Covariance and Correlation
Class 7b, TF, AID-M
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Understand the meaning of covariance and correlation.
2. Be able to compute the covariance and correlation of two random variables.

2 Covariance

Covariance is a measure of how much two random variables vary together. For example, height and weight of giraffes have positive covariance because when one is big the other tends also to be big.

Definition: Suppose X and Y are random variables with means μ_X and μ_Y. The covariance of X and Y is defined as

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

2.1 Properties of covariance

1. $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$ for constants a, b, c, d.
2. $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$.
3. $\text{Cov}(X, X) = \text{Var}(X)$
4. $\text{Cov}(X, Y) = E[XY] - \mu_X\mu_Y$.
5. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$ for any X and Y.
6. If X and Y are independent then $\text{Cov}(X, Y) = 0$.
   Warning: The converse is false: zero covariance does not always imply independence.

Notes. 1. Property 4 is like the similar property for variance. Indeed, if $X = Y$ it is exactly that property: $\text{Var}(X) = E[X^2] - \mu_X^2$.

By Property 5, the formula in Property 6 reduces to our earlier formula $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ when X and Y are independent.

We give the proofs below. However, understanding and using these properties is more important than memorizing their proofs.

2.2 Sums and integrals for computing covariance

Since covariance is defined as an expected value we compute it in the usual way as a sum or integral.

Discrete case: If X and Y have joint pmf $p(x_i, y_j)$ then

$$\text{Cov}(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)(x_i-\mu_X)(y_j-\mu_Y) = \left(\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)x_iy_j\right) - \mu_X\mu_Y.$$

Continuous case: If X and Y have joint pdf $f(x,y)$ over range $[a,b]\times[c,d]$ then

$$\text{Cov}(X,Y) = \int_c^d\int_a^b (x-\mu_x)(y-\mu_y)f(x,y)\,dx\,dy = \left(\int_c^d\int_a^b xyf(x,y)\,dx\,dy\right) - \mu_x\mu_y.$$

2.3 Examples

Example 1. Flip a fair coin 3 times. Let X be the number of heads in the first 2 flips and let Y be the number of heads on the last 2 flips (so there is overlap on the middle flip). Compute $\text{Cov}(X,Y)$.

Solution: We'll do this twice, first using the joint probability table and the definition of covariance, and then using the properties of covariance.

With 3 tosses there are only 8 outcomes {HHH, HHT,...}, so we can create the joint probability table directly.

$X\backslash Y$	0	1	2	$p(x_i)$
0	1/8	1/8	0	1/4
1	1/8	2/8	1/8	1/2
2	0	1/8	1/8	1/4
$p(y_j)$	1/4	1/2	1/4	1

From the marginals we compute $E[X] = 1 = E[Y]$. Now we use use the definition:

$$\text{Cov}(X,Y) = E[(X-\mu_x)(Y-\mu_Y)] = \sum_{i,j} p(x_i,y_j)(x_i-1)(y_j-1)$$

We write out the sum leaving out all the terms that are 0, i.e. all the terms where $x_i = 1$ or $y_i = 1$ or the probability is 0.

$$\text{Cov}(X,Y) = \frac{1}{8}(0-1)(0-1) + \frac{1}{8}(2-1)(2-1) = \frac{1}{4}.$$

We could also have used property 4 to do the computation: From the full table we compute

$$E[XY] = 1\cdot\frac{2}{8} + 2\frac{1}{8} + 2\frac{1}{8} + 4\frac{1}{8} = \frac{5}{4}.$$

So $\text{Cov}(XY) = E[XY] - \mu_X\mu_Y = \dfrac{5}{4} - 1 = \dfrac{1}{4}$.

Next we redo the computation of $\text{Cov}(X, Y)$ using the properties of covariance. As usual, let X_i be the result of the i^{th} flip, so $X_i \sim$ Bernoulli(0.5). We have

$$X = X_1 + X_2 \quad \text{and} \quad Y = X_2 + X_3.$$

We know $E[X_i] = 1/2$ and $\text{Var}(X_i) = 1/4$. Therefore using Property 2 of covariance, we have

$$\text{Cov}(X, Y) = \text{Cov}(X_1+X_2, X_2+X_3) = \text{Cov}(X_1, X_2)+\text{Cov}(X_1, X_3)+\text{Cov}(X_2, X_2)+\text{Cov}(X_2, X_3).$$

Since the different tosses are independent we know

$$\text{Cov}(X_1, X_2) = \text{Cov}(X_1, X_3) = \text{Cov}(X_2, X_3) = 0.$$

Looking at the expression for $\text{Cov}(X, Y)$ there is only one non-zero term

$$\text{Cov}(X, Y) = \text{Cov}(X_2, X_2) = \text{Var}(X_2) = \boxed{\frac{1}{4}}.$$

Example 2. (Zero covariance does not imply independence.) Let X be a random variable that takes values $-2, -1, 0, 1, 2$; each with probability $1/5$. Let $Y = X^2$. Show that $\text{Cov}(X, Y) = 0$ but X and Y are not independent.

Solution: We make a joint probability table:

$Y \backslash X$	-2	-1	0	1	2	$p(y_j)$
0	0	0	1/5	0	0	1/5
1	0	1/5	0	1/5	0	2/5
4	1/5	0	0	0	1/5	2/5
$p(x_i)$	1/5	1/5	1/5	1/5	1/5	1

Using the marginals we compute means $E[X] = 0$ and $E[Y] = 2$.

Next we show that X and Y are not independent. To do this all we have to do is find one place where the product rule fails, i.e. where $p(x_i, y_j) \neq p(x_i)p(x_j)$:

$$P(X = -2, Y = 0) = 0 \quad \text{but} \quad P(X = -2) \cdot P(Y = 0) = 1/25.$$

Since these are not equal X and Y are not independent. Finally we compute covariance using Property 4:

$$\text{Cov}(X, Y) = \frac{1}{5}(-8 - 1 + 1 + 8) - \mu_X\mu_y = 0.$$

Discussion: This example shows that $\text{Cov}(X, Y) = 0$ does not imply that X and Y are independent. In fact, X and X^2 are as dependent as random variables can be: if you know the value of X then you know the value of X^2 with 100% certainty.

The key point is that $\text{Cov}(X, Y)$ measures the linear relationship between X and Y. In the above example X and X^2 have a quadratic relationship that is completely missed by $\text{Cov}(X, Y)$.

Continuous covariance works the same way, except our computations are done with integrals instead of sums. Here is an example.

Example 3. Continuous covariance. Suppose X and Y are jointly distributed random variables, with range on the unit square $[0, 1] \times [0, 1]$ and joint pdf $f(x, y) = 2x^3 + 2y^3$.

(i) Verify the $f(x, y)$ is a valid probability density.

(ii) Compute μ_X and μ_Y.

(iii) Compute the covariance of $\text{Cov}(X, Y)$

Solution: Part of the point of this example is to show how to set up and compute the integrals using a joint density function. Since the pdf here is a polynomial, these computations are relatively easy.

(i) A valid pdf has two properties: it is nonnegative and the total integral over the entire joint range is 1.

Nonnegativity is clear: $f(x, y) \geq 0$. The integral is not hard to compute

$$\int_0^1 \int_0^1 f(x, y)\, dx\, dy = \int_0^1 \int_0^1 2x^3 + 2y^3\, dx\, dy$$

$$\text{Inner integral: } \int_0^1 2x^3 + 2y^3\, dx = \frac{x^4}{2} + 2xy^3 \Big|_0^1 = \frac{1}{2} + 2y^3.$$

$$\text{Outer integral: } \int_0^1 \frac{1}{2} + 2y^3\, dy = \frac{y}{2} + \frac{y^4}{2} \Big|_0^1 = 1.$$

So, the integral over the entire joint range is 1. Thus, $f(x, y) = x + y$ is a valid probability density.

(ii) We need to compute integrals to find the means. We will write down the integrals, but not show the details of their computation. (Also, by symmetry, we know the two means are the same.)

$$\mu_X = \int_0^1 \int_0^1 x f(x, y)\, dx\, dy = \int_0^1 \int_0^1 2x^4 + 2xy^3)\, dx\, dy = \frac{13}{20}$$

$$\mu_Y = \int_0^1 \int_0^1 y f(x, y)\, dx\, dy = \int_0^1 \int_0^1 2yx^3 + 2y^4\, dx\, dy = \frac{13}{20}$$

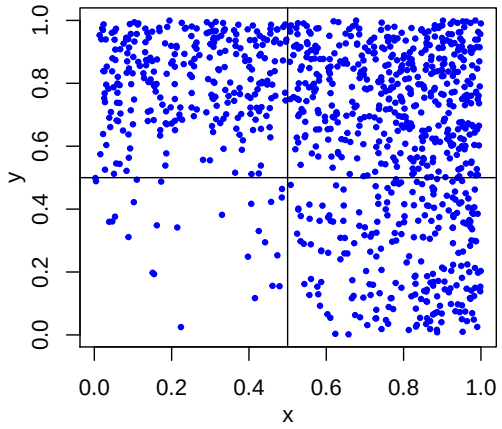
(iii) We know $\text{Cov}(X, Y) = E\left[(X - \mu_x)(Y - \mu_Y)\right]$. This is an integral. Again, we will write down the integral, but not show details of its computation,

$$\text{Cov}(X, Y) = E\left[(X - \mu_x)(Y - \mu_Y)\right] = \int_0^1 \int_0^1 (x - 13/20)(y - 13/20) f(x, y)\, dx\, dy$$

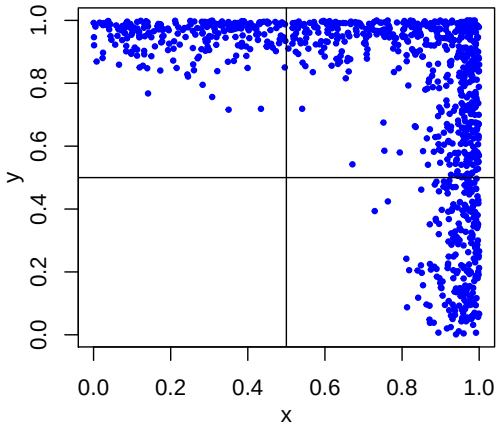
$$= \int_0^1 \int_0^1 (x - 7/12)(y - 7/12)(2x^3 + 2y^3)\, dx\, dy = -\frac{9}{400}$$

(In fact, we wrote down the integral in the most straightforward way, but secretly we did the computation by computing $E[XY] - E[X]E[Y]$.)

Here's a plot of the pseudo-random samples generated from this distribution. Because the R code could do it easily, we also include a plot with a more extreme density function.



Samples from $f(x, y) = 2x^3 + 2y^3$.



Samples from $f(x, y) = 10x^{19} + 10y^{19}$.

3 Correlation

The units of covariance $\mathrm{Cov}(X, Y)$ are 'units of X times units of Y'. This makes it hard to compare covariances: if we change scales then the covariance changes as well. Correlation is a way to remove the scale from the covariance.

Definition: The correlation coefficient between X and Y is defined by

$$\mathrm{Cor}(X, Y) = \rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \, \sigma_Y}.$$

3.1 Properties of correlation

1. ρ is the covariance of the standardizations of X and Y.
2. ρ **is dimensionless** (it's a ratio!).
3. $-1 \le \rho \le 1$. Furthermore,
   $\rho = +1$ if and only if $Y = aX + b$ with $a > 0$,
   $\rho = -1$ if and only if $Y = aX + b$ with $a < 0$.

Property 3 shows that ρ measures the linear relationship between variables. If the correlation is positive then when X is large, Y will tend to large as well. If the correlation is negative then when X is large, Y will tend to be small.

Example 2 above shows that correlation can completely miss higher order relationships.

Example 4. We continue Example 1. To compute the correlation we divide the covariance by the standard deviations. In Example 1 we found $\mathrm{Cov}(X, Y) = 1/4$ and $\mathrm{Var}(X) =$

$2\text{Var}(X_j) = 1/2$. So, $\sigma_X = 1/\sqrt{2}$. Likewise $\sigma_Y = 1/\sqrt{2}$. Thus

$$\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{1/4}{1/2} = \frac{1}{2}.$$

We see a positive correlation, which means that larger X tend to go with larger Y and smaller X with smaller Y. In Example 1 this happens because toss 2 is included in both X and Y, so it contributes to the size of both.

Example 5. Look back at Example 3. See if you can compute the following.

$$\text{Var}(X) = 31/400, \text{ so } \sigma_X = \sqrt{31/400} \approx 0.28$$
$$\text{Var}(Y) = \text{Var}(X), \text{ so } \sigma_Y \approx 0.28$$
$$\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \approx -0.29.$$

3.2 Bivariate normal distributions

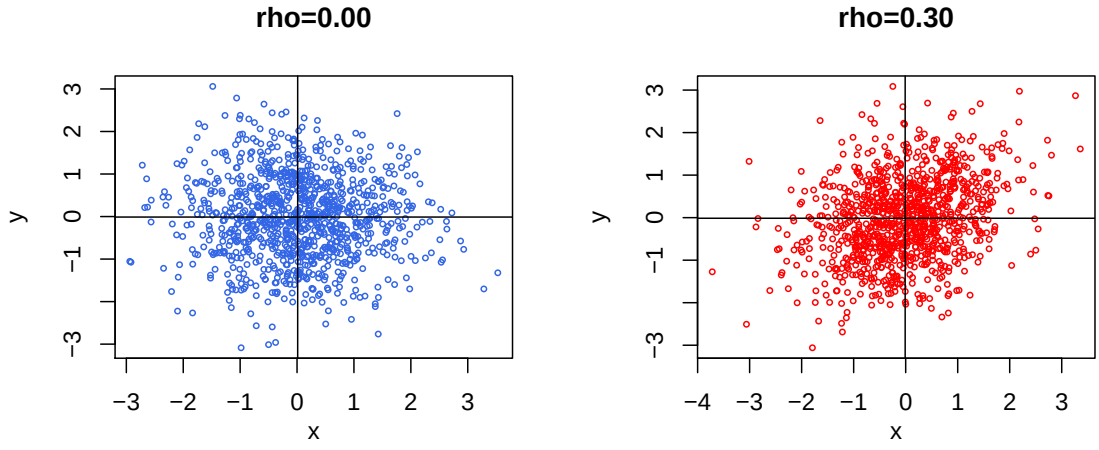
The bivariate normal distribution has density

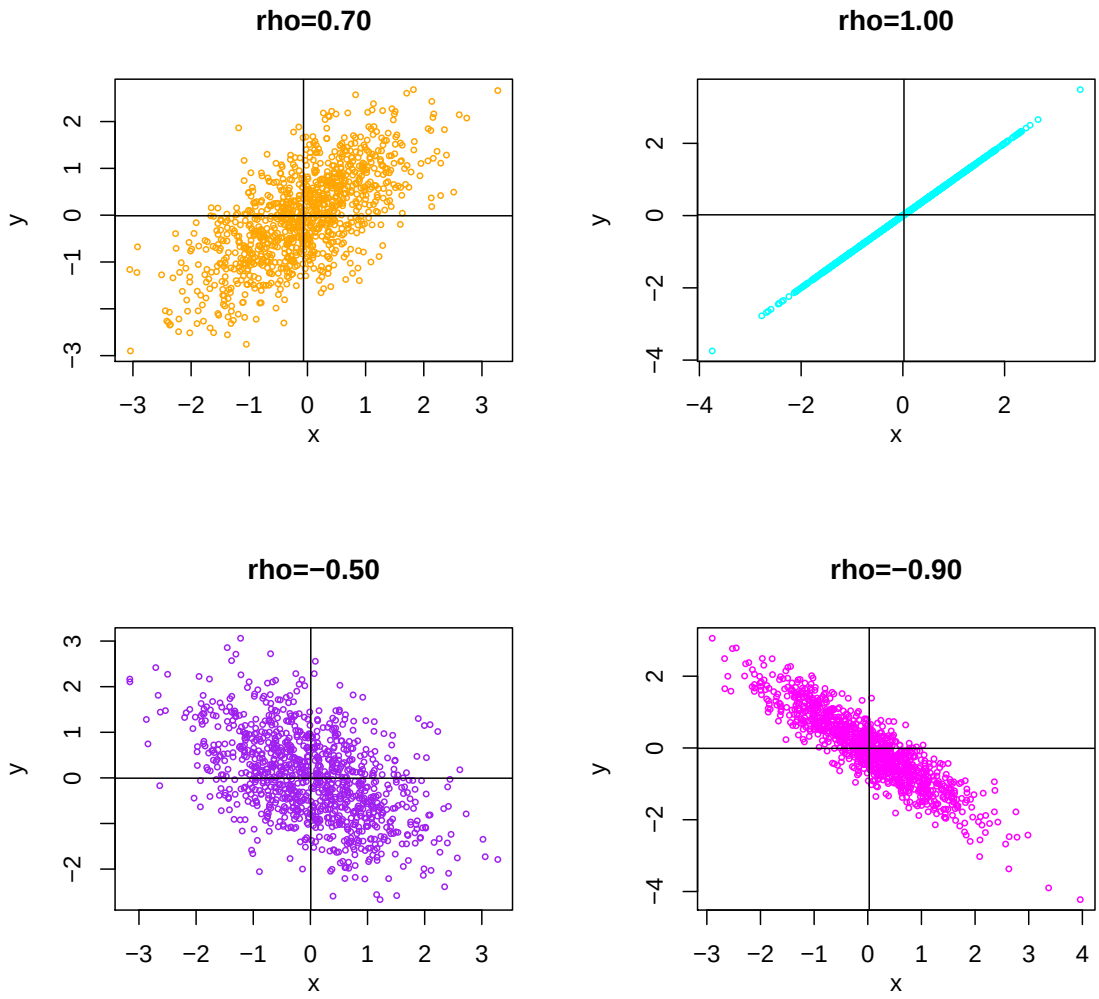
$$f(x,y) = \frac{e^{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y}\right]}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

For this distribution, the marginal distributions for X and Y are normal and the correlation between X and Y is ρ.

In the figures below we used R to simulate the distribution for various values of ρ. Individually X and Y are standard normal, i.e. $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$. The figures show scatter plots of the results.

These plots and the next set show an important feature of correlation. We divide the data into quadrants by drawing a horizontal and a verticle line at the means of the y data and x data respectively. A positive correlation corresponds to the data tending to lie in the 1st and 3rd quadrants. A negative correlation corresponds to data tending to lie in the 2nd and 4th quadrants. You can see the data gathering about a line as ρ becomes closer to ± 1.

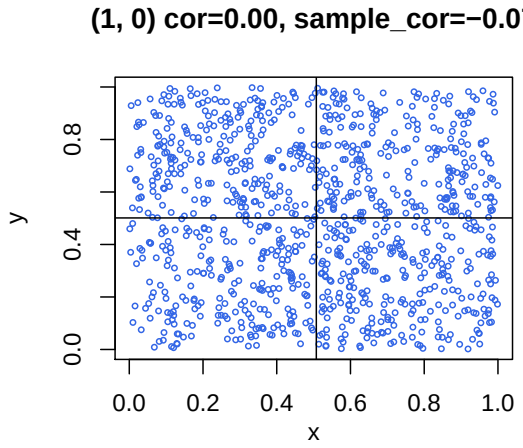


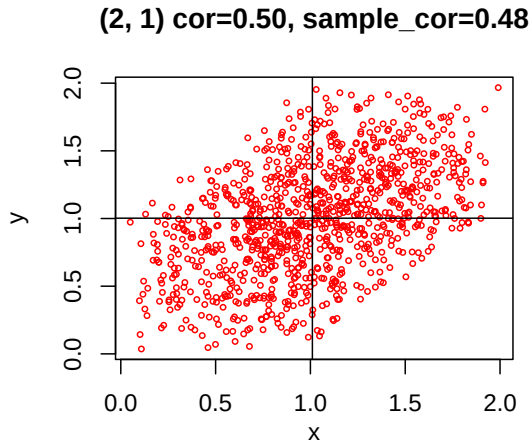


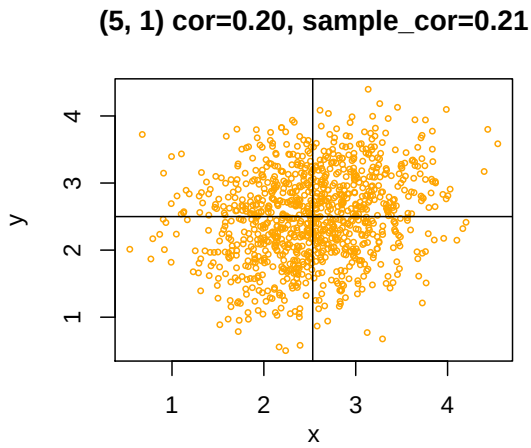
3.3 Overlapping uniform distributions

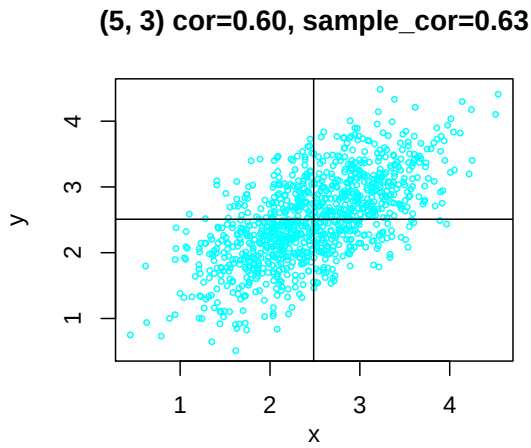
We ran simulations in R of the following scenario. X_1, X_2, ..., X_{20} are i.i.d and follow a $U(0,1)$ distribution. X and Y are both sums of the same number of X_i. We call the number of X_i common to both X and Y the overlap. The notation in the figures below indicates the number of X_i being summed and the number which overlap. For example, 5,3 indicates that X and Y were each the sum of 5 of the X_i and that 3 of the X_i were common to both sums. (The data was generated using `rand(1,1000);`)

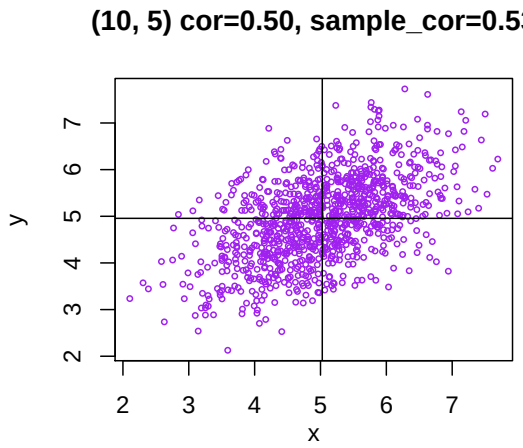
Using the linearity of covariance it is easy to compute the theoretical correlation. For each plot we give both the theoretical correlation and the correlation of the data from the simulated sample.

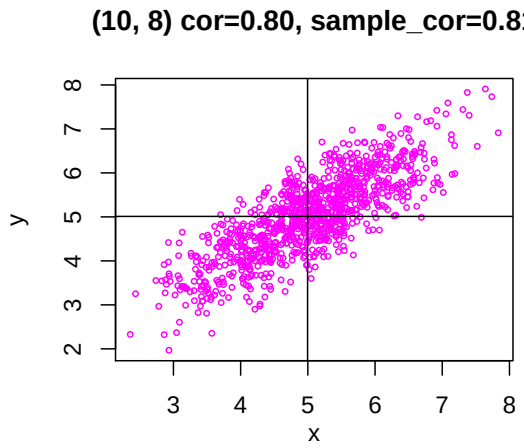

(1, 0) cor=0.00, sample_cor=−0.07
y
x
0.0
0.2
0.4
0.6
0.8
1.0


(2, 1) cor=0.50, sample_cor=0.48
y
x
0.0
0.5
1.0
1.5
2.0


(5, 1) cor=0.20, sample_cor=0.21
y
x
1
2
3
4


(5, 3) cor=0.60, sample_cor=0.63
y
x
1
2
3
4


(10, 5) cor=0.50, sample_cor=0.53
y
x
2
3
4
5
6
7


(10, 8) cor=0.80, sample_cor=0.81
y
x
2
3
4
5
6
7
8

4 Proof of the properties of covariance and correlation

4.1 Proofs of the properties of covariance

1 and 2 follow from similar properties for expected value.

3. This is the definition of variance:

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

4. Recall that $E[X-\mu_x] = 0$. So

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

5. Using properties 3 and 2 we get

$$\mathrm{Var}(X+Y) = \mathrm{Cov}(X+Y,X+Y) = \mathrm{Cov}(X,X)+2\mathrm{Cov}(X,Y)+\mathrm{Cov}(Y,Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y).$$

6. If X and Y are independent then $f(x,y) = f_X(x)f_Y(y)$. Therefore

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \int\int (x-\mu_X)(y-\mu_Y)f_X(x)f_Y(y)\,dx\,dy\\ &= \int (x-\mu_X)f_X(x)\,dx\int (y-\mu_Y)f_Y(y)\,dy\\ &= E[X-\mu_X]E[Y-\mu_Y]\\ &= 0.\end{aligned}$$

4.2 Proof of Property 3 of correlation

(*This is for the mathematically interested.*)

$$0 \le \mathrm{Var}\left(\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}\right) = \mathrm{Var}\left(\frac{X}{\sigma_X}\right)+\mathrm{Var}\left(\frac{Y}{\sigma_Y}\right)-2\mathrm{Cov}\left(\frac{X}{\sigma_X},\frac{Y}{\sigma_Y}\right) = 2-2\rho$$

This implies $\rho \le 1$

Likewise $0 \le \mathrm{Var}\left(\dfrac{X}{\sigma_X}+\dfrac{Y}{\sigma_Y}\right)$, so $-1 \le \rho$.

If $\rho = 1$ then $0 = \mathrm{Var}\left(\dfrac{X}{\sigma_X}-\dfrac{Y}{\sigma_Y}\right) \Rightarrow \dfrac{X}{\sigma_X}-\dfrac{Y}{\sigma_Y} = c.$ ∎

Note by Prof. Mayer: This script is the work of the original authors (Jeremy Orloff and Jonathan Bloom), which appear in the title and all praise for this work must be to them. The only modifications done to fit it to "Theoretical

fundamentals of AI and Data Science" (DIT, Faculty AI, Study course AIN-B) are leaving out certain topics and changing the class numbering.

The complete original script (and more: slides, exams, question sheets, and solutions) can be downloaded at MIT OpenCourseWare: Introduction To Probability And Statistics, 18.05, Spring 2014, Undergraduate

Prof. Mayer thanks Jeremy Orloff and Jonathan Bloom not only for making the course material open to the public, but also for giving him access to the original LaTeX sources to tailor the material better to the DITs needs. Without their help, this course would by far not have the quality of its current state.