St. John's University

Center for Educational Research
Leadership and Accountability

Educational Research and Data Analysis II
EDU 7211

Dr. Francesco Ianni

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

H.G.Wells

Data Analysis Introduction

In this first lecture we will have an introduction to the course and a quick review of SPSS.

We will make use of real data in order to pose relevant questions and learn the appropriate methods of analysis.

The difference between descriptive and inferential statistics will be explained in addition to the different level of measurement.

Descriptive - Inferential

- **Descriptive**
  - **Descriptive Statistics deals with data that describes a specific group of individuals.**
- **Inferential**
  - **Inferential Statistics deals with the estimation of the population parameters using descriptive analysis.**

Let's Look at some terms

- **Data**
  - The measurements obtained from observations.
- **Statistics**
  - All the methods used to interpret a set of data.
- **Statistic**
  - A single number used to describe a set of data from a sample

Let's Look at some terms

- **Variable**
  - Values that change from one individual to another.
- **Constant**
  - A value that doesn't change from subject to subject.
- **Distribution**
  - A set of scores

The Nature of Data

- **Quantitative**
  - Reflects measurement.
- **Qualitative**
  - Describe qualities.

Quantitative Data

- **Discrete**
  - Data that can only be represented with integers.
- **Continuous**
  - Data that can be represented with the use of any value in any given interval.

Another Way to Classify Data

- **Four Levels of Measurement**
  - **Nominal**
  - **Ordinal**
  - **Interval**
  - **Ratio**

Nominal

- **This level of measurement is characterized by data that represents names, labels or categories only.**
- **Examples**

Ordinal

- **Ordinal data that describes order.**
- **Differences between values either cannot be determined or are different from each other.**
- **Example**

Interval

- **Same as ordinal level, but the differences between data is the same.**
- **Zero in this case is just a point on the scale.**
- **Example**

Ratio

- Same of the Interval Level with the additional fact that zero represents the absence of whatever value we are measuring.
- Data at this level have a true zero point.
- Example

"If you cannot convince them, confuse them."

Harry S. Truman

Summary

- Nominal
- Ordinal
- Interval
- Ratio

Quick Practice: Level of Measurement

- Numbers on uniforms that identify players on a cycling team
- Student-teachers rankings of cafeteria food as excellent, ok or poor
- Calendar year of historic events
- Temperatures on Celsius scale
- Runners' times in the Boston Marathon

Additional Practice

Discrete or Continuous?

1. Measurements of the time it takes to walk 3 miles
2. The number of calendar years
3. The number of bikes on a bike store
4. The amount of milk produced in Milk's farm
5. Number of taxi cabs in NYC at 3 pm
6. Numbers of bikes passing through a busy street

Nominal, Ordinal, Interval or Ratio?

- Party affiliation of voters at the pres. Elect.
- Body temperature of a newborn baby
- Type of motorcycle bought by a customer at a motorcycle dealership
- Final grade in STAT (A,B,….D, F)

SPSS

- Open Data
- Data View
- Variable View
- Entering Numerical Data
- Entering Qualitative Data
- The Output Window
- Saving Data

Practice 1

Distribution Analysis

We will learn….

- **Counting the occurrence of data Values**
  - When variables measured at the Nominal level
  - When they are not measured at the Nominal level
  - Describing the shape of the distribution
- **Accumulating Data**
  - Cumulative percent
  - Box Plots
- **How to use SPSS**

Counting the occurrence of data

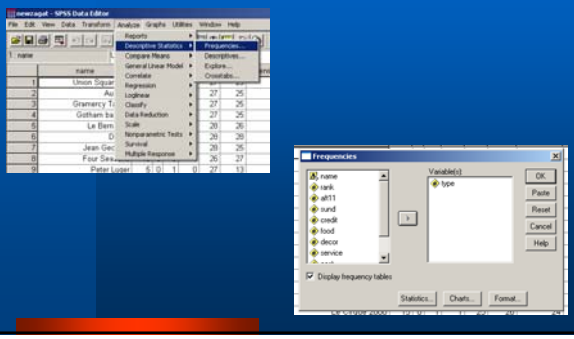
- **When is measured at the nominal level**
  - Frequency and Percent Distribution Table
  - Bar Graph
  - Pie Graph

Frequency and Percent Distribution Tables

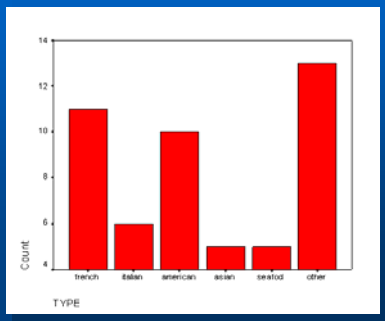
TYPE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	french	11	22.0	22.0	22.0
	italian	6	12.0	12.0	34.0
	american	10	20.0	20.0	54.0
	asian	5	10.0	10.0	64.0
	seafod	5	10.0	10.0	74.0
	other	13	26.0	26.0	100.0
	Total	50	100.0	100.0	

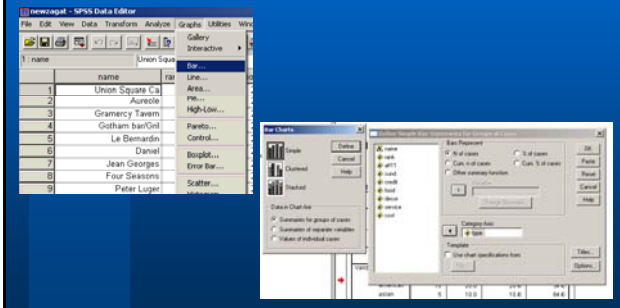
On SPSS



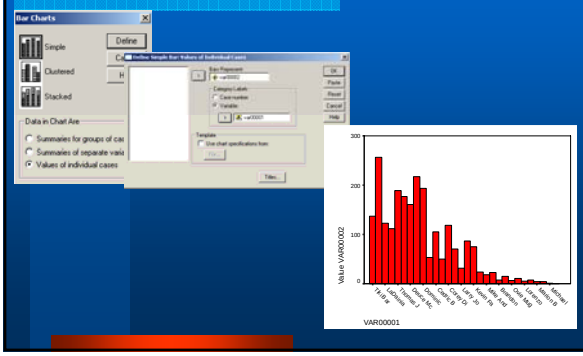
Bar Graphs



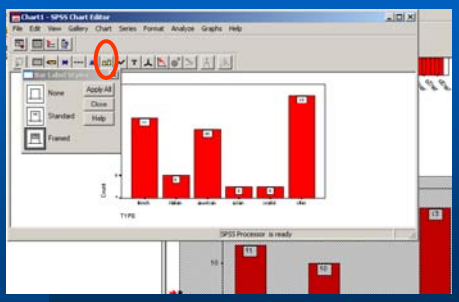
On SPSS



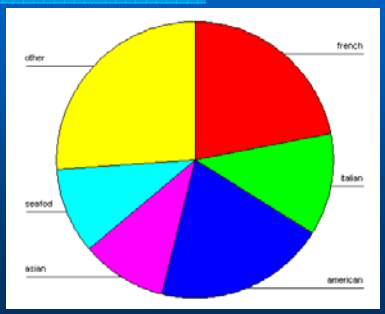
To create a Bar graph for ind. cases



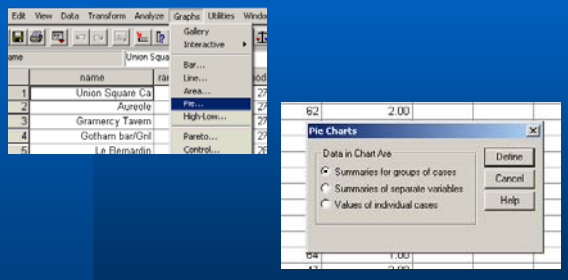
More on Bar Graphs



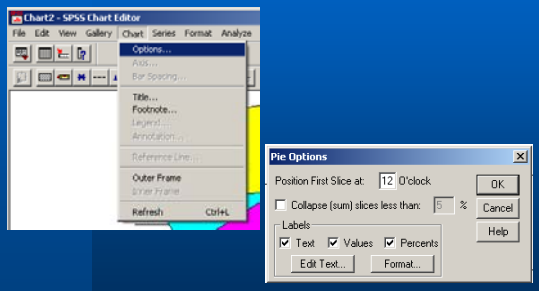
Pie Graphs



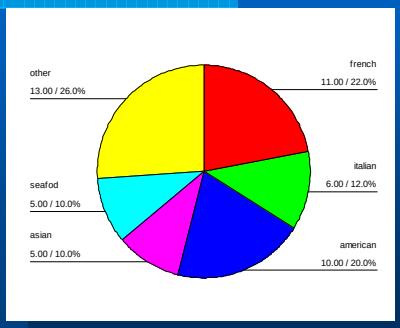
On SPSS



More on Pie Graphs



Pie Graph



Counting the occurrence of data

- **When the variable is not measured at the nominal level**
  - **Stem and Leaf Plot**
  - **Histogram**
  - **Line Graph**

Stem and Leaf Plot

Given:

13,15,16,17,17,17,17 ,18,

20,22,23, 24, 24,24,25,26,26,

32,33,35,

45,46,48,49,

52,53

VAR00001

VAR00001 Stem-and-Leaf Plot

Frequency	Stem &	Leaf
1.00	1 .	3
7.00	1 .	5677778
6.00	2 .	023444
3.00	2 .	566
2.00	3 .	23
1.00	3 .	5
.00	4 .	
4.00	4 .	5689
2.00	5 .	23

Stem width: 10
Each leaf: 1 case(s)

Stem and Leaf Plot : updated

Given:

130,150,160,170, 170,170,170,180,

200,220,230, 240, 240,240,250,260, 260,

320,330,350,

450,460,480,490,

520,530

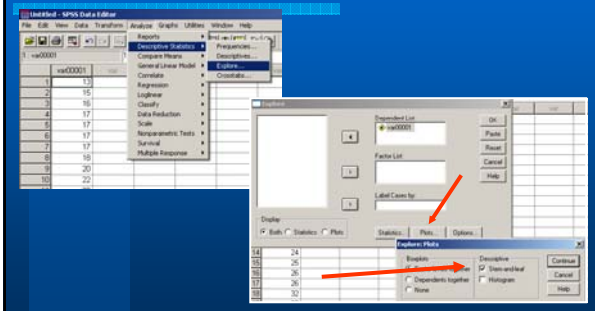
VAR2

VAR2 Stem-and-Leaf Plot

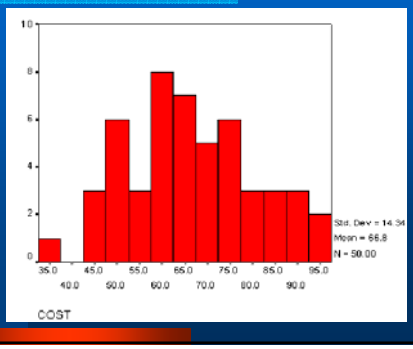
Frequency	Stem &	Leaf
1.00	1 .	3
7.00	1 .	5677778
6.00	2 .	023444
3.00	2 .	566
2.00	3 .	23
1.00	3 .	5
.00	4 .	
4.00	4 .	5689
2.00	5 .	23

Stem width: 100
Each leaf: 1 case(s)

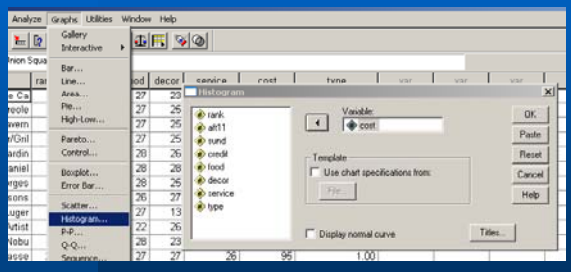
On SPSS



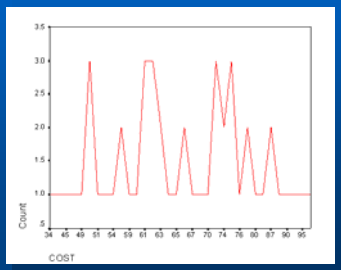
Histograms



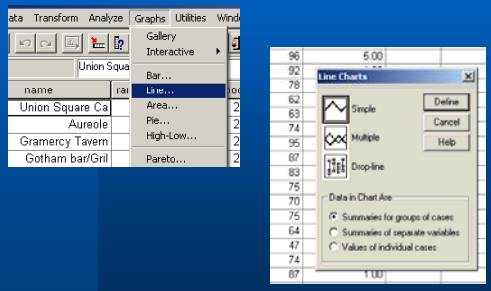
On SPSS



Line Graph



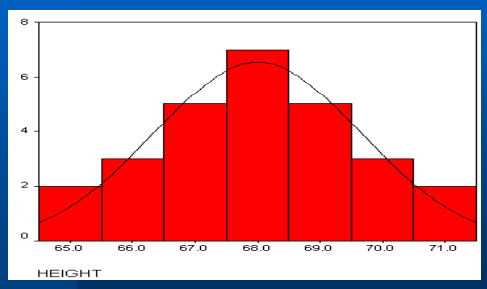
On SPSS



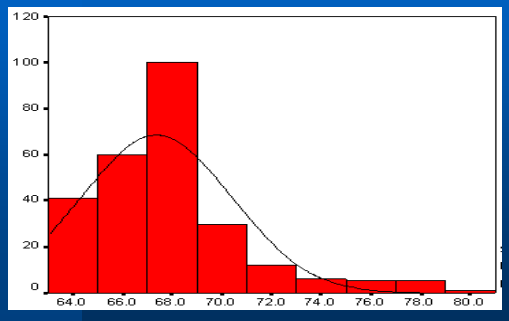
Describing the Shape

- Symmetric
- Positively Skewed
- Negatively Skewed

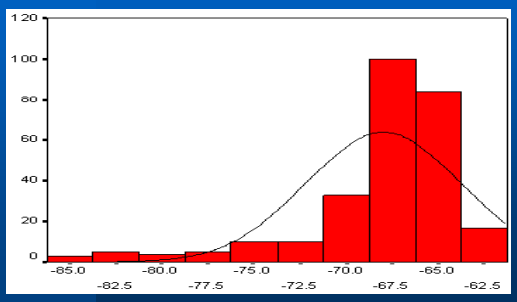
Symmetric



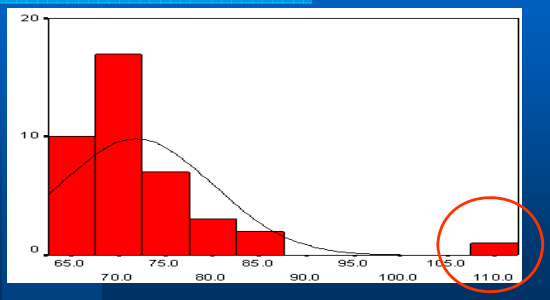
Positive Skew



Negative Skew



Outliers

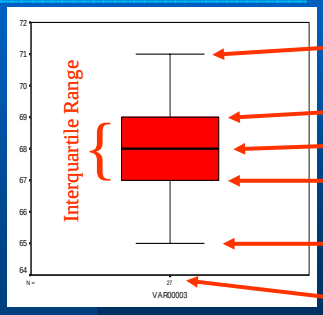


Cumulative Percent Distribution

COST

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	34	1	2.0	2.0	2.0
	43	1	2.0	2.0	4.0
	45	1	2.0	2.0	6.0
	47	1	2.0	2.0	8.0
	49	1	2.0	2.0	10.0
	50	3	6.0	6.0	16.0
	51	1	2.0	2.0	18.0
	52	1	2.0	2.0	20.0
	54	1	2.0	2.0	22.0
	57	2	4.0	4.0	26.0
	59	1	2.0	2.0	28.0
	60	1	2.0	2.0	30.0
	61	3	6.0	6.0	36.0
	62	3	6.0	6.0	42.0
	63	2	4.0	4.0	46.0
	64	1	2.0	2.0	48.0
	65	1	2.0	2.0	50.0
	66	2	4.0	4.0	54.0
	67	1	2.0	2.0	56.0
	68	1	2.0	2.0	58.0
	70	1	2.0	2.0	60.0

Box Plot



Maximum

Quartile 3

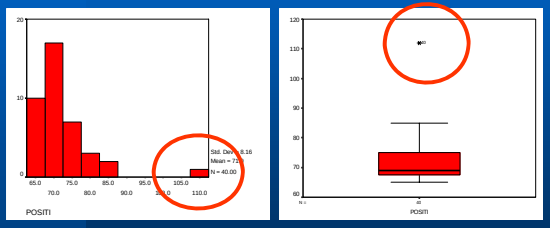
Median or Quartile 2

Quartile 1

Minimum

Total number of cases

Outliers: Box Plot and Histogram



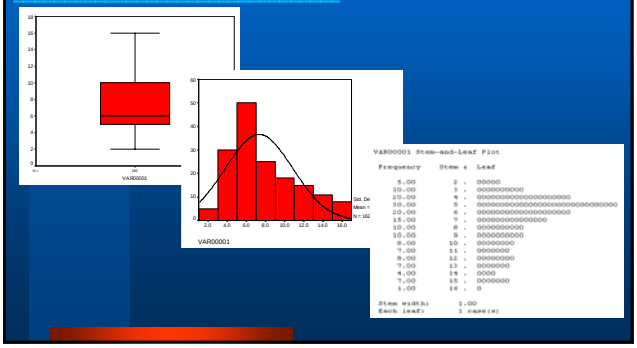
Outliers: Stem and Leaf Plot

POSITI Stem-and-Leaf Plot

Frequency	Stem &	Leaf
2.00	6 .	55
8.00	6 .	66677777
12.nn	6 .	RRRRRRRQQQQQ
5.00	7 .	00011
2.00	7 .	33
4.00	7 .	5555
1.00	7 .	6
2.00	7 .	89
1.00	8 .	0
1.00	8 .	3
1.00	8 .	5
1.00 Extremes		(>=112)

Stem Width: 10.00
Each leaf: 1 case(s)

Positive Skew

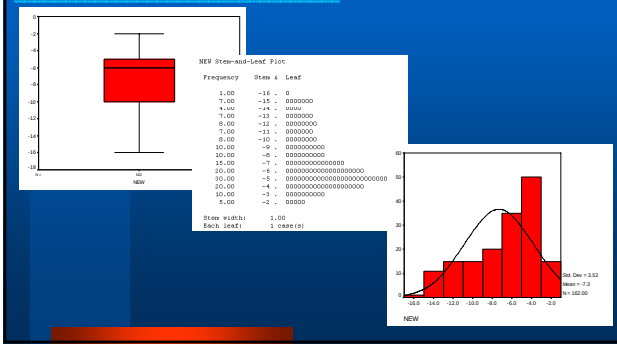


VAR00001 Stem-and-Leaf Plot

Frequency	Stem &	Leaf
5.00	2 .	00000
10.00	3 .	0000000000
20.00	4 .	00000000000000000000
30.00	5 .	000000000000000000000000000000
20.00	6 .	00000000000000000000
15.00	7 .	000000000000000
10.00	8 .	0000000000
10.00	9 .	0000000000
8.00	10 .	00000000
7.00	11 .	0000000
8.00	12 .	00000000
7.00	13 .	0000000
4.00	14 .	0000
7.00	15 .	0000000
1.00	16 .	0

Stem Width: 1.00
Each leaf: 1 case(s)

Negative Skew

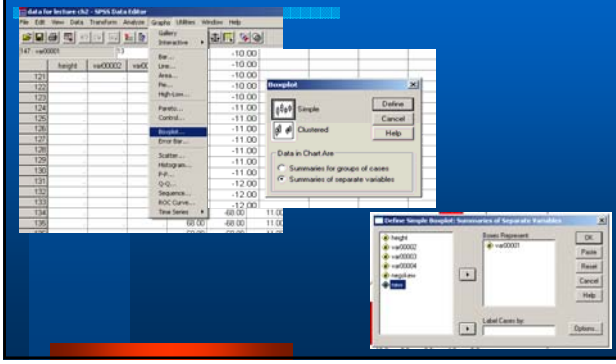


NEW Stem-and-Leaf Plot

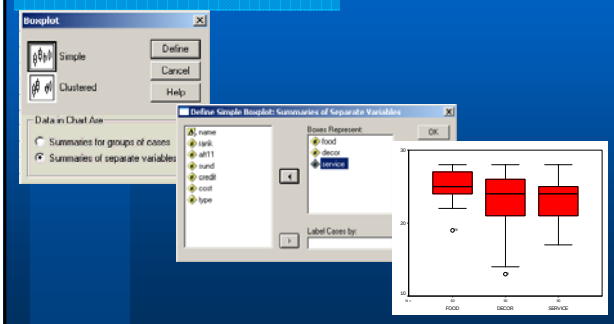
Frequency	Stem &	Leaf
1.00	-16 .	0
7.00	-15 .	0000000
9.00	-14 .	uuuu
7.00	-13 .	0000000
8.00	-12 .	00000000
7.00	-11 .	0000000
8.00	-10 .	00000000
10.00	-9 .	0000000000
10.00	-8 .	0000000000
15.00	-7 .	000000000000000
20.00	-6 .	00000000000000000000
30.00	-5 .	000000000000000000000000000000
20.00	-4 .	00000000000000000000
10.00	-3 .	0000000000
5.00	-2 .	00000

Stem Width: 1.00
Each leaf: 1 case(s)

Box Plot on SPSS



More than one box plot



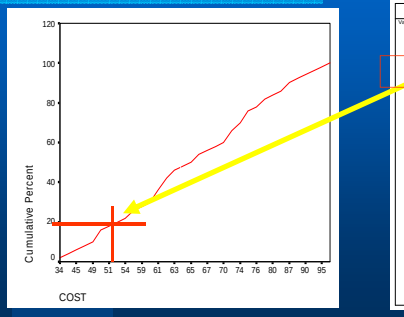
Accumulating Data

- Cumulative Percent Distributions
- Ogive Curves
- Percentile Ranks
- Percentiles
- Quartiles
- Box Plots

Cumulative Percent

- **Definition:**
  - **Percent of values at or below a certain value in the distribution**

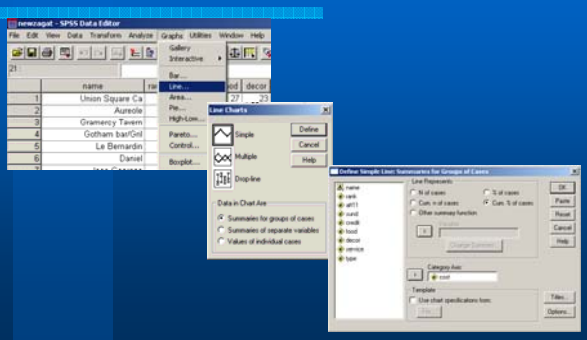
Ogive Curve



COST

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	34	1	2.0	2.0	2.0
	43	1	2.0	2.0	4.0
	45	1	2.0	2.0	6.0
	47	1	2.0	2.0	8.0
	49	1	2.0	2.0	10.0
	50	3	6.0	6.0	16.0
	51	1	2.0	2.0	18.0
	52	1	2.0	2.0	20.0
	54	1	2.0	2.0	22.0
	57	2	4.0	4.0	26.0
	59	1	2.0	2.0	28.0
	60	1	2.0	2.0	30.0
	61	3	6.0	6.0	36.0
	62	3	6.0	6.0	42.0
	63	2	4.0	4.0	46.0
	64	1	2.0	2.0	48.0
	65	1	2.0	2.0	50.0
	66	2	4.0	4.0	54.0
	67	1	2.0	2.0	56.0
	68	1	2.0	2.0	58.0
	70	1	2.0	2.0	60.0
	72	3	6.0	6.0	66.0
	74	2	4.0	4.0	70.0
	75	3	6.0	6.0	76.0
	76	1	2.0	2.0	78.0
	78	2	4.0	4.0	82.0
	80	1	2.0	2.0	84.0
	83	1	2.0	2.0	86.0
	87	2	4.0	4.0	90.0
	88	1	2.0	2.0	92.0
	90	1	2.0	2.0	94.0
	92	1	2.0	2.0	96.0
	95	1	2.0	2.0	98.0
	96	1	2.0	2.0	100.0
	Total	50	100.0	100.0	

How do we create an Ogive Graph



Percentile Ranks

- **Percentile rank (of a raw score) is the percent of scores falling <u>below</u> that raw score in the distribution**

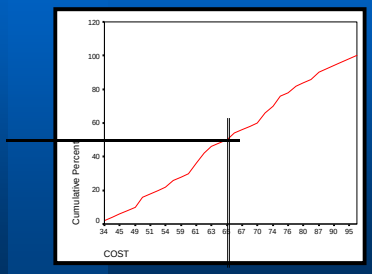
Example

- **A raw score of 780 on a standardized test might have a percentile rank of 80. This means that 80% of the students taking the test received a score <u>less than</u> 780**

Percentiles

- **The Percentile indicates the raw score that would correspond to a given percentile rank.**
- **The values divide the distribution according to the percentage of numbers falling below or above**

Example



50% of the cost falls below 65 dollars

Thus 65 is the 50th percentile with 50% of the cases below and 50% of the cases are above

Percentile Ranks

- **Percentile rank (of a raw score) is the percent of scores falling <u>below</u> that raw score in the distribution**

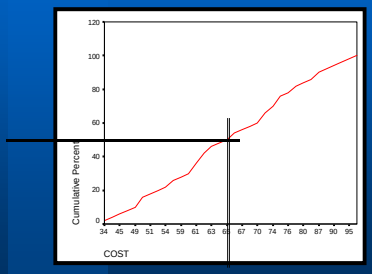
Example

- **A raw score of 780 on a standardized test might have a percentile rank of 80. This means that 80% of the students taking the test received a score <u>less than</u> 780**

Percentiles

- The Percentile indicates the raw score that would correspond to a given percentile rank.
- The values divide the distribution according to the percentage of numbers falling below or above

Example



50% of the cost falls below 65 dollars

Thus 65 is the 50th percentile with 50% of the cases below and 50% of the cases are above

Quartiles

- The three theoretical raw scores that divide the distribution into four equal parts
  - Quartile 1 = Q1
  - Quartile 2 = Q2 = Median
  - Quartile 3 = Q3

Example

- If there is a total of 1000 restaurants in our survey
  - Between Q1 and Q2 = 250
  - Between Q2 and Q3 = 250
  - Above Q3 = 250
  - Below Q1 = 250


Summary