ENGR 123 (2020) Assignment 4 Due 12 midnight 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

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

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?

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;

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!!!

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.

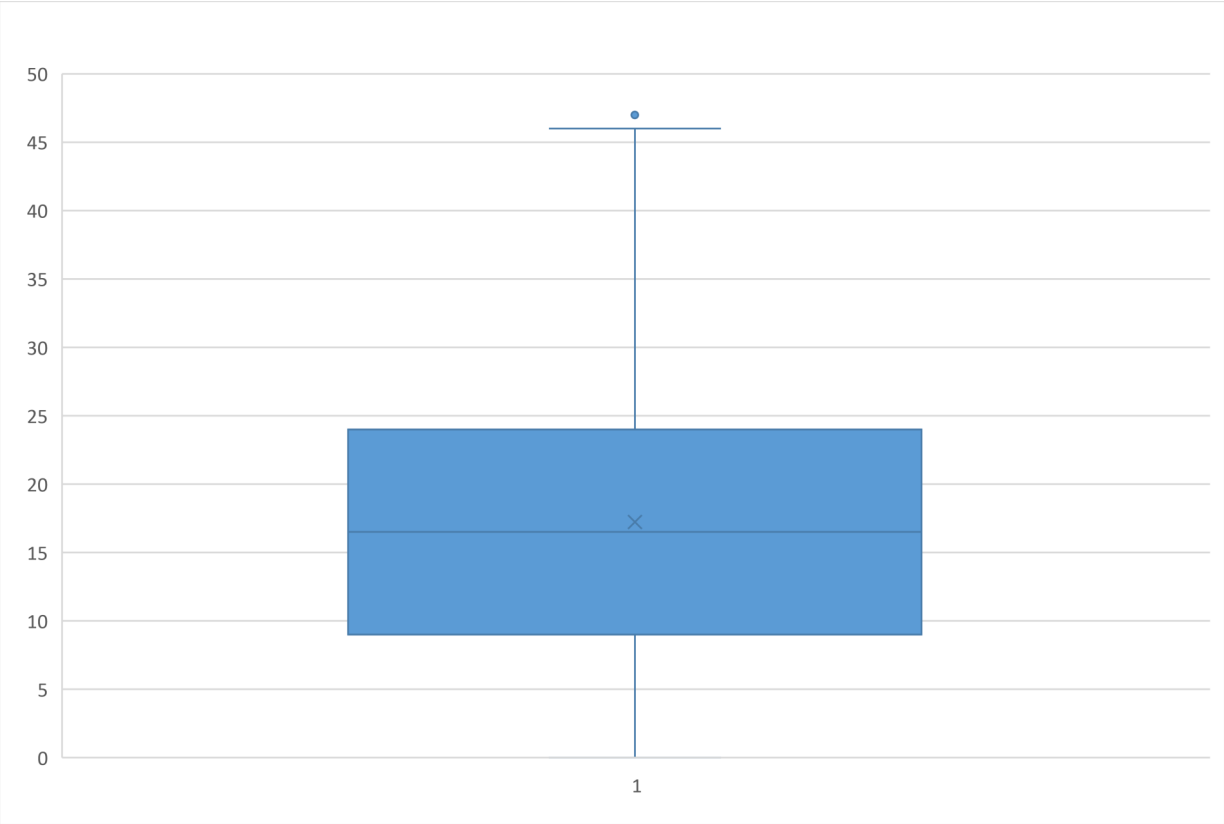

Fig. 1: Assignment Q5.- Box plot

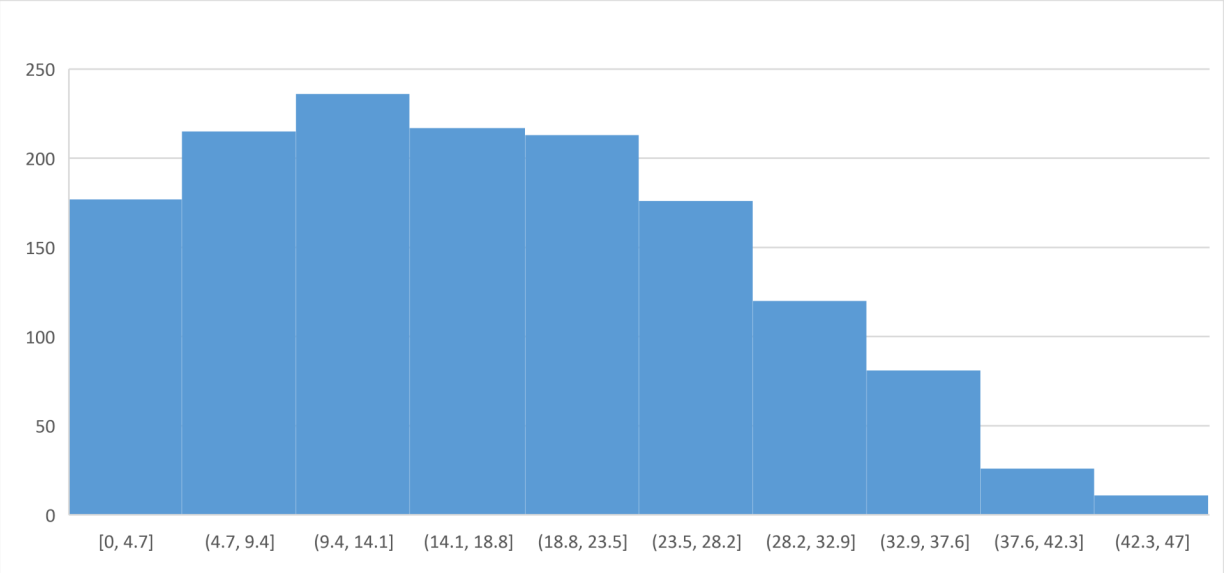

Fig. 2: Assignment Q5.- Histogram

ENGR 123 - Tutorial 4 questions for 14–18 September 2020

1. **PROBABILITY-1** Let S be the sample space of all Wellington males aged 45 years and over. Let A be the subset of overweight males, B be the subset of males with heart disease and C be the subset of males who will die before age 50.

   Express in symbols the following events, and illustrate each by a Venn diagram.

   (a) A randomly selected male is overweight but doesn't have heart disease.

   (b) A randomly selected male is either overweight, has heart disease, or both and will die before age 50.

   (c) A randomly selected male is both overweight and has heart disease and will not die before age 50. epr003

2. **PROBABILITY-2** Given

$$P(A) = 0.4 \qquad P(B) = 0.7 \qquad P(A \cap B) = 0.2$$

   find

   (a) $P(A \mid B)$ (b) $P(A \cup B)$ (c) $P(B \mid A)$ epr015

3. **PROBABILITY-3** Given

$$P(A) = 0.8 \qquad P(B) = 0.7 \qquad P(C) = 0.6$$
$$P(A \mid B) = 0.8 \qquad P(C \mid B) = 0.7 \qquad P(A \cap C) = 0.48$$

   (a) Are A and B are independent?

   (b) Are A and C are independent?

   (c) Are B and C are independent? epr018

4. **SAMPLING**

   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 telecommunications provider is planning a drive test of a newly deployed scheduling algorithm which has been deployed at all of the nationwide set of cellular base stations. Three vehicles are available with measuring equipment to measure the data rates provided by the new system. If you have the budget to perform 24 hours of drive-testing in each of the 3 cars, how would you sample your drive test routes from the network?

   (b) The IT department have been warned that the Okapi Virus may be present on some VUW machines. Due to the ferocity of this particular virus, you have been tasked with trying to find it. Unfortunately, it takes up to 15 hours to fully test that a machine is free of Okapis. You are told to test 50 machines. How do you decide which machines to test?

5. **DATA TYPES**

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

(a) Status of a router (0 = idle, 1 = busy);

(b) Number of frequency channels occupied in one cell in a mobile cellular network;

(c) Number of software faults encountered in a year;

(d) A node in a power network has 3 branches and a fault detector which can only indicate the branch on which a fault occurred;

6. **SUMMARY STATISTICS, GRAPHICS**

Given the box plot below (Fig. 3), answer the following questions:

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

(b) Does the data have any outliers?

(c) Is the data symmetric or skew?

Given the histogram below (Fig. 4), answer the following questions:

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

(e) Does the data have any outliers?

(f) Is the data symmetric or skew? If it is skew, is 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 3 and 4), answer the following question:

(h) Compare the two plots, do you think they come from a same set of data? Briefly explain your judgement?

Given the plot of an ECDF of some dataset below (Fig. 5), answer the following questions:

(h) What are the median, LQ, UQ?

(i) What percentage of the data lies in the interval $[6, 8]$?

7. **Excel/Python/Matlab/etc.**

The aim of this question is to give you a flavour of how to create graphics for a given dataset. There will be no solution given to you from tutors for this question as it is part of the assignment, but you may ask for a little support if their help is available!!!

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-calls 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):

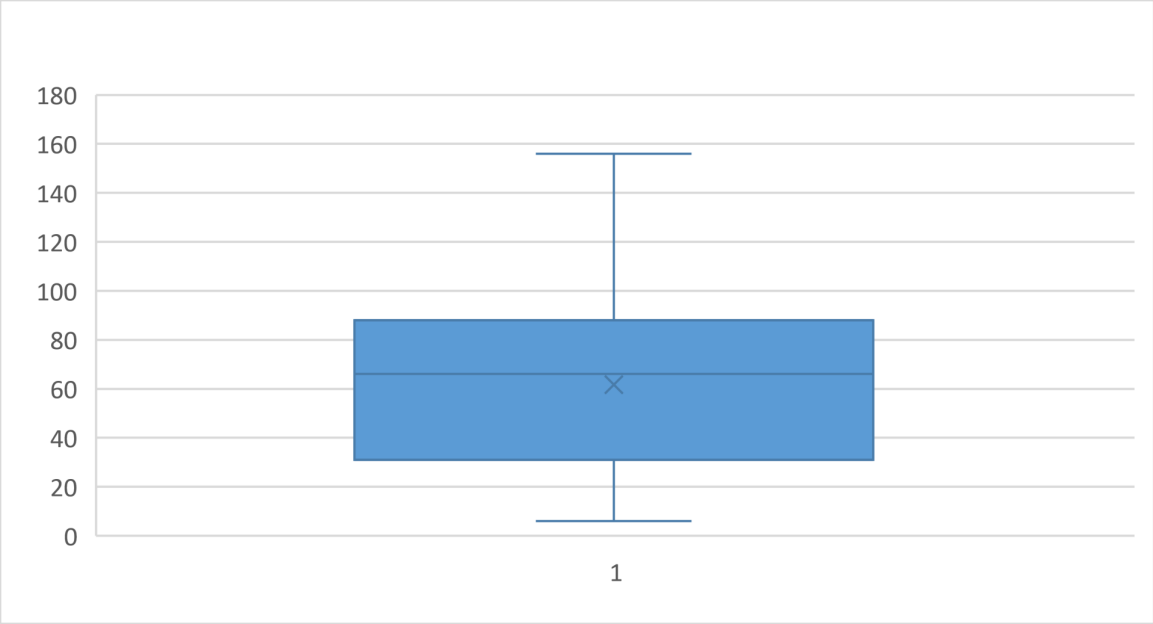

Fig. 3: Tutorial Q6.- Box plot

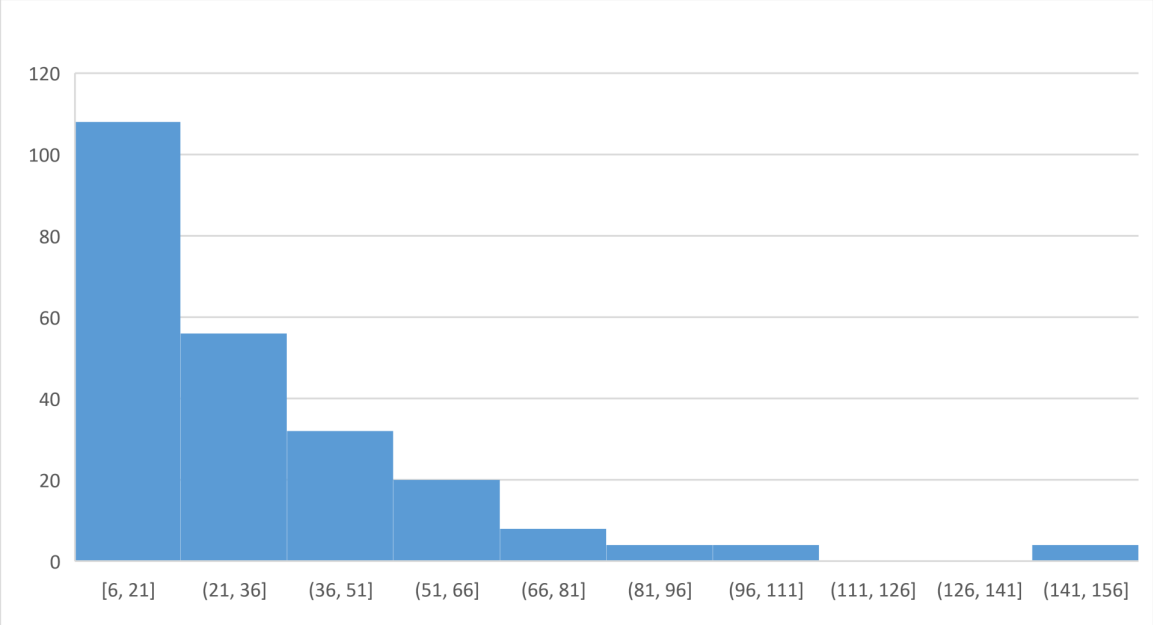

Fig. 4: Tutorial Q6.- Histogram

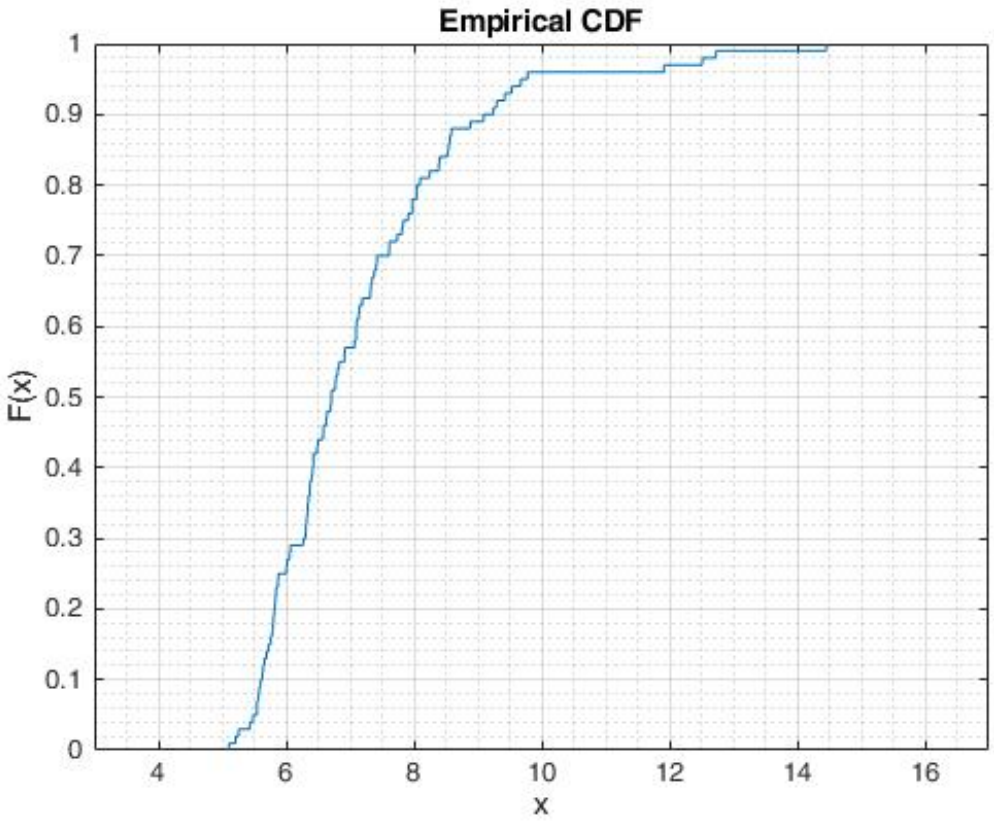

Fig. 5: Tutorial Q6.- ecdf

(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.