3A Standard deviation and correlation

Class problems

Remember that *mean* is another name for expected value.

3A1 (Correlation versus dependence) Discrete random variables X and Y have the following joint distribution.

	Y		
X	-1	0	1
-1	0	$\frac{1}{6}$	$\frac{1}{6}$
0	$\frac{1}{3}$	0	0
1	0	$\frac{1}{6}$	$\frac{1}{6}$

(a) Determine the distribution, mean and standard deviation of X.

(b) Determine the distribution, mean and standard deviation of Y.

(c) Calculate the correlation between X and Y.

(d) Determine whether X and Y are (stochastically) dependent or independent.

Solution.

(a) From the row sums of the joint distribution, we get the distribution of X as

k	-1	0	1
$\mathrm{P}(X=k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Thus its mean is

$$\mathrm{E}(X) \;=\; \sum_k k\,\mathrm{P}(X=k) \;=\; (-1)\times\frac{1}{3} + 0\times\frac{1}{3} + 1\times\frac{1}{3} \;=\; 0.$$

Let us use the alternative formula $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2$. Because

$$\mathrm{E}(X^2) \;=\; \sum_k k^2\,\mathrm{P}(X=k) \;=\; (-1)^2\times\frac{1}{3} + 0^2\times\frac{1}{3} + 1^2\times\frac{1}{3} \;=\; \frac{2}{3},$$

we then have the standard deviation

$$\mathrm{SD}(X) \;=\; \sqrt{\frac{2}{3} - 0^2} \;=\; \sqrt{\mathrm{E}(X^2) - \mathrm{E}(X)^2} \;=\; \sqrt{\frac{2}{3}} \;\approx\; 0.82.$$

(b) From the column sums of the joint distribution, we get the distribution of Y as

k	-1	0	1
$\mathrm{P}(Y=k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

We see that Y has the same distribution as X, so also $\mathrm{E}(Y)=0$ and $\mathrm{SD}(Y)=\sqrt{2/3}\approx 0.82$.

(c) Let use the formula $\mathrm{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\,\mathrm{E}(Y)$. We must calculate, using the joint densities,

$$\mathrm{E}(XY) \;=\; \sum_i \sum_j ij\,\mathrm{P}(X=i,Y=j).$$

We can simply leave out all terms where $i=0$ or $j=0$ (because those terms are zero), so

$$\begin{aligned}\mathrm{E}(XY) \;&=\; \sum_{i\neq 0}\sum_{j\neq 0} ij\,\mathrm{P}(X=i,Y=j)\\ &=\; (-1)\times(-1)\times 0 \;+\; (-1)\times 1\times\frac{1}{6} \;+\; 1\times(-1)\times 0 \;+\; 1\times 1\times\frac{1}{6}\\ &=\; 0.\end{aligned}$$

Thus the correlation is

$$\mathrm{Cor}(X,Y) \;=\; \frac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)} \;=\; \frac{\mathrm{E}(XY)-\mathrm{E}(X)\,\mathrm{E}(Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)} \;=\; 0.$$

(d) X and Y are stochastically *dependent*, because (for example)

$$\mathrm{P}(Y=-1\,|\,X=0) \;=\; 1,$$

while

$$\mathrm{P}(Y=-1) \;=\; 1/3\neq 1.$$

The important message of this exercise is that two variables being uncorrelated *does not mean* they are independent. Being independent is a very strong condition (*all* conditional distributions of Y, for all values $X=x$, need to be exactly the same). If this is true, then correlation is also zero. In contrast, zero correlation is a weaker condition: it simply means that the variables do not have a *linear* stochastic dependence.

3A2 (Average of dice) An ordinary die is rolled many times. The individual results are denoted $X_1, X_2, \ldots$, and they are independent. The average of the first n results is denoted A_n.

(a) Find mean and standard deviation of X_1.

(b) Find the distribution of $A_2=\frac{1}{2}(X_1+X_2)$.

(c) Find mean and standard distribution of A_2.

(d) Find mean and standard distribution of

$$A_{100} \;=\; \frac{1}{100}(X_1 + X_2 + \cdots + X_{100}).$$

Hint: In (c), you can either calculate directly from the distribution of A_2, or you can apply linearity of expectation and covariance. In particular, observe that

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

How does this simplify when X_1 and X_2 are independent? In (d) it would be very laborious to find the exact distribution of A_{100}, so the previous formula is very convenient. How does it work for a sum of many random variables?

Solution.

(a) X_1 is uniformly distributed in the set $\{1, 2, \ldots, 6\}$, so its mean (expected value) is

$$\mathrm{E}(X_1) \;=\; \sum_k k\,\mathrm{P}(X_1 = k) \;=\; \sum_{k=1}^{6} k \times \frac{1}{6} \;=\; \frac{7}{2} \;=\; 3.5.$$

In order to find standard deviation, we first find the expectation of the square

$$\mathrm{E}(X_1^2) \;=\; \sum_k k^2\,\mathrm{P}(X_1 = k) \;=\; \sum_{k=1}^{6} k^2 \times \frac{1}{6} \;=\; \frac{91}{6}.$$

Then we have

$$\mathrm{SD}(X_1) \;=\; \sqrt{\mathrm{E}(X_1^2) - \mathrm{E}(X_1)^2} \;=\; \sqrt{\frac{91}{6} - \left(\frac{7}{2}\right)^2} \;=\; \sqrt{\frac{35}{12}} \;\approx\; 1.7.$$

(b) Because X_1 and X_2 can both take any integer values from 1 to 6, we observe that A_2 can take values in the set $\{1.0, 1.5, 2.0, 2.5, \ldots, 5.5, 6.0\}$. The probabilities of those values can be determined, one by one:

- The event $\{A_2 = 1\}$ occurs if $X_1 = 1$ and $X_2 = 1$. Thus $\mathrm{P}(A_2 = 1) = (\frac{1}{6})^2 = \frac{1}{36}$.
- The event $\{A_2 = 1.5\}$ occurs, if $(X_1, X_2) = (1, 2)$ or $(X_1, X_2) = (2, 1)$. Thus $\mathrm{P}(A_2 = 1.5) = 2 \times (\frac{1}{6})^2 = \frac{2}{36}$.
- ...

After some work we have the distribution of A_2:

k	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
$\mathrm{P}(A_2 = k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

It may be a good idea to draw a graph of the density function, to see how the distribution looks like.

(c) **Method 1.** The distribution of A_2 was calculated in (a). From the distribution we can calculate

$$\mathrm{E}(A_2) \;=\; \sum_x x\,\mathrm{P}(A_2 = x) \;=\; 1.0 \times \frac{1}{36} + 1.5 \times \frac{2}{36} + \cdots \;=\; 3.5$$

and

$$\mathrm{E}(A_2^2) \;=\; \sum_x x^2\,\mathrm{P}(A_2 = x) \;=\; 1.0^2 \times \frac{1}{36} + 1.5^2 \times \frac{2}{36} + \cdots \;\approx\; 13.71,$$

josta

$$\mathrm{SD}(A_2) \;=\; \sqrt{\mathrm{E}(A_2^2) - (\mathrm{E}(A_2))^2} \;\approx\; \sqrt{13.71 - 3.5^2} \;\approx\; 1.2.$$

Since there are so many values to add, one may want to use a computer, for example this R code:

```
a <- seq(1.0,6.0,by=0.5)
p <- c(1:6,5:1)/36
m1 <- sum(x*p)
m2 <- sum(x^2*p)
mu <- m1
sigma <- sqrt(m2-m1^2)
```

Method 2. Because X_2 has the same distribution as X_1, it also has the same mean and standard deviation, calculated in (a). By linearity of expectation,

$$\mathrm{E}(\frac{1}{2}(X_1 + X_2)) = \frac{1}{2}(\mathrm{E}(X_1) + \mathrm{E}(X_2)) = \frac{1}{2}(3.5 + 3.5) = 3.5.$$

Because X_1 and X_2 are independent, their covariance is zero. Following the hint, we find

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2 \cdot \mathrm{Cov}(X_1, X_2) = \frac{35}{12} + \frac{35}{12} + 2 \cdot 0 = \frac{35}{6},$$

so $\mathrm{SD}(X_1 + X_2) = \sqrt{\frac{35}{6}}$. But we are interested in A_2, which is half of $X_1 + X_2$, so we calculate

$$\mathrm{SD}(A_2) = \frac{1}{2}\,\mathrm{SD}(X_1 + X_2) = \frac{1}{2}\sqrt{\frac{35}{6}} \approx 1.2.$$

(d) The sum of the first 100 results can be any integer from 100 to 600, so the average can be any of the numbers $\{1.00, 1.01, 1.02, \ldots, 5.99, 6.00\}$. So A_{100} has a discrete distribution in a set of 501 possible values. In principle, we could tabulate these values and their probabilities. However, it is not quite easy to calculate, for example, the probability of the event $\{A_{100} = 3.97\}$, because there are many combinations of 100 dice results that have sum 397 (thus average 3.97). Note that in total, there are $6^{100} \approx 6 \cdot 10^{77}$ possible result sequences from rolling 100 dice.

By the linearity of expectation, we find easily

$$\mathrm{E}(A_{100}) \;=\; \mathrm{E}\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) \;=\; \frac{1}{100}\sum_{i=1}^{100}\mathrm{E}(X_i).$$

Because each result X_i has the same distribution as X_1, namely the uniform distribution in $\{1,2,3,4,5,6\}$, we have

$$\mathrm{E}(A_{100}) \;=\; \mathrm{E}(X_1) \;=\; 3.5.$$

Furthermore, because all the individual results X_i are independent from each other, their covariances are zero, so using the formula given in the hint (and applying it to a sum of 100 random variables), we find

$$\mathrm{Var}(A_{100}) \;=\; \mathrm{Var}\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) \;=\; \left(\frac{1}{100}\right)^2\sum_{i=1}^{100}\mathrm{Var}(X_i) \;=\; \frac{\mathrm{Var}(X_1)}{100}$$

thus the standard deviation of A_{100} is

$$\mathrm{SD}(A_{100}) \;=\; \sqrt{\frac{\mathrm{Var}(X_1)}{100}} \;=\; \frac{\mathrm{SD}(X_1)}{10} \;=\; \frac{\sqrt{35/12}}{10} \;\approx\; 0.171.$$

This exercise demonstrates that the *average* of independent, identically distributed random variables has the *same mean*, but *smaller standard deviation* than the individual variables. This obervation has many consequences. If one invests money into several assets (e.g. stocks) whose returns are random, but independent, one can reduce the risk of the total return. The fundamendal idea of an insurance company (or a casino) is also based on this. Furthermore, statistical estimation tasks are also based on this, because the bigger sample you take, the smaller is the standard deviation of the average.

Home problems

3A3 (Predicting temperatures) A meteorologist is modelling the relation between today's temperature T_0 and tomorrow's temperature T_1 with the equation

$$T_1 \;=\; T_0 + \Delta T$$

where ΔT is a random variable indicating the change in temperature. The random variables T_0 and ΔT are assumed independent. Moreover, we know that $\mathrm{E}(T_0) = \mu$, $\mathrm{Var}(T_0) = \sigma^2$, $\mathrm{E}(\Delta T) = 0$, and $\mathrm{Var}(\Delta T) = \theta^2$. The model parameters μ, σ and θ are known (and $\sigma > 0$ and $\theta \geq 0$).

(a) Find $\mathrm{E}(T_1)$.

(b) Find $\mathrm{SD}(T_1)$. Recall from exercise 3A2 how you can calculate the variance of a sum.

(c) Find $\mathrm{Cov}(T_1, T_0)$. Use the linearity of covariance.

(d) Find $\mathrm{Cor}(T_1, T_0)$. Before you calculate it, try to guess, by thinking about the meaning of correlation, how the correlation should be if θ is small or zero, and if it is very large (much larger than σ).

The results should be formulas expressed in terms of the model parameters (μ, σ, θ).

Grading. 0.5 points per item, total rounded up. Points are awarded for correct calculations, even if they use incorrect numerical values from earlier items.

Solution.

(a) By the linearity of expectation

$$\mathrm{E}(T_1) = \mathrm{E}(T_0 + \Delta T) = \mathrm{E}(T_0) + \mathrm{E}(\Delta T) = \mu + 0 = \mu.$$

So both days have the same expected temperature.

(b) By the variance summation formula (hint in problem 3A2),

$$\begin{aligned} \mathrm{Var}(T_1) &= \mathrm{Var}(T_0 + \Delta T) \\ &= \mathrm{Var}(T_0) + 2\,\mathrm{Cov}(T_0, \Delta T) + \mathrm{Var}(\Delta T) \\ &= \sigma^2 + 0 + \theta^2 \\ &= \sigma^2 + \theta^2, \end{aligned}$$

where $\mathrm{Cov}(T_0, \Delta T) = 0$ because T_0 and ΔT are independent. Thus $\mathrm{SD}(T_1) = \sqrt{\sigma^2 + \theta^2}$.

(c) Applying the linearity of covariance, we find

$$\begin{aligned} \mathrm{Cov}(T_1, T_0) &= \mathrm{Cov}(T_0 + \Delta T,\, T_0) \\ &= \mathrm{Cov}(T_0, T_0) + \mathrm{Cov}(\Delta T, T_0) \\ &= \mathrm{Var}(T_0) + 0 \\ &= \sigma^2. \end{aligned}$$

(d) If θ is small, then the temperature will (probably) change very little. The temperatures of both days are random, but there is a strong dependence between them (if today's temperature is high, so is tomorrow's), and it seems correlation should be high.

If θ is large, then temperatures can change a lot, most of the variance in tomorrow's temperature is caused by the change and not by today's temperature, and the correlation should be low.

Let us calculate:

$$\mathrm{Cor}(T_1, T_0) = \frac{\mathrm{Cov}(T_1, T_0)}{\mathrm{SD}(T_1)\,\mathrm{SD}(T_0)} = \frac{\sigma^2}{(\sqrt{\sigma^2+\theta^2})\cdot\sigma} = \frac{\sigma}{\sqrt{\sigma^2+\theta^2}}.$$

Indeed, the result is a decreasing function of θ; if $\theta = 0$, then the correlation is $+1$, indicating deterministic linear dependence (tomorrow's temperature equals today's temperature). If θ is much bigger than θ, then the correlation is almost zero.

3A4 (Minimizing loss functions) Eastham is a town that stretches along a straight road, two kilometers from west to east, so we model it as a line segment of length 2. There are more people living in the east end than in the west end; the location of a randomly chosen inhabitant is a random variable X that has density function $f(x) = x/2$, when $0 \le x \le 2$. (Although in reality the population would be finite, here we think there are so many inhabitants that we can treat them as a continuous mass.)

(a) Find the mean $\mu = \mathrm{E}(X)$, and the median m, which is a point such that $\mathrm{P}(X \le m) = \frac{1}{2}$.

(b) Find $\mathrm{E}(X^2)$ and $\mathrm{SD}(X)$.

(c) Abel is a planner who wants to choose a location c for the town hall, somewhere in town. He tries to minimize the *quadratic loss* $q(c) = \mathrm{E}\left((X-c)^2\right)$. In other words, he tries to minimize the average of all inhabitants' *squared distances* to the town hall. Express q as a simple function of c. What is the shape of this function? Find the value of c where $q(c)$ is minimized. Is it one of the values μ and m?

(d) Bertha is another planner that wants to choose a location c for the library, somewhere in town. She tries to minimize the *linear loss* $\ell(c) = \mathrm{E}(\,|X-c|\,)$. In other words, she wants to minimize the average of all inhabitants' *distances* to the library. Express $\ell(c)$ as a simple function of c. Find the value of c where $\ell(c)$ is minimized. Is it one of the values μ and m? Compare the locations that Abel and Bertha chose, and try to give an intuitive explanation. Hint: Express the linear loss as an integral over the whole town. Then break it into two integrals, one for $x < c$ and one for $x \ge c$.

Grading. 0.5 points per item, rounded up. Points are awarded for correct calculations even if using incorrect numeric values from previous items. In (d), intuitive explanation is not required.

Solution.

(a)

$$\mu = \mathrm{E}(X) = \int_0^2 x f(x) dx = \int_0^2 \frac{x^2}{2} dx = \left[\frac{x^3}{6} \right]_0^2 = \frac{8}{6} = \frac{4}{3} \approx 1.333.$$

To find the median, we observe that

$$\mathrm{P}(X \le m) = \int_0^m f(x) dx = \int_0^2 \frac{x}{2} dx = \left[\frac{x^2}{4} \right]_0^m = \frac{m^2}{4}.$$

From the condition $m^2/4 = \frac{1}{2}$ we easily solve $m = \sqrt{2} \approx 1.414$.

(b)

$$\mathrm{E}(X^2) = \int_0^2 x^2 f(x) dx = \int_0^2 \frac{x^3}{2} dx = \left[\frac{x^4}{8} \right]_0^2 = \frac{16}{8} = 2.$$

Now we can calculate

$$\mathrm{SD}(X) = \sqrt{\mathrm{E}(X^2) - (\mathrm{E}(X))^2} = \sqrt{2 - (4/3)^2} = \sqrt{2/9} \approx 0.471.$$

(c) If Abel places the town hall at c, then an inhabitant at x has squared distance $(x-c)^2$. The quadratic loss is

$$q(c) = \mathrm{E}((X-c)^2) = \mathrm{E}(X^2 - 2cX + c^2) = \mathrm{E}(X^2) - 2c\,\mathrm{E}(X) + c^2 = 2 - \frac{8}{3}c + c^2,$$

where we applied linearity of expectation. Now $q(c)$ is an upward-opening parabola, and its minimum is found where its derivative

$$q'(c) = 2c - \frac{8}{3}$$

is zero. Solving c, we find that $c = \frac{4}{3} \approx 1.333$. It turns out that Abel places the town hall at μ, the mean location of the inhabitants.

(d) If Bertha places the library at c, then an inhabitant at x has distance $|x-c|$, that is, $x - c$ if $x \ge c$, and $c - x$ if $x < c$. The linear loss is

$$\begin{aligned}
\ell(c) = \mathrm{E}\left(|X-c| \right) &= \int_0^2 |x-c| \; f(x) dx \\
&= \int_0^c (c-x) f(x) dx \; + \; \int_c^2 (x-c) f(x) dx \\
&= \int_0^c \left(\frac{cx}{2} - \frac{x^2}{2} \right) dx + \int_c^2 \left(\frac{x^2}{2} - \frac{cx}{2} \right) dx \\
&= \frac{c^3}{6} - c + \frac{4}{3}.
\end{aligned}$$

By inspecting the shape of this function (e.g. by drawing it, or from its first and second derivatives), we find that $\ell(c)$ curves upward everywhere in $[0, 2]$, and its minimum is found where its first derivative is zero. The first derivative is

$$\ell'(c) = \frac{1}{2}c^2 - 1,$$

and we easily find that this is zero when $c = \sqrt{2} \approx 1.414$. It turns out that Bertha places the library at m, the median location of the inhabitants.

Both Abel and Bertha place the facilities somewhere in the east part of the town, because most people live there. However, Abel's quadratic loss function places a stronger penalty on people living very far from town hall (because the distances are squared). Trying to avoid such large penalties from people living in the west, Abel places the town hall slightly to the west of the library.

The same thing would happen for other distributions. In general, one can prove that *mean* is the location that minimizes quadratic loss, and *median* is the location that minimizes linear loss. You can try proving these general statements for an arbitrary density function f. (For mean, you find it in Ross p. 118; median not covered there). — If you want a challenge, try to generalize these notions to a town that spans a *two-dimensional area*. Minimizing quadratic loss will be relatively easy, but minimizing linear loss will be difficult (look up "geometric median").