Midterm (21.11.2018) (10:00-12:00)

BLM3590 – Statistical Data Analysis

Q1(10) □	Q2(10) □	Q3(15) □	Q4(15) □	Q5(15) □	Q6(15) □	Q7(20) □	Q8(00) □	Q9(00) □	Total

Q1. Determine whether the following statements are correct or not by placing F(alse) or T(rue) in parenthesis. (10)

a. Using sample statistics to infer some “phenomena” of population parameters is known as descriptive statistics.(F)
b. Data can be defined as unprocessed facts and figures without any added interpretation or analysis. (T)
c. In retrospective studies, researchers look into the histories of the participants. (T)
d. Categorical data are numerical measurements where the numbers are associated with a scale measure (F)
e. Histograms are defined as a frequency distribution commonly used to visualize numerical variables. (T)

==

Q02. *X* is a binomial distribution with the probability of the outcome of interest 0.45 and 100 Bernoulli trials.

What are the **mean** and **variance** for the distribution of *X*?

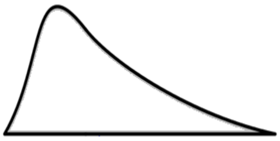
$\theta = 0.45$ $n = 100$

The theoretical (population) mean of a random variable X with Binomial(n,θ) distribution is
$\mu = n\theta = 0.45 \times 100 = 45$.

The theoretical (population) variance of X is
$\sigma^2 = n\theta(1 - \theta) = 100 \times 0.45 \times (1 - 0.45) = 45 \times 0.55 = 24.75$.

==

Q3. A *central tendency* is a central or typical value for a probability distribution. It is also called a center or location of the distribution. Measures of central tendency are often called *averages*. There are several measures that reflect the *central tendency*: sample *mean*, sample *median*, and sample *mode*. If a data set has the following probability distribution,



a. which measure you use to determine the *central tendency* of the data? (07)

I would use sample median, which is also known as the mid-point. It is used to represent the average when the data are not symmetrical (skewed distribution).

b. Explain your choice with an example if possible. (08)

Because the given distribution is skewed distribution, the sample median represents the average value of the data better than the sample mean.

Assume that we have the following two datasets: $x_1 = \{74, 80, 79, 85, 81\}$ *and* $x_2 = \{47, 80, 79, 85, 81\}$

Sample Mean(x_1) = *79.8*, *Sample Mean*(x_2) = *74.4;* *Sample Median*(x_1) = *80*, *Sample Median*(x_2) = *80*

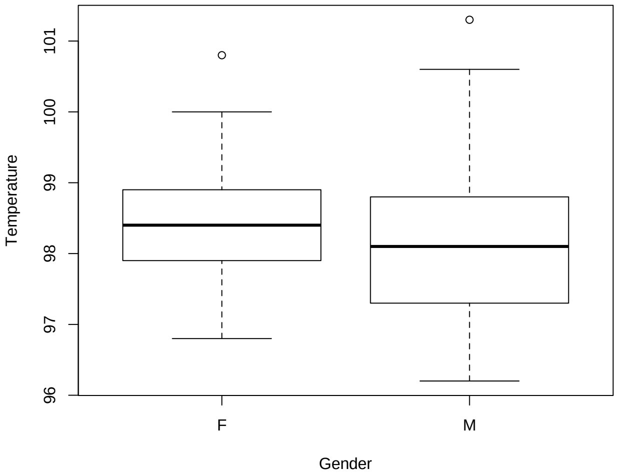
So, the median is more robust against outliers.

==

Q04. For the given boxplots of **temperature** variable (in Fahrenheit) for **M**(ale) and **F**(emale) groups in **BodyTemperature** dataset:

a. Determine the **five-number summary, range** and **IQR** for each group. (05)

	M	F
Min	96.2	96.8
Q1	97.3	97.9
Q2	98.1	98.4
Q3	98.8	98.9
Max	101.3	100.8
Range	4.9	4.0
IQR	1.5	1.0



b. Comment on the **mean** and **dispersion** in both groups (by comparing the values obtained in (a) and (b)?(05)

The mean of the F variable is larger than the mean of the M variable. However, as seen from the table (Range, IQR), the variable M has a greater range than the variable F. That is, the standard deviation of the variable M is greater.

c. According to the boxplot, are there any outliers? Explain (05)

The value ~100.8 in the female variable and the value ~101.3 in the Male data are exceptions. Because these values are larger than Q3 + 1.5 IQR values. Values greater than the upper horizontal line or smaller than the lower horizontal line are considered exceptions and are shown as dots in the boxplot chart.

==

Q5. We want to emphasize sudden changes in the image (sharpen the image).

Considering the same values given in Q4 (x = { 2, 4, 5, 7, 5, 6, 4, 2, 7, 6, 5, 4, 8, 3, 7, 5 }),

a. Suggest a filtering method to apply to this image and justify your choice. (07)

I would suggest moving difference filter, which is a numerical implementation of derivation function.

A simple moving difference filter is a high pass filter, which could be used as sharpening filter.

b. What will be the resulting values of the row after applying the filter? Assume that the first value (x_{-1}) is zero. (08)

$Dx_n=x_n-x_{n-1}$. *So, new values will be:*

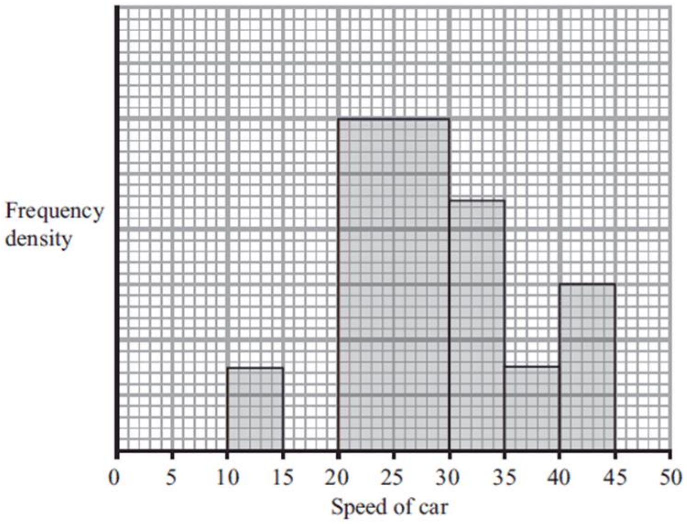
$y = \{x_0-x_{-1}, x_1-x_0, x_2-x_1, x_3-x_2, x_4-x_3, x_5-x_4, x_6-x_5, x_7-x_6, x_8-x_7, x_9-x_8, x_{10}-x_9, x_{11}-x_{10}, x_{12}-x_{11}, x_{13}-x_{12}, x_{14}-x_{13}, x_{15}-x_{14}, x_{16}-x_{15}\}$

y = { 4-0, 6-4, 5-6, 7-5, 5-7, 8-5, 4-8, 4-4, 7-4, 6-7, 6-6, 4-6, 8-4, 4-8, 7-4, 5-7,0-5}

y = { 4, 2, -1, 2, -2, 3, -4, 0, 3, -1, 0, -2, 4, -4, 3, -2, -5}

==

Q06. A policeman records the speed of the traffic on a busy road with a 30 kmh speed limit. He records the speed of a sample of 450 cars. The histogram in the following figure represents the results.



a. Calculate the number of cars that were exceeding the speed limit by at least 5 kmh in the sample. (08)

Hint: Try to estimate number of cars in each bar.

Considering the area under the histogram in terms of the number of cars, number of cars falling in each box can be worked out as;

Total # of cars / # of boxes = 450 / 22.5 = 20 cars

The number of cars that were exceeding the speed limit by at least 5 kmh in the sample can be found by finding number of cars exceeding 35 kmh; 20×4.5 = 90 cars

b. Estimate the value of the mean speed of the cars in the sample. (07)

Hint: Number of cars in each bar should be considered as weight for the mean of each bars.

Mean speeds of each bars (intervals) are mid-points. A weighted mean approach should be used. Weights are the number of cars in each bar; 30 cars for interval 10-15, 240 cars for interval 20-30, 90 cars for interval 30-35, 30 cars for interval 30-35, and 60 cars for interval 40-55.

$$\text{The mean speed of the cars} = \frac{30 \times 12.5 + 240 \times 25 + 90 \times 32.5 + 30 \times 37.5 + 60 \times 42.5}{450} = 28.83$$

==

Q07. A very large bin contains 3 different types of disposable flashlights. The probabilities that **type 1, type 2**, and **type 3** flashlights will give over 100 hours of use are 0.7, 0.4, and 0.3 respectively. Suppose that 20% of the flashlights in the bin are **type 1**, 30% are **type 2** and 50% are **type 3**.

a. What is the probability that a randomly chosen flashlight will give more than 100 hours of use? (08)

Let A be the event {the flashlight will give more than 100 hours of use } and *Bi* be the event {we choose flashlight of type *i*}. Then we have:

$$P(A) = \sum_{i=1}^{3} P(A/B_i)P(B_i) = 0.7 \cdot 0.2 + 0.4 \cdot 0.3 + 0.3 \cdot 0.5 = 0.41$$

b. Given the flashlight lasted over 100 hours, what is the conditional probability that it was a type 1 flashlight?(07)

$$P(B_1) = \frac{P(B_1/A \cdot P(A)}{P(A)} = \frac{0.7 \cdot 0.2}{0.41} = 0.34$$

==

Q08. For the given following multiple alignment of four nucleotide (A, C, G, T) sequences;

Position	p1	p2	p3	p4
Sequence a	A	T	G	T
Sequence b	A	A	G	A
Sequence c	T	A	C	T
Sequence d	C	G	C	A

a. What are the ***type*** of the ***Position*** and ***Nucleotide*** (*residue*) variables? (05)

Both are categorical data type.

b. Determine the frequencies of each residues for each position (contingency table). (05)

c. Create a position specific relative frequency table for each nucleotides. (05)

----

b. Raw frequency table

Pos. Nuc.	p1	p2	p3	p4	Total.
A	2	2	0	2	6
C	1	0	2	0	3
G	0	1	2	0	3
T	1	1	0	2	4

c. Relative frequency table

Pos. Nuc.	p1	p2	p3	p4	Overall freq.
A	0.5	0.5	0	0.5	0.375
C	0.25	0	0.5	0	0.1875
G	0	0.25	0.5	0	0.1875
T	0.25	0.25	0	0.5	0.25