variables.

Random variable	Description	Probability function $P[\ X = x\ ]$	Expected value (Mean)	Standard deviation
Binomial	Number of successes in n independent trials, each having success probability p	$\binom{n}{x} p^x (1-p)^{n-x}$	np	$\sqrt{np(1-p)}$
Hyper-geo-metric	Number of special items obtained in a sample of n from a population of N containing M special items	$\dfrac{\binom{M}{x} \times \binom{N-M}{n-x}}{\binom{N}{n}}$	$n\dfrac{M}{N}$	$\sqrt{n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}}$
Poisson	Number of events observed over a specified period of time (or space) at event rate λ	$e^{-\lambda}\dfrac{\lambda^x}{x!}$	λ	$\sqrt{\lambda}$

☝ ☝ ☝ COVARIANCE FOR DISCRETE RANDOM VARIABLES ☝ ☝ ☝

Suppose that the two random variables *X* and *Y* have this probability structure:

	Y = 1	*Y* = 2
X = 8	0.12	0.18
X = 10	0.20	0.40
X = 12	0.02	0.08

We can check that the probability sums to 1. The easiest way to do this comes in appending one row and one column to hold totals:

	Y = 1	*Y* = 2	Total
X = 8	0.12	0.18	0.30
X = 10	0.20	0.40	0.60
X = 12	0.02	0.08	0.10
Total	0.34	0.66	1.00

Thus P(*Y* = 1) = 0.34 and P(*Y* = 2) = 0.66. Then

$$\text{E } Y = 0.34 \times 1 \ + \ 0.66 \times 2 \ = \ 1.66 \ = \ \mu_Y$$

$$\text{E } Y^2 = 0.34 \times 1^2 \ + \ 0.66 \times 2^2 \ = \ 2.98$$

$$\sigma_Y^2 \ = \ \text{Var}(Y) \ = \ \text{E } Y^2 \ - \ (\text{ E } Y)^2 \ = \ 2.98 \ - 1.66^2 \ = \ 2.98 \ - \ 2.7556 \ = 0.2244$$

$$\sigma_Y \ = \ \text{SD}(Y) \ = \ \sqrt{0.2244} \ \approx \ 0.4737$$

Using P(*X* = 8) = 0.30, P(*X* = 10) = 0.60, and P(*X* = 12) = 0.10. Then

$$\text{E } X = 0.30 \times 8 \ + \ 0.60 \times 10 \ + \ 0.10 \times 12 \ = \ 9.6 \ = \ \mu_X$$

$$\text{E } X^2 = 0.30 \times 8^2 \ + \ 0.60 \times 10^2 \ + \ 0.10 \times 12^2 \ = 93.6$$

$$\sigma_X^2 \ = \ \text{Var}(X) \ = \ \text{E } X^2 \ - (\text{ E } X)^2 \ = \ 93.6 \ - \ 9.6^2 \ = \ 93.6 - 92.16 = 1.44$$

$$\sigma_X = \ \text{SD}(X) = \ \sqrt{1.44} \ = \ 1.2$$

☝ ☝ ☝ COVARIANCE FOR DISCRETE RANDOM VARIABLES ☝ ☝ ☝

Let’s introduce the calculation of Covariance(X, Y) = Cov(X, Y). This is defined as

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

but it is more easily calculated as

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

Here E[X Y] =

$$\begin{aligned} &\quad 0.12 \times\ \ 8 \times 1 \quad + \quad 0.18 \times\ \ 8 \times 2 \\ &+\ 0.20 \times 10 \times 1 \quad + \quad 0.40 \times 10 \times 2 \\ &+\ 0.02 \times 12 \times 1 \quad + \quad 0.08 \times 12 \times 2 \qquad = \quad 16.00 \end{aligned}$$

Then Cov(X, Y) = 16 – 9.6 × 1.66 = 16 – 15.936 = 0.064.

We can then find the *correlation* of X and Y as

$$\rho = \text{Corr}(X, Y) \;=\; \frac{\text{Cov}(X,Y)}{\sigma_X\ \sigma_Y} \;=\; \frac{0.064}{1.2 \times 0.4737} \;\approx\; 0.1126$$

≜ ≜ ≜ ≜ TWO PERSPECTIVES ON LOTTERIES ≜ ≜ ≜ ≜

This section considers two different ways of thinking about a game of chance based on tickets or number selection.

As the first example, consider a lottery in which there are 500 tickets. Let's suppose that each ticket costs \$10 and that there is a single \$3,000 prize. Notice that the lottery organizer will make money; that's the whole point of lotteries in the first place.

Suppose that Zoe has purchased 5 tickets. We'd like to find the probability that Zoe will win the \$3,000.

From Zoe's perspective, her purchase has made 5 of the tickets special and left 495 as ordinary. The lottery operation will now select one ticket. The probability that this will be chosen from the 5 special tickets is given by the hypergeometric probability

$$\text{P[Zoe wins]} \;=\; \frac{\binom{5}{1}\binom{495}{0}}{\binom{500}{1}} \;=\; \frac{5 \times 1}{500} \;=\; \frac{1}{100} \;=\; 0.01$$

This thinks of taking a sample of $n = 1$ from a population of $N = 500$; in the population are $M = 5$ special tickets and $N - M = 495$ ordinary tickets. This is a very obvious result.

Now consider this from the standpoint of the lottery operator. For the lottery operator, one ticket is special and 499 are ordinary. Now Zoe's purchase represents five drawings from the set of 500 tickets and the probability that her five drawings manage to capture the special ticket is again hypergeometric, but now given by

$$\text{P[Zoe wins]} \;=\; \frac{\binom{1}{1}\binom{499}{4}}{\binom{500}{5}} \;=\; \frac{1 \times \dfrac{499 \times 498 \times 497 \times 496}{4 \times 3 \times 2 \times 1}}{\dfrac{500 \times 499 \times 498 \times 497 \times 496}{5 \times 4 \times 3 \times 2 \times 1}} \;=\; \frac{5}{500} = 0.01$$

This thinks of taking a sample of $n = 5$ from a population of $N = 500$; in the population are $M = 1$ special ticket and $N - M = 499$ ordinary tickets. These are of course the same.

This example is a little too transparent, so let's extend this to a common KENO-type game in which the player selects ten numbers from the set {1, 2, …, 80} and then the lottery operator selects 20 numbers from the same set. Let X be the number of matches. The player wins according to the number of matches common to both selections.

From the perspective of the player, $M = 10$ numbers are special and $N - M = 70$ are ordinary. In thinking of the probability P[$X = 4$], the player imagines that the lottery operator will select $n = 20$ from the set of $N = 80$, getting 4 of the 10 special numbers and 20 – 4 = 16 of the ordinary numbers. The probability is then

$$\text{P}[\ X = 4\ ] \ = \ \frac{\binom{10}{4}\binom{70}{16}}{\binom{80}{20}} \ \approx \ 0.147319$$

From the perspective of the lottery operator, $M = 20$ numbers are special and $N - M = 60$ are ordinary. A particular player, such as the one we are considering, will be making $n = 10$ selections from the set of $N = 80$. The probability of getting exactly four matches can be thought of as getting 4 of the 20 special numbers and 6 of the 60 ordinary numbers. The value is

$$\text{P}[\ X = 4\ ] \ = \ \frac{\binom{20}{4}\binom{60}{6}}{\binom{80}{10}} \ \approx \ 0.147319$$

This shows that the probabilities will be computed consistently from the two perspectives.

This particular situation for the hypergeometric can be summarized as

$$\text{P}[\ X = x\ ] \ = \ \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \ = \ \frac{\binom{n}{x}\binom{N-n}{M-x}}{\binom{N}{M}}$$

LINEAR COMBINATIONS OF RANDOM VARIABLES

There are many situations in which we needs to consider *linear combinations* of random variables. If $X_1, X_2, X_3 , \ldots, X_n$ are random variables, a linear combination is any expression of the form $T = a_0 + \sum_{i=1}^{n} a_i X_i = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$. In this notation, the symbols $a_0, a_1, \ldots, a_n$ are assumed to be non-random constants. These a_i's may be known numbers or they may be unknown quantities to be handled as ordinary algebra symbols. In some problems the objective is to find values for the a_i's to satisfy some properties.

If all the random variables $X_1, X_2, \ldots, X_n$ are discrete, then T is discrete. If one or more of the X_i's is continuous, then T is continuous also.

The properties which will be discussed here are means, standard deviations, and correlations. This discussion is quite general, and it has nothing to do with whether the random variables are discrete, continuous, or some of each.

Let's suppose that $\text{E } X_i = \mu_i$ is the mean (or expected value) of X_i . The values $\mu_1, \mu_2, \ldots, \mu_n$ may be known numbers or they may be unknown and treated as algebra symbols.

Let's suppose also that $\text{SD}(X_i) = \sigma_i$ is the standard deviation of X_i .

Finally, let σ_{ij} be the covariance of X_i with X_j . Then define $\rho_{ij} = \text{Corr}(X_i , X_j) = \dfrac{\sigma_{ij}}{\sigma_i \, \sigma_j}$ as the correlation of X_i with X_j .

As with the means, the values for the σ_i's, the σ_{ij}'s, and the ρ_{ij}'s may be known or unknown.

We have three important formulas, noted as [1], [2], and [3]. An important formula, less frequently used, is given later as [4].

Formula [1]: $\text{E } T = \mu_T = \text{E}\left(a_0 + \sum_{i=1}^{n} a_i X_i \right) = a_0 + \sum_{i=1}^{n} a_i \mu_i$

In words, the expected value of a linear combination in the X_i's is the same linear combination of the μ_i's.

- This result holds whether the X_i's are statistically independent of each other or not.
- As a technical note (which we worry about only rarely), if some of the X_i's have infinite or undefined expected values, then T will (likely) have an infinite or undefined expected value.
- There is nothing random associated with the a_0 term, but a_0 still appears on the right side of [1].

Example 1a: If $X_1, X_2, \ldots, X_{10}$ are the final amounts in 10 plays at roulette, each time betting one dollar on a color, find $\mathrm{E}\left(\sum_{i=1}^{10} X_i\right)$.

To solve this, you will need P[X_i = -1] = $\frac{10}{19}$ and P[X_i = +1] = $\frac{9}{19}$. This leads to E X_i = μ_i = $\frac{-1}{19} \approx$ -0.0526. For one-dollar bets, the expected yield is about -5.26¢. [1] says that $\mathrm{E}\left(\sum_{i=1}^{10} X_i\right) = \sum_{i=1}^{10} \mathrm{E}\left(X_i\right) = \mathrm{E}\left(X_1\right) + \mathrm{E}\left(X_2\right) + \ldots + \mathrm{E}\left(X_{10}\right) = 10 \times (-0.0526)$ = -0.526. This is, of course, -52.6¢. This used [1] with $a_0 = 0$ and $a_1 = a_2 = \ldots = a_{10} = 1$.

Example 1b: Suppose that you invest \$1,000 in stock A, for which the expected return per dollar invested is 2.4¢ and that you also invest \$3,000 in stock B, for which the expected gain per dollar invested is 3.8¢. (We can describe 2.4¢ per dollar as a 2.4% expected return.)

Find your expected gain in dollars. In this context, "gain" is the amount by which your initial \$4,000 will change. That is,

final amount = initial amount + gain

This is an easy intuitive problem, and the solution is certainly \$1,000 × (0.024) + \$3,000 × (0.038) = \$24 + \$114 = \$138. Let's be careful about the notation, however.

Let A be the random gain from one dollar invested in stock A. We are told that E A = 0.024. Similarly let B be the random gain from one dollar invested in stock B; we know that E B = 0.038. Your gain from the combined investment should be expressed as G = 1,000 A + 3,000 B. Using [1] (with $a_0 = 0$, $a_1 = 1{,}000$, and $a_2 = 3{,}000$) gives E G = \$138.

Formula [2]: If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$\text{SD}(T) = \sigma_T = \text{SD}\left(a_0 + \sum_{i=1}^{n} a_i X_i\right) = \sqrt{\sum_{i=1}^{n} a_i^2 \sigma_i^2}$$

- * This formula requires that the X_i's be independent of each other. Formula [3] below covers the case in which this does not happen.
- * The a_0 term does not appear on the right side of Formula [2].
- * This expression looks a little cleaner in terms of the variance:

$$\text{Var}\left(a_0 + \sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

- * The condition "independent random variables" is sometimes replaced by the weaker condition "uncorrelated random variables." These are not exactly the same thing, as independence implies uncorrelated (but uncorrelated does not imply independence).

Example 2a: If $X_1, X_2, \ldots, X_{10}$ are your final amounts in 10 plays at roulette, each time betting one dollar on a color, find $\text{SD}\left(\sum_{i=1}^{10} X_i\right)$.

This is an extension of Example 1a above. We noted P[X_i = -1] = $\frac{10}{19}$ and P[X_i = +1] = $\frac{9}{19}$. This led to E X_i = μ_i = $\frac{-1}{19}$ ≈ -0.0526.

Now we note also Var(X_i) = $\sigma_i^2 = \text{E}\left[\left(X_i - \mu_i\right)^2\right] = \text{E}\left(X_i^2\right) - \mu_i^2 = 1 - \left(\frac{-1}{19}\right)^2$ = $\frac{360}{361}$ ≈ 0.99722992. Then SD(X_i) = $\sigma_i = \sqrt{\text{Var}\left(X_i\right)} = \sqrt{0.99722992} \approx 0.9986$.

This calculation used several facts.

(1) SD(X_i) = $\sqrt{\text{Var}\left(X_i\right)}$, and it's easy to get Var(X_i).

(2) The step $\text{E}\left[\left(X_i - \mu_i\right)^2\right] = \text{E}\left(X_i^2\right) - \mu_i^2$ is true in general, and it was used here as an easier computational method. The alternative would have been

$$\left(-1 - \left(\frac{-1}{19}\right)\right)^2 \times \frac{10}{19} + \left(+1 - \left(\frac{-1}{19}\right)\right)^2 \times \frac{9}{19}$$

(3) Since the only values for X_i in this problem are -1 and +1, it happens that X_i^2 is always 1. Thus $\text{E}\left(X_i^2\right) = 1$.

Now use [2] with $a_0 = 0$, $a_1 = a_2 = \ldots. = a_{10} = 1$. This will give

$$\text{SD}\left(\sum_{i=1}^{10} X_i\right) = \sqrt{\sum_{i=1}^{10} \sigma_i^2} = \sqrt{10 \times 0.99722992} = \sqrt{9.9722992} \approx 3.1579$$

This is in money units, of course, and it represents about \$3.16.

Example 2b: Suppose that you invest \$1,000 in stock A, for which the expected gain per dollar invested is 2.4¢, with a standard deviation of 6.7¢. Suppose also that you also invest \$3,000 in stock B, for which the expected gain per dollar invested is 3.8¢, with a standard deviation of 8.2¢. Stocks A and B are assumed to have independent gains.

In Example 1b, we found that the expected gain of this investment is \$138. We will now find the standard deviation of the gain.

As in Example 1b, we let $G = 1{,}000\ A + 3{,}000\ B$. Using [2] (with $a_0 = 0$, $a_1 = 1{,}000$, and $a_2 = 3{,}000$) we find

$$\text{SD}(G) = \sqrt{1{,}000^2 \times (0.067)^2 + 3{,}000^2 \times (0.082)^2}$$

$$= \sqrt{4{,}489 + 60{,}516} = \sqrt{65{,}005} \approx 254.96$$

Thus, the standard deviation of this scheme is \$254.96.

Example 2c: Suppose that $X_1, X_2, \ldots, X_n$ are n independent random variables, each from same population. (It's usually said that these X_i's constitute a *random sample*.) Suppose that the population mean is μ and that the population standard deviation is σ. Let $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ be the usual average. Find the mean and standard deviation of $\overline{X}$.

We will use [1] for the mean and [2] for the standard deviation. These are done with $a_0 = 0$, $a_1 = a_2 = \ldots = a_n = \frac{1}{n}$. Formula [1] quickly gives E $\overline{X} = \mu$, so that the mean of the sample average is the same as the mean of any single value and in turn is the mean of the population.

Formula [2] gives us

$$\text{SD}(\overline{X}) \;=\; \sqrt{\left(\tfrac{1}{n}\right)^2 \sigma^2 \;+\; \left(\tfrac{1}{n}\right)^2 \sigma^2 \;+\; \left(\tfrac{1}{n}\right)^2 \sigma^2 \;+\; \ldots \;+\; \left(\tfrac{1}{n}\right)^2 \sigma^2} \;=\; \sqrt{\frac{\sigma^2}{n}} \;=\; \frac{\sigma}{\sqrt{n}}$$

Example 2d: Hank buys a lottery ticket for \$5. The amount that he can win (gain) is the random variable W, with the probability structure

$$P[\ W = 0\ ] = 0.94 \qquad P[\ W = 25\ ] = 0.04 \qquad P[\ W = 100\ ] = 0.02$$

On the same day places a \$20 bet on a football game, and the amount that he can win (gain) is the random variable X, with

$$P[\ X = 0\ ] = 0.54 \qquad P[\ X = 40\ ] = 0.46$$

Hank will have the gain given by $T = W + X$. He's also interested in his final amount for the day, meaning $U = T - 25 \;=\; W + X - 25$. Find the mean and standard deviation of both T and U.

First for W

$$E(W) \;=\; 0 \times 0.94 \;+\; 25 \times 0.04 \;+\; 100 \times 0.02 \;=\; 0 \;+\; 1 \;+\; 2 \;=\; 3 \;= \mu_W$$

$$\text{Var}(W) \;=\; E[\ (W - \mu_W)^2\ ] \;=\; E[\ W^2\ ] \;-\; \mu_W^2$$

$$=\; [0^2 \times 0.94 \;+\; 25^2 \times 0.04 \;+\; 100^2 \times 0.02] \;-\; \mu_W^2$$

$$=\; [\ 0 \;+\; 25 \;+\; 200\ ] \;-\; \mu_W^2 \;=\; 225 \;-\; 3^2 \;=\; 216 \;=\; \sigma_W^2$$

This could have been done also as

$$(0 - 3)^2 \times 0.94 \;+\; (25 - 3)^2 \times 0.04 \;+\; (100 - 3)^2 \times 0.02$$

$$=\; 9 \times 0.94 \;+\; 484 \times 0.04 \;+\; 9{,}409 \times 0.02 \;=\; 216 \;=\; \sigma_W^2$$

$$\text{SD}(W) \;=\; \sigma_W \;=\; \sqrt{216} \;\approx\; 14.6969$$

Then for X

$$\mathrm{E}(X) \;=\; 0 \times 0.54 \;+\; 40 \times 0.46 \;=\; 18.4 \;= \mu_X$$

$$\mathrm{Var}(X) \;=\; \mathrm{E}[\,(X - \mu_X)^2\,] \;=\; \mathrm{E}[\,X^2\,] \;-\; \mu_X^2$$

$$= \;[\,0^2 \times 0.54 \;+\; 40^2 \times 0.46\,] \;-\; \mu_X^2$$

$$= \;[\;0 \;+\; 736\,] \;-\; \mu_X^2 \;=\; 736 - 18.4^2 \;=\; 397.44$$

This could have been done also as

$$(0 - 18.4)^2 \times 0.54 \;+\; (40 - 18.4)^2 \times 0.46$$

$$= \;338.56 \times 0.54 \;+\; 466.56 \times 0.46 \;=\; 397.44 \;=\; \sigma_X^2$$

$$\mathrm{SD}(X) \;=\; \sigma_X \;=\; \sqrt{397.44} \;\approx\; 19.9359$$

These two bets are certainly independent.

Formula [1] gives immediately

$$\mathrm{E}[\,T\,] = \mathrm{E}[\,W + X\,] \;=\; \mathrm{E}[\,W\,] + \mathrm{E}[\,X\,] \;=\; 3 + 18.4 \;=\; 21.4$$

$$\mathrm{E}[\,U\,] \;=\; \mathrm{E}[W + X - 25\,] \;=\; \mathrm{E}[\,W\,] + \mathrm{E}[\,X\,] - 25 \;=\; 3 + 18.4 - 25 = -3.6$$

It is no surprise that E[U] < 0. Hank is playing \$25 in total, and his expected final value is -\$3.60.

Formula [2] gives

$$\mathrm{SD}[\,T\,] = \sigma_T \;=\; \mathrm{SD}[\,W + X\,] \;=\; \sqrt{\sigma_W^2 \;+\; \sigma_X^2} \;=\; \sqrt{216 \;+\; 397.44}$$

$$= \;\sqrt{613.44} \;\approx\; 24.77$$

It happens also that SD[U] = 24.77, since T and U are distinguished only by the -25 summand, and this -25 does not contribute to the standard deviation. Here the -25 plays the role of a_0 in formula [2].

Formula [3]: If $X_1, X_2, \ldots, X_n$ are random variables with Correlation(X_i , X_j) = ρ_{ij} , then

$$\text{SD}(T) = \sigma_T = \text{SD}\left(a_0 + \sum_{i=1}^{n} a_i \, X_i \right)$$

$$= \sqrt{\sum_{i=1}^{n} a_i^2 \, \sigma_i^2 + 2 \sum_{i<j}\sum a_i \, a_j \, \rho_{ij} \, \sigma_i \, \sigma_j} \quad = \sqrt{\sum_{i=1}^{n} a_i^2 \, \sigma_i^2 + \sum_{i \neq j}\sum a_i \, a_j \, \rho_{ij} \, \sigma_i \, \sigma_j}$$

- * If all the ρ_{ij} values are zero, this reduces to Formula [2].
- * The decision between the last two terms depends on whether it's easier to count cases with i less than j or to count cases with $i \neq j$.
- * These forms can also be expressed with Cov(X_i , X_j) = $\sigma_{ij} = \rho_{ij} \, \sigma_i \, \sigma_j$, perhaps as SD(T)

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

$$= \sqrt{\sum_{i=1}^{n} a_i^2 \, \sigma_i^2 + 2 \sum_{i<j}\sum a_i \, a_j \, \sigma_{ij}} \ ,$$

Example 3a: A private lottery has 20 tickets, sold at $100 each. There are three prizes, in amounts $800, $400, and $200. Fran has purchased five tickets, and she uses $X_1, \ldots, X_5$ to represent her gains. Thus, for her first ticket,

P[X_1 = 0] = 0.85 P[X_1 = 200] = 0.05

P[X_1 = 400] = 0.05 P[X_1 = 800] = 0.05

The probability distributions for $X_2, \ldots X_5$ are identical.

If $G = X_1 + \ldots + X_5$ is Fran's total gain, find E(G) and SD(G). In this example, the X_i's are correlated, and part of the problem is finding the correlation.

Find first μ_X = E(X_1) = E(X_2) = … = E(X_5). (Since each X_i has the same distribution, there is no need to put a subscript on the X in the symbol μ_X .) This is

$$0 \times 0.85 + 0.05 \times 200 + 0.05 \times 400 + 0.05 \times 800$$

$$= 0 + 10 + 20 + 40 = 70 = \mu_X$$

It is not a surprise that $\mu_X < 100$, the ticket cost.

Formula [1] gives very quickly E(R) = μ_R = 5 × 70 = 350. Remember that Fran spent $500 for these tickets.

Next identify σ_X = SD(X_1) = SD(X_2) = … = SD(X_5). It's easier to start with Var(X_1). This is

$$\text{Var}(X_1) \;=\; \text{E}\left[\left(X_1 - \mu_X\right)^2\right] \;=\; \text{E}\left(X_1^2\right) \;-\; \mu_X^2$$

$$= \; 0^2 \times 0.85 \;+\; 200^2 \times 0.05 \;+\; 400^2 \times 0.05 \;+\; 800^2 \times 0.05 \;-\; 70^2$$

$$= \; 0 \;+\; 2{,}000 \;+\; 8{,}000 \;+\; 32{,}000 \;-\; 70^2 \;=\; 37{,}100 \;=\; \sigma_X^2$$

This could have been done also as

$$(0 - 70)^2 \times 0.85 \;+\; (200 - 70)^2 \times 0.05$$

$$+ \; (400 - 70)^2 \times 0.05 \;+\; (800 - 70)^2 \times 0.05$$

$$= \; (4{,}900) \times 0.85 \;+\; (16{,}900) \times 0.05$$

$$+ \; (108{,}900) \times 0.05 \;+\; (532{,}900) \times 0.05$$

$$= \; 37{,}100 \;=\; \sigma_X^2$$

Then SD(X_1) = $\sigma_X \;=\; \sqrt{37{,}100} \;\approx\; 192.61$.

In the actual execution of this lottery, there will be 20 slips of paper, each with the name of a ticket purchaser, in a large bowl, and then three of these will be drawn out in sequence. From the other probability perspective, let's imagine instead that the bowl contains 20 tickets with the composition $\left\{ \underbrace{\$0, \$0, ..., \$0}_{\text{17 times}}, \$200, \$400, \$800 \right\}$. Now Fran gets to make five selections from the bowl. In this way of thinking, X_1 represents Fran's first selection, X_2 her second selection, and so on.

The complication in this problem is that the X_i's are not independent. After all, if Fran's first selection gets the \$800 prize, this prize is not going to be available to her other selections. For X_1 through X_5 there are 10 correlations, but by symmetry they must all the same. Let's just find ρ_{12} . This needs the joint distribution of (X_1, X_2). Here are the probabilities:

		X_2			
		0	200	400	800
X_1	0	$\frac{17}{20}\times\frac{16}{19}$	$\frac{17}{20}\times\frac{1}{19}$	$\frac{17}{20}\times\frac{1}{19}$	$\frac{17}{20}\times\frac{1}{19}$
	200	$\frac{1}{20}\times\frac{17}{19}$	0	$\frac{1}{20}\times\frac{1}{19}$	$\frac{1}{20}\times\frac{1}{19}$
	400	$\frac{1}{20}\times\frac{17}{19}$	$\frac{1}{20}\times\frac{1}{19}$	0	$\frac{1}{20}\times\frac{1}{19}$
	800	$\frac{1}{20}\times\frac{17}{19}$	$\frac{1}{20}\times\frac{1}{19}$	$\frac{1}{20}\times\frac{1}{19}$	0

These probabilities are symmetric. For example, the value in the box ($X_1 = 200$, $X_2 = 0$) is the same as that in the box ($X_1 = 0$, $X_2 = 200$).

The covariance Cov(X_1, X_2) = E[(X_1 - μ_X) (X_2 - μ_X)] is needed. It's easiest to calculate from this form:

$$\text{E}[\ (X_1 - \mu_X)\ (X_2 - \mu_X)\ ] \ = \ \text{E}[X_1\, X_2\ ] \ - \ \mu_X^2$$

This is easy because many of the $X_1\, X_2$ products are zero, and the probabilities are equal to the same $\frac{1}{20}\times\frac{1}{19}$ for all the non-zero products. Thus

$$\text{Cov}(X_1, X_2) \ = \ \text{E}[\ (X_1 - \mu_X)\ (X_2 - \mu_X)\ ] \ = \ \text{E}[X_1\, X_2\ ] \ - \ \mu_X^2$$

$$= \begin{bmatrix} 0 & + \ 200\times 400 & + \ 200\times 800 \\ + \ 400\times 200 & + \quad 0 & + \ 400\times 800 \\ + \ 800\times 200 & + \ 800\times 400 & + \quad 0 \end{bmatrix} \times\frac{1}{20}\times\frac{1}{19} \ - \ 70^2$$

$$\approx \ \text{-1,952.6316}$$

This gives Corr(X_1, X_2) = ρ_{12} $= \dfrac{\text{Cov}\left(X_1\,,\,X_2\right)}{\text{SD}\left(X_1\right)\times\text{SD}\left(X_2\right)} \ = \ \dfrac{-1{,}952.3615}{192.61\times 192.61} \ \approx$ -0.0526.

Then Var(G) = Var($X_1 + X_2 + \ldots + X_5$) $= \sum_{i=1}^{5} a_i^2\ \sigma_i^2 \ + \ 2\sum_{i<j}\sum a_i\ a_j\ \sigma_{ij}$, and this is to be used with $a_1 = \ldots = a_5 = 1$, $\sigma_1^2 = \ldots = \sigma_5^2 \ = \ \sigma_X^2 \ = \ 37{,}100$, and σ_{ij} = -1,952.6316.

The result is

$$\text{Var}(G) \;=\; 5 \times 37{,}100 \;\;+\;\; 2 \times 10 \times (-1{,}952.6316) \;=\; 146{,}447.3680$$

This leads to SD(G) $= \sqrt{146{,}447.3680} \approx 382.68$.

Thus, Fran's expected gain is E(G) = 350 (which is less than the $500 she spent on the tickets) and her standard deviation of gain is SD(G) = $382.68.

Example 3b:
This is a continuation of Examples 1b and 2b, except that we now allow the stocks to have correlated gains. Suppose that you invest $1,000 in stock *A*, for which the expected gain per dollar invested is 2.4¢, with a standard deviation of 6.7¢. Suppose also that you also invest $3,000 in stock *B*, for which the expected gain per dollar invested is 3.8¢, with a standard deviation of 8.2¢. The gains here are correlated, with $\rho = 0.28$. Find the mean and standard deviation of the overall gains.

As before, let *A* be the random gain from one dollar invested in stock *A* and let *B* be the random gain from one dollar invested in stock *B*. The gain is $R = 1{,}000\, A \;+\; 3{,}000\, B$. Using [3] (with $a_0 = 0$, $a_1 = 1{,}000$, and $a_2 = 3{,}000$) and with $\sigma_A = 0.067$, $\sigma_B = 0.082$, and $\rho = 0.28$, we get

$$\text{SD}(R) \;=\; \sqrt{\left(1{,}000\right)^2 \sigma_A^2 \;+\; \left(3{,}000\right)^2 \sigma_B^2 \;+\; 2 \times 1{,}000 \times 3{,}000 \times \rho \sigma_A \sigma_B} \;=$$

$$\sqrt{\left(1{,}000\right)^2 \left(0.067\right)^2 + \left(3{,}000\right)^2 \left(0.082\right)^2 + 2 \times 1{,}000 \times 3{,}000 \times 0.28 \times 0.067 \times 0.082}$$

$$= \sqrt{4{,}489 \;+\; 60{,}516 \;+\; 9{,}229.92} \;=\; \sqrt{74{,}234.92} \;\approx\; 272.46$$

This represents $272.46. This is slightly larger than the $254.96 found in Example 2b, in which the stocks were assumed to be uncorrelated. In Example 3b, the stocks were assumed to have a positive correlation, meaning that they tend to move together. As a result, the variability increases.

Finally, we'll note one additional result related to two different linear combinations. In this story, we still have the random variables X_1, X_2, …, X_n . This time we consider two different linear combinations,

$$T = a_0 + \sum_{i=1}^{n} a_i X_i \qquad \text{and} \qquad U = b_0 + \sum_{j=1}^{n} b_j X_j$$

The symbols b_0, b_1, …, b_n are also assumed to be non-random constants. They may be known numbers or they may be treated as ordinary algebra symbols. It is not critical to use the counter j in the definition of U, but it makes the work a little cleaner.

Formula [4]: If X_1, X_2, …, X_n are random variables with Correlation(X_i , X_j) = ρ_{ij} , then

$$\text{Cov}(T, U) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i\, b_j\, \rho_{ij}\, \sigma_i\, \sigma_j = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i\, b_j\, \sigma_{ij}$$

- * Note that a_0 and b_0 do not appear in the covariance.
- * It follows from Formula [4] that Corr(T, U) = $\dfrac{\text{Cov}(T,U)}{\text{SD}(T)\times\text{SD}(U)}$.

Formula [4] applies directly to two different portfolio strategies over the same set of n stocks. If stock 6 is in the T portfolio but not in the U portfolio, then the formula will have $a_6 > 0$ and $b_6 = 0$.

If the two portfolios involve completely different sets of stocks (or in general any non-overlapping sets of random variables), it may be more convenient to use a slightly different setup. Let X_1, X_2, …, X_n be the stocks for the T portfolio and let Y_1, Y_2, …, Y_q be the stocks for the U portfolio. Note the use of q here; in general $n \neq q$. Now consider

$$T = a_0 + \sum_{i=1}^{n} a_i X_i \qquad \text{and} \qquad U = b_0 + \sum_{j=1}^{q} b_j Y_j$$

With this notation, Formula [4] is

$$\text{Cov}(T, U) = \sum_{i=1}^{n}\sum_{j=1}^{q} a_i\, b_j\, \rho_{ij}\, \sigma_i\, \sigma_j = \sum_{i=1}^{n}\sum_{j=1}^{q} a_i\, b_j\, \sigma_{ij}$$