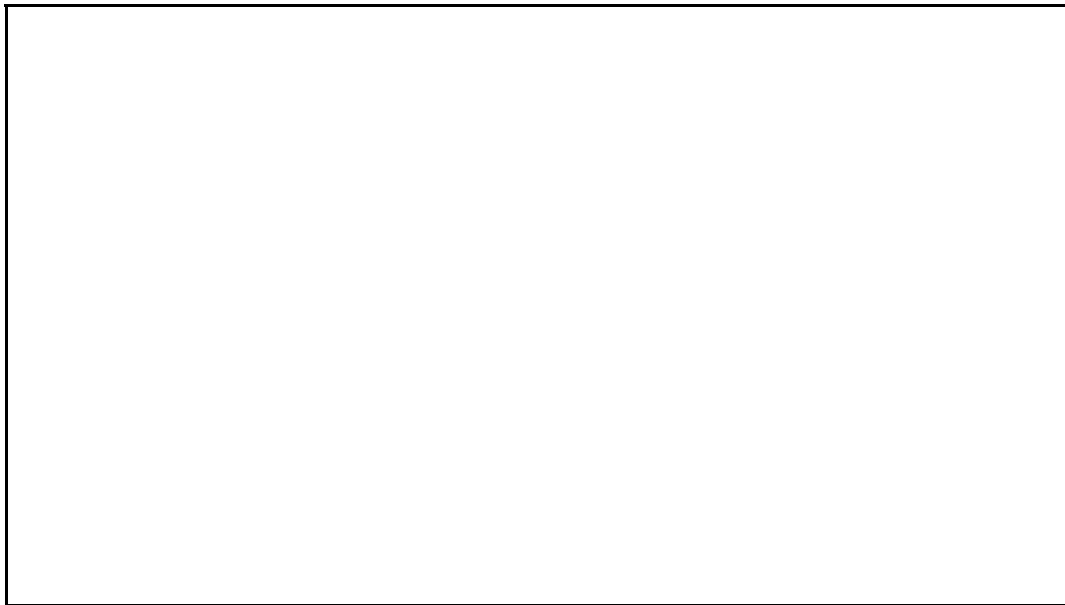


LINEARITY AND THE CORRELATION COEFFICIENT



SPEAKER: MICHAEL J. MAHOMETA, Ph.D.

One of the main things that we'll hear or read

is that the Pearson correlation coefficient - the r value - is only

appropriate to use when there is a LINEAR relationship to our data.

Now one of the first ways we examine a bivariate continuous data relationship

is by plotting it on a scatterplot.

Not only to see if there is some relationship visually

but to see if that relationship is in fact linear.

But what happens to r when that relationship is NOT linear?

Let's go back and take a look at our quadrant set up

in our scatterplot to find out.

Here's some data based on a 2009 survey study of university students in Sri

Lanka and the relationship between Post-traumatic Stress Disorder

and Post-traumatic Growth inventory score.

What do we see?

I think it's clear that there is a relationship between the two

variables, but unfortunately not a linear one.

Now you might be thinking: "What's going on here?"

This is actually an example of something called the Yerkes-Dodson curve.

It's an inverted-U shape to a relationship.

Basically, the authors of this study found

that low levels of post-traumatic distress

showed low levels of personal growth from the traumatic event.

As the level of post-traumatic distress increased, the growth also increased - up to a specific point.

Too much distress, and the growth from the event drops off;

the event was in fact too taxing on the participant for them

to have any psychological benefit from it.

Let's examine what happens if we find a correlation.

It's 0.048; that's really low.

Basically saying that there's no relationship, right?

Well here's why we actually need to visualize our data.

We can see that there's a relationship when we make the scatterplot,

but it's not linear.

So the correlation value is low because the Pearson correlation coefficient

is based in fact on a LINEAR relationship.

Remember those quadrants?

In the non-linear relationship, they actually hurt us.

In the case of our Post-traumatic Growth data,

the quadrant values for quadrants 1 through 4

are 10.47, -12.982, 26.42, and -19.455.

If we sum these values, we actually get a pretty small number - only 4.453, which we then have to divide by $n-1$ (or 92). That's why our correlation value is so small.

The correlation "sees" data in every quadrant.

And the product of the those quadrants are effectively canceling each other out.

Now, what if the original investigators didn't actually

do a good job like these investigators did in selecting study participants.

And only, say, selected people who did not have really high traumatic events.

We might see only this data.

This is a plot of the participants in the same study,

but only those who had a Post-traumatic Stress Disorder score of 40 or less.

We see only about half the data, and we see a mainly linear fit to the data.

Now, if we calculate our quadrants again, we'll get 13.491, -4.162, 17.133, and -0.403.

If we sum up those values, we get a score of 26.059 -

much larger than the quadrant sum of the

non-linear plot.

And now we're only dividing by 46.

Would you like to take a guess at the correlation value?

It's actually 0.567.

So what have we found out through this little experiment?

It turns out that the Pearson correlation value cannot be used when

we don't have a linear relationship.

Well, we COULD use it, but it would in fact give us a very incorrect finding.

And we've seen why this actually happens, because the Pearson

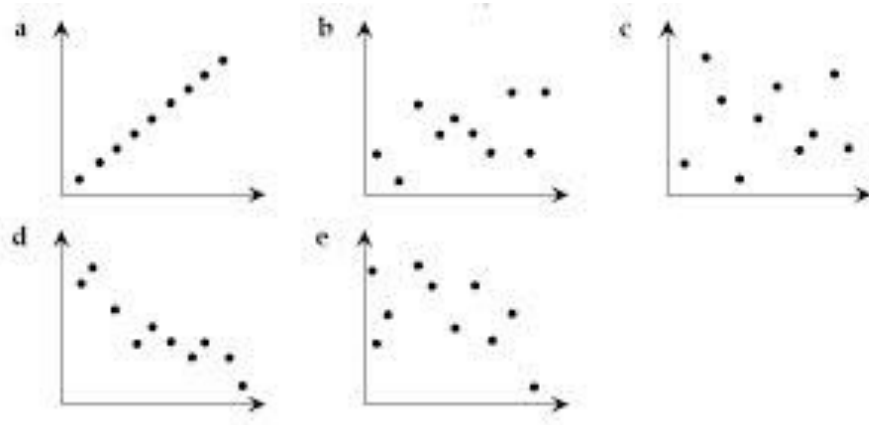
correlation formula is set up to handle just linear data.

The take home message here: draw your picture; find the visualization;

and determine if the Pearson correlation value is appropriate for your data.

Comprehension Check

Below are several scatterplots depicting relationships between bivariate data:



(5/5 points)

Please match each description with the most appropriate graph above. The five different descriptions are each most appropriate to a different graph; therefore, answers should not repeat.

Perfect, positive linear relationship

Graph A

Non-linear relationship that should not be measured with a correlation coefficient

Graph E

Strong negative relationship

Graph D

$r = 0.35$

Graph B

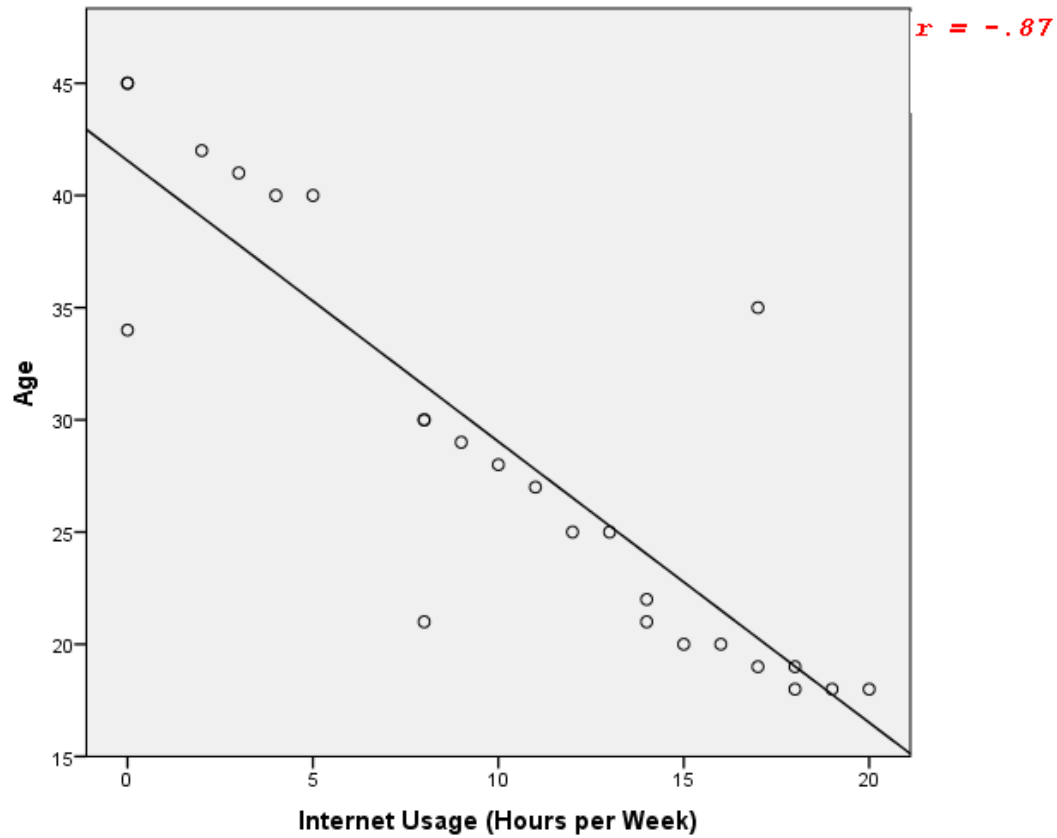
No relationship

Graph C

Check

Show Answer

Is there a relationship between age and internet usage each week? Here is bivariate data collected to examine this question:



(3/3 points)

What makes the data in this scatterplot "bivariate data"?

- ☐ One variable is quantitative, and one variable is qualitative.
- ☐ The data is collected on two different populations of people: those that use the internet, and those that don't.
- ☒ For each subject, we know both their age and how often they use the internet each week. ✓
- ☐ There are multiple data points being graphed using the same two axes.

There are two ~35-year-olds in this dataset. One uses the internet not at all, and the other uses it more than 15 hours per week. Which individual is the stronger outlier, and why?

- ☐ The 35-year-old that doesn't use the internet at all, because very few people did not use the internet.
- ☒ The 35-year-old that uses the internet 15+ hours per week. His data point is farther away from the linear trend. ✓
- ☐ They are both equally outliers because their data points are both at the same height on the graph.

Another researcher was only interested in individuals that use the internet 18+ hours per week. He calculated the correlation coefficient from the same dataset and got $r = 0.02$, showing no relationship. What happened?

- ☒ He created a restriction of range that made it look like there was no relationship. ✓
- ☐ He must have added or subtracted incorrectly; the same data should produce the same correlation coefficient.
- ☐ He should have selected the people that used the internet less than 18 hours per week.
- ☐ The strong non-linear relationship in that part of the graph cannot be described by r .

[Check](#)[Show Answer](#)[Help](#)

EdX offers interactive online classes and MOOCs from the world's best universities. Online courses from MITx, HarvardX, BerkeleyX, UTx and many other universities. Topics include biology, business, chemistry, computer science, economics, finance, electronics, engineering, food and nutrition, history, humanities, law, literature, math, medicine, music, philosophy, physics, science, statistics and more. EdX is a non-profit online initiative created by founding partners Harvard and MIT.

© 2014 edX, some rights reserved.


[Terms of Service and Honor Code](#)

[Privacy Policy \(Revised 4/16/2014\)](#)

About edX

[About](#)[News](#)[Contact](#)[FAQ](#)[edX Blog](#)[Donate to edX](#)[Jobs at edX](#)

Follow Us

 [Twitter](#) [Facebook](#) [Meetup](#) [LinkedIn](#) [Google+](#)