Chpt 2.2

Frequency Distributions and Graphs

2-2 Data Collection and Sampling Techniques.

Chpt 2.2

Homework

- 2-2 Read pages 34-43
  - p43 Applying the Concepts
  - p43 1-7, 12, 13

Chpt 2.2

Objective

2.2 Organizing data into frequency distributions

Organizing Data

Objective: Organizing data into frequency distributions

- Once you have collected a bunch of **raw data**, what should you do with it?
- The first step in statistics is to draw a picture of the data. To do that the data must be organized in some fashion.
  - The most basic organization is a **frequency distribution**.
  - An **ungrouped frequency distribution** is simply each score listed in a table next to the **frequency** of the occurrence of that datum value.

Organizing Data

Objective: Organizing data into frequency distributions

- **Ungrouped Frequency Distributions** - used for data that can be enumerated and when the range of values in the data set is not large.
  - number of miles teachers drive from home to school
  - number of girls in a 4 child family
  - number of pets owned by student families, etc.

Organizing Data

Objective: Organizing data into frequency distributions

If we record the number of hours spent watching TV last night we would not get many results. An ungrouped frequency distribution might look like:

hours	tally	Frequency (f)
0	𝍸 | | | |	9
1	𝍸 𝍸 |	11
2	| | | |	4
3		0
4	| |	2

Organizing Data

Objective: Organizing data into frequency distributions

If our variables are categorical and the data is nominal, or ordinal, the distribution is called, appropriately, a **Categorical Frequency Distribution**.

If we list subjects by personality type we might see:

Personality	tally	Frequency (f)
A	𝍸 𝍸 𝍸 | | |	17
B	𝍸 𝍸 | |	11
A/B	𝍸 𝍸 𝍸 𝍸 | |	22

Organizing Data

Objective: Organizing data into frequency distributions

- Twenty-five army inductees were given a blood test to determine blood type.

Raw Data: ~~A, B, B, AB~~, O
O, O, B, AB, B
B, B, O, A, O
A, O, O, O, AB
AB, A, O, B, A

Type	tally	Frequency (f)
O	IIIIIIIII	9
A	IIIII	5
B	IIIIIII	7
AB	IIII	4

- The tally marks are not a necessary part of the frequency table. They are only to help with accuracy.

Organizing Data

Objective: Organizing data into frequency distributions

- With just a relatively low number of possible data values we can list all the individual data values, but with a large number of data values it becomes unwieldy and overwhelming to list every value.

- A **Grouped Frequency Distribution** is used when the range of data values in the data set is large. The data is grouped into **classes that are more than a single unit in width.**
  - age of spectators at a 49ers game
  - cranking power of a car battery in ampere/hours
  - number of robbery victims in large U.S. cities
  - etc.

- For a **Grouped Frequency Distribution** we create intervals that capture a range of values bounded by class limits or class boundaries.

Organizing Data

Objective: Organizing data into frequency distributions

- Class **limits** represent the largest and smallest data values that an be included in the class.
  - Class limits **are** actual data values.

- Class **boundaries** provide values that eliminate gaps between the classes in the frequency distribution. Class boundaries are one decimal place more accurate than the data.
  - Class boundaries are **not actual data values**.
  - For this course, class boundaries will always end with a **5**.

    - To find a class **boundary**, average the upper class **limit** of one class and the lower class **limit** of the next class.

Objective: Organizing data into frequency distributions

Organizing Data

- The **class limits** define the class and are actual data values.
- The **class boundaries** are ½ an interval above and below the class limits.
  - **Class limits** are possible data values.
  - **Class boundaries** are **not** possible data values but define the intervals within which the data values must fall.

Organizing Data

Objective: Organizing data into frequency distributions

Age groups are often used as classes. The **class limits** are listed. The values 0 and 4 are the **limits** of the data in that class, 5 and 9 are the limits of the next interval (class), etc.

Age	Tally	Frequency
0 - 4		
5 - 9		
10 - 14		
15 - 19		
20 - 24		
25 - 29		

Objective: Organizing data into frequency distributions

Age	Tally	Frequency
0 - 4		
5 - 9		
10 - 14		
15 - 19		
20 - 24		
25 - 29		

- In the age class 0 – 4, the **class limits** 0 and 4 are actual data values.

- Age 4 represents the ages 3.5 to 4.5, including 3.5 and not including 4.5. (i.e. [3.5, 4.5)).
  - Thus the **class boundaries** of 0 – 4 are actually -0.5 to 4.5.
  - The **class boundaries** for the class 5 – 9 are 4.5 and 9.5.

- Relax … the -0.5 is a boundary and cannot be a data value.

Organizing Data

Objective: Organizing data into frequency distributions

Now our Grouped Frequency Distribution might look like:

Age - limits	Age - Boundaries	Tally	Frequency
0 - 4	-0.5 - 4.5		
5 - 9	4.5 - 9.5		
10 - 14	9.5 - 14.5		
15 - 19	14.5 - 19.5		
20 - 24	19.5 - 24.5		
25 - 29	24.5 - 29.5		

Organizing Data

Objective: Organizing data into frequency distributions

- Remember class limits are values the variable might actually take.
- The class boundaries are typically one decimal place more accurate (average upper and lower limits of adjacent classes) and are not actual data values. The class boundaries are the result of the interval each datum represents.
  - The class width can be calculated by subtracting
    - successive lower class limits (or boundaries)
    - successive upper class limits (or boundaries)
    - upper and lower class boundaries.

Age - limits	Age - Boundaries	Tally	Frequency
0 - 4	-0.5 - 4.5		
5 - 9	4.5 - 9.5		
10 - 14	9.5 - 14.5		
15 - 19	14.5 - 19.5		
20 - 24	19.5 - 24.5		
25 - 29	24.5 - 29.5		

- The class midpoint X_m is calculated by averaging upper and lower class limits (or boundaries).

$$\frac{\text{lower limit} + \text{upper limit}}{2} = \frac{10+14}{2} = 12$$

Organizing Data

Age - limits	Age - Boundaries	Tally	Frequency
0 - 4	-0.5 - 4.5		
5 - 9	4.5 - 9.5		
10 - 14	9.5 - 14.5		
15 - 19	14.5 - 19.5		
20 - 24	19.5 - 24.5		
25 - 29	24.5 - 29.5		

Objective: Organizing data into frequency distributions

Rules for setting class width

1. No overlapping intervals (classes). The classes must be mutually exclusive and exhaustive. One datum cannot belong in two classes and all data is included.
2. Class width must be consistent (with one exception).
   Class width = upper **limit** – lower **limit** + 1 interval (14 - 10 + 1 = 5)
   Class width = upper **boundary** – lower **boundary**. (14.5 - 10.5 = 5)
3. The class width should be small enough to accurately portray the data but large enough to keep the number of classes manageable.

Readers of the data cannot assimilate too many classes, the forest gets lost in the trees. The number of classes is best kept between **5** and **15**. Fewer than **5** and the trends get lost, more than **15** and the information becomes lost. This rule is not cast in stone. I try to limit to between **7** and **12** depending on the data.

Organizing Data

Objective: Organizing data into frequency distributions

Rules for setting class width

4. The whole idea of a frequency distribution is to provide an accurate and easily understood picture of the data as simply as possible.

5. If the class width is an odd number the midpoint of the class is an actual datum value. That can be useful for graphical representations of the data. Not a necessary condition but often useful because we like pretty pictures of our data.

Organizing Data

Objective: Organizing data into frequency distributions

Procedure for constructing a grouped frequency

1. Find the highest and lowest values in your data.
2. Find the range of the data.
3. Select the number of classes desired (7-12).
4. Find the width of each interval by dividing the range by the number of classes and rounding up.
5. Select a starting point (usually the lowest value); add the width to get the lower limits.
6. Determine the upper class limits.
7. Determine the boundaries
8. Tally and record the frequencies in each class.

Example

In a survey of 20 patients who smoked, the following data were obtained. Each value represents the number of cigarettes the patient smoked per day. Construct a frequency distribution using **six** classes.

10	8	6	14
22	13	17	19
11	9	18	14
13	12	15	15
5	11	16	11

10	8	6	14
22	13	17	19
11	9	18	14
13	12	15	15
5	11	16	11

Organizing Data

Objective: Organizing data into frequency distributions

Step 1: Find the highest and lowest values: High = 22 and Low = 5.

Step 2: Find the range: R = 22 - 5 = 17.

Step 3: Select the number of classes desired. In this case we are told to create 6 classes.

Step 4: Find the class width by dividing the range by the number of classes.
Width = 17/6 = 2.83. This value is rounded up to 3.

Step 5: Select the starting point, we will start with the lowest value 5.

10	8	6	14
22	13	17	19
11	9	18	14
13	12	15	15
5	11	16	11

Organizing Data

Objective: Organizing data into frequency distributions

Step 5: Select the starting point, we will start with the lowest value 5.
Starting at 5, the subsequent lower limits will be;

5,
5+3=8,
8+3=11,
11+3=14,
14+3=17,
and
17+3=20.

Limits	Boundaries	Tally	Frequency	Cumulative Frequency
5				
8				
11				
14				
17				
20				

10	8	6	14
22	13	17	19
11	9	18	14
13	12	15	15
5	11	16	11

Organizing Data

Objective: Organizing data into frequency distributions

Step 6: The upper class limits will be 7, 10, 13, 16, 19, and 22.
i.e. the upper limit for the first class is computed as 8 - 1, etc.

Note that there is no overlap in limits.

Limits	Boundaries	Tally	Frequency	Cumulative Frequency
5 - 7				
8 - 10				
11 - 13				
14 - 16				
17 - 19				
20 - 22				

10	8	6	14
22	13	17	19
11	9	18	14
13	12	15	15
5	11	16	11

Organizing Data

Objective: Organizing data into frequency distributions

Step 7: Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to the upper class limit. Note: the upper boundary of a class equals the lower boundary of the next class.

Limits	Boundaries	Tally	Frequency	Cumulative Frequency
5 - 7	4.5 - 7.5			
8 - 10	7.5 - 10.5			
11 - 13	10.5 - 13.5			
14 - 16	13.5 - 16.5			
17 - 19	16.5 - 19.5			
20 - 22	19.5 - 22.5			

10	8	6	14
22	13	17	19
11	9	18	14
13	12	15	15
5	11	16	11

Organizing Data

Objective: Organizing data into frequency distributions

Step 8: Tally the data, write the numerical frequency value for the tally in the frequency column.

Limits	Boundaries	Tally	Frequency	Cumulative Frequency
5 - 7	4.5 - 7.5	II	2	
8 - 10	7.5 - 10.5	III	3	
11 - 13	10.5 - 13.5	IIIIII	6	
14 - 16	13.5 - 16.5	IIIII	5	
17 - 19	16.5 - 19.5	III	3	
20 - 22	19.5 - 22.5	II	1	

10	8	6	14
22	13	17	19
11	9	18	14
13	12	15	15
5	11	16	11

Organizing Data

Objective: Organizing data into frequency distributions

Occasionally a researcher finds **cumulative frequency** useful. The cumulative frequency is simply the sum of all previous classes with the current class.

Limits	Boundaries	Tally	Frequency	Cumulative Frequency
5 - 7	4.5 - 7.5	II	2	2
8 - 10	7.5 - 10.5	III	3	5
11 - 13	10.5 - 13.5	IIIIII	6	11
14 - 16	13.5 - 16.5	IIIII	5	16
17 - 19	16.5 - 19.5	III	3	19
20 - 22	19.5 - 22.5	II	1	20

Organizing Data

Objective: Organizing data into frequency distributions

Additional Things to Consider

- Leave no gaps in your classes. A class with no data must be included.
- Boundaries overlap, limits do not overlap.
- Do NOT leave out any data. If you have data far away from most (outliers) you must include that data as well.
- Which leads to Open-ended classes. Instead of 0 – 4, < 4. Or instead of 60 – 64, > 30.
  - Use open ended classes sparingly, only when unavoidable.
    - We will not be using open ended classes.