Combinations of random variables

Tan Do

Vietnamese-German University

Lecture 11

In this lecture

- Linear combinations of random variables

Linear functions of a random variable

Let $a, b \in \mathbb{R}$. A function of the form $aX + b$ is called a linear function of the random variable X.

It can be shown that

$$E(aX + b) = aE(X) + b$$

and

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

It follows that

$$\sigma_{aX+b} = |a| \sigma_X.$$

Given a random variable X of mean μ and variance σ^2.

We can apply a linear transformation on X to standardize it to have mean 0 and variance 1 as follows.

Let

$$Y = \frac{X - \mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma}.$$

Then

- $E(Y) = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = 0$
- $\mathrm{Var}(Y) = \frac{1}{\sigma^2}\mathrm{Var}(X) = 1$

Back to the general case where $Y = aX + b$, the cumulative distribution of Y is

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P(aX \leq y - b).$$

- If $a > 0$, then

$$F_Y(y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

- If $a < 0$, then

$$F_Y(y) = P\left(X \geq \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right).$$

Example (Test score standardization) Suppose that the raw scores X from a particular testing procedure are distributed between -5 and 20 with an expected value of 10 and a variance of 7.

In order to *standardize* the scores so that they lie between 0 and 100, the linear transformation

$$Y = 4X + 20$$

is applied to the scores. This means, for example, that a raw score of $x = 12$ corresponds to a standardized score of $y = 4 \times 12 + 20 = 68$.

The expected value of Y is

$$E(Y) = 4E(X) + 20 = 4 \times 10 + 20 = 60.$$

The variance of Y is

$$\mathrm{Var}(Y) = 4^2\,\mathrm{Var}(X) = 16 \times 7 = 112.$$

Example (Chemical reaction temperatures) The temperature X in degrees Fahrenheit of a particular chemical reaction is distributed between 220 and 280 with a probability density function

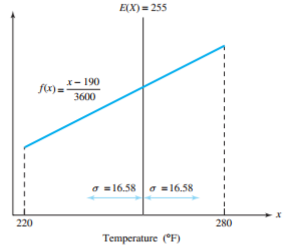
$$f_X(x) = \frac{x - 190}{3600}.$$

Direct calculations give

$$E(X) = 255 \quad \text{and} \quad \mathrm{Var(X)} = 275.$$

In addition,

$$F_X(x) = \int_{220}^{x} f_X(t)\,dt = \frac{(x-190)^2}{7200} - \frac{1}{8}.$$



Suppose that a chemist wishes to convert the temperatures to degrees Centigrade. Let Y be the random variable that measures the reaction temperature in degrees Centigrade, then

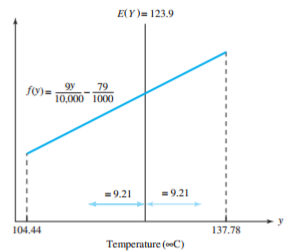
$$Y = \frac{5}{9}X - \frac{160}{9}.$$

Note that $x = 220°$F corresponds to

$$y = \frac{5}{9} \times 220 - \frac{160}{9} = 104.44°\mathrm{C}$$

and $x = 280°$F corresponds to

$$y = \frac{5}{9} \times 280 - \frac{160}{9} = 137.78°\mathrm{C}.$$



Since $a = 5/9$ is positive, the cumulative distribution function of Y is

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right) = F_X\left(\frac{y+160/9}{5/9}\right) = \frac{(9y-790)^2}{180000} - \frac{1}{8}.$$

The probability density function is then

$$f_Y(y) = F'(y) = \frac{9y}{10000} - \frac{79}{1000}$$

for $104.44 \le y \le 137.78$.

The expectation of Y is

$$E(Y) = \frac{5}{9}E(X) - \frac{160}{9} = \frac{5}{9} \times 255 - \frac{160}{9} = 123.9$$

and

$$\mathrm{Var}(Y) = \left(\frac{5}{9}\right)^2 \mathrm{Var}(X) = \left(\frac{5}{9}\right)^2 \times 275 = 84.88.$$

Exercise Re-calculate $E(Y)$ and $\mathrm{Var}(Y)$ using $f_Y(y)$.

Sum of random variables

Let X_1 and X_2 be 2 random variables. Then

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

and

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

where

$$\mathrm{Cov}(X_1, X_2) = E\big((X_1 - E(X_1))\,(X_2 - E(X_2))\big) = E(X_1X_2) - E(X_1)E(X_2).$$

If X_1 and X_2 are independent random variables, then $\mathrm{Cov}(X_1, X_2) = 0$, which in turns implies

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2).$$

Linear combinations of random variables

Let $X_1, X_2, \ldots, X_n$ be random variables.

Let $a_1, a_2, \ldots, a_n$ and b be numbers.

Then

$$E(a_1X_1+a_2X_2+\ldots+a_nX_n+b) = a_1E(X_1)+a_2E(X_2)+\ldots+a_nE(X_n)+b.$$

If $X_1, X_2, \ldots, X_n$ are independent, then

$$\mathrm{Var}(a_1X_1+a_2X_2+\ldots+a_nX_n+b) = a_1^2\mathrm{Var}(X_1)+a_2^2\mathrm{Var}(X_2)+\ldots+a_n^2\mathrm{Var}(X_n).$$

Averaging independent random variables

Suppose that $X_1, X_2, \ldots, X_n$ is a sequence of independent random variables, each of which has expectation μ and variance σ^2.

Let X be the average of $X_1, X_2, \ldots, X_n$, i.e.,

$$X = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Then

$$E(X) = \mu$$

and

$$\mathrm{Var}(X) = \frac{\sigma^2}{n}.$$

Example (Piston head construction) A circular piston head designed to slide smoothly within a cylinder. The company that manufactures them is interested in how well the piston heads actually fit within the cylinders.

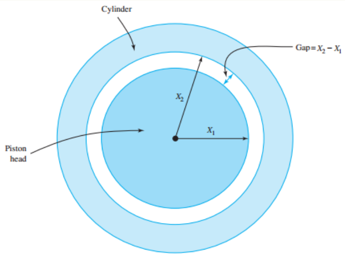
Suppose that the random variable X_1 measures the radius of a piston head, and that it has an expected value of 30 mm and a standard deviation of 0.05 mm.

Also, suppose that the random variable X_2 measures the inside radius of a cylinder, with an expected value of 30.25 mm and a standard deviation of 0.06 mm.

The "gap" between the piston head and the cylinder is $Y = X_2 - X_1$.

The expected value of the gap is

$$E(Y) = E(X_2) - E(X_1) = 30.25 - 30 = 0.25\,\text{mm}$$



We can reasonably assume that the sizes of a piston head and a cylinder are independent. Then

$$\text{Var}(Y) = \text{Var}(X_2 - X_1) = \text{Var}(X_2) + (-1)^2\text{Var}(X_1)$$

$$= 0.06^2 + 0.05^2 = 0.0061.$$

Therefore, $\sigma_Y = \sqrt{0.0061} = 0.078\,\text{mm}.$

Example (Test score standardization) Suppose that in a certain examination procedure, candidates must take 2 tests.
Let X_1 and X_2 measure the score of a candidate on Test 1 and Test 2 respectively.
Suppose that

- the scores on Test 1 are distributed between 0 and 30 with an expectation of 18 and a variance of 24;
- the scores on Test 2 are distributed between -10 and 50 with an expectation of 30 and a variance of 60.

The examining board wishes to standardize each test score to lie between 0 and 100, and then to calculate a final score out of 100 that is weighted 2/3 from Test 1 and 1/3 from Test 2.
The standardized scores of the 2 tests are

$$Y_1 = \frac{10}{3}X_1 \quad \text{and} \quad \mathrm{Y}_2 = \frac{5}{3}\mathrm{X}_2 + \frac{50}{3}.$$

The final score is

$$Z = \frac{2}{3}Y_1 + \frac{1}{3}Y_2 = \frac{20}{9}X_1 + \frac{5}{9}X_2 + \frac{50}{9}.$$

Thus a candidate receiving a score of $x_1 = 11$ on Test 1 and a score of $x_2 = 2$ on Test 2 receives a final score of

$$z = \frac{20}{9} \times 11 + \frac{5}{9} \times 2 + \frac{50}{9} = 31.11.$$

The expected value of the final score is

$$E(Z) = \frac{20}{9}E(X_1) + \frac{5}{9}E(X_2) + \frac{50}{9} = \frac{20}{9} \times 18 + \frac{5}{9} \times 30 + \frac{50}{9} = 62.22.$$

What about the variance of the final score?

To calculate the variance, it is important to check whether the scores on Test 1 and Test 2 are independent or not.

- If Test 1 measures a candidate's proficiency at *probability* and Test 2 measures a candidate's proficiency at *statistics*, then the scores should not be independent.
  So the calculation of the variance of the final score actually requires knowledge of the covariance between the 2 test scores.
- If Test 1 measures a candidate's proficiency at *probability* and Test 2 measures a candidate's proficiency at *athletic* abilities, then the scores can reasonably be assumed to be independent.
  In this case,

$$\mathrm{Var}(Z) = \mathrm{Var}\left(\frac{20}{9}X_1 + \frac{5}{9}X_2 + \frac{50}{9}\right) = \left(\frac{20}{9}\right)^2 \mathrm{Var}(X_1) + \left(\frac{5}{9}\right)^2 \mathrm{Var}(X_2)$$

$$= \left(\frac{20}{9}\right)^2 24 + \left(\frac{5}{9}\right)^2 60 = 137.04.$$

Games of chance

Coin tossing Suppose that a fair coin is tossed 100 times.

Question What are the expectation and the variance of the number of heads obtained?

Answer An easy way to solve this problem is by defining the random variable X_i to be 1 if the ith toss is a head and 0 otherwise.
The number of heads obtained is therefore

$$Y = X_1 + X_2 + \ldots + X_{100}.$$

Exercise For each X_i, show that $E(X_i) = \frac{1}{2}$ and $\mathrm{Var}(X_i) = \frac{1}{4}$.
It follows that

$$E(Y) = E(X_1) + E(X_2) + \ldots + E(X_{100}) = \frac{100}{2} = 50$$

and

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_{100}) = \frac{100}{4} = 25.$$

Next suppose that a die is rolled 10 times and the *average* of the 10 scores is recorded.

The probability mass function of the average score is tedious to compute.

However, we can easily calculate the expectation and the variance.

Recall that the score from the roll of a single die has an expected value of $\mu = 3.5$ and a variance $\sigma^2 = \frac{35}{12}$, so the average score has

- an expected value of 3.5 and
- a variance of $\frac{\sigma^2}{10} = \frac{35}{120} = \frac{7}{24}$.

Nonlinear functions of a random variable

Let X be a random variable. Let g be a nonlinear function.

Then $Y = g(X)$ is a nonlinear function of X.

Some examples are:

- $Y = X^2$.
- $Y = \sqrt{X}$.
- $Y = e^X$.
- $Y = \frac{1}{1+X}$.

Generally, there are no results that relate the expectation and variance of Y to those of X.

The easiest way to construct the probability distribution of Y is to first construct its cumulative distribution function.

Example Let X be a random variable which is distributed between 0 and 1 with $p.d.f$

$$f_X(x) = 1$$

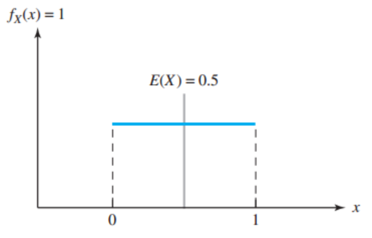
for $0 \leq x \leq 1$ and $f_X(x) = 0$ elsewhere.

Since $f_X(x)$ is symmetric about $x = 0.5$, we deduce that $E(X) = 0.5$.

Also,

$$F_X(x) = x$$

for $0 \leq x \leq 1$.



Now consider $Y = e^X$.

Y takes values between 1 and $e = 2.718$.

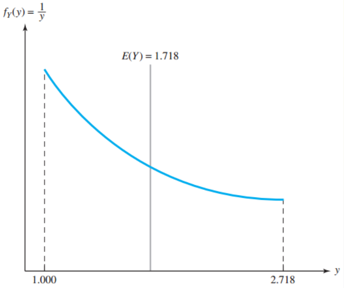
Its cumulative distribution function is

$$F_Y(y) = P(Y \le y) = P(e^X \le y) = P(X \le \ln(y)) = F_X(\ln(y)) = \ln(y).$$

So the probability density function of Y is

$$f_Y(y) = F'(y) = \frac{1}{y}$$

for $1 \le y \le 2.718$.



Using this, we have

$$E(Y) = \int_1^e t\, f_Y(t)\, dt = \int_1^e 1\, dt = e - 1 = 1.718.$$

Notice that

$$E(Y) \ne e^{E(X)} = e^{0.5} = 1.649.$$

Example (Piston head construction) Suppose that X_1, the radius of the piston head, actually takes values between 29.9 mm and 30.1 mm with p.d.f $f_{X_1}(x) = 5$ for $29.9 \leq x \leq 30.1$ and $f_{X_1}(x) = 0$ elsewhere.

The c.d.f of X_1 is

$$F_{X_1}(x) = \int_{29.9}^{x} f_{X_1}(t)\, dt = 5x - 149.5$$

for $29.9 \leq x \leq 30.1$.

The piston manufacturers are particularly interested in the *area* of the piston head since this directly affects the performance of the the piston.

The area of the piston head is

$$Y = \pi X_1^2$$

which takes values between $\pi \times 29.9^2 = 2808.6$ mm^2 and $\pi \times 30.1^2 = 2846.3$ mm^2.

The c.d.f of the pistol head area is

$$F_Y(y) = P(Y \leq y) = P(\pi X_1^2 \leq y) = P(X_1 \leq \sqrt{y/\pi}) = F_{X_1}(\sqrt{y/\pi}) = 2.821\sqrt{y} - 149.5.$$

So the p.d.f is

$$f_Y(y) = F'(y) = \frac{1.41}{\sqrt{y}}$$

for $2808.6 \leq y \leq 2846.3$.