ENGR 123 (2020) Assignment 4 Solutions (Thursday, 17 September)

Topics: PROBABILITY, SAMPLING, DATA, SUMMARY STATISTICS AND GRAPHICS

1. **Probability-1** (3 points) In which of the following are events A and B mutually exclusive? In the cases where the events are not mutually exclusive, list the outcomes in $A \cap B$.

   (a) Toss a coin twice. A is the event of a head on the first toss, and B is the event of a head on the second toss.

   (b) Roll two dice. A is the event of a sum of 7, B is the event of a double (same value on both dice).

   (c) Roll two dice. A is the event of a 2 on at least one of the dice. B is the event of a 3 on one of the dice. epr002

   Solution:

   (a) $\boxed{\text{no}}$ — not mutually exclusive
   $A \cap B = \{(H, H)\}$.

   (b) $\boxed{\text{yes}}$ — mutually exclusive
   $A \cap B = \emptyset$

   (c) $\boxed{\text{no}}$ — not mutually exclusive
   $A \cap B = \{(2, 3), (3, 2)\}$

2. **Probability-2** (2 points) Given that C and D are independent and

   $$P(C \mid D) = \tfrac{2}{3} \qquad P(C \cap D) = \tfrac{1}{3}$$

   find

   (a) $P(C)$ (b) $P(D)$ (c) $P(C \cup D)$ (d) $P(D \mid C)$ epr019

   Solution:

   (a) Since C and D are independent

   $$P(C) = P(C \mid D) = \boxed{\tfrac{2}{3}}$$

   (b) $P(D \mid C) = \frac{P(D \cap C)}{P(C)} = \frac{(\frac{1}{3})}{(\frac{2}{3})} = \frac{1}{2}$
   Since C and D are independent

   $$P(D) = P(D \mid C) = \boxed{\tfrac{1}{2}}$$

   (c) $P(C \cup D) = P(C) + P(D) - P(C \cap D)$
   $= \frac{2}{3} + \frac{1}{2} - \frac{1}{3} = 5/6$

   (d) $P(D|C) = P(D) = 1/2$

3. **SAMPLING** (6 points)
   The idea here is that you do not necessarily have all the facts - you may not know the full situation. This is a classic scenario for engineers to operate in. Hence, your thinking has to cover several possibilities with explanation of your reasoning!

(a) A software company is concerned about the market penetration of Python and other free software products. They plan a customer satisfaction survey of their current customers to see how they might maintain their own market share. Their customers include industry, government, academia as well as private individuals. The bulk of their sale are to industry. Discuss a sampling methodology if you are asked to generate 500 responses.

(b) A network provider is interested in the traffic profile at its network switches. In order to evaluate the traffic statistics, detailed analysis of the data at each switch is required. As a result, only 20 switches will be chosen out of the total network which includes 457 switches. Discuss the issues which could affect your sampling of 20 switches?

Solution:

(a) An obvious (but not the only approach) would be to stratify the market into IND / GOVT / ACAD / PRIVATE. Take SRSs of size n_1, n_2, n_3, n_4 from each ($n_1+n_2+n_3+n_4 = 500$). Choose n_1, n_2, n_3, n_4 according to size of strata perhaps, i.e., number of clients, or financial importance of strata, etc.

Many practical questions remain. Once you select an organisation - there may be many users. Do you ask all of them (a bit like cluster sampling) or just a central person? What do you do with non-response, etc.

(b) If you have 457 homogeneous switches operating in similar conditions, a SRS is fine. If switches are different or environments are different, you might stratify to make sure your sample covers all types.

4. **DATA TYPES** (4 points)

For the following situations, is the data described categorical (nominal), categorical (ordinal), measurement (discrete) or measurement(continuous)? ("measurement" is another terminology for "numerical")

(a) Packet delay time;

(b) Number of error bits in a packet of length N bits;

(c) Signal strength as recorded on a 5 bar display;

(d) Error in a GPS measurement;

Solution:

(a) Measurement (continuous);

(b) Measurement (discrete);

(c) Categorical (ordinal) as $1, 2, 3, 4, 5$ really covers range of signal strength, not particular values;

(d) Measurement (continuous).

5. **SUMMARY STATISTICS, GRAPHICS** (12 points)

Given the box plot below (Fig. 1), answer the following questions:

(a) What is the median? LQ, UQ, IQR? (You don't need to give an exact value, just a rough estimate should be fine)

(b) Does the data have any outliers?

(c) What is missing in the box plot?

Given the histogram below (Fig. 2), answer the following questions:

(d) Is this a histogram with frequencies or relative frequencies?

(e) What is missing in the histogram?

(f) Is data symmetric or skew? If it is skew, does it skew to the left or the right?

(g) Can you guess which interval the median should lie in?

Now looking at both plots (Figures 1 and 2), answer the following question:

(h) Compare the two plots, do you think they come from the same dataset? Please provide your reasons!!!

Solution:

(a) Median $= 17$, $LQ = 9$, $UQ = 24$, $IQR = UQ - LQ = 24 - 9 = 15$ (these are just some rough estimates by looking at the box plot, not the exact values, so other similar values are also fine);

(b) Yes, one outlier which is not too far from the rest of the data;

(c) The title and axis labels for the plot;

(d) It is a frequency histogram as shown by the numbers on the $y-$axis;

(e) The title and axis labels for the histogram;

(f) Data is not symmetric, it is slightly skew to the right;

(g) $(14.1, 18.8]$ as roughly 50% of the data is on the left side of this interval.

It looks like the two plots come from the same dataset. Ranges are the same, from 0 to approximately 47. Median in the box plot is approximate 17 which lies in $(14.1, 18.8]$. Both box plot and histogram show some slight skewness to the right. If we think that the values in the bin $(42.3, 47]$ of the histogram are outliers then these outliers are not too far from other observations. You can also check on the LQ and UQ from the histogram.

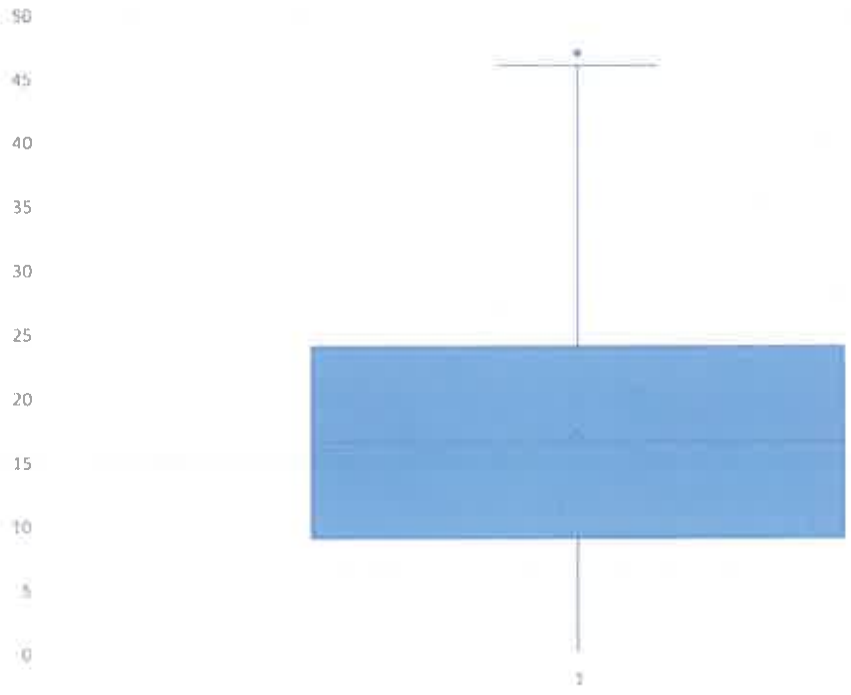

Fig. 1: Assignment Q5.- Box plot

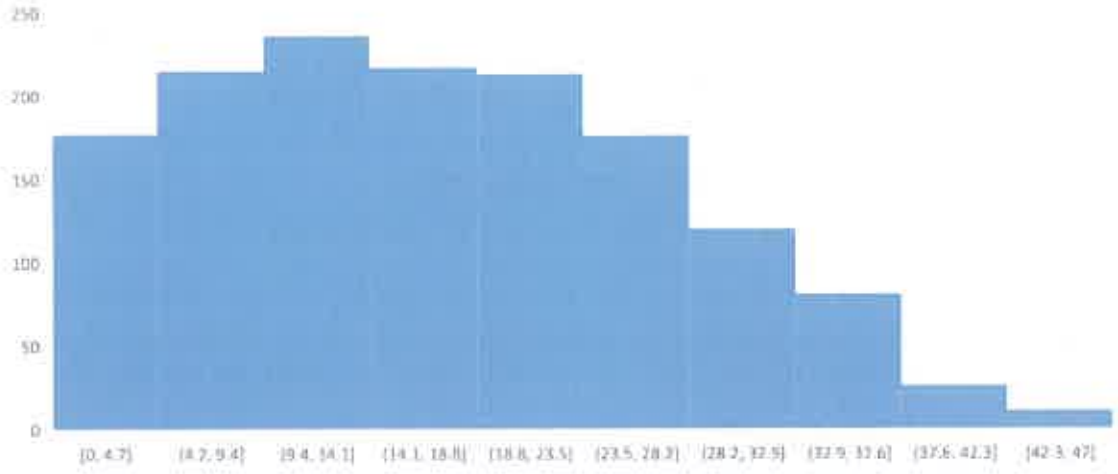

Fig. 2: Assignment Q5.- Histogram

6. **Excel/Python/Matlab/etc.** (3 points)

The aim of this question is to give you a flavour of how to create graphics for a given dataset.

Use your favourite software to produce properly labelled plots for the data set available in IAT.txt (on Blackboard) which contains actual measurements of inter-arrival times between voice-call on a mobile network (normalized to unit mean for confidentially reasons). Due to the normalization, the data is dimensionless and not measured in seconds. You can choose to plot **one of the following** (or even better - have a go at all of them):

(a) Histogram: explain your choice of the number of bins and the type of histogram used?

(b) Box plot: annotating by hand - show the values of the key components? (median, LQ, etc.)

(c) ECDF: annotating by hand - show the min value, max value and the 70%ile?

Hint: If you are familiar with Excel then you can easily create a Histogram with that. Or you may Google search how to use Python/Matlab to create those plots with only a few lines.

Solution:

(a) A relative frequency or probability histogram are more informative. In terms of the number of bins, there is no right or wrong number. If you use too few, you obscure the shape of the data. If you use too many you start to see a very spiky histogram, which again loses the visual impression of the main shape. Ideally, you experiment and pick the number of bins as some compromise. Here, around 10 bins works fine.

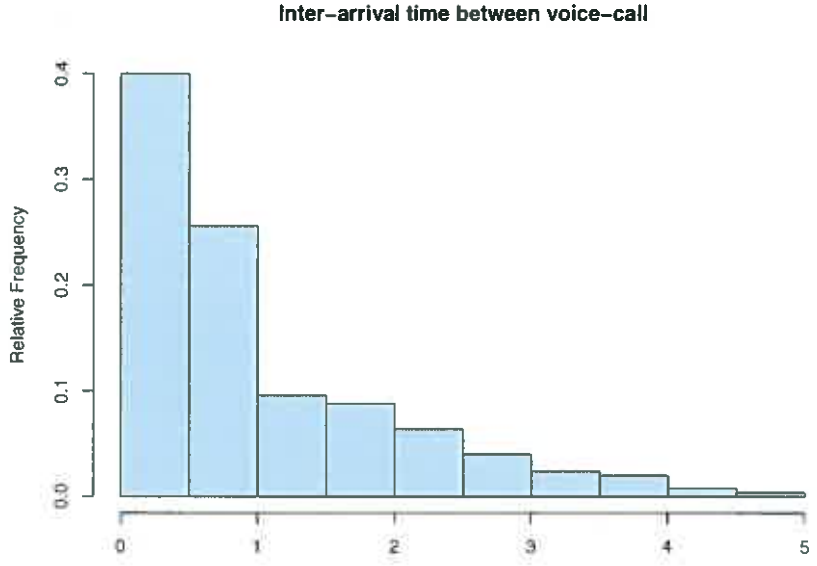

Fig. 3: Assignment Q6.- IAT-hist

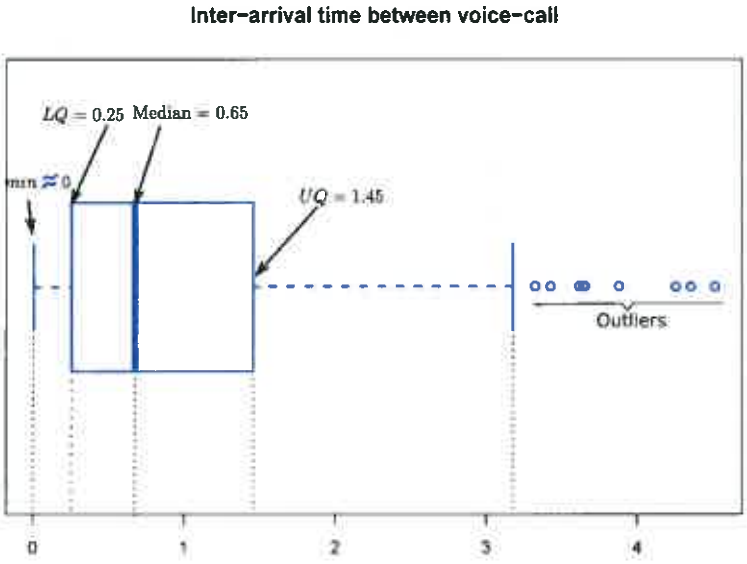

Fig. 4: Assignment Q6.- IAT-box plot

(b) Key values are shown in the plot (Fig. 4),

(c) The min value, max value and the 70%ile are shown in the plot (Fig. 5).

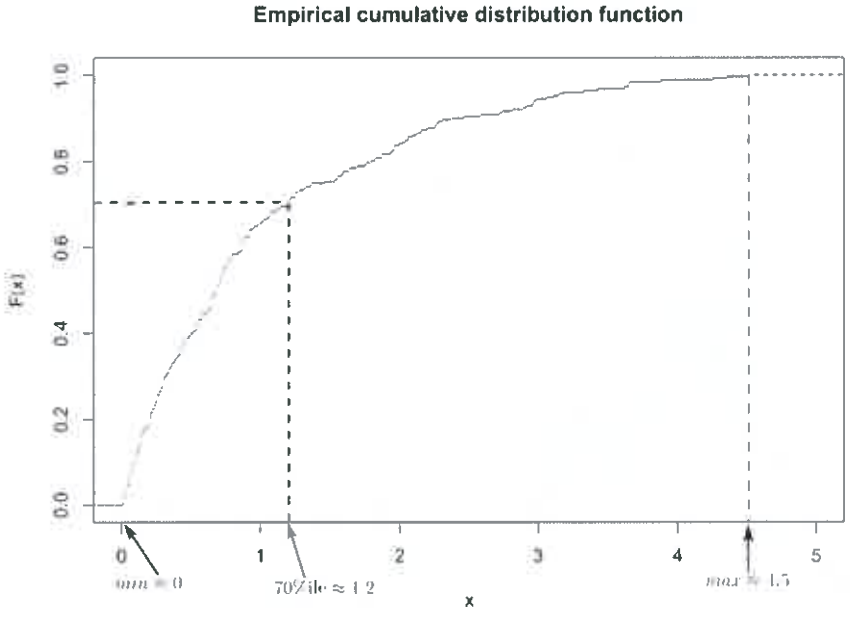

Fig. 5: Assignment Q6.- IAT-ecdf