Midterm PRACTICE Exam
Introduction to Data Science

Bachelor in Data Science

Name: ____
Student ID: ____
Duration: 90 minutes
Instructions: Answer all questions. Show all your work for full credit. Calculators are allowed, but notes and books are not.

Question 1: Concepts

(a) Explain the difference between a population and a sample. (5 points)

(b) Explain the difference between simple random sampling and stratified random sampling. Which method is better and why? (5 points)

Question 2: Data

A biological study is performed involving plants. Six healthy plants — 3 of type A and 3 of type B — are subjected to water deprivation for 45 days. The data recorded are the type of plant and the survival time, starting from the beginning of the study. Towards the last days of the study the temperature is gradually increased to make sure that at day 45 are dead. For two of the plants of type A this is the case.

(a) Make a hypothetical data record table describing the outcome of this data collection.

(b) What are the statistical units and the variables in this study?

(c) For each of the variables describe the type of data (nominal/ordinal categorical or discrete/continuous numerical) that is recorded. Where relevant include whether or not the data are censored or truncated.

Question 3: Random Sampling and Estimation

The world population consists of almost 8 billion individuals. A small fraction of them have a “perfect pitch,” meaning that they can correctly identify any musical note upon hearing it. Two studies are performed to identify the fraction of individuals with perfect pitch.

- Pollster 1 takes a simple random sample of 10,000 individuals and records

  $$X = \text{number of people with perfect pitch in random sample.}$$

- Pollster 2 takes a smaller, stratified random sample, consisting of 5 people who sing in a choir and 995 people that do not sing in a choir, reflecting the global ratio of choir (i.e. 0.5%) and non-choir (i.e. 99.5%) people. She records,

  $$\begin{aligned} X_c &= \text{number with perfect pitch in choir sample} \\ X_n &= \text{number with perfect pitch in non-choir sample} \end{aligned}$$

(a) What are the distributions of X, X_c and X_n?

(b) What are the simple random sample estimate $\hat{p}_1$ and the (proportional) stratified random sample estimate $\hat{p}_2$ of the fraction of people in the world with perfect pitch, as a function of X, X_c and X_n.

(c) What is the mean of $\hat{p}_1$ and $\hat{p}_2$?

(d) What is the standard deviation of $\hat{p}_1$ and $\hat{p}_2$?

(e) What is the approximate distribution of $\hat{p}_1$ and $\hat{p}_2$?

(f) The data are collected. Pollster 1 obtains $X = 19$, whereas pollster 2 observes $X_c = 4$ and $X_n = 2$.

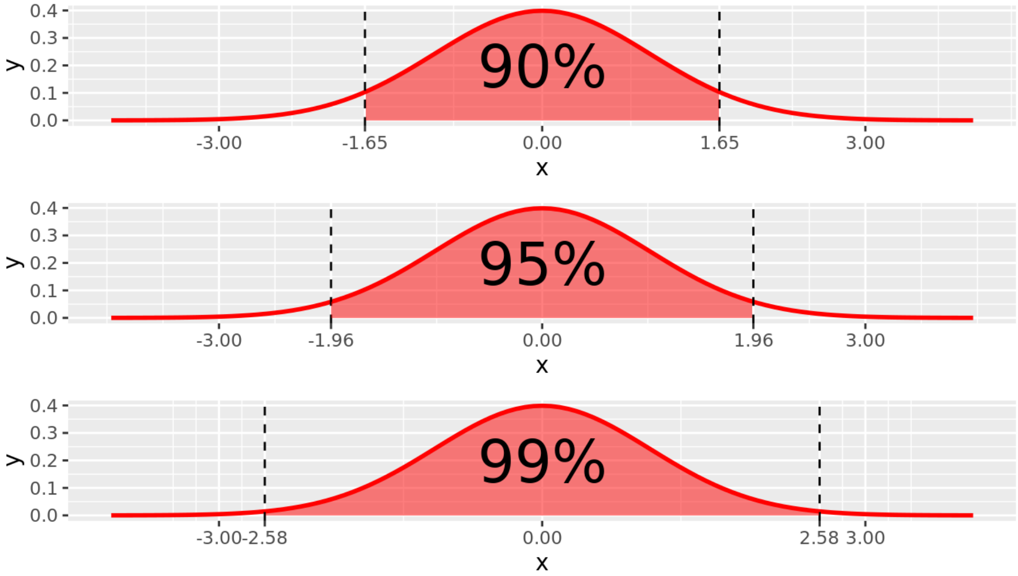
(a) What are the estimates of pollster 1 and pollster 2 of the faction of people in the world with perfect pitch?

(b) Use the values in (a) to determine the standard deviation for both estimators.

(c) Calculate a 90% confidence interval using the data of pollster 1?

(d) Calculate a 90% confidence interval using the data of pollster 2?

(e) Are the estimates of the two pollster consistent with each other?


0.4
0.3
0.2
0.1
0.0
y
90%
-3.00
-1.65
0.00
1.65
3.00
x
0.4
0.3
0.2
0.1
0.0
y
95%
-3.00
-1.96
0.00
1.96
3.00
x
0.4
0.3
0.2
0.1
0.0
y
99%
-3.00
-2.58
0.00
2.58
3.00
x