Kenya Medical Training College
Department of Clinical Medicine
Year Two Semester One
Guide to Calculations in Statistics
5th November 2020

Willis J. Opalla

Guide to Calculations in Statistics

- Learning Objective

  To apply formulae for statistical measures in making statistical inferences.

Learning Outcomes

- By the end of this session, you should be able to
  1. Construct a frequency distribution table for a given data set.
  2. Explain the relationship between various parameters and statistics.
  3. Select the appropriate formulae to solve problems in statistics and make correct inferences.

Frequency Distribution Table

- Consider the data set 3, 5, 2, 4, 0, 1, 3, 5, 2, 3, 2, 3, 3, 2, 4, 1 for number of decayed teeth among adolescents with a craving for sweets and other confectionaries.
- To construct a frequency table:
  - Step 1: Rearrange the values from the smallest to the largest. i.e. 0,1,1,2,2,2,2,3,3,3,3,3,4,4,5,5
  - Step 2: Count the numbers that are the same.
- A question may require calculation of mean, mode, variance and standard deviation.

Frequency Distribution Table

- The next step in solving this type of question is to construct a frequency distribution table.
- Mandatory captions for the table are those for *x* (each random observation or item or random number), tally (a score for the number of times each observation appears in the data set), frequency, *f*, (number of times each observation appears in the data set).

x	Tally	Frequency

Frequency Distribution Table

- Other captions to be included e.g. cumulative frequency, CF, relative frequency, e.t.c. depend on what the question is asking for.
- Each stub represents each of the random numbers or observation in the data set, *i.e.* x_1, x_2, x_3, ……$x_{n,}$ (in ascending order), where n is the total number of items in the sample.
- For grouped data, the stubs represent classes into which the data has been grouped.

Frequency Distribution Table

- From this data set 3, 5, 2, 4, 0, 1, 3, 5, 2, 3, 2, 3, 3, 2, 4, 1, the highest number of decayed teeth is 5 while the least is 0 and the sample size, n, is 16.
- *x* therefore represents number of decayed teeth:

No. of Decayed Teeth x	Tally	Frequency f	Cumulative Frequency CF	Relative Frequency, RF
0	/	1	1	0.0625
1	//	2	3	0.125
2	////	4	7	0.25
3	/////	5	12	0.3125
4	//	2	14	0.125
5	//	2	16	0.125
n = 16	Total = 16	Total = 16		Sum = 1

Frequency Distribution Table

- From the table, note that:
  - n = 16 and $\sum f_i = 16$. Hence $n = \sum f_i$
  - Relative Frequency $= \dfrac{\text{Frequencies}}{\sum \text{ of Cumulative Frequencies}}$

    $= \text{Frequency} \div \text{All outcomes}$

    $= \dfrac{\text{Frequency}}{n}$ or $\dfrac{\text{Frequency}}{\text{Sample size}}$

  - $\sum RF = 1$

Frequency Table for Grouped Data

- The simple frequency table is not used to present the data for large samples.
- The data should be grouped into classes with equal class intervals.

1. The number of classes (k).
   - Large class intervals (class size) hence fewer classes are not desirable as information will be lost.
   - Many classes complicate summarization of data.

- As a rule, number of classes, k:
  - $6 \leq k \leq 15$, or $k = 1 + 3.322 (\log n)$

Frequency Table for Grouped Data

2. For a quantitative data set,
   - Range (R) = x_{max} - x_{min}

3. The Class interval (w).
   - Class intervals are the width, w, of the classes.
   - Classes should have the same intervals, i.e. they should have equal width.
   - To determine the class interval:

$$\text{Class interval} \geq \frac{\text{Range}}{\text{Number of Classes}} \quad \text{i.e.} \quad w \geq R / k.$$

Frequency Table for Grouped Data

- Use of the formula k = 1+3.322(log n)
  - Example: Assume the sample size, n = 100, then

    k = 1+3.322(log 100)

    = 1 + 3.3222 (2) = 7.6 ≅ 8.
- Assuming the smallest value = 5 and largest= 61, and given that R = $x_{max} - x_{min}$ then,

  R = 61 – 5 = 56 and using the formula w ≥ R / k,

  w = 56 / 8 = 7.
- For more comprehensible summarization, the class width should be 5 or 10 or multiples of 10.

Frequency Table for Grouped Data

- Worked Example
  - To determine the number of classes appropriate for grouping data in the frequency distribution of the ages of 189 subjects in a study on smoking cessation:

30 34 35 37 37 38 38 38 38 39 39 40 40 42 42 43 43 43 43 43
43 44 44 44 44 44 44 44 45 45 45 46 46 46 46 46 46 47 47 47
47 47 47 48 48 48 48 48 48 48 49 49 49 49 49 49 49 50 50 50
50 50 50 50 50 51 51 51 51 52 52 52 52 52 52 53 53 53 53 53
53 53 53 53 53 53 53 53 53 53 53 54 54 54 54 54 54 54 54 54
54 54 54 55 55 55 56 56 56 56 56 56 57 57 57 57 57 57 57 58
58 59 59 59 59 59 59 60 60 60 60 61 61 61 61 61 61 61 61 61
61 61 62 62 62 62 62 62 62 63 63 64 64 64 64 64 64 65 65 66
66 66 66 66 66 67 68 68 68 69 69 69 70 71 71 71 71 71 71 71
72 73 75 76 77 78 78 78 82

Frequency Table for Grouped Data

- Solution:
  - Since the number of observations is 189,

$$k = 1+3.322\ (\log 189)$$
$$= 1 + 3.3222\ (2.276)$$
$$\cong 9,$$
$$R = 82 - 30 = 52$$
$$w = 52 / 9 = 5.778$$

- However, a class interval of 10 makes the summarization more clear.
- Increasing w, causes a reduction in k.

Frequency Table for Grouped Data

- With a class interval, w = 10, then:

Class	Frequency, f.
30 – 39	11
40 – 49	46
50 – 59	70
60 – 69	45
70 – 79	16
80 – 89	1
n = 189 k = 6 w = 10	$\sum f_i = 189$

- Note that n = $\sum f_i$ = 189 i.e. Sum of frequency = sample size.

Frequency Table for Grouped Data

- Cumulative Frequency, CF:
  - CF = $f_1 + f_2 + f_3 \ldots\ldots + f_n$
    $= \sum f_i$
- Relative Frequency, RF:
  - Frequency ÷ all outcomes
- The Cumulative Relative Frequency:
  - CRF = RF_1 +RF+ $RF_3 \ldots\ldots + RF_n$
    $= \sum RF_i$
- The Mid-interval:
  - (Lower border of a class + its upper border) ÷ 2.

Mean for Grouped Data

- The arithmetic mean for grouped data

$$m = \frac{\sum fx}{f}$$

Where

x is the mid-point for each class and f, frequency.

Median for Grouped Data

- The arithmetic mean for grouped data

$$\text{median} = L = \frac{(n/2 - F)\, c}{f}$$

Where,

L is the lower limit of the median class.

n the total number of observations.

F the number of observations up to the median class and

f the frequency in the median class.

c the class interval in the median class.

Median for Grouped Data

- Worked Example:
  - For the data set 8 8 9 9 9 9 10 10 10 10 10 10 10 10 11 11 11 11 11 11 11 11 11 11 11 11 11 11 12 12 12 12 12 12 12 12 13 13 14 16 16 , n = 41
  - If odd numbers, then median = $\frac{n + 1}{2}$

    $= \frac{41 + 1}{2} = 21$
  - Therefore the median lies in the 21st observation.

Mid-point for Grouped Data, m

- Worked Example:

  Midpoint = $(\sum x_i) \div 2$

  - For the data set 8 8 9 9 9 9 10 10 10 10 10 10 10 10 11 11 11 11 11 11 11 11 11 11 11 11 11 12 12 12 12 12 12 12 12 13 13 14 16 16 , n = 41

- The data be grouped in classes 8 – 10, 11 -13, 14 – 16

  The mid-point of class 8 – 10 = (8 + 10) ÷ 2

  = 9

  For class 7.5 – 10.5, midpoint = (7.5 + 10.5) ÷ 2

  = 9

Mode for Grouped Data

- Mode = 3 median – 2 mean

  or

- Mode = $L_m + \frac{d_1 C}{d_1 + d_2}$

Where,

L_m is the lower limit of the modal class.

d_1 frequency in the modal class minus the frequency in the preceding class.

d_2 frequency in the modal class minus the frequency in the succeeding class.

C class interval of the modal class.

Grouped data: Frequency distribution table for the ages of 189 smoking cessation respondents.

R.f= freq/n

Class k	Mid- Point. m	Frequency, f	Cumulative Frequency, CF	Relative Frequency, RF	Cumulative Relative Frequency, CRF
30 – 39	34.5	11	11	0.0582	0.0582
40 – 49	44.5	46	57	0.2434	---------
50 – 59	54.5	----	127	--------	0.6720
60 – 69	------	45	-----	0.2381	0.9101
70 – 79	74.5	16	188	0.0847	0.9948
80 – 89	84.5	1	189	0.0053	1
n = 189 k = 6 w = 10	$\sum m = 357$	$\sum f_i = 189$		1	

Practice Questions

- From the table, determine the following:
  1. The number of observations with age less than 50 years.
  2. The number of observations with age between 40-69 years.
  3. Relative frequency of observations with age between 70-79 years.
  4. Relative frequency of observations with age more than 69 years.
  5. The percentage of observations with age between 40-49 years.

Practice Questions

- From the table, determine the following:
  6. The percentage of observations with age less than 60 years.
  7. The Range (R).
  8. Number of classes (K).
  9. The width of the interval (W).

Frequency Distribution Table

- Consider the data set 8, 5, 4, 12, 15, 5, 7 for number of decayed teeth among adolescents with a craving for sweets and other confectionaries.
- If the question requires calculation of mean, mode, variance and standard deviation then foremost, the formulae for those measures must be scribbled in order to identify the factors to assign captions in the frequency distribution table to be constructed.

Frequency Distribution Table

- Consider the data set 8, 5, 4, 12, 15, 5, 7 …
- Since variance, $s^2 = \frac{\Sigma(X - \overline{X})^2}{n-1}$ and SD, $S = \sqrt{S^2}$
- The frequency distribution table to be constructed should have captions for $x - \overline{x}$ and $(x - \overline{x})^2$.

x	Frequency f	$x - \overline{x}$	$(x - \overline{x})^2$

Calculations for Grouped Data

- Short cut formula for Variance

$$S^2 = \sum x^2 - \frac{(\sum x)^2 / n}{n - 1}$$

i.e. S^2 = Sum of x^2 - Square of sum of x / n

- From the data set 8, 5, 4, 12, 15, 5, 7
- $\sum x = 56$, $\sum x^2 = 548$

$$\text{Variance, } S^2 = \sum x^2 - \frac{(\sum x)^2 / n}{n - 1} = \frac{548 - (56^2 \div 7)}{n - 1}$$

$$= \frac{548 - (3136 \div 7)}{7 - 1} = \frac{548 - (448)}{6} = 16.67$$

Modified Formula for Variance

- When data is grouped in the frequency table, Variance is calculated using a modified shortest formula:

$$S^2 = \frac{\sum fm^2 - (\sum fm)^2 / n}{n - 1}$$

where f is frequency and m is the mid-point of each class.

-

Strategies for Answering Statistics Questions

- Step 1: Read the whole question.
- Step 2: Identify the statistical measures required to be calculated.
- Step 3: Write down the formula for calculating each measure that is asked for in the question.
- Step 4: Construct a frequency distribution table
- for the set of data given in the question

Strategies for Answering Statistics Questions

- Step 4: Construct a frequency distribution table for the set of data given in the question

  or

  If a table has been given, then complete it by inserting columns for the factors that will help solve the equations for calculating each statistical measure.

Strategies for Answering Statistics Questions

- Step 5: Pick values from the table that is now completed and apply them in the formula for calculating each statistical measure.
- One frequency distribution table is enough to provide all columns for the values required to calculate each of the variables in the question asked.

Question 1

- The Erythrocyte Sedimentation Rate (ESR) in normal individuals was 2 4 5 4 2 4 5 and 3.

a) Identify the type of series of the above data.
b) Calculate the mean ESR.
c) What is the median of the ESR?
d) What is the mode?
e) Calculate the variance.
f) Calculate the mean deviation.
g) Calculate the Coefficient of Variation.
h) Calculate the Standard Error.

Question 2

- The following scores of marks were recorded among second year students in their Health Statistics exam.

Use the data to calculate

a) The mean score.
b) The variance.
c) The standard deviation.
d) The coefficient of variation.
e) The standard error.

x	Frequency
60 - 61	10
62 - 63	20
64 – 65	45
66 – 67	50
68 – 69	40
70 - 71	15

Question 2

- Step 2:
  - The statistical measures of interest are mean, variance, standard deviation, coefficient of variation and standard error.
- Step 3:

Mean , $\overline{x} = \frac{\sum x_i}{n}$ but $n = \sum f_i$

Variance, $S^2 = \frac{\sum (x - \overline{x})^2}{n - 1}$ ….which has a round off error.

or

Variance, $S^2 = \frac{\sum x^2 - (\sum x)^2/n}{n - 1}$ …which has no round off error.

Step 3 Cont…

Standard deviation, $S = \sqrt{S^2}$

Coefficient of variance, $CV = \frac{S}{\overline{x}} \times 100.$

Standard error of mean, $\delta\overline{x} = \frac{\delta s}{\sqrt{n}}$

or Standard error, $SE = \frac{S}{\sqrt{n}}$

Step 4

- Complete the frequency distribution table that has been given by inserting columns for factors that will be required to calculate the statistical measures, based on formula for each

x	Frequency, f_i	Mid –point x_i	$f_i x_i$	$x - \bar{x}$	$(x - \bar{x})^2$
60 - 61	10	60.5	605	-5.5	30.25
62 - 63	20	62.5	1250	-3.5	12.25
64 – 65	45	64.5	2902.5	-1.5	2.25
66 – 67	50	66.5	3325	0.5	0.25
68 – 69	40	68.5	2740	2.5	6.25
70 - 71	15	70.5	1057.5	4.5	20.25

Step 4 Cont….

x	Frequency, f_i	Mid –point x_i	$f_i x_i$	$x - \bar{x}$	$(x - \bar{x})^2$
60 - 61	10	60.5	605	-5.5	30.25
62 - 63	20	62.5	1250	-3.5	12.25
64 – 65	45	64.5	2902.5	-1.5	2.25
66 – 67	50	66.5	3325	0.5	0.25
68 – 69	40	68.5	2740	2.5	6.25
70 - 71	15	70.5	1057.5	4.5	20.25
	$\sum f_i = 180$ hence n = 180	$\sum(x_i/2 = 393$ $\sum x_i = 393$	$\sum f_i x_i = 11880$		$\sum(x - \bar{x})^2 = 71.5$

Mean = 11880/180

Variance, S^2 = 71.5/n-1
But n-1 = 180 -1
n - 1= 179
S^2 = 0.399

Step 5

- Answers as worked out from the completed table:

Mean, $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$ ….Already available in the completed table. Their columns were added because they are necessary for calculation of mean.

$= \frac{11880}{180}$

$= 66$

Step 5 Cont…

- Or

Mean, $\bar{x} = \frac{\sum x_i}{n}$ ….Because this is grouped data, if $\sum x_i$ is applied to calculate the mean, then n will be the classes into which the set of data has been grouped.

$= \frac{393}{6 \text{ classes}}$

$= 65.5$

$= 66$

Step 5 Cont…

- Variance, $s^2 = \frac{\Sigma(X - \overline{X})^2}{n-1}$ but $n = \sum f_i = 180$

$$S^2 = 71.5 \div (180 - 1)$$
$$= 71.5 \div 179$$
$$= 0.399$$
$$= 0.4$$

- Standard deviation, $S = \sqrt{S^2}$
$$= \sqrt{0.4}$$
$$= 0.632$$

Step 5 Cont…

- Coefficient of Variation, CV, $= \frac{S}{\bar{x}} \times 100.$

$$= \frac{0.632}{66} \times 100$$

$$= 0.958 = 0.96$$

- Standard error of mean, $SE = \frac{S}{\sqrt{n}}$

$$= 0.632 \div \sqrt{180}$$

$$= 0.049$$

Summary

- Calculations in statistics demands a good grasp of the formulae for various statistical measure and clear understanding of the question.
- Construction of the frequency distribution table is not only a mandatory step in solving problems in statistics, but also organizes data and simplifies calculations.

References

- Heumann, C., Schomaker, M. and Shalabh (2016) *Introduction to Statistics and Data Analysis With Exercises, Solutions and Applications in R*, Springer International Publishing Switzerland.
- Holmes, L, Jr. (2018) *Applied Biostatistical Principles and Concepts: Clinician's Guide to Data Analysis and Presentation*, Routledge, Taylor & Francis group, New York.