1 Lecture 2 Notes

Important Distinction:

1. **Descriptive Statistics**: The objective is to summarize or describe the data

2. **Inferential Statistics**:The objective is to make inference of the population from the sample

Summarizing the Data

1. **Frequency**: - the number or count a number appears

2. **Frequency Distribution**: - shows how data is broken up into classes (bins) and number the number of occurrences that appear within each bin based on data

Example 1: Frequency distribution of cotinine (a metabolite of nicotine) level of smokers. A sample of 40 smokers and their cotinine level) in ng/ml (1st edition)

1	0	131	173	265	210	44	277	32	3
35	112	477	289	227	103	222	149	313	491
130	234	164	198	17	253	87	121	266	290
120	167	250	245	48	86	284	1	208	173

Procedure for Constructing a Frequency Distribution

1. Select number of bins (between 5-20), lets choose 5

2. Calculate Width:

$$\text{Class Width} = \frac{Max\text{–}Min}{\text{\# of bins}} = \frac{491\text{–}0}{5} = 98.2 \approx 100$$

   Round up to make life easier.

3. **Find the Lower limits (LL)** for each bin. Choose the lowest number in the data set and add the Class Width

4. **Find Upper limit (UL)** Use the Lower Limit of the next bin to find the UL

5. Make a list of the LL and UL, as follows:

6. Go through the data and determine the occurrences within each bin:

7. Determine Relative Frequency

8. Determine Cumulative Frequency

LL	UL	Frequency	Relative Frequency	Cumulative Frequency
0	99	11	11/40=0.275	11
100	199	12	12/40=0.3	23
200	299	14	14/40=0.35	37
300	399	1	1/40=0.025	38
400	499	2	2/40=0.05	40

Types of Plots (purposes)

1. Histograms – visually displays the shape of the distribution of the data, shows location of the center, spread, if there are outliers (i.e., gas prices)

2. Frequency Polygons – uses line segments connected to points located directly above class midpont values for each bin (i.e., IOP)

$$\text{Mid Point} = \frac{UL - LL}{2}$$

3. Bar Graphs & Bar Plot - used of equal width to show frequencies of categories (i.e., Political Party)

4. Pareto Charts - bar graph for categorical data, bars are arranged in descending order per frequencies, decrease left to right (i.e., Favorite Ice Cream)

5. Scatter Plots - shows the relationship between two variables (i.e., study hours vs. gpa)

6. Time Serie Plots - data collected at different time points (i.e., weather, finances, blood pressure)

7. Others: Dot Plots, Stem-and-Leaf Plots, and Pie Charts

Central Tendency (Measures of the Center) New Notation
N : Population Size
n : Sample Size
x_i i^{th} observation within population/sample
$\sum$:

1. Mean - the central or typical value in a set of data (vulnerable to outliers)

$$\mu = \sum_{i=1}^{N} = \frac{x_i}{N}$$

$$\bar{x} = \sum_{i=1}^{n} = \frac{x_i}{n}$$

2. Median - Is the middle value of the original data values when they are arranged in increasing order.

   - - Case n is odd: The median is exactly the center value

   - - Case n is even: The median is the average of the two middle values

3. Mode - Value that occurs most frequently

4. Midrange – maximum value plus minimum value divided by two

$$\text{Midrange} = \frac{\text{Max} + \text{Min}}{2}$$

5. Weighted Mean - Each value has a different level of importance:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Problem 1 5.40,1.10, 0.42, 0.73, 0.48, 1.10

1. Mean 1.538
2. Median 0.915
3. Mode 1.10

Problem 2 27, 27, 27, 55, 55, 55, 88, 88, 99

1. Mean 57.89

2. Median 55
3. Mode 27, 55

Problem 3 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

1. Mean 5.5
2. Median 5.5
3. Mode NA

Skewness

Left Skew: Mean < Median

Right Skew: Mean > Median