ECON 3818: Introduction to Statistics with Computer Applications

Kyle Butts

November 9, 2020

Chapter 4: Correlation

Multiple Variables

Almost everything we've done so far has been *univariate* statistics, but often we're interested in how multiple random variables are related?

- How does education affect earnings?
- How does race affect earnings?
- How does experience affect earnings?

Many events are *dependent* on other random variables. In this chapter we'll formalize this concept.

Probability Theory

Recall that with single random variables we characterized probabilities with

- PMF (probability mass function), $P(X = x)$, in discrete case
- PDF (probability density function), $f(x)$, in continuous case

When we have multiple random variables we use the joint distribution

- $P(X = x, Y = y)$ in the discrete case
- $f(x, y)$ in the continuous case

Properties for Joint Distribution

For short hand, $P(x, y) = P(X = x, Y = y)$

In this class we'll focus solely on the discrete case

- $0 \leq P(x, y) \leq 1$
- $\sum_x \sum_y P(x, y) = 1$

As long as X and Y are not independent

$$P(x, y) \neq P(x)P(y)$$

Example

Suppose that X is the number of girls born out of three kids and Y is whether the first child is a girl.

Outcome	X	Y
BBB	0	0
GBB	1	1
BGB	1	0
BBG	1	0
GGB	2	1
GBG	2	1
BGG	2	0
GGG	3	1

Example

Notice that the sample spaces are $S_X = \{0, 1, 2, 3\}$ and $S_Y = \{0, 1\}$. The associated joint probabilities are:

$X \backslash Y$	0	1
0	1/8	0
1	2/8	1/8
2	1/8	2/8
3	0	1/8

Example

Let's check this table satisfies the definition of a joint distribution

- $0 \leq P(x,y) \leq 1\checkmark$
- $\sum_x \sum_y P(x,y) = 1$

$$\begin{aligned}\sum_{x \in S_X} \sum_{y \in S_Y} Pr(x,y) =& Pr(0,0) + Pr(0,1) + Pr(1,0) + Pr(1,1) \\ &+ Pr(2,0) + Pr(2,1) + Pr(3,0) + Pr(3,1) \\ &= 1/8 + 0 + 2/8 + 1/8 \\ &+ 1/8 + 2/8 + 0 + 1/8 = 1\checkmark\end{aligned}$$

Clicker Question

Given the following joint probability mass function, what is the probability of the NASDAQ increasing in value and your portfolio loses value?

NASDAQ \ Portfolio	Increases	Decreases
Increases	0.40	.05
Decreases	.15	0.40

(a) 0.40

(b) 0.05

(c) 0.15

(d) 0.45

Clicker Question

Given the following joint probability mass function, what is the probability that the NASDAQ increases in value?

NASDAQ \ Portfolio	Increases	Decreases
Increases	0.40	.05
Decreases	.15	0.40

(a) 0.40

(b) 0.05

(c) 0.15

(d) 0.45

Clicker Question

Given the following joint probability mass function, what is the probability that the NASDAQ increases in value, conditional on the portfolio value decreases?

NASDAQ \ Portfolio	Increases	Decreases
Increases	0.40	.05
Decreases	.15	0.40

(a) 0.111

(b) 0.889

(c) 0.05

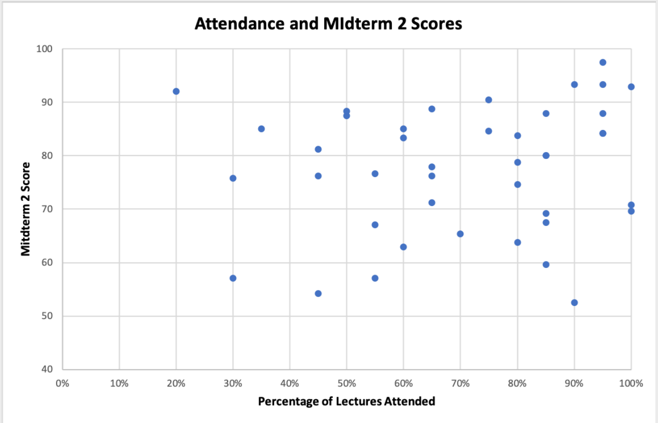
(d) 0.40

Visualizing a Joint Distribution

The most useful for displaying the relationship between two quantitative variables is a scatterplot

- Shows relationship between two quantitative variables
  - Each axis represents a variable
  - Individual data appear as a point, fixed by the values of both variables

Scatterplot Example



Interpreting a Scatterplot

- Looking for patterns, and deviations from that pattern
  - Direction, form, strength of relationship
  - Any outliers?
- Describing the association
  - Positive Association: above-average values of one tend to accompany above-average values of the other, and below-average values also tend to occur together
  - Negative Association: above-average values of one tend to accompany below-average values of the other, and vice versa
- In general, if one variable is explanatory (influences change) and one is a response variable (outcome), then the explanatory variable is plotted on the x-axis

Correlation

We need to supplement the graph with a numerical measure, generally we use correlation.

Definition (Correlation)
The correlation measures the direction and strength of the linear relationship between two quantitive variables. Correlation is usually written as r

Covariance

In order to understand correlations, we must first discuss covariance

Recall: $V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab \cdot cov(X, Y)$

Covariance measures the joint variability of two random variables

- Sign of covariance explains direction of relationship
- Magnitude of covariance is hard to interpret – hence the usage of correlation coefficient
- Covariance equals zero whenever X and Y are independent

Covariance

We use the following formula to calculate covariance

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

Note: $E(XY) \neq E(X)E(Y)$ unless X & Y are independent and then cov(X,Y)=0

The magnitude of covariance depends on the units of X & Y

- This means $cov(A, B) > cov(C, D)$ does not imply that A&B have stronger relationship than C&D
- In order to compare relationships we must find a way to normalize their covariances

Correlation

Definition (Correlation)
The correlation measures the direction and strength of the linear relationship between two quantitive variables. Correlation is usually written as r

To calculate correlation, we normalize the covariance as so:

$$r = \frac{cov(X, Y)}{\sqrt{V(X)} \cdot \sqrt{V(Y)}}$$

Correlation

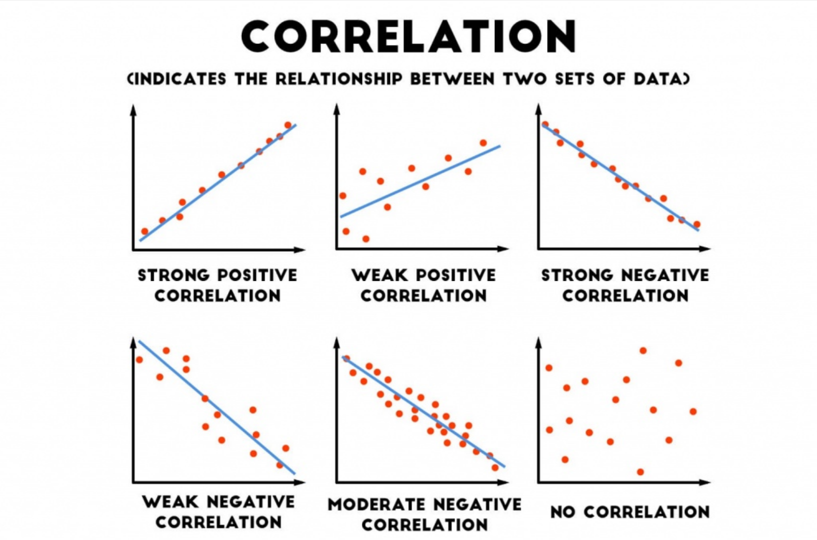
Notes on correlation

- Values are always between -1 and 1
  - $1 \rightarrow$ perfectly linear positive relationship (variables move same direction and same magnitude)
  - $-1 \rightarrow$ pefectly linear negative relationship (variables move in opposite direction but same magnitude)
- Correlations are unit-less
- Doesn't imply a causal relationship

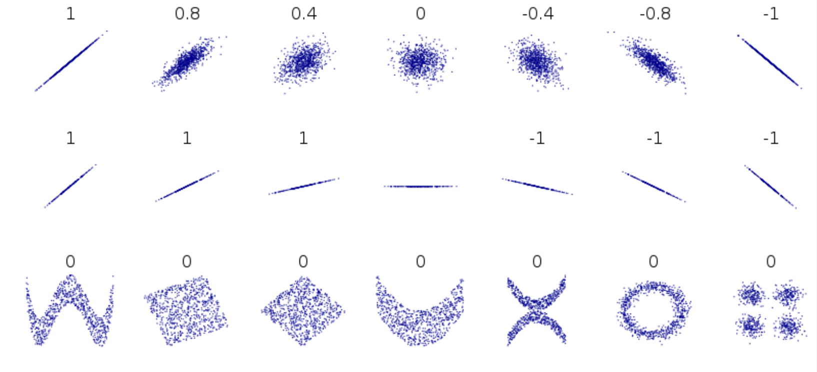
Drawbacks of correlation

- Only measures *linear relationships*
  - Just because correlation is zero doesn't necessarily mean variables are independent
- Not resistant to outliers

Correlations - Visualized



Why Correlation isn't Perfect



Covariance and Independence

Since covariance (and correlations) only measure linear relationships:

$$cov(X, Y) = 0 \nrightarrow X\&Y \text{ are independent}$$

However, since $E(XY) = E(X)E(Y)$ when X & Y are independent:

$$X\&Y \text{ are independent} \rightarrow cov(X, Y) = 0$$

Joint Distributions

When calculating the covariance we use equation $cov(X, Y) = E(XY) - E(X)E(Y)$

$X \backslash Y$	0	1
0	1/8	0
1	2/8	1/8
2	1/8	2/8
3	0	1/8

$$E(XY) = x \cdot y \cdot P(x, y)$$

In this example:

$$E(XY) = (0 \cdot 0 \cdot 1/8) + (0 \cdot 1 \cdot 0) + (1 \cdot 0 \cdot 2/8) + (1 \cdot 1 \cdot 1/8) +$$
$$(2 \cdot 0 \cdot 1/8) + (2 \cdot 1 \cdot 2/8) + (3 \cdot 0 \cdot 0) + (3 \cdot 1 \cdot 1/8) = 1$$

Marginal Probabilties

In order to calculate $E(X)$ and $E(Y)$ from a joint distribution we must first calculate the marginal probabilities of both X and Y.

$X \backslash Y$	0	1	$Pr(X)$
0	1/8	0	1/8
1	2/8	1/8	3/8
2	1/8	2/8	3/8
3	0	1/8	1/8
$Pr(Y)$	4/8	4/8	1

These marginal probabilities, $P(X = x)$are calculated adding up the probabilities across each scenario where $X = x$

Marginal Probabilities

We can use these marginal probabilities to calculate $E(X)$ and $E(Y)$.

$X \backslash Y$	0	1	$Pr(X)$
0	1/8	0	1/8
1	2/8	1/8	3/8
2	1/8	2/8	3/8
3	0	1/8	1/8
$Pr(Y)$	4/8	4/8	1

$$E(X) = (0 \cdot 1/8) + (1 \cdot 3/8) + (2 \cdot 3/8) + (3 \cdot 1/8) = 1.5$$
$$E(Y) = (0 \cdot 4/8) + (1 \cdot 4/8) = 0.5$$

Covariance of Joint Distribution

All of that work leads us here:

$$E(XY) = 1$$

$$E(X) = 1.5$$

$$E(Y) = 0.5$$

$$cov(X, Y) = E(XY) - E(X)E(Y) = 1 - (1.5 \cdot 0.5) = 0.25$$

Covariance to Correlation

Again, we often use correlation instead of covariance because correlation does not depend on the units

To find correlation from covariance we use the following equation:

$$r = \frac{\text{cov}(X,Y)}{\sqrt{V(X) \cdot V(Y)}}$$

So we need to calculate the variance of X and Y, using information about the joint probabilities

Covariance to Correlation

Recall the joint probabilities we gathered from the table

X	P(X)	Y	P(Y)
0	1/8	0	4/8
1	3/8	1	4/8
2	3/8		
3	1/8		

$$E(X^2) = (0^2 \cdot 1/8) + (1^2 \cdot 3/8) + (2^2 \cdot 3/8) + (3^2 \cdot 1/8) = 3$$
$$E(Y^2) = (0^2 \cdot 4/8) + (1^2 \cdot 4/8) = 0.5$$

Covariance to Correlation

$E(X) = 1.5$ and $E(X^2) = 3 \rightarrow V(X) = 3 - 1.5^2 = 0.75$

$E(Y) = 0.5$ and $E(Y^2) = 0.5 \rightarrow V(Y) = 0.5 - 0.5^2 = 0.25$

$cov(X, Y) = 0.25$

$$r = \frac{\text{cov(X,Y)}}{\sqrt{\text{V(X)}} \cdot \sqrt{\text{V(Y)}}} = \frac{0.25}{\sqrt{0.75} \cdot \sqrt{0.25}} = 0.577$$

Clicker Question

What can be said of the correlation between the brand of an automobile and its quality?

(a) The correlation is negative, because smaller cars tend to have higher quality and larger cars tend to have lower quality.

(b) The correlation is positive, because better brands have higher quality.

(c) If the correlation is negative, an arithmetic mistake was made; correlation must be positive.

(d) Correlation makes no sense here, because brand is a categorical variable.

Clicker Question

Which of the following statements is false?

(a) Older men tend to have lower muscle density, so the correlation between age and muscle density in older men must be negative.

(b) Older children tend to be taller than younger children, so the correlation between age and height in children must be positive.

(c) A researcher finds that the correlation between two variables is close to 0, so the two variables must be unrelated.

(d) Taller people tend to be heavier than shorter people, so the correlation between height and weight must be positive.