1. Let us reconsider the earlier example about automobile and homeowner policy deductibles. The joint and marginal pdfs are given as:

$p(x, y)$		y: 0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

x	100	250
$p_X(x)$	.5	.5

y	0	100	200
$p_Y(y)$	.25	.25	.5

What is the $Cov(X, Y)$?

Solution:

We note that $\mu_X = \sum xp_X(x) = 175$ and $\mu_Y = 125$. Thus,

$$Cov(X,Y) = \sum_x \sum_y (x-175)(y-125)p(x,y)$$

$$= (100-175)(0-125)(0.20)+...+(250-175)(200-125)(0.30) = 1875$$

2. Returning to the earlier problem about three different chocolate types A, B, and C in a 1 lb box, the joint and marginal pdf's of X= weight of type A chocolates and Y= weight of type B chocolates are:

$$f(x,y) = \begin{cases} 24xy & 0 \leq x \leq 1, 0 \leq y \leq 1, x+y \leq 1 \\ 0, & \text{otherwsie} \end{cases}$$

$$f_X(x) = \begin{cases} 12x(1-x)^2 & 0 \leq x \leq 1 \\ 0, & \text{otherwsie} \end{cases}$$

$$f_Y(y) = \begin{cases} 12y(1-y)^2 & 0 \leq y \leq 1 \\ 0, & \text{otherwsie} \end{cases}$$

(a) What is $Cov(X,Y)$?

(b) Does the covariance indicate a stronger or weaker relationship than what we found in the last problem about the insurance policies?

Solution:

(a) We can easily find that $\mu_x = \mu_Y = \frac{2}{5}$ and

$$E(XY) = \int_0^1 \int_0^{1-x} xy \cdot 24xy \quad dydx = 8 \int_0^1 x^2(1-x)^3 dx = \frac{2}{15}$$

Thus

$$Cov(X,Y) = \frac{2}{15} - (\frac{2}{5})(\frac{2}{5}) = \frac{2}{15} - \frac{4}{25} = -\frac{2}{75}$$

The negative sign indicates that if there are more type A chocolates in the box, there will be less type B chocolates.

(b) Given the dependence of covariance calculation on units of measurement, it is impossible to compare two or more covariance values to decide which is a stronger relationship!

Solution:

We note that $\mu_X = \sum xp_X(x) = 175$ and $\mu_Y = 125$. Thus,

$$Cov(X,Y) = \sum\sum_{(} x,y)(x-175)(y-125)p(x,y)$$

$= (100-175)(0-125)(0.20) + ... + (250-175)(200-125)(0.30) = 1875$

Now we can easily verify that

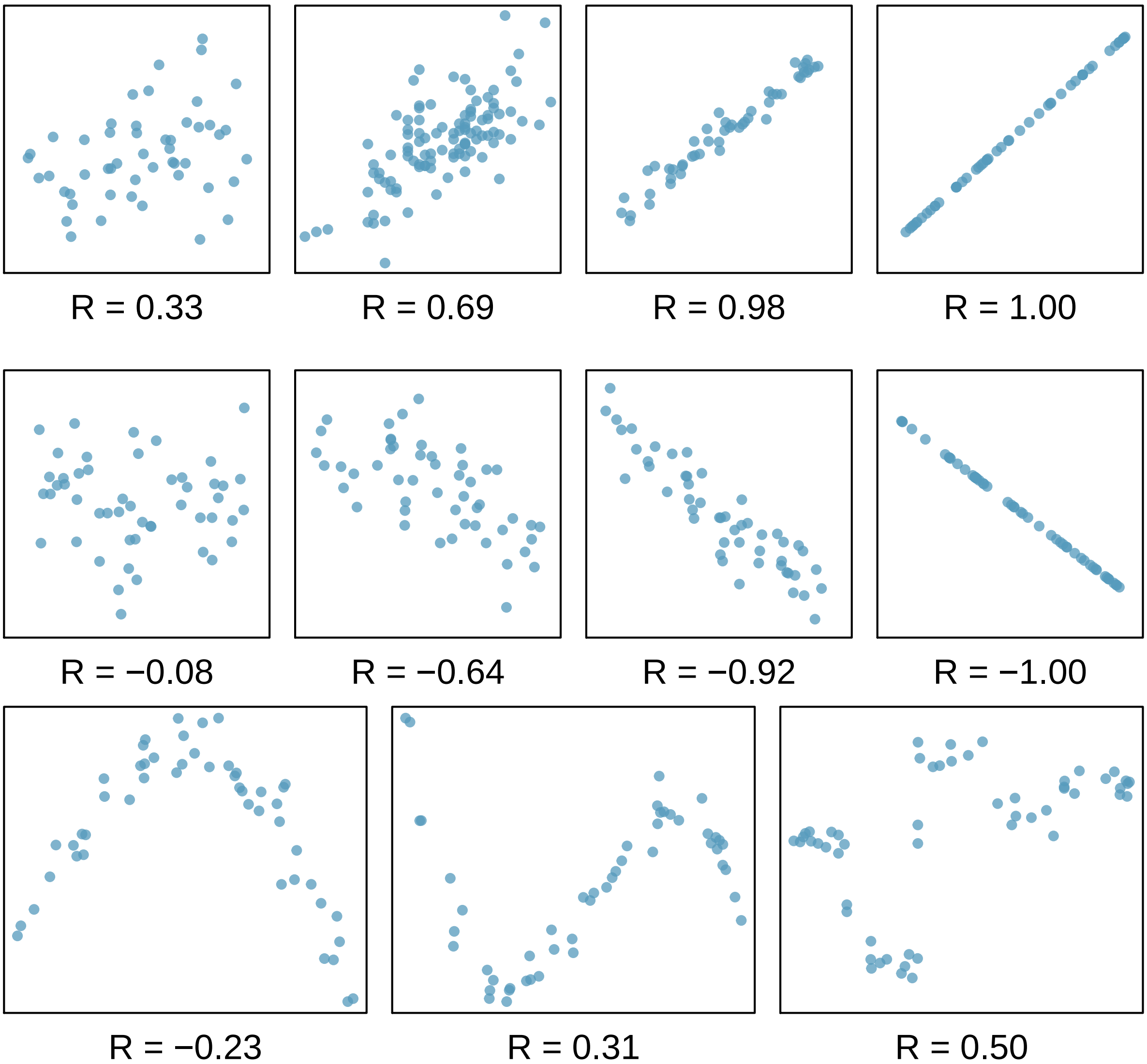
$$E(X^2) = 36250, \sigma_X^2 = 36250 - (175)^2 = 5625$$

$$\sigma_X = 75, E(Y^2) = 22500, \sigma_Y^2 = 6875, \sigma_Y = 82.92$$

Thus,

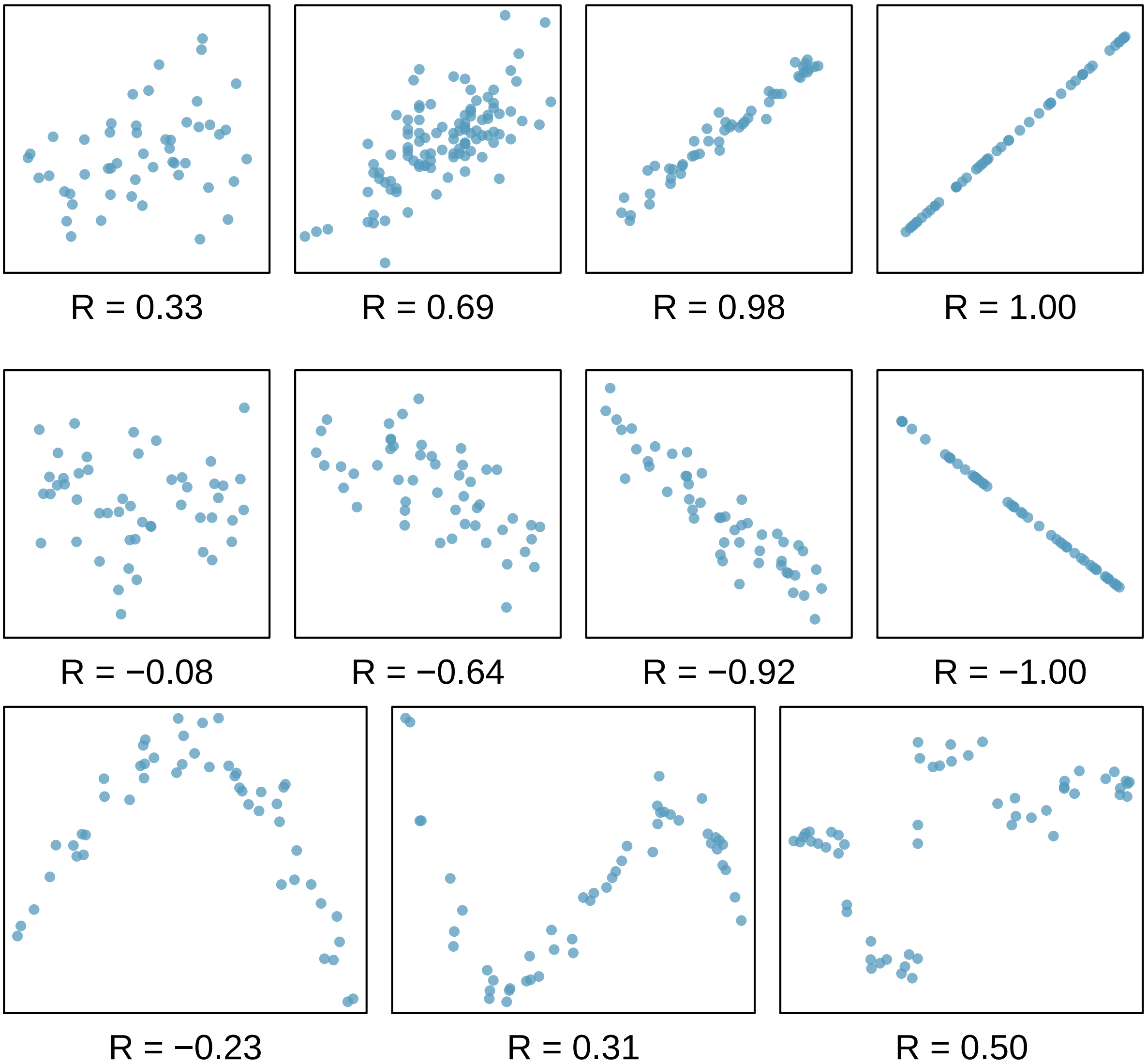
$$\rho = \frac{1875}{(75)(82.92)} = 0.301$$

We have plotted pairs of (x, y) values in *scatterplots*[1]. What do you see in these relationships?



[1]`http://en.wikipedia.org/wiki/Scatter\_plot`

Solution: First row - positive relationship; second row negative relationship; third row - nonlinear relationships.



Let us assume the random variables X and Y follow the joint distribution:

$$f(x, y) = \begin{cases} 2 & 0 \leq x \leq y < 1 \\ 0, & \text{otherwsie} \end{cases}$$

What is the correlation coefficient between X and Y?

Solution:

We first compute the marginal distributions:

$$f_X(x) = \begin{cases} 2(1-x) & 0 < x < 1 \\ 0, & \text{otherwsie} \end{cases}$$

$$f_Y(y) = \begin{cases} 2y & 0 < y < 1 \\ 0, & \text{otherwsie} \end{cases}$$

Then applying the relevant formulas, we find

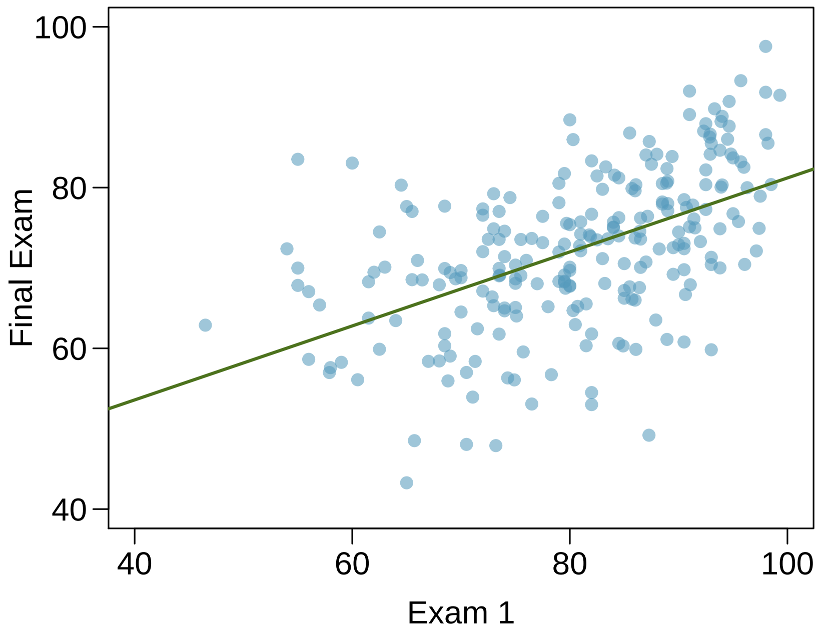
$$E(X) = \frac{1}{3}, E(Y) = \frac{2}{3}, Var(X) = Var(Y) = \frac{1}{18}, E(XY) = \frac{1}{4}$$

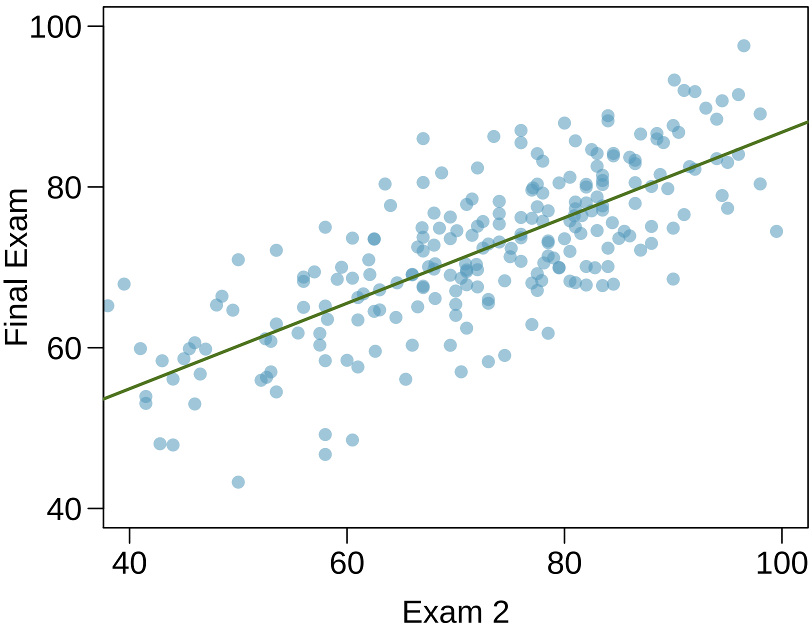
Thus, $\rho = \frac{1}{2}$

Finals week is coming close and I am worrying about how to set the "right" questions. The two scatterplots below show the relationship between the Final and two intermediate examination (Exam 1 and Exam 2; Exam 1 was conducted earlier in the term than Exam 2) grades recorded during a similar course I have taught at another university.

(a) Based on these plots, which of the two exams has the strongest correlation with the final exam grade? Explain.

(b) Can you think of a reason why the correlation between the exam you chose in part (a) and the Final exam is higher?





100
80
60
40
Final Exam
40
60
80
100
Exam 2

Solution:

(a) Exam 2; since there is less of a scatter in the plot of Final exam grade versus Exam 2. Notice that the relationship between Exam 1 and the Final Exam appears to be slightly nonlinear.

(b) There can be different interpretations. One interpretation can be that Exam 2 and the Final are relatively close to each other chronologically, so students who prepared well for Exam 2 were also better prepared for the Final exam.

Zero Correlation does not imply independence. (however, this is true for normal distributions ;-))
Consider the following probability density for X_1, X_2 for a fixed value $n \geq 1$

$$f(X_1 = x_1, X_2 = x_2) = \tag{1}$$

$$= \begin{cases} \frac{1}{2\pi} & \text{if } (x_1, x_2) = (r\cos(t), r\sin(t)),\ t \in [0, 2\pi), r \in [1, \sqrt{3}] \\ 0 & \text{for all other points } (x_1, x_2) \in \mathbb{R}^2 \end{cases} \tag{2}$$

- Validate that each point X_1, X_2 with a non-zero probability is a point on a circle with radius $r \in [1, \sqrt{3}]$.
- If you wonder what it is, quickly plot it in matplotlib
- Show that its Covariance and Correlation Coefficient are zero.
- Show that X_1 and X_2 are not independent

The moral of all the math below: any circle disc with radii in some interval is an example for zero correlation but non-independence.

Solution:

Validation:

points (x_1, x_2) on a circle with radius r must satisfy the equation: $x_1^2 + x_2^2 = r^2$.

$$x_1^2 + x_2^2 = r^2 \cos^2(t) + r^2 \sin^2(t) = r^2, r \in [1, \sqrt{3}) \text{ by definition} \tag{3}$$

We used here $\cos^2(x) + \sin^2(x) = 1$.

Covariance:

We have

$$Cov(X_1, X_2) = E[X_1 \cdot X_2] - E[X_1] \cdot E[X_2] \tag{4}$$

Lets consider what $E[X_1]$ and $E[X_2]$ must be.

It is clear that the mean of all x coordinates on a circle with a fixed radius r around the point $(0, 0)$ is zero ... because it is symmetric around the vertical line given by $y = 0$. For every point (x_1, x_2) on the circle with fixed radius r, there is exactly one point $(-x_1, x_2)$. Why is that so ? If $x_1^2 + x_2^2 = r^2$, then it must also hold that $(-x_1)^2 + x_2^2 = r^2$. :) !!

Integrating all these points over $t \in [0, 2\pi)$ will make them cancel out: we have for each (x_1, x_2) a point $(-x_1, x_2)$. This cancellation effect holds for every fixed radius r (if you do not believe it, see how the term EX_1X_2 is calculated below ... the computation is very similar). So the mean of all x coordinates of all points on a circle around $(0, 0)$ is zero. Same argument holds for the y-coordinates. As a consequence:

$$E[X_1] = E[X_2] = 0 \tag{5}$$

what is $E[X_1 \cdot X_2]$?

$$E[X_1 \cdot X_2] = \int_{(x_1,x_2)\in A} x_1 \cdot x_2 \cdot f(x_1,x_2) dx_1 dx_2 \tag{6}$$

We use here the polar transform $(x_1, x_2) = f(r, \alpha) = (r\cos(\alpha), r\sin(\alpha))$. We have $det(Df) = r$. By integral transformation theorem we have

$$\int_{f(U)} g(z)dz = \int_U g(f(u))|det(Df)(u)|du \tag{7}$$

Applying this with $det(Df) = r$ we obtain:

$$\begin{aligned}
E[X_1 \cdot X_2] &= \int_{(x_1,x_2)\in A} x_1 \cdot x_2 \cdot f(x_1,x_2) dx_1 dx_2 && (8) \\
&= \int_{(r,\alpha)\in[1,\sqrt{3}]\times[0,2\pi)} r\cos(\alpha) r\sin(\alpha) \frac{1}{2\pi} r \; dr \; d\alpha && (9) \\
&= \int_{r=1}^{r=\sqrt{3}} r^3 \frac{1}{2\pi} \left(\int_{\alpha=0}^{\alpha=2\pi} \cos(\alpha)\sin(\alpha) \; d\alpha \right) \; dr && (10) \\
& && (11)
\end{aligned}$$

We can show (using: $2\sin(x)\cos(x) = \sin(2x)$) that

$$\begin{aligned}
&\int_{\alpha=0}^{\alpha=2\pi} \cos(\alpha)\sin(\alpha) \; d\alpha && (12) \\
=&\int_{\alpha=0}^{\alpha=2\pi} \frac{1}{2}\sin(2\alpha) \; d\alpha && (13) \\
=&\frac{1}{2} \cdot (-1) \cdot \cos(2\alpha)\Big|_0^{2\pi} && (14) \\
=&-\frac{1}{2}(\cos(4\pi) - \cos(0)) = 0 && (15)
\end{aligned}$$

So $E[X_1 \cdot X_2] = 0 = E[X_1] = E[X_2]$.

So we have: $Cov(X_1, X_2) = 0 - 0 \cdot 0$.

The variables (X_1, X_2) are uncorrelated. That is not surprising Covariance measures a linear relationship. A circle is not that linear. Compare also: "straight to the goal" versus "Walking in Circles".

So how about independence or the lack thereof? It is enough to show that the conditional density $f_{X_1|X_2}(x_1 \mid x_2)$ is not the same for two different values of x_2.

Intuitively this is clear. For $x_2 = 0$, the set of x_1 which can satisfy $x_1^2 + x_2^2 \in [1^2, \sqrt{3}^2]$ are not the same as for $x_2 = 0.5$.
For $x_2 = 0$ we have that $x_1^2 + x_2^2 \in [1^2, \sqrt{3}^2]$ if $x_1 \in [-\sqrt{3}, -1]$ or $x_1 \in [1, \sqrt{3}]$. So the conditional density for x_1 is non-zero on these two intervals.
For $x_2 = 0.5$ we have that

$x_1^2 \in [1 - 0.25, 3 - 0.25] = [0.75, 2.75]$, so the conditional density for x_1 is non-zero on $x_1 \in [-\sqrt{2.75}, -\sqrt{0.75}]$ or $x_1 \in [\sqrt{0.75}, \sqrt{2.75}]$