Q1 – Thinking scientifically

Some adolescents using social media (e.g. Facebook) to follow their peers' lives can have feelings of low self-esteem, inadequacy and inferiority about their own life. As described in this article, "When kids scroll through their feeds and see how great everyone seems, it only adds to the pressure." What statistical artifact from Chapter 1 could contribute to an *entirely normal* adolescent feeling this way? Write a paragraph explaining your answer.

In my opinion, the statistical artifact that could contribute to the cognitive biases of the normal adolescent is selection bias. When kids are scrolling through their friends' feeds on social media, they can see others' lives are filled with excitement and adventures however they neglect the fact that people tend to share the best aspects of their life on social media. It is also possible the information others shared are exaggerated, fabricated, or untrue. What normal kids see in this case is under-representative of others' life, and as a result, the kids may think their friends all have amazing life which is very different from the ordinary life they live. Hence, the entirely normal adolescents feel low self-esteem and inferiority as they drew biased conclusions from social media.

Q2 – Thinking scientifically

Risk factors for rare diseases are often assessed using case-control studies (see slide 2.23 if these are unfamiliar) based in *referent databases*. Referent databases collect family histories of cases, and the entire family – i.e. all cases and controls – are included in the database if at least one family member was diagnosed with the disease of interest.
In this design, family size can often *appear* to be a risk factor for the disease – families containing cases tend to be larger than families (in the general population, not the referent database) who are all controls.
Which statistical artifact from Chapter 1 does this illustrate? Write a paragraph explaining your answer. You may include diagrams if you find them helpful – or not, if you don't.

The statistical artifact that resulted in this illustration is selection bias. When calculating key factors of the study, such as the marginal probability of additional cases in the family, the value of family size is crucial to the calculation. Since the family is selected from the referent databases, the larger the family size, the greater the possibility of finding such family in the referent databases. Hence, the mean family size in the referent databases is greater than the mean family size in the population. As a result, family size can appear to be a risk factor for the disease of interest.

Q3 – Quantifying distributions

Weather forecasters often make statements like "there is an 80% chance of rain tomorrow in Seattle".

a. Describe carefully what this *might* mean; what exactly could "rain tomorrow" mean? And 80% is a proportion of exactly what? (Note there is no single uniquely right answer to either question; state clearly what you think the potential answer(s) could be.)
b. State a numerical, rain-related outcome that can be measured every day and described and estimated less ambiguously than "rain tomorrow".

a. I think the statement "there is an 80% chance of rain tomorrow in Seattle" is trying to summarize the tendency of rain (which is 80% probability) occurs on the day after today in Seattle. "Rain tomorrow" simply means the possibility of a weather event – rain that could happen on the next day. 80% is the tendency of

the raining event will occur, so basically 8 out of 10 similar situations (taking account into the temperature, historical data, and location, etc.) will result in a raining outcome on the next day.

b. One thought to better explain "rain tomorrow" we could do a simulation on a computer to test whether it will rain tomorrow (let's say 10/10) in Seattle. We set up all the known factors or predictions for tomorrow, including temperature, humidity, cloud patterns, wind speed, and others, and use this as our baseline for testing if it would rain on 10/10 in Seattle. We then run the simulation ten times and observe eight outcomes of a rainy day in the setting. By applying this method, we can measure "rain tomorrow" and better described the weather outcome.

Q4 – Clustering variables

Colon cancer can in many cases be prevented through screening, and various screening tools are available. Some are more/less invasive than others, and these tend to be administered more/less frequently. To help schedule upcoming screenings for patients at risk of colon cancer, it helps to know what screens have recently been done. One detailed way to find out a patient's history is to have them complete a checklist of procedures they have undergone, e.g.

Which of these was your previous screening test?

- ☐ *Endoscopy*
- ☐ *Flexible Sigmoidoscopy*
- ☐ *Fecal-occult blood test (FOBT)*
- ☐ *Fecal immunochemical test (FIT)*

Unfortunately, questions like this produce unreliable responses, as patients confuse the various screening procedures, and forget details of them including their names. By thinking carefully about what those procedures involve, give a much simpler question with a *single* Yes/No answer, that would capture *most* of the information the detailed approach seeks to obtain. Hint: look up the screening procedures, if you don't know what they involve, and in particular how they are administered.

One possible question:

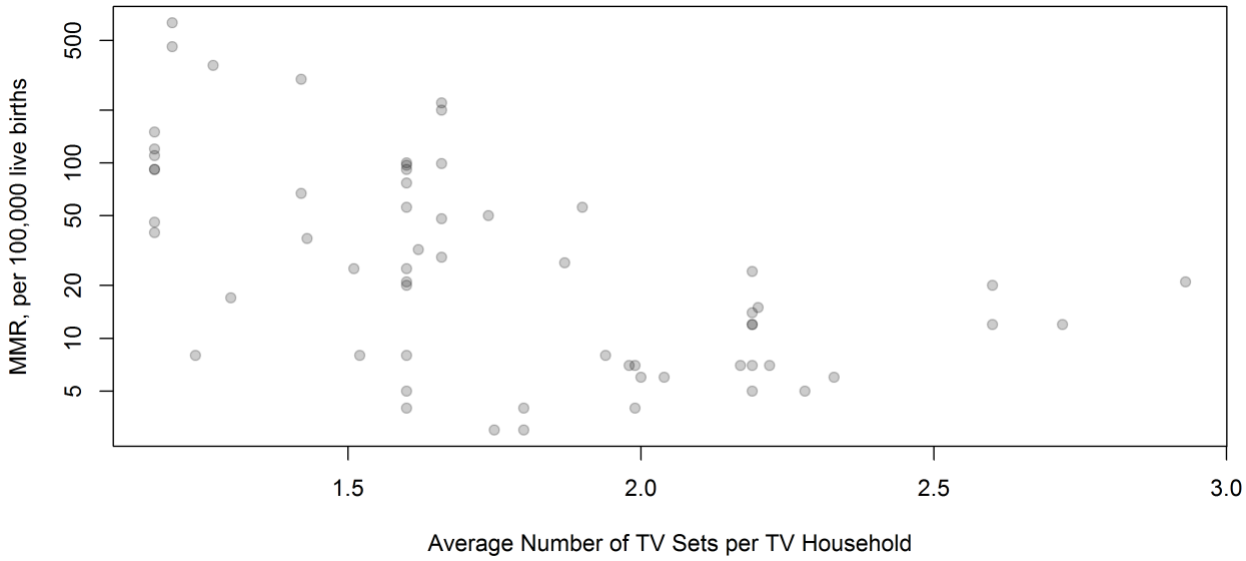
Have you had an examination with an endoscope (looks like a long, flexible tube) at two different spots (down your throat and up the butt), also had your stool sample collected with results back?

Q5 – Prediction versus causation

a. The number of ambulances attending a road traffic accident is a very good predictor of how many fatalities will occur due to that accident; more ambulances tend to be present at accidents with more fatalities. Briefly (i.e. in a sentence or two) explain why we don't therefore reduce fatalities by sending fewer ambulances.

There is no causal relationship between the number of ambulances and the number of road traffic accidents. Therefore, reducing the number of ambulances would not reduce fatalities and will only result in the inefficacy of saving lives.

b. The graph below shows data from 61 countries' *maternal mortality rate* (MMR, the annual number of female deaths per 100,000 live births from any cause related to or aggravated by pregnancy or its management excluding accidental or incidental causes) versus television ownership per household.



The data is from 2010 and can be accessed here and here. Looking at the original data may help you answer the questions below.

i. Give one reason, with explanation, why donating TV sets to the countries with worst MMR rates *would not* improve maternal mortality there.

Donating TV sets to countries with the worst MMR rates would not improve maternal mortality because there is no causation between the two factors.

ii. Give one reason, with explanation, why donating TV sets to the countries with worst MMR rates *might* improve maternal mortality there.

Donating TV sets to countries with the worst MMR rates would improve maternal mortality because women in those countries would have access to more health information through watching TV.

iii. Which of i), ii) do you think is the more plausible outcome of donating TV sets?

I think ii) is the more plausible outcome of donating TV sets as the graph above shows a negative relationship between the number of TV sets and MMR rates. The MMR rates are lower at higher numbers of TV sets per household.

Q6 – Looking ahead

In preparation for next week's lectures (Oct 11th-15th) read slides 2.31-2.98. Also read Vittinghoff Chapter 1 and Chapter 2, up to and including Section 2.3.

Done!