Handout 1: Descriptive Statistics

Frequency Distribution

1. List all values of the variable (sorted from smallest to largest if applicable)
2. For each value of the variable, indicate how many cases have this value
3. Calculate % of cases for each value by dividing its number of cases by the total

Cumulative Frequency Distribution

1. Construct the frequency distribution
2. For each value, add percentages for that value plus for all the values smaller than it

Measures of Central Tendency

A. Mean:

Mean = sum of all numbers / how many numbers; $\bar{X} = (\sum X)/n$

To make this calculation easier when the list of numbers is long, use the frequency distribution:

1. Construct a frequency distribution
2. Multiply each value by its frequency
3. Add those products up and divide by the total number of cases

B. Median

Median is the number that splits the sorted list of values into two lists of equal sizes: half of the numbers are above it, half are below.

1. Sort the list from the smallest numbers to the largest (list duplicates as many times as they occur).
2. Divide the list in half.
3. If the number of values is odd→ the number in the middle is the median. If the number of values is even → the median is the average between the two values in the middle.

Or using cumulative distribution, find the value where the percentage first becomes 50% or larger.

C. Mode

It is the most frequently occurring value in the list (the value with the highest percentage).

Measures of Variation

A. Variation Ratio.

Rarely used, only for nominal variables. Represents the percentage of cases NOT in modal category.
To calculate: subtract the percentage of cases in modal category from 100%.

B. Range

Describes the spread of the distribution – from the smallest to largest value.
To calculate: subtract the smallest value from the largest.

C. Interquartile Range (IQR)

$IQR = Q_3 - Q_1$

1. Sort the distribution (smallest to largest; list duplicates as many times as they occur).
2. Divide in half (that's Q2, the median); then divide each half again -- that's Q1 and Q3.
3. Subtract Q1 from Q3

D. Standard Deviation

1. Calculate distance, or deviation, between every number in the list and the mean of the list.
   Deviation of a number = the number minus the mean.
2. Square all the deviations
3. Add up all the squared deviations and divide by (number of cases minus one)
4. Take square root of the result.

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

E. Variance

Variance = squared standard deviation

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Example:

Variable "child's age" is measured for 10 children: 3, 1, 2, 3, 6, 7, 6, 2, 8, 2.

Frequency Distribution and Cumulative Frequency Distribution

Step 1: values are 1, 2, 3, and 6, 7, 8 years old
Step 2: 1 – 1 case, 2 – 3 cases, 3 – 2 cases, 6 – 2 cases, 7 - 1 case, 8 – 1 case.
Step 3:

Value of X	Frequency	Percent of cases	Cumulative Percent
1	1	10%	10% [10%]
2	3	30%	40% [10%+30%]
3	2	20%	60% [10%+30%+20%]
6	2	20%	80% [10%+30%+20%+20%]
7	1	10%	90% [10%+30%+20%+20%+10%]
8	1	10%	100% [10%+30%+20%+20%+10%+10%]

Central Tendency:

Mean:
3+1+2+3+6+7+6+2+8+2 = 40/10 = 4
Or using frequency distribution, create products of value and frequency:

Value of X	Frequency	Percent of cases	Value * Frequency
1	1	10%	1
2	3	30%	6
3	2	20%	6
6	2	20%	12
7	1	10%	7
8	1	10%	8

Sum all these products and divide by total number of cases:
1*1 + 2*3 + 3*2 + 6*2 + 7*1 + 8*1 = 1 + 6 + 6 + 12 + 7 + 8 = 40/10 = 4

Median:
Sorted list split in half: 1, 2, 2, 2, 3 | 3, 6, 6, 7, 8. Median=3
Or using the cumulative frequency distribution, the 50% threshold is crossed at 3 (where percentage goes up to 60%), so median is 3.

Mode:
In the frequency distribution, value 2 has the highest percentage (30%). Mode=2.

Variation:

Variation Ratio:
3 cases (30%) in modal category (2), the others 7 (70%) in other categories. Variation ratio: 100%-30%=70%,

Range:
Highest value – lowest value = 8-1 = 7

Interquartile Range:
Divide the list in half, then again in half:
1, 2, 2, 2, 3 | 3, 6, 6, 7, 8
Q1=2 and Q3=6.
Or using the cumulative frequency distribution, the 25% threshold is crossed at 2 (where percentage goes up to 40%), and 75% threshold is crossed at 6 (where percentage goes to 80%), so Q1=2 and Q3=6.
IQR = Q3 - Q1 = 6 - 2 = 4

Standard Deviation and Variance:
We know the mean = 4. Subtract the mean from each value, then square these deviations:

X	$X - \bar{X}$	$(X - \bar{X})^2$
3	-1	1
1	-3	9
2	-2	4
3	-1	1
6	2	4
7	3	9
6	2	4
2	-2	4
8	4	16
2	-2	4
sum	0	56

Or do the same using frequency distribution:

Value of X	Frequency	$X - \bar{X}$	$(X - \bar{X})^2$	$(X - \bar{X})^2 *$ Frequency
1	1	-3	9	9
2	3	-2	4	12
3	2	-1	1	2
6	2	2	4	8
7	1	3	9	9
8	1	4	16	16
sum				56

Variance = 56/9 = 6.22
Standard deviation = sqrt(Variance) = sqrt(6.22) = 2.49