Data analysis for engineering science students II

December 2nd 2024

Exercises

Prof. Dr. NDOH MBUE I.

……………………………………………..

1. The following table shows data collected from forestry survey:

	Tree ID	Species	Height (m)	Elevation (m)	# of scars	Damage rating
	B-11	Pine	23.5	347	1	Low
	B-15	Pine	15.6	960	5	Severe
	C-42	Eucalyptus	22.3	826	3	Medium
	F-01	Mahogany	45.2	450	0	Low

Copy and complete the table below:

Variable	Quantitative or Qualitative	Discrete or Continuous	Level of Measurement
Tree ID			
Species			
Height			
Elevation			
# scars			
Damage rating			

2. Classify each of the following variables as either quantitative or qualitative. If a variable is qualitative, state the possible categories.

(a) Geographical region
(b) Price of a house
(c) Temperature
(d Fuel consumption
(e) Employment rate
(f) Number of children in a family
(g) Race
(h) Political party preference

3 In each of the following sets of variables, identify which of the variables can be regarded as a response variable and which can be used as predictors? (Explain)

(a) Number of cylinders and gasoline consumption of cars.

(b) SAT scores, grade point average, and college admission.

(c) Supply and demand of certain goods.

(d) Company’s assets, return on a stock, and net sales.

(e) The distance of a race, the time to run the race, and the weather conditions

(f) The weight of a person, whether or not the person is a smoker, and whether

4

a) Make a sketch of a normal distribution that has been positively skewed.
b) What is the area under a positively skewed normal distribution?
c) If the mean, median, and mode for a data set are not the same, what can you conclude about the data's distribution?

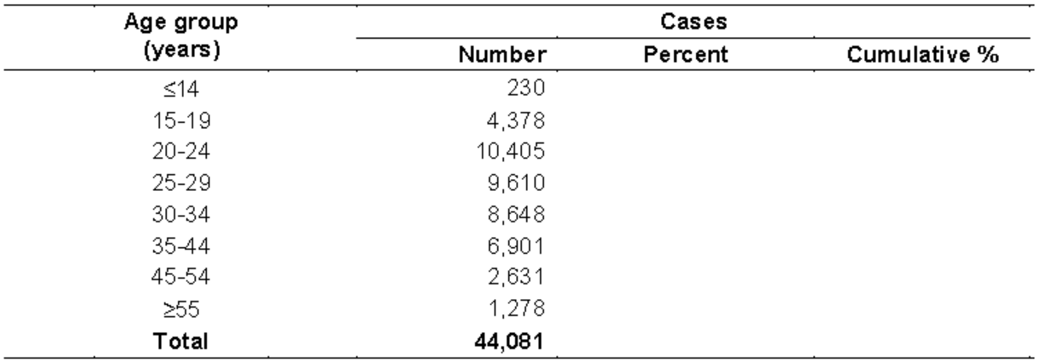
1. Arithmetic mean is 12 and number of observations are 20 then sum of all values is

A. 8 B. 32 C. 240 D. 1.667

5. The presence of extreme observations does not affect

A. AM B. Median C. Mode D. Any of these

6. Which of the following statements is true?
   A. Usually mean is the best measure of central tendency
   B. Usually median is the best measure ‘of central tendency
   C. Usually mode is the best measure of central tendency
   D. Normally, GM is the best measure of central tendency
7. For a moderately skewed distribution, which of he following relationship holds?
   A. Mean - Mode = 3 (Mean - Median)
   B. Median - Mode = 3 (Mean - Median)
   C. Mean - Median = 3 (Mean - Mode)
   D. Mean - Median = 3 (Median - Mode)
8. Primary and secondary syphilis morbidity by age, Mankon Village, 1989 is shown in the table that follows



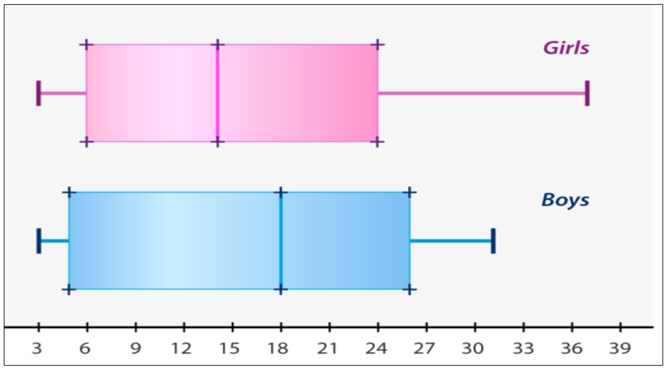
Age group (years)	Cases		
	Number	Percent	Cumulative %
≤14	230		
15-19	4,378		
20-24	10,405		
25-29	9,610		
30-34	8,648		
35-44	6,901		
45-54	2,631		
≥55	1,278		
Total	**44,081**		

Copy and complete the table.

9. A university offers only two degree programs: English and Computer Science. Admission is competitive and there is a suspicion of discrimination against women in the admission

process. Of the 80 males, 35 were admitted, of the 60 females, 40 were denied admission. Present this data in a tabular form, showing the row and column percentages.

10. The double box-and-whisker plot compares the number of CDs owned by boys and girls in the seventh grade. Analyze the data to answer the questions below.



a) Compare the median number of CDs owned by boys and girls.

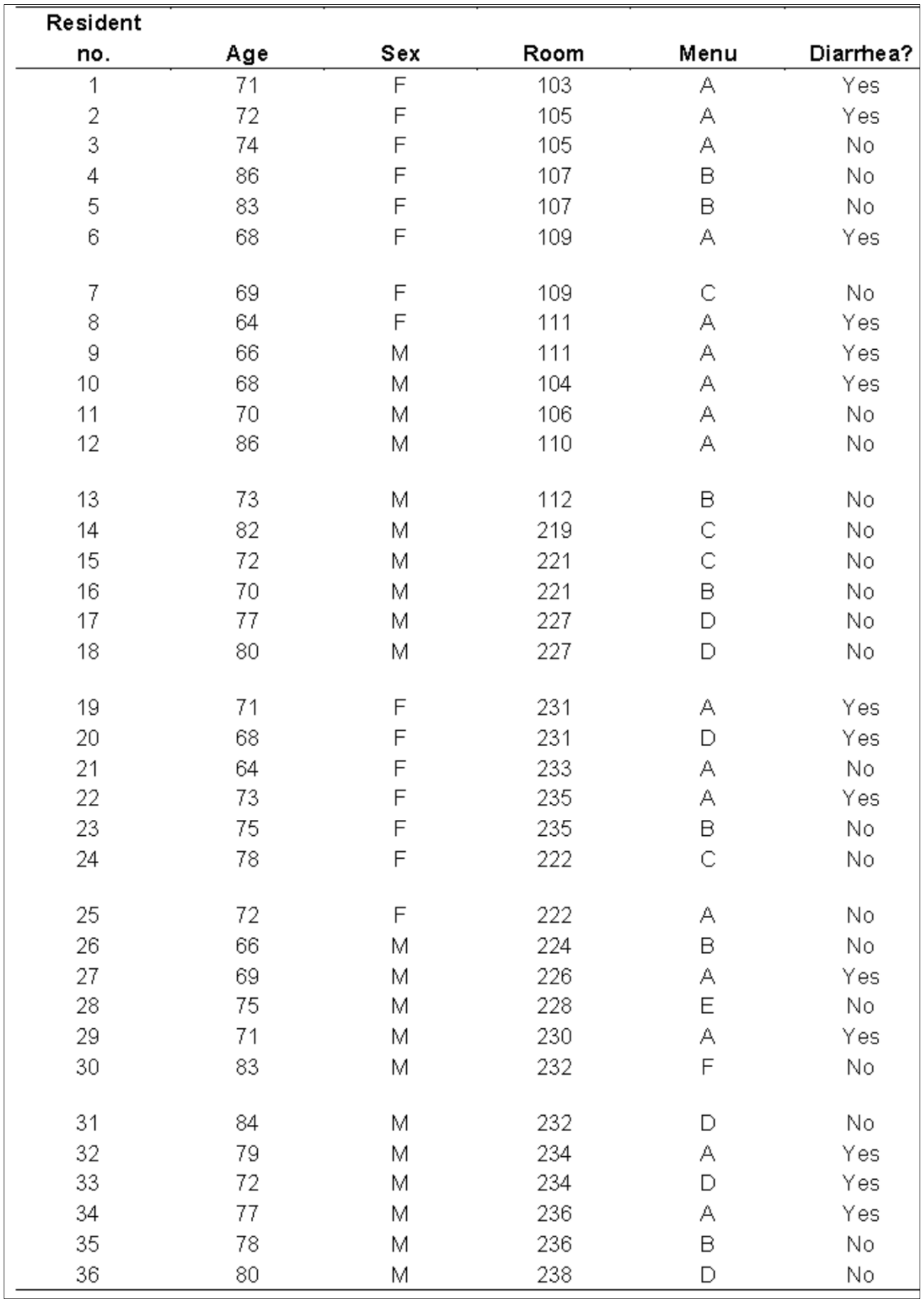
b) What is the least number of CDs owned by boys and girls? The smallest amount of CDs owned by both boys and girls is 3.

c) Identify the greatest number of CDs owned by both girls and boys. The greatest amount of CDs owned by girls is 37. The greatest amount of CDs owned by boys is 31.

d) What conclusions can be drawn about the number of CDs owned by girls? You can see that a larger range of the data values falls in the upper quartile. Therefore, the half of the girls that own more than the median number of CDs has a wider spread in the number of CDs they own.

e) What conclusions can be drawn about the number of CDs owned by boys?

11. The data in the table that follows describes characteristics of the 36 residents of a nursing home during an outbreak of diarrheal disease.



Resident no.	Age	Sex	Room	Menu	Diarrhea?
1	71	F	103	A	Yes
2	72	F	105	A	Yes
3	74	F	105	A	No
4	86	F	107	B	No
5	83	F	107	B	No
6	68	F	109	A	Yes
7	69	F	109	C	No
8	64	F	111	A	Yes
9	66	M	111	A	Yes
10	68	M	104	A	Yes
11	70	M	106	A	No
12	86	M	110	A	No
13	73	M	112	B	No
14	82	M	219	C	No
15	72	M	221	C	No
16	70	M	221	B	No
17	77	M	227	D	No
18	80	M	227	D	No
19	71	F	231	A	Yes
20	68	F	231	D	Yes
21	64	F	233	A	No
22	73	F	235	A	Yes
23	75	F	235	B	No
24	78	F	222	C	No
25	72	F	222	A	No
26	66	M	224	B	No
27	69	M	226	A	Yes
28	75	M	228	E	No
29	71	M	230	A	Yes
30	83	M	232	F	No
31	84	M	232	D	No
32	79	M	234	A	Yes
33	72	M	234	D	Yes
34	77	M	236	A	Yes
35	78	M	236	B	No
36	80	M	238	D	No

a) Enter the data into SPSS

b) Recode the variable, “Age”, into three (03) categories: <60 = Old adults; 60-70: Seniors; >70: Wisemen

c) Is there any relationship between: Age category and diarrhea status?

d) Construct a table of the illness (diarrhea) by menu type. Use diarrhea status as column labels and menu types as row labels.

e) Construct a two-by-two table of the illness (diarrhea) by exposure to menu A.Interpret your results.

12. The number of disk Input/Output's (I/O's) and processor times of seven programs were measured as:

number of disk (x)	14	16	27	42	39	50	83
processor times (y)	2	5	7	9	10	13	20

Required:

a) Scattered diagram

b) Regression line relating x and y.

c) Interpret your result

Coefficient of Determination and explain your results

13. In a study of physical fitness and cardiovascular risk factors in children, blood pressure and recovery index (post exercise recovery rate, an indicator of fitness) were measured (Hoffman and Walter 1989). Multiple regression was used to look at the relationship between systolic blood pressure and recovery index, adjusted for age, race, area of residence and ponderal index (wt/ht^2). For the boys, the adjusted regression coefficient of systolic blood pressure on recovery index was given as follows:

b = –0.086, SE b = 0.039, 95% CI = –0.162 to –0.010.

a) What is meant by 'multiple regression analysis'?

b) What is meant by the terms 'b', 'SE b' and '95% CI'?

c) What assumptions about the variables are required for these analyses to be valid?

d) Why was the regression adjusted and what does this mean?

e) What would be the effect of adjusting for race if systolic blood pressure were related to race and recovery index were not?

f) What would be the effects of adjusting for ponderal index if blood pressure and recovery index were both related to ponderal index?

14. Match the statements below with the corresponding terms from the list.

A) R^2 adjusted B) Residual plots C) R^2 D) Residual E) Influential points F) outliers

___ Worst kind of outlier, can totally reverse the direction of association between x and y

____ Used to check the assumptions of the regression model.

____ Used when trying to decide between two models with different numbers of predictors.

____Proportion of the variability in y explained by the regression model.

____ ls the observed value of y minus the predicted value of y for the observed x.

____ A point that lies far away from the rest.

15. In regression analysis, the variable that is used to explain the change in the outcome of an experiment, or some natural process, is called

a. the x-variable

b. the independent variable

c. the predictor variable

d. the explanatory variable

e. all of the above (a-d) are correct

f. none are correct

16. In a regression and correlation analysis if $r^2 = 1$, then

a. SSE = SST
b. SSE = 1
c. SSR = SSE
d. SSR = SST

17. In a regression analysis if SSE = 200 and SSR = 300, then the coefficient of determination is

a. 0.6667

b. 0.6000

c. 0.4000

d. 1.5000

18. If the correlation coefficient is 0.8, the percentage of variation in the response variable explained by the variation in the explanatory variable is

a. 0.80% b. 80% c. 0.64% d. 64%

19. A residual plot:

a. displays residuals of the explanatory variable versus residuals of the response variable.

b. displays residuals of the explanatory variable versus the response variable.

c. displays explanatory variable versus residuals of the response variable.

d. displays the explanatory variable versus the response variable.

e. displays the explanatory variable on the x axis versus the response variable on the y axis.

Complete the statement by filling in the blank. When constructing a confidence interval, if the level of confidence increases the margin of error will …………..and the confidence interval will be……………………… A larger sample size will improve the accuracy of the confidence interval, therefore margin of error will ………………..and the confidence interval will be …………….

A) Decrease, wider. Increase, narrower B) Increase, narrower. Decrease, wider.

C) Increase, wider. Decrease, narrower. D) Decrease, narrower. Increase, wider.