STA 371H Exercises 1 · Warm up

Due Monday, January 26, 2015

(1) Assessing causal claims

Read the following three short articles about research on healthy eating. All were published in the last week on the New York Times website, all refer to published scientific research, and all purport to give health advice.

1. "A 12-hour window for a healthy weight."[1] Key claim: "Contain your eating to 12 hours a day or less. And pay attention to when you begin." (Note: you'll see when you read the article that the advice is NOT to eat for 12 continuous hours, as the "key claim" might seem to suggest. I'm just saying.)
2. "How vitamin D may fight colon cancer."[2] Key claim: "The higher the blood levels of vitamin D, the less likely people were to develop colorectal tumors."
3. "Blueberries may lower blood pressure."[3] Key claim: "There is something very special about the composition of blueberries that is responsible for their effect on blood pressure."

On the basis of the evidence described in these articles, which of these key claims do you think is the strongest? Which is the weakest? Why? Write a few substantial paragraphs, no more than 1 single-spaced page in length, justifying your answers.[4]

(2) The sample mean

Suppose we have a data set with N observations $y_1, \ldots, y_N$, and we want to summarize these N numbers with a single number $\hat{\theta}$ that represents the center of the data set. A standard choice is the sample mean: $\hat{\theta} = \bar{y}$. In this question, you'll provide a justification of this choice.

Consider the following function of θ:

$$f(\theta) = \sum_{i=1}^{N} (y_i - \theta)^2 \, .$$

This function represents the total "error" for a particular choice for θ. The quantity $(y_i - \theta)$ is the error in using θ to approximate the ith data point. The function $f(\theta)$ just squares these quantities (so that positive and negative errors count equally) and sums them up to measure the

[1] `http://well.blogs.nytimes.com/2015/01/15/a-12-hour-window-for-a-healthy-weight/`

[2] `http://well.blogs.nytimes.com/2015/01/15/how-vitamin-d-may-fight-colon-cancer/`

[3] `http://well.blogs.nytimes.com/2015/01/14/blueberries-may-lower-blood-pressure/`

[4] This is a maximum length, not a minimum.

total approximation error. Show mathematically that $f(\theta)$ is minimized when you choose θ to be the sample mean:

$$\hat{\theta} = \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \, .$$

(3) Paint by numbers

The goal of this question is to walk you through a few basics in R. It's a "paint by numbers" question because all of the necessary commands can be found in the R script `heights.R`, available from the class website. You'll be able to use this script as a starting point, modifying it as appropriate to answer the questions.

A famous statistical analysis was carried out in 1885 by Sir Francis Galton, who wanted to see the effect of parents' height on the height of their children. We will reconsider this problem and take a look at a sample of heights for University of Texas students and their parents, collected from the students in the very first class I ever taught. The data is in the file "heights.csv" available on the course website. The variables are:

SHGT: Student's height in inches
MHGT: Mother's height in inches
FHGT: Father's height in inches.

Part A: Create two scatter plots of child height versus mother's height and versus father's height. Based on the visual evidence, does the mother's height or father's height appear to be a stronger predictor of the child's height? Turn in a copy of whichever plot seems to show the stronger correlation. (You can copy and paste an R plot directly into a Word/Mac Pages/etc document.)

Part B: Compute the mean and sample standard deviation of the fathers' heights.