1. (10 points total)

The classification of variables by type distinguishes qualitative versus quantitative and also continuous versus discrete.

In this exercise, for one point each, classify each of the following measurements as categorical, ordinal or quantitative, by placing an **"x"** in the appropriate column.

		Classification (choose ONE) = Qualitative categorical	Qualitative ordinal	Quantitative discrete OR continuous
1a	Response to treatment coded as 1=no response, 2=minor response, 3=major improvement, 4=complete recovery		X	
1b	Annual income (pre-tax dollars)			X
1c	Body temperature (degrees Celsius)			X
1d	Area of parcel of land (acres)			X
1e	Population density (people per acre)			X
1f	Political party affiliation coded 1=Democrat, 2=Republican/non Tea Party, 3=Tea Party, 4=Independent, 5=Other	X		
1g	Presence of type II diabetes mellitus codes as "yes" or "no"	X		
1h	White blood cells per deciliter of whole blood			X
1i	Leukemia rates in geographic regions (cases per 100,000 people)			X
1j	Blood cholesterol level classified as either 1=hypercholesterolemic, 2=borderline hypercholesterolemic, 3=normocholesterolemic		X	

2. (10 points total)

The following table lists length of stay in hospital (days) for a sample of 25 patients.

5	10	6	11	5	14	30	11	17	3
9	3	8	8	5	5	7	4	3	7
9	11	11	9	4					

2a. (4 points).

Construct a frequency/relative frequency table for these data using 5-day class intervals. Include columns for the frequency counts, relative frequencies, and cumulative frequencies.

Dear class - I will accept either of two solutions here -cb

Hospital Stay Length (days)	Frequency Count	Relative Frequency	Cumulative Frequency
1-5	9	0.36	9
6-10	9	0.36	18
11-15	5	0.2	23
16-20	1	0.04	24
21-25	0	0	24
26-30	1	0.04	25
total	25	1	

Hospital Stay Length (days)	Frequency Count	Relative Frequency	Cumulative Frequency
0–4	5	0.20	5
5-9	12	0.48	17
10-14	6	0.24	23
15-19	1	0.04	24
20-24	0	0.00	24
25-29	0	0.00	24
30-34	1	0.04	25
total	25	1	

2b. (2 points).
What percentage of hospital stays were less than 5 days?

Data values less than 5: 3, 4, 4, 3, 3 →
% of hospital stays less than 5 days = (5/25)*100 = 20%

2c. (2 points).
What percentage of hospital stays were less than 15 days?

Data values less than 15: 5, 9, 9, 10, 3, 11, 6, 8, 11, 11, 8, 9, 5, 5, 4, 14, 5, 7, 11, 4, 3, 3, 7 →
% of hospital stays less than 15 days = (23/25)*100 = 92%

2d. (2 points).
What percentage of hospital stays were at least 15 days in length?

Data values greater than or equal to 15: 30, 17 →
% of hospital stays greater than or equal to 15 days = (2/25)*100 = 8%
***Note*: Solution is alternatively = 100% - % stays less than 15 days = (1 - .92)*100% = 8%**

3. (10 points total)

The following values are values of months between bacterial meningitis and the onset of seizures (induction time) for a sample of 13 cases.

0.10	0.25	0.50	4	12	12	24	24	31	36
42	55	96							

3a. (2 points).
Calculate the mean induction time.

Mean Induction time
= (0.01 + 0.25 + 0.50 + 4+ 12 + 12 + 24 + 24 + 31 + 36 + 42 + 55 + 96)/13
= 25.91 months

3b. (2 points)
Calculate the median induction time

Ordered value to take is (n+1)/2th largest value
= (13 + 1) /2
= 7th largest value

Median
= 7th largest value
= 24 months.

3c. (3 points)
In 1-2 sentences, compare the mean and the median. What does this comparison tell you about the shape of the distribution of induction times?

25.91 > 24. Thus, in this sample, mean > median. This suggests that the shape of the distribution of induction times is positively skewed (tail to the right).

3d. (3 points)
In 1-2 sentences, which measure of central location would you use to describe the distribution's center? Explain your preference.

For skewed data, often, the median is a better summary of the distribution with respect to summarizing central location. This is because the median identifies the separation of the lower and upper halves of the distribution.

4. (10 points total)

Consider a sample of 40 women comprised of two subgroups, **"cancer"** and **"normal."** Women in the "cancer" subgroup have a diagnosis of a malignant breast lump. Those in the "normal" subgroup do not. The following is a separate listing of parity (here, parity is defined as the number of births) values for each of the two subgroups of the same data.

Subgroup: Cancer

					2				
	2	8	3			1			2
							3		
			5						

Subgroup: Normal

4	0	2	3	2		3	3	0	3
1				4	2		2	2	
2	2	3	2	2	3	0		2	4
0	1	3		1	1	0	3	2	1

4a. (5 points).
Construct two 2x2 table that cross-classification tables, one showing counts and the other showing suitably defined percentages. You may choose to report row percentages or column percentages. In each table, cross-classify according to parity recoded as "two or fewer children" or "three or more children" and diagnosis coded as "cancer" or "normal".

Counts:

	Cancer	Normal	
≤ 2 children	4	21	25
3+ children	4	11	15
	8	32	40

ROW Percentages:

	Cancer	Normal	
≤ 2 children	4/25 = 16%	21/25 = 84%	100%
3+ children	4/15 = 26.7%	11/15 =73.3%	100%
	8/40 = 20%	32/40 = 80%	100%

COLUMN Percentages:

	Cancer	Normal	
≤ 2 children	4/8 = 50%	21/32 =65.6%	25/40 = 62.5%
3+ children	4/8 = 50%	11/32=34.4%	15/40 =37.5%
	100%	100%	100%

4b. (5 points).
In 1-2 sentences, comment on your cross-classification. Does it suggest any possible association between parity and diagnosis?

While the numbers are small, these data are suggestive of a positive association of parity with malignancy, meaning that giving birth to more children is associated with a higher occurrence of malignancy.

Note my choice of word "occurrence" here. Because these are cross-sectional data means that I have no basis for using the word "probability".

Comment using row percentages: The relative frequency of a malignant lump is higher among women with 3 or more children compared to that seen among women with 2 or fewer children (26.7% versus 16%).

Comment using column percentages: While women with 3 or more children represent 37.5% of the total sample, among the women with a malignant lump, a disproportionate percentage, 50%, have 3 or more children.

5. (20 points total)

The following table summarizes the distribution of n=1000 cases of acute gastroenteritis in a hypothetical community over the course of one calendar year.

Cases of Acute Gastroenteritis (Number) in 2010 in a Hypothetical Community	
Period	Number of cases
January 1, 2010 – March 31, 2010	240
April 1, 2010 – June 30, 2010	180
July 1, 2010 – July 31, 2010	0
August 1, 2010 – September 30, 2010	320
October 1, 2010 – December 31, 2010	260

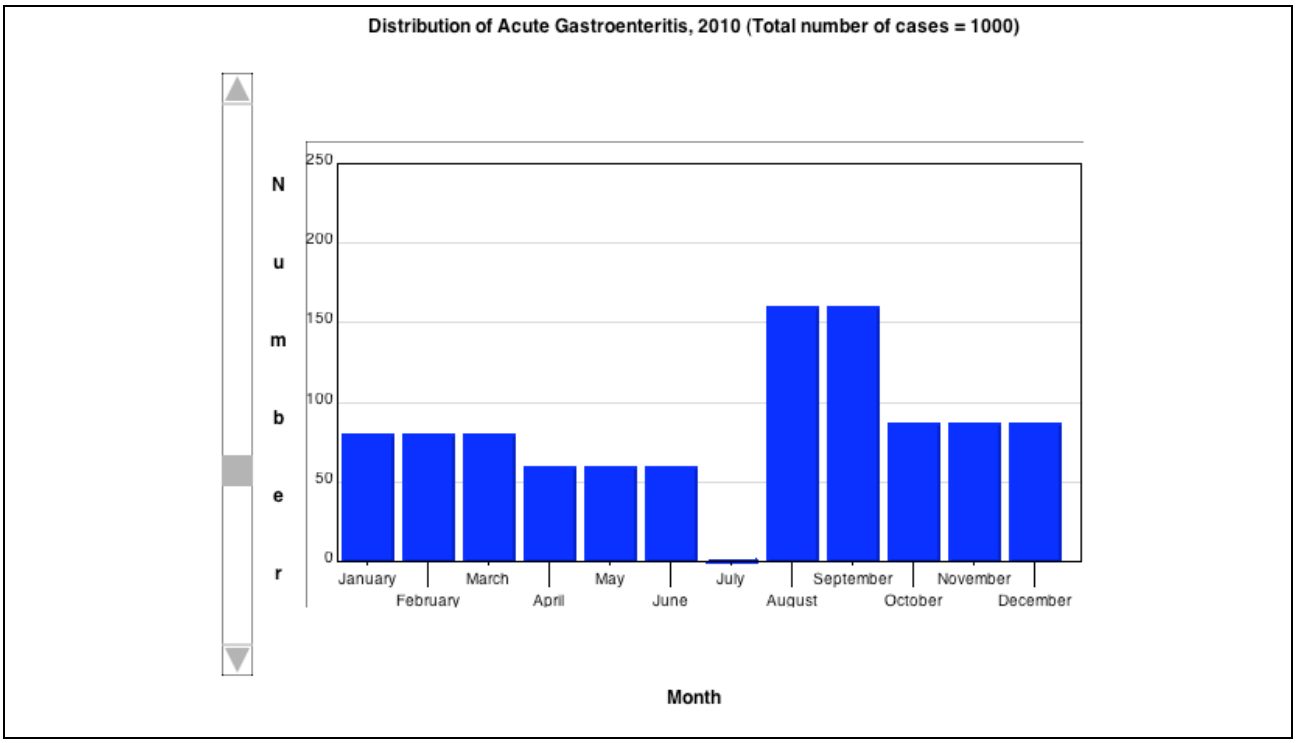
5a. (10 points)

By hand or by any means you like, construct an appropriate graphical summary of the distribution of cases of acute gastroenteritis over one calendar year. ***Tip!!*** **Notice that the calendar periods are NOT of equal duration.**

Dear Class,

I will accept a variety of graphs here. The best would be a histogram for varying interval lengths. Alas, Excel does not do this easily. Nor Stata. **Bar Graph Plot using Shodor:**

Note – To address the unequal calendar periods, I input numbers for each of the 12 months using appropriate division. For example 240 cases over 3 months represents 80 cases over each single month, for 3 months.



5b. (10 points)
In 1-2 sentences, state in words the findings of this surveillance.

Answer:
The number of cases of acute gastroenteritis was highest in the third quarter of 2010, while the lowest was in the spring months. The peak in late summer is interesting and worth pursuing further – perhaps an event was the point source of an outbreak, or the virus may have spread more quickly during the back-to-school period for young children.

6. 20 points total)

The Irish Department of Public Health publishes health care information on its websites. An issue in the development of these materials is their reading level. The recommended reading level is that of 12-14 year olds.

The following is a stem and leaf plot of the reading levels of health information postings on n=46 websites.

08	0
09	
10	0
11	0 0
12	0
13	0 0 0 0
14	0 0 0
15	0 0 0 0 0 0 0
16	0 0 0
17	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

6a. (5 points)
Which measures of location and spread would you use to describe these data? In 1-2 students, state your preference and explain your selection.
Answer:
As the data are skewed, the median would be the best measure of location and indicate the clustering of data around 17 (the mean would be 15.34, far left of the median due to the fact that the data is skewed to the left). The median absolute deviation from the median (MADM) would be the best measure of spread, as the data is skewed.

6b. (5 points)
Calculate the **mean** and **standard deviation** of these data.

Solution:

Data value	Freq	Value * Freq	(Xi - $\bar{X}$)	(Xi - $\bar{X}$)2	(Xi - $\bar{X}$)2* Freq
8	1	8	-7.37	54.3169	54.3169
9	0	0	-6.37	40.5769	0
10	1	10	-5.37	28.8369	28.8369
11	2	22	-4.37	19.0969	38.1938
12	1	12	-3.37	11.3569	11.3569
13	4	52	-2.37	5.6169	22.4676
14	3	42	-1.37	1.8769	5.6307
15	7	105	-0.37	0.1369	0.9583
16	3	48	0.63	0.3969	1.1907
17	24	408	1.63	2.6569	63.7656
	46	**707**			**226.7174**

Mean = $\sum X_i/n$ =707/46 = 15.37
Standard deviation = v($\sum (X_i -)^2$/(n-1)) = v(226.72/(46-1)) = 2.24

6c. (10 points)
By hand or by any means you like, construct a box-plot summary of these data. In developing your answer, fill in the following table.

P25 = **14**	(1.5)* IQR = **4.5**
P50 = **17**	Lower Fence = **10**
P75 = **17**	Upper Fence = **17**
Interquartile Range (IQR) = **3**	Extreme values (if any) = **8**

Solution:

Observation # for P25 = 0.25 * n = 0.25 * 46 = 11.5 → round up to 12
12th observation = 14

Observation # for P50 = 0.50 * n = 0.50 * 46 = 23
→ Because 23 is an integer, find values for observation numbers 23, 24: both are 17
P50 is the average of these: (17 +17)/2 = 17

Observation # for P25 = 0.75 * n = 0.75 * 46 = 34.5 → round up to 35
35th observation = 17

Interquartile Range = P75 – P25 = 17 – 14 = 3

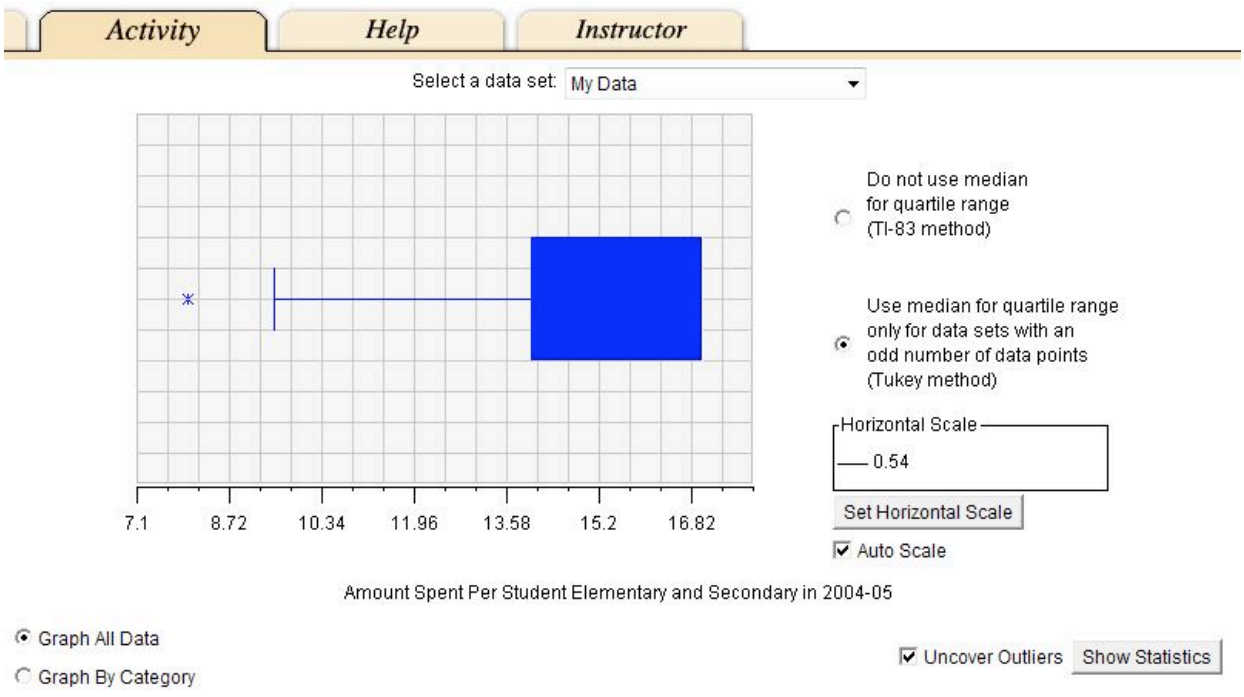
(1.5) * IQR = (1.5) * 3 = 4.5

Lower Fence = [smallest actual data value] that is larger than [P25 – (1.5)*IQR]
= [smallest actual data value] that is larger than [14 – 4.5]
= [smallest actual data value] that is larger than 9.5
= 10

Upper Fence = [largest actual data value] that is smaller than [P75 + (1.5)*IQR]
= [largest actual data value] that is smaller than [17 + 4.5]
= [largest actual data value] that is smaller than 21.5
= 17

Extreme values: since the minimum value of 8 falls below the lower fence, it is an extreme value.

Box Plot using Shodor:

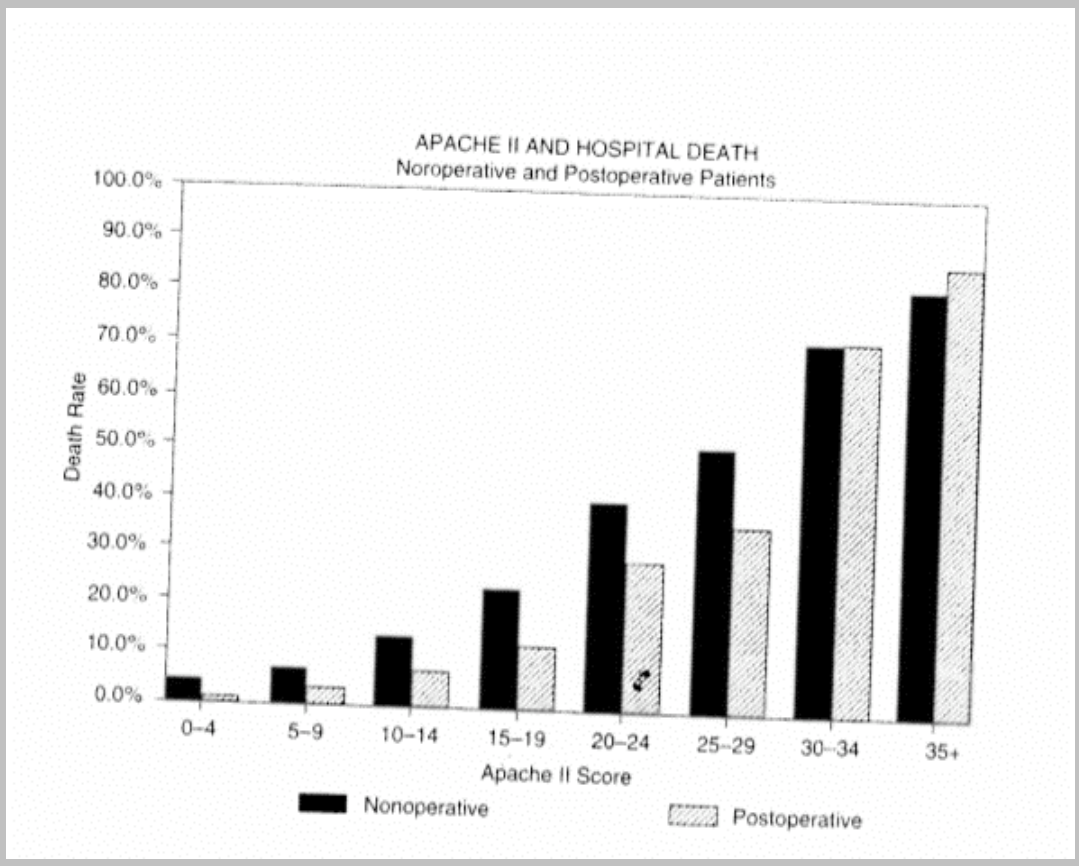


7. (20 points total)

7a. (10 points)

The figure below is an example of a **clustered bar chart**. It summarizes data on in-hospital mortalilty (Y-axis) in hospital in relationship to assessed risk of death (X-axis) using the APACHE II scale in two groups of patients: (1) nonoperative group- these patients were admitted into a medical intensive care unit without surgery; and (2) postoperative – these patients underwent surgery and were then admitted to the intensive care unit.

APACHE II scores have valid range 0 to 71, with 0=least risk of death and 71=greatest risk of death. In 1-2 sentences, interpret this clustered bar chart, focusing on a comparison of the two groups of patients.



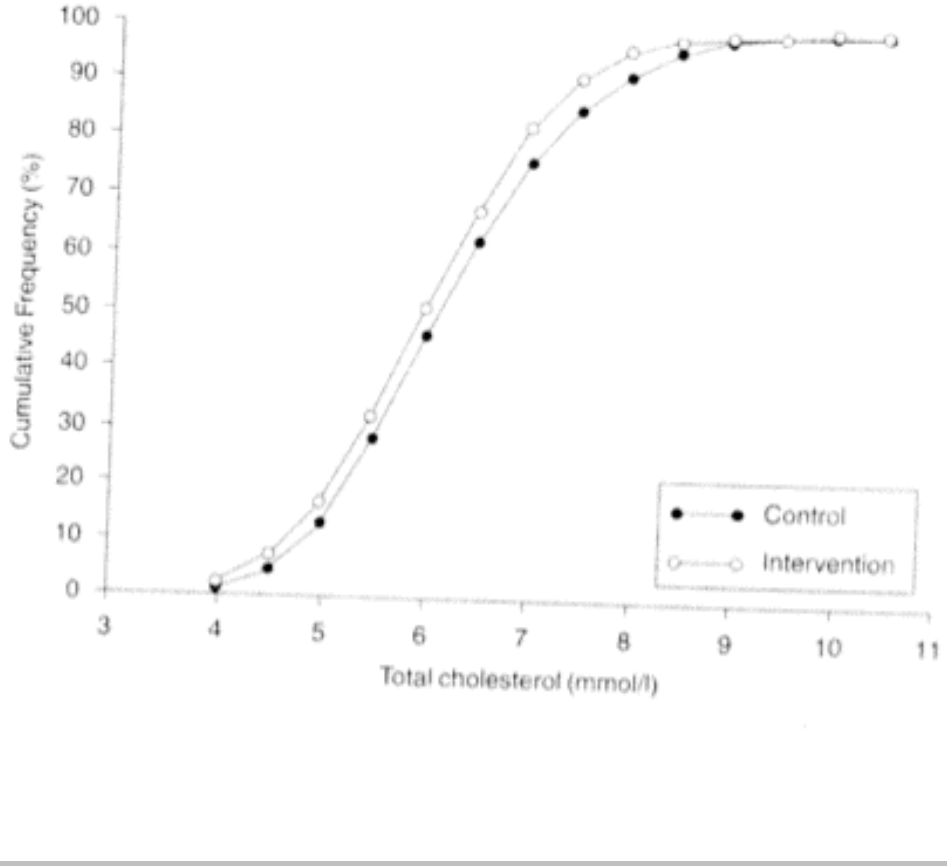
Answer:

Risk of death increases for both groups in relation to increasing Apache II score. Because the two groups have a different relationship between Apache II score and risk of death (as n be seen in the different rate of death for the two groups at each Apache II score category), the Apache II scores do not estimate the same actual risk of death for non-operative and postoperative patients.

7b. (10 points)

The figure below is a comparison of two cumulative frequency polygons. It is a summary of data on total cholesterol (mmol/L) in a randomized controlled trial that investigated the effectiveness of health checks (intervention group) relative to standard care (control group).

In 1-2 sentences, comment on the two total cholesterol distributions. Speculate on the comparison of the two groups with respect to the benefit of intervention.



Answer:

At each cholesterol measurement the intervention group has a higher cumulative frequency than the control group. This indicates that the intervention is associated with *lower* cholesterol measurements. However, we don't know if this is significant (not yet anyway!)