Welcome to Week #3!

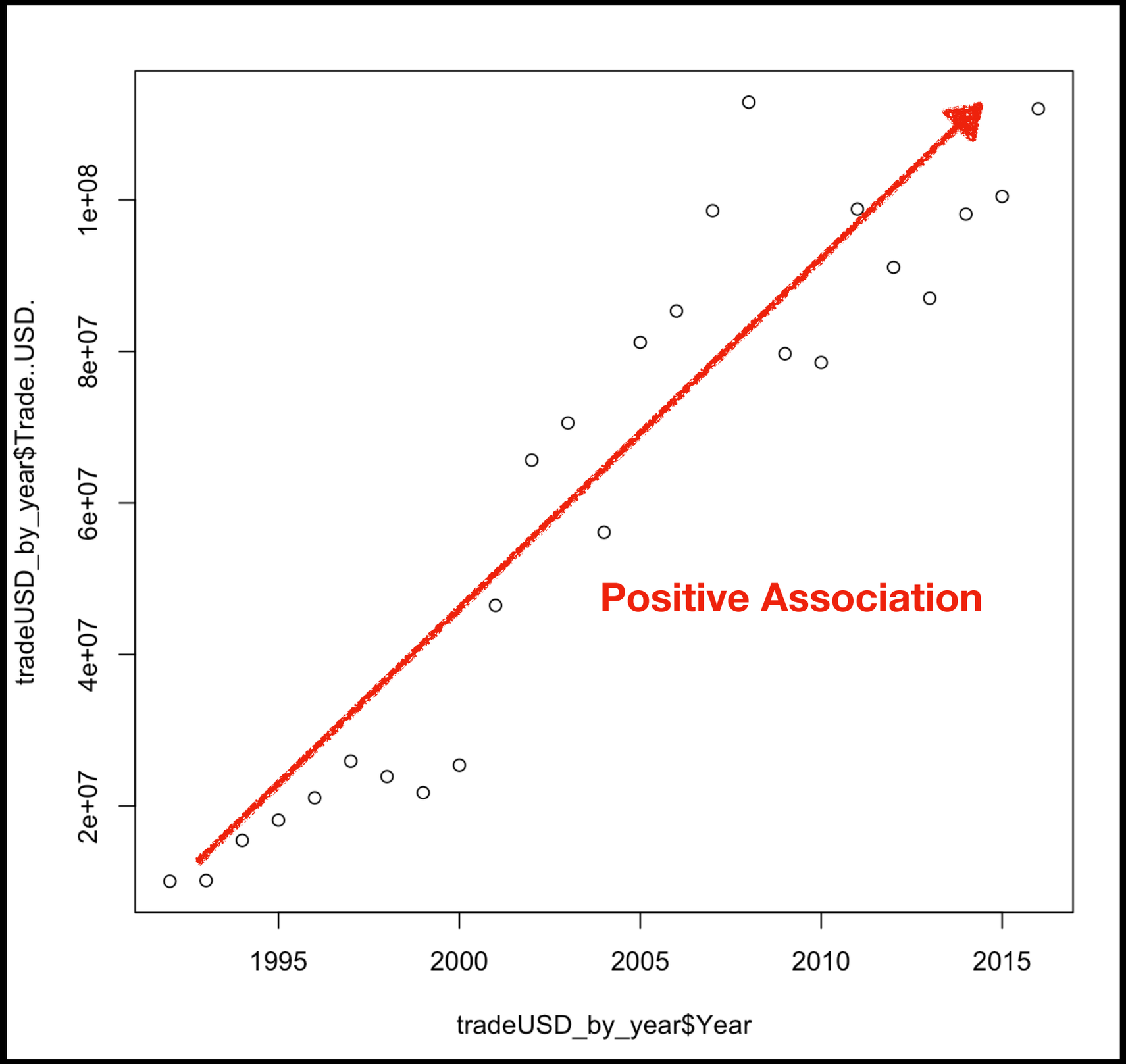
Finish up a few things with FISH DATA in R

Week	Topic	Reading
1	• Data, Models, and Information • Elementary statistics: Definitions • Overview of R	OIS 1 (ISL 1)
2	• Elementary statistics: Applications & Plots	OIS 1 (ISL 1)
3	• Introduction to data analysis with R • Review of tabular and graphical displays of data	ITR 1, 2, 5, 6, 7, 12
4	• Random variables: expectation and variance • Joint and conditional probability • Bayes rule	OIS 2
5	• Random variables: distributions (normal, binomial, poisson)	OIS 3

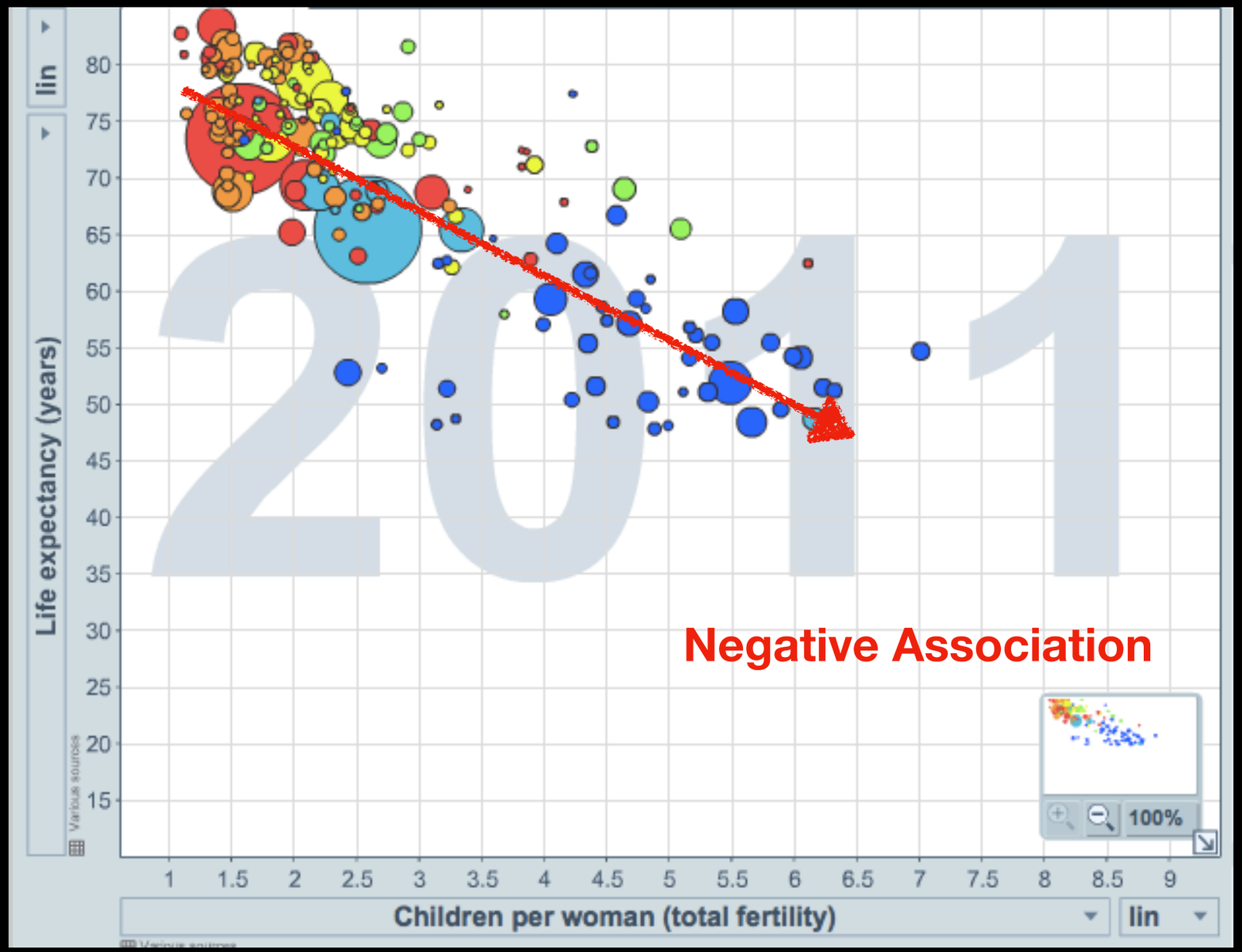
Definitions, basic concepts, R practice

Lots of definitions and pen-and-paper practice

Fish Dataset: Croatian Import Trade in USD vs. Time



Fish Dataset: Life expectancy vs. Children Birthed per Woman



http://www.gapminder.org/world

Fish Dataset: Life expectancy vs. Children Birthed per Woman

Associated or independent, not both

A pair of variables are either related in some way (associated) or not (independent).
No pair of variables is both associated and independent.

Designing Studies: Observational & Experimental

There are more airplane deaths now than there were 50 years ago. Does that mean airplane travel is becoming more dangerous?

No, there are a lot more people flying now.
Compare rates, not absolute numbers.

The death rate in the Navy during the Spanish-American war was 9/1000. For civilians in the same time period in New York City it was 16/1000. Does that mean it is safer to be in the Navy?

No, civilians include the old and sick while the Navy was comprised of healthy, young men.
Make sure to compare like to like.

Designing Studies: Observational & Experimental

1975 study of 1286 British women.

- ★ 23% of smokers died by 20-year follow-up
- ★ 29% of nonsmokers died by 20-year follow-up

Explanatory ➡ Response

- So do smokers tend to live longer?
- What is a potential confounding factor?

TIP: Explanatory and response variables

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other and plan an appropriate analysis.

explanatory variable → (might affect) → response variable

Designing Studies: Table Proportions

	Died in 20 yrs	Alive	Total Smoker or Non
Non-smokers	205	499	704
Smokers	138	444	582
Total Alive or Dead	343	943	1286

Designing Studies: Table Proportions

	Died in 20 yrs	Alive	Total Smoker or Non
Non-smokers	205	499	704
Smokers	138	444	582
Total Alive or Dead	343	943	1286

What percentage of the study participants are smokers? Non-smokers?

Designing Studies: Table Proportions

	Died in 20 yrs	Alive	Total Smoker or Non
Non-smokers	205	499	704
Smokers	138	444	582
Total Alive or Dead	343	943	1286

What percentage of the study participants are alive? Dead?

Designing Studies: Table Proportions

	Died in 20 yrs	Alive	Total Smoker or Non
Non-smokers	205	499	704
Smokers	138	444	582
Total Alive or Dead	343	943	1286

What percentage of non-smokers (and smokers) are alive?

Designing Studies: Table Proportions

	Died in 20 yrs	Alive	Total in Row
Non-smokers Age 18-64	65	474	539
Non-smokers Age 65+	140	25	165
Smokers Age 18-64	96	437	533
Smokers Age 65+	42	7	49
Total In Column	343	943	1286

What percentage of non-smokers (and smokers) are alive from the <65 age group?

Designing Studies: Table Proportions

	Died in 20 yrs	Alive	Total in Row
Non-smokers Age 18-64	65	474/539 = 88%	539
Non-smokers Age 65+	140	25	165
Smokers Age 18-64	96	437/533 = 82%	533
Smokers Age 65+	42	7	49
Total In Column	343	943	1286

What percentage of non-smokers (and smokers) are alive from the <65 age group?

Now looks like non-smokers fair better than smokers.

Designing Studies: Table Proportions

Age is a confounding factor

	Died in 20 yrs	Alive	Total in Row
Non-smokers Age 18-64	65	474/539 = 88%	539
Non-smokers Age 65+	140	25	165
Smokers Age 18-64	96	437/533 = 82%	533
Smokers Age 65+	42	7	49
Total In Column	343	943	1286

What percentage of non-smokers (and smokers) are alive from the <65 age group?

Now looks like non-smokers fair better than smokers.

Confounding factors

- A confounding factor is a variable associated with the both the explanatory and the response variable.

- Because of this, the response could be due to the supposed explanatory variable or to the confounding factor - the two are confounded.

Confounding factors

- In the previous example the supposed explanatory variable is:
  - ⋆ Smoking

- The response variable is:
  - ⋆ Dying within 20 years

- The confounding factor is:
  - ⋆ Age - there were more older people in the non-smoking group and older people are more likely to pass away within 20 years.

A case study

Eating more fruit, particularly blueberries, apples and grapes, is linked to a reduced risk of developing type-2 diabetes, suggests a study in the British Medical Journal.

http://www.bbc.com/news/health-23880701

"Blueberries cut the risk (of type-2 diabetes) by 26%..."

"**The research** looked at the diets of more than 187,000 people in the US..."

"The studies used food frequency questionnaires to follow up the participants every four years, asking how often, on average, they ate a standard portion of each fruit…"

A case study

- Can we conclude that eating blueberries will reduce risk of type-2 diabetes?

- Identify a possible confounding factor and explain how it could confound. Make sure to explicitly connect it with risk of type-2 diabetes.

Observation study vs. Experiment

- In an observational study researches watch/record information without imposing any treatment.

- In an experiment, researchers impose a treatment to try to draw a causal conclusion about the effect of the treatment.

- Why would we carry out one or the other?

A case study

- You have a hypothesis:
  Prolonged smoking leads to dental problems.
- How will you collect data to test this hypothesis?
- Should we do an experiment?
- You could survey people and ask about smoking habits and dental health, but will this prove a causal relationship?
- Lots of possible confounding factors. Give an example.
- The best we do is an observational study and try to compare like to like - older people to older people, women to women, etc.

Women, age, etc are blocking factors

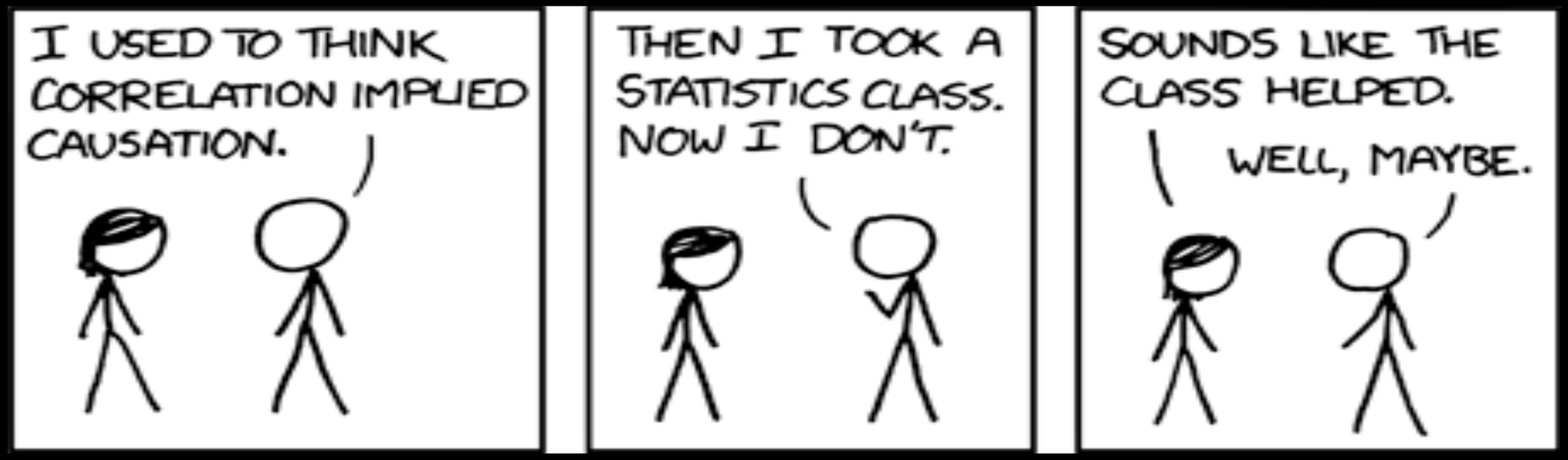
Blocking Variables

Blocking variables are characteristics that the experimental units come with, that we would like to control for.

Blocking is an effort to minimize confounding factors.

Correlation is not causation

- Correlation is not causation!



http:// xkcd.com/ 552/

- Observational studies alone cannot prove causation; only well designed experiments can prove causation.

More Experimental Design Terminology...

Placebo: fake treatment, often used as the control group for medical studies

Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment

Blinding: when experimental units do not know whether they are in the control or treatment group

Double-blind: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Practice #1

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?

(1) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

(2) There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)

(3) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

(4) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Practice #2

Protein May Hold the Key to Who Gets Alzheimer's
By PAM BELLUCK, MARCH 19, 2014, *New York Times*

It is one of the big scientific mysteries of Alzheimer's disease: Why do some people whose brains accumulate the plaques and tangles so strongly associated with Alzheimer's not develop the disease?

Now, a series of studies by Harvard scientists suggests a possible answer, one that could lead to new treatments if confirmed by other research….

The research, published on Wednesday in the journal Nature, focuses on a protein previously thought to act mostly in the brains of developing fetuses. The scientists found that the protein also appears to protect neurons in healthy older people from aging-related stresses. But in people with Alzheimer's and other dementias, the protein is sharply depleted in key brain regions.

Experts said if other scientists could replicate and expand upon the findings, the role of the protein, called REST, could spur development of new drugs for dementia, which has so far been virtually impossible to treat. But they cautioned that much more needed to be determined…

Practice #2

Protein May Hold the Key to Who Gets Alzheimer's
By PAM BELLUCK, MARCH 19, 2014, *New York Times*

- What type of study is this, observational study or an experiment?

Is there an association between lack of REST protein and having Alzheimer's?