Lec 4

Biostatistics

Data summarization

Data summarization: Method by which the people organize, summarize and communicate information using a variety of tools such as tables, graphs and diagrams.

Uses of data presentation

- ❑ Easy and better understanding of the subject
- ❑ Provides first-hand information about data
- ❑ Helpful in future analysis
- ❑ Easy for making comparisons
- ❑ Very attractive

Tabular presentation of data

It is important to know how to present data in meaningful tables that are easy to understand**.**

Nominal Variables

• Nominal variables: Frequency

We can present them as frequencies, the number of individuals in each category.

For example, the nationalities of participants:

Nominal Variables:

Frequency

Nationality	**Frequency**(n=180)
Bahraini	22
Egyptian	42
Iraqi	36
Lebanese	17
Qatari	8
Saudi	55

Here, the categories are arranged alphabetically, but as they don't have an order, it may be more comfortable for the reader to arrange them according to the frequencies.

We start with the nationality with the highest frequency to the lowest as follows:

Nationality	**Frequency** (n= 180)
Saudi	55
Egyptian	42
Iraqi	36
Bahraini	22
Lebanese	17
Qatari	8

Nominal Variables: Relative frequency:

Although reporting of frequencies is easy to understand, reporting the percentages (relative frequencies) is more comfortable for most people to get a sense of the data.

It is calculated easily by dividing the number of individuals in each category and dividing it by the total number. **Then we multiply it by 100 to get the percentage as follows:**

Nationality	Frequency (n= 180)	Relative frequency	How to calculate?
Saudi	55	**30.6**	55÷180 *100
Egyptian	42	**32.3**	42÷180 *100
Iraqi	36	**20.0**	36÷180 *100
Bahraini	22	**12.2**	22÷180 *100
Lebanese	17	**9.4**	17÷180 *100
Qatari	8	**4.4**	8÷180 *100

Ordinal Variables: Frequency

Satisfaction level	Frequency (n= 140)
Very satisfied	43
Satisfied	55
Neutral	15
Dissatisfied	19
Very dissatisfied	8

Satisfied	55
Ver…ied	43
Dissatisf…	19
Neu…	15
Very …issatisfied	8
Satisfied	55

Ordinal Variables: The same as nominal variables, percentages (relative frequencies) are calculated and presented as follows:

Relative frequency

Satisfaction level	Frequency (n= 140)	Relative frequency	How to calculate?
Very satisfied	43	30.7	=43÷140X 100
Satisfied	55	39.3	=55÷140X 100
Neutral	15	10.7	=15÷140X 100
Dissatisfied	19	13.6	=19÷140X 100
Very dissatisfied	8	5.7	=81÷40X 100

Ordinal Variables: Cumulative relative frequency**:** Sometimes we use the cumulative relative frequency to present the ordinal variables making benefit from the order. They are presented and calculated as follows:

Satisfaction level	Frequency (n= 140)	Relative frequency	Cumulative relative frequency	How to calculate?
Very satisfied	43	30.7	30.7	30.7
Satisfied	55	39.3	70.0	30.7+39.3=70
Neutral	15	10.7	80.7	70.0+10.7=80.7
Dissatisfied	19	13.6	94.3	80.7+13.6=94.3
Very dissatisfied	8	5.7	100.0	94.3+5.7=100

The cumulative relative frequency at one level is calculated simply by adding the relative frequency at this level to all relative frequencies before this level.

For example, if the cumulative relative frequency at the "satisfied" level is 70%, this means that 70% of the individuals are either satisfied or very satisfied. While the cumulative relative frequency at the "neutral" level is 80.7% meaning that 80.7% of the participants are very satisfied, satisfied, or neutral

Numerical Discrete Variables:

- **Numerical Discrete Variables: Frequency, relative frequency, and cumulative relative frequency**

If the numerical discrete variable is of few levels, we can represent it in frequencies, relative frequencies, and cumulative relative frequencies in the same way as in ordinal variables.

For example, the number of kids in the family:

Frequency

Number of kids	Frequency (n= 240)
0	32
1	64
2	83
3	42
4	13
5	6

Numerical Discrete Variables: Relative frequency

Number of kids	Frequency (n= 240)	Relative frequency	How to calculate?
0	32	13.3	=32÷240X 100
1	64	26.7	=64÷240X 100
2	83	34.6	=83÷240X 100
3	42	17.5	=42÷240X 100
4	13	5.4	=13÷240X 100
5	6	2.5	=6÷240X 100

Numerical Discrete Variables: Cumulative relative frequency

Number of kids	Frequency (n= 240)	Relative frequency	Cumulative relative frequency	How to calculate?
0	32	13.3	13.3	13.3
1	64	26.7	40.0	13.3+26.7=40
2	83	34.6	74.6	40.0+34.6=74.6
3	42	17.5	92.1	74.6+17.5=92.1
4	13	5.4	97.5	92.1+5.4=97.5
5	6	2.5	100.0	97.5+2.5

Here, for example, 74.6% of the families have two kids or less (2, 1, or 0).

• Numerical Continuous Variables: Frequency, relative frequency, and cumulative relative frequency

If we are dealing with a continuous variable as the birth weight in grams, it is impractical and useless to present the frequencies for each birth weight we observe in grams.

Instead, we can group the variable into groups of equal width: (2000-2499, 2500-2999, 3000-3499, 3500-3999, and 4000-4500).

For those groups, we can present the frequency, relative frequency, and cumulative relative frequency as we did before

Numerical Continuous Variables: relative frequency:

Birth weight(g)	Frequency (n= 45)	Relative frequency	How to calculate?
2000-2499	3	6.7	=345X 100
2500-2999	13	28.9	=1345X 100
3000-3499	18	40.0	=1845X 100
3500-3999	7	15.6	=745X 100
4000-4499	4	8.9	=445X 100

Numerical Continuous Variables:

Cumulative relative frequency

Birth weight(g)	Frequency (n= 45)	Relative frequency	Cumulative relative frequency	
2000-2499	3	6.7	6.7	6.7
2500-2999	13	28.9	35.6	6.7+28.9=35.6
3000-3499	18	40.0	75.6	35.6+40.0=75.6
3500-3999	7	15.6	91.1	75.6+15.6=91.1
4000-4499	4	8.9	100.0	91.1+8.9=100

Sometimes, instead of having some groups with very few frequencies at the lower or the upper end, we group them into one group less than a specific value, or one group that is higher than a specific value and call them “open-ended groups ”as in the following table representing the age of patients:

Open ended groups

Age of patient	Frequency (n= 120)
≤19	5
20-24	42
25-29	36
30-34	30
≥ 35	7

Two Categorical Variables

• Cross- tabulation: two-way table

Sometimes we are interested in presenting two categorical variables in the same table, which we call the two-way table (as we have two variables). A table representing the relationship between sex and the disease status can be as flows:

		Sex		
		Male	Female	total
Disease	Diseased	24	18	42
	Not diseased	41	35	76
	total	65	53	118

From this table we can get the following information:

- Total number of participants: 118 (cell in the right lower corner)
- Total number of males: 65 (lower margin)
- Total number of females: 53 (lower margin)
- Total number of diseased: 42 (right margin)
- Total number of not diseased: 76 (right margin)
- Males and diseased: 24
- Females and diseased: 18
- Males and not diseased: 41
- Females and not diseased: 35

We can even make the table more informative by adding percentages by rows or columns.

Adding percentages by rows gives us the following table:

		Sex		
		Male	Female	total
Disease	Diseased	24	18	42
		57%	43%	100%
	Not diseased	41	35	76
		54%	46%	100%
	total	65	53	118
		55%	45%	100%

From the percentages presented in the table we can see that:

• The total percentage of males is 55% while that of females is 45% (last row)

• The percentage of males among diseased is 57% while that of females is 43%.

• The percentage of males among not diseased is 54% while that of females is 46%.

Adding percentages by columns gives us the following table:

		Sex		
		Male	Female	total
Disease	Diseased	24	18	42
		37%	34%	36%
	Not diseased	41	35	76
		63%	64%	63%
total		65	53	118
	total	100%	100%	100%

From the percentages presented in the table we can see that:

• The total percentage of diseased is 36% while that of not diseased is 64% (last column).

• The percentage of diseased among males is 37% while that of not diseased is 63%.

• The percentage of diseased among females is 34% while that of not diseased is 66%.

Three Categorical Variables

• Cross- tabulation: Three-way table

		sex	
		male	female
Smoker	Diseased	36	42
	Not diseased	22	18
Non smoker	Diseased	24	18
	Not diseased	41	35

Three categorical variables can be presented in the same table such as sex, disease status, and smoking status as follows:

In this table the three variables are presented, we can add more numbers as total numbers and percentages, but we prefer to keep it simple. The arrangement of the variables can be also changed. It all depends on what information we want to tell the reader