Probability and Statistics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 6: A GENTLE INTRODUCTION TO EVERYDAY STATISTICS

1Introduction

1.1 Difference between probability and statistics

1.2 Relevant questions in a statistics context

1.3 Relation between descriptive and inferential statistics

1.4 General flow in statistics

1.5 Statistics from an engineering perspective

2 Looking at data

2.1 Exploratory Data Analysis (EDA)

2.2 Graphical representation of a single variable (univariate)

2.3 Graphical representation of two variables (bivariate)

2.4 Graphical representation of multiple variables (multivariate)

2.5 Things to look out for

3 Highlight: Graphical techniques according to problem identification

3.1 Univariate

3.2 Regression

3.3 One factor

3.4 Multi-factor / Comparative

3.5 Multi-factor / Screening

3.6 Interlab

3.7 Multivariate

1 Introduction

1.1 Difference between probability and statistics

Probability

- In probability, we start with a model describing what events we think are going to occur, with what likelihoods.
- The events may be random, in the sense that we don't know for sure what will happen next, but we do quantify our degree of surprise when various things happen.

- The standard example is flipping a fair coin. “Fair” means, technically, that the probability of heads on a given flip is 50%, and the probability of tails on a given flip is 50%. This doesn't mean that every other flip will give a head — after all, three heads in a row is no surprise. Five heads in a row would be more surprising, and when you've seen twenty heads in a row you're sure that something fishy is going on. What the 50% probability of heads does mean is that, as the number of flips increases, we expect the number of heads to approach half the number of flips. Seven heads on ten flips is no surprise; 700,000 heads on 1,000,000 tosses is highly unlikely.
- Another example would be flipping an unfair coin, where we know ahead of time that there's a 60% chance of heads on each toss, and a 40% chance of tails.
- A third example would be rolling a loaded die, where (for example) the chances of rolling 1, 2, 3, 4, 5, or 6 are 25%, 5%, 20%, 20%, 20%, and 10%, respectively. Given this setup, you'd expect rolling three 1's in a row to be much more likely than rolling three 2's in a row.

Probability

- As these examples illustrate, the probabilist starts with a probability model (something which assigns various percentage likelihoods of different things happening), then tells us which things are more and less likely to occur.

- Key points about probability:

1. Rules → data: Given the rules, describe the likelihoods of various events occurring.
2. Probability is about prediction — looking forward.
3. Probability is mathematics.

Statistics

- The statistician turns this around:

1. Rules ← data: Given only the data, try to guess what the rules were. That is, some probability model controlled what data came out, and the best we can do is guess — or approximate — what that model was. We might guess wrong; we might refine our guess as we get more data.
2. Statistics is about looking backward.
3. Statistics is an art. It uses mathematical methods, but it is more than math.
4. Once we make our best *statistical* guess about what the probability model is (what the rules are), based on looking *backward*, we can then use that *probability* model to predict the *future*. (In a sense, probability doesn't need statistics, but statistics uses probability.)

Statistics

- Example: suppose you are given a list of heads and tails. You, as the statistician, are in the following situation:

    - You do not know ahead of time that the coin is fair. Maybe you've been hired to decide whether the coin is fair (or, more generally, whether a gambling house is committing fraud).
    - You may not even know ahead of time whether the data come from a coin-flipping experiment at all.

    o Suppose the data are three heads. Your first guess might be that a fair coin is being flipped, and these data don't contradict that hypothesis. Based on these data, you might hypothesize that the rules governing the experiment are that of a fair coin: your probability model for predicting the future is that heads and tails each occur with 50% likelihood.

- If there are ten heads in a row, though, or twenty, then you might start to reject that hypothesis and replace it with the hypothesis that the coin has heads on both sides. Then you'd predict that the next toss will certainly be heads: your new probability model for predicting the future is that heads occur with 100% likelihood, and tails occur with 0% likelihood.
- If the data are “heads, tails, heads, tails, heads, tails”, then again, your first fair-coin hypothesis seems plausible. If on the other hand you have heads alternating with tails not three pairs but 50 pairs in a row, then you reject that model. It begins to sound like the coin is not being flipped in the air, but rather is being flipped with a spatula. Your new probability model is that if the previous result was tails or heads, then the next result is heads or tails, respectively, with 100% likelihood.

- For a historical overview of probability and statistics, the following reference is a good starting point:

  http://www.economics.soton.ac.uk/staff/aldrich/Figures.htm

1.2 Relevant questions in a statistics context

- Related to understanding the intuition behind concepts [Chapters 1-4]
  - What is the difference between "statistic" and "parameter"? [What is a sampling distribution?]

Statistic: characteristic of a sample

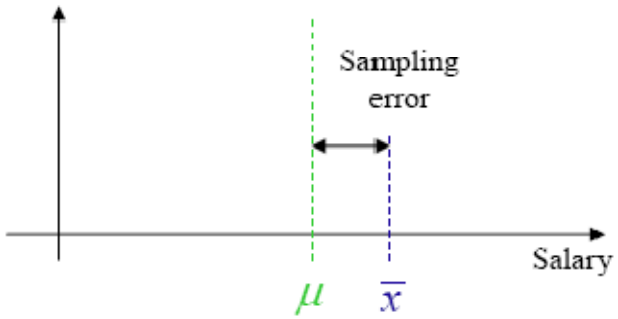
What is the average salary of 2000 people randomly sampled in Spain?

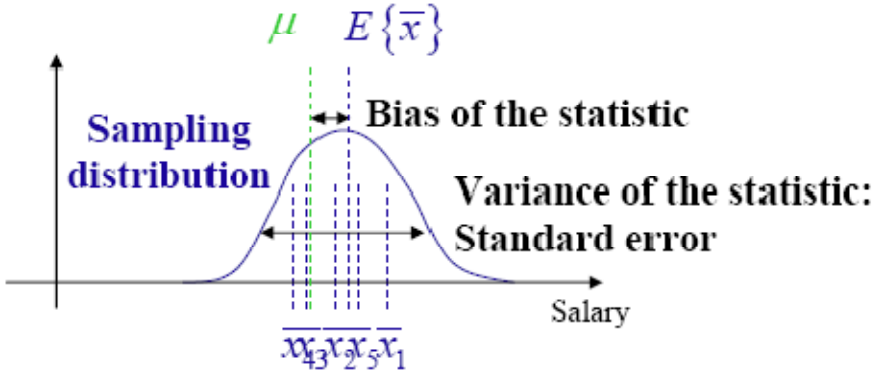
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Parameter: characteristic of a population

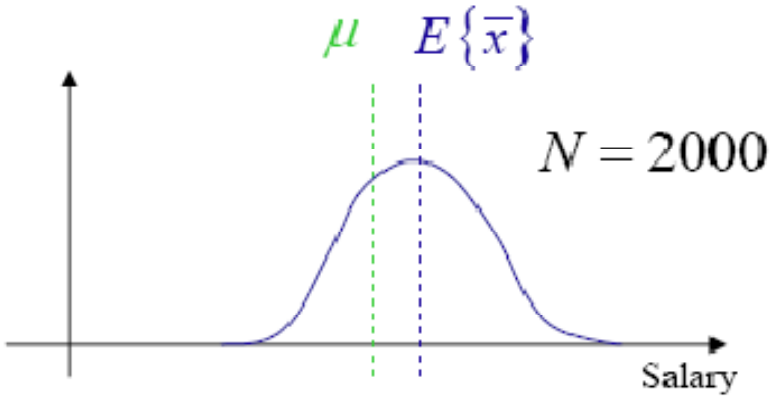
What is the average salary of all Spaniards?

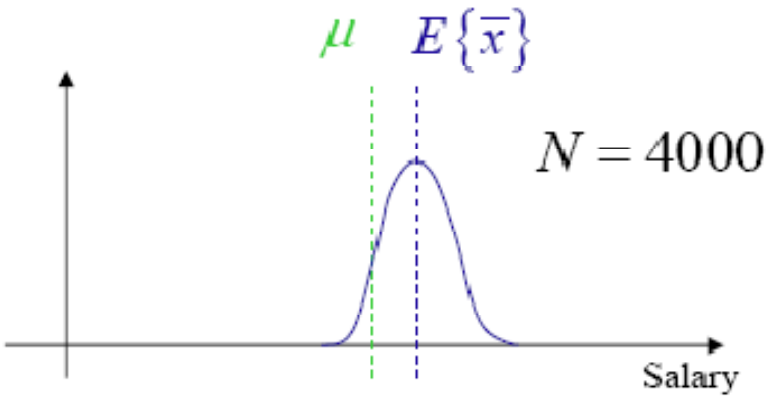
μ





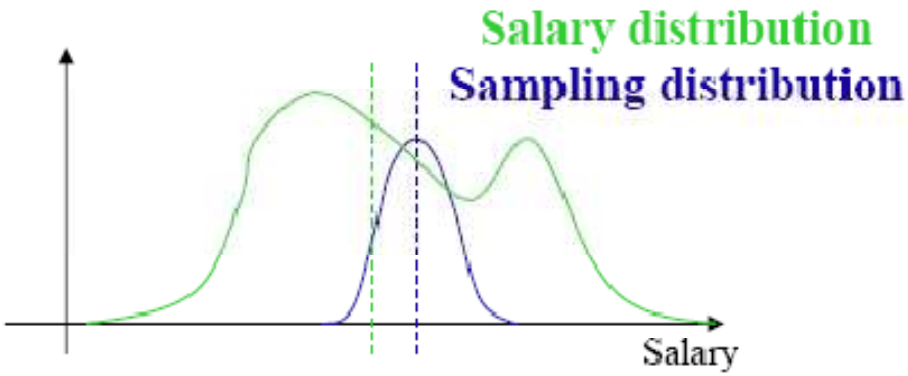
  -



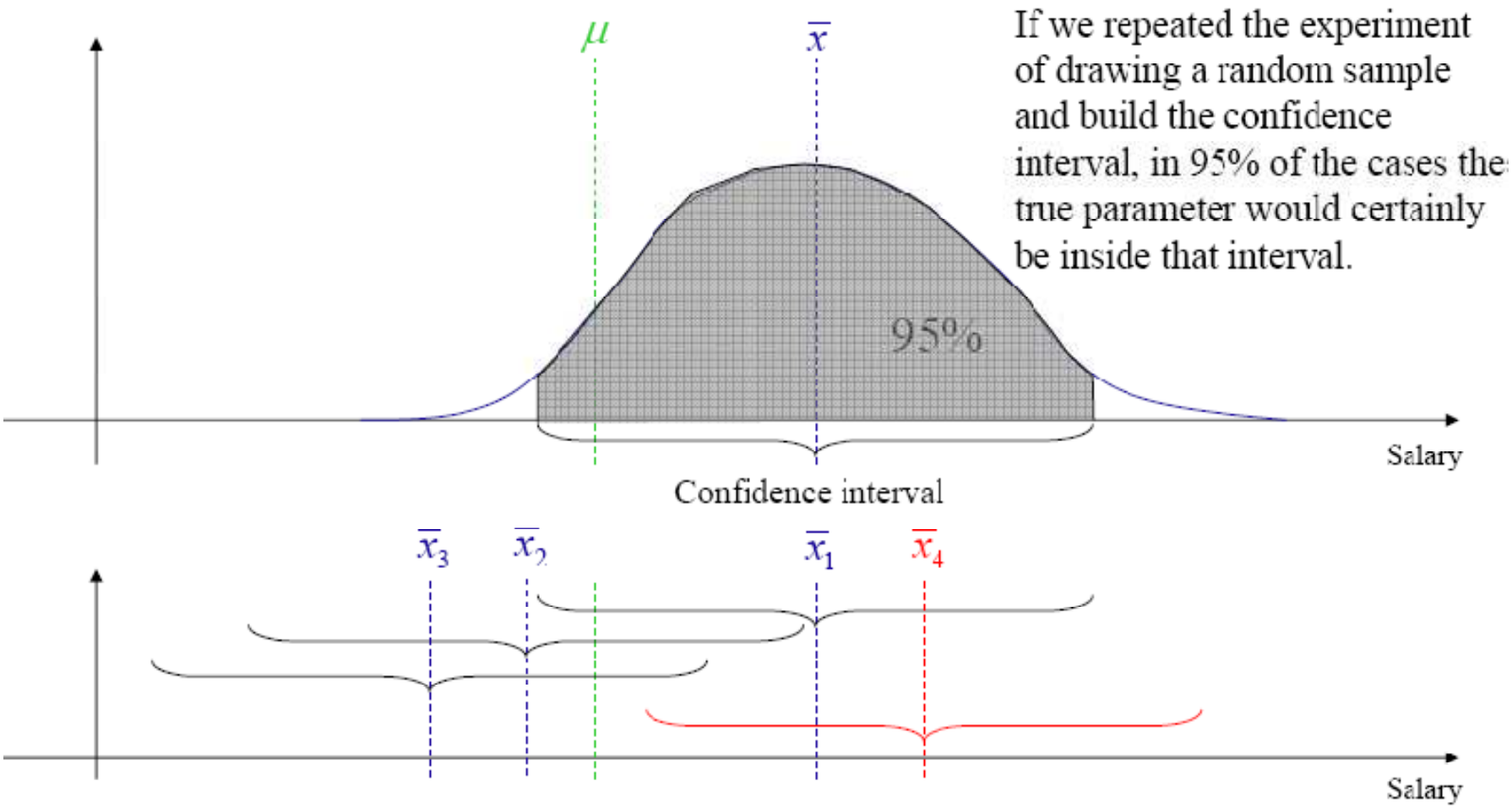


Unbiased $\mu - E\{\bar{x}\} = 0$

Asymptotically unbiased $\lim_{N\to\infty} \mu - E\{\bar{x}\} = 0$



Sampling distribution: distribution of the statistic if all possible samples of size N were drawn from a given population

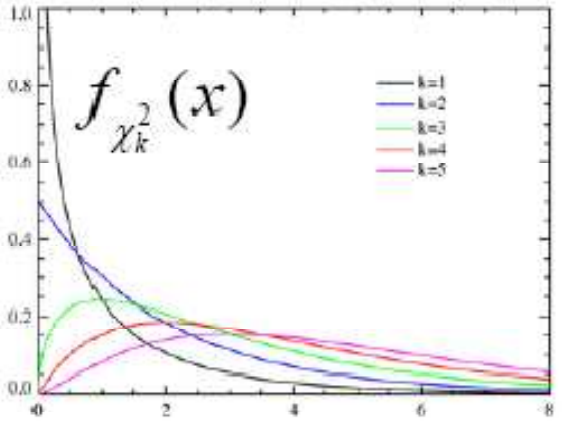


If we repeated the experiment of drawing a random sample and build the confidence interval, in 95% of the cases the true parameter would certainly be inside that interval.

Sometimes the distribution of the statistic is known

$$X_i \sim N(\mu,\sigma^2) \Rightarrow \frac{1}{N}\sum_{i=1}^{N} X_i \sim N\left(\mu,\frac{\sigma^2}{N}\right)$$

$$\sum_{i=1}^{N}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_N$$



Sometimes the distribution of the statistic is NOT known, but still the mean is well behaved

$$E\{X_i\} = \mu$$
$$Var\{X_i\} = \sigma^2$$
$$\Rightarrow \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} X_i \sim N\left(\mu,\frac{\sigma^2}{N}\right)$$

Central limit theorem!!

But:

- The sample must be truly random
- Averages based on samples whose size is more than 30 are reasonably Gaussian

- Which types of variables occur? [Data identification]
- What to measure? Central tendency, differences, variability, skewness and kurtosis, association

During the last 6 months the rentability of your account has been:
5%, 5%, 5%, -5%, -5%, -5%. Which is the average rentability of your account?

Arithmetic mean

(-) Very sensitive to large outliers, not too meaningful for certain distributions
(+) Unique, unbiased estimate of the population mean,
better suited for symmetric distributions

$$x^*_{AM} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$x^*_{AM} = \frac{1}{6}(5+5+5-5-5-5) = 0\%$$

Property $E\left\{x^*_{AM}\right\} = \mu$

$$x^*_{AM} = \frac{1}{6}(1.05+1.05+1.05+0.95+0.95+0.95) = 1 = 0\%$$

Geometric mean

(-) Very sensitive to outliers
(+) Unique, used for the mean of ratios and percent changes,
less sensitive to asymmetric distributions

$$x^*_{GM} = \sqrt[N]{\prod_{i=1}^{N} x_i} \Rightarrow \log x^*_{GM} = \frac{1}{N}\sum_{i=1}^{N} \log x_i$$

$$x^*_{GM} = \sqrt[6]{1.05 \cdot 1.05 \cdot 1.05 \cdot 0.95 \cdot 0.95 \cdot 0.95} = 0.9987 = -0.13\%$$

Harmonic mean

(-) Very sensitive to small outliers
(+) Usually used for the average of rates,
less sensitive to large outliers

$$x^*_{HM} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{x_i}} \Rightarrow \frac{1}{x^*_{HM}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{x_i}$$

A car travels 200km. The first 100 km at a speed of 60km/h, and the second 100 km at a speed of 40 km/h.

100 km | 100 km

$60km/h \Rightarrow t = 100min$ $40km/h \Rightarrow t = 150min$

$$x^*_{HM} = \frac{1}{\frac{1}{2}\left(\frac{1}{60}+\frac{1}{40}\right)} = 48km/h$$

$$x^*_{AM} = \frac{1}{2}(60+40) = 50km/h$$

Which is the right average speed?

Property: For positive numbers $\boxed{x^*_{HM} \le x^*_{GM} \le x^*_{AM}}$

More affected by extreme large values
Less affected by extreme small values

Less affected by extreme values
More affected by extreme small values

Generalization: Generalized mean $x^* = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^p \right)^{\frac{1}{p}}$

Minimum	$p = -\infty$
Harmonic mean	$p = -1$
Geometric mean	$p = 0$
Arithmetic mean	$p = 1$
Quadratic mean	$p = 2$
Maximum	$p = \infty$

Measures of robust central tendency

During the last 6 months the rentability of your account has been:
5%, 3%, 7%, -15%, 6%, 30%. Which is the average rentability of your account?

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6$

$x_{(3)} \quad x_{(2)} \quad x_{(5)} \quad x_{(1)} \quad x_{(4)} \quad x_{(6)}$

Trimmed mean, truncated mean, Windsor mean:

Remove p% of the extreme values on each side

$$x^* = \frac{1}{4}\left(x_{(2)} + x_{(3)} + x_{(4)} + x_{(5)}\right) = \frac{1}{4}\left(3+5+6+7\right) = 5.25\%$$

Median

Which is the central sorted value? (50% of the distribution is below that value) It is not unique

Any value between $x_{(3)} = 5\%$ and $x_{(4)} = 6\%$

Winsorized mean:

Substitute p% of the extreme values on each side

$$x^* = \frac{1}{6}\left(x_{(2)} + x_{(2)} + x_{(3)} + x_{(4)} + x_{(5)} + x_{(5)}\right) = \frac{1}{6}\left(3+3+5+6+7+7\right) = 5.1\widehat{6}\%$$

M-estimators

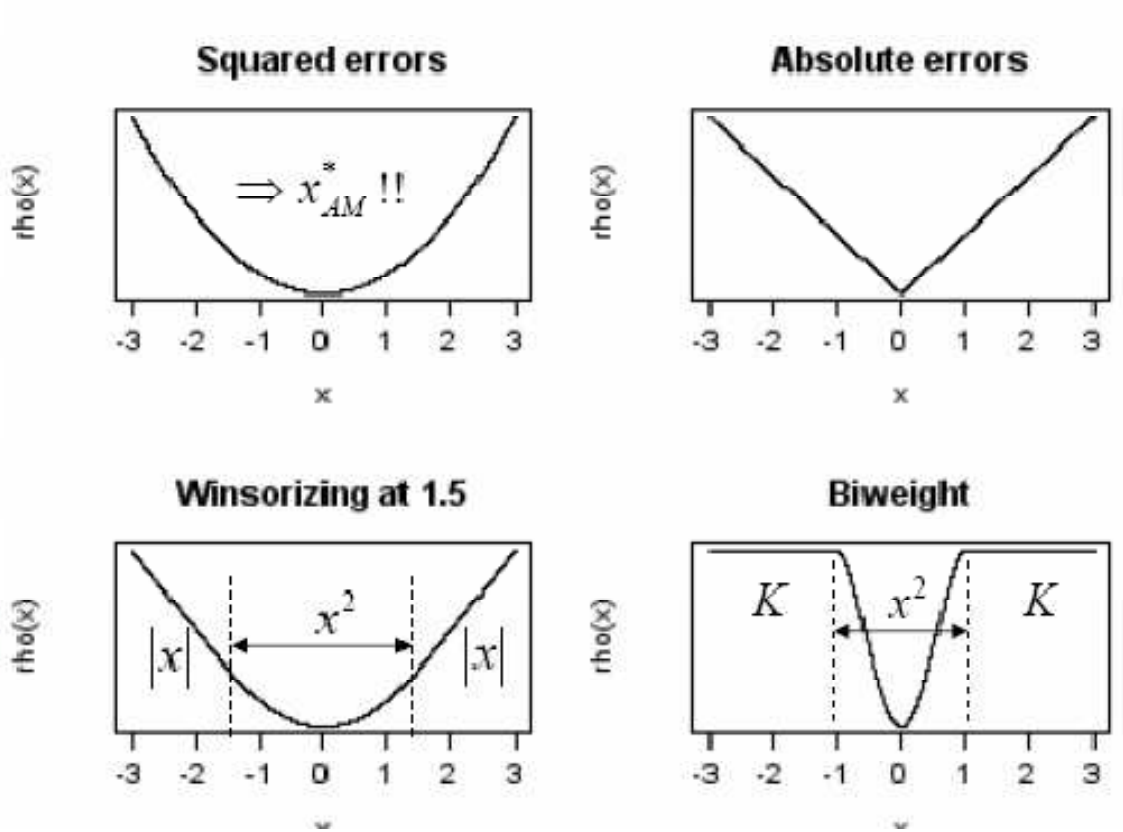
Give different weight to different values

$$x^* = \arg\min_x \frac{1}{N}\sum_{i=1}^{N}\rho(x_i - x)$$

R and L-estimators

Now in disuse

The distribution of robust statistics is usually unknown and has to be estimated experimentally (e.g., bootstrap resampling)



Mode:

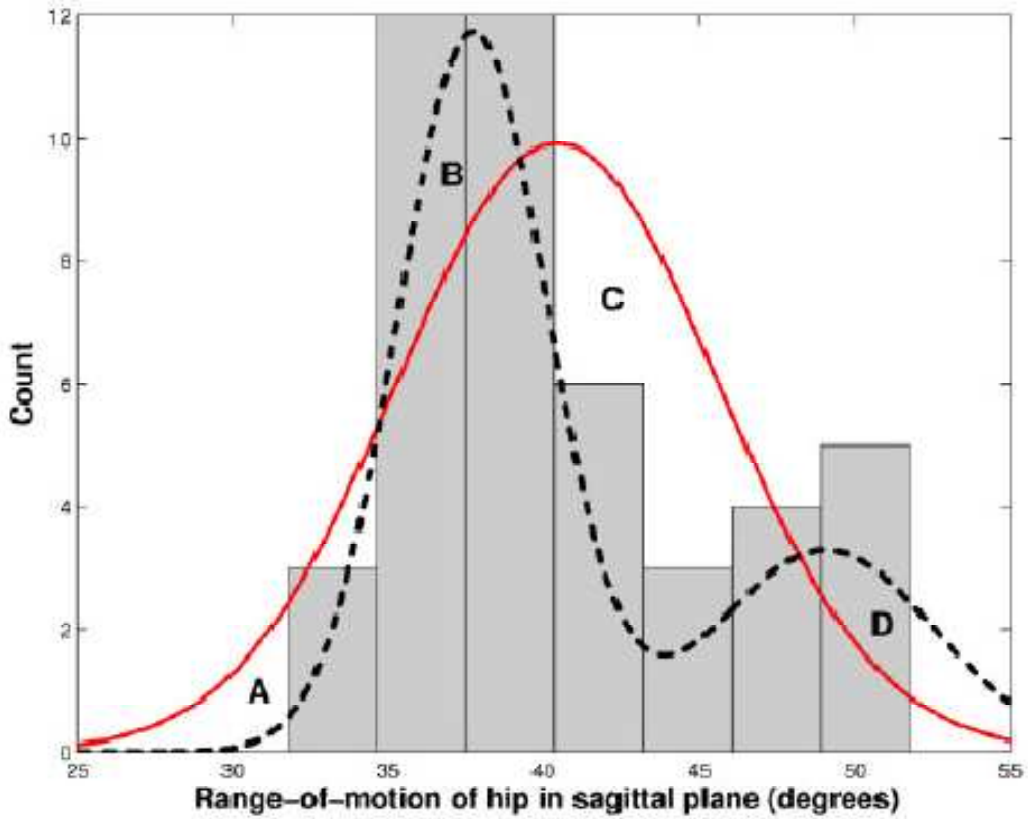
Most frequently occurring

(-) Not unique (multimodal)

(+) representative of the most "typical" result

$$x^* = \arg\max f_X(x)$$

If a variable is multimodal, most central measures fail!



Differences

An engineer tries to determine if a certain modification makes his motor to waste less power. He makes measurements of the power consumed with and without modifications (the motors tested are different in each set). The nominal consumption of the motors is 750W, but they have from factory an unknown standard deviation around 20W. He obtains the following data:

Unmodified motor (Watts): 741, 716, 753, 756, 727 $\quad \bar{x} = 738.6$

Modified motor (Watts): 764, 764, 739, 747, 743 $\quad \bar{y} = 751.4$

Not robust measure of unpaired differences $\quad d^* = \bar{y} - \bar{x}$

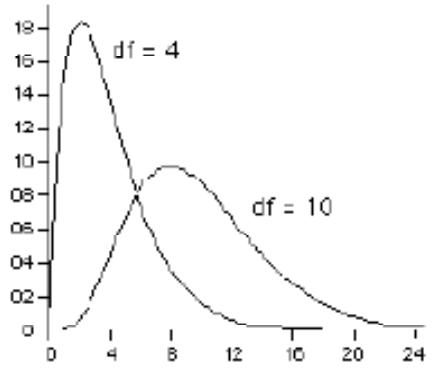
Robust measure of unpaired differences $\quad d^* = median\{y_i - x_j\}$

If the measures are paired (for instance, the motors are first measured, the modified and remeasured), then we should first compute the difference.

Difference: 23, 48, -14, -9, 16 $\quad d^* = \bar{d}$

Variability

During the last 6 months the rentability of an investment product has been:
-5%, 10%, 20%, -15%, 0%, 30% (geometric mean=5.59%)
The rentability of another one has been: 4%, 4%, 4%, 4%, 4%, 4%
Which investment is preferrable for a month?



Variance

(-) In squared units
(+) Very useful in analytical expressions

$$\sigma^2 = E\left\{(X \quad \mu)^2\right\}$$

$$s_N^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad E\left\{s_N^2\right\} = \frac{N-1}{N}\sigma^2$$

Subestimation of the variance

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad E\left\{s^2\right\} = \sigma^2$$

$$X_i \sim N(\mu, \sigma^2) \Rightarrow (N-1)\frac{s^2}{\sigma^2} \sim \chi^2_{N-1}$$

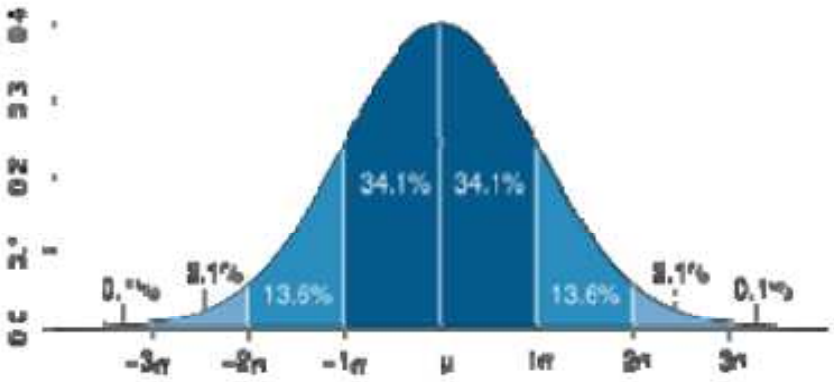
Standard deviation

(+) In natural units,
provides intuitive information about variability
Natural estimator of measurement precision
Natural estimator of range excursions

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Rentability=5.59±√0.0232=5.59±15.23%

$$X_i \sim N(\mu, \sigma^2) \Rightarrow \sqrt{N-1}\,\frac{S}{\sigma} \sim \chi_{N-1}$$



Tchebychev's Inequality

$$\Pr\{\mu - K\sigma \le X \le \mu + K\sigma\} = 1 - \frac{1}{K^2}$$

At least 50% of the values are within √2 standard deviations from the mean.
At least 75% of the values are within 2 standard deviations from the mean.
At least 89% of the values are within 3 standard deviations from the mean.
At least 94% of the values are within 4 standard deviations from the mean.
At least 96% of the values are within 5 standard deviations from the mean.
At least 97% of the values are within 6 standard deviations from the mean.
At least 98% of the values are within 7 standard deviations from the mean.

For any distribution!!!

Percentiles

(-) Difficult to handle in equations
(+) Intuitive definition and meaning
(+) Robust measure of variability

$$\Pr\left\{X \le x^*\right\} = q$$

Someone has an IQ score of 115. Is he clever, very clever, or not clever at all?

Deciles

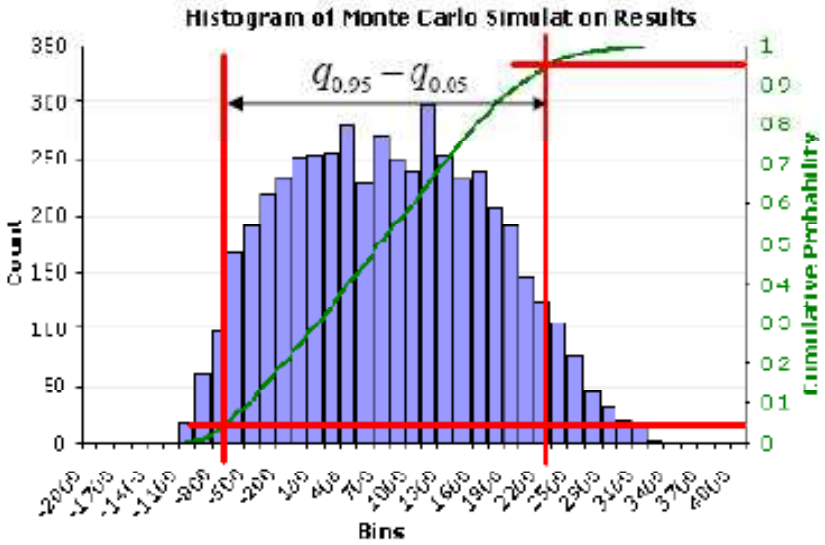
$q_{0.10}, q_{0.20}, q_{0.30}, q_{0.40}, q_{0.50}$
$q_{0.60}, q_{0.70}, q_{0.80}, q_{0.90}$

$q_{0.90} \quad q_{0.10}$

Quartiles

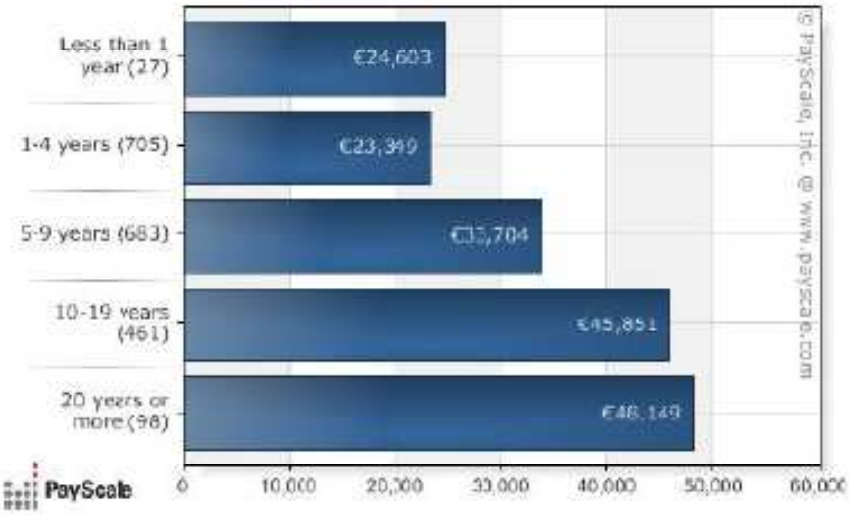
$q_{0.25}, q_{0.50}, q_{0.75}$

$q_{0.75} - q_{0.25}$

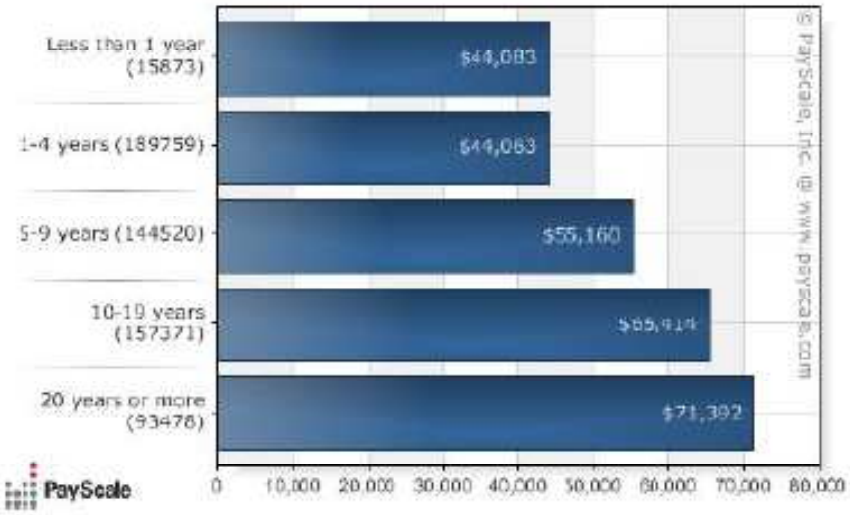


Coefficient of variation

Median salary in Spain by years of experience



Median salary in US by years of experience



In which country you can have more progress along your career?

Skewness

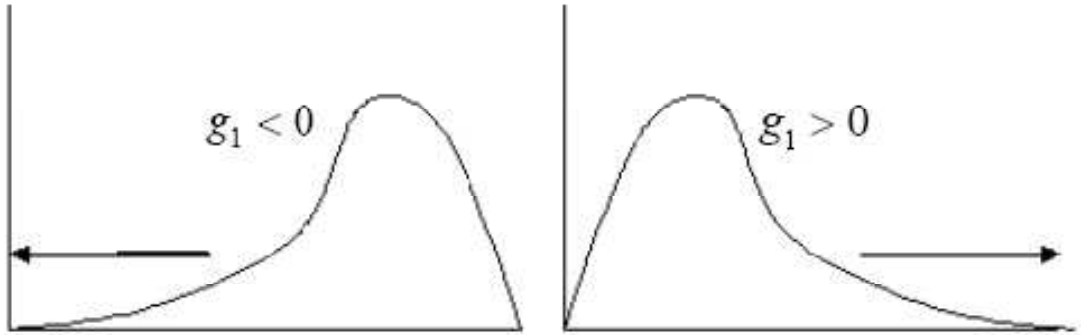
Skewness: Measure of the assymetry of a distribution

$$\gamma_1 = \frac{E\left\{(X-\mu)^3\right\}}{\sigma^3}$$

Unbiased estimator

$$g_1 = \frac{m_3}{s^3}$$

$$m_3 = \frac{N\sum_{i=1}^{N}(x_i-\bar{x})^3}{(N-1)(N-2)}$$



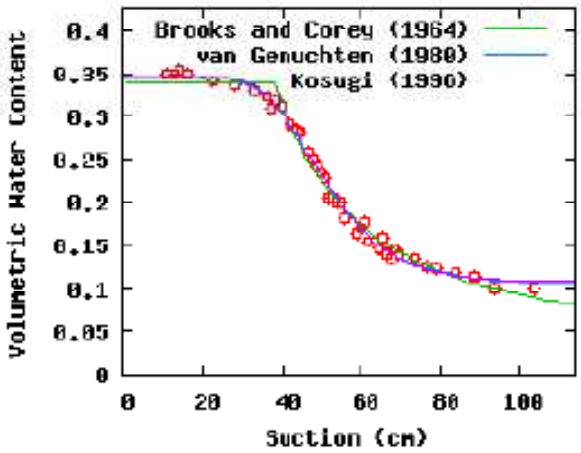
Negative Skew
Elongated tail at the **left**
More data in the left tail than would be expected in a normal distribution

$\mu < Med < Mode$

Positive Skew
Elongated tail at the **right**
More data in the right tail than would be expected in a normal distribution

$Mode > Med > \mu$

The residuals of a fitting should not be skew! Otherwise, it would mean that positive errors are more likely than negative or viceversa. This is the rationale behind some goodness-of-fit tests.



Correlation / association [see also later]

Is there any relationship between education, free-time and salary?

Person	Education (0-10)	Education	Free-time (hours/week)	Salary $	Salary
A	10	High	10	70K	High
B	8	High	15	75K	High
C	5	Medium	27	40K	Medium
D	3	Low	30	20K	Low

Pearson's correlation coefficient

$$\rho = \frac{E\{(X-\mu_X)(Y-\mu_Y)\}}{\sigma_X \sigma_Y} \in [-1,1]$$

$$r = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{s_X s_Y}$$

Salary ↑ ⇒ *FreeTime* ↓ *Education* ↑ ⇒ *Salary* ↑

Correlation	Negative	Positive
Small	−0.3 to −0.1	0.1 to 0.3
Medium	−0.5 to −0.3	0.3 to 0.5
Large	−1.0 to −0.5	0.5 to 1.0

Correlation between two ordinal variables? Kendall's tau

Is there any relationship between education and salary?

Person	Education	Salary $
A	10	70K
B	8	75K
C	5	40K
D	3	20K

Person	Education	Salary $
A	1st	2nd
B	2nd	1st
C	3rd	3rd
D	4th	4th

P=Concordant pairs

Person A
Education: (A>B) (A>C) (A>D)
Salary: (A>C) (A>D)
} 2

Person B
Education: (B>C) (B>D)
Salary: (B>A) (B>C) (B>D)
} 2

Person C
Education: (C>D)
Salary: (C>D)
} 1

Person D
Education:
Salary:
} 0

$$\tau = \frac{P}{\frac{N(N-1)}{2}}$$

$$\tau = \frac{2+2+1+0}{\frac{4(4-1)}{2}} = \frac{5}{6} = 0.83$$

Correlation between two ordinal variables? Spearman's rho

Is there any relationship between education and salary?

Person	Education	Salary $
A	10	70K
B	8	75K
C	5	40K
D	3	20K

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}$$

Person	Education	Salary $	di
A	1st	2nd	-1
B	2nd	1st	1
C	3rd	3rd	0
D	4th	4th	0

$$\rho = 1 - \frac{6((-1)^2 + 1^2 + 0^2 + 0^2)}{4(4^2-1)} = 0.81$$

Other correlation flavours:

- Correlation coefficient: How much of Y can I explain given X?
- Multiple correlation coefficient: How much of Y can I explain given X_1 and X_2?
- Partial correlation coefficient: How much of Y can I explain given X_1 once I remove the variability of Y due to X_2?
- Part correlation coefficient: How much of Y can I explain given X_1 once I remove the variability of X_1 due to X_2?

o Use and abuse of the normal distribution

Univariate $$X \sim N(\mu, \sigma^2) \Rightarrow f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

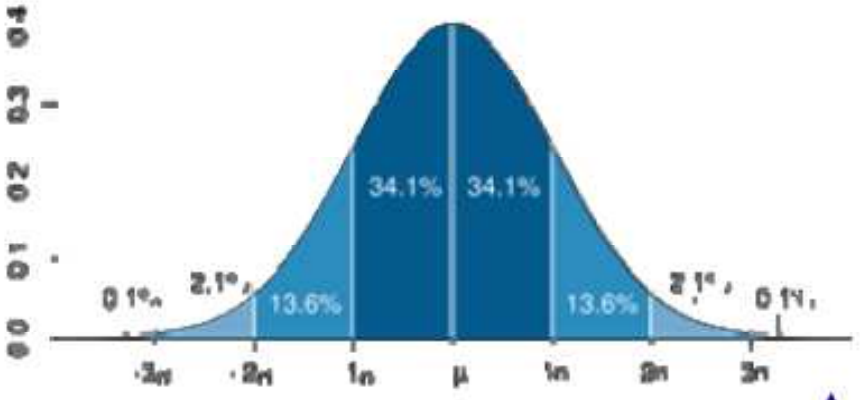
Multivariate $$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \Rightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

Covariance matrix

Use: Normalization

$$X \sim N(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$$

Z-score



Compute the z-score of the IQ ($\mu = 100, \sigma = 15$) of:
Napoleon Bonaparte (emperor): 145
Gary Kasparov (chess): 190

$$z_{Napoleon} = \frac{145-100}{15} = 3$$

$$z_{Kasparov} = \frac{190-100}{15} = 6$$

Use: Computation of probabilties IF the underlying variable is normally distributed

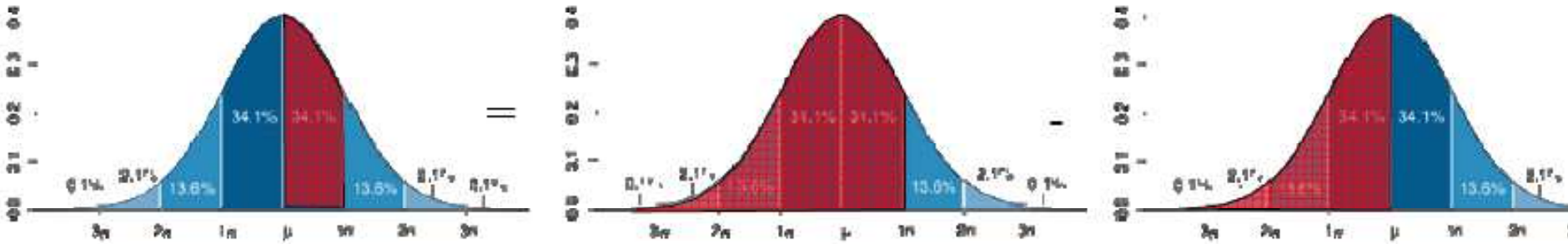
$$X \sim N(\mu, \sigma^2)$$

What is the probability of having an IQ between 100 and 115?

$$\Pr\{100 \le IQ \le 115\} = \int_{100}^{115} \frac{1}{\sqrt{2\pi 15^2}} e^{-\frac{1}{2}\left(\frac{x-100}{15}\right)^2} dx = \int_{0}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx - \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.341$$

Normalization

Use of tabulated values



Use: Computation of probabilties IF the underlying variable is normally distributed

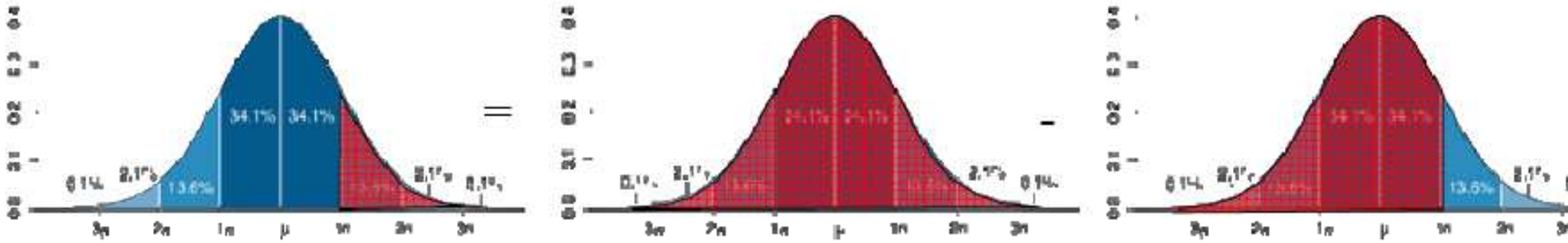
$$X \sim N(\mu, \sigma^2)$$

What is the probability of having an IQ larger than 115?

$$\Pr\{100 \le IQ \le 115\} = \int_{115}^{\infty} \frac{1}{\sqrt{2\pi 15^2}} e^{-\frac{1}{2}\left(\frac{x-100}{15}\right)^2} dx = \int_{1}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 1 - \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.159$$

Normalization

Use of tabulated values

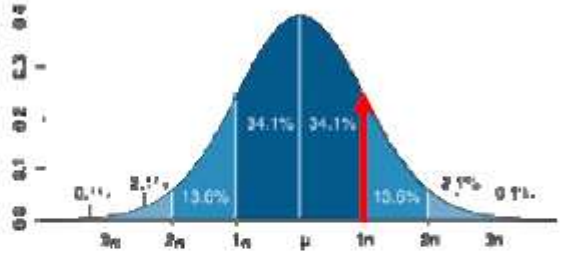


Abuse: Computation of probabilties of a single point

What is the probability of having an IQ exactly equal to 115?

$$\Pr\{IQ = 115\} = 0$$

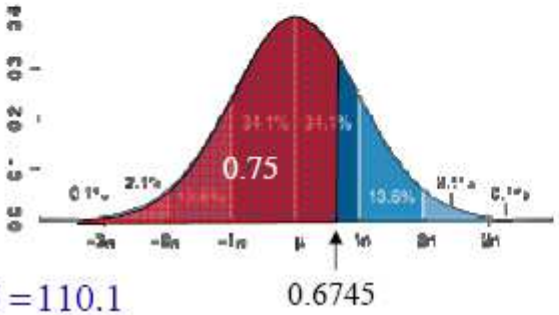
$$Likelihood\{IQ = 115\} = Likelihood\{z_{IQ} = 1\} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}}$$



Use: Computation of percentiles

Which is the IQ percentile of 75%?

$$\int_{-\infty}^{q_{0.75}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 75 \Rightarrow q_{0.75} = 0.6745$$



$$IQ_{0.75} = \mu_{IQ} + q_{0.75}\sigma_{IQ} = 100 + 0.6745 \cdot 15 = 110.1$$

Abuse: Assumption of normality

Many natural fenomena are normally distributed (thanks to the central limit theorem):

- Error in measurements
- Light intensity
- Counting problems when the count number is very high (persons in the metro at peak hour)
- Length of hair
- The logarithm of weight, height, skin surface, … of a person

But many others are not

- The number of people entering a train station in a given minute is not normal, but the number of people entering all the train stations in the world at a given minute is normal.

Many distributions of mathematical operations are normal

$$X_i \sim N \longrightarrow aX_1 + bX_2; a + bX_1 \sim N$$

But many others are not

$$X_i \sim N \longrightarrow \frac{X_1}{X_2} \sim Cauchy; e^X \sim LogNormal; \frac{\sum X_i^2}{\sum X_j^2} \sim F - Snedecor$$

$$\sum X_i^2 \sim \chi^2; \sqrt{\sum X_i^2} \sim \chi; \sqrt{X_1^2 + X_2^2} \sim Rayleigh$$

Some distributions can be safely approximated by the normal distribution

Binomial $np > 10$ *and* $np(1-p) > 10$, Poisson $\lambda > 1000$

Abuse: banknotes

O Are my data really independent?

Independence is different from mutual exclusion

In general,

$$p(A \cap B) = p(A)\,p(B \mid A)$$

$$p(B \mid A) = 0$$

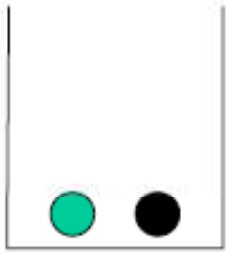
$$p(A \cap B) = 0$$

Knowing A does not give any information about the next event

Mutual exclusion is when two results are impossible to happen at the same time.

Independence is when the probability of an event does not depend on the results that we have had previously.

$$p(A \cap B) = p(A)\,p(B)$$



Example: Sampling with and without replacement

What is the probability of taking a black ball as second draw, if the first draw is green?

Sampling without replacement

In general samples are not independent except if the population is so large that it does not matter.

Sampling with replacement

Samples may be independent. However, they may not be independent (see Example 1)
Examples: tossing a coin, rolling a dice

Random sample: all samples of the same size have equal probability of being selected

Example 1: Study about child removal after abuse, 30% of the members were related to each other because when a child is removed from a family, normally, the rest of his/her siblings are also removed. Answers for all the siblings are correlated.

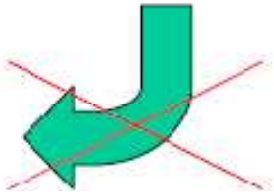
Example 2: Study about watching violent scenes at the University. If someone encourages his roommate to take part in this study about violence, and the roommate accepts, he is already biased in his answers even if he is acting as control watching non-violent scenes.

Consequence: The sampling distributions are not what they are expected to be, and all the confidence intervals and hypothesis testing may be seriously compromised.

- What is the difference between “descriptive” and “inferential statistics”? [see later]
- What is there to know about “parametric vs non-parametric statistics”? [see later]

Parameter estimation

$$X_i \sim N(\mu, \sigma^2) \Rightarrow \frac{1}{N}\sum_{i=1}^{N} X_i \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$



Solution: Resampling (bootstrap, jacknife, …)

Hypothesis testing

Cannot use statistical tests based on any assumption about the distribution of the underlying variable (t-test, F-tests, χ^2-tests, …)

Solution:

- discretize the data and use a test for categorical/ordinal data (non-parametric tests)
- use randomized tests

- How do I collect the data? [Experimental design – Chapter 5]
  - Methodology
  - Design types
  - Basics of experimental design
  - Some designs: Randomized Complete Blocks, Balanced Incomplete Blocks, Latin squares, Graeco-latin squares, Full 2k factorial, Fractional 2k-p factorial
  - What is a covariate?

A covariate is variable that affects the result of the dependent variable, can be measured but cannot be controlled by the experimenter.

We want to measure the effect of 3 different car lubricants. To perform a statistical study we try with 3 different drivers, 3 different cars, and 3 different driving situations (city, highway, mixture). All these are variables that can be controlled. However, the atmospheric temperature also affects the car consumption, it can be measured but cannot be controlled.

Covariates are important in order to build models, but not for designing experiments.

- Now I have data, how do I extract information? [Parameter estimation]
  - How to estimate a parameter of a distribution?

In a class of 20 statisticians, 4 of them smoke.
What is the proportion of smokers among statisticians?

The height of the statisticians in this class is:
1.73, 1.67, 1.76, 1.76, 1.69, 1.81, 1.81, 1.75, 1.77, 1.76,
1.74, 1.79, 1.72, 1.86, 1.74, 1.76, 1.80, 1.75, 1.75, 1.71
What is the average height of statisticians?

The height of 4 Spanish statisticians is:
1.73, 1.79, 1.76, 1.76
What is the average height of Spanish statisticians
knowing that the average should be around 1.70
because that is the average height of Spaniards?

Statistic: characteristic of a sample

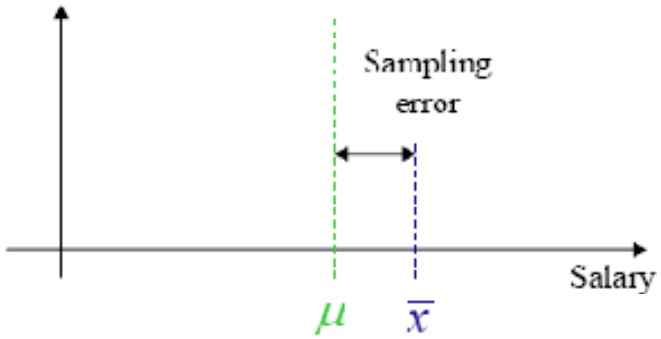
What is the average salary of 2000 people randomly sampled in Spain?

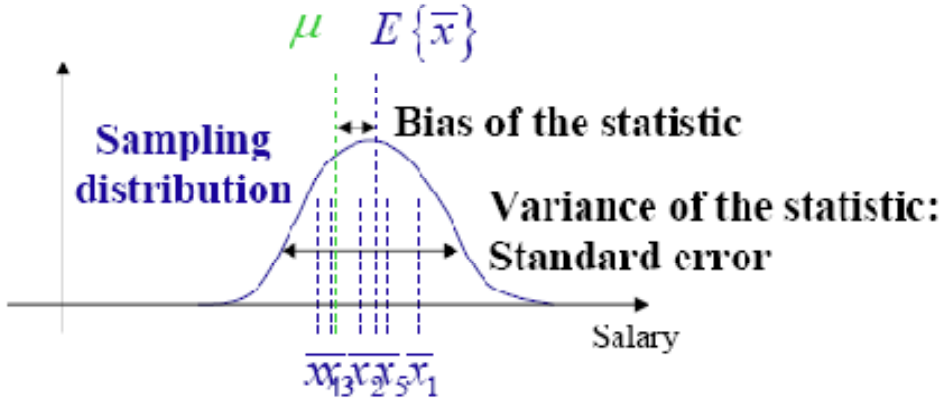
$$x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Parameter: characteristic of a population

What is the average salary of all Spaniards?

μ





- How to report on a parameter of a distribution? What are confidence intervals?

o What if my data are "contaminated"? [Robust statistics]

A "contamination" is anything that keeps your data away from the assumptions of the methods to compute confidence intervals (mainly "normality"):

- The data is not normal
  - → Use non-parametric estimators of confidence intervals (e.g., bootstrap)
  - → Fit a general probability density function (e.g. Gaussian mixture)
- The data is normal but there are outliers
  - → Remove outliers if possible
  - → Use robust estimators of the mean, variance, …

Bootstrap estimate of the confidence interval for the mean of N(>100) samples:

1. Take a random subsample of the original sample of size N (with replacement)
2. Estimate the mean of your subsample.
3. Repeat steps 1 and 2 at least 1000 times.

This gives you the empirical distribution of the mean from which the confidence interval can be computed. This empirical distribution can be computed with any statistic (median, mode, regression coefficient, …)

- Can I see any interesting association between two variables, two populations, …? [Chapter 8]
  - What are the different measures available?
  - Use and abuse of the correlation coefficient
  - Can I deduce a model for my data?
    - What kind of models are available?
    - How to select the appropriate model?
  - Regression as a model
    - What are the assumptions of regression
    - Are there other kind of regressions?
    - How reliable are the coefficients? Confidence intervals
    - How reliable are the coefficients? Validation

- How can I know if what I see is “true”? [Hypothesis testing – Chapter 7]
  - The basics: What is a hypothesis test? What is the statistical power? What is a p-value? How to use it? What is the relationship between sample size, sampling error, effect size and power? What are bootstraps and permutation tests?
  - What are the assumptions of hypothesis testing?
  - How to select the appropriate statistical test
    - Tests about a population central tendency
    - Tests about a population variability
    - Tests about a population distributions
    - Tests about differences randomness
    - Tests about correlation/association measures
  - Multiple testing

- How many samples do I need for my test? [Sample size – Chapter 7]
  - Basic formulas for different distributions
  - Formulas for samples with different costs
  - What if I cannot get more samples? Resampling:
  - Bootstrapping, jackknife

1.3 Relation between descriptive and inferential statistics

Statistics (="state arithmetic")

Descriptive: describe data

- How rich are our citizens on average? → Central Tendency
- Are there many differences between rich and poor? → Variability
- Are more intelligent people richer? → Association
- How many people earn this money? → Probability distribution
- Tools: tables (all kinds of summaries), graphs (all kind of plots), distributions (joint, conditional, marginal, …), statistics (mean, variance, correlation coefficient, histogram, …)

Inferential: derive conclusions and make predictions

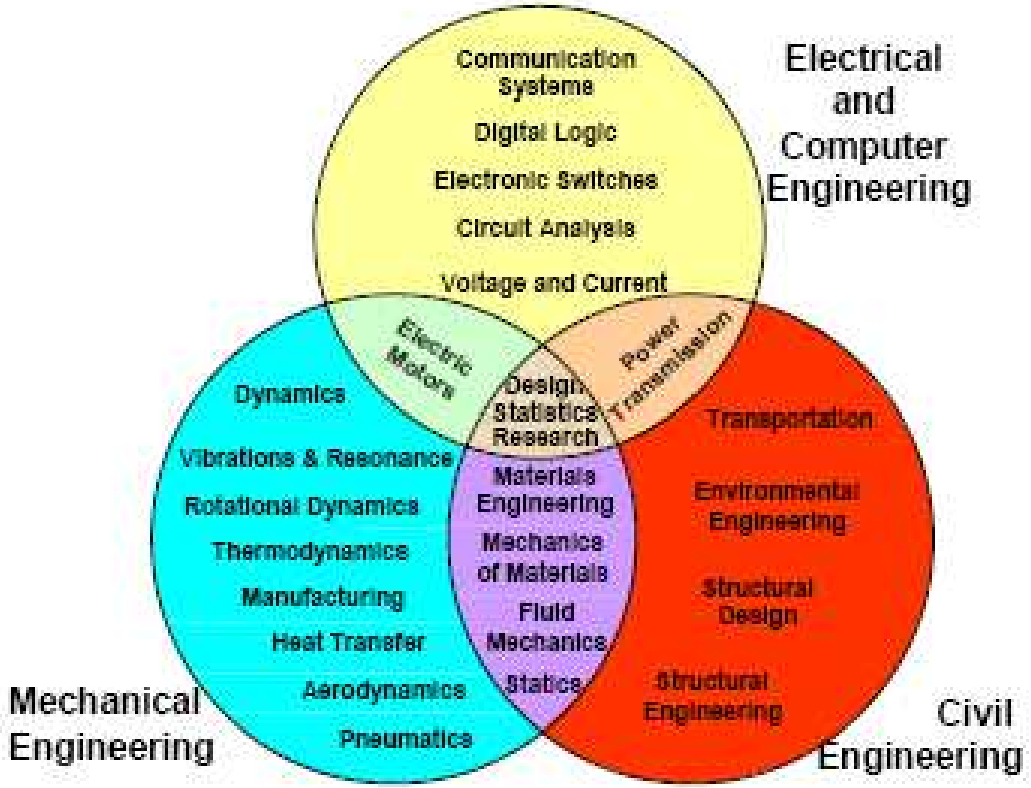
- Is my country so rich as my neighbors? → Inference
- To measure richness, do I have to consider EVERYONE? → Sampling
- If I don't consider everyone, how reliable is my estimate? → Confidence
- Is our economy in recession? → Prediction
- What will be the impact of an expensive oil? → Modelling
- Tools: Hypothesis testing, Confidence intervals, Parameter estimation, Experiment design, Sampling, Time models, Statistical models (ANOVA, Generalized Linear Models, …)

- Statistics is defined as a process by which numerical data are transformed into a usable form for scientific interpretation.
- This entails manipulating data to summarize the findings (descriptive statistics).
- It can also be used to develop general conclusions from the data (inferential statistics)

1.4 General flow in statistics

- The aforementioned questions give a good idea about the potential road to travel by when "doing statistics"
  - Collect the data to address your question of interest
  - Identify and describe the data (exploratory data analysis – EDA / descriptive statistics) – Chapter 6
  - Retrieve interesting relationships from the data by
    - using graphics (EDA) or by
    - using appropriate parametric models (check the validity of the models) – Chapter 6, 8
  - Extract additional information from the data via parameter estimation and confidence intervals – Chapter 7
  - Formally test if what you observe is "true" (plausible) via hypothesis testing, hereby accounting for the "power" of your test – Chapter 7

1.5 Statistics from an engineering perspective



- From an engineering point of view, the statistical work flow can also be organized in the following way:
  - **Explore**:
    - Involves assumptions, principles, and techniques necessary to gain insight into the data via EDA-exploratory data analysis
  - **Measure**:
    - Understanding the measurement process in terms of the errors that affect the process,
    - Which steps can be taken to exercise statistical control over the measurement process and demonstrate the validity of the uncertainty statement,
    - Develop procedures for calibrating artifacts and instruments while guaranteeing the 'goodness' of the calibration results

- Exploit uncertainty as a measure of the 'goodness' of a result. Without such a measure, it is impossible to judge the fitness of the value as a basis for making decisions relating to health, safety, commerce or scientific excellence.

o **Characterize**:

- How to plan and conduct a *Production Process Characterization Study* (PPC) on manufacturing processes.
- How to model manufacturing processes and use these models to design a data collection scheme and to guide data analysis activities.
- How to analyze the data collected in characterization studies and how to interpret and report the results.

- **Model:**
  - Involves specific analysis techniques needed to construct a statistical model that describes a particular scientific or engineering process
  - Involves using different types of models for prediction of process outputs, for calibration, or for process optimization
- **Improve:**
  - What are the engineering and mathematical assumptions that are typical for design of experiments in engineering applications?
- **Monitor:**
  - Understand techniques for monitoring and controlling processes and signaling when corrective actions are necessary

- **Reliability:**
  - Involves understanding terms, models and techniques used to evaluate and predict product reliability.

- More details about this categorization can be retrieved from the online Engineering Statistics Handbook (http://www.itl.nist.gov/div898/handbook/stoc.htm)
- There exist “selective guides to literature on statistical information for engineers”, such as the equally called reference with ISBN 0-87823-155-2

2 Looking at data

1.2 Exploratory Data Analysis (EDA)

- What is exploratory data analysis?
- How did it begin?
- How and where did it originate?
- How is it differentiated from other data analysis approaches, such as classical and Bayesian?
- Is EDA the same as statistical graphics?
- What role does statistical graphics play in EDA?
- Is statistical graphics identical to EDA?

What is EDA?

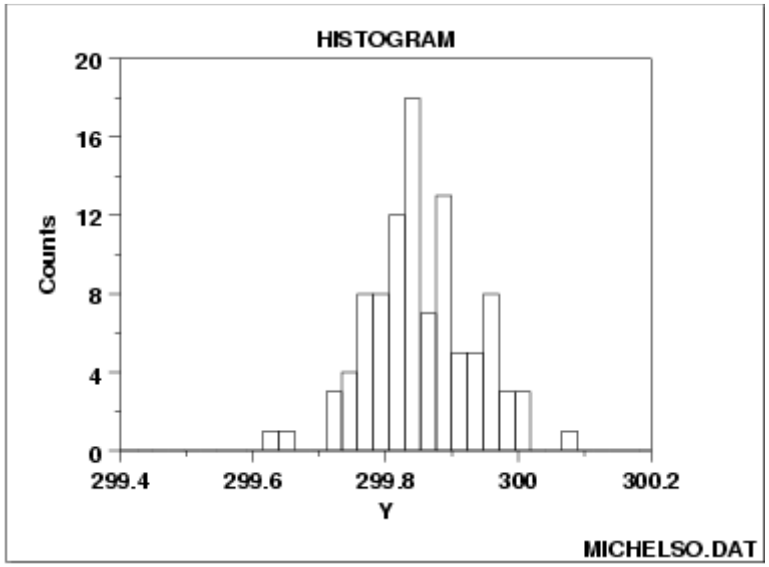
- Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

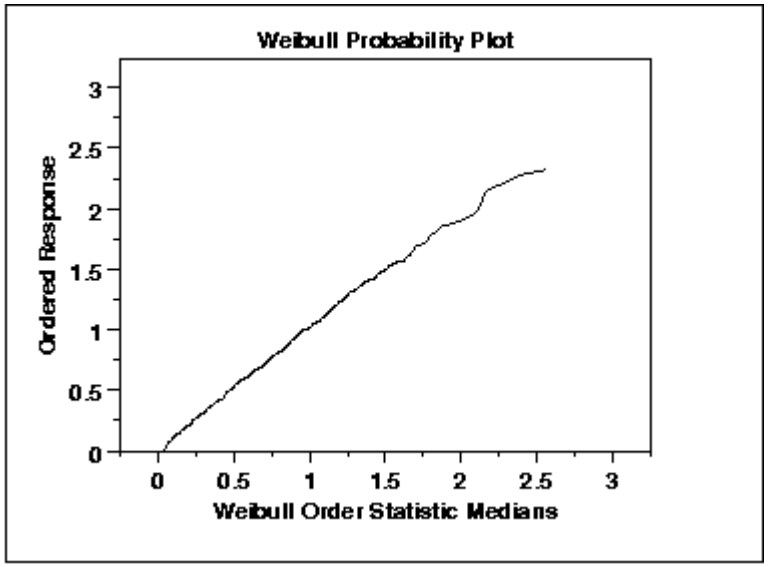
  - maximize insight into a data set;
  - uncover underlying structure;
  - extract important variables;
  - detect outliers and anomalies;
  - test underlying assumptions;
  - develop parsimonious models; and
  - determine optimal factor settings.

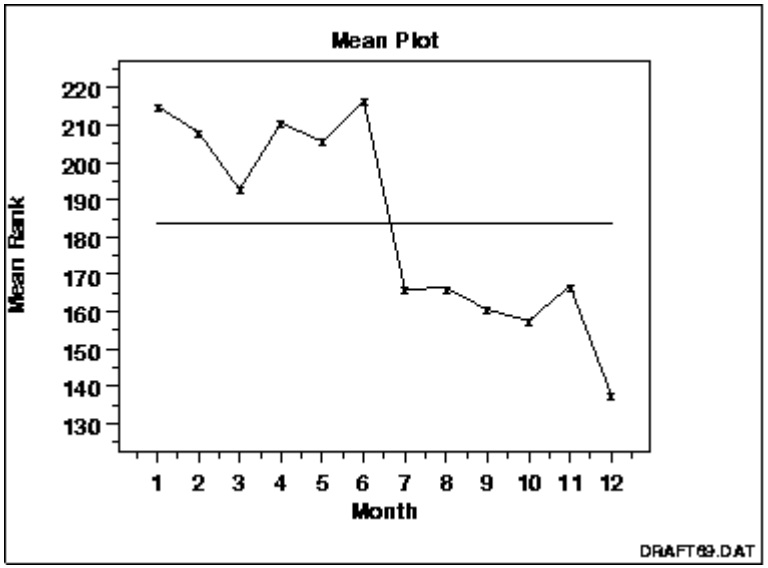
- The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out
- EDA is not identical to statistical graphics although the two terms are used almost interchangeably.
    - Statistical graphics is a collection of techniques--all graphically based and all focusing on one data characterization aspect.
    - EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model.

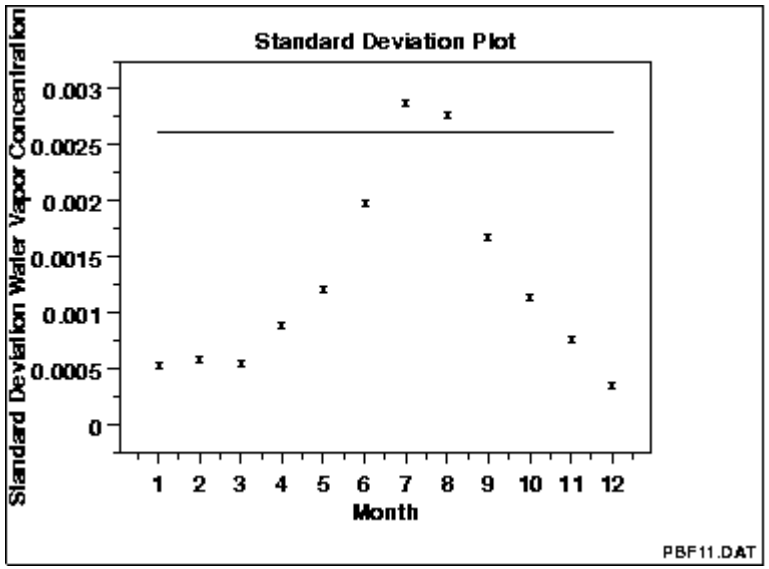
- Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

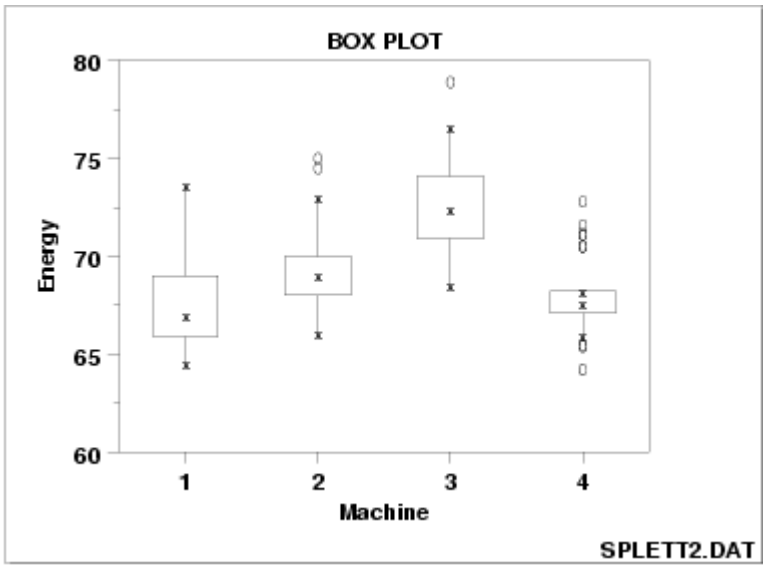
- The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:
  - Plotting the raw data such as histograms, probability plots
  - Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
  - Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.




Weibull Probability Plot
Ordered Response
Weibull Order Statistic Medians
0
0.5
1
1.5
2
2.5
3


Mean Plot
Mean Rank
Month
DRAFT69.DAT
130
140
150
160
170
180
190
200
210
220
1 2 3 4 5 6 7 8 9 10 11 12


Standard Deviation Plot
Standard Deviation Water Vapor Concentration
Month
PBF11.DAT
0
0.0005
0.001
0.0015
0.002
0.0025
0.003
1 2 3 4 5 6 7 8 9 10 11 12


BOX PLOT
Energy
Machine
SPLETT2.DAT
60
65
70
75
80
1 2 3 4

How does EDA differ from classical data analysis?

- Three popular data analysis approaches are:
  - Classical
  - Exploratory (EDA)
  - Bayesian
- These three approaches are similar in that they all start with a general science/engineering problem and all yield science/engineering conclusions. The difference is the sequence and focus of the intermediate steps.

  - For classical analysis, the sequence is

    Problem => Data => Model => Analysis => Conclusions
  - For EDA, the sequence is

    Problem => Data => Analysis => Model => Conclusions
  - For Bayesian, the sequence is

    Problem => Data => Model => Prior Distribution
    => Analysis => Conclusions

- Thus for classical analysis, the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model.
- For EDA, the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a goal of inferring what model would be appropriate.
- Finally, for a Bayesian analysis, the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data-independent distribution on the parameters of the selected model; the analysis thus consists of formally combining both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.

- Focusing on EDA versus classical, these two approaches differ as follows:
  - Models
  - Focus
  - Techniques
  - Rigor
  - Data Treatment
  - Assumptions

How does EDA differ from a summary analysis?

- A summary analysis is simply a numeric reduction of a historical data set. It is quite passive. Its focus is in the past. Quite commonly, its purpose is to simply arrive at a few key statistics (for example, mean and standard deviation) which may then either replace the data set or be added to the data set in the form of a summary table.
- In contrast, EDA has as its broadest goal the desire to gain insight into the engineering/scientific process behind the data. Whereas summary statistics are passive and historical, EDA is active and futuristic. In an attempt to "understand" the process and improve it in the future, EDA uses the data as a "window" to peer into the heart of the process that generated the data. There is an archival role in the research and manufacturing world for summary statistics, but there is an enormously larger role for the EDA approach.

EDA goals

- The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

  1. a good-fitting, parsimonious model
  2. a list of outliers
  3. a sense of robustness of conclusions
  4. estimates for parameters
  5. uncertainties for those estimates
  6. a ranked list of important factors
  7. conclusions as to whether individual factors are statistically significant
  8. optimal settings

- Insight implies detecting and uncovering underlying structure in the data. Such underlying structure may not be encapsulated in the list of items above; such items serve as the specific targets of an analysis, but the real insight and "feel" for a data set comes as the analyst judiciously probes and explores the various subtleties of the data. The "feel" for the data comes almost exclusively from the application of various graphical techniques, the collection of which serves as the window into the essence of the data. Graphics are irreplaceable--there are no quantitative analogues that will give the same insight as well-chosen graphics.
- To get a "feel" for the data, it is not enough for the analyst to know what is in the data; the analyst also must know what is not in the data, and the only way to do that is to draw on our own human pattern-recognition and comparative abilities in the context of a series of judicious graphical techniques applied to the data.

The role of graphics

- Statistics and data analysis procedures can broadly be split into two parts:
  - quantitative
  - graphical
- Quantitative techniques are the set of statistical procedures that yield numeric or tabular output. Examples of quantitative techniques include:
  - hypothesis testing
  - analysis of variance
  - point estimates and confidence intervals
  - least squares regression

These and similar techniques are all valuable and are mainstream in terms of classical analysis.

- On the other hand, there is a large collection of statistical tools that we generally refer to as graphical techniques. These include:
  - scatter plots
  - histograms
  - probability plots
  - residual plots
  - box plots
  - block plots

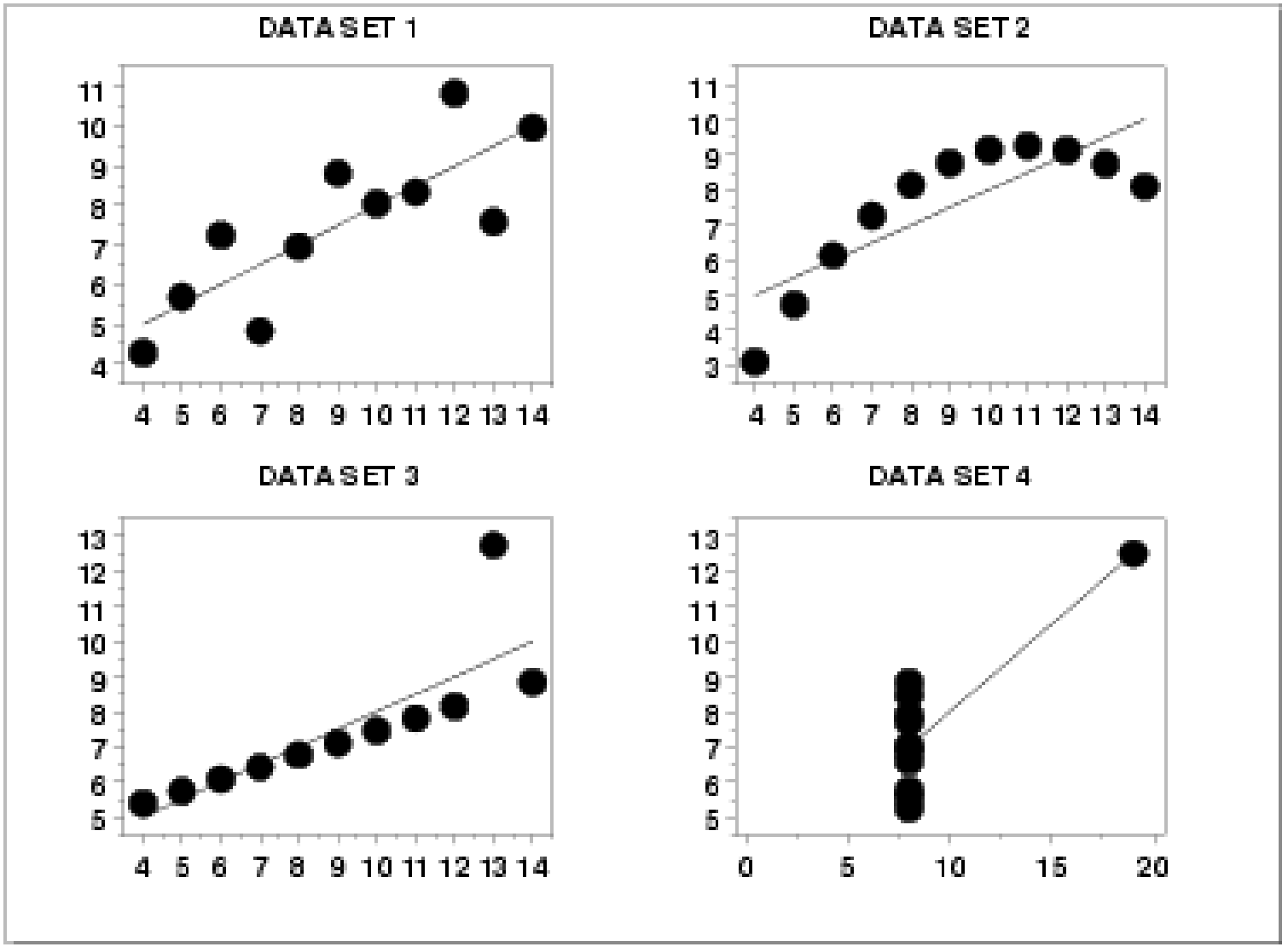
- The EDA approach relies heavily on these and similar graphical techniques. Graphical procedures are not just tools that we could use in an EDA context, they are tools that we must use.
- Such graphical tools are the shortest path to gaining insight into a data set in terms of:
  - testing assumptions
  - model selection
  - model validation
  - estimator selection
  - relationship identification
  - factor effect determination
  - outlier detection
- If one is not using statistical graphics, then one is losing insight into one or more aspects of the underlying structure of the data.

An EDA graphics example

- Given 4 data sets (actual data omitted), for which

  N = 11
  Mean of X = 9.0
  Mean of Y = 7.5
  Intercept = 3
  Slope = 0.5
  Residual standard deviation = 1.236
  Correlation = 0.816 (0.817 for data set 4)

- This implies that in some quantitative sense, all four of the data sets are "equivalent".
- In fact, the four data sets are far from "equivalent" and a scatter plot of each data set, which would be step 1 of any EDA approach, would tell us that immediately.


DATA SET 1
DATA SET 2
DATA SET 3
DATA SET 4

General problem categories

Univariate and Control

UNIVARIATE	CONTROL
Data: A single column of numbers, *Y*. Model: *y* = constant + error Output: 1. A number (the estimated constant in the model). 2. An estimate of uncertainty for the constant. 3. An estimate of the distribution for the error. Techniques: • Probability Plot	Data: A single column of numbers, *Y*. Model: *y* = constant + error Output: A "yes" or "no" to the question "Is the system out of control?" Techniques: • Control Charts

Comparative and Screening

COMPARATIVE	SCREENING
Data: A single response variable and k independent variables (Y, X_1, X_2, ... , X_k), primary focus is on *one* (the primary factor) of these independent variables. Model: $y = f(x_1, x_2, ..., x_k)$ + error Output: A "yes" or "no" to the question "Is the primary factor significant?". Techniques: • Block Plot • Scatter Plot • Box Plot	Data: A single response variable and k independent variables (Y, X_1, X_2, ... , X_k). Model: $y = f(x_1, x_2, ..., x_k)$ + error Output: 1. A ranked list (from most important to least important) of factors. 2. Best settings for the factors. 3. A good model/prediction equation relating Y to the factors. Techniques: • Block Plot • Probability Plot • Bihistogram

Optimization and Regression

OPTIMIZATION	REGRESSION
Data: A single response variable and k independent variables (Y, X_1, X_2, ... , X_k). Model: $y = f(x_1, x_2, ..., x_k)$ + error Output: Best settings for the factor variables. Techniques: • Block Plot • Least Squares Fitting • Contour Plot	Data: A single response variable and k independent variables (Y, X_1, X_2, ... , X_k). The independent variables can be continuous. Model: $y = f(x_1, x_2, ..., x_k)$ + error Output: A good model/prediction equation relating Y to the factors. Techniques: • Least Squares Fitting • Scatter Plot • 6-Plot

Time Series and Multivariate

TIME SERIES	MULTIVARIATE
Data: A column of time dependent numbers, *Y*. In addition, time is an indpendent variable. The time variable can be either explicit or implied. If the data are not equi-spaced, the time variable should be explicitly provided. Model: $y_t = f(t)$ + error The model can be either a time domain based or frequency domain based.	Data: *k* factor variables (X_1, X_2, ... , X_k). Model: The model is not explicit. Output: Identify underlying correlation structure in the data. Techniques: • Star Plot • Scatter Plot Matrix • Conditioning Plot • Profile Plot • Principal Components • Clustering • Discrimination/Classification .

Output: A good model/prediction equation relating *Y* to previous values of *Y*. Techniques: • Autocorrelation Plot • Spectrum • Complex Demodulation Amplitude Plot • Complex Demodulation Phase Plot • ARIMA Models	

Assumptions of EDA

- There are four assumptions that typically underlie all measurement processes; namely, that the data from the process at hand "behave like":
  - random drawings;
  - from a fixed distribution;
  - with the distribution having fixed location; and
  - with the distribution having fixed variation
- The "fixed location" referred to in item 3 above differs for different problem types. The simplest problem type is univariate; that is, a single variable. For the univariate problem, the general model

$$response = deterministic\ component + random\ component$$

becomes

$$response = constant + error$$

- For this case, the "fixed location" is simply the unknown constant. We can thus imagine the process at hand to be operating under constant conditions that produce a single column of data with the properties that
  - the data are uncorrelated with one another;
  - the random component has a fixed distribution;
  - the deterministic component consists of only a constant; and
  - the random component has fixed variation.
- The universal power and importance of the univariate model is that it can easily be extended to the more general case where the deterministic component is not just a constant, but is in fact a function of many variables, and the engineering objective is to characterize and model the function.

- The key point is that regardless of how many factors there are, and regardless of how complicated the function is, if the engineer succeeds in choosing a good model, then the differences (residuals) between the raw response data and the predicted values from the fitted model should themselves behave like a univariate process. Furthermore, the residuals from this univariate process fit will behave like:
  - random drawings;
  - from a fixed distribution;
  - with fixed location (namely, 0 in this case); and
  - with fixed variation.

- Thus if the residuals from the fitted model do in fact behave like the ideal, then testing of underlying assumptions becomes a tool for the validation and quality of fit of the chosen model. On the other hand, if the residuals from the chosen fitted model violate one or more of the above univariate assumptions, then the chosen fitted model is inadequate and an opportunity exists for arriving at an improved model.
- How can I tell that my model fits the data well? [see Chapter 8]

    - Often the validation of a model seems to consist of nothing more than quoting the R^2 statistic from the fit (which measures the fraction of the total variability in the response that is accounted for by the model). Unfortunately, a high R^2 value does not guarantee that the model fits the data well!!!

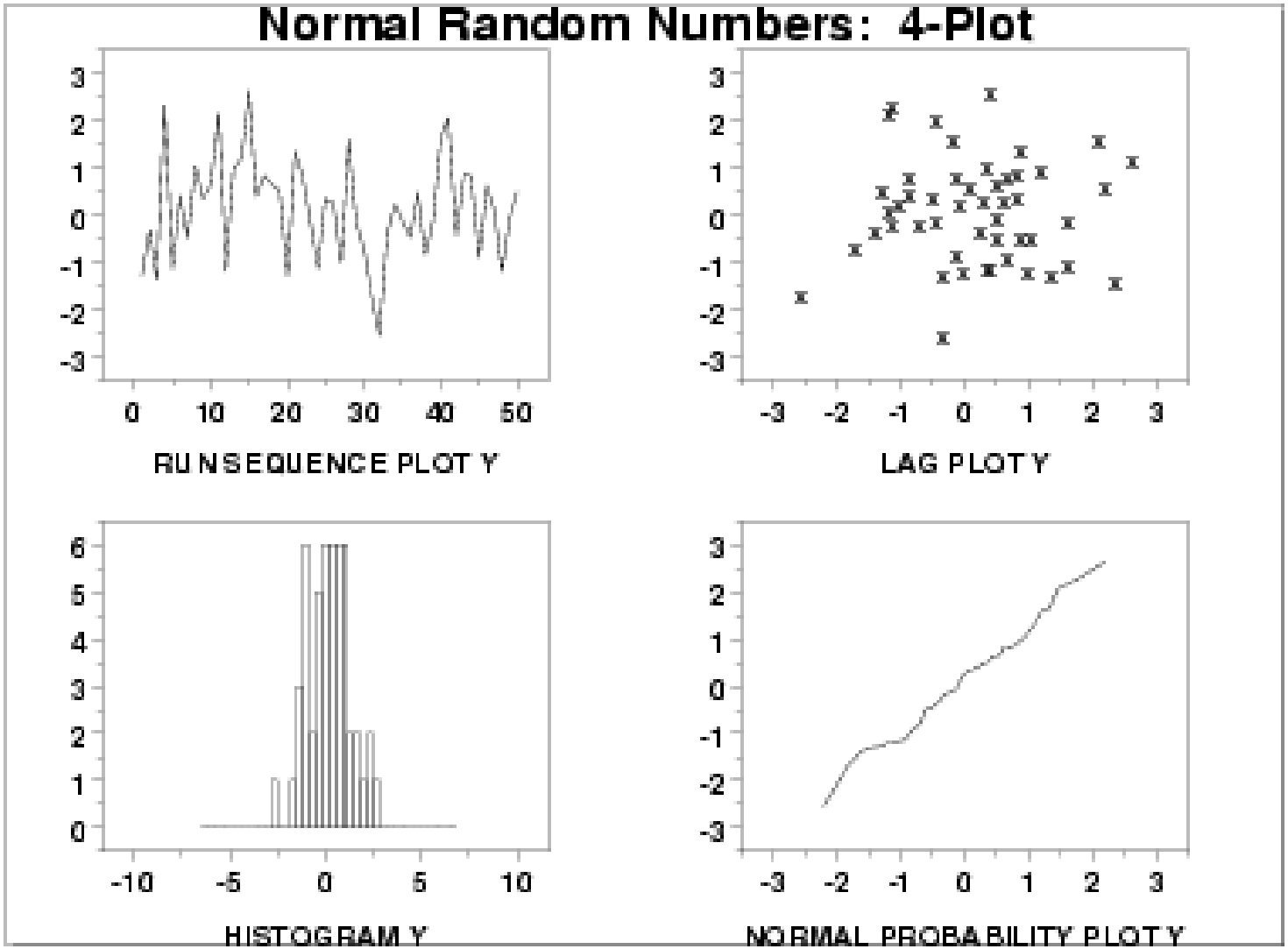
Techniques for testing assumptions

- Because the validity of the final scientific/engineering conclusions is inextricably linked to the validity of the underlying univariate assumptions, it naturally follows that there is a real necessity that each and every one of the above four assumptions be routinely tested.
- The following EDA techniques are simple, efficient, and powerful for the routine testing of underlying assumptions:
    - o run sequence plot (Y_i versus i)
    - o lag plot (Y_i versus Y_{i-1})
    - o histogram (counts versus subgroups of Y)
    - o normal probability plot (ordered Y versus theoretical ordered Y)

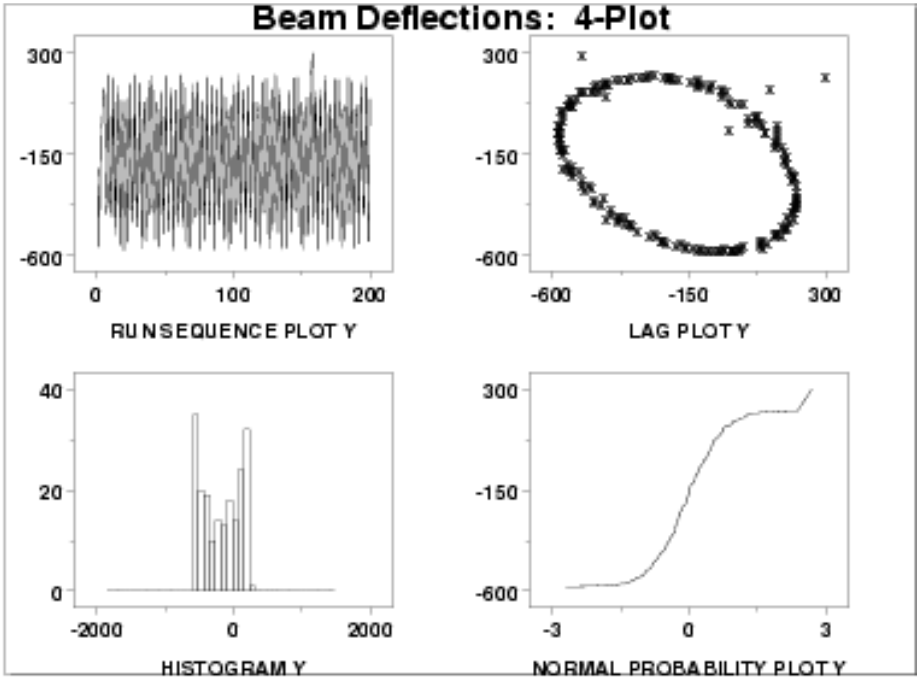
- The four EDA plots can be juxtaposed for a quick look at the characteristics of the data. The plots on the next slide are ordered as follows:

  1. Run sequence plot - upper left
  2. Lag plot - upper right
  3. Histogram - lower left
  4. Normal probability plot - lower right

- This 4-plot reveals a process that has fixed location, fixed variation, is random, apparently has a fixed approximately normal distribution, and has no outliers.


Normal Random Numbers: 4-Plot
RUNSEQUENCE PLOT Y
LAG PLOT Y
HISTOGRAM Y
NORMAL PROBABILITY PLOT Y

- If one or more of the four underlying assumptions do not hold, then it will show up in the various plots as demonstrated in the following example.



- This 4-plot reveals a process that has fixed location, fixed variation, is non-random (oscillatory), has a non-normal, U-shaped distribution, and has several outliers

Interpretation of 4-plots

- The four EDA plots discussed on the previous page are used to test the underlying assumptions:
  - **Fixed Location:**
    If the fixed location assumption holds, then the run sequence plot will be flat and non-drifting.
  - **Fixed Variation:**
    If the fixed variation assumption holds, then the vertical spread in the run sequence plot will be the approximately the same over the entire horizontal axis.

- **Randomness:**
  If the randomness assumption holds, then the lag plot will be structureless and random.
- **Fixed Distribution:**
  If the fixed distribution assumption holds, in particular if the fixed normal distribution holds, then
  - the histogram will be bell-shaped, and
  - the normal probability plot will be linear.

- Conversely, the underlying assumptions are tested using the EDA plots:
  - **Run Sequence Plot:**
    If the run sequence plot is flat and non-drifting, the fixed-location assumption holds. If the run sequence plot has a vertical spread that is about the same over the entire plot, then the fixed-variation assumption holds.
  - **Lag Plot:**
    If the lag plot is structureless, then the randomness assumption holds.
  - **Histogram:**
    If the histogram is bell-shaped, the underlying distribution is symmetric and perhaps approximately normal.
  - **Normal Probability Plot:**
    If the normal probability plot is linear, the underlying distribution is approximately normal.

If all 4 assumptions hold, then the process is said to be "in statistical control".

- If some of the underlying assumptions do not hold, what can be done about it? What corrective actions can be taken?
    - The positive way of approaching this is to view the testing of underlying assumptions as a framework for learning about the process. Assumption-testing promotes insight into important aspects of the process that may not have surfaced otherwise.

- The primary goal is to have correct, validated, and complete scientific/engineering conclusions flowing from the analysis.
    - This usually includes intermediate goals such as the derivation of a good-fitting model and the computation of realistic parameter estimates.
    - It should always include the ultimate goal of an understanding and a "feel" for "what makes the process tick". There is no more powerful catalyst for discovery than the bringing together of an experienced/expert scientist/engineer and a data set ripe with intriguing "anomalies" and characteristics.

- The following sections discuss in more detail the consequences of invalid assumptions:
  - Consequences of non-randomness
  - Consequences of non-fixed location parameter
  - Consequences of non-fixed variation
  - Consequences related to distributional assumptions

(http://www.itl.nist.gov/div898/handbook/eda/section2/eda25.htm)

Consequences of non-randomness

- If the randomness assumption does not hold, then
    - All of the usual statistical tests are invalid.
    - The calculated uncertainties for commonly used statistics become meaningless.
    - The calculated minimal sample size required for a pre-specified tolerance becomes meaningless.
    - The simple model: y = constant + error becomes invalid.
    - The parameter estimates become suspect and non-supportable

Consequences of non-fixed location parameter

- If the run sequence plot does not support the assumption of fixed location, then
  - The location may be drifting.
  - The single location estimate may be meaningless (if the process is drifting).
  - The choice of location estimator (e.g., the sample mean) may be sub-optimal.
  - The usual formula for the uncertainty of the mean:

$$s(\bar{Y}) = \frac{1}{\sqrt{N(N-1)}}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

    may be invalid and the numerical value optimistically small.
  - The location estimate may be poor.
  - The location estimate may be biased.

Consequences of non-fixed variation

- If the run sequence plot does not support the assumption of fixed variation, then

    - The variation may be drifting.
    - The single variation estimate may be meaningless (if the process variation is drifting).
    - The variation estimate may be poor.
    - The variation estimate may be biased.

Consequences related to distributional assumptions

- Problems that may flow from issues wrt distributional assumptions include:

At the level of distributions

- The distribution may be changing.
- The single distribution estimate may be meaningless (if the process distribution is changing).
- The distribution may be markedly non-normal.
- The distribution may be unknown.
- The true probability distribution for the error may remain unknown.

At the level of models

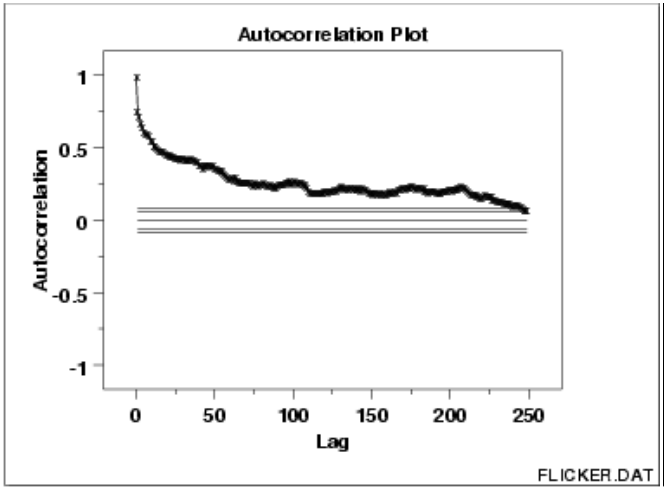
- The model may be changing.
- The single model estimate may be meaningless.
- The default model Y = constant + error may be invalid.
- If the default model is insufficient, information about a better model may remain undetected.
- A poor deterministic model may be fit.
- Information about an improved model may go undetected.

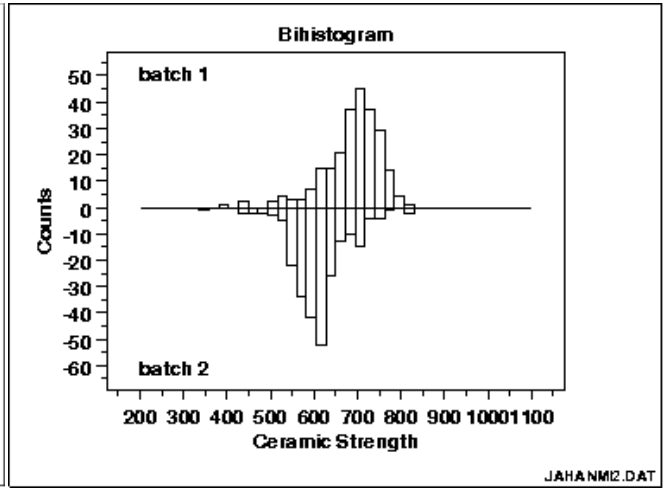
Typical EDA analysis questions

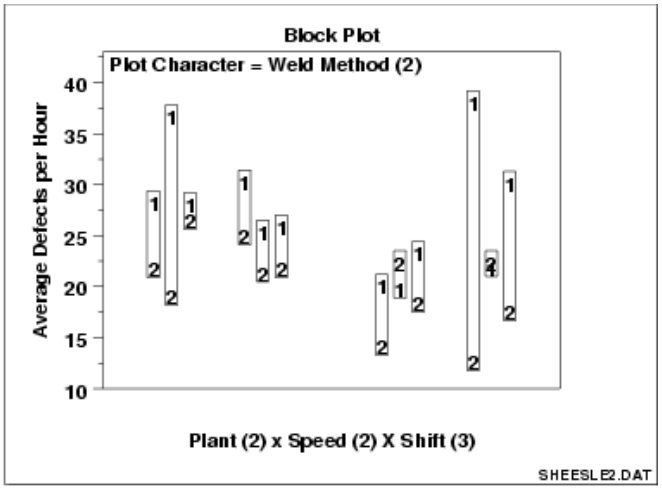
- Some common questions that exploratory data analysis is used to answer are:
  - What is a typical value?
  - What is the uncertainty for a typical value?
  - What is a good distributional fit for a set of numbers?
  - What is a percentile?
  - Does an engineering modification have an effect?
  - Does a factor have an effect?
  - What are the most important factors?
  - Are measurements coming from different laboratories equivalent?
  - What is the best function for relating a response variable to a set of factor variables?
  - What are the best settings for factors?
  - Can we separate signal from noise in time dependent data?
  - Can we extract any structure from multivariate data?
  - Does the data have outliers?

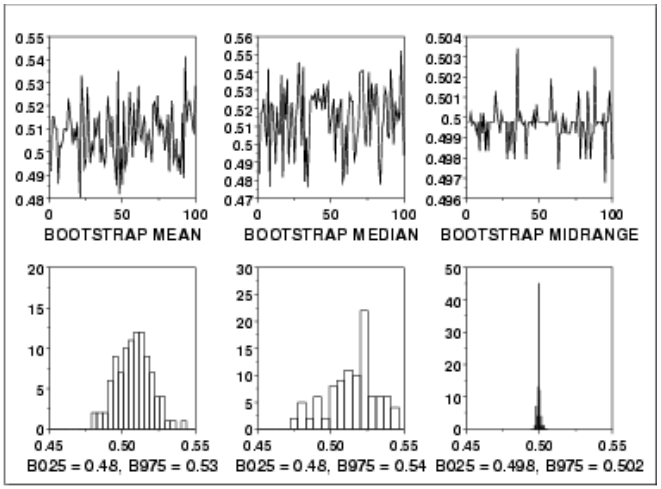
- A critical early step in any analysis is to identify (for the engineering problem at hand) which of the above questions are relevant. That is, we need to identify which questions we want answered and which questions have no bearing on the problem at hand.
- After collecting such a set of questions, an equally important step, which is invaluable for maintaining focus, is to prioritize those questions in decreasing order of importance.
- EDA techniques are tied in with each of the questions. There are some EDA techniques (e.g., the scatter plot) that are broad-brushed and apply almost universally. On the other hand, there are a large number of EDA techniques that are specific and whose specificity is tied in with one of the above questions.

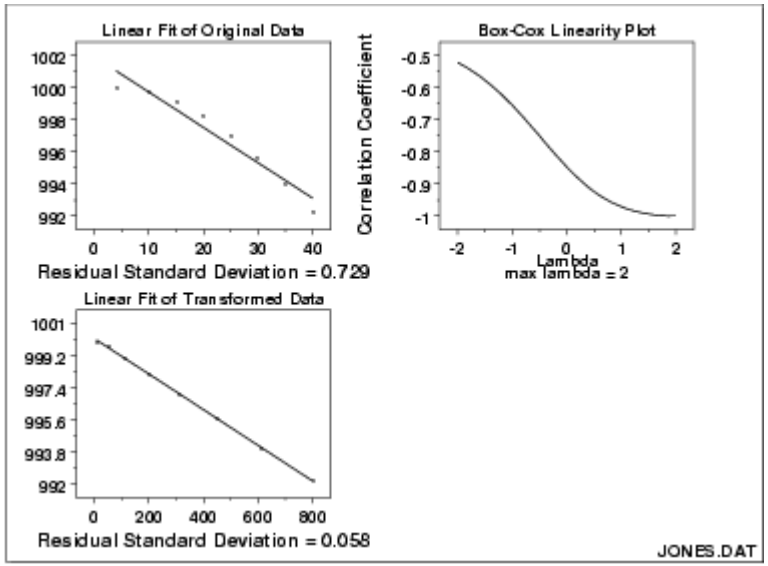
Graphical EDA techniques

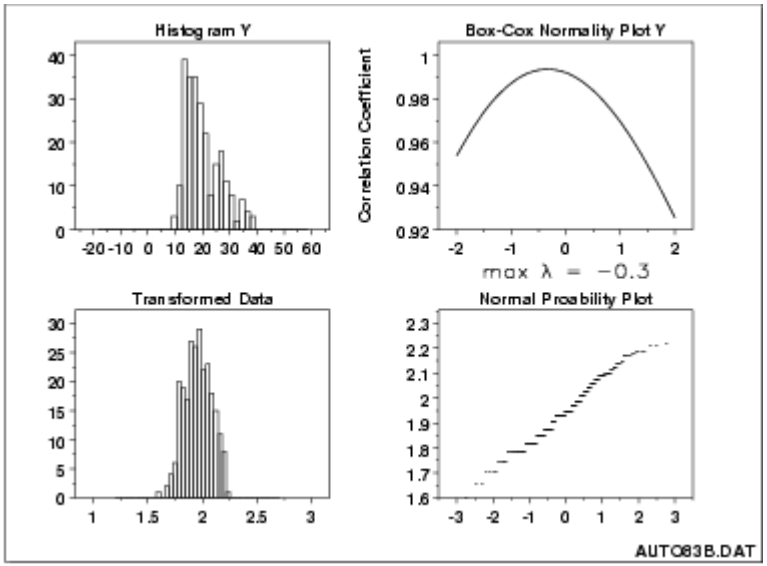


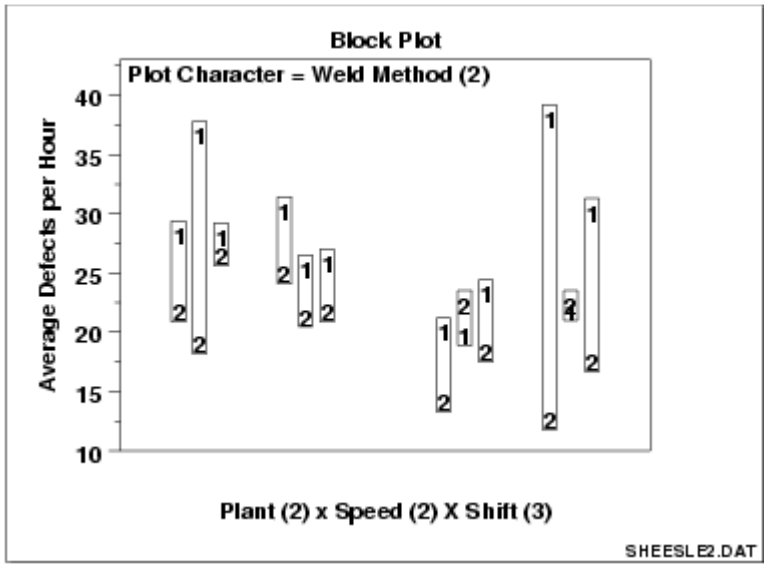


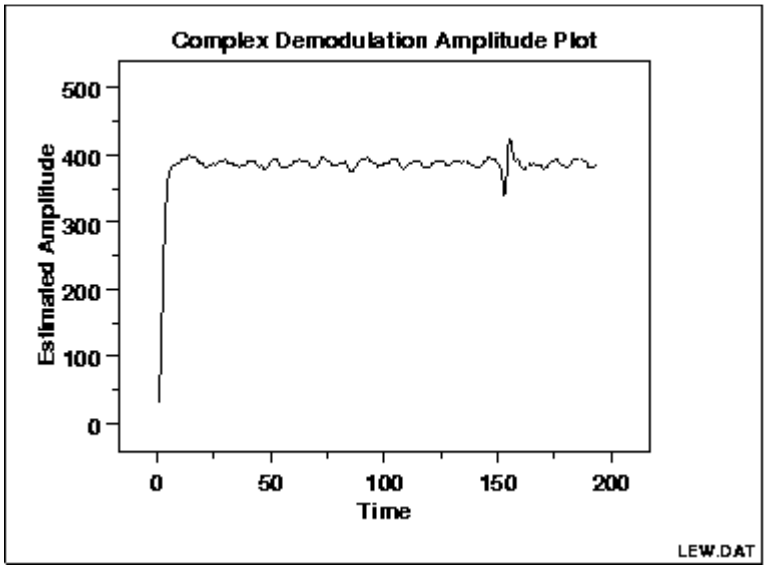


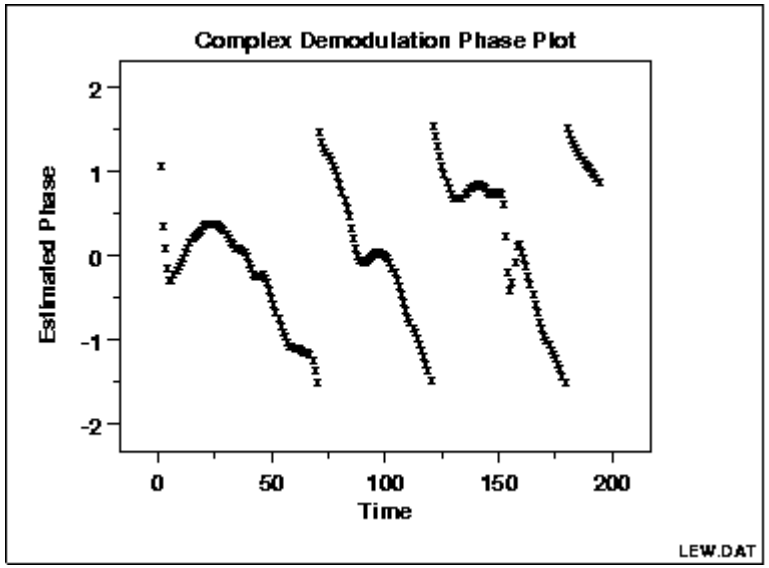


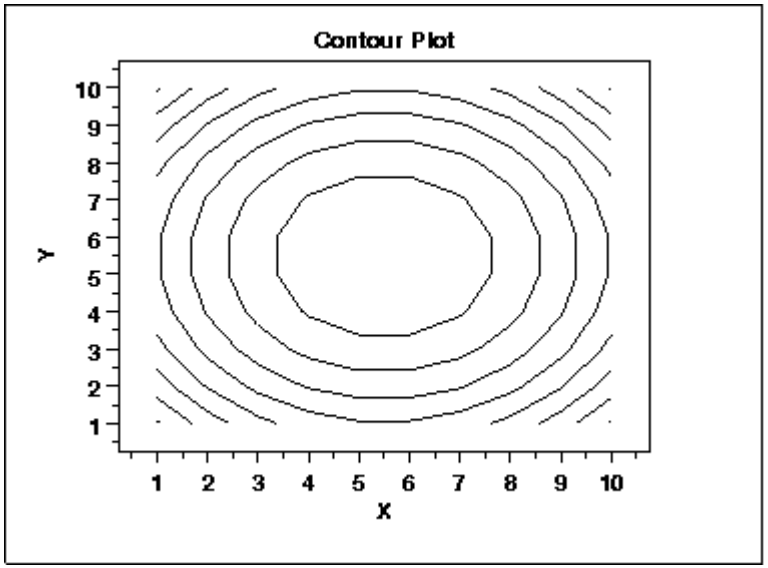

Linear Fit of Original Data
Residual Standard Deviation = 0.729
Box-Cox Linearity Plot
Correlation Coefficient
Lambda
max lambda = 2
Linear Fit of Transformed Data
Residual Standard Deviation = 0.058
JONES.DAT

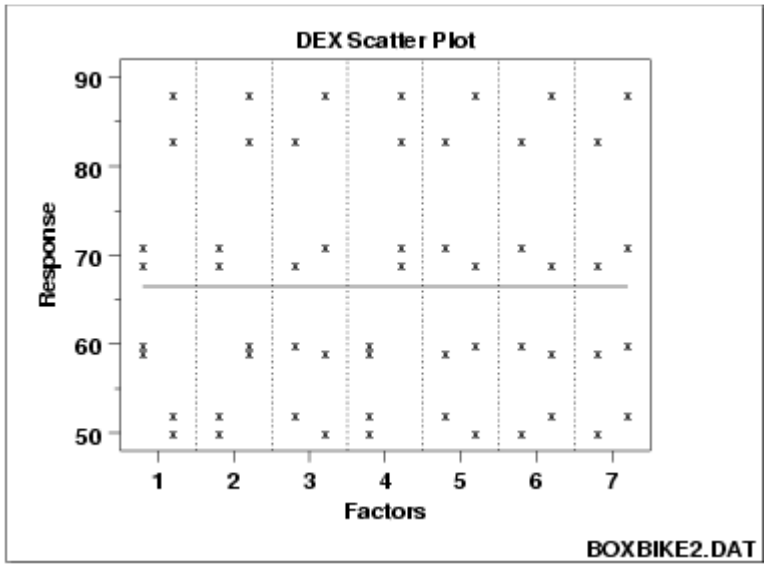

Histogram Y
Box-Cox Normality Plot Y
Correlation Coefficient
max λ = −0.3
Transformed Data
Normal Probability Plot
AUTO83B.DAT

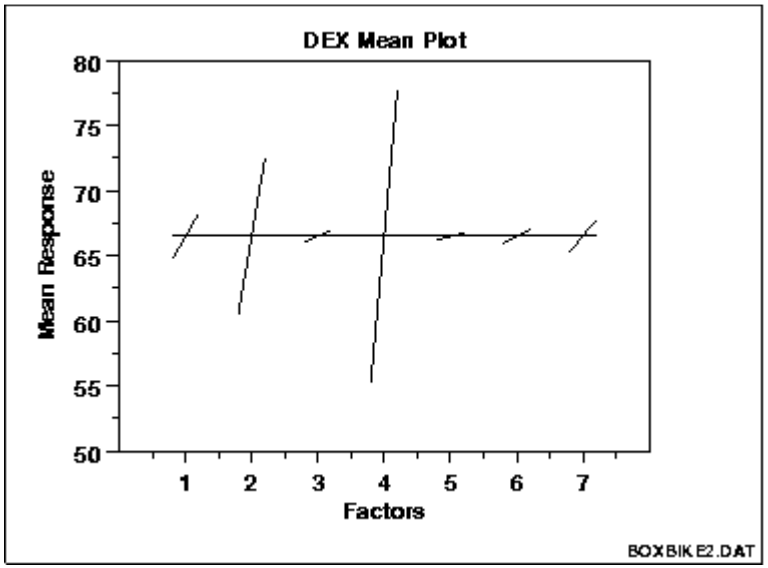

Block Plot
Plot Character = Weld Method (2)
Average Defects per Hour
Plant (2) x Speed (2) X Shift (3)
SHEESLE2.DAT

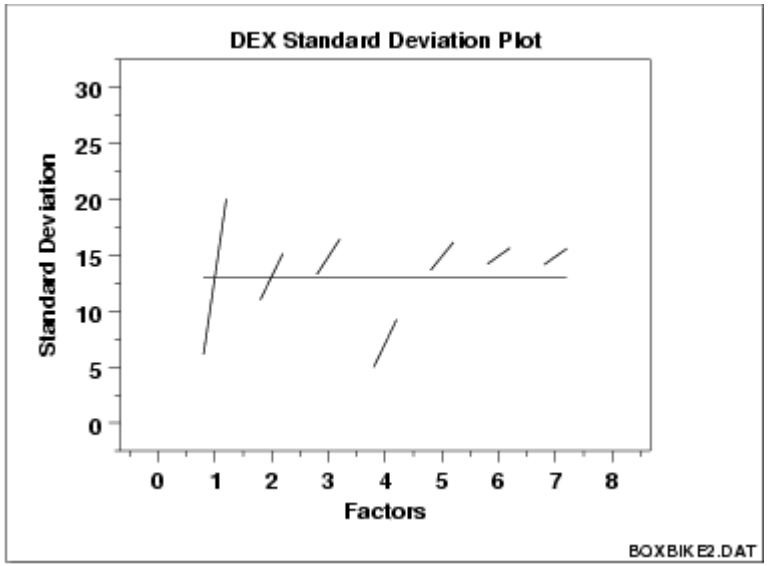

Complex Demodulation Amplitude Plot
Estimated Amplitude
Time
LEW.DAT

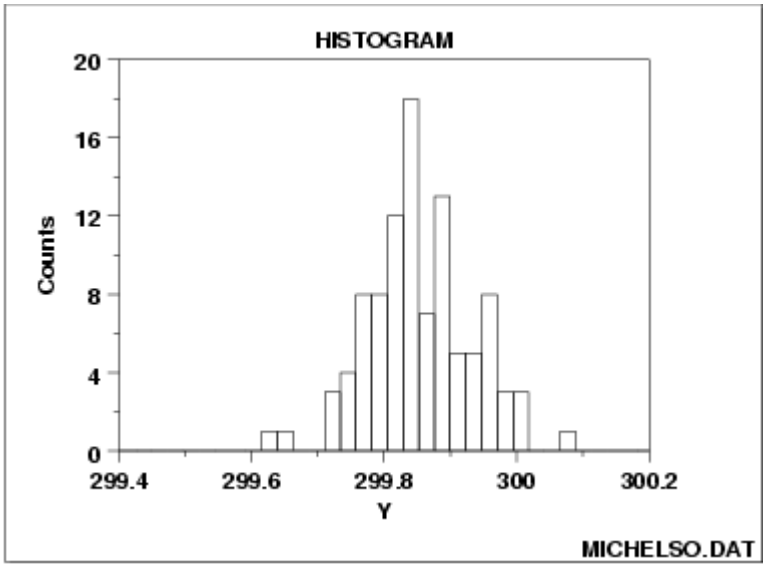

Complex Demodulation Phase Plot
Estimated Phase
Time
LEW.DAT

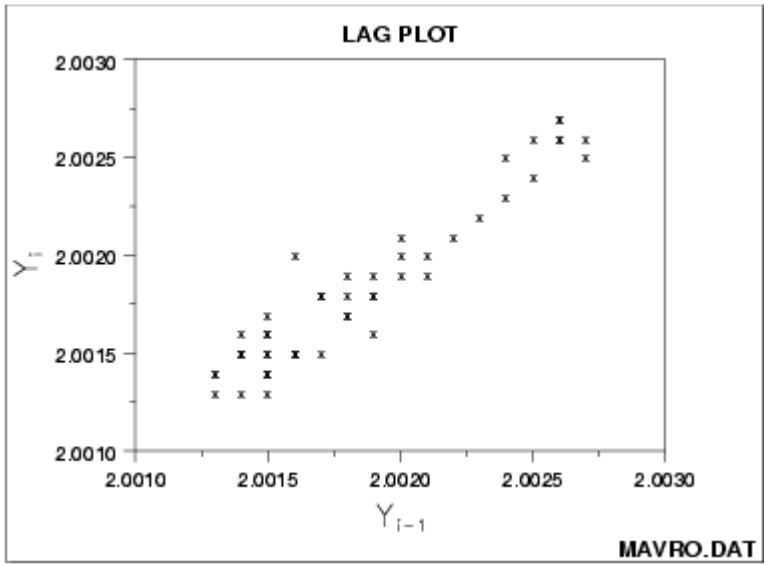

Contour Plot
Y
X

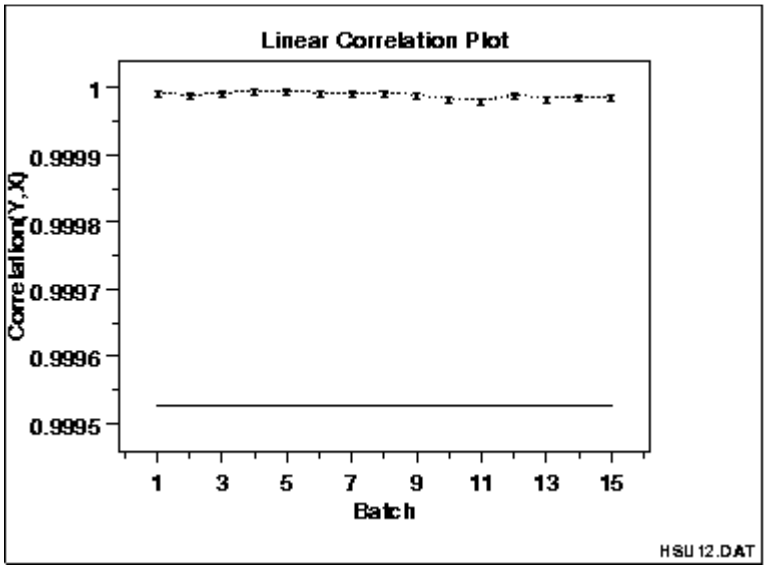

DEX Scatter Plot
Response
Factors
BOXBIKE2.DAT

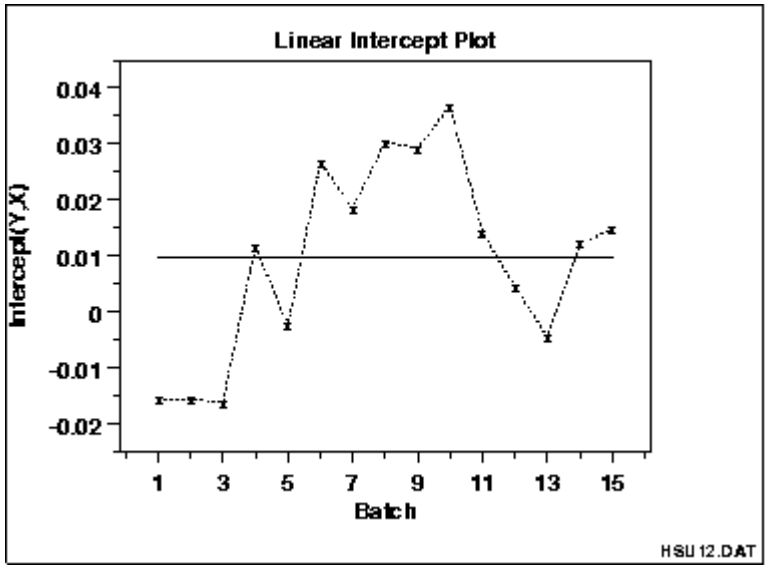

DEX Mean Plot
Mean Response
Factors
BOXBIKE2.DAT

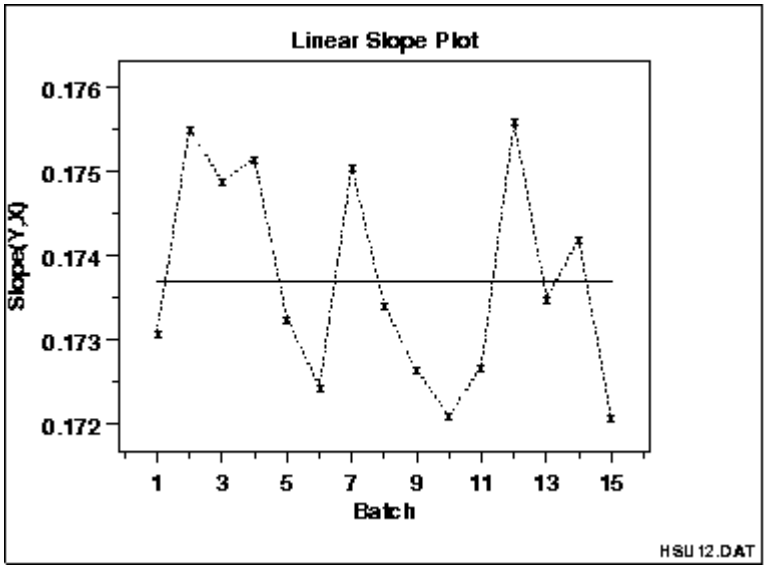

DEX Standard Deviation Plot
Standard Deviation
Factors
BOXBIKE2.DAT

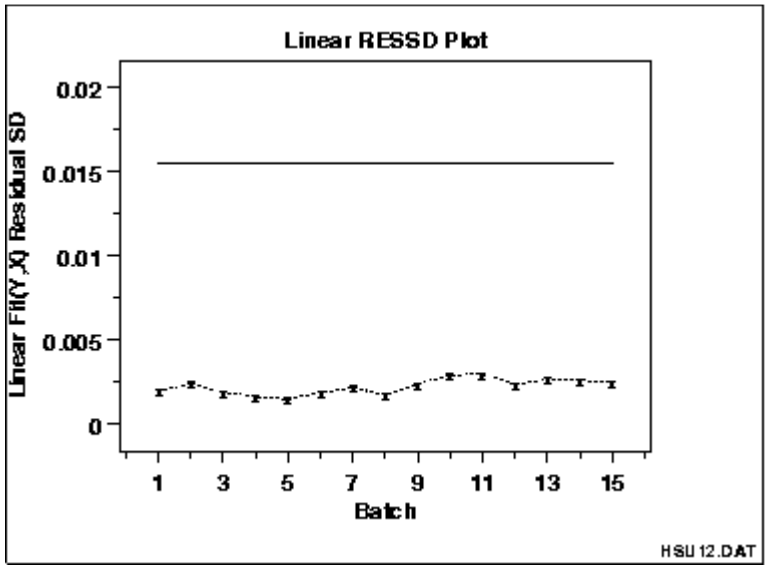

HISTOGRAM
Counts
Y
MICHELSO.DAT

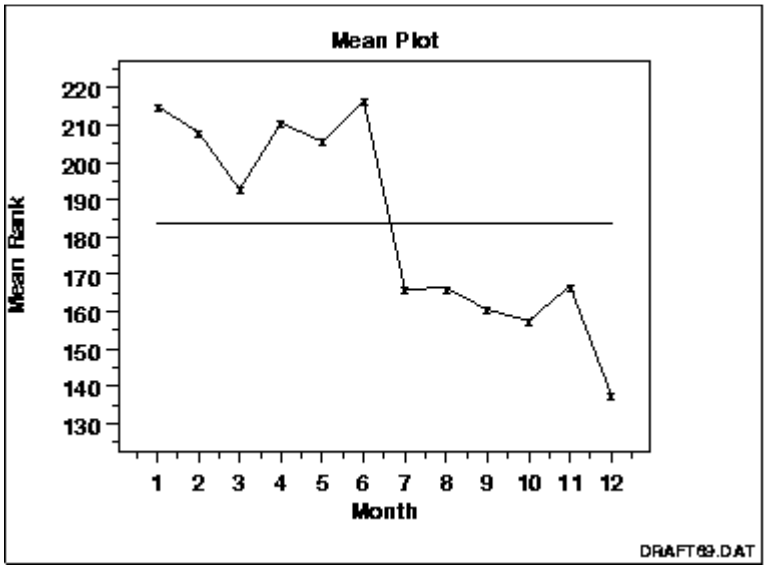

LAG PLOT
MAVRO.DAT

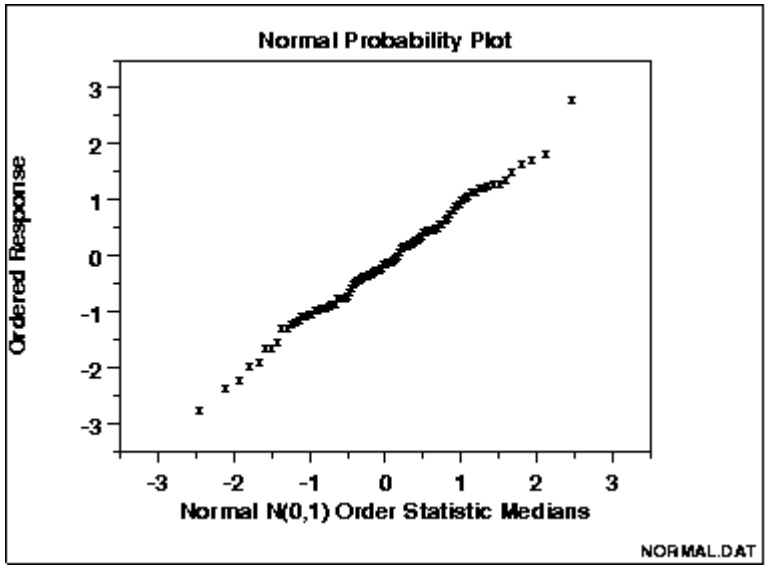

Linear Correlation Plot
Correlation(Y,X)
Batch
HSU12.DAT

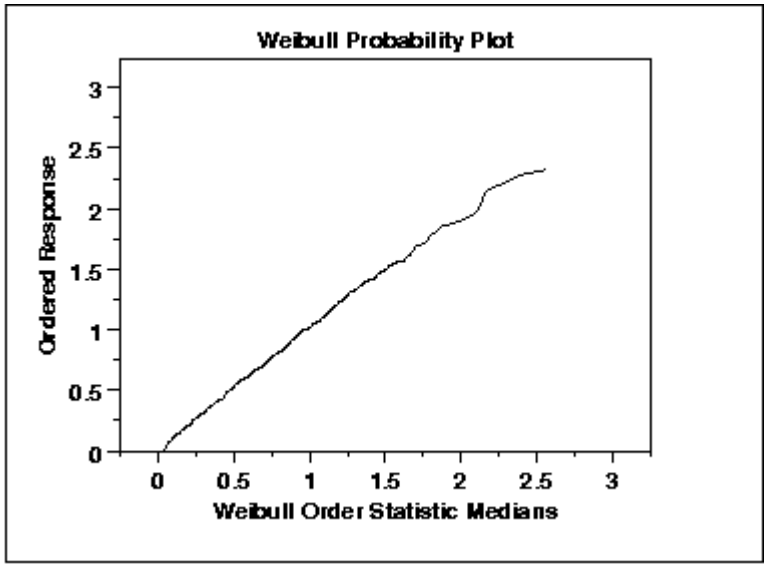

Linear Intercept Plot
Intercept(Y,X)
0.04
0.03
0.02
0.01
0
-0.01
-0.02
1 3 5 7 9 11 13 15
Batch
HSU12.DAT

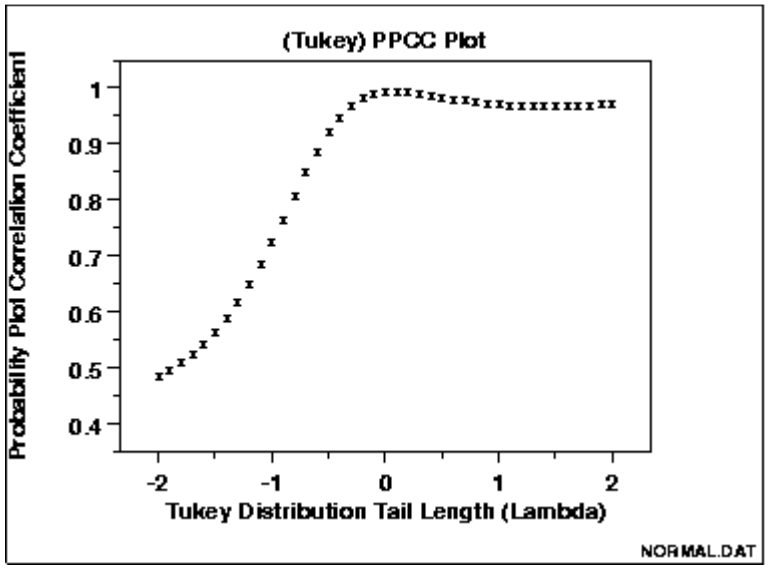

Linear Slope Plot
Slope(Y,X)
0.176
0.175
0.174
0.173
0.172
1 3 5 7 9 11 13 15
Batch
HSU12.DAT

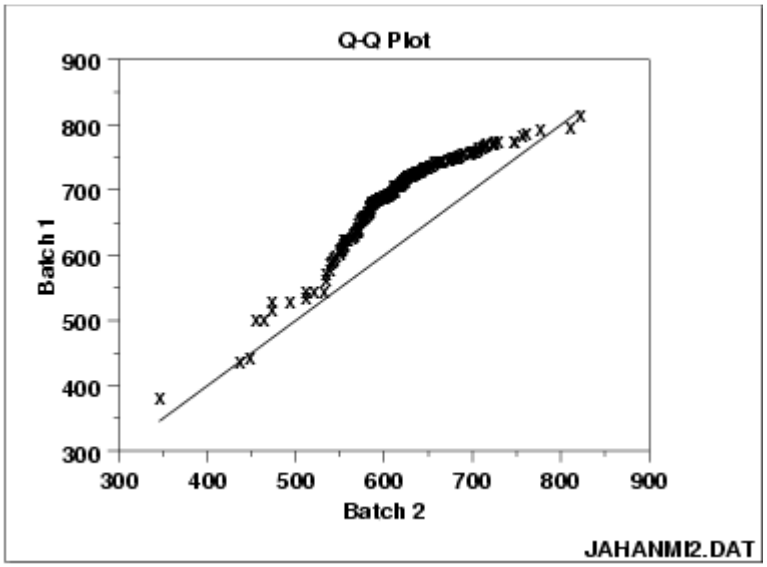

Linear RESSD Plot
Linear Fit(Y,X) Residual SD
0.02
0.015
0.01
0.005
0
1 3 5 7 9 11 13 15
Batch
HSU12.DAT

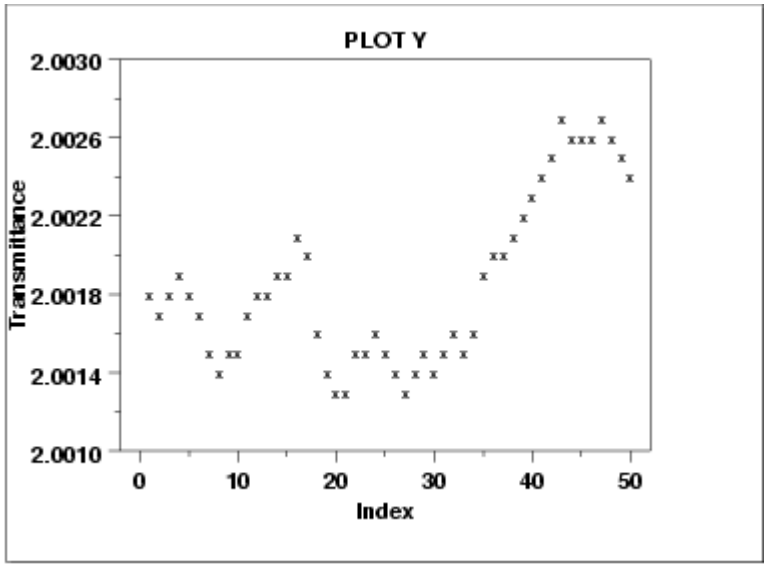

Mean Plot
Mean Rank
220
210
200
190
180
170
160
150
140
130
1 2 3 4 5 6 7 8 9 10 11 12
Month
DRAFT69.DAT

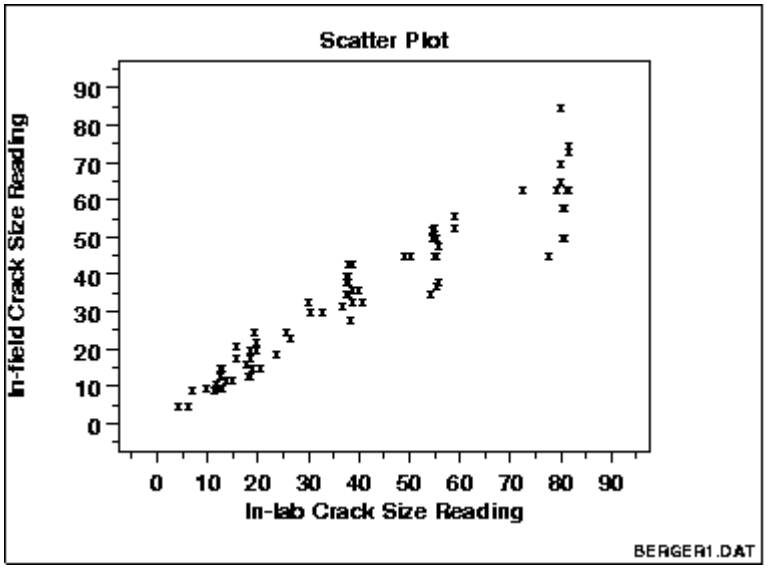

Normal Probability Plot
Ordered Response
Normal N(0,1) Order Statistic Medians
NORMAL.DAT

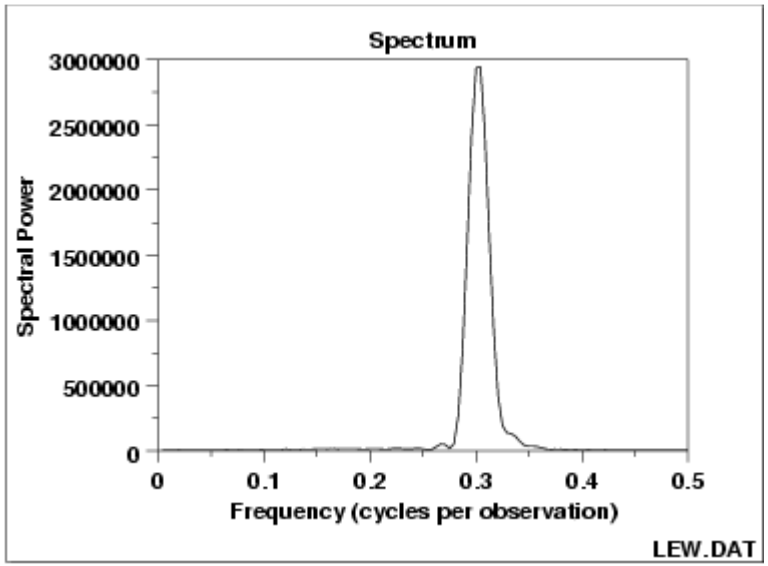

Weibull Probability Plot
Ordered Response
Weibull Order Statistic Medians

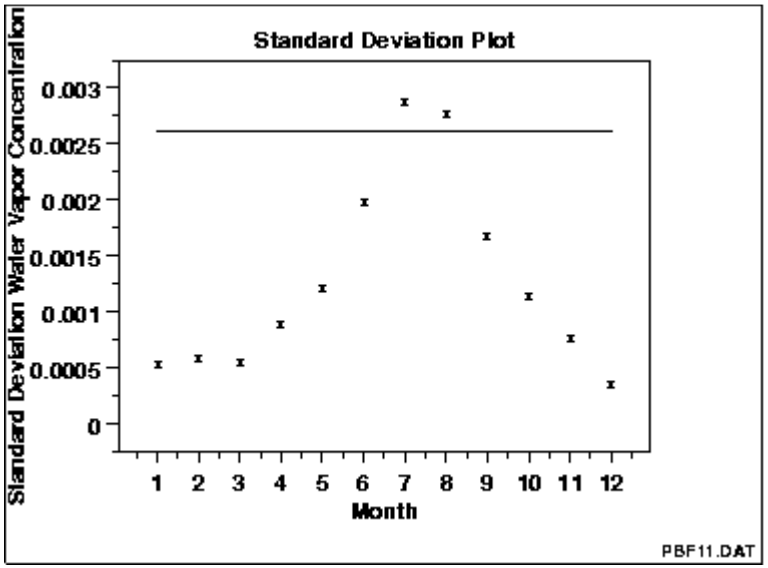

(Tukey) PPCC Plot
Probability Plot Correlation Coefficient
Tukey Distribution Tail Length (Lambda)
NORMAL.DAT

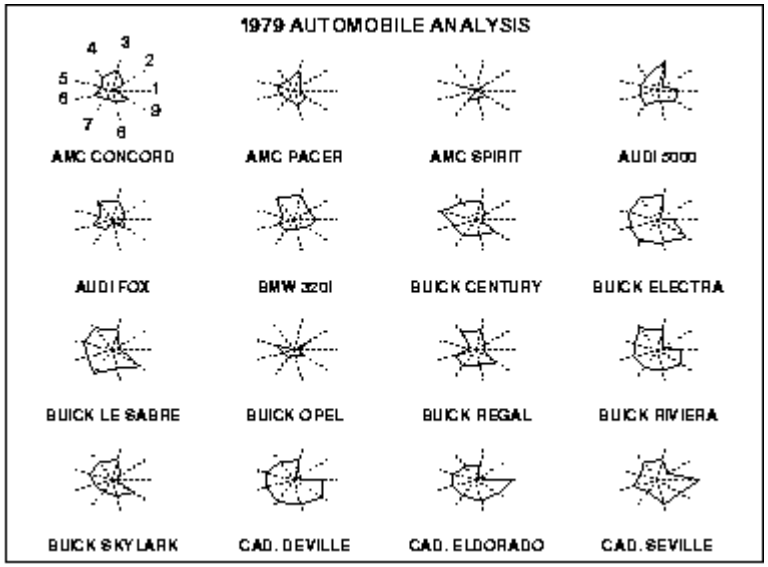

Q-Q Plot
Batch 1
Batch 2
JAHANMI2.DAT

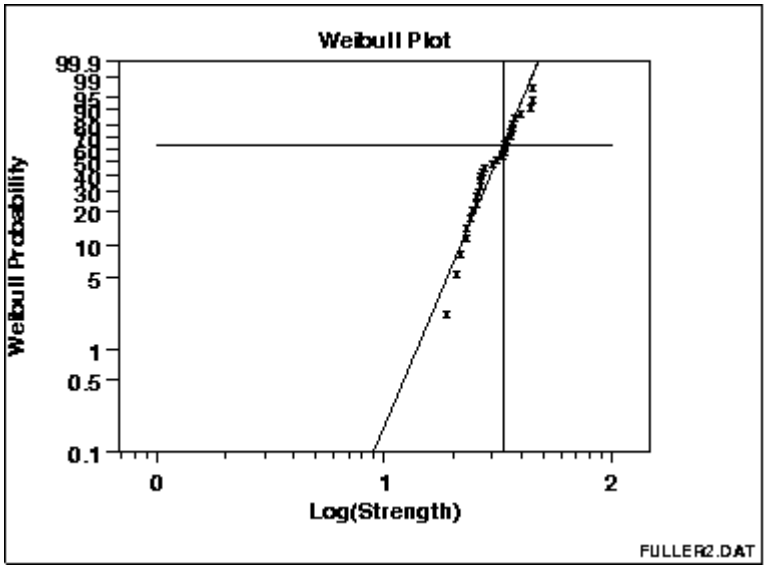

PLOT Y
Transmittance
2.0030
2.0026
2.0022
2.0018
2.0014
2.0010
0
10
20
30
40
50
Index

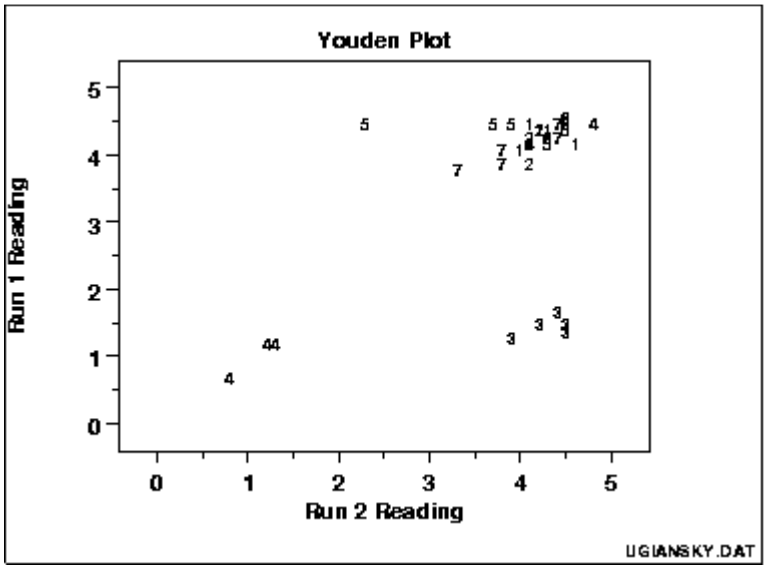

Scatter Plot
In-field Crack Size Reading
90
80
70
60
50
40
30
20
10
0
0 10 20 30 40 50 60 70 80 90
In-lab Crack Size Reading
BERGER1.DAT

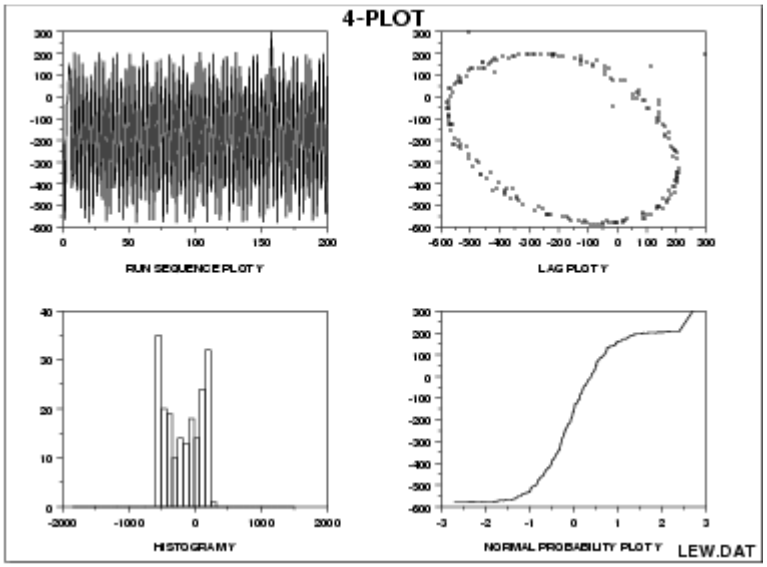

Spectrum
Spectral Power
3000000
2500000
2000000
1500000
1000000
500000
0
0 0.1 0.2 0.3 0.4 0.5
Frequency (cycles per observation)
LEW.DAT

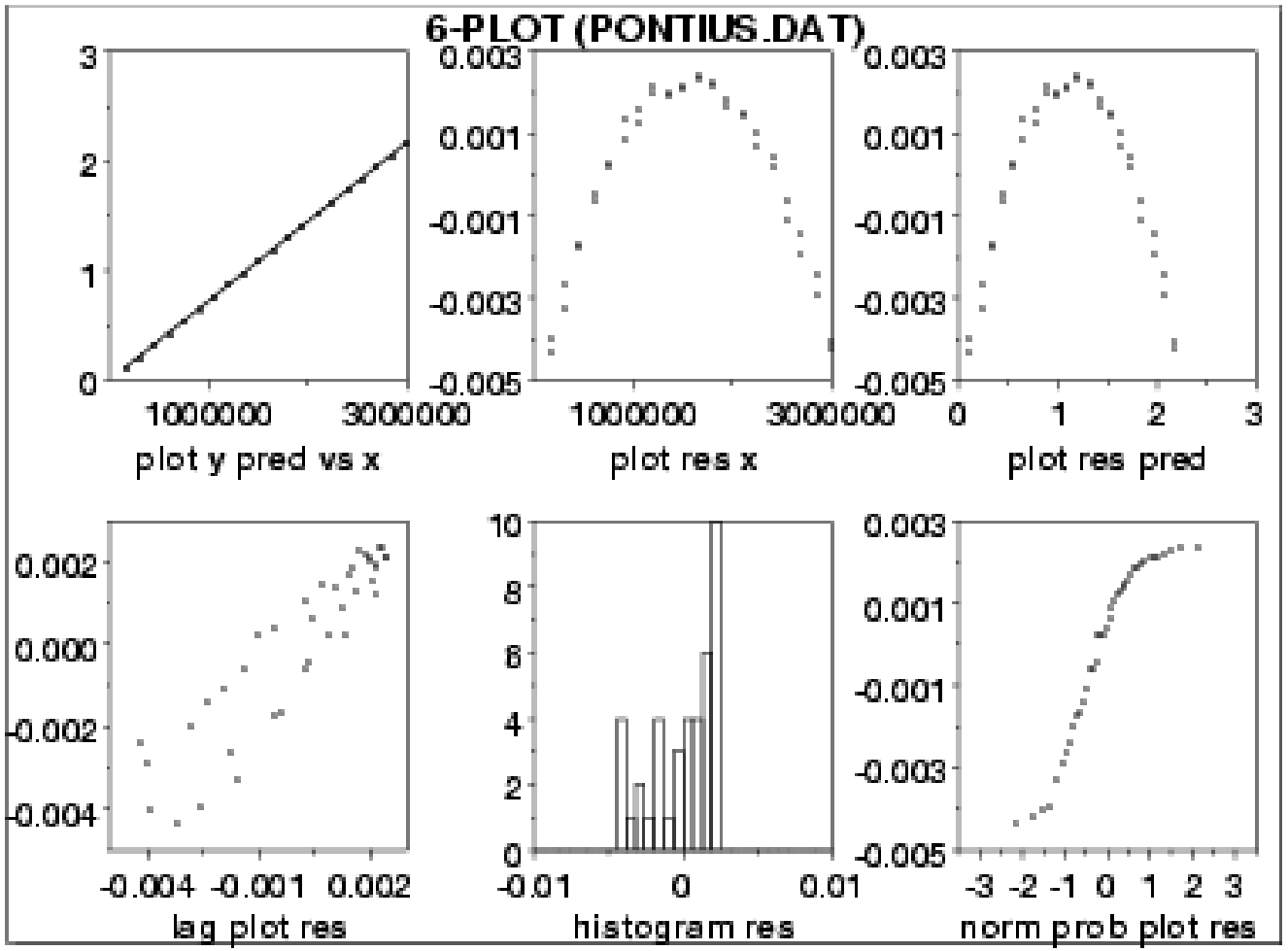

Standard Deviation Plot
Standard Deviation Water Vapor Concentration
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
1 2 3 4 5 6 7 8 9 10 11 12
Month
PBF11.DAT


1979 AUTOMOBILE ANALYSIS
AMC CONCORD
AMC PACER
AMC SPIRIT
AUDI 5000
AUDI FOX
BMW 320I
BUICK CENTURY
BUICK ELECTRA
BUICK LE SABRE
BUICK OPEL
BUICK REGAL
BUICK RIVIERA
BUICK SKYLARK
CAD. DEVILLE
CAD. ELDORADO
CAD. SEVILLE


Weibull Plot
Weibull Probability
99.9
99
20
10
5
1
0.5
0.1
0
1
2
Log(Strength)
FULLER2.DAT

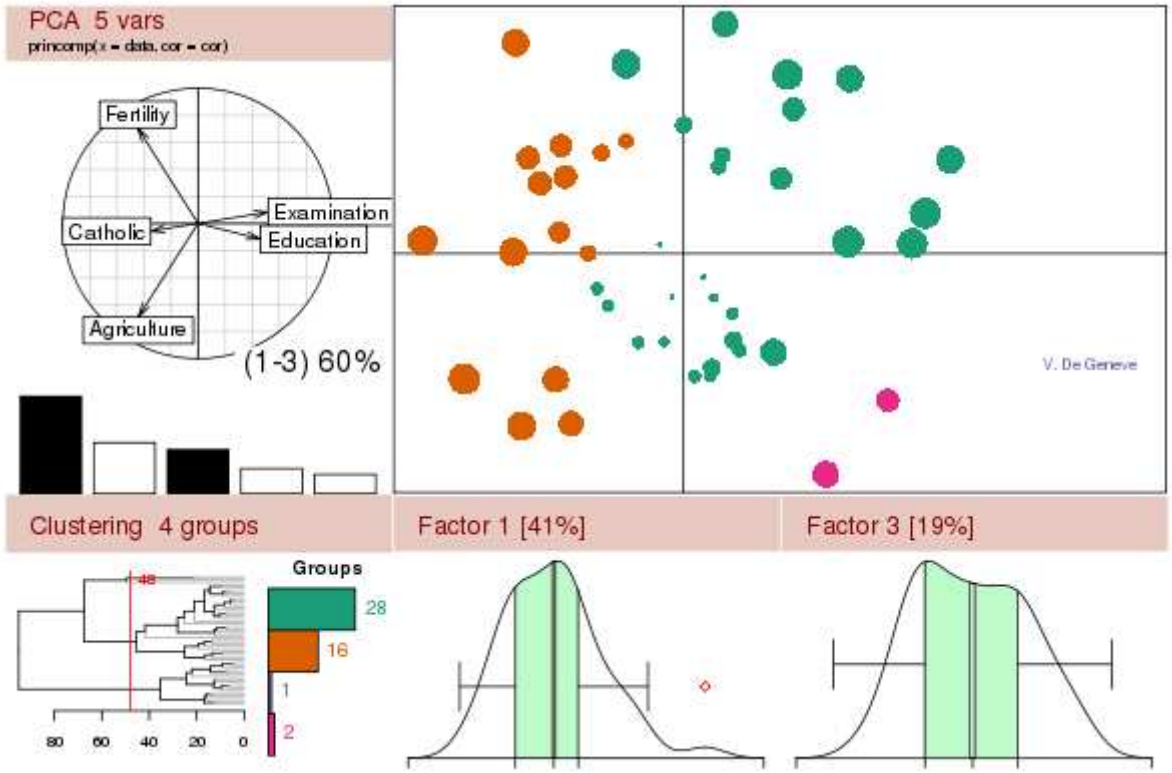

Youden Plot
Run 1 Reading
Run 2 Reading
0
1
2
3
4
5
UGIANSKY.DAT


4-PLOT
RUN SEQUENCE PLOT Y
LAG PLOT Y
HISTOGRAM Y
NORMAL PROBABILITY PLOT Y
LEW.DAT


6-PLOT (PONTIUS.DAT)
plot y pred vs x
plot res x
plot res pred
lag plot res
histogram res
norm prob plot res

- The 6-plot can be used to answer the following questions:
  - Are the residuals approximately normally distributed with a fixed location and scale?
  - Are there outliers?
  - Is the fit adequate?
  - Do the residuals suggest a better fit?
- Time permitting, we will revise this plot in Chapter 8 (model validation)

The R Project for Statistical Computing



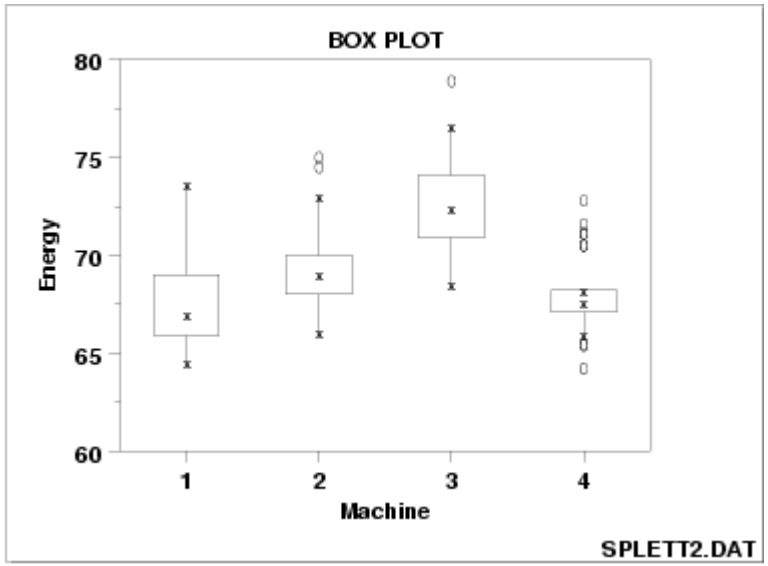
(http://www.r-project.org/)

2.2 Graphical representation of a single variable (univariate)

Introduction

- We have seen in Chapter 2 several ways to look at data, in particular at a single variable
- One additional useful representation is the box plot.
- Box plots (Chambers 1983) are an excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data.

Box plots



This box plot, comparing four machines for energy output, shows that machine has a significant effect on energy with respect to both location and variation. Machine 3 has the highest energy response (about 72.5); machine 4 has the least variable energy response with about 50% of its readings being within 1 energy unit.

- Box plots are formed by
  - Vertical axis: Response variable
  - Horizontal axis: The factor of interest
- More specifically, we
  - Calculate the median and the quartiles (the lower quartile is the 25th percentile and the upper quartile is the 75th percentile).
  - Plot a symbol at the median (or draw a line) and draw a box (hence the name--box plot) between the lower and upper quartiles; this box represents the middle 50% of the data--the "body" of the data.
  - Draw a line from the lower quartile to the minimum point and another line from the upper quartile to the maximum point. Typically a symbol is drawn at these minimum and maximum points, although this is optional.
- Thus the box plot identifies the middle 50% of the data, the median, and the extreme points.

Box plots

- There is a useful variation of the box plot that more specifically identifies outliers. To create this variation:

  - Calculate the median and the lower and upper quartiles.
  - Plot a symbol at the median and draw a box between the lower and upper quartiles.
  - Calculate the interquartile range (the difference between the upper and lower quartile) and call it IQ.
  - Calculate the following points:

    L1 = lower quartile - 1.5*IQ

    L2 = lower quartile - 3.0*IQ

    U1 = upper quartile + 1.5*IQ

    U2 = upper quartile + 3.0*IQ

- The line from the lower quartile to the minimum is now drawn from the lower quartile to the smallest point that is greater than L1. Likewise, the line from the upper quartile to the maximum is now drawn to the largest point smaller than U1.
- Points between L1 and L2 or between U1 and U2 are drawn as small circles. Points less than L2 or greater than U2 are drawn as large circles.

Uses of box plots

- The box plot can provide answers to the following questions:
  - Is a factor significant?
  - Does the location differ between subgroups?
  - Does the variation differ between subgroups?
  - Are there any outliers?
- Box plots are related to mean plots and Analysis of Variance (ANOVA – a regression technique for categorical predictor variables)
- Not however that mean plots are typically used in conjunction with standard deviation plots. The mean plot checks for shifts in location while the standard deviation plot checks for shifts in scale.

Single or multiple box plots

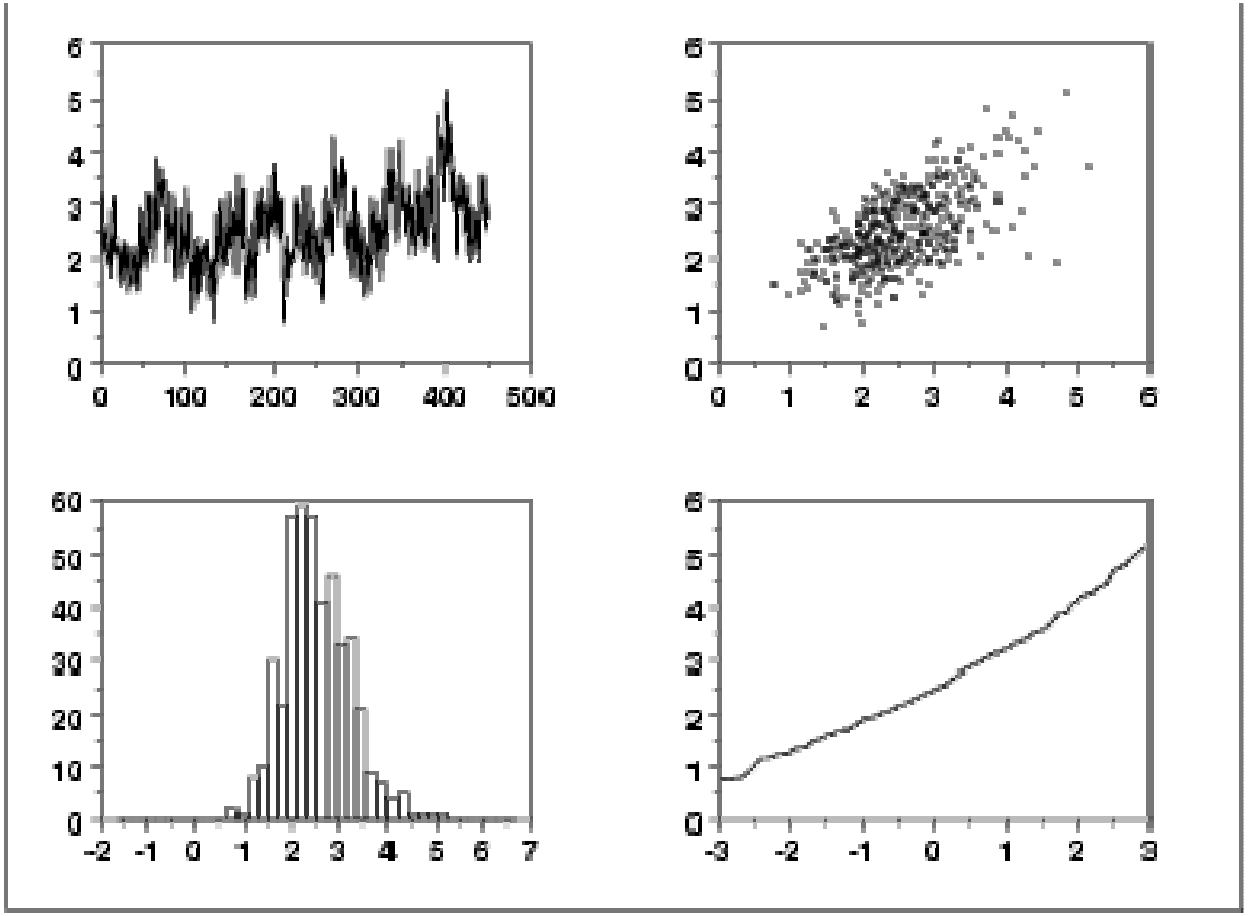
- A single box plot can be drawn for one batch of data with no distinct groups. Alternatively, multiple box plots can be drawn together to compare multiple data sets or to compare groups in a single data set.
- For a single box plot, the width of the box is arbitrary.
- For multiple box plots, the width of the box plot can be set proportional to the number of points in the given group or sample (some software implementations of the box plot simply set all the boxes to the same width).

2.3 Graphical representation of two variables (bivariate)

Introduction

- The first step in analyzing the (bivariate) data is to generate some simple plots of the *response* (outcome of interest) and then of the response versus the various *predictor variables, explanatory variables* (which may or may not be *factors*).

Four-plot of data

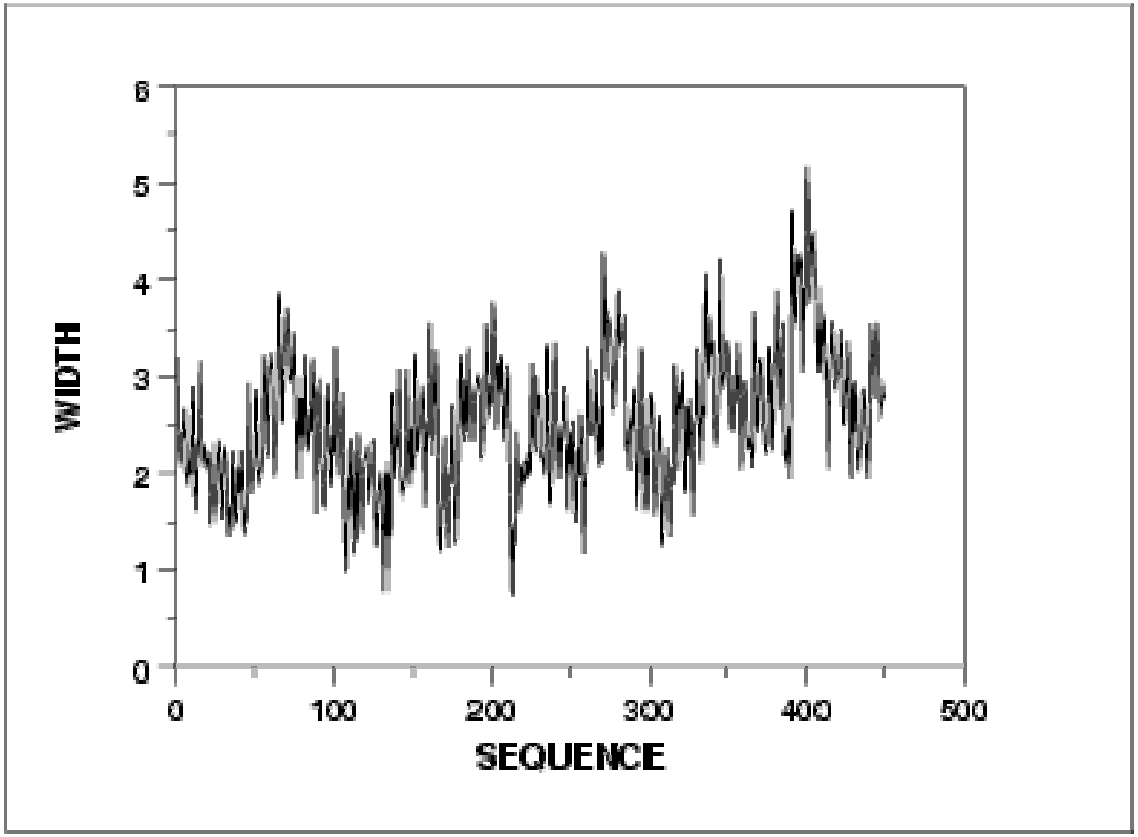


Interpretation

- This 4-plot shows the following.
    - The run sequence plot (upper left) indicates that the location and scale are not constant over time. This indicates that the three factors do in fact have an effect of some kind.
    - The lag plot (upper right) indicates that there is some mild autocorrelation in the data. This is not unexpected as the data are grouped in a logical order of the three factors (i.e., not randomly) and the run sequence plot indicates that there are factor effects.
    - The histogram (lower left) shows that most of the data fall between 1 and 5, with the center of the data at about 2.2.
    - Due to the non-constant location and scale and autocorrelation in the data, distributional inferences from the normal probability plot (lower right) are not meaningful.

Run sequence plot

- The run sequence plot is shown at full size to show greater detail. In addition, a numerical summary of the data is generated.



Numerical summary

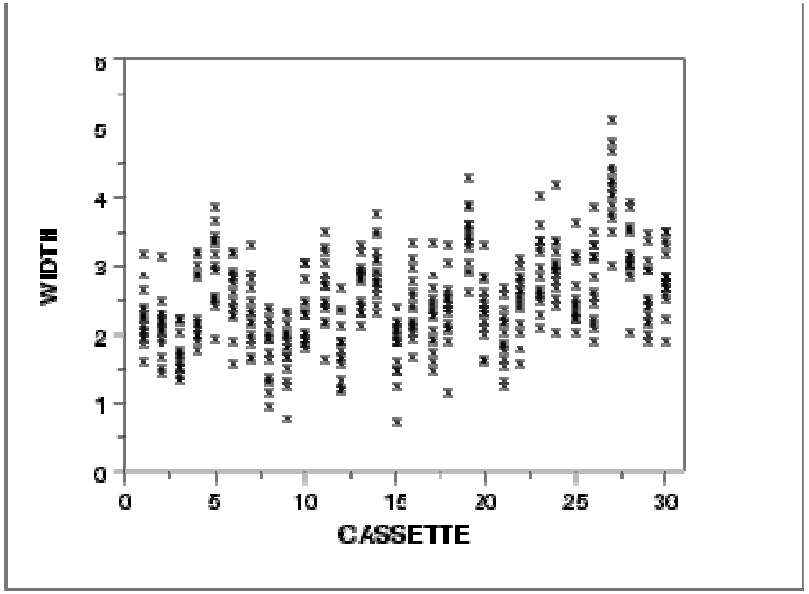
```
                    NUMBER OF OBSERVATIONS =      450

**************************************************************************
*        LOCATION MEASURES          *       DISPERSION MEASURES          *
**************************************************************************
*  MIDRANGE     =   0.2957607E+01  *  RANGE        =   0.4422122E+01  *
*  MEAN         =   0.2532284E+01  *  STAND. DEV.  =   0.6937559E+00  *
*  MIDMEAN      =   0.2393183E+01  *  AV. AB. DEV. =   0.5482042E+00  *
*  MEDIAN       =   0.2453337E+01  *  MINIMUM      =   0.7465460E+00  *
*               =                  *  LOWER QUART. =   0.2046285E+01  *
*               =                  *  LOWER HINGE  =   0.2048139E+01  *
*               =                  *  UPPER HINGE  =   0.2971948E+01  *
*               =                  *  UPPER QUART. =   0.2987150E+01  *
*               =                  *  MAXIMUM      =   0.5168668E+01  *
**************************************************************************
*       RANDOMNESS MEASURES         *     DISTRIBUTIONAL MEASURES       *
**************************************************************************
*  AUTOCO COEF  =   0.6072572E+00  *  ST. 3RD MOM. =   0.4527434E+00  *
*               =   0.0000000E+00  *  ST. 4TH MOM. =   0.3382735E+01  *
*               =   0.0000000E+00  *  ST. WILK-SHA =   0.6957975E+01  *
*               =                  *  UNIFORM PPCC =   0.9681802E+00  *
*               =                  *  NORMAL  PPCC =   0.9935199E+00  *
*               =                  *  TUK -.5 PPCC =   0.8528156E+00  *
*               =                  *  CAUCHY  PPCC =   0.5245036E+00  *
**************************************************************************
```

- This summary generates a variety of statistics. In this case, we are primarily interested in the mean and standard deviation. From this summary, we see that the mean is 2.53 and the standard deviation is 0.69.

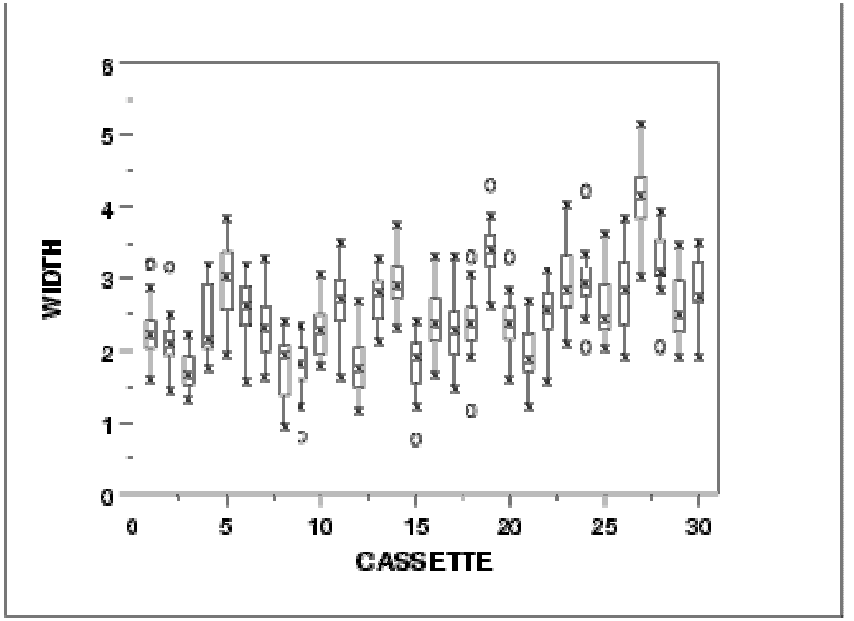
Plot response against individual factors

- The next step is to plot the response against each individual factor. For comparison, we generate both a scatter plot and a box plot of the data. The scatter plot shows more detail. However, comparisons are usually easier to see with the box plot, particularly as the number of data points and groups become larger

Scatter plot of width vs cassette



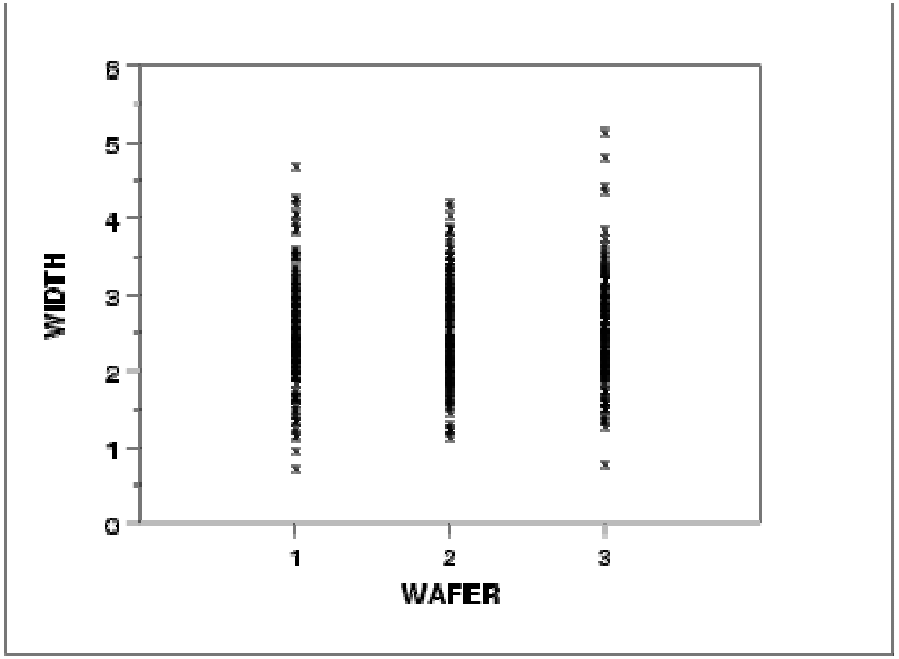
Box plot of width vs cassette



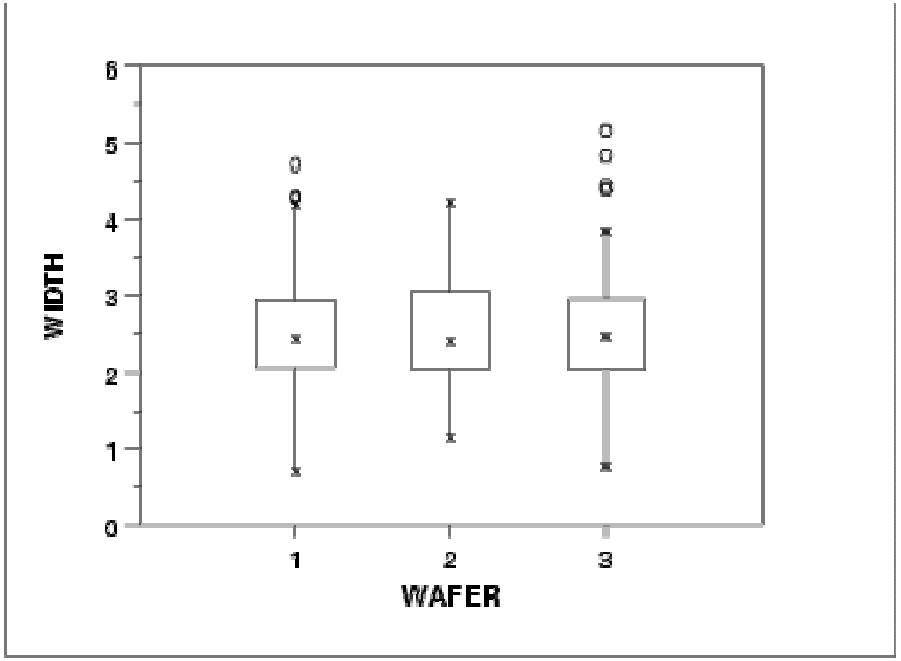
Interpretation

- We can make the following conclusions based on the above scatter and box plots.
  - There is considerable variation in the location for the various cassettes.
  - The medians vary from about 1.7 to 4.
  - There is also some variation in the scale.
  - There are a number of outliers.

Scatter plot of width vs wafer



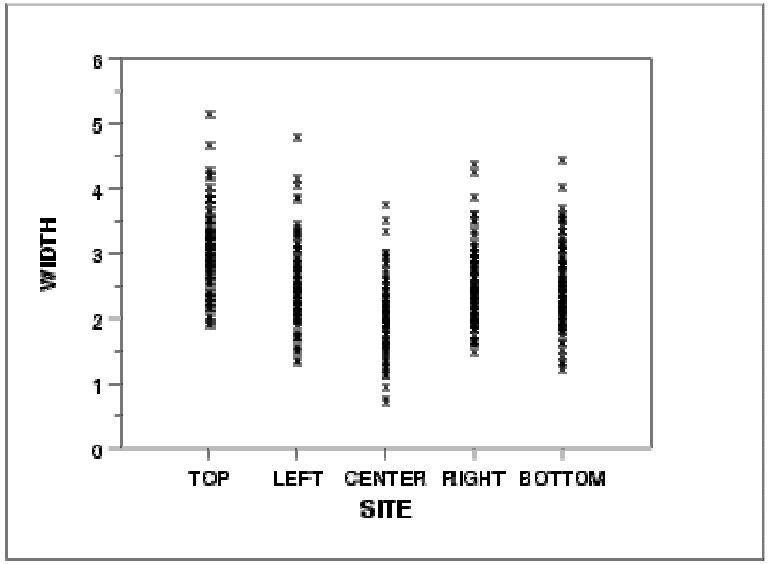
Box plot of width vs wafer



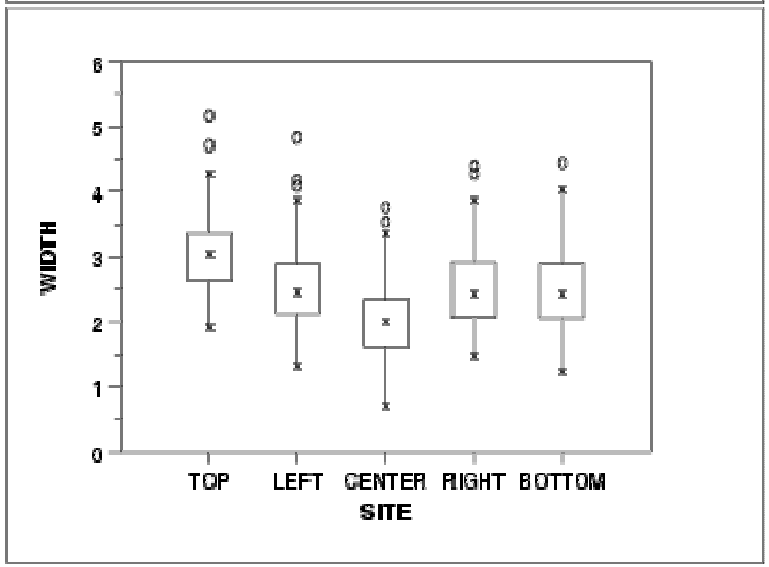
Interpretation

- We can make the following conclusions based on the above scatter and box plots.
  - The locations for the 3 wafers are relatively constant.
  - The scales for the 3 wafers are relatively constant.
  - There are a few outliers on the high side.
  - It is reasonable to treat the wafer factor as homogeneous.

Scatter plot of width vs site



Box plot of width vs site

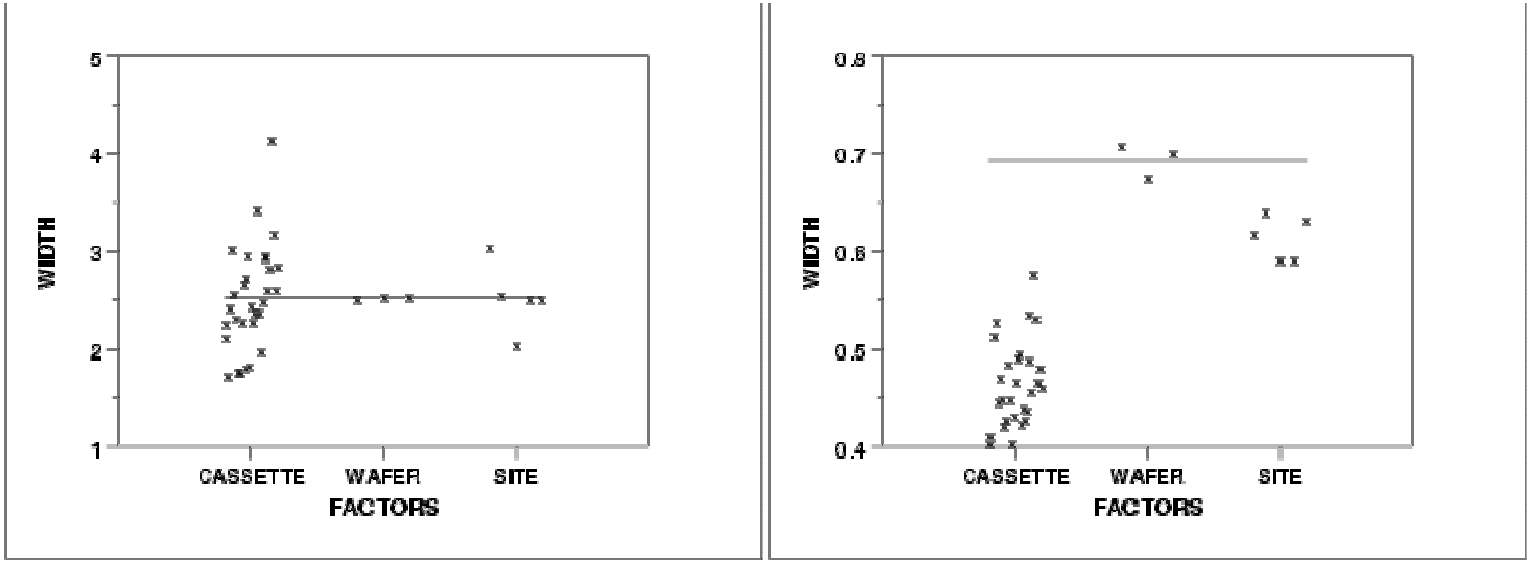


Interpretation

- We can make the following conclusions based on the above scatter and box plots.
  - There is some variation in location based on site. The center site in particular has a lower median.
  - The scales are relatively constant across sites.
  - There are a few outliers.

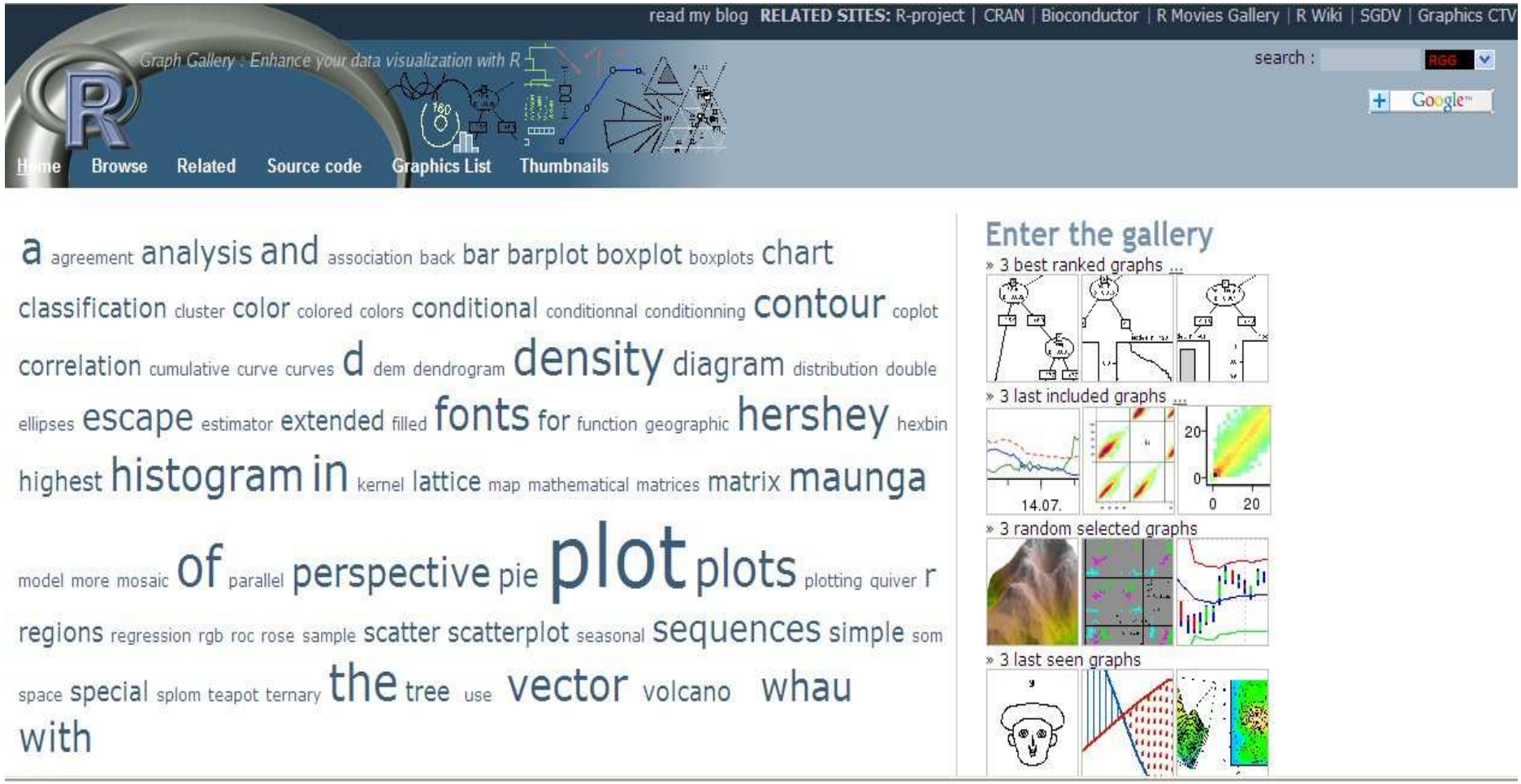
Dex mean and dex sd plot

- We can use the dex mean plot (left) and the dex standard deviation plot (right) to show the factor means and standard deviations together for better comparison.



- While summarizing all of the above graphs, we can postulate that there are differences between the lots and the sites.

2.4 Graphical representation of multiple variables (multivariate)



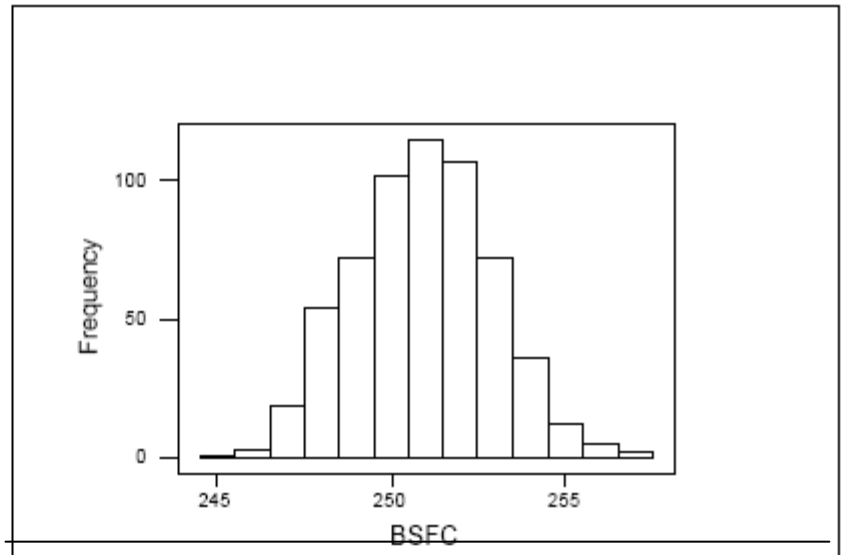
(http://addictedtor.free.fr/graphiques/)

2.5 Things to ALWAYS look out for

Bar charts and histograms provide an easily understood illustration of the distribution of the data. As well as showing where most observations lie and how variable the data are, they also indicate certain "danger signals" about the data.

Normally distributed data

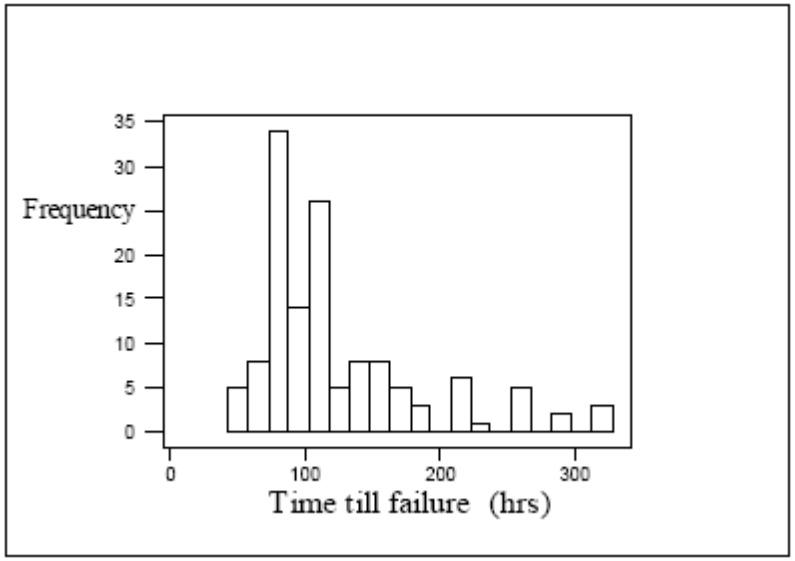
The histogram is bell-shaped, like the probability density function of a Normal distribution. It appears, therefore, that the data can be modelled by a Normal distribution. (Other methods for checking this assumption are available.)



Similarly, the histogram can be used to see whether data look as if they are from an Exponential or Uniform distribution.

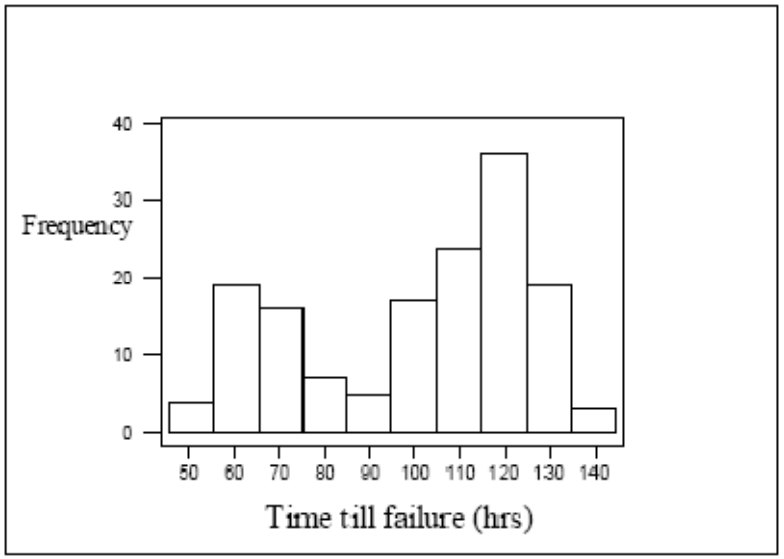
Very skew data

The relatively few large observations can have an undue influence when comparing two or more sets of data. It might be worthwhile using a transformation e.g. taking logarithms.



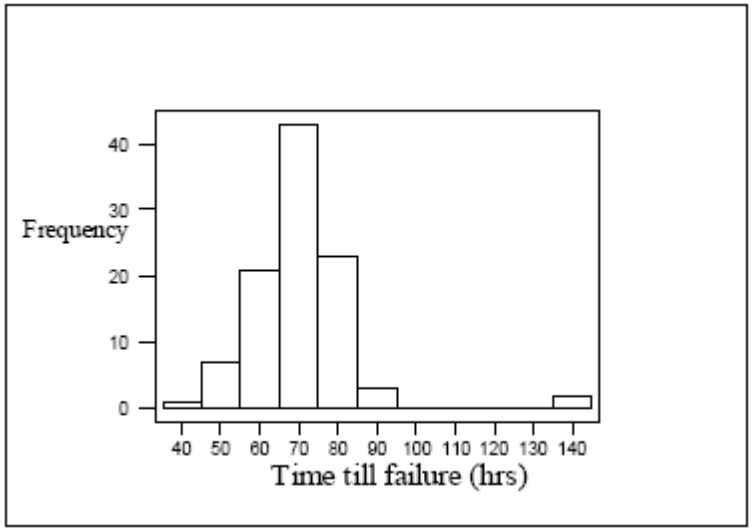
Bimodality

This may indicate the presence of two sub-populations with different characteristics. If the subpopulations can be identified it might be better to analyse them separately.



Outliers

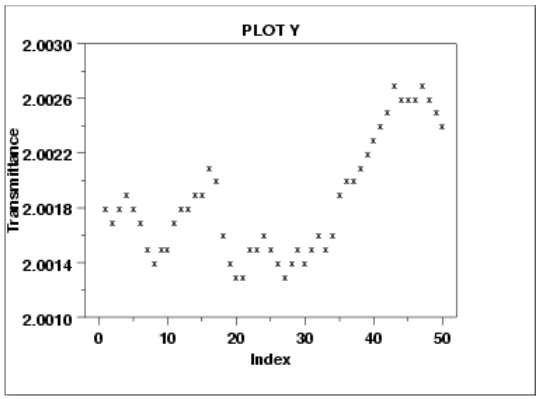
The data appear to follow a pattern with the exception of one or two values. You need to decide whether the strange values are simply mistakes, are to be expected or whether they are correct but unexpected. The outliers may have the most interesting story to tell.



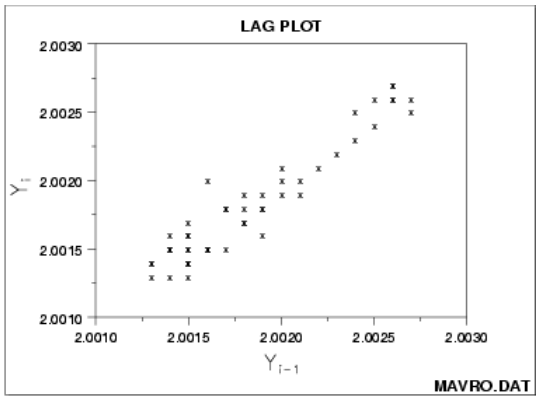
3Highlight: Graphical techniques according to problem identification

3.1 Univariate

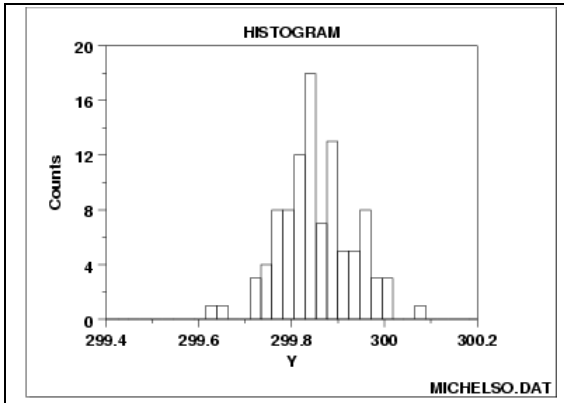
$$y = c + e$$



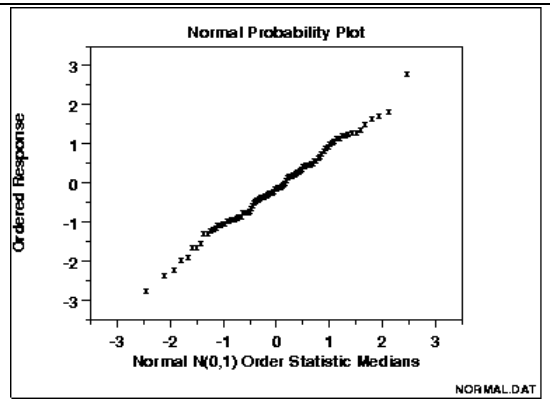
a: Run Sequence Plot



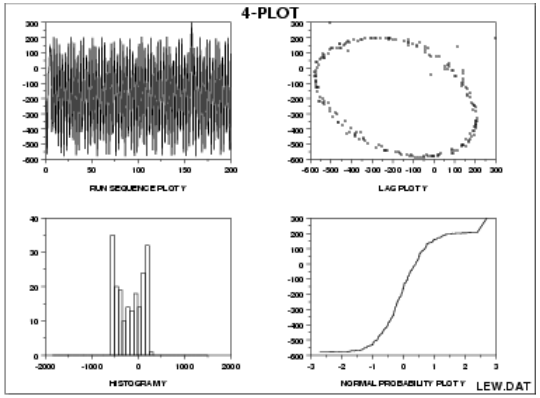
b: Lag plot



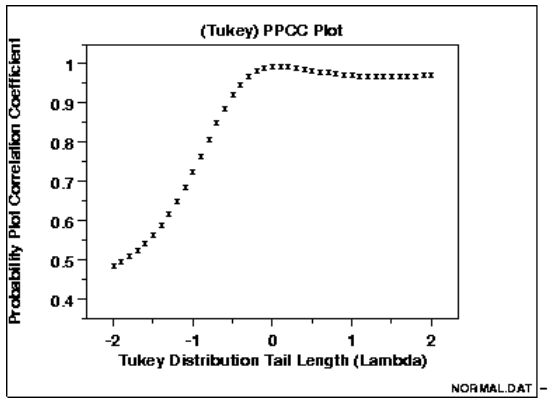
c: Histogram



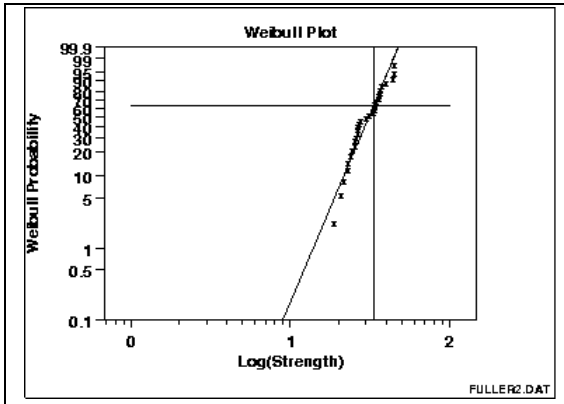
d: Normal Probability Plot



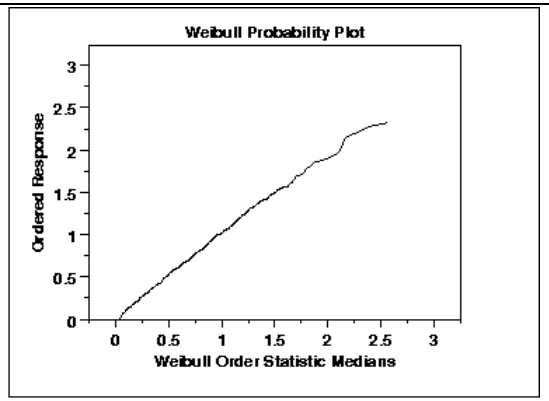
e: 4 – Plot



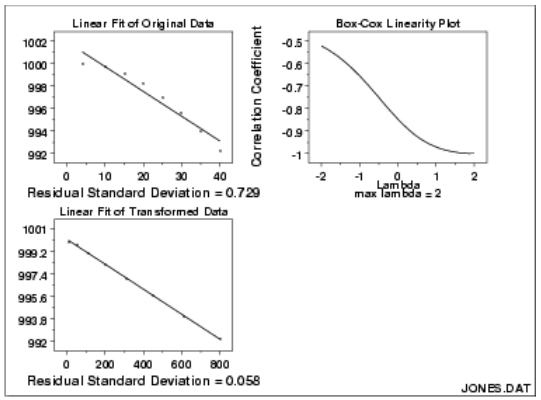
f: PPCC Plot



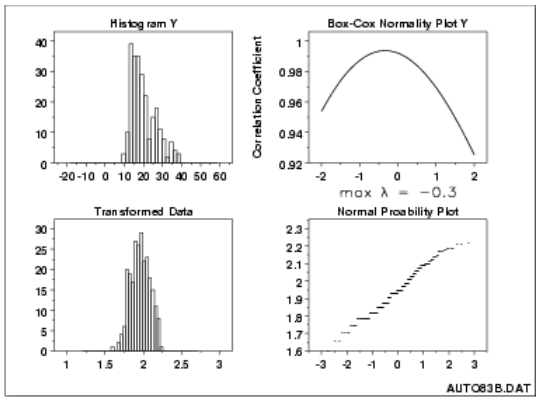
g: Weibull Plot



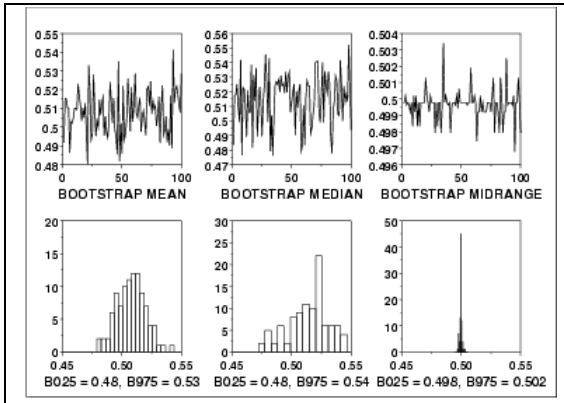
h: Probability Plot



i: Box-Cox Linearity Plot



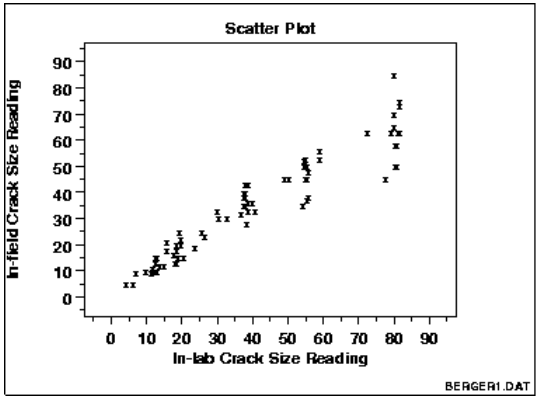
j: Box-Cox Normality Plot

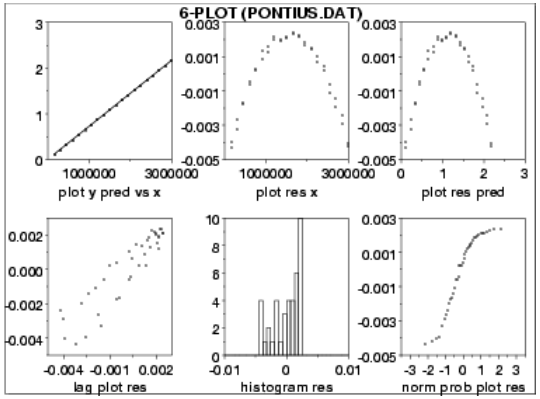


k: Bootstrap Plot

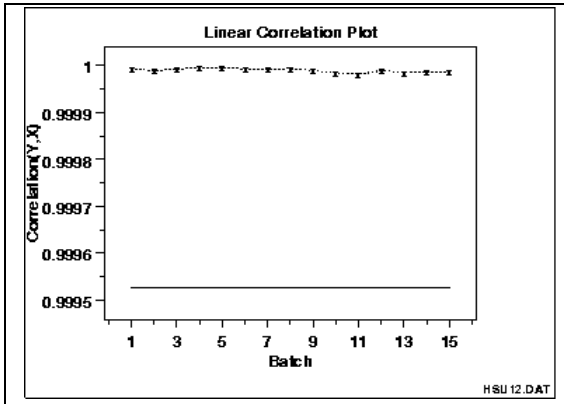
3.2 Regression

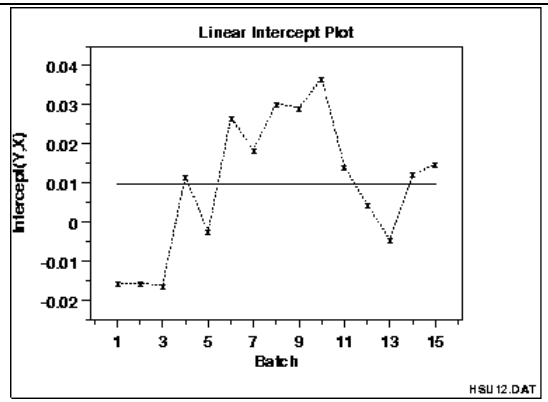
$$y = f(x_1, x_2, \ldots, x_k) + e$$



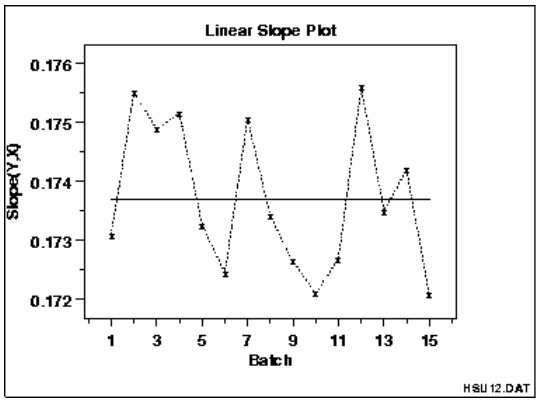


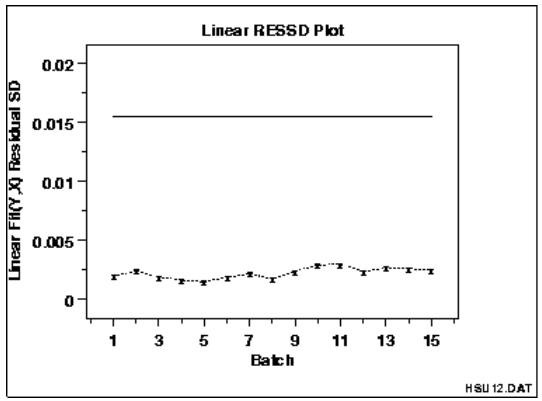
a :Scatter Plot b :6 – Plot





c:Linear Correlation Plot d:Linear Intercept Plot

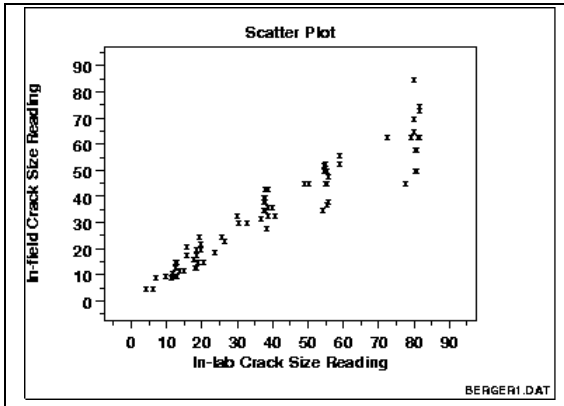




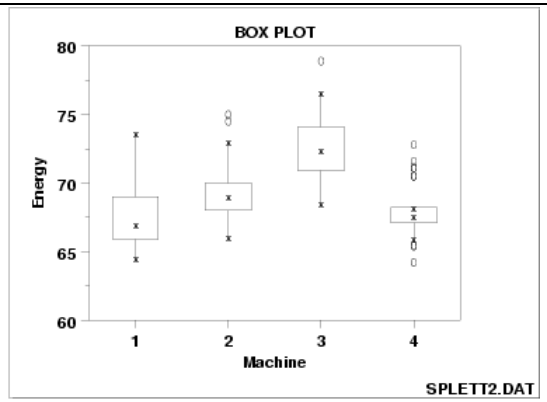
e:Linear Slope Plot f:Linear Residual Standard Deviation Plot

3.3 One Factor

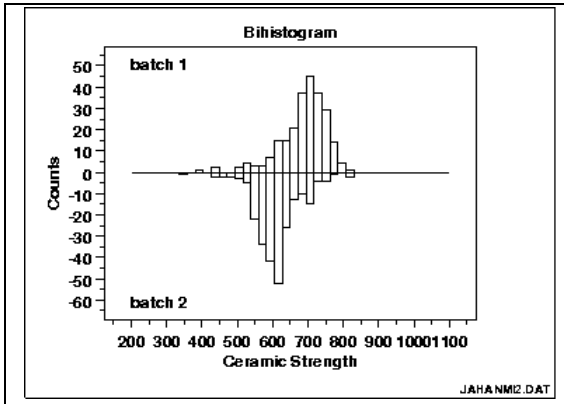
$$y = f(x) + e$$



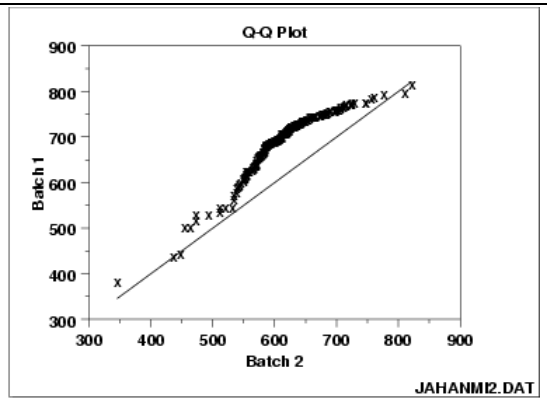
a:Scatter Plot



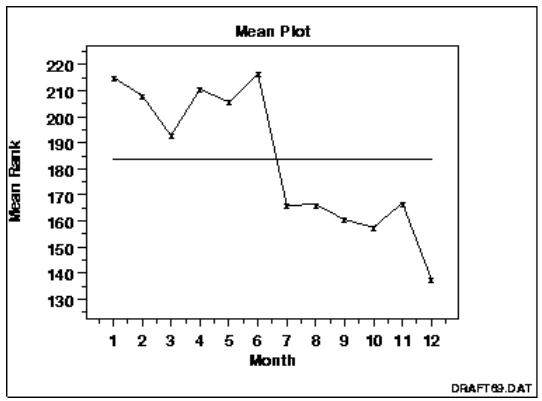
b:Box Plot



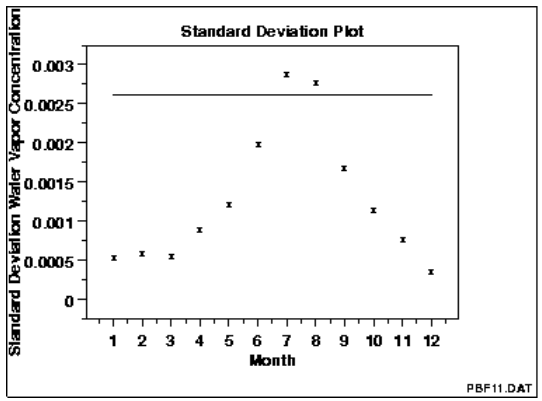
c:Bihistogram



d :Quantile-Quantile Plot



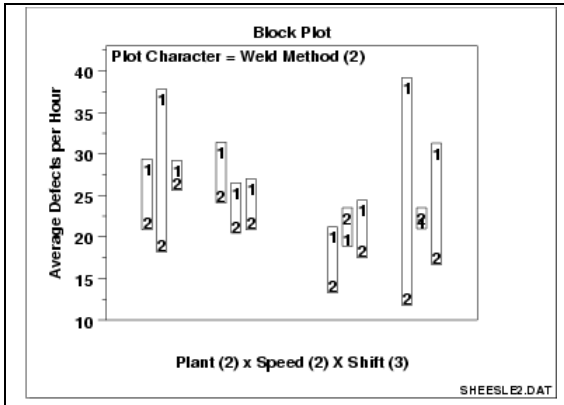
e:Mean Plot



f:Standard Deviation Plot

3.4 Multi-factor/Comparative

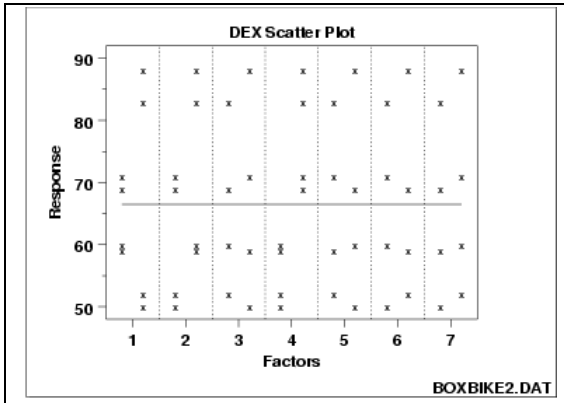
$$y = f(x_p, x_1, x_2, ..., x_k) + e$$



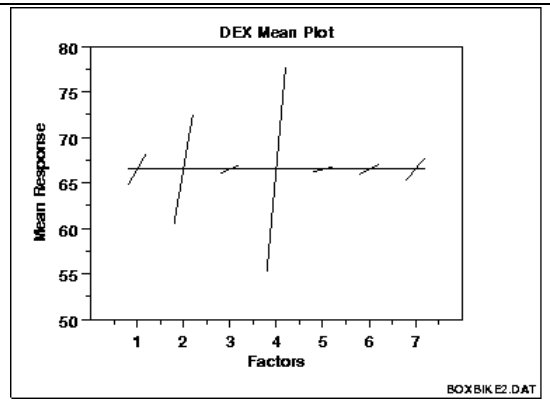
Block Plot

3.5 Multi-factor/Screening

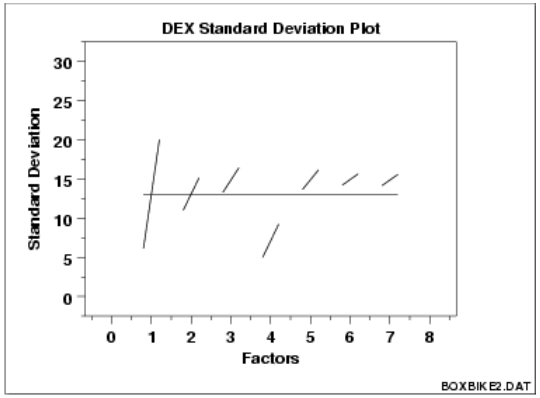
$$y = f(x_1, x_2, ..., x_k) + e$$



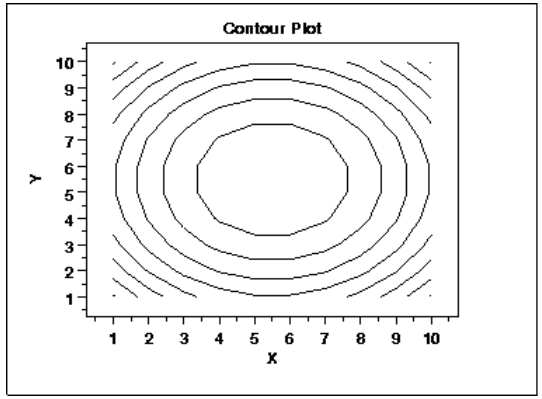
a: Dex Scatter Plot



b: Dex mean plot



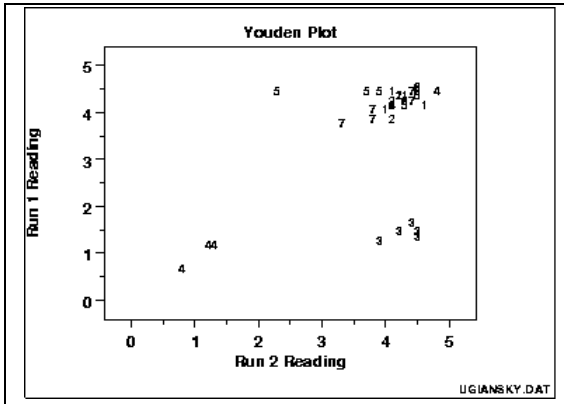
c:Dex standard Deviation Plot



d :Contour plot

3.6 Interlab

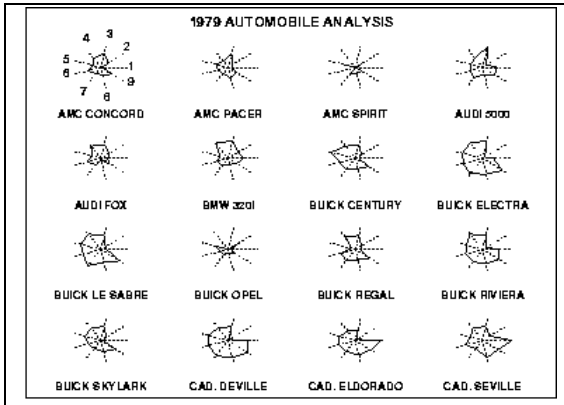
$$(y_1, y_2) = f(x) + e$$



Youden plot

3.7 Multivariate

$(y_1, y_2, \ldots, y_p)$



Star Plot