Learning Goal: Distinguish between association and causation. Identify lurking variables that may explain an observed relationship.

A *lurking variable* is a variable that is not measured in the study. It is a third variable that is neither the explanatory nor the response variable, but it affects your interpretation of the relationship between the explanatory and response variable.

1) To understand the above ideas, read this excerpt from *A Mathematician Reads the Newspaper* by John Allen Paulos.

"A more elementary widespread confusion is that between correlation and causation. Studies have shown repeatedly, for example, that children with longer arms reason better than those with shorter arms, but there is no causal connection here. Children with longer arms reason better because they're older! Consider a headline that invites us to infer a causal connection: BOTTLED WATER LINKED TO HEALTHIER BABIES. Without further evidence, this invitation should be refused, since affluent parents are more likely both to drink bottled water and to have healthy children; they have the stability and wherewithal to offer good food, clothing, shelter, and amenities. Families that own cappuccino makers are more likely to have healthy babies for the same reason. Making a practice of questioning correlations when reading about "links" between this practice and that condition is good statistical hygiene." (p. 137)

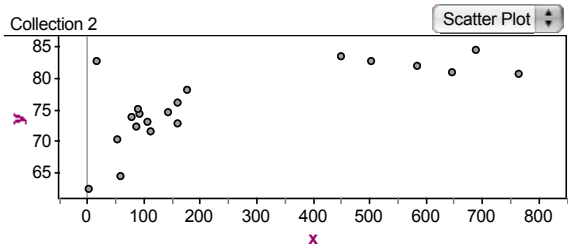
a) Pick one of the Paulos' examples and identify the explanatory and response variables.

b) Explain what it means to say "there is no causal connection" between these two variables.

c) These two variables have a strong association, but there is not a cause-and-effect relationship. Identify the lurking variable that is responsible for the relationship we see between the explanatory and response variables.

d) What is "good statistical hygiene" to Paulos?

2) For the 20 countries with the largest population for 2009 the scatterplot shows
x = internet users per 1000 people
y = life expectancy (years)
(World Almanac Book of Facts, 2009)



a) In 2009, which country had the largest number of internet users per 1,000 people? What is the life expectancy for people in this country?

country	x	y
Banglade...	2	62.5
Brazil	160	76.1
China	92	74.5
Egypt	79	73.9
France	449	83.5
Germany	583	82.0
India	53	70.4
Indonesia	87	72.4
Iran	113	71.7
Italy	503	82.9
Japan	688	84.7
Mexico	177	78.3
Pakistan	58	64.4
Philippines	105	73.2
Russia	160	72.9
Thailand	143	74.7
Turkey	88	75.2
United Ki...	646	81.1
United St...	765	80.8
Nigeria	17	82.9

b) In 2009, which country had the lowest life expectancy? What can we say about internet use in this country?

c) Based on the scatterplot, describe the form, direction, and strength of the relationship between life expectancy and the number of internet users per 1000 people.

d) The association between these two variables is positive. Explain what this means for this context. (Your answer should include a precise reference to the meanings of the variables.)

e) The correlation coefficient is 0.72, which is strong. Larger numbers of internet users per 1,000 correlate with longer life expectancy. Someone who confuses correlation with causation might suggest that an easy way to improve a country's life expectancy is to get more people onto the internet, which is a ridiculous cause-and-effect statement. Identify a lurking variable that might be explaining the strong association between life expectancy and the number of internet users per 1,000.