Unit 1 - Summarizing Data
Homework #1 (Unit 1 – Summarizing Data)

Solutions

1.

a) Are the exams "in-class"/proctored or are they take-home?
The exams are take home. In particular, they are 2 week, open book exams.

b) How are the exam grades weighted in the final course grade determination?
Best exam -40%, Second best – 20%, Third best – 15%

c) How are the homeworks graded?
The homeworks are graded pass/fail. Pass is a score of 100.

d) Is attendance in Zoom classes required?
No.

e) Your course score is ***not determined by the columns in Blackboard.*** How is the course score determined?
Course score = (.25)[Homework score] + (.40)[Best test] + (.20)[2nd best test] + (.15)[3rd best test]

f) How are the final course letter grades determined?

Course Score	Letter Grade
95 and over	A
90-94	A MINUS
87-89	B PLUS
83-86	B
80-82	B MINUS
77-79	C PLUS
70-76	C
Below 70	F

g) Is it possible to obtain the exam questions early?
No

h) Are late homework and exam submissions allowed (yes or no)?
Yes, per the late policy for this course

i) What is the policy on late homework and late exam submissions?
Late submission within 48 hours of due date: - 10 points
Late submission, post 48 hours but within 1 week: -20 points
Submissions after 1 week are not accepted.

2. For each of the following variables indicate whether it is quantitative or qualitative and specify the measurement scale that is employed when taking measurements on each:

a. Class standing of members of this class relative to each other
Categorical/Qualitative ordinal
b. Admitting diagnosis of patients admitted to a mental health clinic
Categorical/Qualitative nominal
c. Weights of babies born in a hospital during a year
Numerical/Quantitative continuous ratio
d. Gender of babies born in a hospital during a year
Categorical/Qualitative nominal
e. Range of motion of elbow joint of students enrolled in a university health sciences curriculum
Numerical/Quantitative continuous ratio
a. f) Under-arm temperature of day-old infants born in a hospital
Numerical/Quantitative continuous interval

3. Let $x_1=3$, $x_2=1$, $x_3=4$, and $x_4=6$

3a. Express the following sum in sigma notation and evaluate numerically.

$$(x_1 + x_2 + x_3 + x_4)^2$$

Answer = 196

$$\begin{aligned}(X_1 + X_2 + X_3 + X_4)^2 &= \left[\sum_{i=1}^{4} X_i\right]^2 \\ &= (3+1+4+6)^2 \\ &= 14^2 \\ &= 196.\end{aligned}$$

3b. Express the following sum in sigma notation and evaluate numerically.

$x_1^2 + x_2^2 + x_3^2 + x_4^2$

Answer = 62

$$
\begin{aligned}
X_1^2 + X_2^2 + X_3^2 + X_4^2 &= \sum_{i=1}^{4} X_i^2 \\
&= 3^2 + 1^2 + 4^2 + 6^2 \\
&= 9 + 1 + 16 + 36 \\
&= 62.
\end{aligned}
$$

3c. Evaluate the following numerically.

$\Sigma (X_i - 1)^2$ for $i=1\ldots4$.

Answer = 38

$$
\begin{aligned}
\sum_{i=1}^{4}(X_i - 1)^2 &= (3-1)^2 + (1-1)^2 + (4-1)^2 + (6-1)^2 \\
&= 2^2 + 0^2 + 3^2 + 5^2 \\
&= 4 + 0 + 9 + 25 \\
&= 38.
\end{aligned}
$$

Note:

$$
\begin{aligned}
\sum_{i=1}^{4}(X_i - 1)^2 &= \sum_{i=1}^{4}\left[X_i^2 - 2X_i + 1\right] \\
&= \sum_{i=1}^{4} X_i^2 \; - \; 2\sum_{i=1}^{4} X_i + 1\sum_{i=1}^{4} \\
&= 62 \; - \; (2)(14) + (1)(4) \\
&= 38.
\end{aligned}
$$

3d. Evaluate the following numerically.

$\Sigma\, 3X_i$ for $i=1\ldots4$.

Answer = 42

$$\sum_{i=1}^{4} 3X_i = 3\sum_{i=1}^{4} X_i$$
$$= 3(14)$$
$$= 42$$

4. The following are behavioral ratings as measured by the Zang Anxiety Scale (ZAS) for 26 persons with a diagnosis of panic disorder:

53	51	46	45	40	35
59	51	45	60	35	
45	38	53	43	31	
36	40	41	41	38	
69	41	46	38	36	

4a. Compute the mean, median, mode, range, variance, and standard deviation, and the 25th and 75th percentiles.

Mean = 44.5
Median = 42
Mode (there are 3 actually) = 38, 41, 45
Range = 38
25^{th} Percentile = 38
75^{th} Percentile = 51

MEAN $\bar{x} = \frac{1}{n}\sum_{i=1}^{26} X_i$

$= \frac{1}{n}(1156) \qquad = 44.46 \quad \text{so} \quad \bar{x} = 44.5$

MEDIAN First solve $\left(\frac{n+1}{2}\right) = \left(\frac{26+1}{2}\right) = 13.5$

Median is midpoint of 13^{th} and 14^{th} observation.

$\tilde{x} = \frac{1}{2}(41+43) \qquad \text{so} \quad \tilde{x} = 42$

MODE This sample is tri - modal 38,41,45

RANGE Maximum - Minimum

$= 69 - 31$ so range $= 38$

VARIANCE Let's save ourselves the trouble of a very long brute force formula by using the formula for grouped data.

Let j index the unique values. There are 14 unique values.

j	X_j	f_j	$(x_j - \bar{x})^2$	$f_j(x_j - \bar{x})^2$
1	31	1	182.25	182.25
2	35	2	90.25	180.50
3	36	2	72.25	144.50
4	38	3	42.25	126.75
5	40	2	20.25	40.50
6	41	3	12.25	36.75
7	43	1	2.25	2.25
8	45	3	0.25	0.75
9	46	2	2.25	4.50
10	51	2	42.25	84.50
11	53	2	72.25	144.50
12	59	1	210.25	210.25
13	60	1	240.25	240.25
14	69	1	600.25	600.25
TOTALS		26		1998.50

$$S^2 = \frac{\sum_{j=1}^{14} f_j \left(x_j - \bar{x}\right)^2}{\left(\sum_{j=1}^{14} f_j\right) - 1} = \frac{1998.50}{25} \qquad So \;\; S^2 = 79.94$$

Standard deviation $S = \sqrt{S^2}$ *So* $S = 8.94$

25th Percentile

First solve (.25) (n) = (.25) (26) = 6.5
So 25th percentile is the 7th observation $P_{25} = 38$

75th Percentile

First solve (.75) (n) = (.75) (26) = 19.5
So 75th percentile is the 20^{th} observation $P_{75} = 51$

4b. The following are behavioral ratings as measured by the Zang Anxiety Scale (ZAS) for 21 healthy controls:

26	26	25	25	25
28	26	26	25	
34	30	31	28	
26	34	25	25	
25	28	25	25	

Compute the mean, median, mode, range, variance, and standard deviation, and the 25th and 75th percentiles.

Mean = 27.0
Median = 26
Mode = 25
Variance = 8.35
Standard Deviation = 2.89
25^{th} Percentile = 25
75^{th} Percentile = 28

MEAN: $\bar{x} = \frac{1}{n}\sum_{i=1}^{21} x_i = \frac{1}{21}(568) = 27.04$

MEDIAN: Solving $\left(\frac{n+1}{2}\right) = \left(\frac{21+1}{2}\right) = 11 \rightarrow$ Median is the 11^{th} ordered observation = 26

MODE: Most frequently occurring observation = 25

RANGE: Maximum - minimum = 34 - 25 = 9

Variance There are 6 unique values.

j	X_j	f_j	$(x_j - \bar{x})^2$	$f_j(x_j - \bar{x})^2$
1	25	9	4	36
2	26	5	1	5
3	28	3	1	3
4	30	1	9	9
5	31	1	16	16
6	34	2	49	98
TOTALS		21	80	167

$$S^2 = \frac{\sum_{j=1}^{6} f_j (x_j - \bar{x})^2}{\left(\sum_{j=1}^{6} f_j\right) - 1} = \frac{167}{20} \qquad \text{So} \qquad S^2 = 8.35$$

Standard deviation $S = \sqrt{S^2} = \sqrt{8.35}$ So $S = 2.89$

25th Percentile

Solving (.25) (n) = (.25) (21) = 5.25
So 25th percentile is 6th observation $P_{25} = 25$
Note - I get this by noticing from the table above that the smallest value (=25) occurs with a frequency of 9 times in the sample.

75th Percentile

Solving (.75) (n) = (.75) (21) = 15.75
So 75th percentile is 16th observation $P_{75} = 28$
Note – I get this by noticing in the table that the value = 28 occurs with a frequency of 3 times in the sample and comes after the first 9 observations all equal to 25 and after the next 5 observations all equal to 26, so that the value of 28 is the 15th, 16th and 17th observations in the ordered sample.

5. The following table shows the age distribution of cases of a certain disease reported during a year in a particular state.

Age	Number of Cases
5-14	5
15-24	10
25-34	20
35-44	22
45-54	13
55-64	5
TOTAL	75

5a. Construct a frequency table with columns for class endpoints, class midpoint, frequency, relative frequency, cumulative frequency, and cumulative relative frequency.

Class Endpoints	Class Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Freq.
5-14.99	10	5	.067	5	.067
15-24.99	20	10	.133	15	.200
25-34.99	30	20	.267	35	.467
35-44.99	40	22	.293	57	.760
45-54.99	50	13	.173	70	.933
55-64.99	60	5	.067	75	1.000
TOTALS			1.000		

5b. Estimate the values of the mean, median, variance, and standard deviation. Tip - Use the midpoints of each age interval as your values and use number of cases as their frequencies. For example, the value 10 has an estimated frequency of 5, the value 20 has an estimated frequency of 10, and so on.

Midpoint X_j	Frequency f_j	$X_j f_j$	$(x_j - \bar{x})$	$f_j(x_j - \bar{x})^2$
10	5	50	-25.7	3302.45
20	10	200	-15.7	2464.90
30	20	600	-5.7	649.80
40	22	880	4.3	406.78
50	13	650	14.3	2658.37
60	5	300	24.3	2952.45
Total	75	2680		12434.75

$$MEAN \quad \bar{x} = \frac{\sum_{j=1}^{6} f_j x_j}{\sum_{j=1}^{6} f_j} = \frac{2680}{75} \quad \text{So} \quad \bar{x} = 35.7$$

MEDIAN *Note to reader* – I've consulted a number of texts on this. There is no single correct answer. With interval data, whatever median you calculate is an approximation. Here is what is suggested in <u>Think and Explain with Statistics (Lincoln E. Moses, page 64)</u>

$$\text{First solve } \frac{n+1}{2} = \frac{75+1}{2} = 38th \text{ observation}$$

Examination of the table reveals that the 38th observation is in the interval 35 to 44.99

Set the following quantities:

The letter l = lower limit of interval = 35
The letter u = upper limit of interval = 44.99
R = cumulative frequency up to the lower limit of interval = 35
M = # observations contained in interval = 22
N = total # observations = 75

An approximate solution for the median is calculated as

$$\tilde{x} = l + \left[\frac{N/2 - R}{M}\right](u-l) \qquad = 35 + \left[\frac{75/2 - 35}{22}\right](44.99 - 35) = 36.135 \text{ or } \mathbf{37}$$

VARIANCE

$$S^2 = \frac{\sum_{j=1}^{6} f_j \left(x_j - \bar{x}\right)^2}{\left(\sum_{j=1}^{6} f_j\right) - 1} = \frac{12434.75}{74} \text{ so } S^2 = 168.04$$

Standard deviation $S = \sqrt{S^2}$ so $S = 13.0$