DATA 100: Vitamin 1 Solutions

February 14, 2019

1 Study Design

Which of the following describe a proxy variable in study design? A proxy variable is ____ (check all that apply)

□ A variable of interest

☑ Not a variable of interest

☑ Easy to measure

□ Difficult to measure

☑ Related to a variable of interest

□ Related to a population of interest

Explanation: When the variable(s) of interest may not be measurable or easily measurable, a proxy variable can be used instead. This is done because the proxy variable is easy to measure and relates to the variable of interest.

2 Census

Why are p values meaningless when taking a census?

□ The chance that a census was drawn at random is large.

□ A p value expresses the chance that some measured difference (or a larger one) would arise due to random sampling.

□ There is a better alternative to a p value when taking a census.

☑ There is no need for statistical inference when directly measuring an entire population.

Explanation: A census consists of collecting data on an entire population. There is no need to perform any statistical inference to learn about the population's underlying probability distribution. Thus, performing any kind of hypothesis test is not required.

3 Big Data

What are the four V's of big data, according to IBM?

- □ Visualization
- ☑ Velocity
- □ Validity
- ☑ Veracity
- ☑ Volume
- □ Variability
- ☑ Variety
- □ Variance

Explanation: See infographic presented in slides on study design.

4 Sampling

4.1 Probability Sampling

If everyone in a probability sample has an equal chance of being included in the sample, this guarantees that the sample is a simple random sample (SRS).

- □ True.
- ☑ False.

Explanation: Consider performing a cluster sampling on the clusters [A, B] and [C, D], where the sample consists of one randomly selected cluster. Then the probability that the sample contains any of the letters A, B, C or D is equal, i.e. $P(\textit{sample contains } A) = P(\textit{sample contains } B) = P(\textit{sample contains } C) = P(\textit{sample contains } D) = \frac{1}{2}$.

4.2 Simple Random Sample

In a simple random sample (SRS), everyone has an equal chance of being included in the sample.

- ☑ True
- □ False

Explanation: A simple random sample consists of a subset (a sample) of a set (a population). Each observation in the sample is selected randomly by chance such that each member of the population has an equal probability of being selected at each stage of the sampling process. Additionally, each subset of observations of size n has an equal probability of being selected from the population as any other subset of size n.