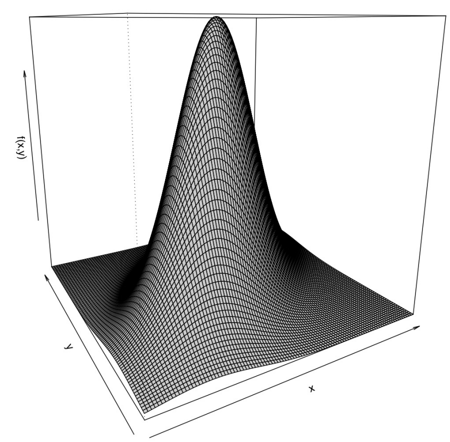
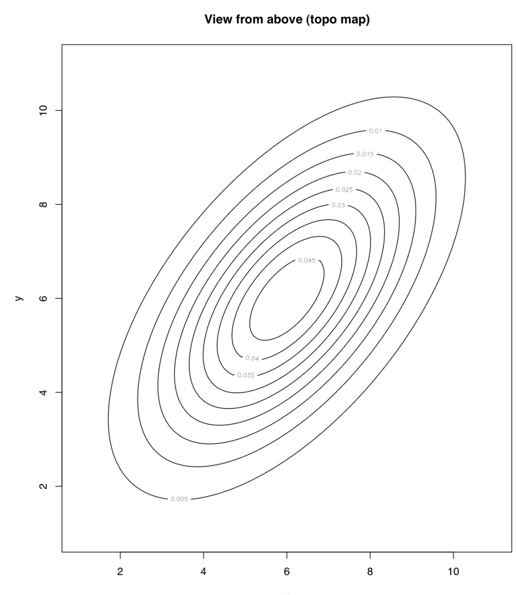
Chapter 5: JOINT PROBABILITY DISTRIBUTIONS

Part 2: Covariance and Correlation

Section 5-2

Consider the joint probability distribution $f_{XY}(x, y)$.

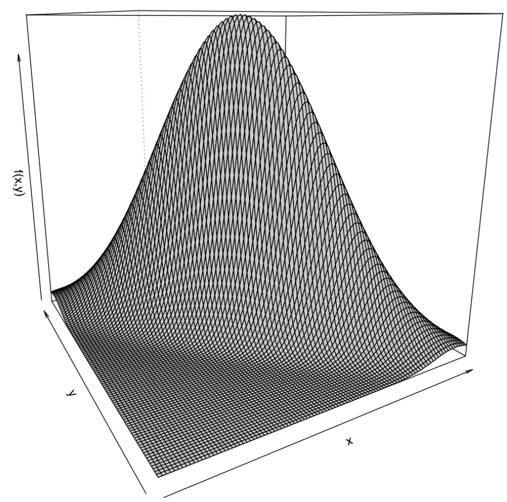


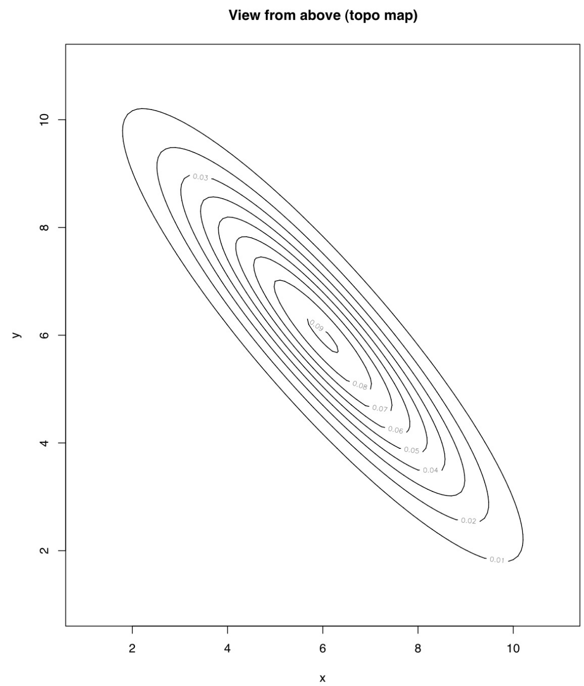


Is there a relationship between X and Y? If so, what kind?

If you're given information on X, does it give you information on the distribution of Y? (Think of a conditional distribution). Or are they independent?

Below is a different joint probability distribution for X and Y.





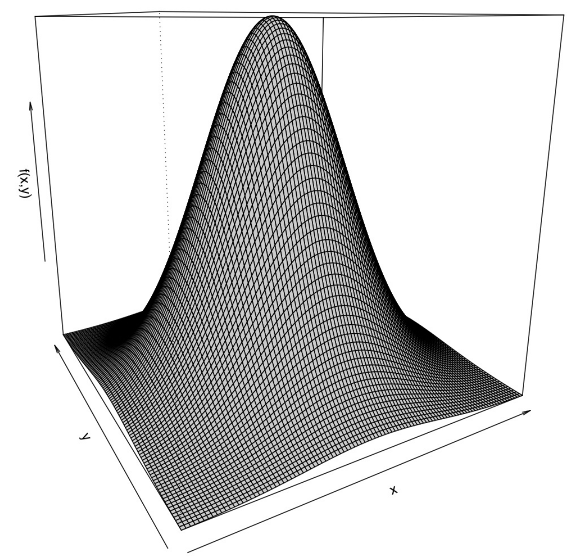
Does there seem to be a relationship between X and Y? Are they independent?

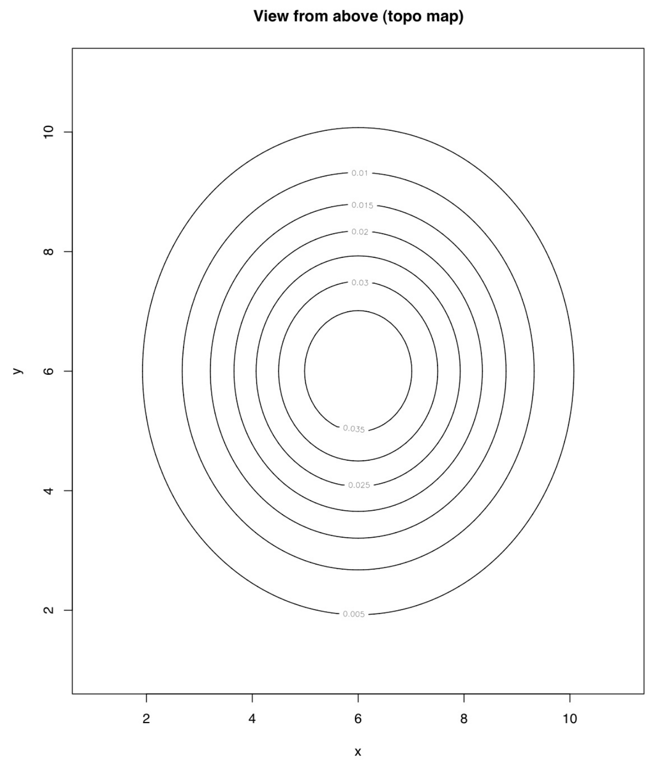
If you're given information on X, does it give you information on the distribution of Y?

How would you describe the relationship?

Is it stronger than the relationship on the previous page? Do you know MORE about Y for a given X?

Below is a joint probability distribution for an <u>independent</u> X and Y.





$\uparrow$

This picture is the give-away that they're independent.

Does there seem to be a relationship between X and Y?

If you're given information on X, does it give you information on the distribution of Y?

Covariance

When two random variables are being considered simultaneously, it is useful to describe how they relate to each other, or how they *vary* together.

A common measure of the relationship between two random variables is the **covariance**.

- **Covariance**
  The covariance between the random variables X and Y, denoted as $cov(X, Y)$, or σ_{XY}, is

$$\sigma_{XY} = E[(X - E(X))(Y - E(Y))]$$

$$= E[(X - \mu_X)(Y - \mu_Y)]$$

$$= E(XY) - E(X)E(Y)$$

$$= E(XY) - \mu_X \mu_Y$$

To calculate covariance, we need to find the expected value of a function of X and Y. This is done similarly to how it was done in the univariate case...

For X, Y discrete,

$$E[h(x,y)] = \sum_x \sum_y h(x,y) f_{XY}(x,y)$$

For X, Y continuous,

$$E[h(x,y)] = \int \int h(x,y) f_{XY}(x,y) dx dy$$

Covariance (i.e. σ_{XY}) is an expected value of a function of X and Y over the (X, Y) space, if X and Y are continuous we can write

$$\sigma_{XY} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{XY}(x,y)\, dx\, dy$$

To compute covariance, you'll probably use...

$$\boxed{\sigma_{XY} = E(XY) - E(X)E(Y)}$$

When does the **covariance** have a positive value?

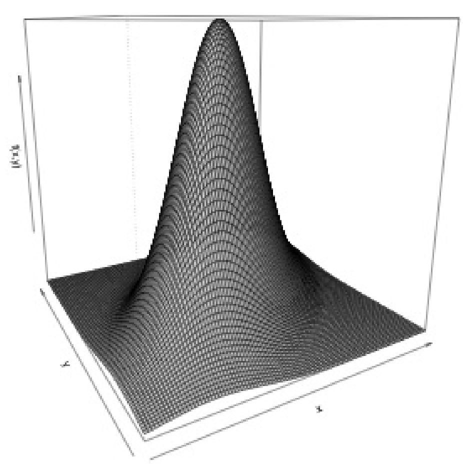
In the integration we're conceptually putting 'weight' on values of $(x - \mu_X)(y - \mu_Y)$.

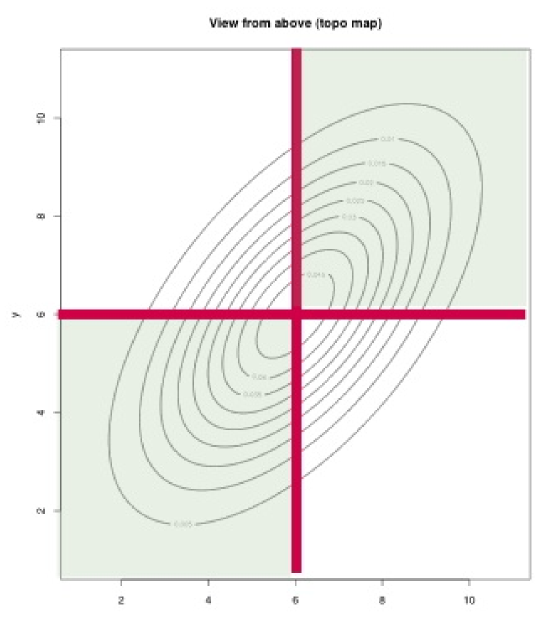
What regions of (X, Y) space has...

$$(x - \mu_X)(y - \mu_Y) > 0?$$

- Both X and Y are above their means.
- Both X and Y are below their means.
- $\Rightarrow$ Values along a line of positive slope.

A distribution that puts high probability on these regions will have a **positive covariance**.





When does the **covariance** have a <u>negative</u> value?

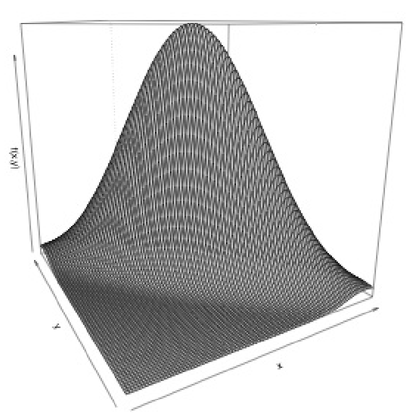
In the integration we're conceptually putting 'weight' on values of $(x - \mu_X)(y - \mu_Y)$.

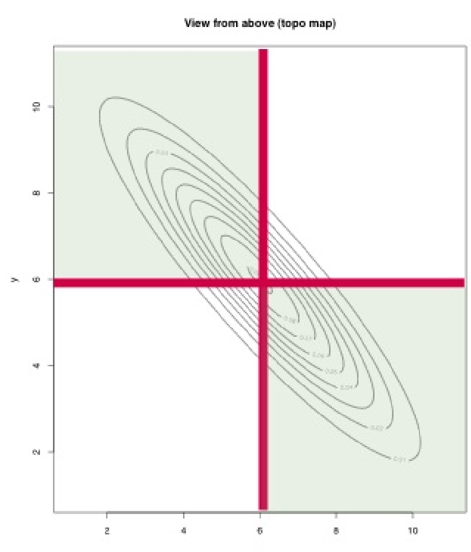
What regions of (X, Y) space has...

$$(x - \mu_X)(y - \mu_Y) < 0?$$

- X is above its mean, and Y is below its mean.
- Y is above its mean, and X is below its mean.
- $\Rightarrow$ Values along a line of negative slope.

A distribution that puts high probability on these regions will have a **negative covariance**.





Covariance is a measure of the linear relationship between X and Y.

If there is a non-linear relationship between X and Y (such as a quadratic relationship), the covariance may not be sensitive to this.

When does the **covariance** have a zero value?

This can happen in a number of situations, but there's one situation that is of large interest... when X and Y are independent...

When X and Y are independent, $\sigma_{XY} = 0$.

If X and Y are independent, then...

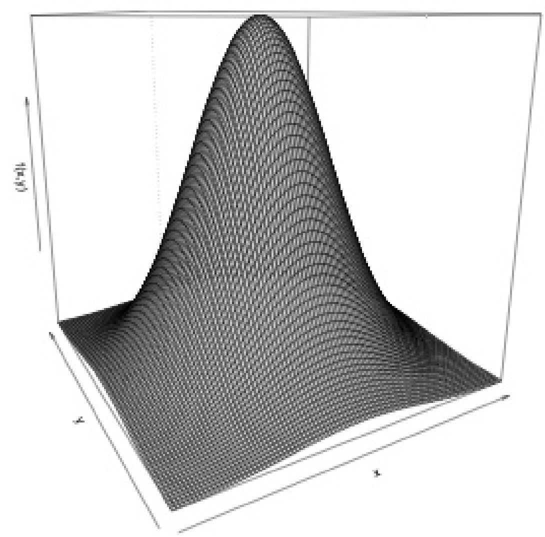
$$\begin{aligned}
\sigma_{XY} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)\boldsymbol{f_{XY}(x,y)}\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)\boldsymbol{f_X(x)f_Y(y)}\,dx\,dy \\
&= \left(\int_{-\infty}^{\infty}(x-\mu_X)f_X(x)dx\right)\cdot\left(\int_{-\infty}^{\infty}(y-\mu_Y)f_Y(y)dy\right) \\
&= \left(\int_{-\infty}^{\infty}xf_X(x)dx-\mu_X\right)\cdot\left(\int_{-\infty}^{\infty}yf_Y(y)dy-\mu_Y\right) \\
&= (E(X)-\mu_X)\cdot(E(Y)-\mu_Y) \\
&= (\mu_X-\mu_X)\cdot(\mu_Y-\mu_Y) \\
&= 0
\end{aligned}$$

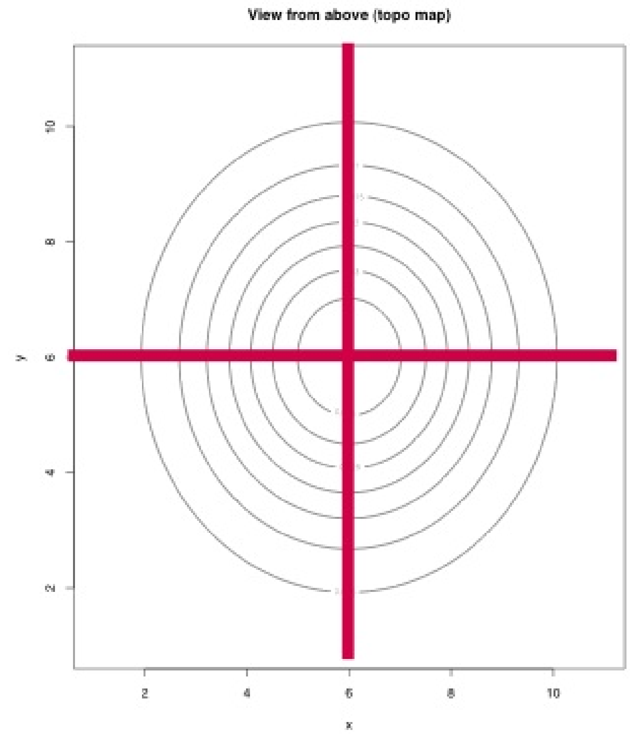
This <u>does NOT mean</u>... If covariance=0, then X and Y are independent.

We can find cases to the contrary of the above statement, like when there is a strong quadratic relationship between X and Y (so they're not independent), but you can still get $\sigma_{XY} = 0$.

Remember that covariance specifically looks for a linear relationship.

When X and Y are independent, $\sigma_{XY} = 0$.





For this distribution showing independence, there is equal weight along the positive and negative diagonals.

A couple comments...

- You can also define covariance for discrete X and Y:

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

$$= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f_{XY}(x, y)$$

- And recall that you can get the expected value of any function of X and Y:

$$E[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f_{XY}(x, y) \, dx \, dy$$

or

$$E[h(X, Y)] = \sum_x \sum_y h(x, y) f_{XY}(x, y)$$

Correlation

Covariance is a measure of the linear relationship between two variables, but perhaps a more common and more easily interpretable measure is correlation.

- **Correlation**
  The correlation (or correlation coefficient) between random variables X and Y, denoted as ρ_{XY}, is

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

  Notice that the numerator is the covariance, but it's now been scaled according to the standard deviation of X and Y (which are both > 0), we're just *scaling* the covariance.

  NOTE: Covariance and correlation will have the same sign (positive or negative).

Correlation lies in $[-1, 1]$, in other words,

$$-1 \leq \rho_{XY} \leq +1$$

Correlation is a *unitless* (or dimensionless) quantity.

Correlation...

- $-1 \leq \rho_{XY} \leq +1$
- If X and Y have a strong positive linear relation ρ_{XY} is near $+1$.
- If X and Y have a strong negative linear relation ρ_{XY} is near -1.
- If X and Y have a non-zero correlation, they are said to be correlated.
- Correlation is a measure of linear relationship.
- If X and Y are independent, $\rho_{XY} = 0$.

- **Example**: Recall the particle movement model

An article describes a model for the movement of a particle. Assume that a particle moves within the region A bounded by the x axis, the line $x = 1$, and the line $y = x$. Let (X, Y) denote the position of the particle at a given time. The joint density of X and Y is given by

$$f_{XY}(x, y) = 8xy \quad \text{for} \quad (x, y) \in A$$

a) Find $cov(X, Y)$

ANS: Earlier, we found $E(X) = \frac{4}{5}$ ...

- **Example**: Book problem 5-43 p. 179.

  The joint probability distribution is

x	-1	0	0	1
y	0	-1	1	0
f_{XY}	0.25	0.25	0.25	0.25

  Show that the correlation between X and Y is zero, but X and Y are not independent.