STAT 206: Chapter 2 (Organizing and Visualizing Variables)

- **Methods to Organize and Visualize Variables**
  - For **Categorical** Variables:
    - Summary Table; contingency table (2.1)
    - Bar chart, pie chart, Pareto chart, side-by-side bar chart (2.2)
  - For **Numerical** Variables
    - (Array), Ordered Array, frequency distribution, relative frequency distribution, percentage distribution, cumulative percentage distribution (2.3)
    - Stem-and-Leaf display, histogram, polygon, cumulative percentage polygon (2.4)
    - Other methods later…

- **2.1 Organizing Categorical Variables**
  - Must identify variable type to determine the appropriate organization and visualization tools
    ➔**Recall Variable Types**
    - Categorical (Category)
      - Nominal – Name of a Category
      - Ordinal – Has a natural ordering
    - Numerical / Quantitative (Quantity)
      - Discrete – distinct cutoffs between values
      - Continuous – on a continuum
  - **Definitions**:
    - **Summary Table**

    - **Contingency Table**

  - Each response counted/tallied into **one and only one** category/cell
  - Example (Problem 2.2, p. 40): The following data represent the responses to two questions asked in a survey of 40 college students majoring in business:
    - What is your gender? (M=male; F=female)
    - What is your major? (A=Accounting; C=Computer Information; M=Marketing)

Gender:	M	M	M	F	M	F	F	M	F	M
Major:	A	C	C	M	A	C	A	A	C	C
Gender:	F	M	M	M	M	F	F	M	F	F
Major:	A	A	A	M	C	M	A	A	A	C
Gender:	M	M	M	M	F	M	F	F	M	M
Major:	C	C	A	A	M	M	C	A	A	A
Gender:	F	M	M	M	M	F	M	F	M	M
Major:	C	C	A	A	A	A	C	C	A	C

Summary Table (Gender):			
value	frequency	relative frequency	percentage
Male (M)	25	0.625	62.5
Female (F)	15	0.375	37.5
TOTALS	40	1.000	100.0

Summary Table (Major):			
value	frequency	relative frequency	percentage
A (Accounting)	20	0.500	50.0
C (Computer)	15	0.375	37.5
M (Marketing)	5	0.125	12.5
TOTALS	40	1.000	100.0

- Now to combine the two variables (**Gender and Major**):

	MAJOR CATEGORIES			
GENDER	**A (Accounting)**	**C (Computer)**	**M (Marketing)**	**TOTALS**
Male (M)	14	9	2	25
Female (F)	6	6	3	15
TOTALS	20	15	5	40

  - Table based on **Total** percentages:

	MAJOR CATEGORIES			
GENDER	**A (Accounting)**	**C (Computer)**	**M (Marketing)**	**TOTALS**
Male (M)				
Female (F)				
TOTALS				

  - Table based on **Row** percentages:

	MAJOR CATEGORIES			
GENDER	**A (Accounting)**	**C (Computer)**	**M (Marketing)**	**TOTALS**
Male (M)				
Female (F)				
TOTALS				

  - Table based on **Column** percentages:

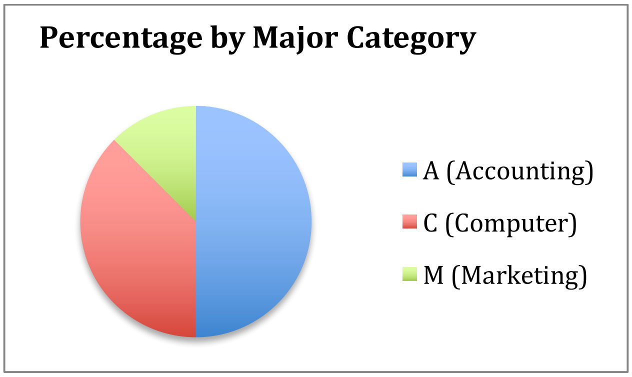
	MAJOR CATEGORIES			
GENDER	**A (Accounting)**	**C (Computer)**	**M (Marketing)**	**TOTALS**
Male (M)				
Female (F)				
TOTALS				

- **Questions:**
  - How many of the surveyed students were females majoring in Marketing?
  - What percentage of the surveyed students were females majoring in Marketing?
  - What percentage of the male students surveyed were majoring in Computer?
  - Of the students majoring in Accounting, what percentage was male?

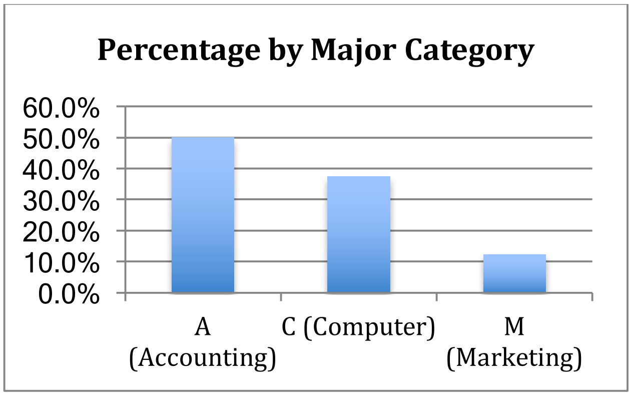
- **2.3 Visualizing Categorical Variables**
  - **Pie chart** – uses sections of a circle to represent the tallies/frequencies/percentages for each category
  - **Bar chart** – a series of bars, with each bars representing the tallies/frequencies/percentages for a single category
  - Consider our previous example for Major Category:

Summary Table (Major):			
value	**frequency**	**relative frequency**	**percentage**
A (Accounting)	20	0.500	50.0
C (Computer)	15	0.375	37.5
M (Marketing)	5	0.125	12.5
TOTALS	40	1.000	100.0

    - Pie Chart:



    - Bar Chart:



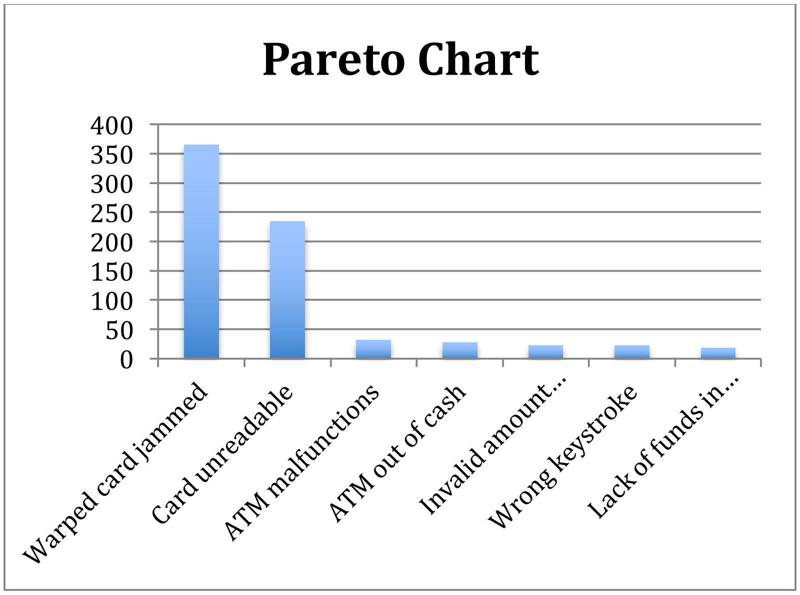
    - **Question**: Which major has the lowest concentration of students?

    - Discussion: Preference for type of chart?

- **Pareto chart** – a series of vertical bars showing tallies/frequencies/percentages in **descending order**
- Example:

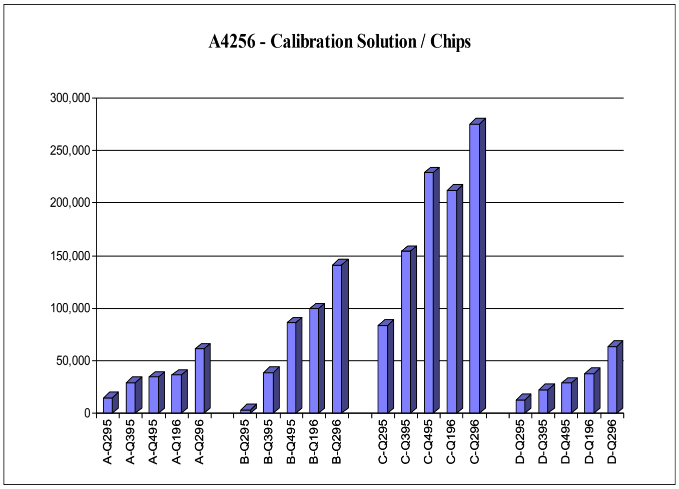
Summary Table of Causes of Incomplete ATM Transactions

Cause	Frequency	Percentage
ATM malfunctions	32	4.42%
ATM out of cash	28	3.87%
Invalid amount requested	23	3.18%
Lack of funds in account	19	2.62%
Card unreadable	234	32.32%
Warped card jammed	365	50.41%
Wrong keystroke	23	3.18%
TOTAL	724	100.00%

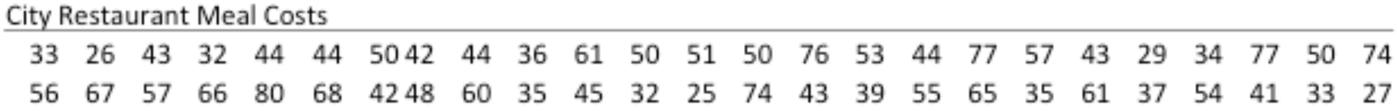


- **Discussion**: How or why do you think that a Pareto chart would be useful in the business world?

- **Side-by-Side Bar charts** – Uses sets of bars to show the joint response from two categorical variables
  - Example:
  - **Discussion**: What can you determine about product utilization for this side-by-side bar chart that you might not be able to tell otherwise?

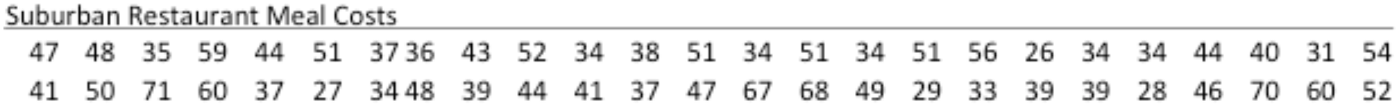


- **2.2 Organizing Numerical Variables**
  - **Ordered array** arranges the values of a numerical variable in rank order (smallest value to largest value) Array ➔ Ordered Array
  - Example (Table 2.8 A & B, p. 42):



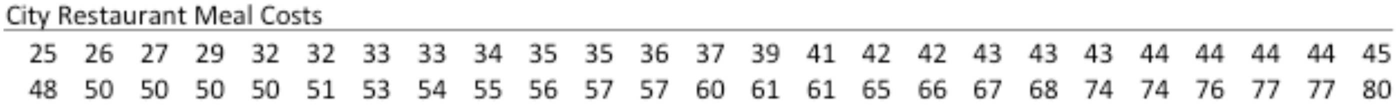
City Restaurant Meal Costs

33	26	43	32	44	44	50	42	44	36	61	50	51	50	76	53	44	77	57	43	29	34	77	50	74
56	67	57	66	80	68	42	48	60	35	45	32	25	74	43	39	55	65	35	61	37	54	41	33	27



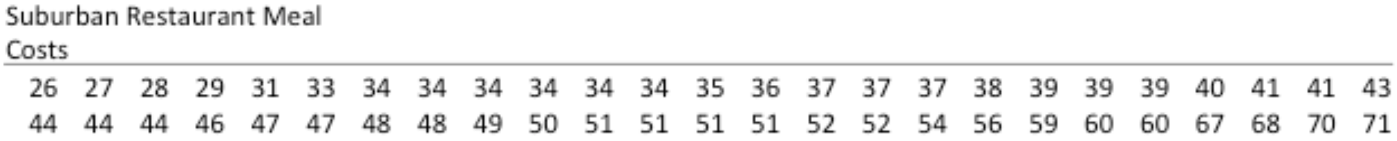
Suburban Restaurant Meal Costs

47	48	35	59	44	51	37	36	43	52	34	38	51	34	51	34	51	56	26	34	34	44	40	31	54
41	50	71	60	37	27	34	48	39	44	41	37	47	67	68	49	29	33	39	39	28	46	70	60	52



City Restaurant Meal Costs

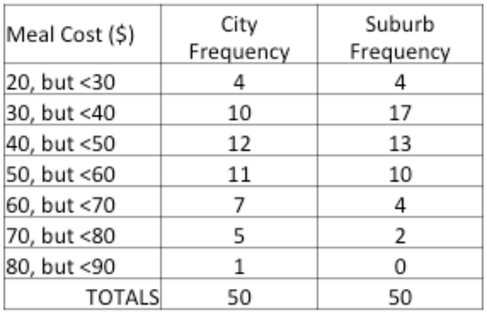
25	26	27	29	32	32	33	33	34	35	35	36	37	39	41	42	42	43	43	43	44	44	44	44	45
48	50	50	50	50	51	53	54	55	56	57	57	60	61	61	65	66	67	68	74	74	76	77	77	80



Suburban Restaurant Meal Costs

26	27	28	29	31	33	34	34	34	34	34	34	35	36	37	37	37	38	39	39	39	40	41	41	43
44	44	44	46	47	47	48	48	49	50	51	51	51	51	52	52	54	56	59	60	60	67	68	70	71

- **Frequency Distribution** tallies the values of a numerical variable into a set of numerically ordered **classes**, called a ***class interval***
  - How many classes?
  - Determine the interval width by the following:
  - Using our Meal Cost data, we estimate that we want __________ classes so the interval width is:



Meal Cost ($)	City Frequency	Suburb Frequency
20, but <30	4	4
30, but <40	10	17
40, but <50	12	13
50, but <60	11	10
60, but <70	7	4
70, but <80	5	2
80, but <90	1	0
TOTALS	50	50

- **Relative Frequency Distribution** presents relative frequency, or proportion of the total for each group
- **Proportion** or **relative frequency**, in each group is equal to the number of values in each class divided by the total number of values
  - **Example:**

	CITY			**SUBURBAN**		
Meal Cost ($)	Frequency	Relative Frequency	Percentage	Frequency	Relative Frequency	Percentage
20, but <30	4			4		
30, but <40	10			17		
40, but <50	12			13		
50, but <60	11			10		
60, but <70	7			4		
70, but <80	5			2		
80, but <90	1			0		
TOTALS	50	1	100.00%	50	1	100.00%

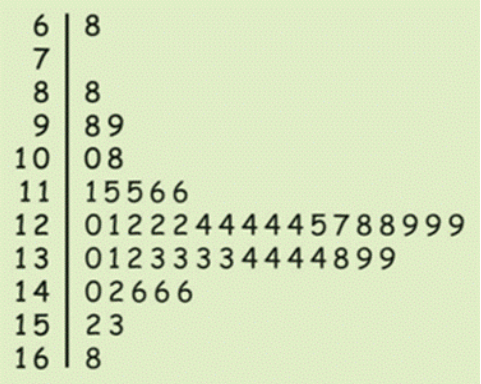
  - **TOTAL** of the **relative frequency** column MUST BE **1.00**
  - **TOTAL** of the **percentage** column MUST BE **100.00**

- **Cumulative Percentage Distribution** provides a way of presenting information about the percentage of values that **less than a specific amount**

	CITY and SUBURBAN				
Meal Cost ($)	Frequency	Relative Frequency	Percentage	< lower boundary	Cumulative Percentage < lower boundary
20, but <30	8	0.08	8.0%	<20	0 (no meals cost less than $20)
30, but <40	27	0.27	27.0%	<30	8% = 0 + 8%
40, but <50	25	0.25	25.0%	<40	35% = 0 + 8% +27%
50, but <60	21	0.21	21.0%	<50	60% = 0 + 8% +27% + 25%
60, but <70	11	0.11	11.0%	<60	81% = 0 + 8% +27% + 25% + 21%
70, but <80	7	0.07	7.0%	<70	92% = 0 + 8% +27% + 25% + 21% + 11%
80, but <90	1	0.01	1.0%	<80	99% = 0 + 8% +27% + 25% + 21% + 11% + 7%
TOTALS	100	1.00	100.0%	<90	100% = 0 + 8% +27% + 25% + 21% + 11% + 7% + 1%

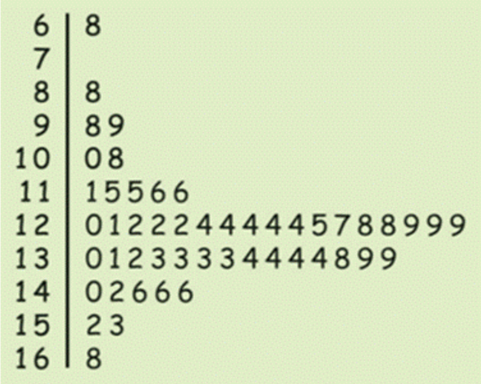
  - **Question:** What percentage of meal costs was less than $50?

- **2.4 Visualizing Numerical Variables**
  - **Stem-and-Leaf Display** – How to create:
    1. Separate each observation into
       - Stem (all but final digit(s)) and
       - Leaf (final digit(s)).
    2. Write stems in vertical column – smallest on top
    3. Write each leaf, *in increasing numerical order*, in row next to appropriate stem
  - Example: For each state, percentage (with one decimal place) of residents 65 and older
  - Notice stem of "7" does not have a leaf ➔ we conclude no value of 7.x there should be the same number of ***leaves*** as observations! Include ALL ***stems*** even if no values/***leaves***
    - Leave a space holder if no leaf for a stem
    - No punctuation (i.e., no decimal points, no commas)
    - Leaves should be lined on top of one another to determine ***SHAPE***
    - Simple way to deliver a lot of detailed information



6	8
7	
8	8
9	89
10	08
11	15566
12	01222444445788999
13	01233334444899
14	02666
15	23
16	8

<u>FOR THIS EXAMPLE read data values as</u>:

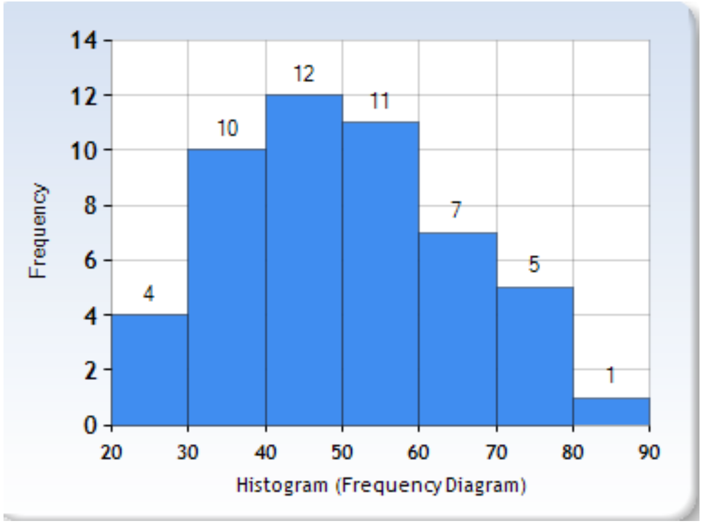


6	8
7	
8	8
9	89
10	08
11	15566
12	01222444445788999
13	01233334444899
14	02666
15	23
16	8

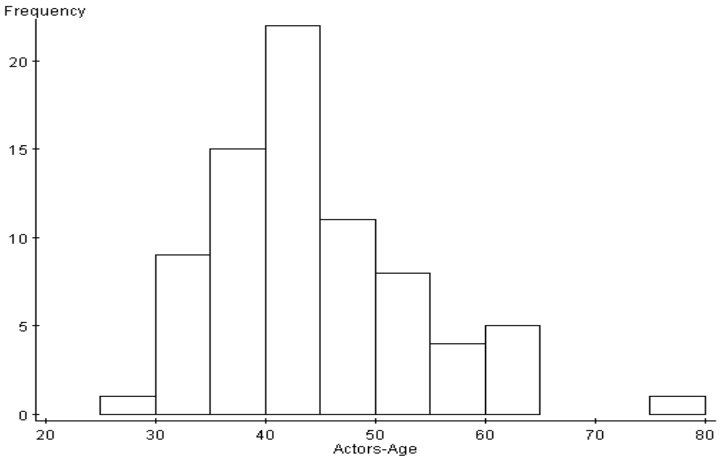
6.8
-
8.8
9.8, 9.9
10.0,10.8
11.1,11.5,11.5,11.6,11.6
12.0,12.1,12.2,12.2,12.2,12.4,12.4,12.4,12.4,12,4,12.5,12.7,12.8,12.8,12.9,12.9,12.
13.0,13.1,12.3,13.3,13.3,13.3,13.4,13.4,13.4,13.4,13.8,13.9,13.9
14.0,14.2,14.6,14.6.14.6
15.2,15.3
16.8

- **Histogram**: Displays a **quantitative** variable across different groupings of values
  - Careful when choosing how to group together values!
    - Groupings must cover the same range so have of equal width
    - Height used to compare the frequency of each range of values
    - Steps to create a frequency histogram:
      - Create equal width classes (groupings)
      - Count number of values in each class
      - Draw histogram with a bar for each class
      - Height of a bar represents the count for that bar’s class
      - Bars touch since there are NO GAPS between classes
    - Be careful:
      - Number of categories can’t be too large or too small
      - Don’t skip any categories
      - Be clear about contents of each category

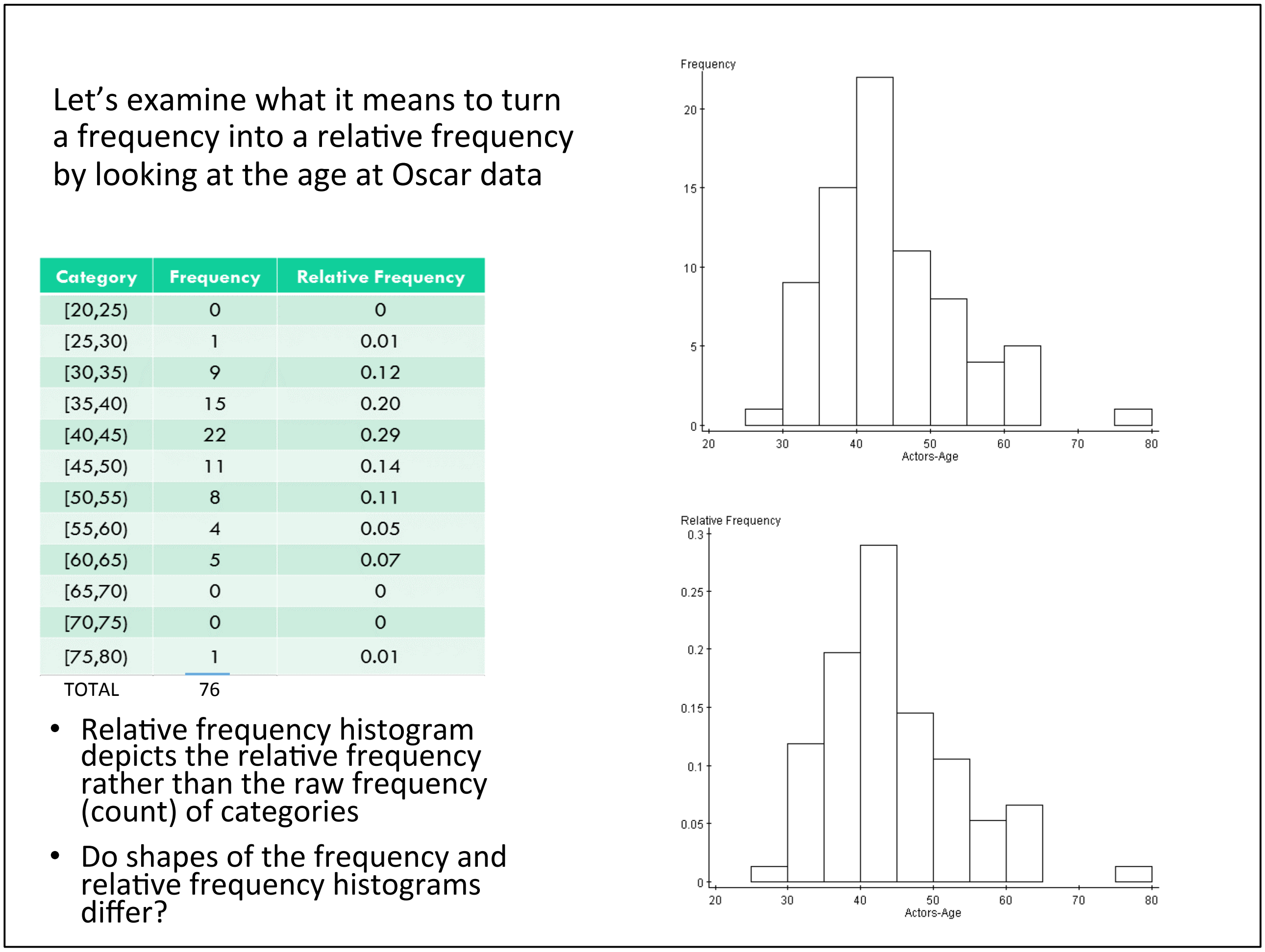
HISTOGRAM of Meal Cost Location = CITY



- **Histogram Example**: using Age at Time of First Oscar Award:



  - Groupings chosen here are:
    [20,25)   [25,30)   [30,35)   [35,40)   [45,50), …
  - Where “**[**“ means the number is INCLUDED in the interval,
    but “**)**” means the number is NOT included in the interval
  - **Question**: If Jack Nicholson won Best Actor at age 70, which category frequency would increase?
    A. [60,65)
    B. [65,70)
    C. [70,75)
    D. [75,80)

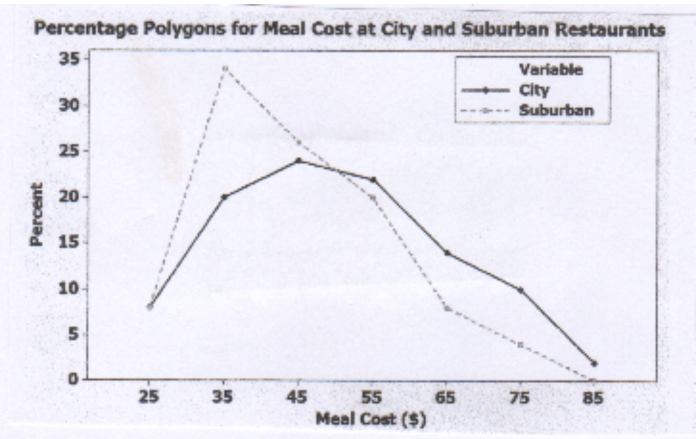


Let’s examine what it means to turn a frequency into a relative frequency by looking at the age at Oscar data

Category	Frequency	Relative Frequency
[20,25)	0	0
[25,30)	1	0.01
[30,35)	9	0.12
[35,40)	15	0.20
[40,45)	22	0.29
[45,50)	11	0.14
[50,55)	8	0.11
[55,60)	4	0.05
[60,65)	5	0.07
[65,70)	0	0
[70,75)	0	0
[75,80)	1	0.01
TOTAL	76	

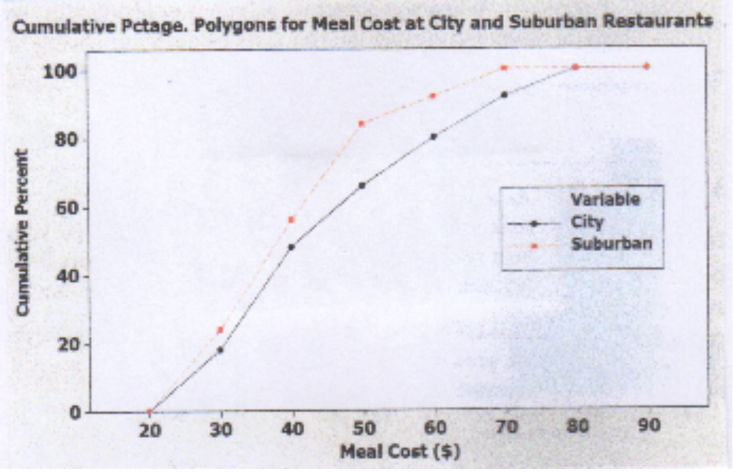
- Relative frequency histogram depicts the relative frequency rather than the raw frequency (count) of categories
- Do shapes of the frequency and relative frequency histograms differ?

- **Percentage polygon** – used for visualization when dividing the data of a numerical variable into two or more groups
  - Uses midpoints of each class to represent the data in the class
  - Combines data from two groups to allow easier comparison



Conclusions?

- **Cumulative Percentage Polygon (Ogive)** uses the cumulative percentage distribution (discussed previously) to plot the cumulative percentages along the ***Y*** axis
    - LOWER BOUNDS of the class intervals are plotted on the ***X*** axis



Conclusions?