Joint probability is the probability that the RVs X & Y take values x & y. like the PDF of the two events, x and y. We will denote a joint probability function as

$P_{X,Y}(x,y) = P(X=x \cap Y=y)$

- **Marginal probability** of X is the probability that RV X has the value x regardless of the value of Y. That is

$$P_X(x) = P(X = x) = \sum_y P_{X,Y}(x,y)$$

Example: Let X represent the number of weekly credit card purchases a person makes, and Y the number of credit cards a person owns. Suppose the bivariate table for the two variables looks as follows:

		X				
		0	1	2	3	
	1	.08	.10	.10	.02	.30
Y	2	.08	.05	.22	.05	.40
	3	.04	.04	.04	.18	.30
		.20	.19	.36	.25	1.0

So, using the new notation, $P_{X,Y}(0,1) = .08$ This is the value which the joint probability function for X and Y takes when X=0 and Y=1.

The marginal probability of X is the probability that a randomly selected person makes a certain number of credit card purchases per week, for example $P_X(2)$ = the probability that a randomly selected person makes 2 credit card purchases per week,

$$P_X(2) = \sum_y P_{X,Y}(2, y) = .10 + .22 + .04 = .36$$

The complete marginal probability function of X is:

$P_X(0) = .2$, $P_X(1) = .19$, $P_X(2) = .36$, and $P_X(3) = .25$.

- We can also rewrite the definition of conditional probability using this new notation:

$$P_{Y|X}(2,3)=P(Y=2 \mid X=3)=\frac{P_{X,Y}(2,3)}{P_X(3)} = \frac{.05}{.25} =.2.$$

The interpretation of this is, as we know from previous discussion of conditional probability, that the probability that someone who makes three weekly credit card purchases owns two credit cards is .2.

- We can also rewrite the definitions of mutually exclusive events using this notation,

  If $P_{X,Y}(x,y)=0$ for all pairs x and y, then X and Y are mutually exclusive, otherwise they are not.

- And the definition of independent events,

  If $P_{X,Y}(x,y)=P_X(x)P_Y(y)$ for all pairs x and y then X and Y are independent, otherwise they are not.

For single variable distributions, we defined cumulative probability functions. We may also be interested in the probability over some range for bivariate distributions. For example, we might want to know what proportion of the population owns 2 or fewer credit cards and makes 2 or fewer credit card purchases per week, i.e. we might be interested in the following cells:

	0	1	2	3
1	.08	.10	.10	.02
2	.08	.05	.22	.05
3	.04	.04	.04	.18

- The **joint cumulative probability function** defines the probability that X takes a value less than or equal to x and Y takes a value less than of equal to y,

  $$F_{X,Y}(a,b) = P_{X,Y}(X\le a,\ Y\le b) = \sum_{X\le a}\sum_{Y\le b} P_{X,Y}(x,y)$$

Example: In the above example, $F_{X,Y}(2,2) = .08+.10+.10+.08+.05+.22 = .63$. The interpretation is that 63% own two or fewer cards and make 2 or fewer purchases per week. Or, the probability that a randomly selected person owns two or fewer cards and used them to make two or fewer purchases per week is .63.

- Joint cumulative probability is a crucial concept when we are interested in the joint probability of two or more continuous RV's. It defines the probability that the each of the variables falls into some given interval.

Remember that the probability that a continuous RV is equal to any particular value is zero.

B. Covariance and correlation

We want to use bivariate probability distributions to talk about the relationship between two variables. The test for independence tells us whether or not two variables are independent. We also want to know how two variables are related if they are not independent, e.g. if income and education are not independent, do people with more education tend to have higher incomes or lower incomes?

- **Covariance** measures the average product of deviations of two variables from their respective means. It offers a measure of how two variables co-vary, whether one is above (below) its mean when the other is above (below) its mean.

$$Cov(X,Y) = \sigma_{XY} = E[(X-\mu_x)(Y-\mu_y)] = \sum_y \sum_x (x-\mu_x)(y-\mu_y)\, p_{X,Y}(x,y)$$

- A simpler formula to use to calculate the covariance is

$$Cov(X,Y) = \sigma_{XY} = E(XY) - \mu_x\mu_y = \sum_x \sum_y xyp_{X,Y}(x,y) - \mu_x\mu_y$$

- Note that covariance is also a descriptive statistic which can be used to describe data:

Population: $$\sigma_{XY} = \frac{1}{N}\sum_y \sum_x (x-\mu_x)(y-\mu_y)$$

Sample: $$s_{XY} = \frac{1}{n-1}\sum_y \sum_x (x-\bar{x})(y-\bar{y})$$

- The simplified equations for the descriptive statistic are

Population: $$\sigma_{XY} = \frac{1}{N}\sum_y \sum_x xy - \mu_x\mu_y$$

Sample: $$s_{XY} = \sum_x \sum_y \frac{xy}{n-1} - \frac{n}{n-1}\overline{xy}$$

Covariance is a measure of simultaneous movements in X and Y. In other words, the value of the covariance answers the question “when X is above its mean, where do we expect Y to lie with respect to its mean?”

- Example: consider the credit card example again

		X				
		0	1	2	3	
	1	.08	.10	.10	.02	.30
Y	2	.08	.05	.22	.05	.40
	3	.04	.04	.04	.18	.30
		.20	.19	.36	.25	1.0

The easiest equation to use to calculate covariance is: Cov(X,Y)=

$$= \sum_x \sum_y xyp_{x,y}(x,y) - \mu_x \mu_y .$$

What we need to do is, working from top to bottom and left to right, multiply for each cell in the table the specific value of X associated with that cell times the specific value for Y associated with that cell times the probability of the intersection of those specific values for X and Y.

For the first column, X=0, Y can be either 1, 2, or 3. Since X=0, each product of X and Y is going to be 0, so we’re done. In the next column, X = 1. That’s another convenient shortcut because for the cells in this column, when we sum up x*y*P_{XY}(x,y), we can forget about the X term. Now, once again Y can be either 1, 2, or 3, so we have:

(1*.1) + (2*.05) + (3*.04) = .1 + .1 + .12 = .32

In the next column, X = 2. No short-cut here, but since your getting the idea I'll just sum up:

(2*1*.1) + (2*2*.22) + (2*3*.04) = .2 + .88 + .24 = 1.32.

We have one more column to do, and that's where X = 3:

(3*1*.02) + (3*2*.05) + (3*3*.18) = .06 + .3 + 1.62 = 1.98.

Summing up the expressions for all four columns, we have: 0 + .32 +.1.32 + 1.98 = 3.62. That's the first half of the covariance term, or

$\sum_x \sum_y xyp_{x,y}(x,y) = 3.62.$ To complete the calculation of the covariance, all we have to do is subtract the product of the means from 3.62. Now, before we can do that, we have to calculate the means:

$\mu_X = \sum_x xp_X(x) =$ (0*.2) +(1*.19) + (2*.36) + (3*.25) = 0 + .19 + .72 + .75 = 1.66

$\mu_Y = \sum_y yp_Y(y) =$ (1*.3) +(2*.4) + (3*.3) = .3 + .8 + .9 = 2.0

Now, the product of the means is: $\mu_X\mu_Y$ = 1.66*2.0 = 3.32, so

Cov(X,Y) = 3.62 - 3.32 = 0.3

What does this mean? We want to answer questions like "when X is above its mean, where do we expect Y to lie with respect to its mean?"

- A positive covariance, like in our example, means that there is a **positive** relationship between X and Y.

What that means is that X and Y tend to move in the same direction. When X is high (above its mean), Y tends to be high as well. When Y is low (below its mean), X tends to be low as well. Some examples of things with positive covariances are: income and education, height and weight, and political power or influence and wealth.

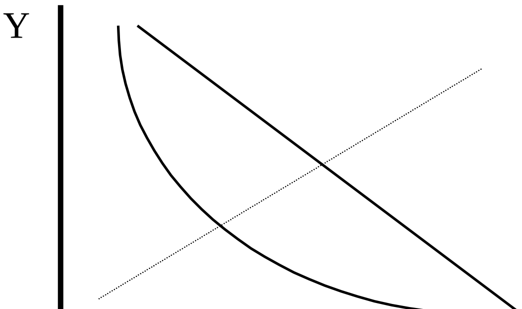
- When the covariance is negative, we say that there is a **negative** relationship between X and Y.

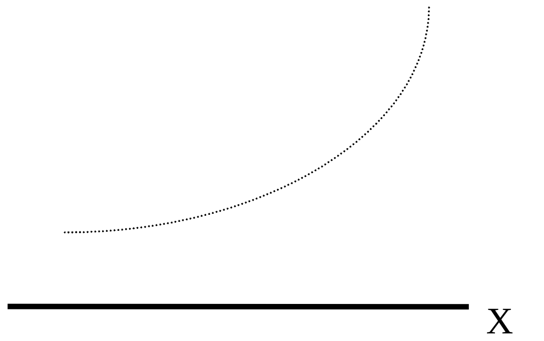
That means that X and Y tend to move in the opposite direction. When X is high (above its mean), Y tends to be low. When Y is low (below its mean), X tends to be high. Some examples of things with negative covariances are: education and poverty status, household wealth and the number of children in the family, temperature and the number of winter coats sold.

- The final possibility is that the value of the covariance is zero. In that case, we say that there is **no linear relationship** between X and Y. Note the qualification of the word **linear**.

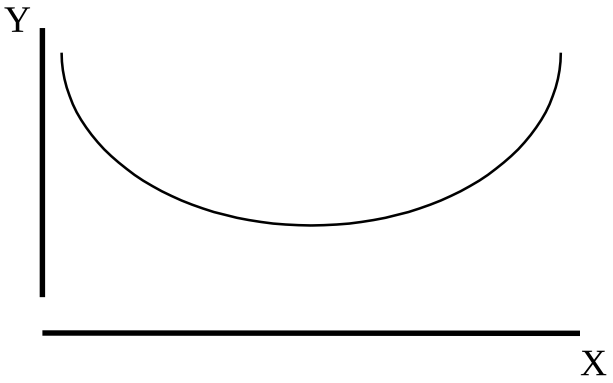
Let me explain. Sometimes a zero covariance will mean that there is no relationship between X and Y. Notice the lack of a qualification. That means that knowledge about the location of X vis-à-vis the mean of X tells us nothing about the probable location of Y. This not the only possible source of a zero covariance, and hence the qualification of linear.

Graphically, if we graphed values of X and Y, a positive relationship would look like one of the dotted lines, and a negative relationship would have a downward slope as depicted by the solid lines. What do we mean by the term linear? well, consider the following lines:





Linear means straight, so a linear relationship is a straight line. Now, even the two curved lines are moving basically in one direction. That's different from the U-shaped curve below:



This line is non-linear (or curved), and describes a specific non-linear relationship between X and Y. When X is small, as X increases, Y tends to fall, and when X is big, as X gets bigger, Y tends to get bigger as well. An example of this type of relationship is age (X) and net tax payments and government transfers (Y). So what should the sign of the covariance be? Well, depending on how negative and how positive these ranges are, we might get a covariance of zero.

So we might find that the covariance is zero even though information about one variable relative to its mean does actually tell us something about the position of the other variable relative to its mean. Covariance might be zero simply because this relationship is non-linear in such a way that positive covariance over some range exactly cancels out the negative covariance over another range.

So how can we tell these two cases apart? Well, with the tools we have now, we can't. We can plot the values and eyeball the graph, but that isn't a very precise tool. We won't discuss non-linear tools in this course, but you should be aware of the problem and remember to interpret a zero covariance as no linear relationship rather than no relationship.

- So, to sum up, the covariance tells us the direction of the linear relationship **if there is a linear or near-linear relationship**, and whether or not there is a linear relationship.
- If two variables X and Y are independent then their covariance is zero. However, if the covariance of X and Y is zero, the **variables are not necessarily independent** (this follows from the discussion above).

- Another problem with covariance is that it is **not bounded**. What that means is that the covariance can take on any value between negative infinity and positive infinity. The actual size of the covariance depends on the units of measure of the underlying random variables, X and Y. Therefore it tells us nothing about the strength of the relationship between X and Y, only the direction.
- **Example:** Suppose the average credit card purchase is $100, and we changed X from the number of purchases to the amount spend with credit cards. Instead of 0, 1, 2, and 3, X would take on the values of 0, 100, 200, and 300. The joint probability values would remain the same and

the number of credit cards owned would be the same. But everywhere we had an x in the covariance equation, we would replace it with $100*x. If we repeated the calculations we did before, we would find that now Cov(X,Y) = 30. Before, we had a covariance of .3, and now it is 100 times greater. By changing the units of X by a factor of 100, we've increased the size of the covariance by the same magnitude. So, does this represent a **stronger** positive relationship than before? No. The underlying relationship is the same.

We have a measure of the relationship between two variables which helps us to correct this problem.

- **Correlation** between two variables X and Y is the covariance divided by the product of their standard deviations.

  Population: $\rho_{X,Y} = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$ Sample: $r_{XY} = \hat{\rho}_{X,Y} = \dfrac{Cov(X,Y)}{s_X s_Y}$

- Unlike the covariance, the correlation is **bounded**. The units of the covariance are the units of X times the units of Y. If X is measured in hours and Y in dollars, the units of the covariance are hour*dollar. The correlation is the covariance by the two standard deviations, which have the same units as X and Y. So, the units cancel out and we are left with a "pure number" that has no units whatsoever.
- Correlation varies between 1 and -1. The **size** (in absolute value) of the correlation reflects the strength of the relationship, while the sign (like the covariance) represents the direction.

What this means is that whereas the covariance told us only about the direction, the correlation tells us something about how likely high (low) values of X are to be associated with high (low) values of Y. A correlation

- **Example:** let’s calculate the correlation for the credit card example

$$\sigma_X^2 = \sum x^2 p_X(x) - \mu_X^2 = (0^2 * .2) + (1^2 * .19) + (2^2 * .36) + (3^2 * .25) - 1.66^2 = 0 + .19 + 1.44 + 2.25 - 2.76 = 1.12$$

$$\sigma_Y^2 = \sum y^2 p_Y(y) - \mu_Y^2 = (1^2 * .3) + (2^2 * .4) + (3^2 * .3) - 2^2 = .3 + 1.6 + 2.7 - 4 = .6$$

$$\sigma_X = 1.06, \qquad \sigma_Y = .77$$

This means that: $\rho_{XY} = \dfrac{.3}{(1.06)(.77)} = .367$

We can say that this value indicates a stronger relationship than if the correlation were only .075, and a weaker relationship than if the value were greater in absolute value than .5, say -.73.

- Like the covariance, the correlation only captures linear relationships.

- **Neither the covariance nor the correlation tell us anything about causation.**

Suppose we observe that ice cream cone sales in a given day and the number of violent crimes committed in a given day are positively correlated. Is this a causal relationship? Which way does the causality go? Or is a third factor, e.g. the temperature outside, driving the correlation?