INTRODUCTION

Data Analysis involves several activities and assumptions:

- Data Set -> set of measurements of one or more attributes.
- Software items to be comparable

  eg: compare modules.
- We wish to determine the characteristics and relationships between attributes values.

INTRODUCTION

- Purpose of data analysis:

To make any patterns or relationships more visible and to make judgments about attributes.

ANALYZING THE RESULTS OF EXPERIMENTS

- Relevant data is collected , analyzed
- Items to consider in choosing the analysis techniques:
  - Nature of data
  - Purpose of experiment
  - Type of experimental design

NATURE OF DATA

- Sampling, population and data distribution
- Distribution of software measurements
- Statistical inference and hypothesis testing

SAMPLING, POPULATION AND DATA DISTRIBUTION

- Understand data as a sample from a larger population of all data gathered.
- From sample data, make decision about measured differences using statistical techniques
- Eg: In an experiment if we measure each subject once, the sample size is number of subjects.
- Sample statistics: describe and summarize measures obtained from a finite group of subjects
- Population parameters: represent values obtained if all possible subjects were measured.

We can measure the productivity of a group of programmers before and after training

DISTRIBUTION OF SOFTWARE MEASUREMENTS

- Techniques to assess whether distribution is normal:
- Robust statistical methods: yield meaningful results whether data is normal or not.
- Non parametric statistical techniques: take into account that data is not normal, and allow us to test various hypothesis about data set.

Statistical inference and hypothesis testing

- Statistical inference is the process of drawing conclusions about the population from observations about a sample
- Depends on data distribution
- Determine if a sample is a good representation of larger population based on two possible outcomes:

- Null Hypothesis H0:No real difference between treated subject and untreated ones.
- Alternative Hypothesis H1:measured difference indicate real treatment effects of independent variable.
- **Two errors:**
  - Type II error: accepting H0 when it is actually false.
  - Type I error: incorrectly rejecting Null hypothesis.

Purpose of experiment

- To confirm a theory
- To explore a relationship

CONFIRMING A THEORY

- Explore truth of a theory
- Uses ANOVA-analysis of variance.
- Consider two populations, one that uses old technique and the one that uses new.
- Perform statistical test.
- Analyze variance between 2 sets of data to see if they come from one population or two.
- Student's test: to compare groups
- F statistic: for more than 2 groups

Exploring a relationship

- Among data points describing one variable or across multiple variables.
- Three techniques:

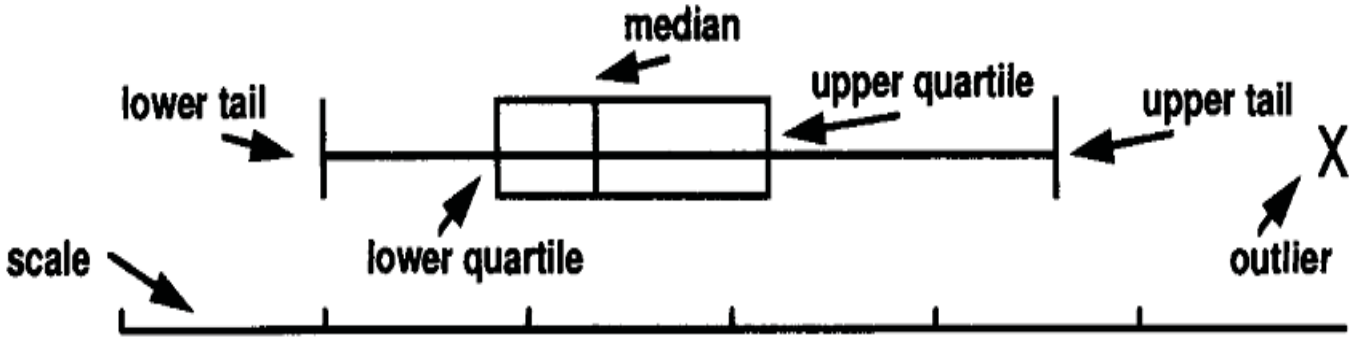
1. Box plot
2. Scatter plot
3. Correlation analysis

TECHNIQUES

- Box plot:
- Depict summary of range of a set of data. shows where most of data is clustered and where any outlier data may be.
- Scatter plot:
- Depict relationship between two variables by viewing their relative positions of pairs of data points.
- Correlation analysis:
- Uses statistical methods to confirm whether there is a true relationship between two attributes.

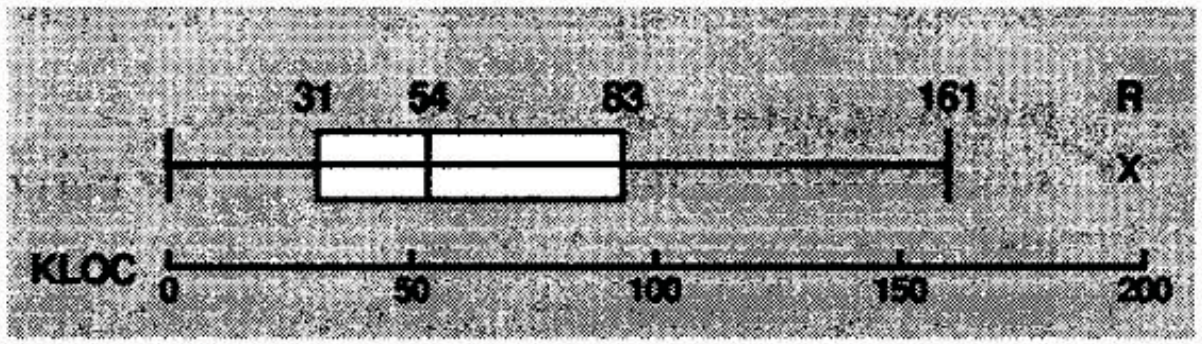
BOX PLOTS

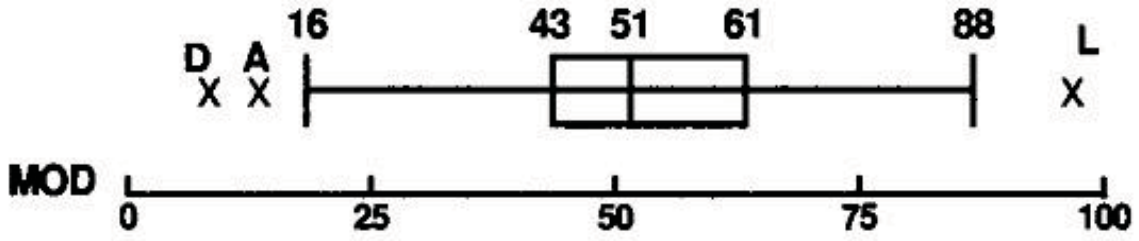
- Uses median, quartile to define central location and spread of values.
- Three summary statistics:

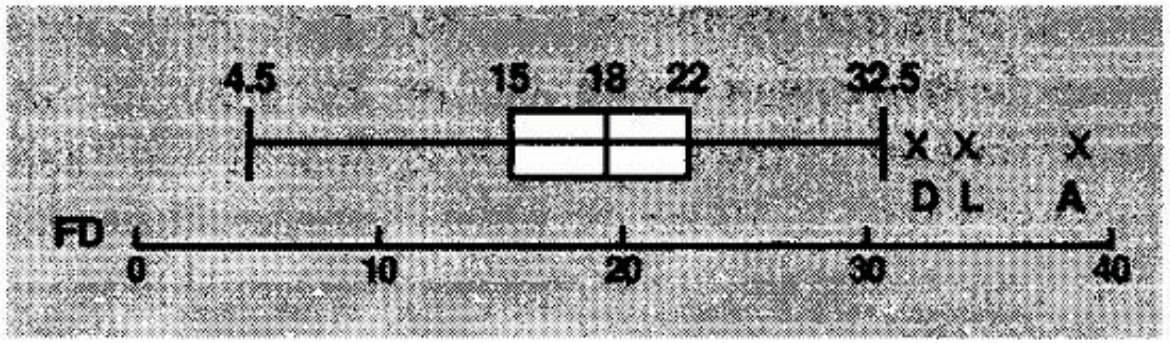


EXAMPLE

System	KLOC	MOD	FD
A	10	15	36
B	23	43	22
C	26	61	15
D	31	10	33
E	31	43	15
F	40	57	13
G	47	58	22
H	52	65	16
I	54	50	15
J	67	60	18
K	70	50	10
L	75	96	34
M	83	51	16
N	83	61	18
P	100	32	12
Q	110	78	20
R	200	48	21


31
54
83
161
R
X
KLOC
0
50
100
150
200

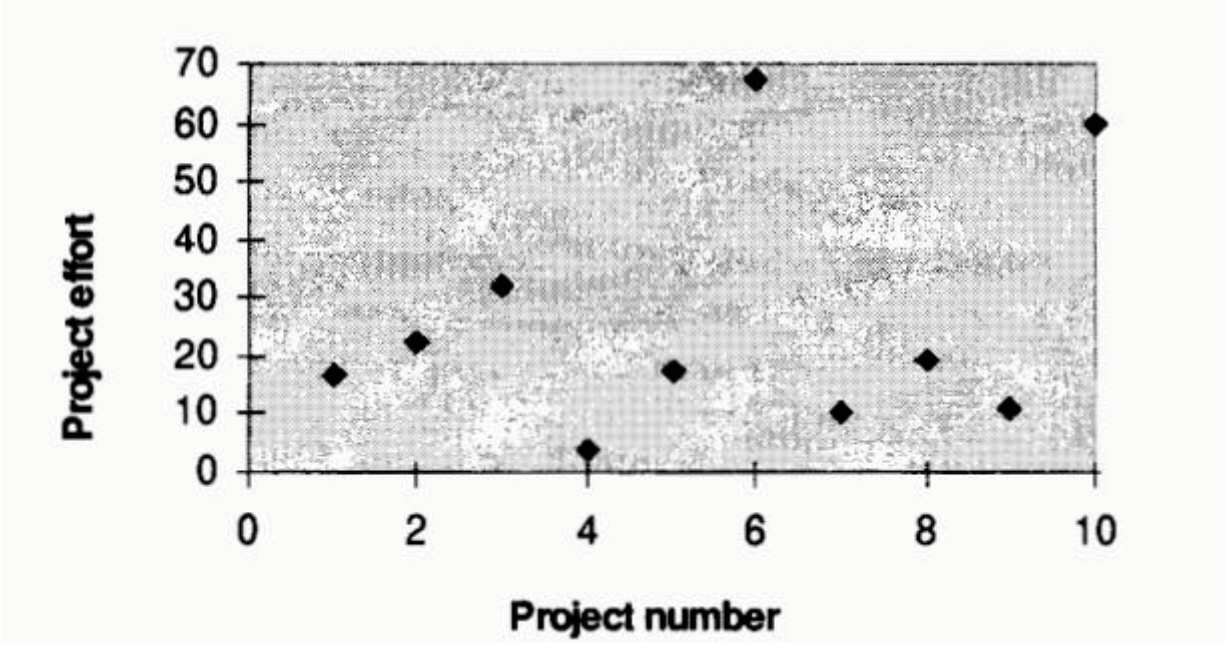

16
43
51
61
88
L
D
A
X
X
X
MOD
0
25
50
75
100


4.5
15
18
22
32.5
X
X
X
D
L
A
FD
0
10
20
30
40

- Examine all systems whose MOD values are outliers in box plot, since these are the ones most likely to be fault prone->D,L,A

SCATTER PLOTS

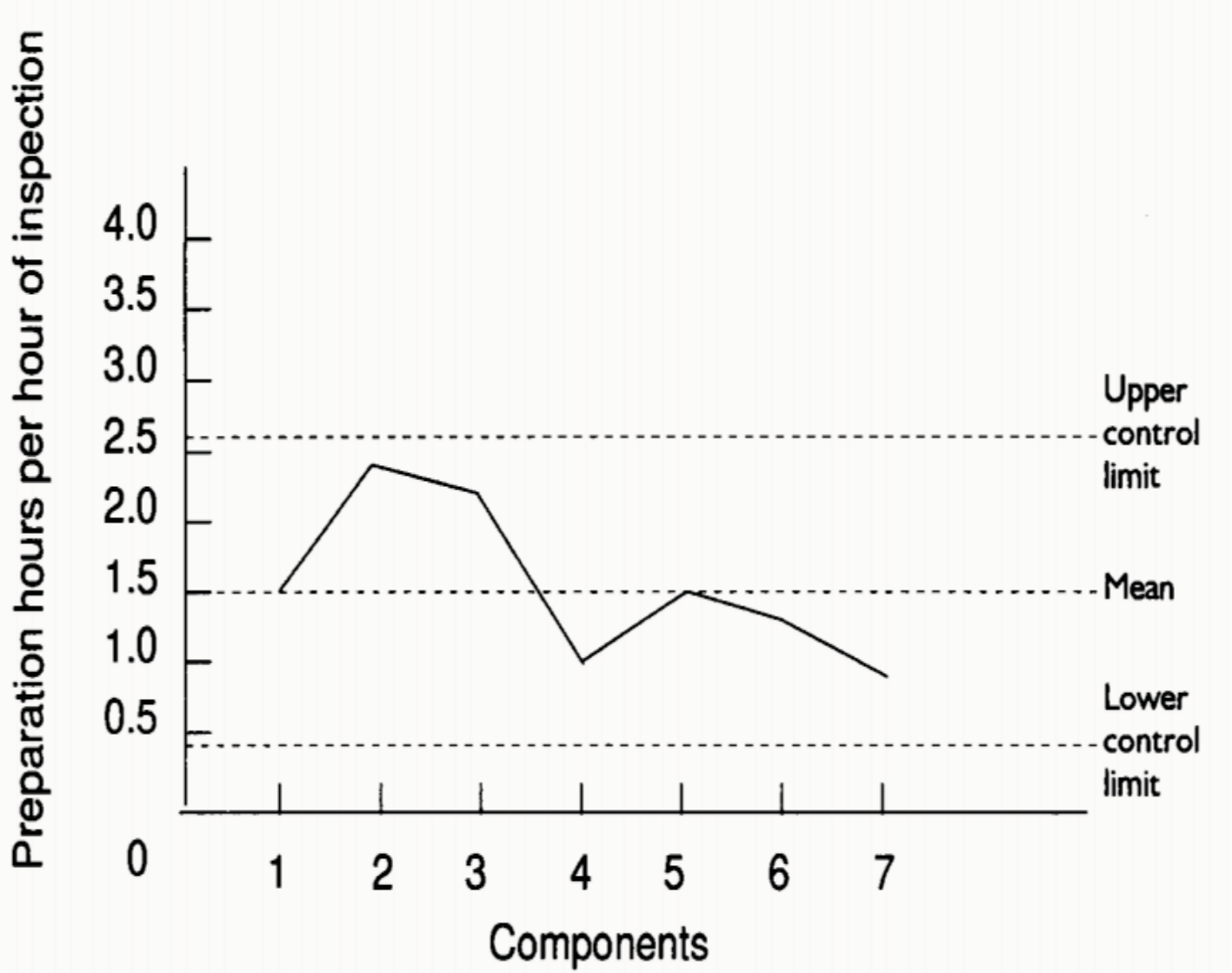
- Box plot hides expected behavior and shows us what is unusual.
- Graph all data to see what patterns may be.
- Relationship between one and other attributes.
- Used for planning purposes.



CONTROL CHARTS

- Help us to check whether data is in the acceptable bounds.
- We can decide to take action to prevent problems before they occur.
- Calculate the mean, standard deviation and two control limits.
- Upper control limit is equal to two std deviations above mean.
- Lower control limit is equal to two std deviations below mean.

Component number	Preparation hours/ inspection hours
1	1.5
2	2.4
3	2.2
4	1.0
5	1.5
6	1.3
7	1.0
Mean	1.6
Standard deviation	0.5
Upper control limit (UCL)	2.6
Lower control limit (LCL)	0.4


Preparation hours per hour of inspection
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
1
2
3
4
5
6
7
Components
Upper control limit
Mean
Lower control limit