Stats with R

Assignment 0

Name: Chirag Bhuvaneshwara; 2571703

Done with: Ayan Majumdar; 2571656

Egla Hajdini; 2571690

1. In image processing,
   - The peak signal to mean noise power ratio (PSNR) between the original image and a noisy version of it is a continuous measurement variable. Gives meaningful numeric values and contains natural zero. Therefore, it is a variable of ratio scale.
   - the pixel locations in a digital image is a discrete measurement variable. It does not contain a natural zero and operations of divisions and multiplications are not possible. Therefore, it is a variable of interval scale.

2. In machine learning we develop an objective function to maximize or minimize which can be done by stochastic gradient descent (SGD). In this case, the training set of all examples forms the population and the SGD algorithm samples this training set to consider a sample of small batch of examples to perform Gradient Descent on. The samples are chosen randomly from the training set of all examples which is the reason for it to be called a "Stochastic algorithm".

3. In SGD algorithm, the sample chosen is a random sample of the population because the sample is a batch of randomly chosen examples from the training set of all samples.

4. Please find a recent research paper form an area you're interested in, which includes a study that reports statistical significance. Write down:
   a) the research :
   This paper explores the potential uses of Facebook ad audience estimates for eHealth by studying the following: (1) for what type of health conditions prevalence estimates can be obtained via social media and (2) what type of marker interests (constraints) are useful in obtaining such estimates.
   b) the population:
   Users on Facebook.
   c) the sample and whether you think it was random
   Audiences on Facebook that meet some constraints (People who like the magazine *Diabetes Daily* in some states in the US). It is not a random sample as all users are not equally likely to be chosen.
   d) the dependent variable
   a set of public health indices, such as the fraction of the adult population that has diabetes (collected from the US Public Health Data)
   e) the independent variable
   It is the set of indices derived from the Facebook's ad audience estimates that consist of some markers or interests in certain topics that may relate to particular health conditions such as diabetes (type II), obesity or alcoholism. It is included as the fraction of the active users who have an interest in a certain topic.
   f) for all variables, whether they're continuous or discrete :
   Both the dependent and the independent variables are continuous variables.

g) the measurement scale of each variable:
Both the dependent and the independent variables are ratio scale.

h) What statistical test did the authors use?
Pearson correlation test. Each pair of indices *f* (Facebook interest index) and *h* (health index) constitutes one hypothesis that is being tested.

i) Can you find out why they used that specific statistical test?
The goal of the experiment is to find if there is significant correlation between the online activity and interest of individuals and the health condition. Hence, the authors chose the Pearson correlation test for a simple measure to check if the 2 variables are correlated.

Link to paper:

Online Health Monitoring using Facebook Advertisement Audience Estimates in the United States: Evaluation Study