BIOSTATS 540 - Introductory Biostatistics
Fall 2022

Exam 1
Unit 1 – Summarizing Data &
Unit 2 – Data Visualization

Due: Monday October 10, 2022

Last Date for Submission with -10 points: Wednesday October 12, 2022
Last Date for Submission with Credit (-20 points): Monday October 17, 2022

This is an "open book" "take-home" exam. You are welcome to use any reference materials you wish. You are welcome to use the computer as you wish, too. However, you MUST work this exam by yourself and you may not consult with anyone (except me!).

How to Submit Your Exam

Please upload a ***SINGLE PDF***
of your completed exam to the Blackboard Learn ASSIGNMENT tab
no later than **Monday October 10, 2022 (11:59 pm EDT or in your time zone, whichever is later).**

… *unless you are opting to submit your exam late per the late submissions policy above.*

BIOSTATS 540 - Introductory Biostatistics
Fall 2022

Exam 1
Unit 1 – Summarizing Data &
Unit 2 – Data Visualization

Due: Monday October 10, 2022 (11:59 pm EDT or in your time zone, whichever is later)

Last Date for Submission with -10 points: Wednesday October 12, 2022
Last Date for Submission with Credit (-20 points): Monday October 17, 2022

Signature

This is to confirm that in completing this exam, I worked independently and did not consult with anyone.

Signature: _____

Printed Name: _____

Date: _____

9/26/2022
Dear BIOSTATS 540 Fall 2022,

The points on this exam total 105

Thus, you can lose 5 points in your work and still score 100 on this test.
Please note, however: **The maximum score you can earn is 100.**

1. (10 points total)

1a. (1 point)
Give *two* examples of a variable that is **quantitative – discrete**.

1b. (1 point)
Give *two* examples of a variable that is **quantitative – continuous**.

The following table lists some variables that might be of interest in your next data analysis. For each variable, complete the associated table indicating whether it is categorical (and if so, is it nominal or ordinal) or quantitative (and if so, is it discrete or continuous).

		Categorical		Quantitative	
	Variable	nominal	ordinal	discrete	continuous
Example	*Eye color*	*X*			
1c (1 point)	2022 Annual income, dollars				
1d (1 point)	A list of the different professions within a profession				
1e (1 point)	The ranking of specialities with respect to income				
1f (1 point)	Staging of breast cancers as Type I, II, III or IV				
1g (1 point)	ST depression that is coded as follows: = 1 if ST depression is < 1 mm 2 if (1 mm ≤ ST depression < 5 mm) 3 if ST depression ≥ 5 mm				
1h (1 point)	ICD-11 Classifications as follows: 0295 = Organic psychosis 0296 = Depression etc.				
1i (1 point)	Diastolic blood pressure , mm Hg				
1j (1 point)	Self -reported pain reported on a ten-point scale .				

2. (10 points total)

2a. (3 points)

When a distribution is skewed to the right,

i) **True or False**. The median is greater than the mean.
ii) **True or False**. The distribution is uni-modal
iii) **True or False**. The majority of observations are less than the mean.

2b. (3 points)

The shape of a frequency distribution can be described using:

i) **True or False**. A box and whisker plot.
ii) **True or False**. A table of frequencies
iii) **True or False**. A histogram

2c. (4 points)

For the sample 3, 1, 7, 2 and 2:

i) **True or False**. The sample mean is 3
ii) **True or False**. The sample median is 7
iii) **True or False**. The range is 1
iv) **True or False**. The sample variance is 5.5

3. (10 points total)

3a. (2 points)
True or False. The sample mean is always one of the data points.

3b. (2 points)
True or False. The sample standard deviation can never be negative.

3c. (2 points)
True or False. When the sample size is odd, the median is always one of the data points.

3d (2 points)
True or False. The sample variance has the same units of measurement as the original observations.

3e. (2 points)
True or False. As the sample size is increased, the range of the data can only stay the same or get larger.

4. (10 points total)

Consumer Reports magazine reported the following data on the number of calories in a hot dog for each of a sample of 17 brands of meat hot dogs:

173	191	182	190	172	147	146	139	175
136	179	153	107	195	135	140	138	

4a. (3 points)

By any approach you like (by hand, Excel, Art of Stat, R, or something else is fine too!), calculate the **sample mean, sample variance,** and **sample standard deviation.** Show your work ("cut and paste" screen capture is fine).

4b. (3 points)

By any approach you like (by hand, Excel, Art of Stat, R, or something else is fine too!), calculate the **standard error** of the sample mean. Show your work ("cut and paste" screen capture is fine).

4c. (4 points)

The **"five-point summary**" is the term that is often used to refer to the following five statistics: minimum, P_{25}, P_{50}, P_{75} and the maximum. By any approach you like (by hand, Excel, Art of Stat, R, or something else is fine too!), calculate the five-point summary**.** Show your work ("cut and paste" screen capture is fine).

5. (10 points total)

5a. (2 points)

You read that the median income of U.S. households in 2010 was $49,455. In 1-2 sentences at most, explain in plain language what *"the median income"* is.

5b. (2 points)

The Census Bureau website gives several choices for ***"average income"*** in its historical income data. In 2010, the median income of American households was $49,455. The mean household income was $67,530. The median income of families was $60,395, and the mean family income was $78,361. The Census Bureau says, *"Households consist of all people who occupy a housing unit. The term 'family' refers to a group of two or more people related by birth, marriage, or adoption who reside together"*. In at most 5 sentences, explain carefully why mean incomes are higher than median incomes and why family incomes are higher than household incomes.

5c. (2 points)

A magazine article reported that the average income for readers of the business magazine *Forbes* was $217,000. In your opinion, is the median wealth of these readers greater or less than $217,000? In at most 1-2 sentences, explain your reasoning.

5d. (2 points)

The distribution of individual incomes in the United States is strongly skewed to the right. In 2008, the mean and median incomes of the top 1% of Americans were $558,726 and $1,137,680. Which of these numbers is the mean and which is the median? In at most 1-2 sentences, explain your reasoning.

5e. (2 points)

By any means you like (by hand is fine) which of the following two data sets is more spread out? Show your work. In at most 1-2 sentences, explain your reasoning.

Data set "A": 4 0 1 4 3 6
Data set "B": 5 3 1 3 4 2

6. (10 points total)

Consider the relationship between the standard deviation and the standard error. Suppose it is known that the standard deviation is 3. How large a sample n should be taken for the standard error of the mean to have a value of 0.5?

7. (10 points total)

7a. (2 points)

Choose one, **bar chart** or **histogram**: Which choice of graph would you use to summarize the distribution of the following variable: **Starting salary for the different specialties in a profession.**

7b. (2 points)

Choose one, **bar chart** or **histogram**: Which choice of graph would you use to summarize the distribution of the following variable: **Elapsed time since last visit to the dentist.**

7c. (2 points)

Choose one, **bar chart** or **histogram**: Which choice of graph would you use to summarize the distribution of the following variable: **Number of hair transplant sessions per person.**

7d. (2 points)

Choose one, **bar chart** or **histogram**: Which choice of graph would you use to summarize the distribution of the following variable: **Number of patients with 0, 1, or 2+ vessels with > 75% stenosis.**

7e. (2 points)

Choose one, **bar chart** or **histogram**: Which choice of graph would you use to summarize the distribution of the following variable: **Expected income at age 35.**

8. (10 points total)

Age is a variable that is important in most if not all epidemiological studies of health because our health is so strongly related to age. The following table is a two-dimensional table that summarizes the distribution of age in a certain population every five years starting with 1990 (the first row) to 2020 (the last row). Each row contains information for one year. Reading across the row (over the column), the table shows the number of individuals in each age group.

	Interval of Age, years				
Year	0-4	5-14	15-44	≥ 44	Total
1990	1,400	3,000	8,000	7,600	20,000
1995	2,700	5,000	12,000	10,300	30,000
2000	4,600	9,000	15,000	11,400	40,000
2005	6,000	11,000	16,500	11,500	45,000
2010	8,000	12,000	18,000	12,000	50,000
2015	10,000	13,500	19,000	12,500	55,000
2020	11,500	15,000	20,500	13,000	60,000

8a. (5 points)

Using the data provided, produce a new two-dimensional table that shows, for each year, the percentage of that year's total population in each age interval (Tip. In each row, the sum of your percentages should be 100%).

8b. (5 points)

Write a brief report (a paragraph or so, no more!) on what these summaries suggest regarding the possibility (or lack thereof) of changes over time in the age distribution of this population.

9. (10 points total)

A **box plot** is the graph of a **five number summary (minimum, P25, P50, P75, and maximum)**. The following table lists the average month's temperature (Farenheit) of Springfield, Massachusetts and San Francisco, California for a given year.

Month	Ave Temp (F) Springfield	Month	Ave Temp (F) San Francisco
January	32	January	49
February	36	February	52
March	45	March	53
April	56	April	55
May	65	May	58
June	73	June	61
July	78	July	62
August	77	August	63
September	70	September	64
October	58	October	61
November	45	November	55
December	36	December	49

9a. (5 points)

Obtain the five number summary for the average monthly temperatures, separately for each data set, Springfield versus San Francisco. Use these values to complete the following table.

	Springfield	San Francisco
Minimum		
Q1		
Q2 = median		
Q3		
Maximum		

9b. (5 points)

By any means you like (by hand, Excel, Art of Stat, R, or something else is fine too!), produce a side-by-side box and whisker plot of the two distributions of average monthly temperatures.

10. (15 points total)

The distribution of the ages of a nation's population has a strong influence on economic and social conditions. The following table sows the age distribution of U.S. residents in 1950 (data obtained by census) and the projections for 2050 (projections made by the Census Bureau).

Age Distribution in the United States (in millions of persons)

Age group, years	1950	2050 (projected)
Under 10	29.3	56.2
10-19	21.8	56.7
20-29	24.0	56.2
30-39	22.8	55.9
40-49	19.3	52.8
50-59	15.5	49.1
60-69	11.0	45.0
70-79	5.5	34.5
80-89	1.6	23.7
90-99	0.1	8.1
100 and over	--	0.6
Total	**150.9**	**438.8**

10a. (5 points)

Because the projected total population in 2050 is so different (larger!) than the 1950 population, comparing percentages in each age group is clearer than comparing counts in each group. Make a table of the percentages in each age group, separately for 1950 and 2050.

Age	1950 #	1950 %	2050 #	2050 %
Under 10	29.3		56.2	
10-19	21.8		56.7	
20-29	24.0		56.2	
30-39	22.8		55.9	
40-49	19.3		52.8	
50-59	15.5		49.1	
60-69	11.0		45.0	
70-79	5.5		34.5	
80-89	1.6		23.7	
90-99	0.1		8.1	
100 and over	0.0		0.6	
Totals =	**150.9**		**438.8**	

10b. (5 points)

By any means you like (including by hand!) Construct a histogram of the 1950 age distribution, in percents. Construct a histogram of the projected 2050 age distribution, in percents.

10c. (2 points)

In 1-2 sentences at most, describe the main features of the histograms you con constructed in question #10b.

10d. (3 points)

In 1-2 sentences at most, what are the most important changes in the U.S. age distribution projected for the 100-year period between 1950 and 2050?