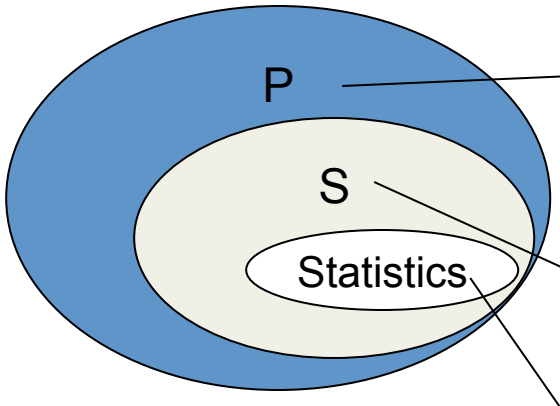
Lecture 4
Descriptive Statistics

Fall 2013

Prof. Yao Xie, yao.xie@isye.gatech.edu
H. Milton Stewart School of Industrial Systems & Engineering
Georgia Tech

Population, Sample, Statistics



Population : all possible observations

Sample : observed measurements taken from population

Statistics : useful summary of the sample data

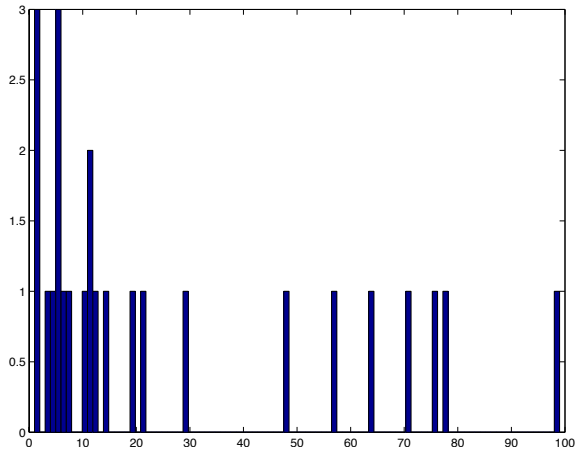
Descriptive Statistics

Descriptive statistics provides simple summaries about the sample (observations) using:

Quantitative values (summary statistics)

e.g. NBA, shooting percentage of a player = 33%, meaning that the player makes approximately one shot in every three.

Visual displays (simple-to-understand graphs)



Histogram of “number guessing” result

Descriptive Statistics vs. Inferential Statistics

Descriptive statistics provides simple summaries about the sample (observations). (typically not based on theory of probability).

→ used to help build intuitive understanding of the data.

e.g. draw a group of sample patients, their average age, proportion of male to female, demographic chart.

Inferential statistics: assuming a probabilistic model for the population, and use data (samples) to learn about the probabilistic model.

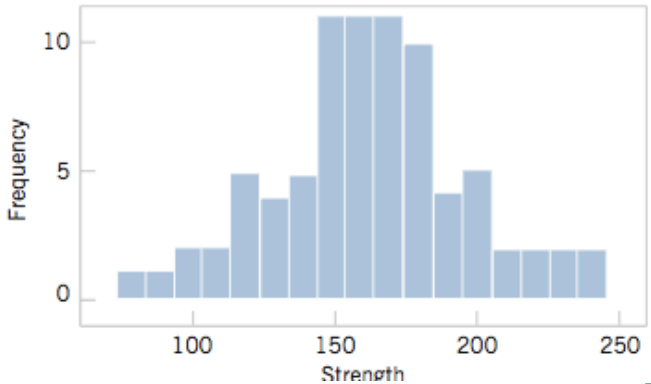
→ used to draw “main conclusion”.

e.g. e.g. patient blood pressure has Gaussian distribution. We use samples to estimate the mean and variance.

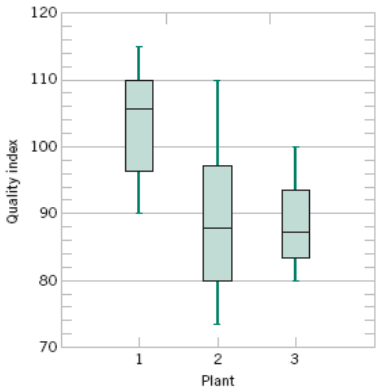
Exploratory Data Analysis (EDA)

A set of methods to generate descriptive statistics: **exploratory data analysis**

- Histogram



- Box plot



- Stem-and-leaf diagram

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Stem : Tens and hundreds digits (psi); Leaf: Ones digits (psi)

Data type

Sample of data $x_1, x_2, \cdots, x_n$

Numerical data

Observed values of x_i are integer, real or complex numbers

Examples: GE scores of GT students (integer values)

Lifetime of a light bulb (real values)

Categorical or nominal data

Observed values of x_i is one of several categories, e.g. “male”, “female”, that are not associated with numerical values.

Often summarized using frequency of occurrence of each category.

Data Features

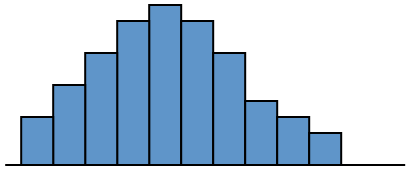
Shape of the data distribution: symmetric, skewed to the right or to the left.

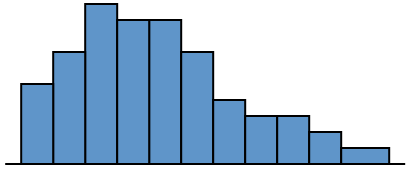
Spread of the data: range, long or short tails

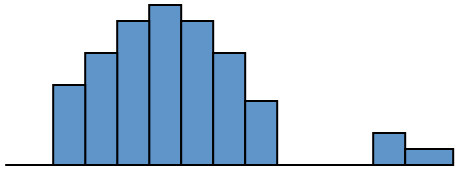
Outliers: extreme values that appear separate from the rest of the data

Modes: concentrations of the data – unimodal, bimodal

Gaps: different subpopulations







Estimating Central Tendency

Sample of data $x_1, x_2, \cdots, x_n$

1. Sample mean: $\bar{x} = \frac{1}{n} x_1 + \frac{1}{n} x_2 + \cdots + \frac{1}{n} x_n$

2. Sample median – numerical value that separates the higher half from the lower half of the samples
   1) rank the data to get $y_1, y_2, \cdots, y_n$
   2) odd number of samples, median = $y_{(n+1)/2}$
      even number of samples, median = $(y_{(n-1)/2} + y_{(n-1)/2})/2$

3. Trimmed mean – take out extremes from both sides and compute mean of the trimmed samples

Example:

Data = {3, 1, 1, 0, 5, 4, 13, 3}

1. Mean: 3.75
2. Median : order data (0, 1, 1, 3, 3, 4, 5, 13)
   3

Data = {3, 1, 1, 0, 5, 4, 3}

1. Mean: 2.4286
2. Median : order data (0, 1, 1, 3, 3, 4, 5)
   3

Estimating Data Variability

1. Sample range = [largest item] – [smallest item]

2. Sample variance (n-1 degree of freedom)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

3. Sample quartile

   The p^{th} quartile (x_p) is such that 100p% of the sample is smaller than x_p

4. Inter quartile range (IQR)

   IQR = Upper quartile - Lower quartile

Example: Data = {3, 1, 1, 0, 5, 4, 13, 3}

1. Sample Range: 10
2. Sample Variance: 16.7857
3. Sample quantiles

   - ordered sample – [0,1,1,3,3,4,5,13]
   - lower (25th) quartile = (1+1)/2 = 1
   - median (50th) quartile = 3
   - upper (75th) quartile = (4+5)/2 = 4.5

4. IQR = upper quartile - lower quartile = 4.5 - 1= 3.5

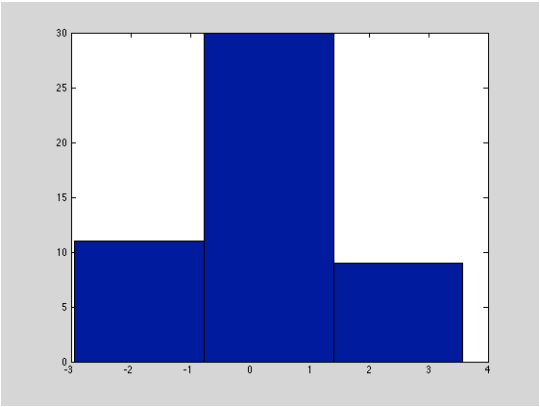
Histograms

Divide observations into groups to construct frequency histogram.

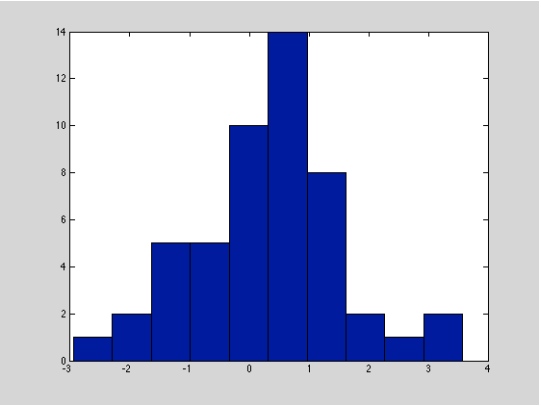
We can (generally) rely on computers to generate histograms

- Choose number of bins ≈ square root of sample size
- Use Excel, R, Minitab, SAS, MATLAB, etc

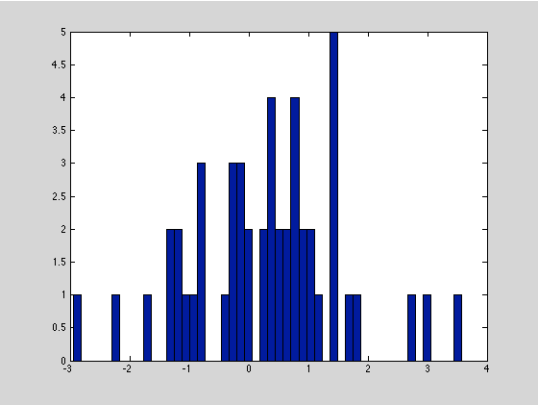
e.g. 50 samples, $\sqrt{50} = 7.07$



5 bins



10 bins



50 bins

Table 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

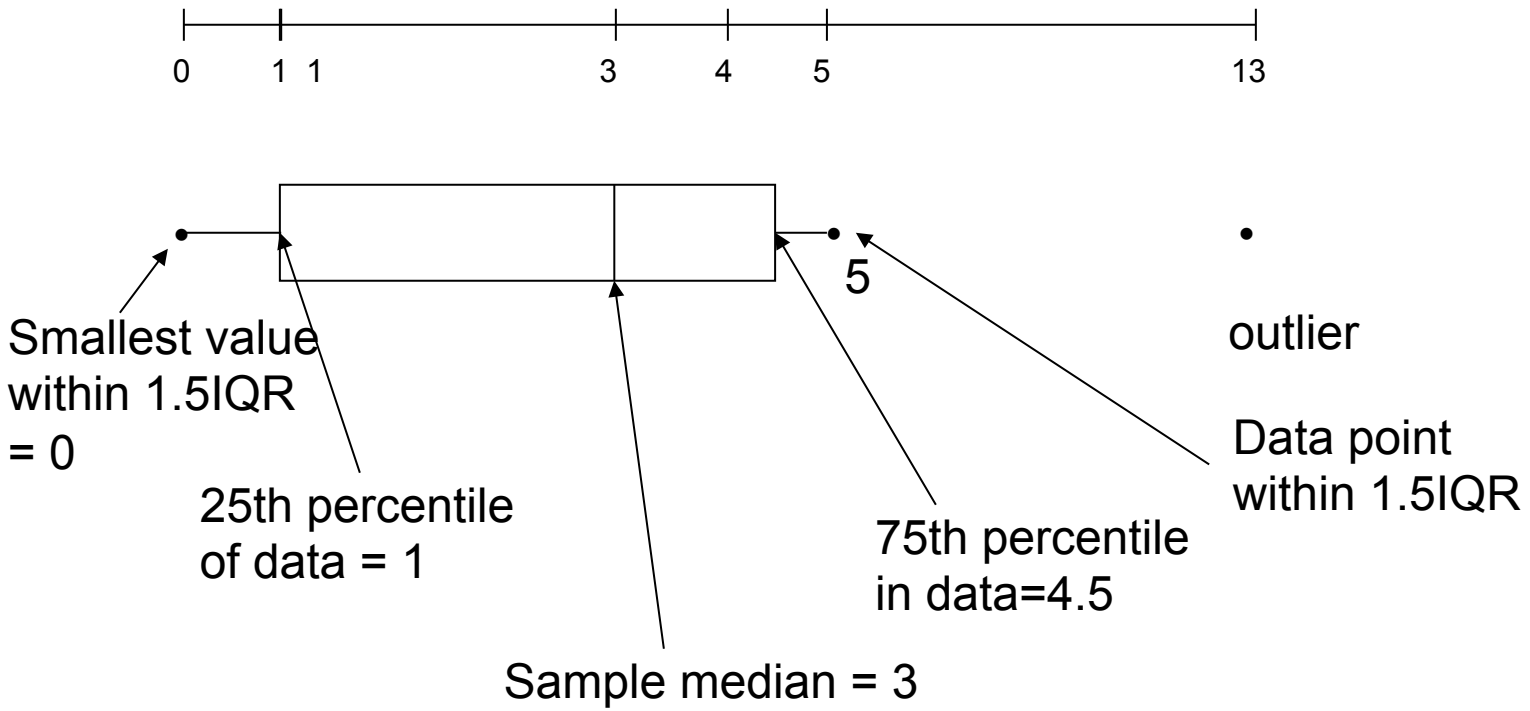
105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Table 6-4 Frequency Distribution for the Compressive Strength Data in Table 6-2

Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$
Frequency	2	3	6	14	22	17	10
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250

Box Plot

The box plot is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry.



Stem and Leaf Plots

Informative visual display of data.

Step 1: Select leading digits for the stem values. The trailing digit values become the leafs.

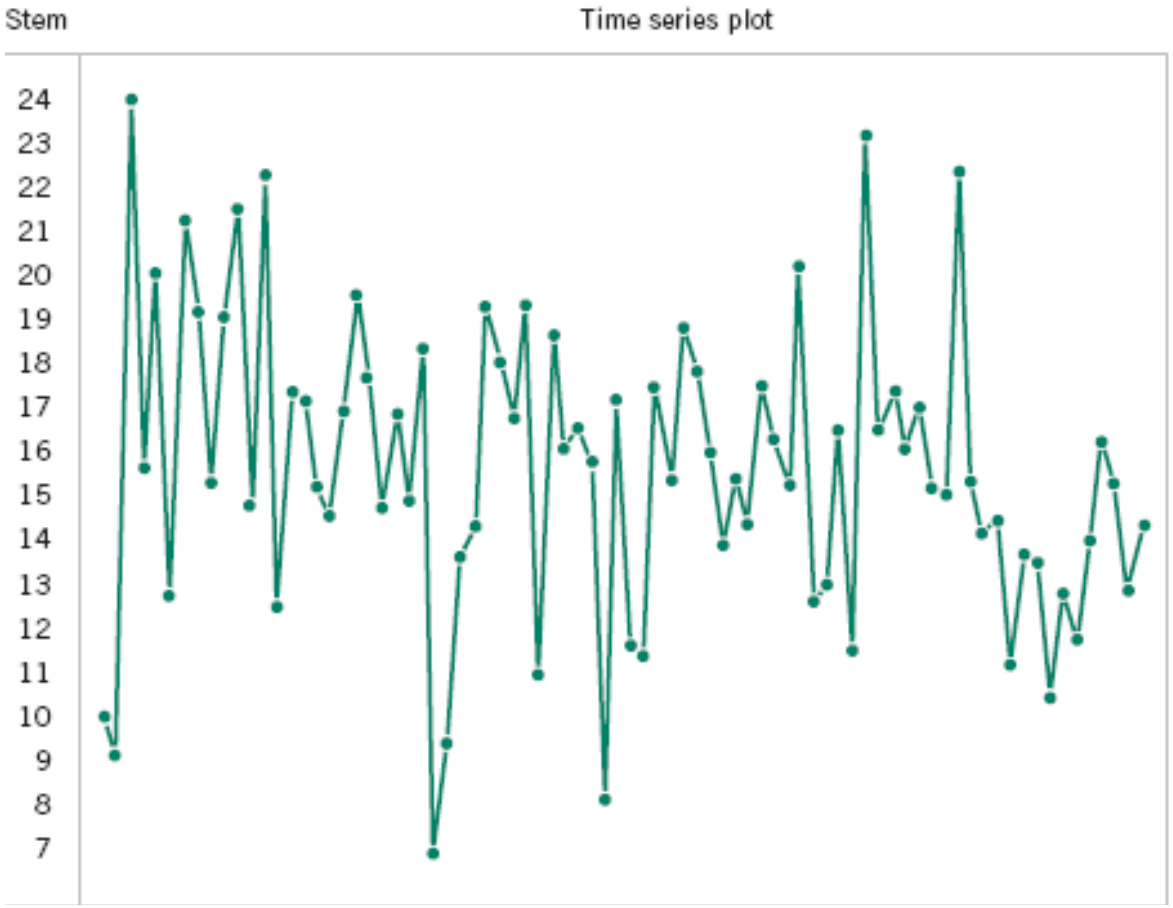
Step 2: List possible stem ordered values in vertical column.

Stem 3: Record the leafs for every observation besides the corresponding stem values.

e.g. breaks data set into 10 groups, according to first digit of observation:
Data = {33, 28, 16, 35, 11, 44, 33, 38}

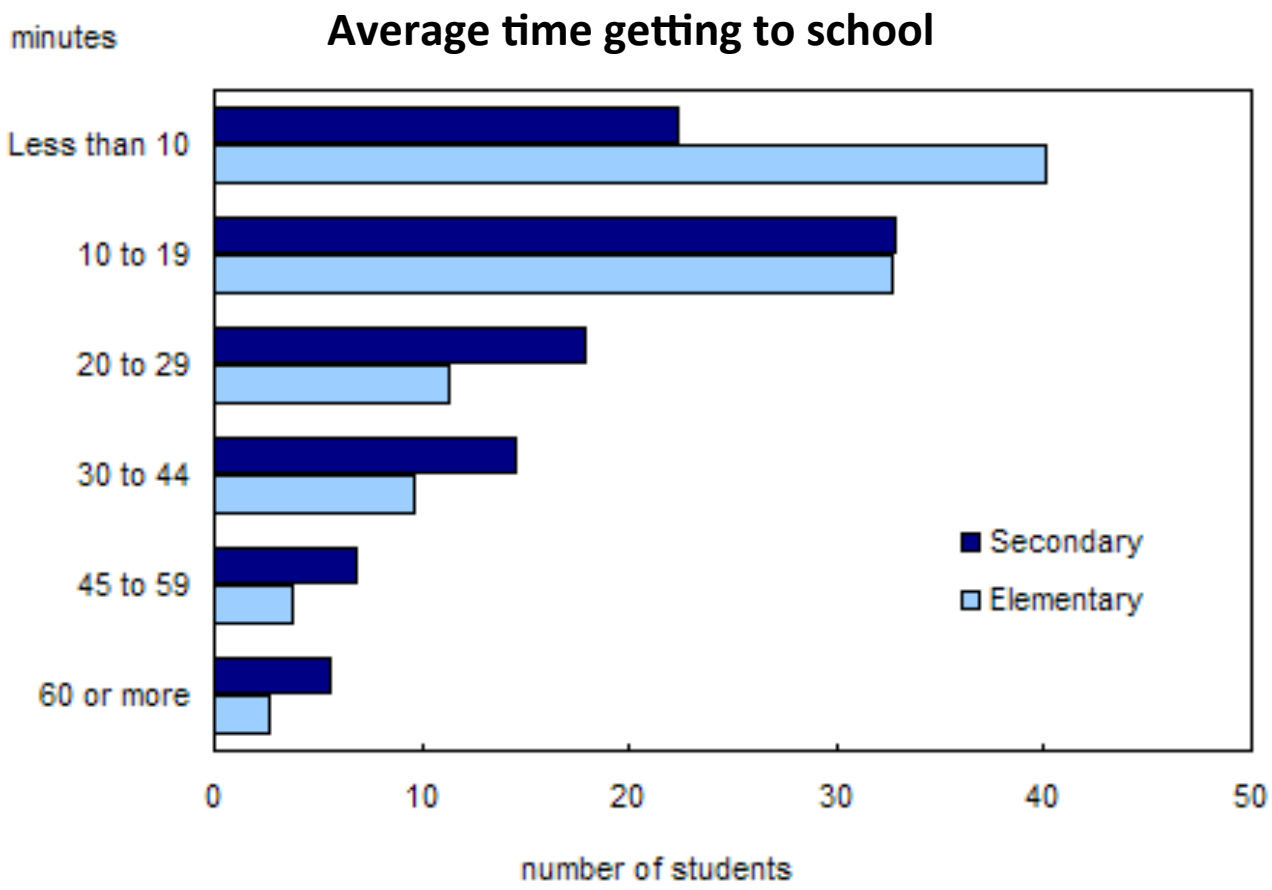
(Stem) 1st digit	(Leaf) 2nd digit
0	-
1	1, 6
2	8
3	3, 3, 5, 8
4	4

Time Series Plot

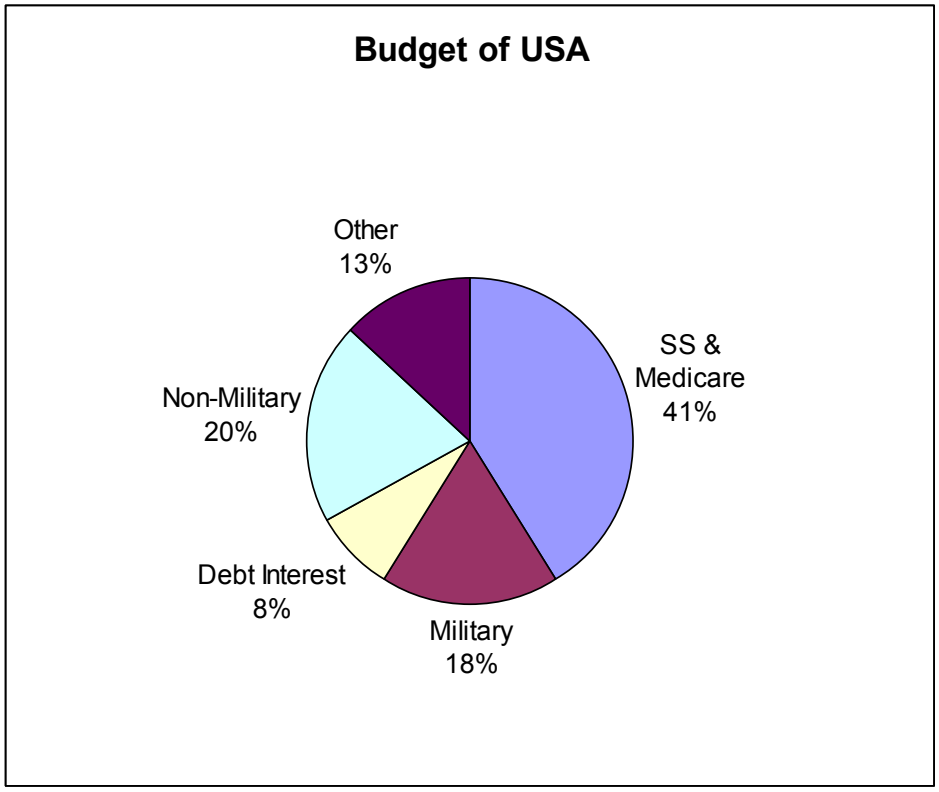


Time Series and Box Plots

Line Graphs


Average time getting to school
minutes
Less than 10
10 to 19
20 to 29
30 to 44
45 to 59
60 or more
Secondary
Elementary
0
10
20
30
40
50
number of students

Pie Chart – Federal Budget



Here is a pie chart of the USA Federal budget for 2004.

Budget – Another point of view