Assignment I

Descriptive Statistics

Data Presentation   (Unit 2)
Numerical Summary Measures   (Unit 3)

For all problems you should use appropriate notation and show your work.

Use your calculator when needed (not the computer) for problems 1 through 6.

1. How do ordinal data differ from nominal data?

2. When you construct a table, when might it be beneficial to use relative rather than absolute frequencies?

3. Classify each of the measurements listed here as one of the following:
   nominal; binary; ordinal; discrete (count); or continuous.

   (a) Population density (people per acre)

   (b) Grade in a course coded: A, B, C, D, or F

   (c) Treatment group: 1 = active treatment, 2 = placebo

   (d) Head circumference (cm)

   (e) Course credit (pass or fail)

   (f) Response to treatment coded as 1 = no response, 2 = minor improvement,
   3 = major improvement, 4 = complete recovery

   (g) Number of hospitalizations in one year for an individual

   (h) Country of birth

4. A study was conducted investigating the long-term prognosis of children who have suffered an acute episode of bacterial meningitis, an inflammation of the membranes enclosing the brain and spinal cord. Listed below are the times to the onset of seizure for 13 children who took part in the study. In months, the measurements are:

   $$0.10, \ 0.25, \ 0.50, \ 4, \ 8, \ 11, \ 20, \ 26, \ 30, \ 38, \ 42, \ 57, \ 96$$

   Find the following numerical summary measures of the data:

   (a) mean

   (b) median

   (c) range

   (d) interquartile range

5. The following data lists the skinfold measurements (in mm) at the midpoint of the triceps in five men with chronic lung disease and six comparably aged controls.

   Chronic lung disease: 9.2, 10.9, 11.4, 15.1, 18.1

   Controls: 10.7, 19.8, 20.5, 23.6, 24.8, 32.5

   (a) Find the mean, median, standard deviation, and range for the men in the chronic lung disease group.

   (b) For the chronic lung disease data, verify that $\sum_{i=1}^{5}(x_i - \bar{x})$ is equal to zero.

   (c) Find the mean, median, standard deviation, and range for the men in the control group.

6. The total number of deaths in the united States in various years are presented below.

Year	Number of Deaths
2020	3,383,729
2015	2,712,630
2010	2,468,435
2000	2,403,351
1990	2,148,463
1980	1,989,841
1970	1,921,031
1960	1,711,982
1950	1,452,454
1940	1,417,269

   The statement is made that since the number of deaths has been increasing over the years, the population as a whole must be growing less healthy. Do you agree with this statement? Why or why not?

🖥 Problems involving the computer:

7. The 2022 obesity rates for youth ages 10 to 17 for each state in the United States are contained in the dataset `Obesity_10-17yo_2022` on ICON. The rates are per 100 children in this age range.

   (a) Which state has the lowest rate of obese youth? Which state has the highest rate of obese youth? What factors might influence the variability among different states?

   (b) Construct a box plot for the rate of obese youth (per 100 children aged 10 to 17).

   (c) Are the observations symmetric or skewed? Are there any states that could be considered outliers?

   (d) Display the rate of obese youth using a frequency histogram. Do you find this graph to be more or less informative than the box plot?

   (e) The average of the 51 state rates is 16.0 obese children per 100. Is it sensible to calculate the mean of these values this way? Why or why not?

8. A study is planned to be conducted in Mexico and the United States. The frequency distibution of population age is displayed below for each country.

Age group (years)	Mexico ('000s)	United States ('000s)
0–14	33,461.0	61,408.9
15–24	21,832.0	42,937.8
25–54	52,827.9	129,458.2
55–64	10,663.2	42,782.5
65+	9,865.4	56,051.6
Total	128,649.5	332,639.0

(a) Is it reasonable to compare the distributions of age for Mexico and the U.S. based on the absolute frequencies in each interval? Why or why not?

(b) Compute the relative age frequencies for each country.

(c) Construct a pair of (overlayed) frequency polygons.

(d) Describe the shape of each polygon. What can you say about the age distribution of each country?

9. The data set for this problem and the next is from the Outcomes for Children with Hearing Loss (OCHL) study. You can download the dataset named `OCHL4120.csv`. There you will also find a document that describes the data elements.

For each of the following variables, determine the data type. If the variable is categorical, specify if it is nominal or ordinal. If the data type is continuous, specify the unit of measurement and the range of values the data may take on.

(a) Location

(b) Sex

(c) Age aid fitted

(d) Income Range

(e) Better ear PTA

(f) Mother's education

10. For each of the following variables, create an appropriate graphical summary of its distribution, (i.e., bar chart or histogram or box plot) with an informative title and informative axis labels. For data summarized with histogram/boxplots indicate if the data are skewed or symmetric and comment on any outliers. You do not need to submit your code, only the figures.

(a) Sex

(b) Age aid fitted

(c) Income Range

(d) Better ear PTA

(e) Mother's education

Due Date: 9:30 a.m., Tuesday, September 9, 2025.

Upload a **single pdf file** of your work to ICON before the start of class.