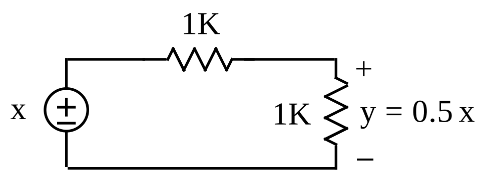
ECE 315 - DISCRETE RANDOM VARIABLES - INVEST 11
CORRELATION AND COVARIANCE

WINTER 2004 **A.P. FELZER**

To do "well" on this investigation you must not only get the right answers but must also do neat, complete and concise writeups that make obvious what each problem is, how you're solving the problem and what your answer is. You also need to include drawings of all circuits as well as appropriate graphs and tables.

In the last Investigation we introduced joint probability distributions $f_{XY}(x,y)$. The objective of this Investigation is to see if we can find how the random variables X and Y are related. Now when we analyze a circuit like the following

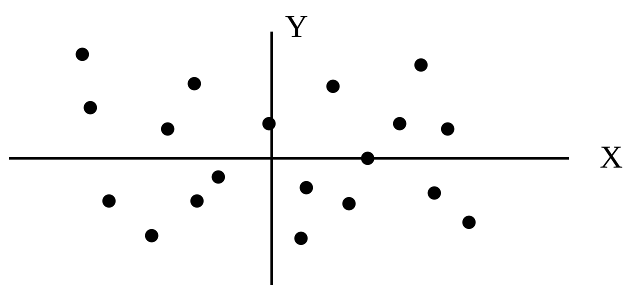


we come up with an exact equation for the relationship between y and x. If we know x we can calculate y exactly. But this is not in general the case for random variables Y and X. The best we can generally do is find the most probable value of Y for a given value of X. We call this **regression analysis**. In this investigation we focus on **linear regression** where we find which straight line as follows

$$Y = aX + b$$

gives us the "most" probable values of Y for given values of X.

1. One way to "see" how two random variables X and Y are related is to draw **scatter plots** as follows

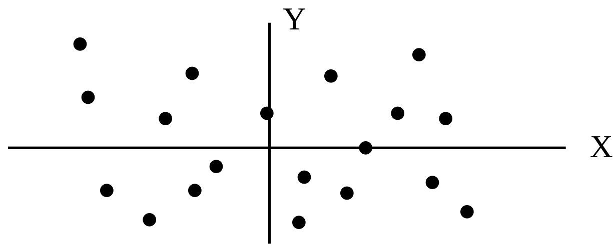


with the coordinates of each dot equal to the values of X and Y when we do a random experiment. Suppose in particular that we flip coin A three times and coin B two times with

X = Number of heads when we flip coin A
Y = Number of heads when we flip coin B

Do this coin flipping experiment five times to obtain five points on a scatter plot

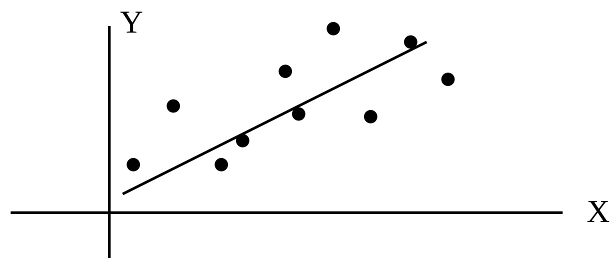
2. Given a scatter plot as follows



with the coordinates of each dot equal to the values of X and Y when we do a random experiment like the coin flipping of Problem (1)

a. How can you tell from the scatter plot that X and Y are not linearly related
b. Draw a scatter plot for X and Y if they are approximately linearly related
c. Draw a scatter plot for X and Y if they're exactly linearly related

3. The objective of this problem is to introduce the criteria of minimum squared error (MSE) for finding the equation for the **line of regression** - the straight line that best approximates the relationship between Y and X in scatter plots like the following

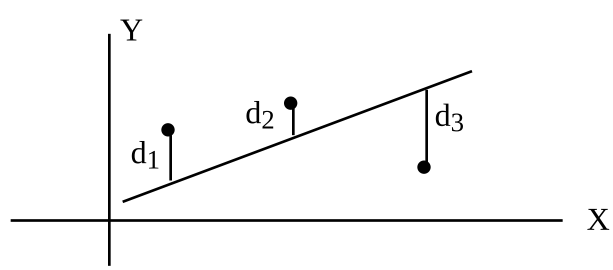


by choosing *a* and *b* in the linear equation Y = aX + b that minimizes the average of the squares of the errors between the actual values of Y and the linear approximation as follows

$$E\left[\left(Y-(aX+b)\right)^2\right] = \text{Average Squared Error}$$

We use this criteria because it not only makes sense but also because the analysis is relatively straightforward. We then call Y = aX + b the **line of regression** for Y and X.

The line of regression for the random variables Y and X with a scatter plot obtained from a whole bunch of random experiments as follows



is the line for which

$$\text{Average Squared Error} = \text{Average}\left[d_1^2, d_2^2, d_3^2\right]$$

is minimum. Given the following data points (x,y) on the scatter plot

(2, 4) (1, 2) (3, 2)

a. Find the average squared error for Y = X + 1
b. Find the average squared error for Y = 2X + 1
c. Which line is better - has a lower average squared error

4. In order to get nice "compact" expressions for the parameters a and b in the equation for the line of regression as follows

$$Y = aX + b$$

it's helpful to define what we call the **covariance** of two random variables as follows

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

a. Describe in words the equation for the covariance of two random variables
b. Calculate the covariance of the random variables X and Y with joint probability distribution function

X \ Y	2	4
1	0.3	0.1
2	0.4	0.2

c. Calculate the covariance of the random variables X and Y with Y = 2X and joint probability distribution function

X \ Y	2	4
1	0.4	0
2	0	0.6

d. Calculate the covariance of the random variables X and Y with Y = -2X and joint probability distribution function

X \ Y	–2	–4
1	0.4	0
2	0	0.6

e. Explain why the magnitude of the covariance is larger when Y and X are linearly related than when they're not linearly related. Hint - look at the signs of the products $(X - E[X])(Y - E[Y])$

5. Derive the following expression and then use it to double check your results in Problem (4)

$$Cov(X,Y) = E\big[(X - E[X])(Y - E[Y])\big] = E[XY] - E[X]E[Y]$$

6. Make use of the fact that independent random variables X and Y can be shown to satisfy

$$E[XY] = E[X]E[Y]$$

to show that $Cov(X,Y) = 0$ for independent random variables

7. If we now go back to our line of regression given by

$$Y = aX + b$$

and go through the analysis to find where $M = E\left[(Y-(aX+b))^2\right]$ is minimum - where

$$\frac{\partial M}{\partial a} = 0 \quad \text{and} \quad \frac{\partial M}{\partial b} = 0$$

we find that the mean squared error will be minimum when

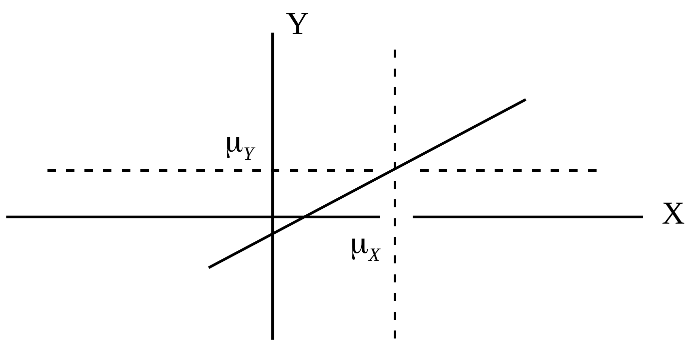
$$a = \rho\,\frac{\sigma_Y}{\sigma_X} \qquad\qquad b = \mu_Y - a\mu_X$$

where $\sigma_X = \sqrt{Var(X)}$ and $\sigma_Y = \sqrt{Var(Y)}$ and ρ is the **correlation coefficient** of X and Y as given by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

Note that the correlation is really just a "normalized" covariance that can be shown to stay in the range $-1 \le \rho \le 1$

a. Verify that the line of regression goes through the point (μ_x, μ_y) as shown in the following graph



b. Make use of the above relations to find the line of regression Y = aX + b for the random variables X and Y with joint probability distribution as follows

X \ Y	2	4
1	0.3	0.1
2	0.4	0.2

8. The correlation coefficient ρ(X, Y) is not only useful because it simplifies the expression for the paramater *a* in the line of regression but also for the information it gives us about how close two random variables X and Y are to being linearly related.

a. First show that if Y = aX with a > 0 then ρ(X,Y) = 1

b. Now show that if Y = aX with a < 0 then ρ(X,Y) = –1

c. Generalizing on the results of parts (a) and (b) it can be shown that if X and Y are linearly related with $Y = aX + b$ then $\rho(X,Y) = \pm 1$ depending on the sign of a. How does knowing that $\rho(X,Y) = \pm 1$ enable you to predict the value of Y from the value of X

d. Now suppose that $\rho(X,Y) = 0$. What's the corresponding equation for the line of regression.

e. How does knowing the value of X help you predict the value of Y when $\rho(X,Y) = 0$. Hint - make use of your result in part (d)

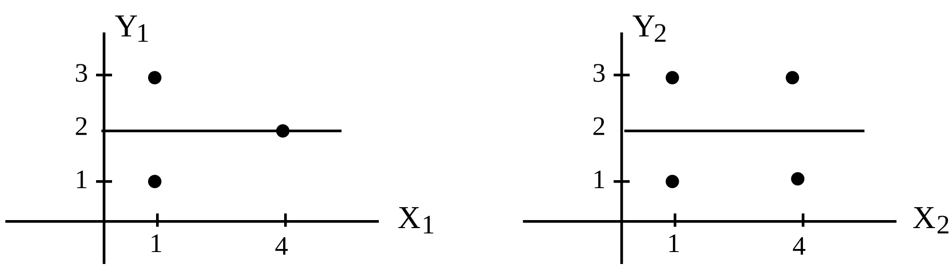
f. Draw a scatter plot for random variables X and Y if $\rho(X,Y) \approx 1$

g. Draw a scatter plot for random variables X and Y if $\rho(X,Y) \approx -1$

h. Draw a scatter plot for random variables X and Y if $\rho(X,Y) = 0$. Note that when $\rho(X,Y) = 0$ then the best estimate of Y is it's average μ_y - no matter what the value of X

9. From the results of Problem (8) we say that two random variables X and Y are *correlated* if $\rho = 1$ or is "close" to 1. And *uncorrelated* if $\rho = 0$. Show that if two random variables X and Y are independent then they're uncorrelated - the correlation coefficient is zero

10. Given the following two scatter plots obtained after doing the random experiments a "whole bunch" of times

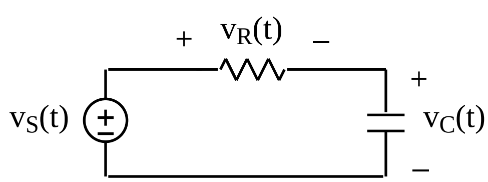


a. Make use of the result from Problem (8h) to explain why both correlations are zero - that the best estimate for Y for any value of X is simply the expectation of Y

b. Which pair of random variables is independent and which is not. Hint - for each pair of random variables test whether $P(Y \mid X) = P(Y)$

c. What can you conclude from parts (a) and (b) about the relationship between correlation and independence - about whether zero correlation implies independence

11. Review Problem: Given the following circuit



Sketch the magnitude of each of the following transfer functions as a function of frequency ω

a. $|G_C(j\omega)| = \left|\dfrac{V_C(j\omega)}{V_S}\right|$

b. $|G_R(j\omega)| = \left|\dfrac{V_R(j\omega)}{V_S}\right|$