Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2017

Lecture 9: Joint Distributions

Relevant textbook passages:

Pitman [3]: Chapter 5; Section 6.4–6.5

Larsen–Marx [2]: Sections 3.7, 3.8, 3.9, 3.11

9.1 Random vectors and joint distributions

Recall that a **random variable** X is a real-valued function on the sample space $(S, \mathcal{E}, P)$, where P is a probability measure on S; and that it induces a probability measure P_X on $\boldsymbol{R}$, called the **distribution** of X, given by

$$P_X(B) = P(X \in B) = P\big(\{s \in S : X(s) \in B\}\big),$$

for every (Borel) subset of $\boldsymbol{R}$. The distribution is enough to calculate the expectation of any (Borel) function of X.

Now suppose I have more than one random variable on the same sample space. Then I can consider the **random vector** (X, Y), or $\boldsymbol{X} = (X_1, \dots, X_n)$.

- A random vector $\boldsymbol{X}$ defines a probability $P_{\boldsymbol{X}}$ on $\boldsymbol{R}^{\mathrm{n}}$, called the distribution of $\boldsymbol{X}$ via:

$$P_{\boldsymbol{X}}(B) = P(\boldsymbol{X} \in B) = P\left\{s \in S : \big(X_1(s), \dots, X_n(s)\big) \in B\right\},$$

for every (Borel) subset of $\boldsymbol{R}^{\mathrm{n}}$. This distribution is also called the **joint distribution** of $X_1, \dots, X_n$.

- We can use this to define a **joint cumulative distribution function**, denoted $F_{\boldsymbol{X}}$, by

$$F_{\boldsymbol{X}}(x_1, \dots, x_n) = P(X_i \leqslant x_i, \text{ for all } i = 1, \dots, n)$$

- **N.B.** While the subscript $_{X,Y}$ or $_{\boldsymbol{X}}$ is often used to identify a joint distribution or cumulative distribution function, it is also frequently omitted. You are supposed to figure out the domain of the function by inspecting its arguments.

9.1.1 Example Let $S = \{SS, SF, FS, FF\}$ and let P be the probability measure on S defined by

$$P(SS) = \frac{7}{12},\ P(SF) = \frac{3}{12},\ P(FS) = \frac{1}{12},\ P(FF) = \frac{1}{12}.$$

Define the random variables X and Y by

$$X(SS) = 1,\ X(SF) = 1,\ X(FS) = 0,\ X(FF) = 0,$$
$$Y(SS) = 1,\ Y(SF) = 0,\ Y(FS) = 1,\ Y(FF) = 0.$$

That is, X and Y indicate Success or Failure on two different experiments, but the experiments are not necessarily independent.

Then

$$P_X(1) = \frac{10}{12},\ P_X(0) = \frac{2}{12},$$
$$P_Y(1) = \frac{8}{12},\ P_Y(0) = \frac{4}{12},$$
$$P_{X,Y}(1,1) = \frac{7}{12},\ P_{X,Y}(1,0) = \frac{3}{12},\ P_{X,Y}(0,1) = \frac{1}{12},\ P_{X,Y}(0,0) = \frac{1}{12}.$$

□

9.2 Joint PMFs

A random vector $\boldsymbol{X}$ on a probability space $(S, \mathcal{E}, P)$ is **discrete**, if you can enumerate its range.

Pitman [3]: Section 3.1; also p. 348
Larsen–Marx [2]: Section 3.7

When $X_1, \dots, X_n$ are discrete, the **joint probability mass function** of the random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ is usually denoted $p_{\boldsymbol{X}}$, and is given by

$$p_{\boldsymbol{X}}(x_1, x_2, \dots, x_n) = P\left(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } \cdots \text{ and } X_n = x_n\right).$$

If X and Y are independent random variables, then $p_{X,Y}(x,y) = p_X(x)p_Y(y)$.
For a function g of X and Y we have

$$\boldsymbol{E}\, g(X,Y) = \sum_x \sum_y g(x,y) p_{X,Y}(x,y).$$

9.3 Joint densities

Pitman [3]: Chapter 5.1
Larsen–Marx [2]: Section 3.7

Let X and Y be random variables on a probability space $(S, \mathcal{E}, P)$. The random vector (X,Y) has a **joint density** $f_{X,Y}(x,y)$ if for every (Borel) subset $B \subset \boldsymbol{R}^2$,

$$P\left((X,Y) \in B\right) = \int\int_B f_{X,Y}(x,y)\, dx\, dy.$$

If X and Y are **independent**, then $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.
For example,

$$P\left(X \geqslant Y\right) = \int_{-\infty}^{\infty} \int_{y}^{\infty} f_{X,Y}(x,y)\, dx\, dy.$$

For a function g of X and Y we have

$$\boldsymbol{E}\, g(X,Y) = \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\, dx\, dy.$$

Again, the subscript ${}_{X,Y}$ or ${}_{\boldsymbol{X}}$ on a density is frequently omitted.

9.4 Recovering marginal distributions from joint distributions

So now we have random variables X and Y, and the random vector (X,Y). They have distributions P_X, P_Y, and $P_{X,Y}$. How are they related? The **marginal distribution** of X is just the distribution P_X of X alone. We can recover its probability mass function from the joint probability mass function $p_{X,Y}$ as follows.

In the discrete case:

$$p_X(x) = P(X = x) = \sum_y p_{X,Y}(x, y)$$

Likewise

$$p_Y(y) = P(Y = y) = \sum_y p_{X,Y}(x, y)$$

If X and Y are independent random variables, then $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.

Larsen–Marx [2]: p. 169

For the density case, the **marginal density** of X, denoted f_X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy,$$

and the marginal density f_Y of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx.$$

The recovery of a marginal density of X from a joint density of X and Y is sometimes described as **"integrating out"** out y.

9.5 ⋆ Expectation of a random vector

Since random vectors are just vector-valued functions on a sample space S, we can add them and multiply them just like any other functions. For example, the sum of random vectors $\boldsymbol{X} = (X_1, \dots, X_n)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ is given by

$$(\boldsymbol{X} + \boldsymbol{Y})(s) = \boldsymbol{X}(s) + \boldsymbol{Y}(s) = \big(X_1(s), \dots, X_n(s)\big) + \big(Y_1(s), \dots, Y_n(s)\big).$$

Thus the set of random vectors is a vector space. In fact, the subset of random vectors whose components have a finite expectation is also a vector subspace of the vector space of all random vectors.

If $\boldsymbol{X} = (X_1, \dots, X_n)$ is a random vector, and each X_i has expectation $\boldsymbol{E}\, X_i$, the expectation of $\boldsymbol{X}$ is defined to be

$$\boldsymbol{E}\,\boldsymbol{X} = (\boldsymbol{E}\, X_1, \dots, \boldsymbol{E}\, X_n).$$

- Expectation is a **linear operator** on the space of random vectors. This means that

$$\boldsymbol{E}(a\boldsymbol{X} + b\boldsymbol{Y}) = a\,\boldsymbol{E}\,\boldsymbol{X} + b\,\boldsymbol{E}\,\boldsymbol{Y}.$$

- Expectation is a **positive operator** on the space of random vectors. For vectors $\boldsymbol{x} = (x_1, \dots, x_n)$, define $\boldsymbol{x} \geqq 0$ if $x_i \geqslant 0$ for each $i = 1, \dots, n$. Then

$$\boldsymbol{X} \geqq 0 \implies \boldsymbol{E}\,\boldsymbol{X} \geqq 0.$$

9.6 The distribution of a sum

We already know to calculate the expectation of a sum of random variables—since expectation is a linear operator, the expectation of a sum is the sum of the expectations.

We are now in a position to describe the *distribution* of the sum of two random variables. Let $Z = X + Y$.

Pitman [3]: p. 147
Larsen–Marx [2]: p. 178ff.

Discrete case:

$$P(Z = z) = \sum_{(x,y):x+y=z} P(x,y) = \sum_{\text{all } x} p_{X,Y}(x, z - x)$$

9.6.1 Density of a sum

Pitman [3]: pp. 372–373

If (X, Y) has joint density $f_{X,Y}(x, y)$, what is the density of $X + Y$? Recall that the density is the derivative of the cdf, so

$$\begin{aligned} f_{X+Y}(t) = \frac{d}{dt} P(X + Y \leqslant t) &= \frac{d}{dt} \int\int_{\{(x,y):x \leqslant t-y\}} f_{X,Y}(x, y)\, dx\, dy \\ &= \frac{d}{dt} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{t-y} f_{X,Y}(x, y)\, dx \right) dy \\ &= \int_{-\infty}^{\infty} \frac{d}{dt} \left(\int_{-\infty}^{t-y} f_{X,Y}(x, y)\, dx \right) dy \\ &= \int_{-\infty}^{\infty} f_{X,Y}(t - y, y)\, dy. \end{aligned}$$

So if X and Y are independent, we get the **convolution**

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(t - y) f_Y(y)\, dy.$$

9.7 Covariance

Pitman [3]: § 6.4, p. 430

When X and Y are independent, we proved

$$\boldsymbol{Var}(X + Y) = \boldsymbol{Var}\, X + \boldsymbol{Var}\, Y.$$

More generally however, since **expectation is a positive linear operator**,

$$\begin{aligned} \boldsymbol{Var}(X + Y) &= \boldsymbol{E}\big((X + Y) - \boldsymbol{E}(X + Y)\big)^2 \\ &= \boldsymbol{E}\big((X - \boldsymbol{E}\, X) + (Y - \boldsymbol{E}\, Y)\big)^2 \\ &= \boldsymbol{E}\big((X - \boldsymbol{E}\, X)^2 + 2(X - \boldsymbol{E}\, X)(Y - \boldsymbol{E}\, Y) + (Y - \boldsymbol{E}\, Y)^2 \\ &= \boldsymbol{Var}(X) + \boldsymbol{Var}(Y) + 2\, \boldsymbol{E}(X - \boldsymbol{E}\, X)(Y - \boldsymbol{E}\, Y). \end{aligned}$$

9.7.1 Definition *The* ***covariance*** *of X and Y is defined to be*

$$\boldsymbol{Cov}(X, Y) = \boldsymbol{E}(X - \boldsymbol{E}\, X)(Y - \boldsymbol{E}\, Y). \tag{1}$$

In general

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

There is another way to write the covariance:

$$\boldsymbol{Cov}(X,Y) = \boldsymbol{E}(XY) - \boldsymbol{E}(X)\,\boldsymbol{E}(Y). \tag{2}$$

Proof: Since **expectation is a positive linear operator**,

$$\begin{aligned}\boldsymbol{Cov}(X,Y) &= \boldsymbol{E}\big((X-\boldsymbol{E}(X))\big)\big(Y-\boldsymbol{E}(Y)\big)\\ &= \boldsymbol{E}\big(XY - X\,\boldsymbol{E}(Y) - Y\,\boldsymbol{E}(X) + \boldsymbol{E}(X)\,\boldsymbol{E}(Y)\big)\\ &= \boldsymbol{E}(XY) - \boldsymbol{E}(X)\,\boldsymbol{E}(Y) - \boldsymbol{E}(X)\,\boldsymbol{E}(Y) + \boldsymbol{E}(X)\,\boldsymbol{E}(Y)\\ &= \boldsymbol{E}(XY) - \boldsymbol{E}(X)\,\boldsymbol{E}(Y).\end{aligned}$$

■

9.7.2 Remark It follows that for any random variable X,

$$\boldsymbol{Cov}(X,X) = \boldsymbol{Var}\,X.$$

If X and Y are independent, then $\boldsymbol{Cov}(X,Y) = 0$.

The converse is not true.

9.7.3 Example (Covariance = 0, but variables are not independent)
(cf. Feller [1, p. 236]) Let X be a random variable that assumes the values ± 1 and ± 2, each with probability $1/4$. ($\boldsymbol{E}\,X = 0$)

Define $Y = X^2$, and let $\bar{Y} = \boldsymbol{E}\,Y$ $(= 2.5)$. Then

$$\begin{aligned}\boldsymbol{Cov}(X,Y) &= \boldsymbol{E}\big(X(Y-\bar{Y})\big)\\ &= \big(1(1-\bar{Y})\big)\frac{1}{4} + \big((-1)(1-\bar{Y})\big)\frac{1}{4} + \big(2(4-\bar{Y})\big)\frac{1}{4} + \big((-2)(4-\bar{Y})\big)\frac{1}{4}\\ &= 0.\end{aligned}$$

But X and Y are not independent:

$$P\,(X = 1 \;\&\; Y = 1) = P\,(X = 1) = 1/2,$$

but $P\,(X = 1) = 1/4$ and $P\,(Y = 1) = 1/2$, so

$$P\,(X = 1) \cdot P\,(Y = 1) = 1/8.$$

□

9.7.4 Example (Covariance = 0, but variables are not independent) Let U, V be independent and identically distributed random variables with $\boldsymbol{E}\,U = \boldsymbol{E}\,V = 0$. Define

$$X = U + V, \qquad Y = U - V.$$

Since $\boldsymbol{E}\,X = \boldsymbol{E}\,Y = 0$,

$$\boldsymbol{Cov}(X,Y) = \boldsymbol{E}(XY) = \boldsymbol{E}\big((U+V)(U-V)\big) = \boldsymbol{E}(U^2 - V^2) = \boldsymbol{E}\,U^2 - \boldsymbol{E}\,V^2 = 0$$

since U and V have the same distribution.

But are X and Y independent?

If U and V are integer-valued, then X and Y are also integer-valued, but more importantly they have the same parity. That is, X is odd if and only if Y is odd. (This is a handy fact for KenKen solvers.)

So let U and V be independent and assume the values ± 1 and ± 2, each with probability $1/4$. ($\boldsymbol{E}\,U = \boldsymbol{E}\,V = 0$.) Then

$$P\,(X \text{ is odd}) = P\,(X \text{ is even}) = P\,(Y \text{ is odd}) = P\,(Y \text{ is even}) = \frac{1}{2},$$

but

$$P\,(X \text{ is even and } Y \text{ is odd}) = 0 \neq \frac{1}{4} = P\,(X \text{ is even})\,P\,(Y \text{ is odd}),$$

so X and Y are not independent. □

Pitman [3]: p. 432

9.7.5 Remark The product $(X - \boldsymbol{E}\,X)(Y - \boldsymbol{E}\,Y)$ is positive at outcomes s where $X(s)$ and $Y(s)$ are either both above or both below their means, and negative when one is above and the other below. So one very loose interpretation of positive covariance is that the random variables are probably both above average or below average rather than not. Of course this is just a tendency.

9.7.6 Example (The effects of covariance) For mean zero random variables that have a positive covariance, the joint density tends to concentrate on the diagonal. Figure 9.1 shows the joint density of two standard normals with various covariances. Figure 9.3 shows random samples from these distributions.

□

9.8 A covariance menagerie

Recall that for independent random variables X and Y, $\boldsymbol{Var}(X+Y) = \boldsymbol{Var}\,X + \boldsymbol{Var}\,Y$, and $\boldsymbol{Cov}(XY) = 0$. For any random variable X with finite variance, $\boldsymbol{Var}\,X = \boldsymbol{E}(X^2) - (\boldsymbol{E}\,X)^2$, so $\boldsymbol{E}(X^2) = \boldsymbol{Var}\,X + (\boldsymbol{E}\,X)^2$. Also, if $\boldsymbol{E}\,X = 0$, then $\boldsymbol{Cov}(XY) = \boldsymbol{E}(XY)$ (Why?).

9.8.1 Theorem (A Covariance Menagerie) *Let $X_1, \dots, X_n$ be independent and identically distributed random variables with common mean μ and variance σ^2. Define*

$$S = \sum_{i=1}^{n} X_i, \quad \text{and} \quad \bar{X} = S/n,$$

and let

$$D_i = X_i - \bar{X}, \quad (i = 1, \dots, n).$$

be the deviation of X_i from $\bar{X}$.

Then

1. $\boldsymbol{E}(X_i X_j) = (\boldsymbol{E}\,X_i)(\boldsymbol{E}\,X_j) = \mu^2$, *for $i \neq j$ (by independence).*
2. $\boldsymbol{E}(X_i^2) = \sigma^2 + \mu^2$.
3. $\boldsymbol{E}(X_i S) = \sum_{\substack{j=1 \\ j \neq i}}^{n} \boldsymbol{E}(X_i X_j) + \boldsymbol{E}(X_i^2) = \boldsymbol{E}(X_i^2) = \sigma^2 + n\mu^2$.
4. $\boldsymbol{E}(X_i \bar{X}) = (\sigma^2/n) + \mu^2$.
5. $\boldsymbol{E}(S) = n\mu$.
6. $\boldsymbol{Var}(S) = n\sigma^2$.

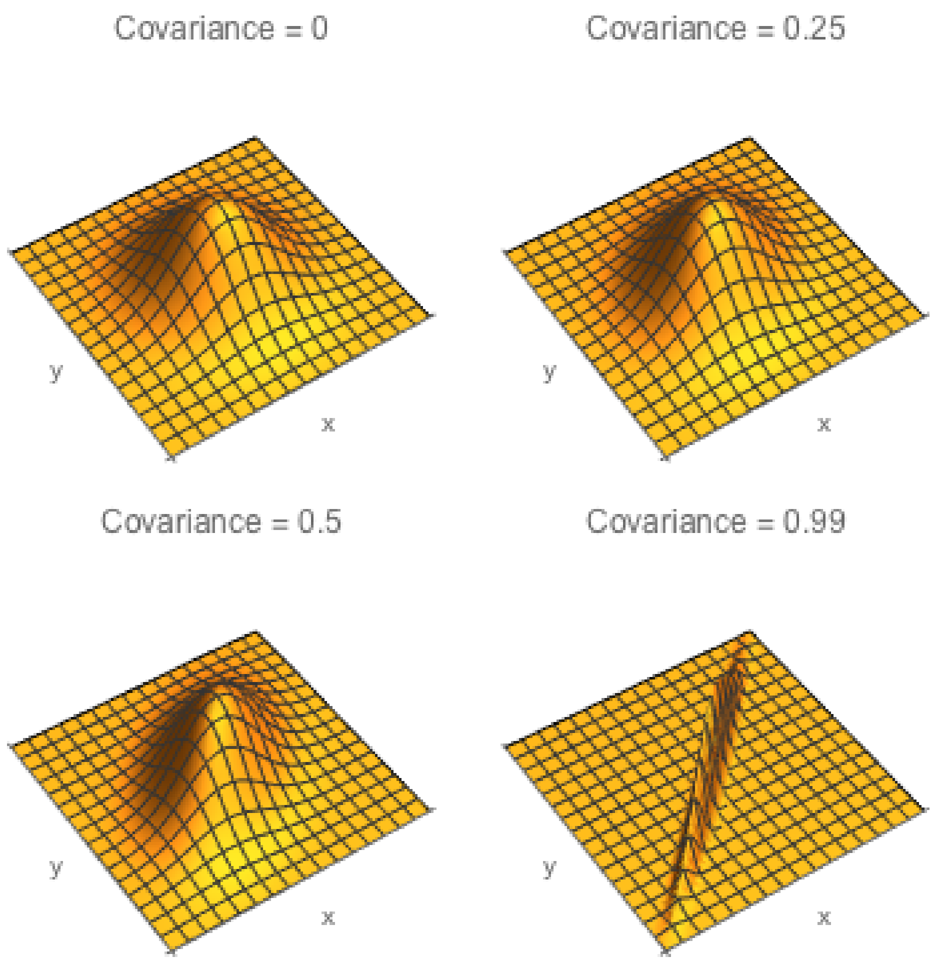


Figure 9.1. Joint density of standard normals, as covariance changes.

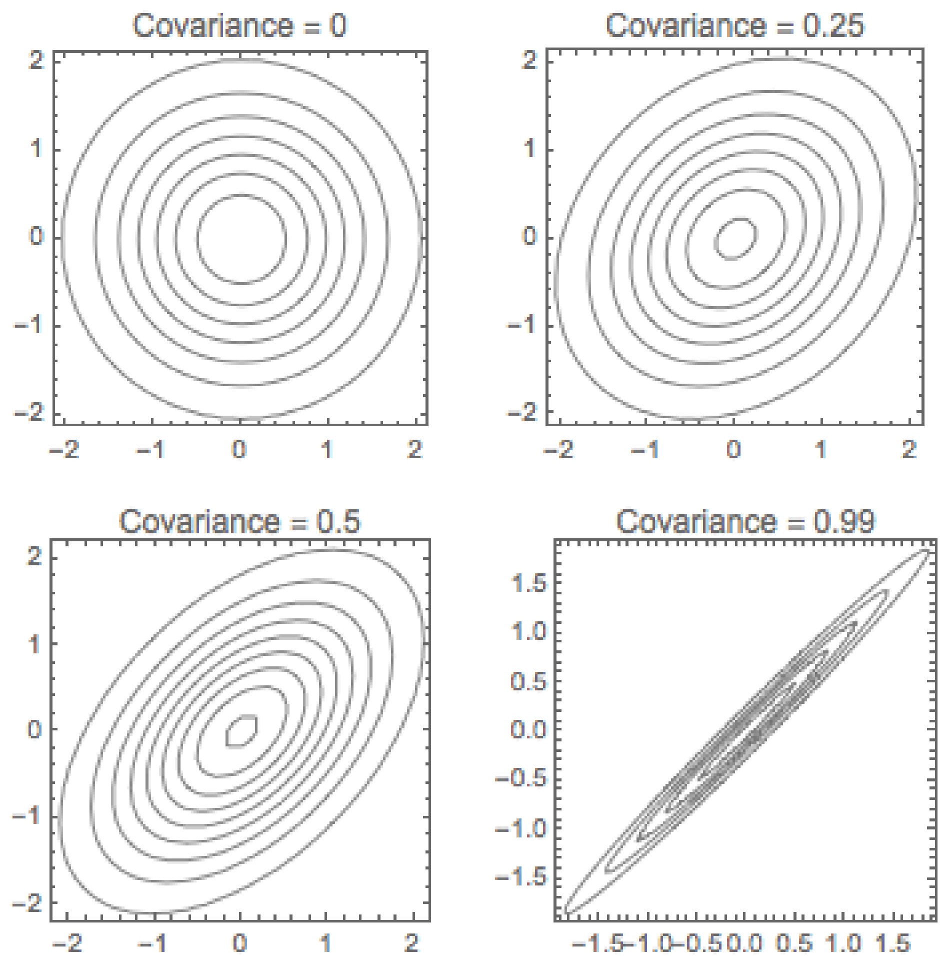
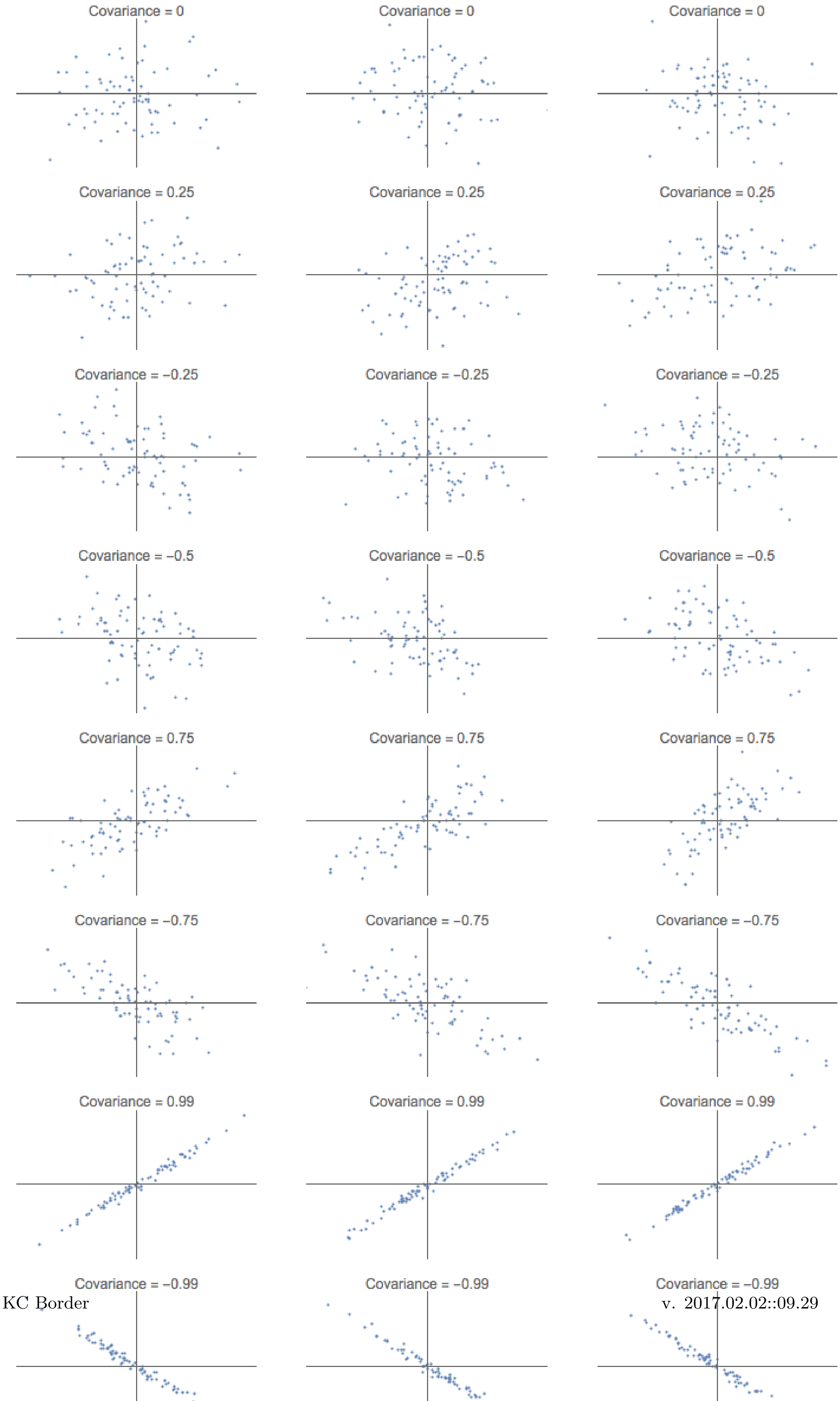


Figure 9.2. Contours of the joint density of standard normals, as covariance changes.


Covariance = 0
Covariance = 0
Covariance = 0
Covariance = 0.25
Covariance = 0.25
Covariance = 0.25
Covariance = −0.25
Covariance = −0.25
Covariance = −0.25
Covariance = −0.5
Covariance = −0.5
Covariance = −0.5
Covariance = 0.75
Covariance = 0.75
Covariance = 0.75
Covariance = −0.75
Covariance = −0.75
Covariance = −0.75
Covariance = 0.99
Covariance = 0.99
Covariance = 0.99
Covariance = −0.99
Covariance = −0.99
Covariance = −0.99

7. $\boldsymbol{E}(S^2) = n\sigma^2 + n^2\mu^2$.

8. $\boldsymbol{E}(\bar{X}) = \mu$.

9. $\boldsymbol{Var}(\bar{X}) = \sigma^2/n$.

10. $\boldsymbol{E}(\bar{X}^2) = (\sigma^2/n) + \mu$.

11. $\boldsymbol{E}(D_i) = 0$, $i = 1, \ldots, n$.

12. $\boldsymbol{Var}(D_i) = \boldsymbol{E}(D_i^2) = (n-1)\sigma^2/n$

$$\begin{aligned}\boldsymbol{Var}(D_i) = \boldsymbol{E}(X_i - \bar{X})^2 = &\\ &\boldsymbol{E}(X_i^2) - 2\,\boldsymbol{E}(X_i\bar{X}) + \boldsymbol{E}(\bar{X}^2) = \\ &(\sigma^2 + \mu^2) - 2\big((\sigma^2/n) + \mu^2\big) + \big((\sigma^2/n) + \mu^2\big) = \left(1 - \frac{1}{n}\right)\sigma^2.\end{aligned}$$

13. $\boldsymbol{Cov}(D_i, D_j) = \boldsymbol{E}(D_i D_j) = -\sigma^2/n$.

$$\begin{aligned}\boldsymbol{E}(D_i D_j) = \boldsymbol{E}\big((X_i - \bar{X})(X_j - \bar{X})\big) = \boldsymbol{E}(X_i X_j) - \boldsymbol{E}(X_i\bar{X}) - \boldsymbol{E}(X_j\bar{X}) + \boldsymbol{E}(\bar{X}^2)&\\ \mu^2 - \big[(\sigma^2/n) + \mu^2\big] - \big[(\sigma^2/n) + \mu^2\big] + \big[(\sigma^2/n) + \mu^2\big] = -\sigma^2/n.&\end{aligned}$$

Note that this means that deviations from the mean are negatively correlated. This makes sense, because if one variate is bigger than the mean, another must be smaller to offset the difference.

14. $\boldsymbol{Cov}(D_i, S) = \boldsymbol{E}(D_i S) = 0$.

$$\begin{aligned}\boldsymbol{E}(D_i S) = \boldsymbol{E}\big(\big(X_i - (S/n)\big)S\big) = \boldsymbol{E}(X_i S) - \boldsymbol{E}(S^2/n)&\\ = (\sigma^2 + n\mu^2) - (n\sigma^2 + n^2\mu^2)/n = 0.&\end{aligned}$$

15. $\boldsymbol{Cov}(D_i, \bar{X}) = \boldsymbol{E}(D_i\bar{X}) = \boldsymbol{E}(D_i S)/n = 0$.

The proof of each is a straightforward plug-and-chug calculation. The only reason for writing this as a theorem is to be able to refer to it easily.

9.9 ⋆ Covariance matrix of a random vector

In general, we define the **covariance matrix** of a random vector by

$$\boldsymbol{Var}\,\boldsymbol{X} = \begin{bmatrix} & \vdots & \\ \cdots & \boldsymbol{E}(X_i - \boldsymbol{E}\,X_i)(X_j - \boldsymbol{E}\,X_j) & \cdots \\ & \vdots & \end{bmatrix} = \begin{bmatrix} & \vdots & \\ \cdots & \boldsymbol{Cov}(X_i, X_j) & \cdots \\ & \vdots & \end{bmatrix}$$

9.10 ⋆ Variance of a linear combination of random variables

9.10.1 Proposition *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random vector with covariance matrix*

$$\Sigma = \begin{bmatrix} & \vdots & \\ \cdots & \boldsymbol{Cov}(X_i, X_i) & \cdots \\ & \vdots & \end{bmatrix}$$

and let $\boldsymbol{a} = (a_1, \dots, a_n)$*. The random variable*

$$Z = \boldsymbol{a} \cdot \boldsymbol{X} = \sum_{i=1} a_i X_i$$

has variance given by

$$\boldsymbol{Var}\, Z = \boldsymbol{a}\Sigma\boldsymbol{a}' = \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{Cov}(X_i, X_j) a_i a_j,$$

where $\boldsymbol{a}$ *is treated as a row vector, and* $\boldsymbol{a}'$ *is its transpose, a column vector.*

Proof: This just uses the fact that **expectation is a positive linear operator**. Since adding constants don't change variance, we may subtract means and assume that each $\boldsymbol{E}\, X_i = 0$. Then $\boldsymbol{Cov}(X_i, X_j) = \boldsymbol{E}(X_i X_j)$. Then Z has mean 0, so

$$\begin{aligned}\boldsymbol{Var}\, Z = \boldsymbol{E}\, Z^2 = \boldsymbol{E}\left(\sum_{i=1} a_i X_i\right)\left(\sum_{j=1} a_j X_j\right) \\ = \boldsymbol{E} \sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j a_i a_j = \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{E}(X_i X_j) a_i a_j \\ = \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{Cov}(X_i, X_j) a_i a_j.\end{aligned}$$

∎

Since $\boldsymbol{Var}\, Z \geqslant 0$ and $\boldsymbol{a}$ is arbitrary, we see that $\boldsymbol{Var}\, \boldsymbol{X}$ is a positive semidefinite matrix.

9.11 ⋆ An inner product for random variables

Since random variables are just functions on the probability space $(S, \mathcal{E}, P)$, the set of random variables is a **vector space** under the usual operations of addition of functions and multiplication by scalars. The collection of random variables that have finite variance is a linear subspace, often denoted $L_2(P)$, and it has a natural inner product.

9.11.1 Fact (Inner product on the space $L_2(P)$ of random variables) *Let* $L_2(P)$ *denote the linear space of random variables that have finite variance. Then*

$$(X, Y) = \boldsymbol{E}\, XY$$

is a real inner product on $L_2(P)$*.*

The proof of this straightforward, and is essentially the same as the proof that the Euclidean inner product on $\boldsymbol{R}^{\mathrm{m}}$ is an inner product.[1]

The next result is just the Cauchy–Schwartz Inequality for this inner product, but I've written out a self-contained proof for you.

9.11.2 Cauchy–Schwartz Inequality

$$\boldsymbol{E}(XY)^2 \leqslant (\boldsymbol{E}\, X^2)(\boldsymbol{E}\, Y^2), \tag{3}$$

with equality only if X *and* Y *are linearly dependent, that is, there exist* a, b *not both zero* $aX + bY = 0$ *a.s..*

[1] As usual, there is a caveat for infinite sample spaces, namely equality is replaced by "equality almost surely."

Proof: If X or Y is zero (almost surely), then we have equality, so assume X, Y are nonzero. Define the quadratic polynomial $Q\colon \boldsymbol{R} \to \boldsymbol{R}$ by

$$Q(\lambda) = \boldsymbol{E}\big((\lambda X + Y)^2\big) \geqslant 0.$$

Since **expectation is a positive linear operator**,

$$Q(\lambda) = \lambda^2\, \boldsymbol{E}(X^2) + 2\lambda\, \boldsymbol{E}(XY) + \boldsymbol{E}(Y^2).$$

Therefore the discriminant of the quadratic Q is nonpositive,[2] that is, $4\,\boldsymbol{E}(XY)^2 - 4\,\boldsymbol{E}(X^2)\,\boldsymbol{E}(Y^2) \leqslant 0$, or $\boldsymbol{E}(XY)^2 \leqslant \boldsymbol{E}(X^2)\,\boldsymbol{E}(Y2)$. Equality in (3) can occur only if the discriminant is zero, in which case Q has a real root. That is, there is some λ for which $Q(\lambda) = \boldsymbol{E}\big((\lambda X + Y)^2\big) = 0$. But this implies that $\lambda X + Y = 0$ (almost surely). ∎

9.11.3 Corollary

$$|\boldsymbol{Cov}(X,Y)|^2 \leqslant \boldsymbol{Var}\, X\ \boldsymbol{Var}\, Y. \tag{4}$$

Proof: Apply the Cauchy–Schwartz inequality to the random variables, $X - \boldsymbol{E}\,X$ and $Y - \boldsymbol{E}\,Y$, and then take square roots. ∎

9.12 Covariance is bilinear

********* Write this up *********

9.13 Correlation

9.13.1 Definition *The* ***correlation*** *between X and Y is defined to be*

$$\operatorname{Corr}(X,Y) = \frac{\boldsymbol{Cov}(X,Y)}{(\mathrm{SD}\,X)(\mathrm{SD}\,Y)}$$

It is also equal to

$$\operatorname{Corr}(X,Y) = \boldsymbol{Cov}(X^*,Y^*) = \boldsymbol{E}(X^*Y^*),$$

where X^* and Y^* are the standardization of X and Y.

Let X have mean μ_X and standard deviation σ_X, and ditto for Y. Recall that

$$X^* = \frac{X - \mu_X}{\sigma_X}$$

has mean 0 and std. dev. 1. Thus by the alternate formula for covariance

$$\boldsymbol{Cov}(X^*,Y^*) = \boldsymbol{E}(X^*Y^*) - \underbrace{\boldsymbol{E}(X^*)}_{=0}\,\underbrace{\boldsymbol{E}(Y^*)}_{=0}.$$

Now

$$\begin{aligned}
\boldsymbol{E}(X^*Y^*) &= \boldsymbol{E}\left(\frac{X-\mu_X}{\sigma_X}\frac{Y-\mu_Y}{\sigma_Y}\right)\\
&= \frac{\boldsymbol{E}(XY) - \boldsymbol{E}(X)\,\boldsymbol{E}(Y)}{\sigma_X\sigma_Y}\\
&= \operatorname{Corr}(X,Y)
\end{aligned}$$

Pitman [3]: p. 433

Corollary 9.11.3 (the Cauchy–Schwartz Inequality) implies:

$$-1 \leqslant \operatorname{Corr}(X, Y) \leqslant 1$$

If the correlation between X and Y is zero, then $X - \boldsymbol{E}\,X$, $Y - \boldsymbol{E}\,Y$ are orthogonal in terms of our inner product.

Bibliography

[1] W. Feller. 1968. *An introduction to probability theory and its applications*, 3d. ed., volume 1. New York: Wiley.

[2] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[3] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.

[2] In case you have forgotten how you derived the quadratic formula in Algebra I, rewrite the polynomial as

$$f(z) = \alpha z^2 + \beta z + \gamma = \frac{1}{\alpha}\left(\alpha z + \frac{\beta}{2}\right)^2 - (\beta^2 - 4\alpha\gamma)/4\alpha,$$

and note that the only way to guarantee that $f(z) \geqslant 0$ for all z is to have $\alpha > 0$ and $\beta^2 - 4\alpha\gamma \leqslant 0$.