Basic Statistics and Probability Theory

in

Civil, Surveying and Environmental

Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zürich, Switzerland

Contents of Todays Lecture

- Overview of descriptive statistics

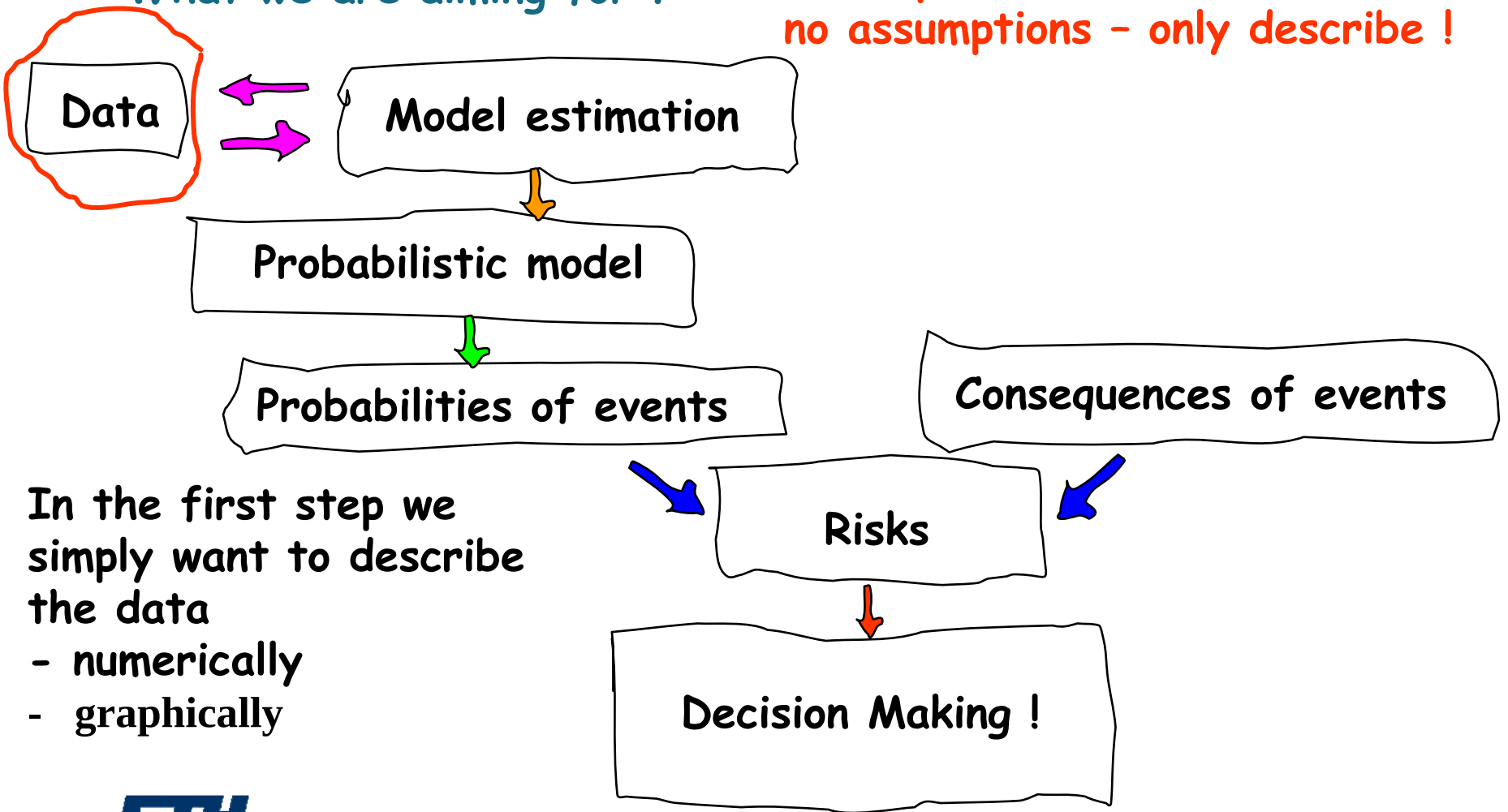
- Numerical summaries
  - Central measures
  - Dispersion measures
  - Other measures
  - Measures of correlation

- Graphical representations
  - One-dimensional scatter plots
  - Histograms
  - Quantile plots
  - Tukey Box plots
  - Q-Q plots and Tukey mean-difference plot

Overview of Descriptive Statistics

- What we are aiming for ?

Descriptive statistics make no assumptions – only describe !



In the first step we simply want to describe the data
- numerically
- graphically

Numerical Summaries

- **Central measures:**

Sample mean : $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

If one number should be given to represent a data set typically the sample mean would be chosen

Median : The 0.5 quantile (obtained from ordered data sets, see quantile plots)

Mode : Most frequent value – obtained from histograms

Numerical Summaries

- **Dispersion measures:**

Sample variance: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ s: standard deviation

Indicator of variability around the sample mean

Sample coefficient of variation (CoV): $\nu = \frac{s}{\bar{x}}$

Indicator of variability relative to the sample mean

Numerical Summaries

- **Other measures:**

Sample skewness:

$$\eta = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

Measure of symmetry

Sample kurtosis

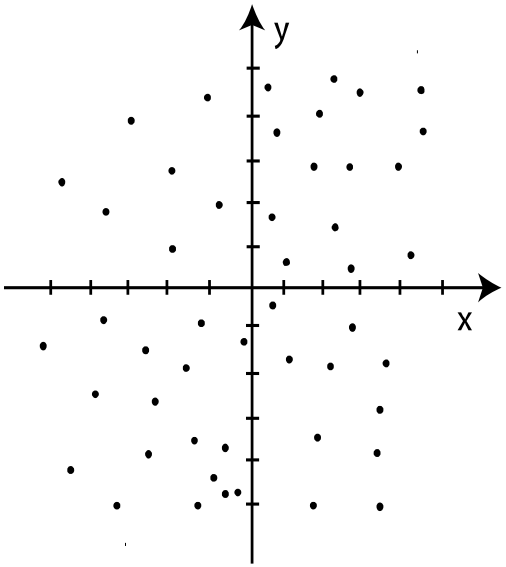
$$\kappa = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4}$$

Measure of peakedness

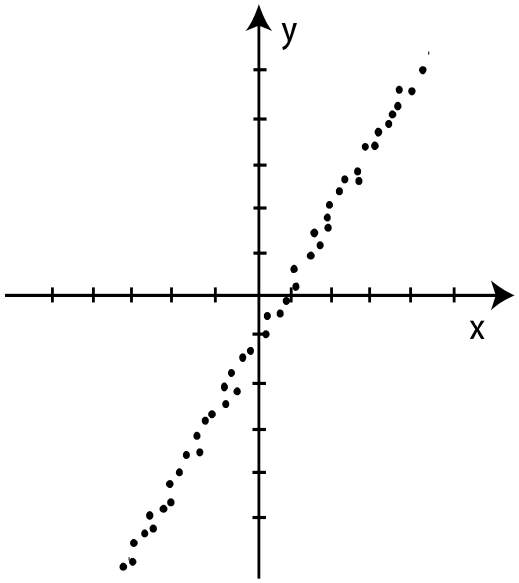
Numerical Summaries

- Measures of correlation (linear dependency between data pairs):

2-dimensional scatter plots



Almost no dependency



Almost full dependency

Numerical Summaries

- **Measures of correlation (linear dependency between data pairs):**

Sample covariance: $$s^2_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})$$

The sum will get positive contributions in case of low-low or high-high data pairs

Sample coefficient of correlation: $$r_{XY} = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{s_X \cdot s_Y}$$

r_{XY} is limited in the interval -1 to +1

Numerical Summaries

- **Summary:**

Central measures:

- sample mean value:	The center of gravity of a data set
- sample median:	The mid value of a data set
- sample mode:	The most frequent value/range of a data set

Dispersion measures:

- sample variance:	The distribution around the sample mean
- sample CoV:	The variability relative to the sample mean

Other measures:

- sample skewness:	The skewness relative to the sample mean
- sample kurtosis:	The peakedness around the sample mean

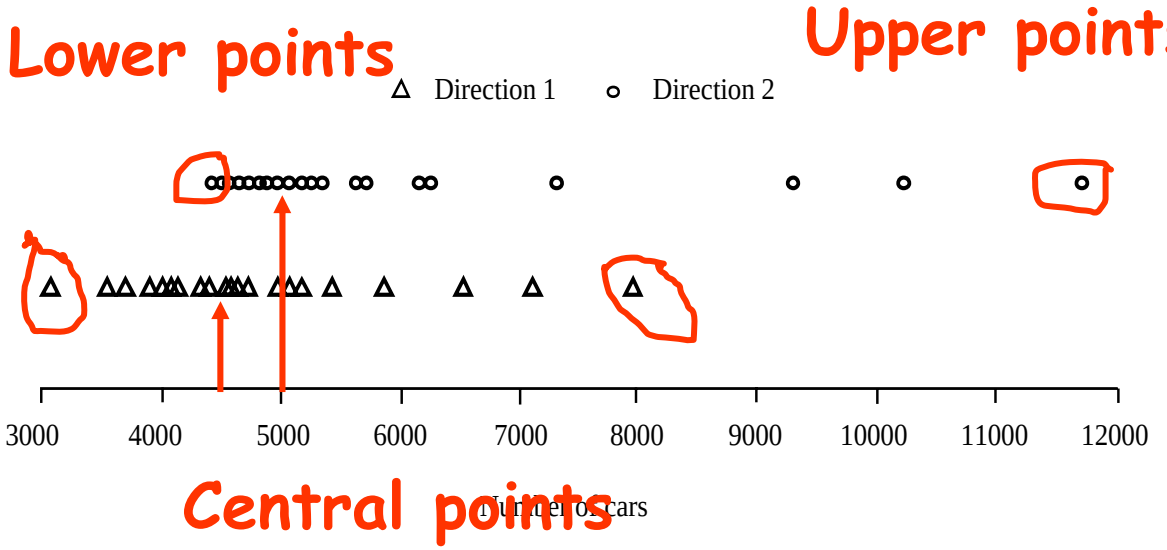
Measures of correlation:

- sample covariance:	Tendency for high-high, low-low and high-low pairs in two data sets
- sample coefficient of correlation :	Normalized coefficient between -1 and +1

Graphical Representations

- Assume that we have a set of data (observations of road way traffic)

The simplest representation of the data is the one-dimensional scatter plot



Date	Direction 1		Direction 2	
	Unordered	Ordered	Unordered	Ordered
01.01	3087	3087	3677	3677
02.01	4664	3578	7357	4453
03.01	4164	3710	9323	4480
04.01	3710	3737	11748	4560
05.01	4029	3906	10256	4635
06.01	4323	4029	4453	4648
07.01	4041	4041	4815	4672
08.01	3737	4085	4757	4757
09.01	4103	4103	4672	4791
10.01	5457	4164	5401	4815
11.01	4563	4323	5688	4880
12.01	3906	4359	6308	4928
13.01	4419	4366	4946	4946
14.01	4359	4368	4635	5005
15.01	4667	4371	5100	5013
16.01	5098	4419	4791	5100
17.01	6551	4563	5235	5220
18.01	4371	4588	4560	5235
19.01	3578	4664	5729	5281
20.01	4366	4667	5005	5318
21.01	4368	4727	4480	5398
22.01	4588	4739	4880	5401
23.01	5001	4741	4928	5679
24.01	7118	5001	5398	5688
25.01	4727	5098	4648	5729
26.01	4085	5193	6183	6183
27.01	4741	5457	5220	6308
28.01	4739	5892	5013	7357
29.01	5193	6551	5281	9323
30.01	5892	7118	5318	10256
31.01	7974	7974	5679	11748

Graphical Representations

- **Histograms**

The data are grouped into intervals

Date	Direction 1		Direction 2	
	Unordered	Ordered	Unordered	Ordered
01.01	3087	3087	3677	3677
02.01	4664	3578	7357	4453
03.01	4164	3710	9323	4480
04.01	3710	3737	11748	4560
05.01	4029	3906	10256	4635
06.01	4323	4029	4453	4648
07.01	4041	4041	4815	4672
08.01	3737	4085	4757	4757
09.01	4103	4103	4672	4791
10.01	5457	4164	5401	4815
11.01	4563	4323	5688	4880
12.01	3906	4359	6308	4928
13.01	4419	4366	4946	4946
14.01	4359	4368	4635	5005
15.01	4667	4371	5100	5013
16.01	5098	4419	4791	5100
17.01	6551	4563	5235	5220
18.01	4371	4588	4560	5235
19.01	3578	4664	5729	5281
20.01	4366	4667	5005	5318
21.01	4368	4727	4480	5398
22.01	4588	4739	4880	5401
23.01	5001	4741	4928	5679
24.01	7118	5001	5398	5688
25.01	4727	5098	4648	5729
26.01	4085	5193	6183	6183
27.01	4741	5457	5220	6308
28.01	4739	5892	5013	7357
29.01	5193	6551	5281	9323
30.01	5892	7118	5318	10256
31.01	7974	7974	5679	11748

Interval (Number of cars $*10^2$)	Interval Midpoint (Number of cars $*10^2$)	Number of observations	Frequency [%]	Cumulative frequency
30-35	32.5	0	0.0000	0.0000
35-40	37.5	1	3.2258	0.0323
40-45	42.5	2	6.4516	0.0968
45-50	47.5	10	32.2581	0.4194
50-55	52.5	9	29.0323	0.7097
55-60	57.5	3	9.6774	0.8065
60-65	62.5	2	6.4516	0.8710
65-70	67.5	0	0.0000	0.8710
70-75	72.5	1	3.2258	0.9032
75-80	77.5	0	0.0000	0.9032
80-85	82.5	0	0.0000	0.9032
85-90	87.5	0	0.0000	0.9032
90-95	92.5	1	3.2258	0.9355
95-100	97.5	0	0.0000	0.9355
100-105	102.5	1	3.2258	0.9677
105-110	107.5	0	0.0000	0.9677
110-115	112.5	1	3.2258	1.0000

$= \frac{1}{31} \cdot 100$

$\Sigma = 31$

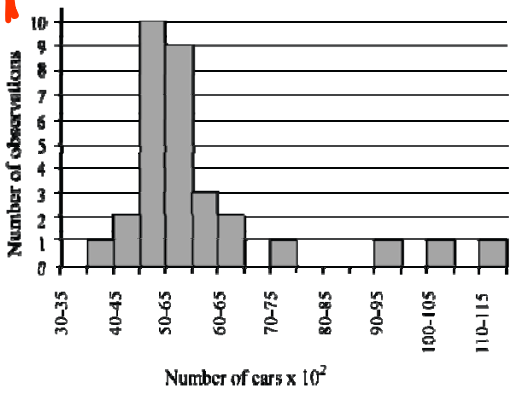
Graphical Representations

- Histograms

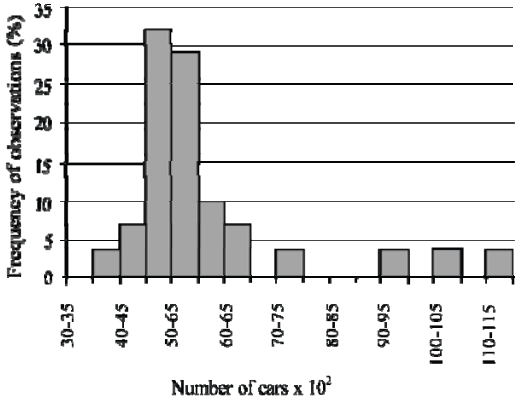
The grouped data are plotted

mode

Interval (Number of cars *10^2)	Interval Midpoint (Number of cars *10^2)	Number of observations	Frequency [%]	Cumulative frequency
30-35	32.5	0	0.0000	0.0000
35-40	37.5	1	3.2258	0.0323
40-45	42.5	2	6.4516	0.0968
45-50	47.5	10	32.2581	0.4194
50-55	52.5	9	29.0323	0.7097
55-60	57.5	3	9.6774	0.8065
60-65	62.5	2	6.4516	0.8710
65-70	67.5	0	0.0000	0.8710
70-75	72.5	1	3.2258	0.9032
75-80	77.5	0	0.0000	0.9032
80-85	82.5	0	0.0000	0.9032
85-90	87.5	0	0.0000	0.9032
90-95	92.5	1	3.2258	0.9355
95-100	97.5	0	0.0000	0.9355
100-105	102.5	1	3.2258	0.9677
105-110	107.5	0	0.0000	0.9677
110-115	112.5	1	3.2258	1.0000



Simple histogram



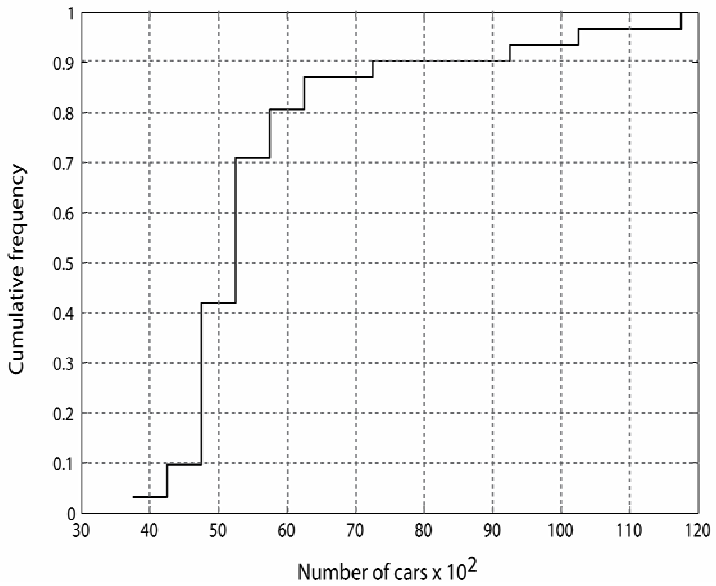
Frequency distribution

Graphical Representations

- **Histograms**

  The grouped data are plotted

Interval (Number of cars *10^2)	Interval Midpoint (Number of cars *10^2)	Number of observations	Frequency [%]	Cumulative frequency
30-35	32.5	0	0.0000	0.0000
35-40	37.5	1	3.2258	0.0323
40-45	42.5	2	6.4516	0.0968
45-50	47.5	10	32.2581	0.4194
50-55	52.5	9	29.0323	0.7097
55-60	57.5	3	9.6774	0.8065
60-65	62.5	2	6.4516	0.8710
65-70	67.5	0	0.0000	0.8710
70-75	72.5	1	3.2258	0.9032
75-80	77.5	0	0.0000	0.9032
80-85	82.5	0	0.0000	0.9032
85-90	87.5	0	0.0000	0.9032
90-95	92.5	1	3.2258	0.9355
95-100	97.5	0	0.0000	0.9355
100-105	102.5	1	3.2258	0.9677
105-110	107.5	0	0.0000	0.9677
110-115	112.5	1	3.2258	1.0000



Cumulative frequency distribution

Graphical Representations

- **Histograms**

  The number of intervals selected will influence the information maintained

No general rule can be given but some suggest the following

$$k = 1 + 3.3 \log n$$

k: **number of intervals**
n: **number of data**

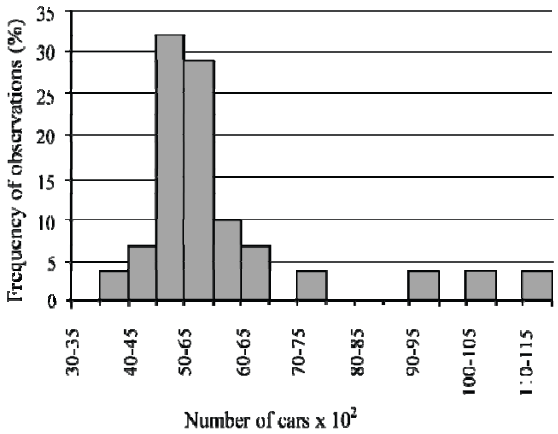
For the traffic flow data set:
$k = 1 + 3.3 \log 31 = 5.92 = 6$

Graphical Representations

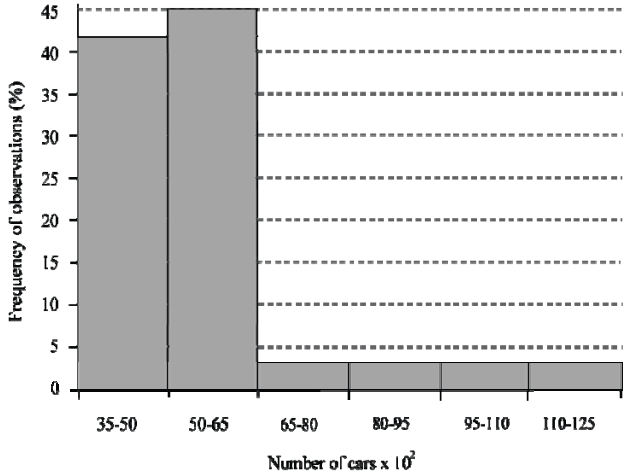
- Histograms

The number of intervals selected will influence the information maintained

$k=17$



$k=6$



Graphical Representations

- Quantile plots

Definition : the Q-quantile corresponds to the value in a data set which is exceeded by 100% - Q x 100% of the data

e.g. the 0.75 quantile is exceeded by 100% - 0.75 x 100% = 25% of the data

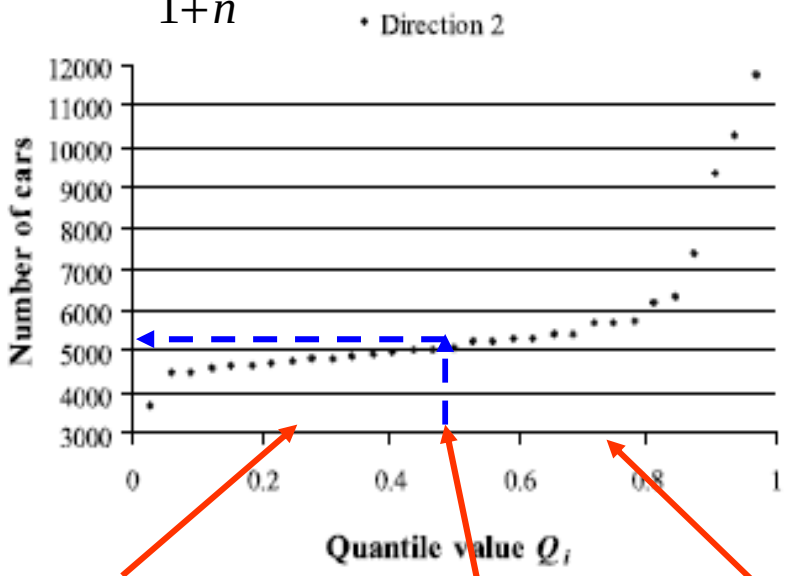
Quantile plots are generated by plotting the data against their quantile values

Graphical Representations

- Quantile plots

The quantiles are calculated from the ordered data set as:

$$Q_i = \frac{i}{1+n}$$



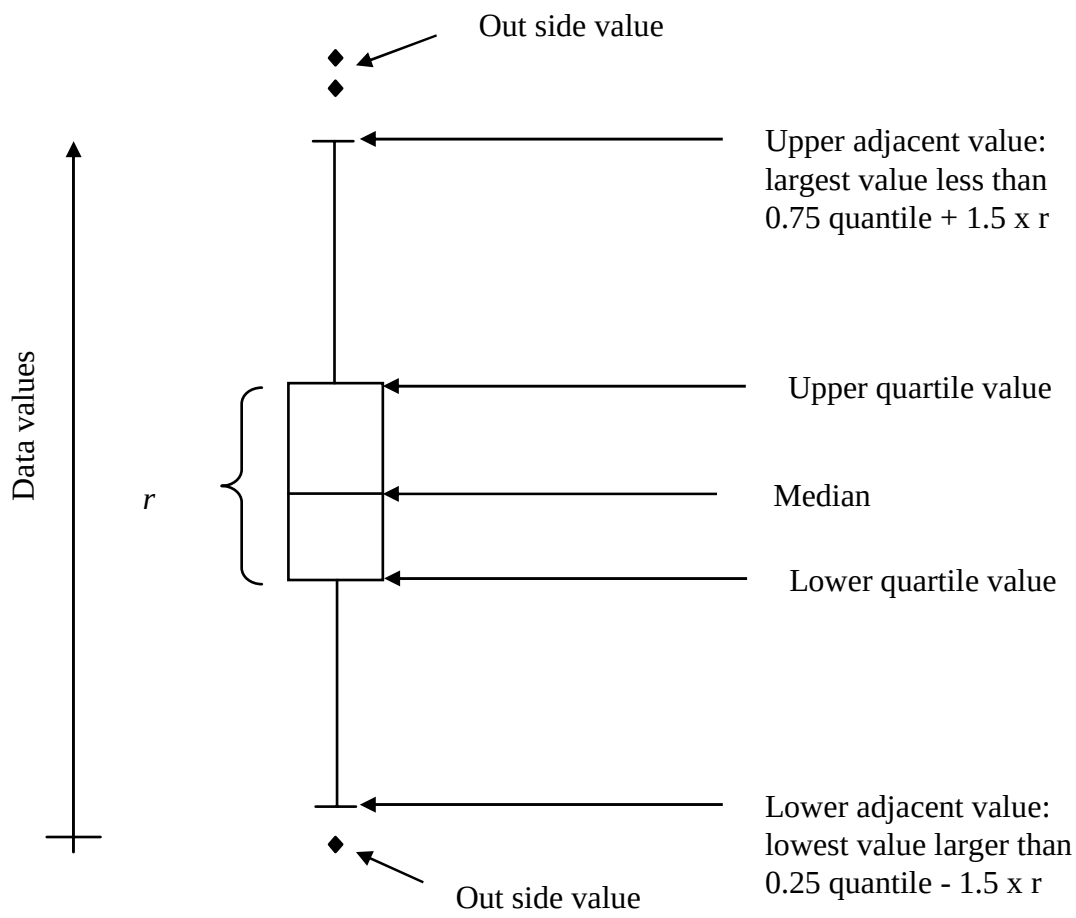
x_i	Direction 1	Direction 2	Q_i quantile
1	3087	3677	0.0313
2	3578	4453	0.0625
3	3710	4480	0.0938
4	3737	4560	0.1250
5	3906	4635	0.1563
6	4029	4648	0.1875
7	4041	4672	0.2188
8	4085	4757	0.2500
9	4103	4791	0.2813
10	4164	4815	0.3125
11	4323	4880	0.3438
12	4359	4928	0.3750
13	4366	4946	0.4063
14	4368	5005	0.4375
15	4371	5013	0.4688
16	4419	5100	0.5000
17	4563	5220	0.5313
18	4588	5235	0.5625
19	4664	5281	0.5938
20	4667	5318	0.6250
21	4727	5398	0.6563
22	4739	5401	0.6875
23	4741	5679	0.7188
24	5001	5688	0.7500
25	5098	5729	0.7813
26	5193	6183	0.8125
27	5457	6308	0.8438
28	5892	7357	0.8750
29	6551	9323	0.9063
30	7118	10256	0.9375
31	7974	11748	0.9688

Lower quartile = 0.25 quantile value

Median = 0.5 quantile value

Upper quartile = 0.75 quantile value

Graphical Representations



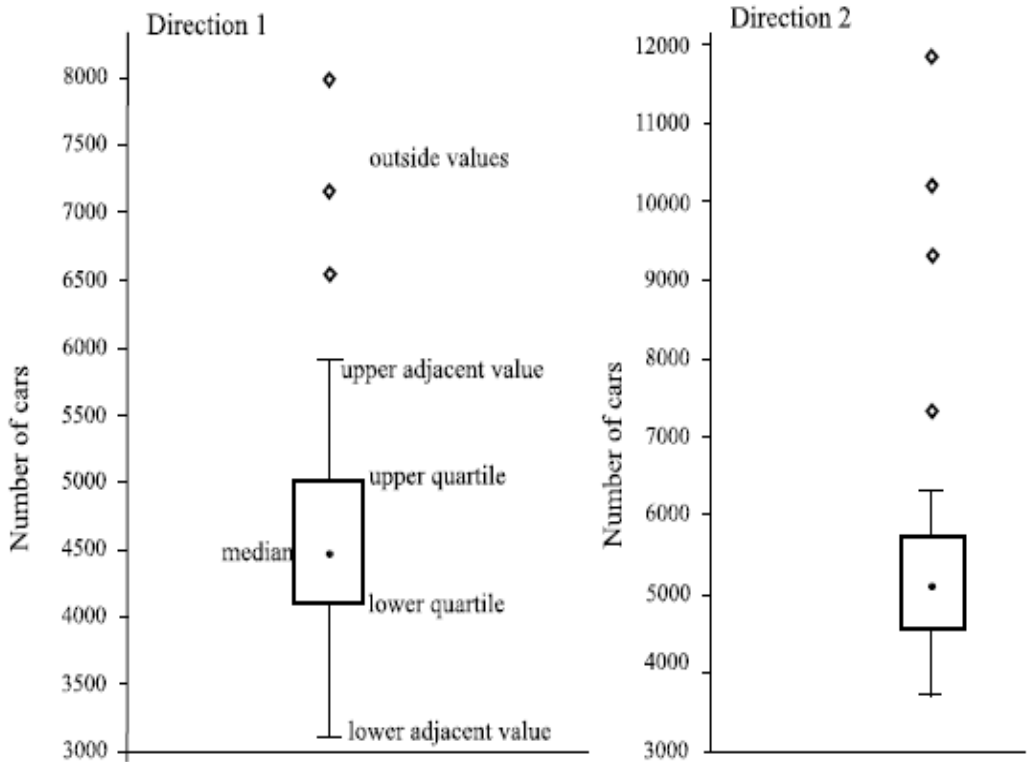
r : Inter-quartile range (50% of data)

Graphical Representations

- Tukey Box plots (traffic data)

Statistic
Lower adjacent value
Lower quartile
Median
Upper quartile
Upper adjacent value
Outside values

Direction 1	Direction 2
3087	3677
4085	4757
4419	5100
5001	5688
5892	6308
6551	7357
7118	9323
7974	10256
	11748

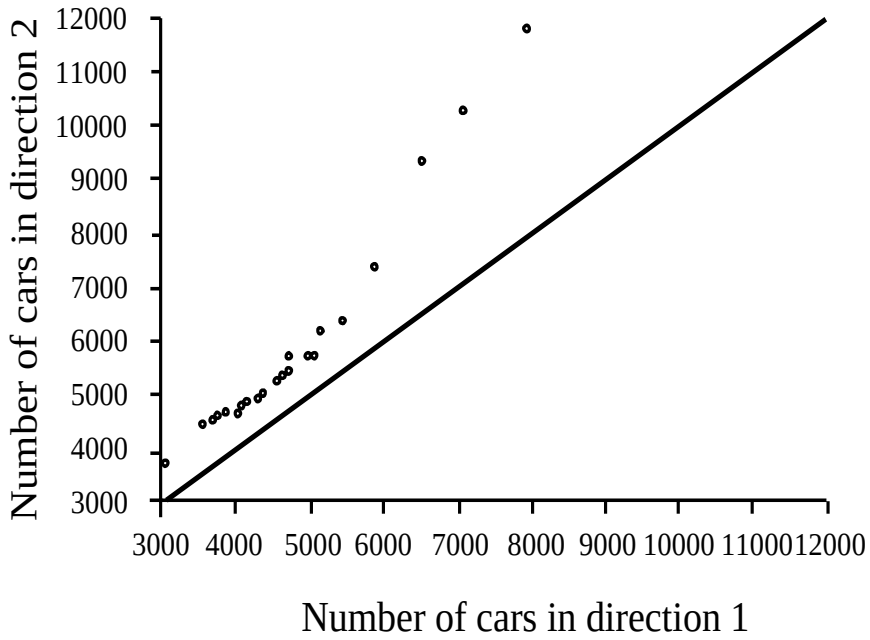


Graphical Representations

- Q-Q plots

Q-Q plots are produced to represent and compare 2 data sets

Data points of the two data sets with the same quantile values are plotted against each other



Graphical Representations

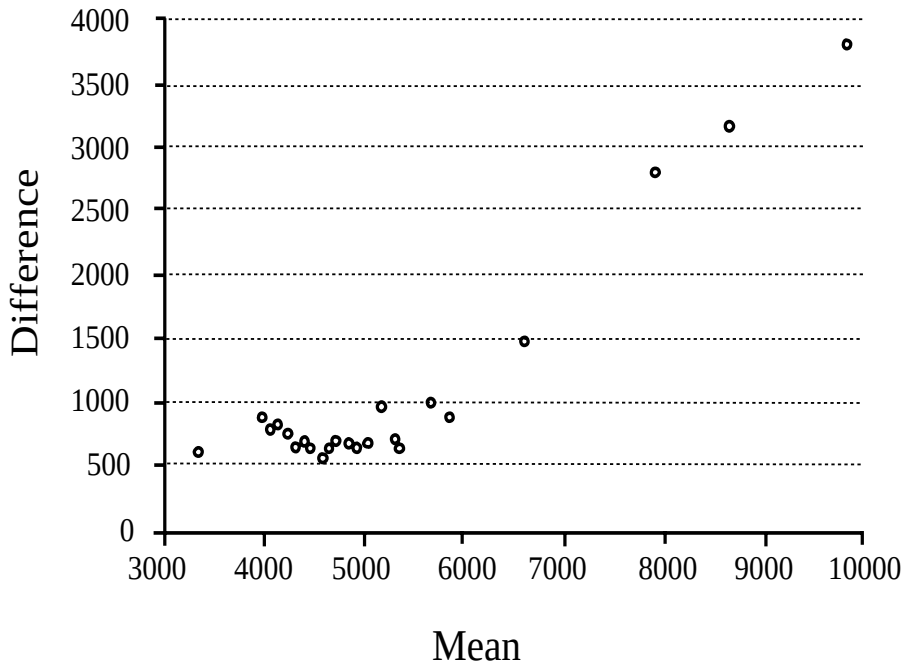
- Mean vs. difference plots

Mean vs. difference plots are produced to represent and compare 2 data sets

$$(y_i + x_i)/2$$

is plotted against

$$y_i - x_i$$



Graphical Representations

- Summary

One-dimensional scatter plots :	illustrate the range and distribution of a data sets along one axis, indicate symmetry.
Histograms:	illustrate how the data are distributed over the range of data, indicate mode and symmetry.
Quantile plots: symmetry	Illustrate median, distribution and
Tukey – Box plots:	Illustrate median, upper/lower quartiles, symmetry and distribution
Q-Q plots:	Compare two data set, relative shapes
Mean vs. difference plots:	Compare two data sets, relative shapes