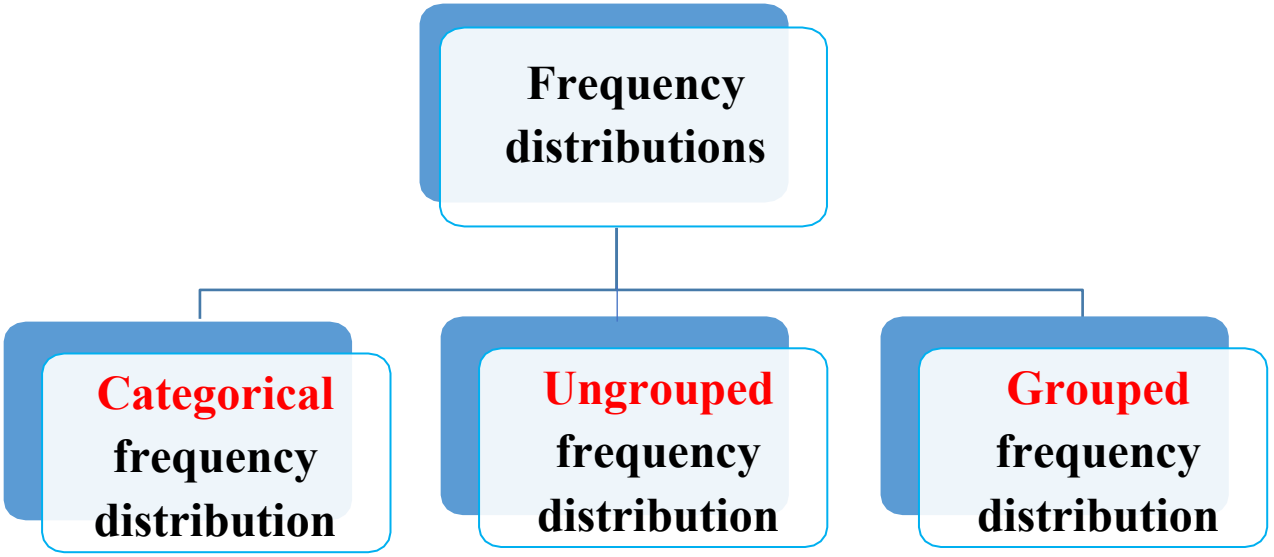
Probability and Statistics (PHM111s)-Lecture 2

2. Organizing Data.



Example 1: Twenty-five army inductees were given a blood test to determine their blood type. The data set is

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

Construct a frequency distribution for the data.

Solution

Since the data are categorical, discrete classes can be used. There are four blood types: A, B, O, and AB. These types will be used as the classes for the distribution.

The frequency distribution for the data is:

A Class	B Tally	C Frequency	D Relative Frequency (Percent)
A	𝍸	5	20
B	𝍸 //	7	28
O	𝍸 ////	9	36
AB	////	4	16
	Total	25	100

For the sample, more people have type O blood than any other type.

Example 2: Find the frequency distribution table for 20 families who own certain number of pets. The data set is: 3, 0, 1, 4, 4, 1, 2, 0, 2, 2, 0, 2, 0, 1, 3, 1, 2, 1, 1, 3.

Solution

The frequency distribution for the data is:

No. of Pets	Tally	Frequency	Relative Frequency (proportion)
0	////	4	4/20
1	𝍸 /	6	6/20
2	𝍸	5	5/20
3	///	3	3/20
4	//	2	2/20
	Total	20	1

For the sample, more families have 1 pet.

Example 3: These data represent the record high temperatures in degrees Fahrenheit (°F) for each of the 50 states. Construct a grouped frequency distribution for the data using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Creating a Grouped Frequency Distribution:

1. Arrange data (in ascending or descending order).
2. Compute the Range = Maximum value (largest value or highest value) – Minimum value (smallest or lowest value). [134-100=34]
3. Select the number of classes desired. This is usually between 5 and 20, (7 in this example).
4. Find the class width by dividing the range by the number of classes and rounding up. There are two things to be careful of here. You must *round up*, not off. Normally 3.2 would round to be 3, but in rounding up, it becomes 4. [34/7≈4.86→5]

Solution

The grouped frequency distribution for the data is:

Class limits	Class boundaries	Class Midpoints	Tally	Frequency	Cumulative Frequency	Relative Freq.(prop.)	Relative Freq.(perc.)
100–104	99.5–104.5	102	//	2	2	2/50	4
105–109	104.5–109.5	107	𝍸///	8	10	8/50	16
110–114	109.5–114.5	112	𝍸𝍸/𝍸𝍸/𝍸𝍸///	18	28	18/50	36
115–119	114.5–119.5	117	𝍸𝍸/𝍸𝍸///	13	41	13/50	26
120–124	119.5–124.5	122	𝍸𝍸//	7	48	7/50	14
125–129	124.5–129.5	127	/	1	49	1/50	2
130–134	129.5–134.5	132	/	1	50	1/50	2

$n = \Sigma f = 50$

If the range divided by the number of classes gives an integer value (or no remainder), add one unit to the class width.

e.g. Let the largest value = 31.5 and the smallest value = 7.5 and the data set is approximated to 1D,

$\Rightarrow$ range = 31.5 – 7.5 = 24

Say we need 5 classes,

The class width = 24 / 5 = 4.8 (1D, i.e. no remainder $\rightarrow$ add one unit 0.1)

$\therefore$ The class width = 4.8 + 0.1 = 4.9

Class limits
7.5–12.3
12.4–17.2
17.3–22.1
22.2–27
27.1–31.9

5. The boundaries (if needed) are half-way between the upper limit of one class and the lower limit of the next class.

If one of the data points meets one of the class boundaries, in which class it will be put?

It's impossible to happen since If data is integer $\rightarrow$ Class boundaries will be 1D, If data is 1D $\rightarrow$ Class boundaries will be 2D, ….

Stem & Leaf

Example 4: At an outpatient testing center, the number of cardiograms performed each day for 20 days is shown. Construct a stem and leaf plot for the data.

25	31	20	32	13
14	43	02	57	23
36	32	33	32	44
32	52	44	51	45

Solution

1- Arrange the data in order:

02, 13, 14, 20, 23, 25, 31, 32, 32, 32, 32, 33, 36, 43, 44, 44, 45, 51, 52, 57

2- Separate the data according to the first digit, as shown.

02 13, 14 20, 23, 25 31, 32, 32, 32, 32, 33, 36

43, 44, 44, 45 51, 52, 57

3- A display can be made by using the leading digit as the stem and the trailing digit as the leaf. For example, for the value 32, the leading digit, 3, is the stem and the trailing digit, 2, is the leaf. For the value 14, the 1 is the stem and the 4 is the leaf. Now a plot can be constructed as shown in Figure:

Leading digit (stem)	**Trailing digit (leaf)**
0	2
1	3 4
2	0 3 5
3	1 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

If there are no data values in a class!!!

you should write the stem number and leave the leaf row blank. Do not put a zero in the leaf row.

If the arranged data is: 50, 51, 51, 52, 53, 53, 55, 55, 56, 57, 57, 58, 59,
62, 63, 65, 65, 66, 66, 67, 68, 69, 69,
72, 73, 75, 75, 77, 78, 79.

Plot the data as shown:

Leading digit (stem)	Trailing digit (leaf)
5	0 1 1 2 3 3
5	5 5 6 7 7 8 9
6	2 3
6	5 5 6 6 7 8 9 9
7	2 3
7	5 5 7 8 9

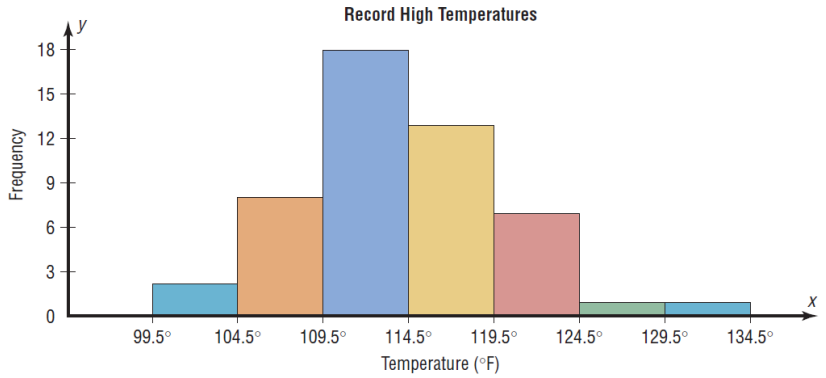
When the data values are in the hundreds, such as 325!!!

The stem and leaf plot for the data values 32, 27, 30, 42, 325, 41, 45, and 47 looks like this

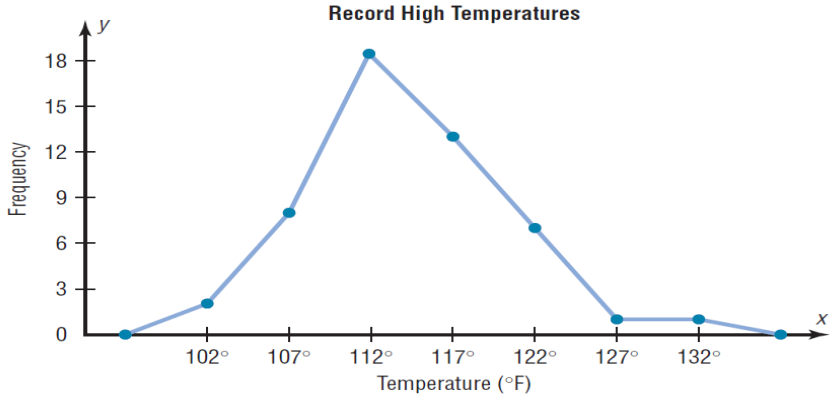
Leading digit (stem)	Trailing digit (leaf)
2	7
3	0 2
4	1 2 5 7
32	5

3. Presenting Data

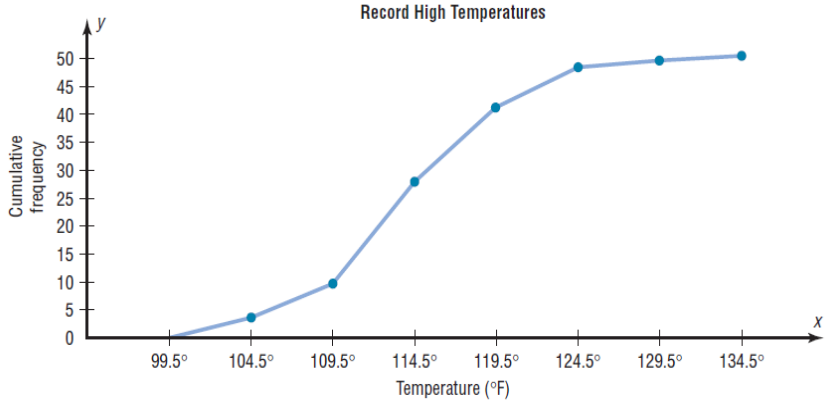
1- Histogram



2- Frequency Polygon



3- Ogive



Part I: Introduction to Statistical Methods.

Part II: Methods of Descriptive Statistics.

1-Collecting Data.
2-Organizing Data.
3-Presenting Data.
4-Summarizing Data.

Part III: Introduction to Probability.

Part IV: Methods of Inferential Statistics.