Applications of Mathematical Statistics in Management

C. Heuchenne

23 March 2012

In particular, three domains of applications will be considered:

- Statistical Process Control in Total Quality Management

- Flexible Modelling in Financial Risk Management

- Survival Analysis for General Duration Data (time to find a new job, insurance contract duration, …)

In each case, new statistical methodologies are proposed to solve given problems for which no or only partial solutions have been provided.

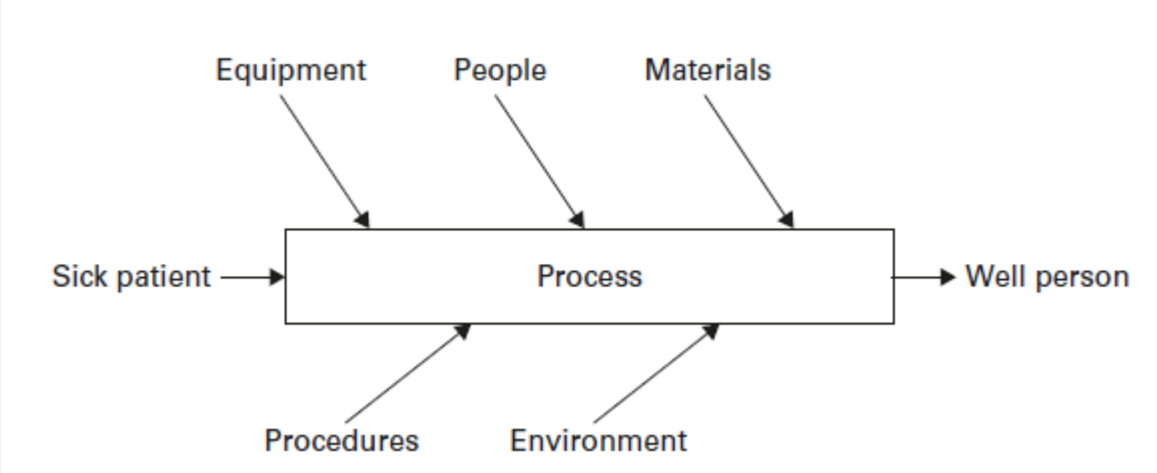
I) Statistical Process Control (SPC)

A) Context and basic ingredients

B) Problems met and proposed solutions

A) Context and basic ingredients

- Question: how to monitor a production or services process? (How to detect failures and their causes, to warn and repair the system sufficiently early…)
- Basic answer: use SPC
- Example of a general process:



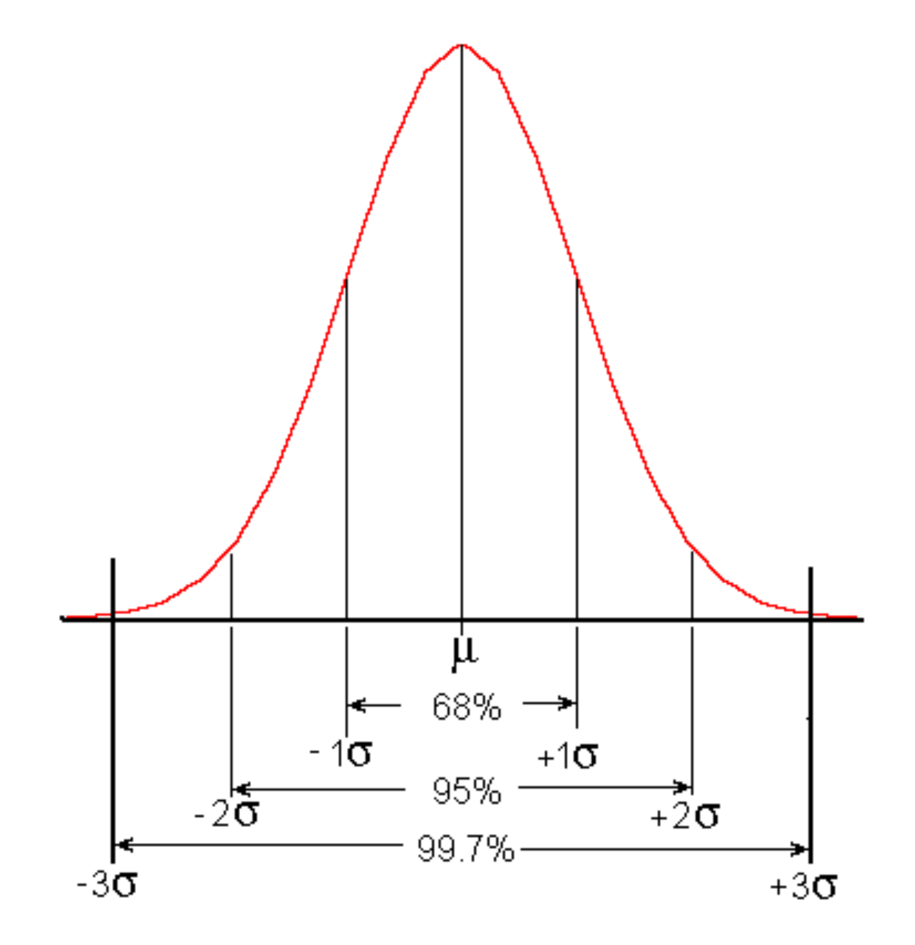
The very simple idea of SPC is to monitor a characteristic of interest of the process (for example the mean time to achieve a task or a measure of the quality of a product) by

1. considering its distribution,
2. comparing observed values of the characteristic (on the basis of successive samples) with the above distribution,
3. defining a decision rule that establishes if the process is « in control » or « out of control ».

Walter Shewhart

Developer of Control Charts in the late 1920’s

Natural distribution:



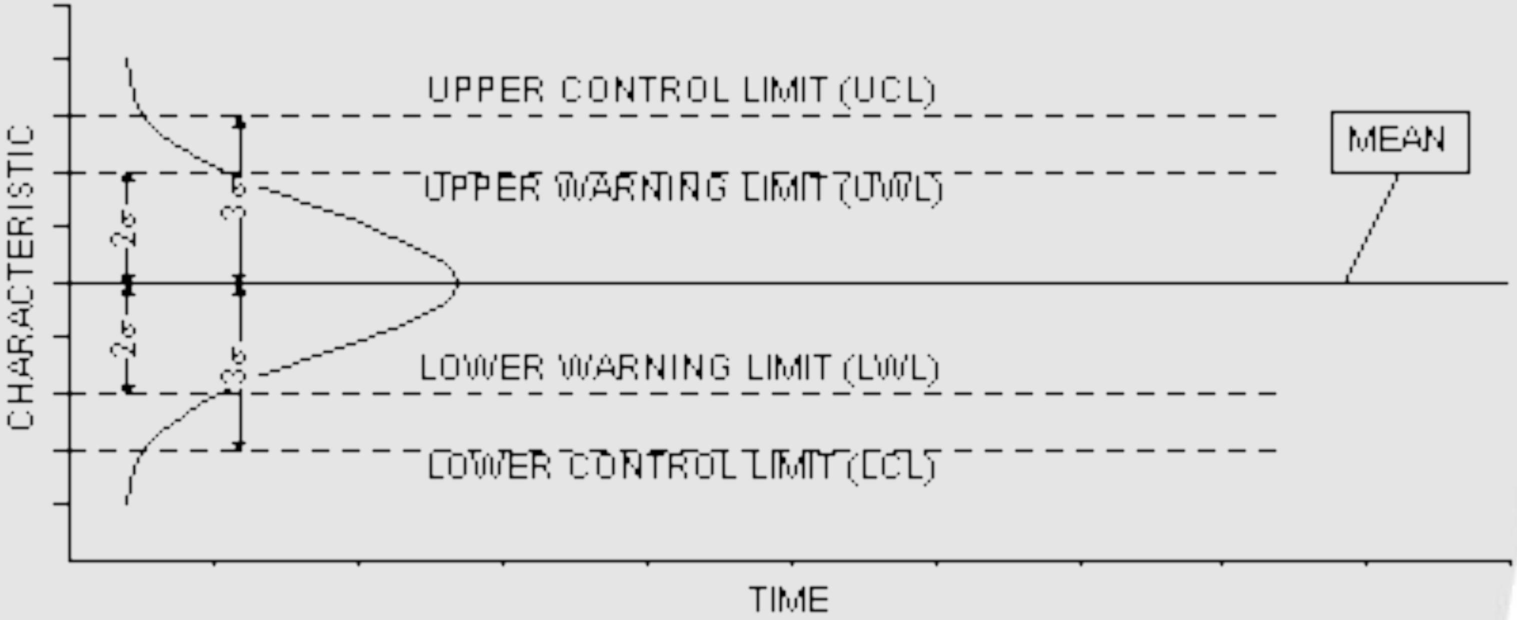
Normal distribution
defined by two parameters:
mean and standard deviation

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

Control charts provide a graphical mean to test hypotheses about the data being monitored.

μ+3σ

μ-3σ


UPPER CONTROL LIMIT (UCL)
MEAN
UPPER WARNING LIMIT (UWL)
CHARACTERISTIC
2σ
3σ
2σ
3σ
LOWER WARNING LIMIT (LWL)
LOWER CONTROL LIMIT (LCL)
TIME

http://lorien.ncl.ac.uk/ming/spc/spc8.htm#interpretation

Design control chart

- sample size
  - larger sample size leads to faster detection
- setting control limits
- time between samples
  - sample more frequently with few items or
  - sample less frequently with more items?
- choice of measurement
- …

Heuristic Designs of control Charts

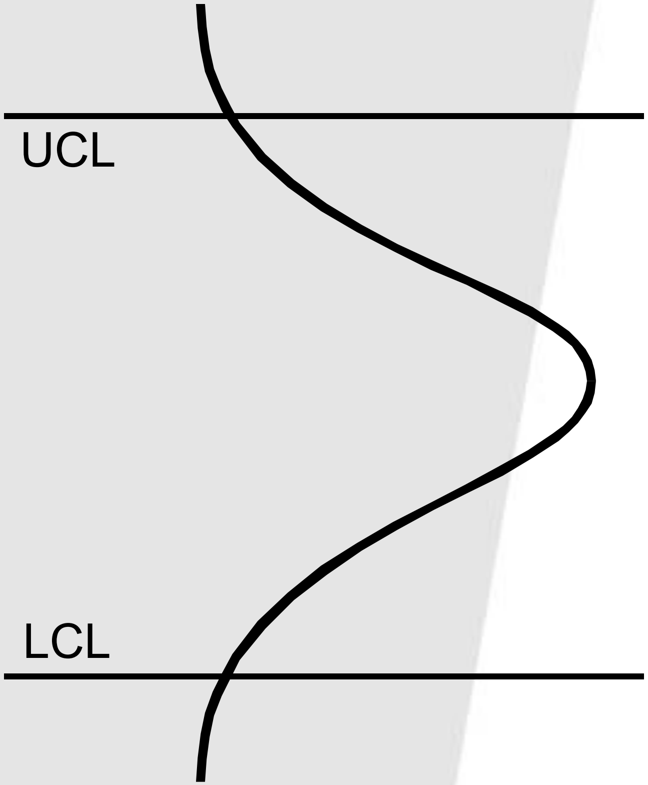
$$P(LCL < \bar{X} < UCL) = 1 - \alpha$$

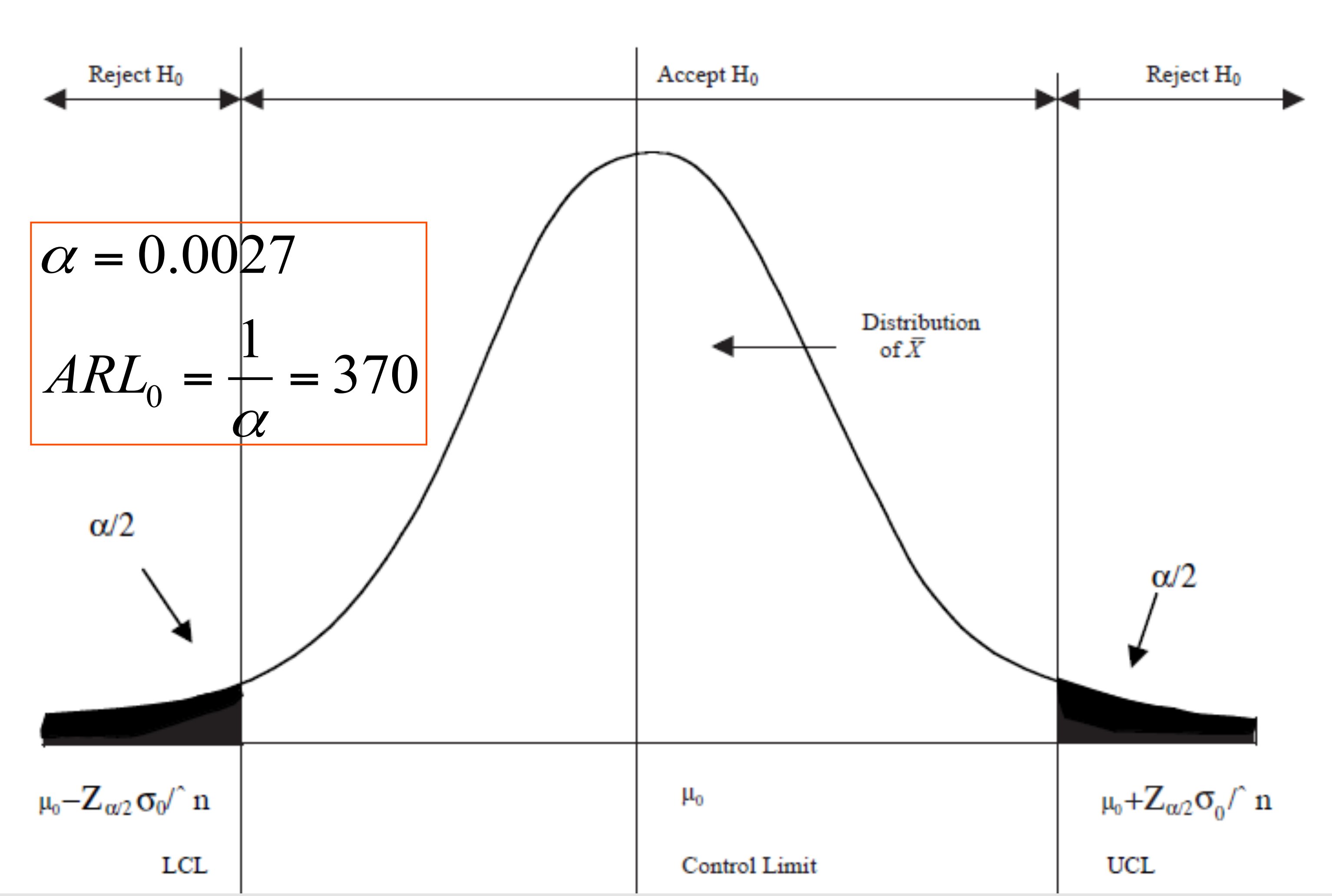
$$LCL = \mu - z_{\alpha/2}\sigma / \sqrt{n}; \quad UCL = \mu + z_{\alpha/2}\sigma / \sqrt{n}$$

$$\alpha = 0.0027$$

$$LCL = \mu - 3\sigma / \sqrt{n}; \quad UCL = \mu + 3\sigma / \sqrt{n}$$

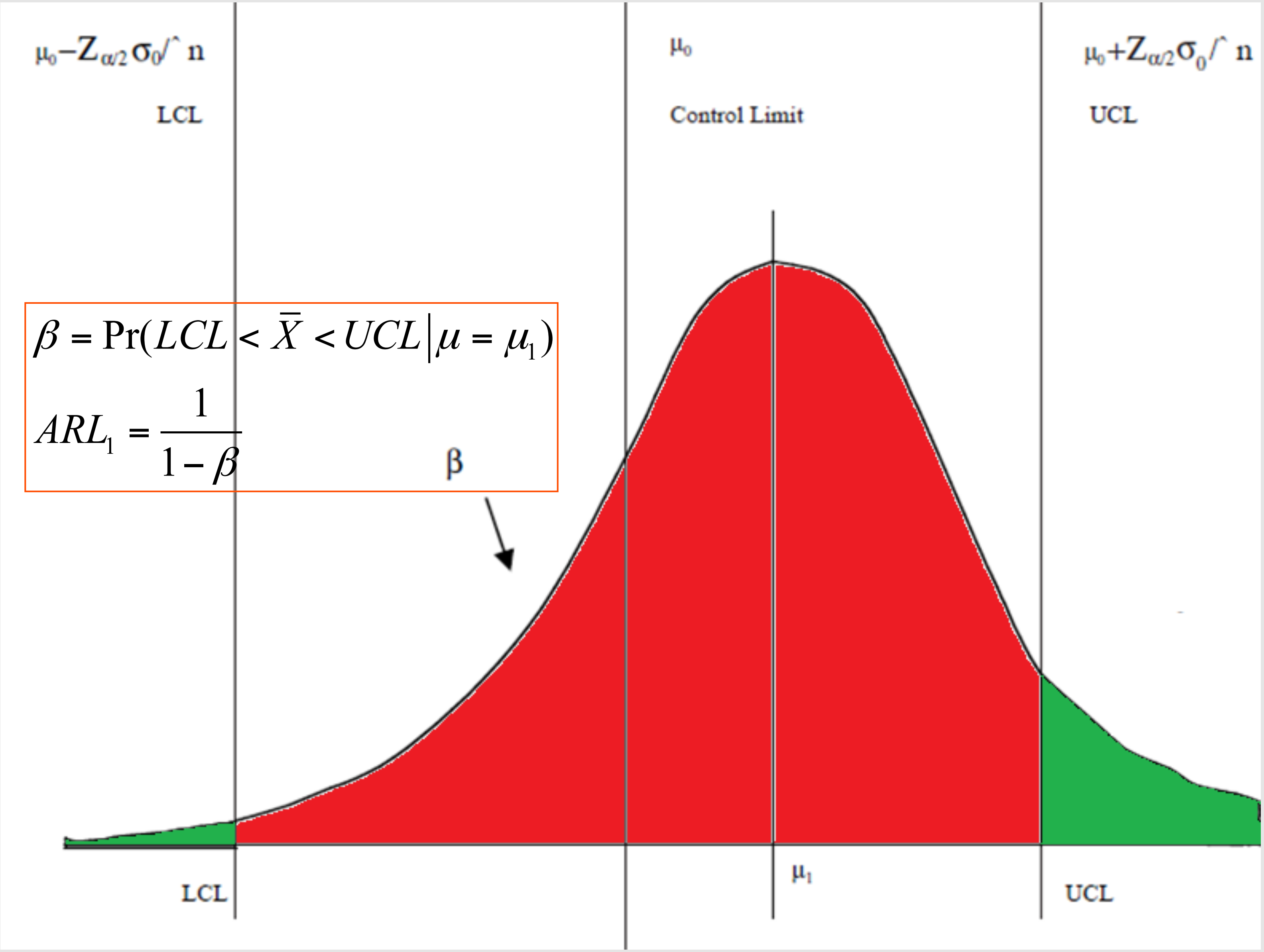
$X \sim N(\mu, \sigma)$




Reject H_0
Accept H_0
Reject H_0
Distribution of $\bar{X}$
$\alpha/2$
$\alpha/2$
$\mu_0 - Z_{\alpha/2}\sigma_0/\hat{}\ n$
μ_0
$\mu_0 + Z_{\alpha/2}\sigma_0/\hat{}\ n$
LCL
Control Limit
UCL

$$\alpha = 0.0027$$

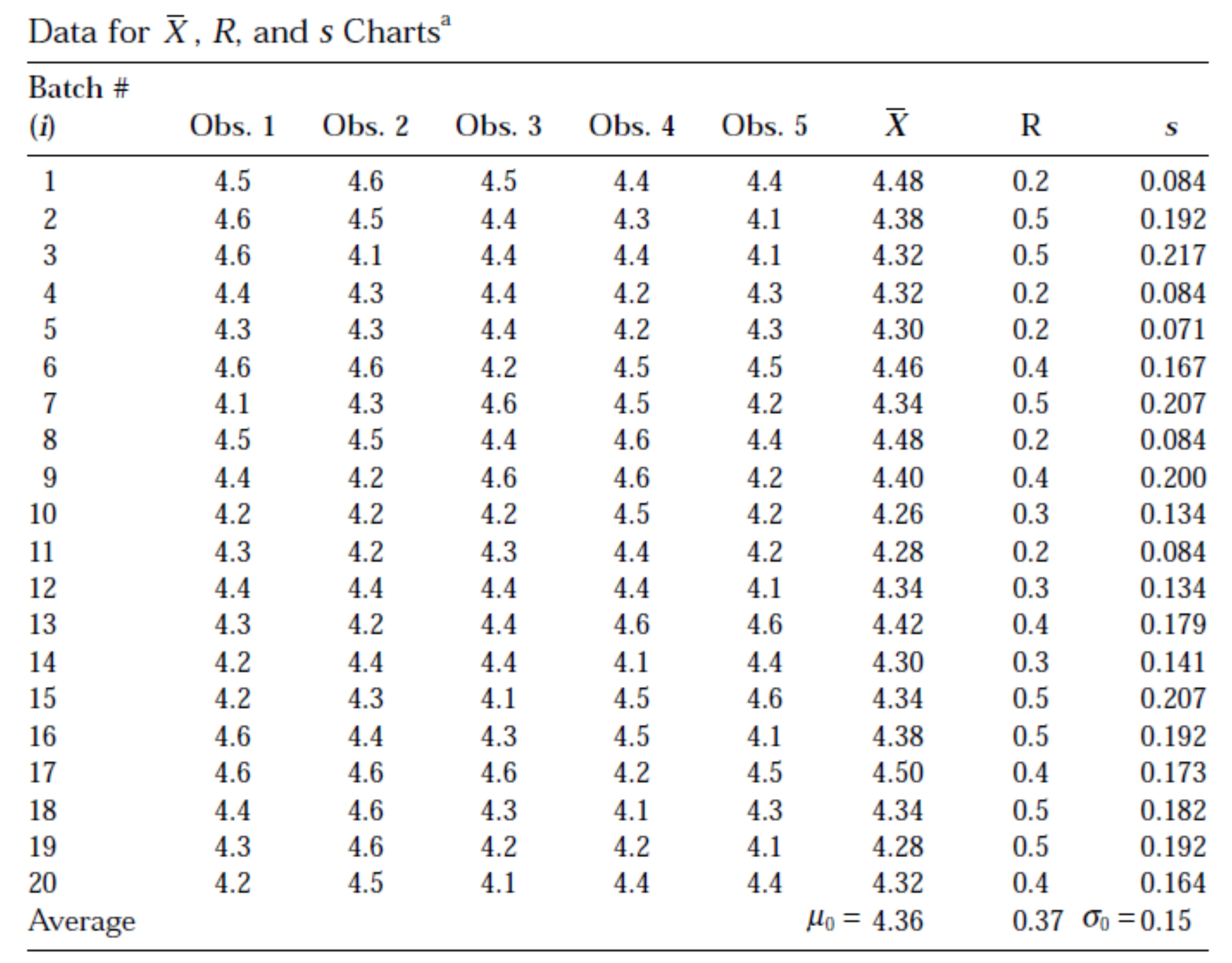
$$ARL_0 = \frac{1}{\alpha} = 370$$



$$\beta = \Pr(LCL < \bar{X} < UCL \,|\, \mu = \mu_1)$$

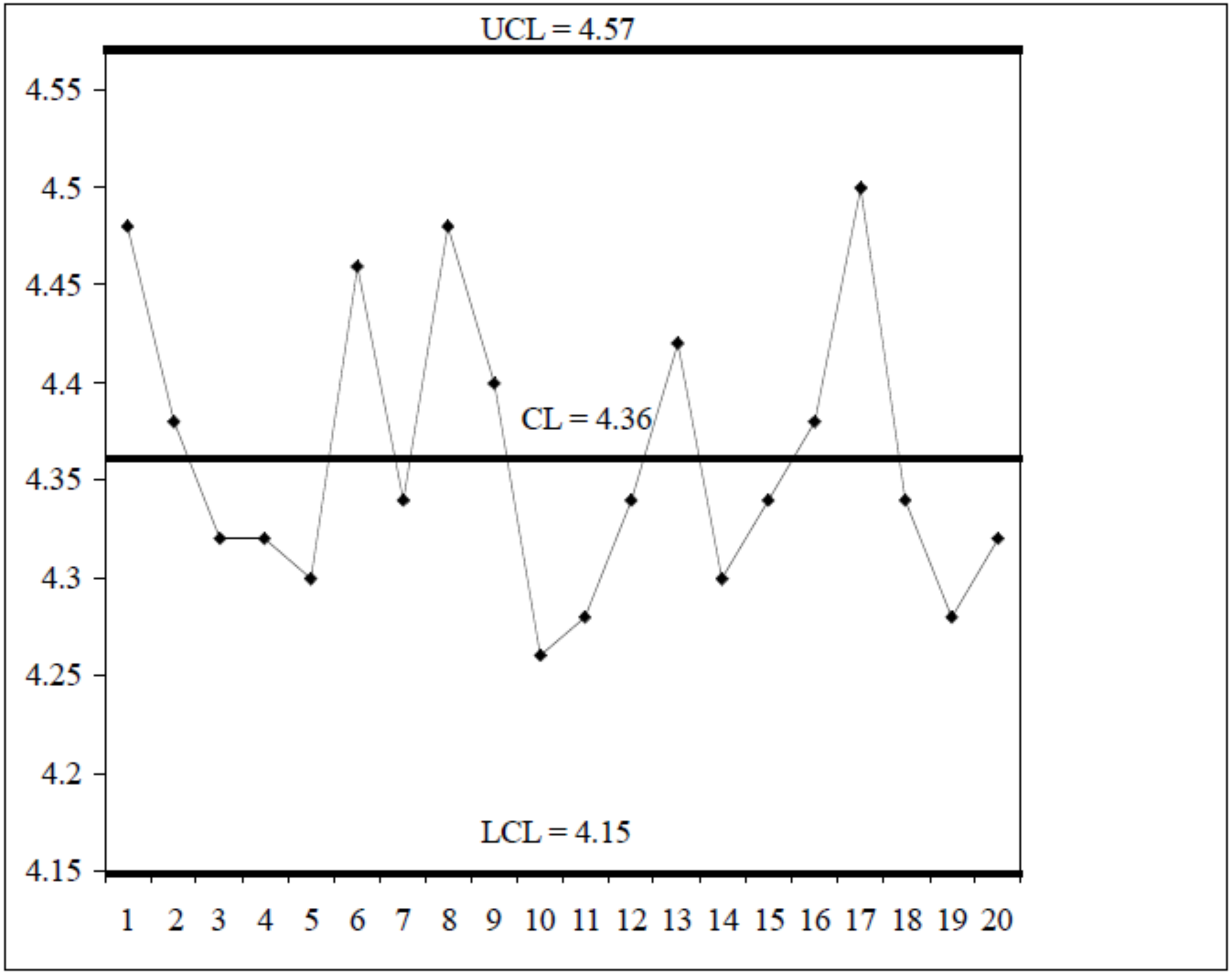
$$ARL_1 = \frac{1}{1-\beta}$$

Example



Data for $\bar{X}$, R, and s Charts[a]

Batch # (i)	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5	$\bar{X}$	R	s
1	4.5	4.6	4.5	4.4	4.4	4.48	0.2	0.084
2	4.6	4.5	4.4	4.3	4.1	4.38	0.5	0.192
3	4.6	4.1	4.4	4.4	4.1	4.32	0.5	0.217
4	4.4	4.3	4.4	4.2	4.3	4.32	0.2	0.084
5	4.3	4.3	4.4	4.2	4.3	4.30	0.2	0.071
6	4.6	4.6	4.2	4.5	4.5	4.46	0.4	0.167
7	4.1	4.3	4.6	4.5	4.2	4.34	0.5	0.207
8	4.5	4.5	4.4	4.6	4.4	4.48	0.2	0.084
9	4.4	4.2	4.6	4.6	4.2	4.40	0.4	0.200
10	4.2	4.2	4.2	4.5	4.2	4.26	0.3	0.134
11	4.3	4.2	4.3	4.4	4.2	4.28	0.2	0.084
12	4.4	4.4	4.4	4.4	4.1	4.34	0.3	0.134
13	4.3	4.2	4.4	4.6	4.6	4.42	0.4	0.179
14	4.2	4.4	4.4	4.1	4.4	4.30	0.3	0.141
15	4.2	4.3	4.1	4.5	4.6	4.34	0.5	0.207
16	4.6	4.4	4.3	4.5	4.1	4.38	0.5	0.192
17	4.6	4.6	4.6	4.2	4.5	4.50	0.4	0.173
18	4.4	4.6	4.3	4.1	4.3	4.34	0.5	0.182
19	4.3	4.6	4.2	4.2	4.1	4.28	0.5	0.192
20	4.2	4.5	4.1	4.4	4.4	4.32	0.4	0.164
Average						μ_0 = 4.36	0.37	σ_0 = 0.15

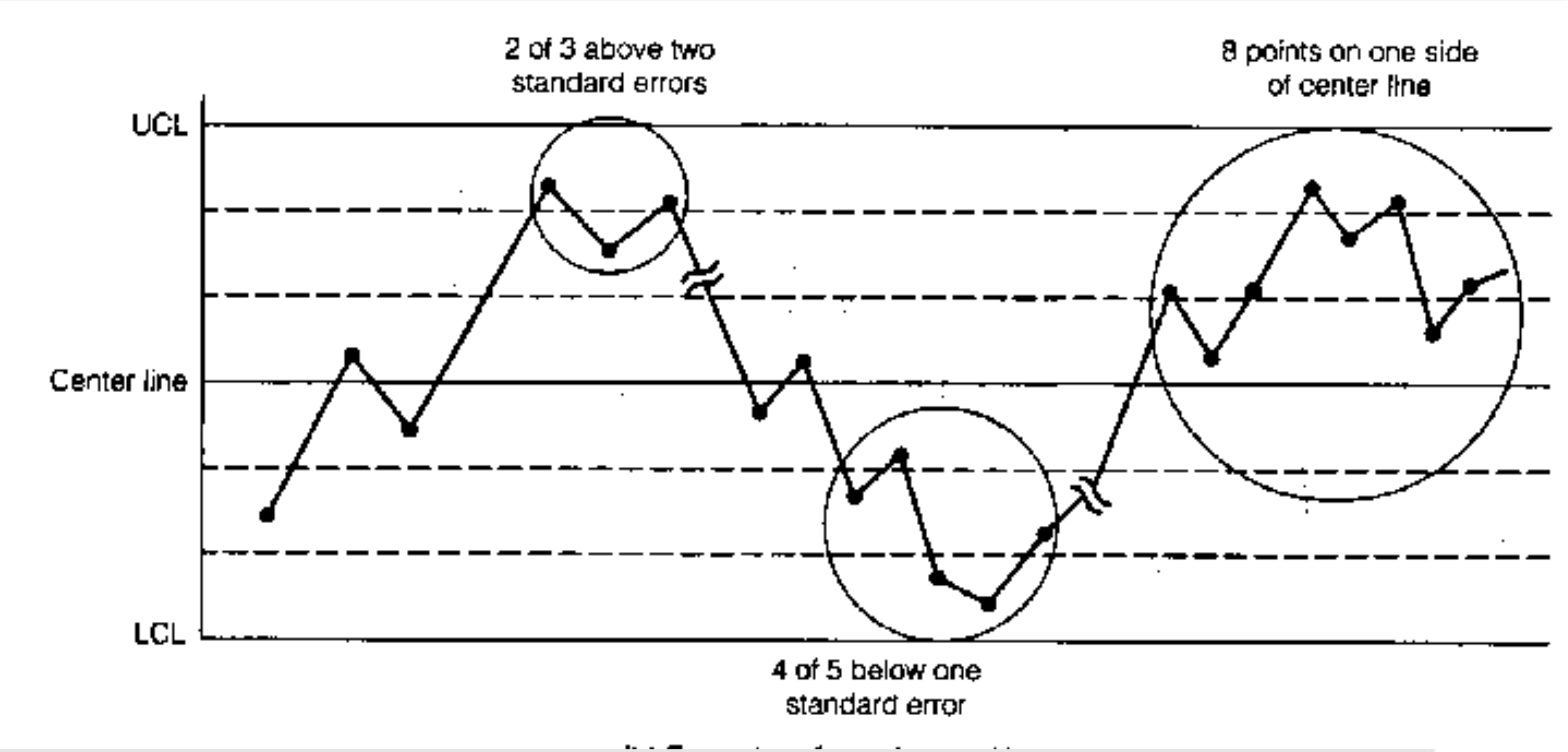

UCL = 4.57
CL = 4.36
LCL = 4.15
4.55
4.5
4.45
4.4
4.35
4.3
4.25
4.2
4.15
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

How do you know a process is “out of control”?

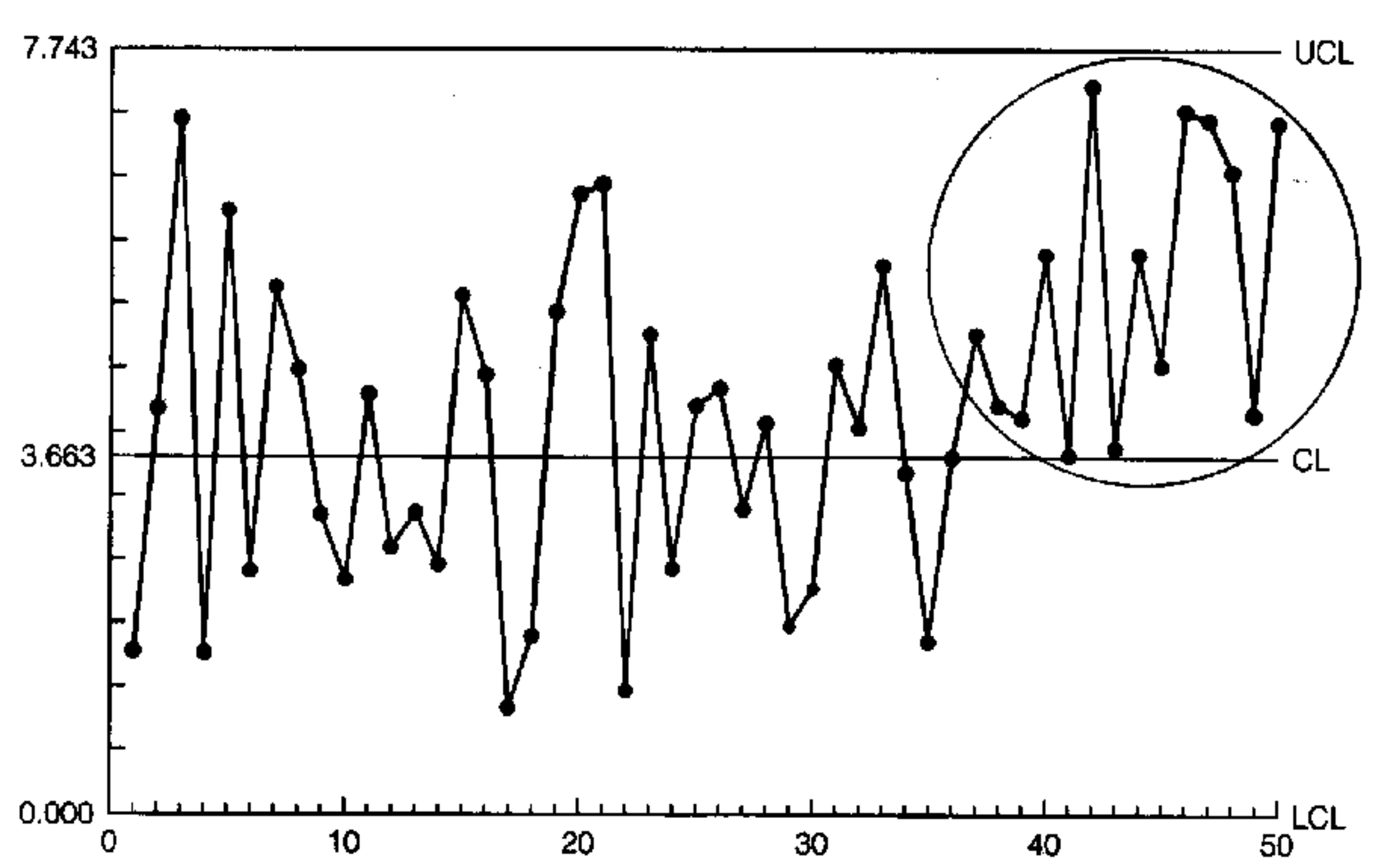
“Out of control” patterns:

- points outside of control limits (±3σ)
- 8 consecutive points on one side of center line
- 2 of 3 consecutive points outside the 2σ limits
- 4 of 5 consecutive points outside the 1σ limits
- 7 consecutive points trending up or down
- sudden shift in process average
- cycles
- trends
- hugging the center line
- hugging the control limits
- instability

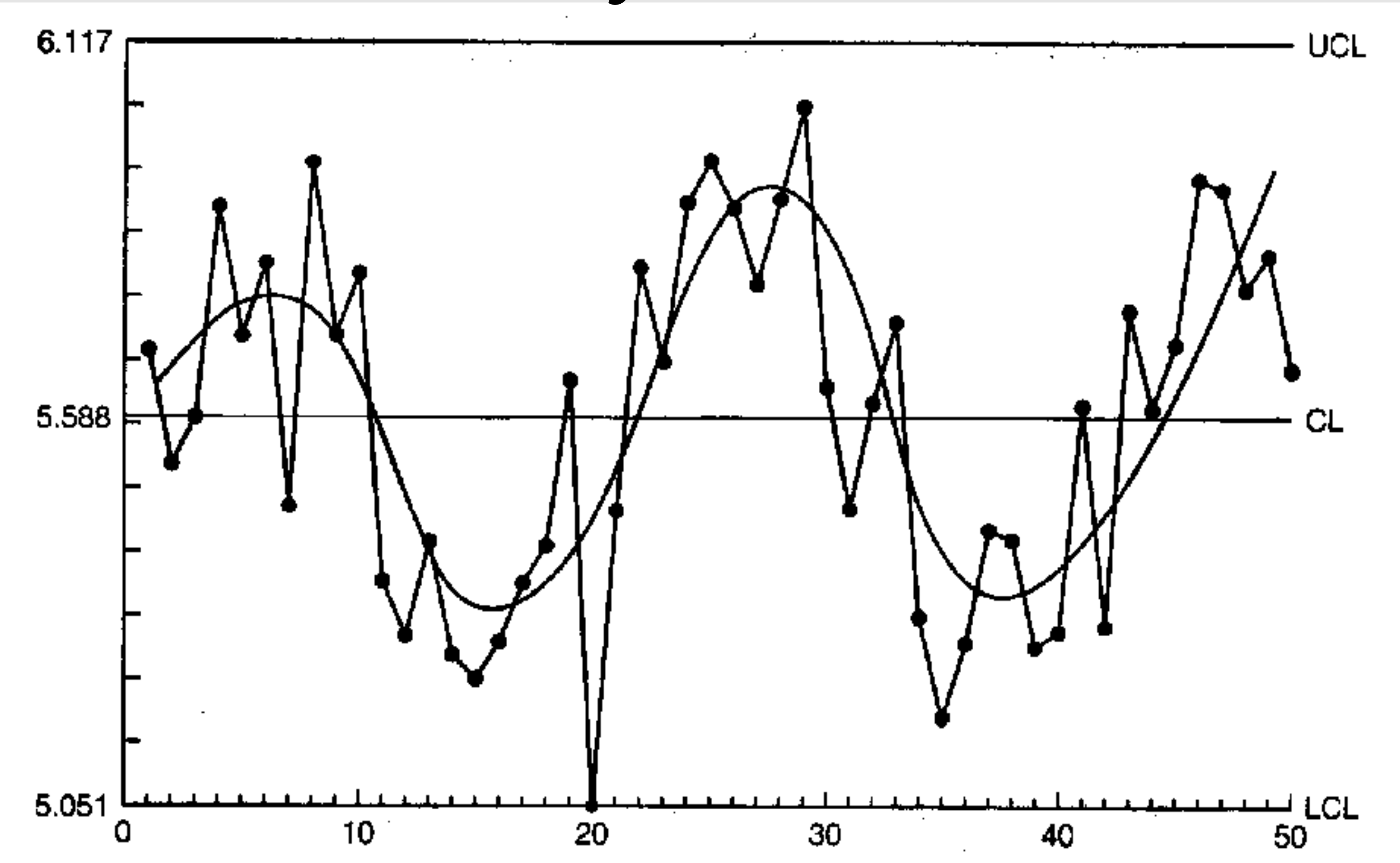
Identifying Potential Shifts



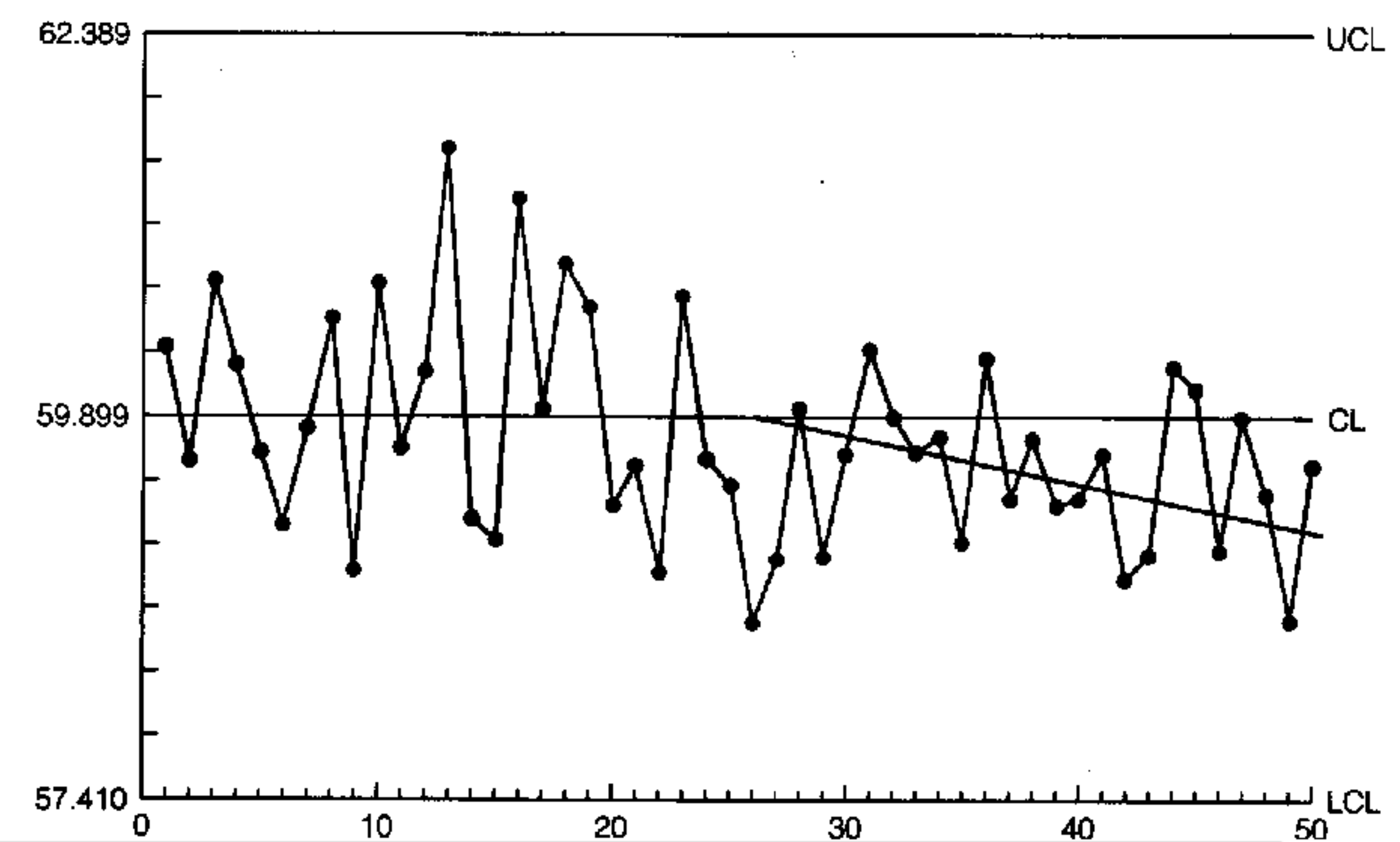
Shift in Process Average

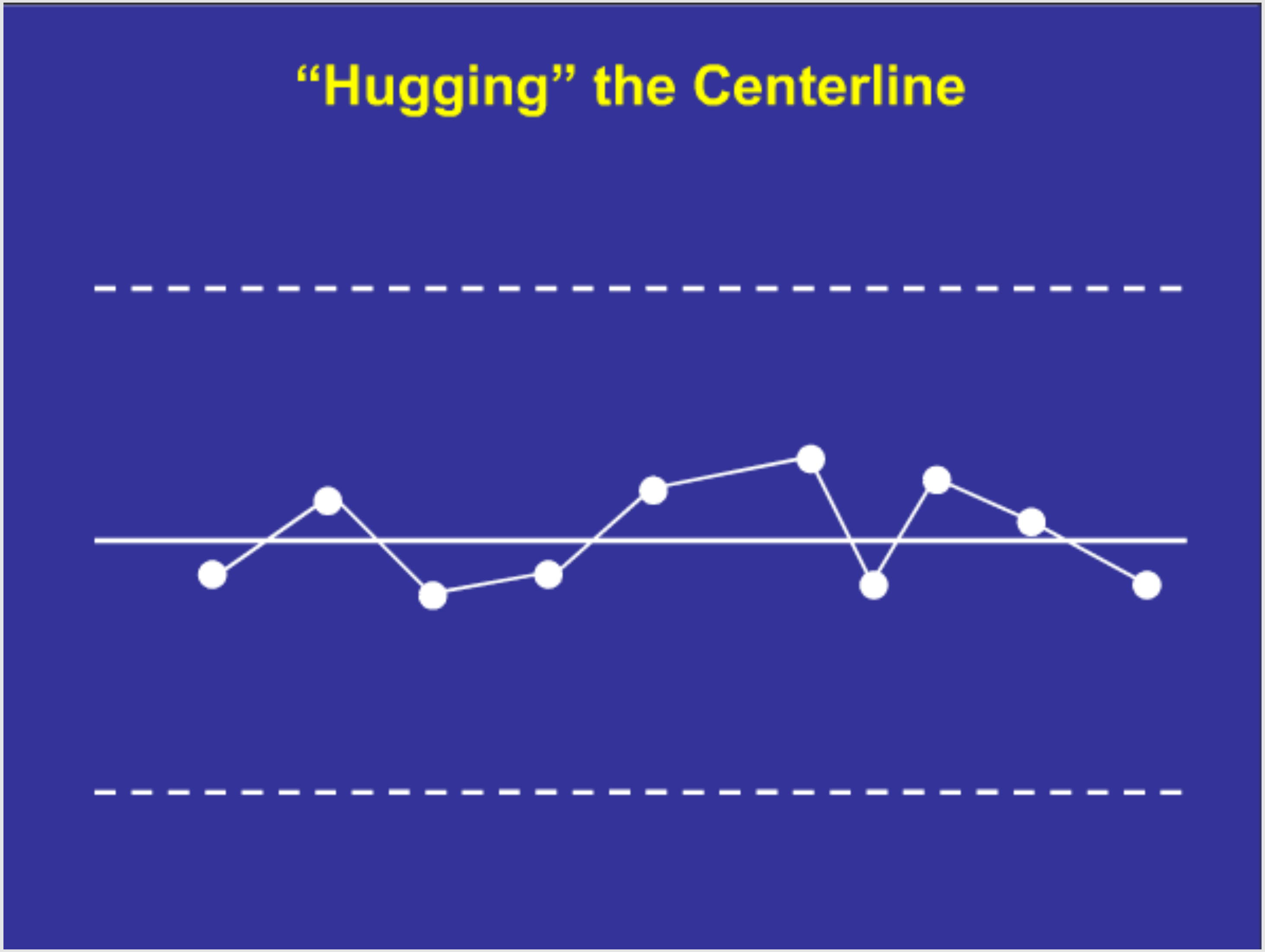


Cycles

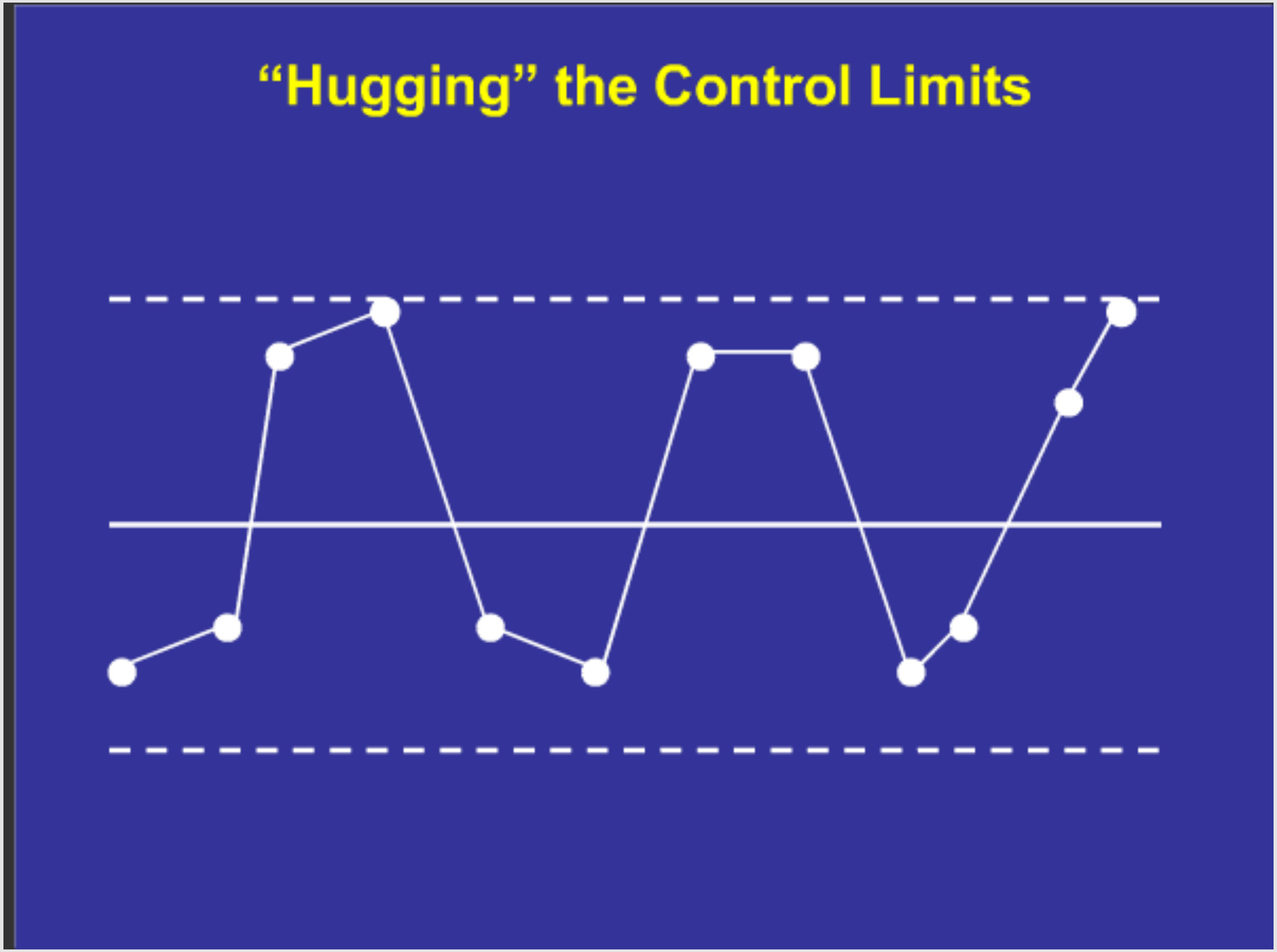


Trend



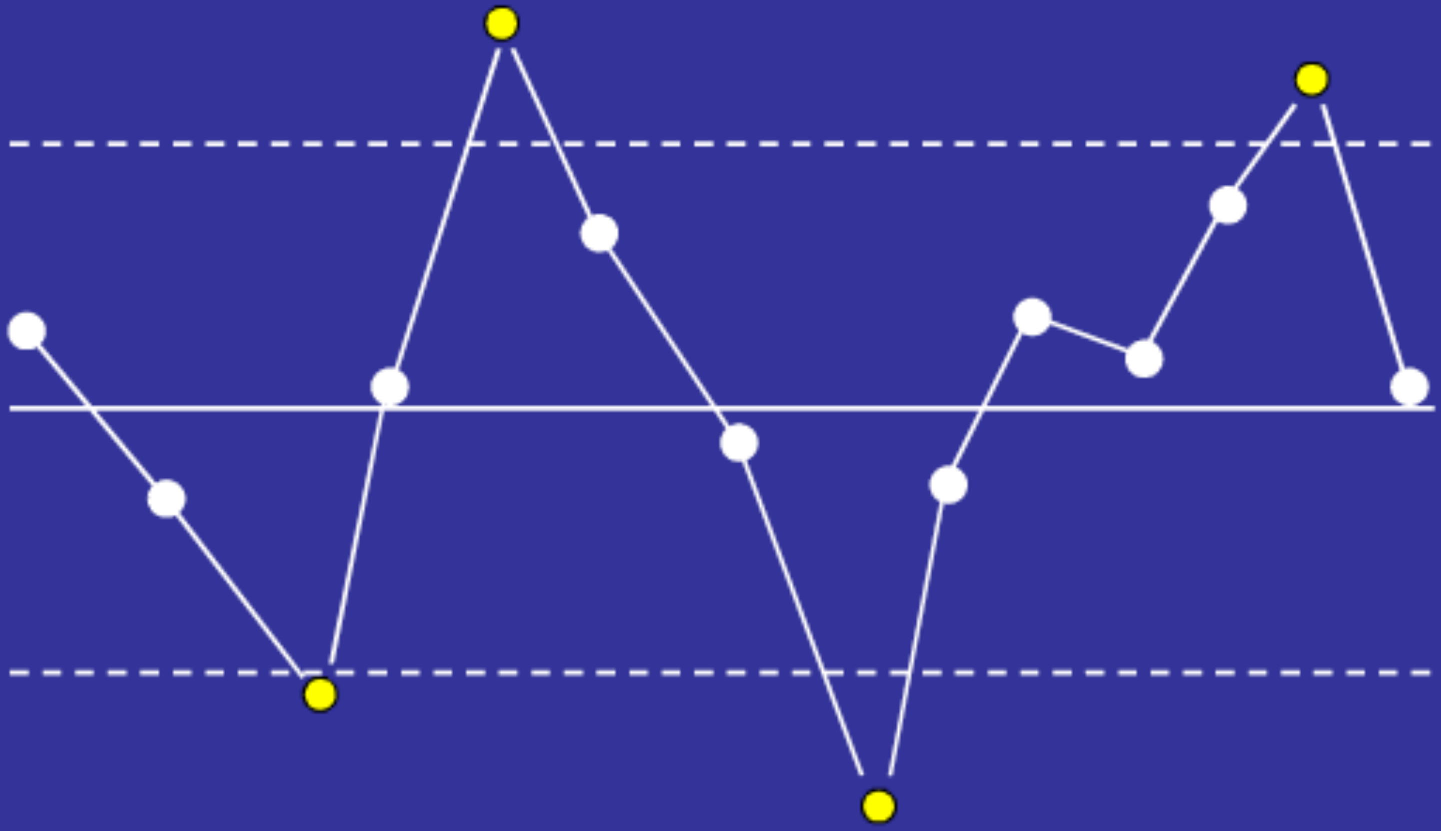


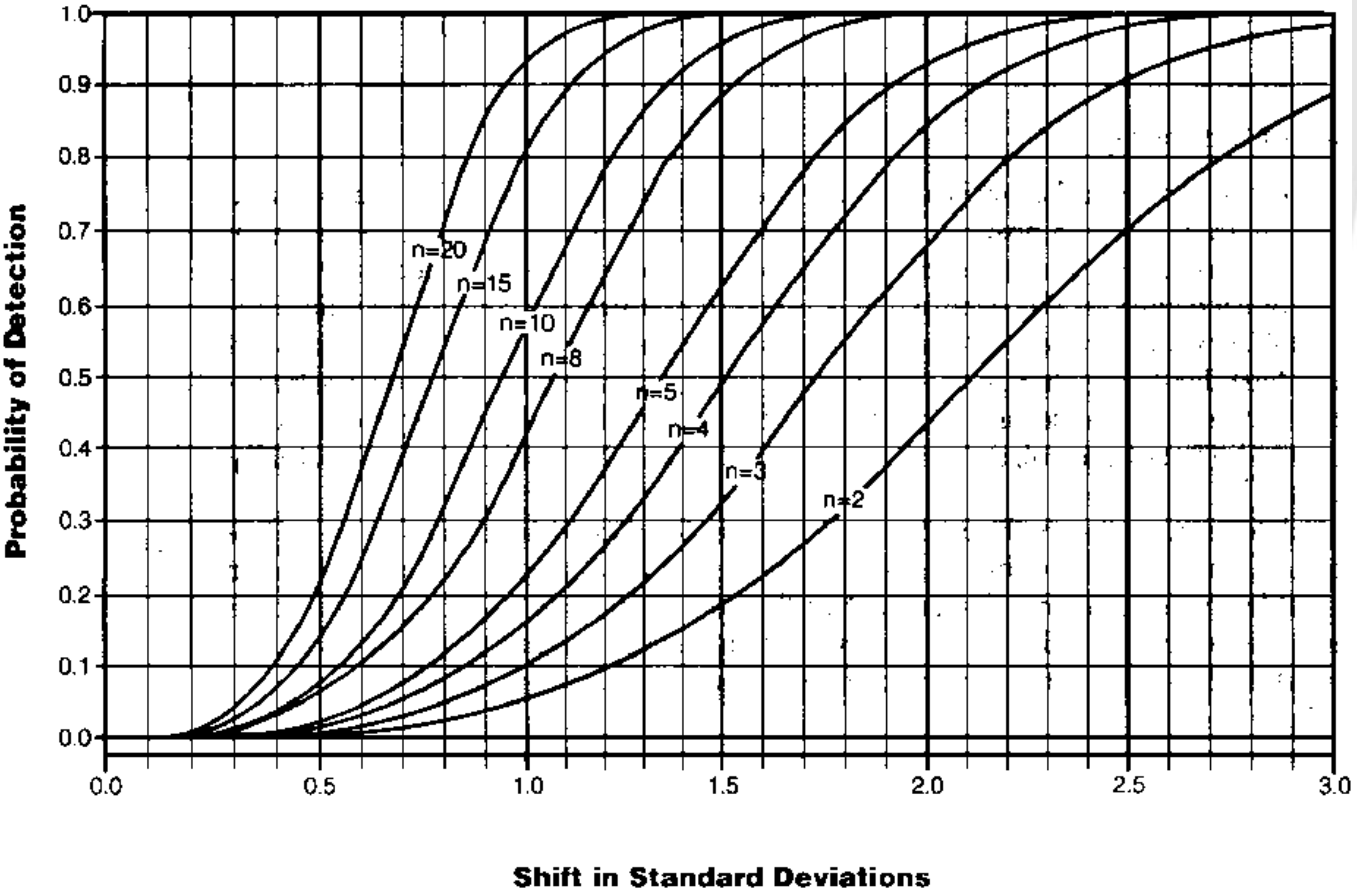
“Hugging” the Centerline



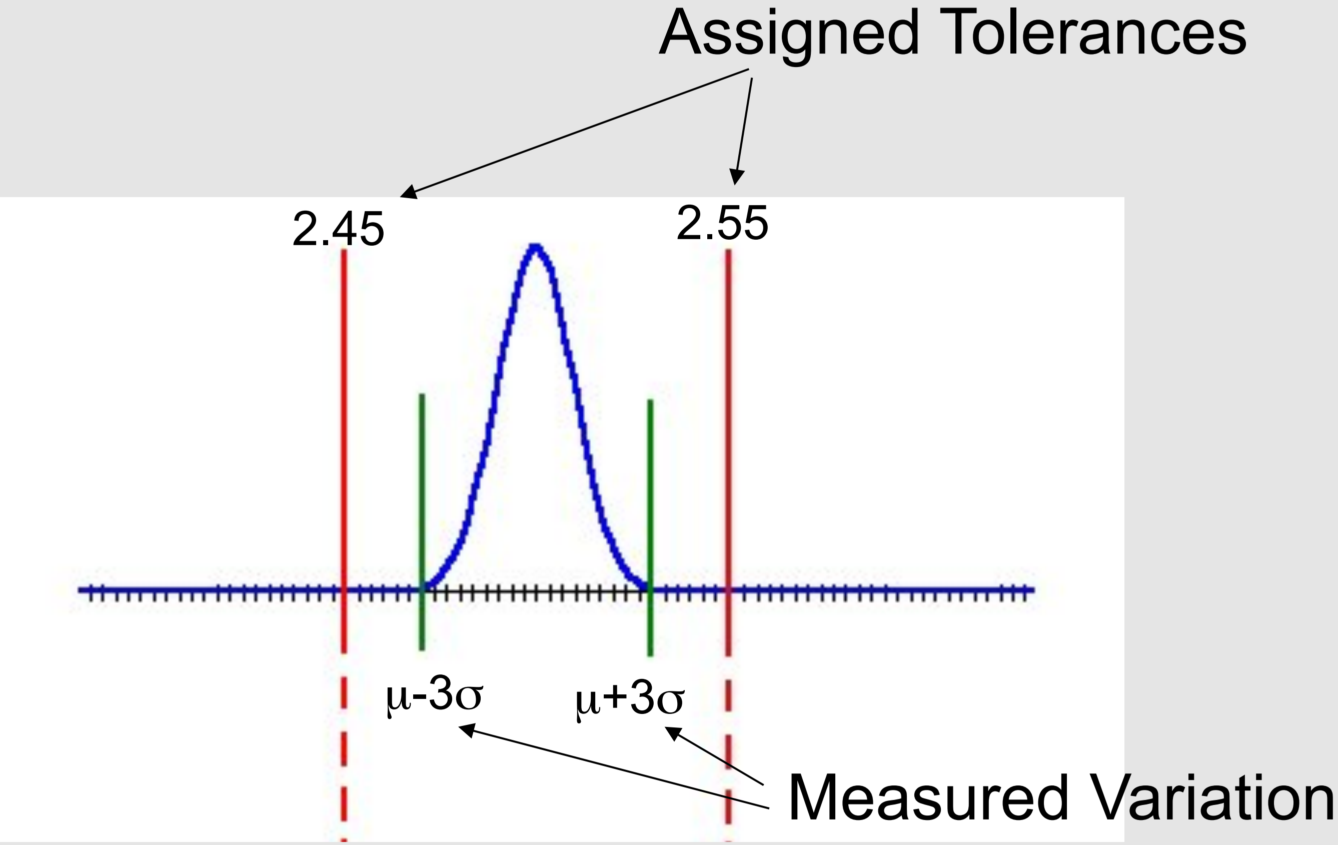
“Hugging” the Control Limits

Instability




Probability of Detection
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
n=20
n=15
n=10
n=8
n=5
n=4
n=3
n=2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Shift in Standard Deviations

How does the control chart relate to the tolerances?



Assigned Tolerances

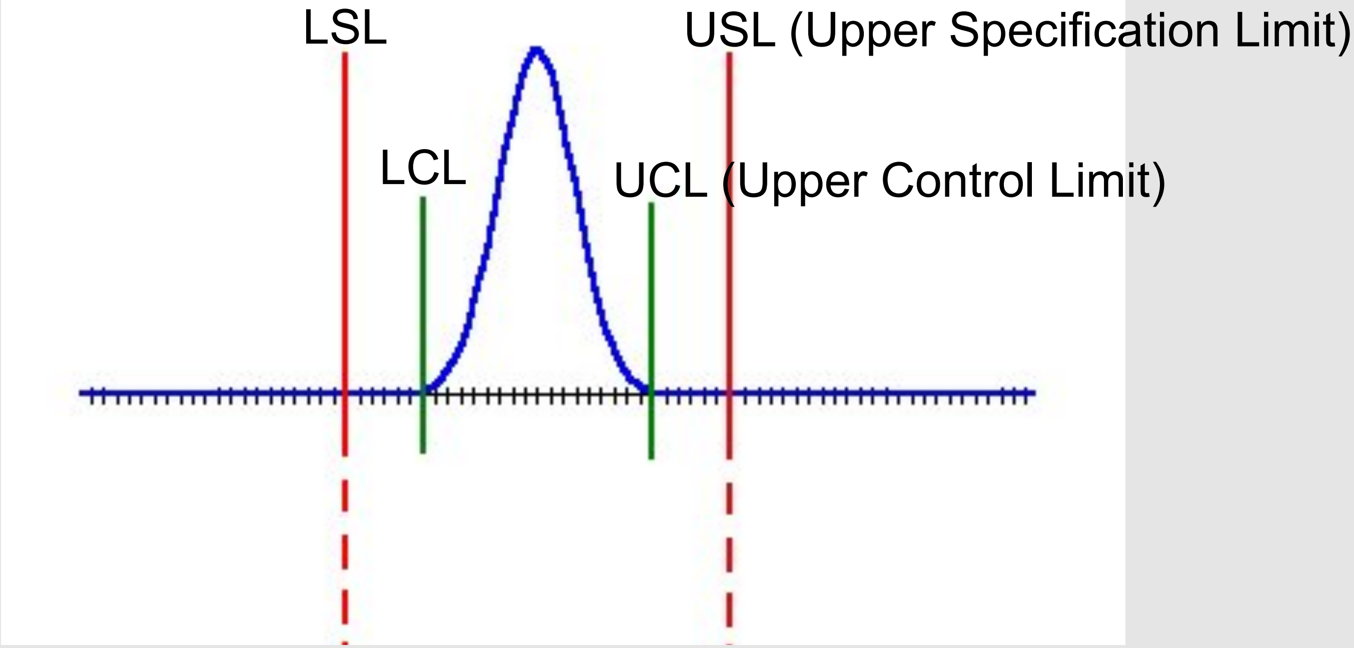
Measured Variation

Process Capability

- Comparing the control chart information with the tolerance specification tells you about the process capability.

The capability index is defined as:

Cp = (allowable range)/6s = (USL - LSL)/6s



LSL

USL (Upper Specification Limit)

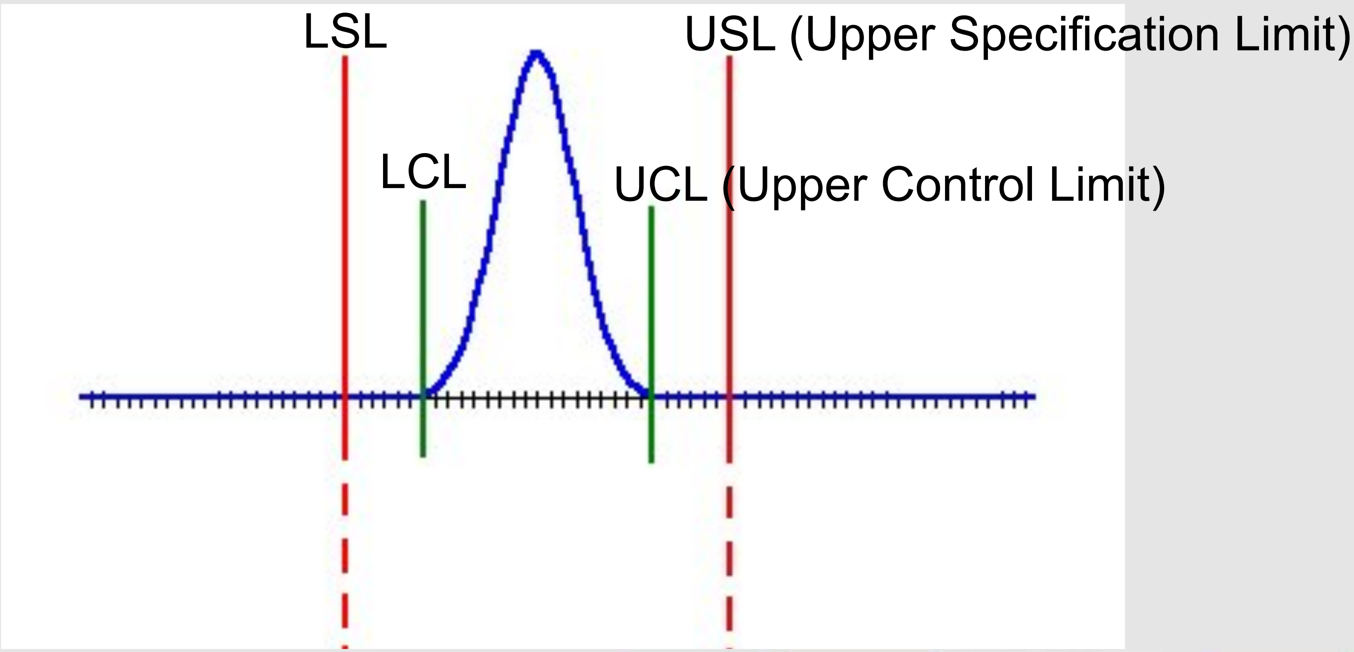
LCL

UCL (Upper Control Limit)

http://lorien.ncl.ac.uk/ming/spc/spc9.htm

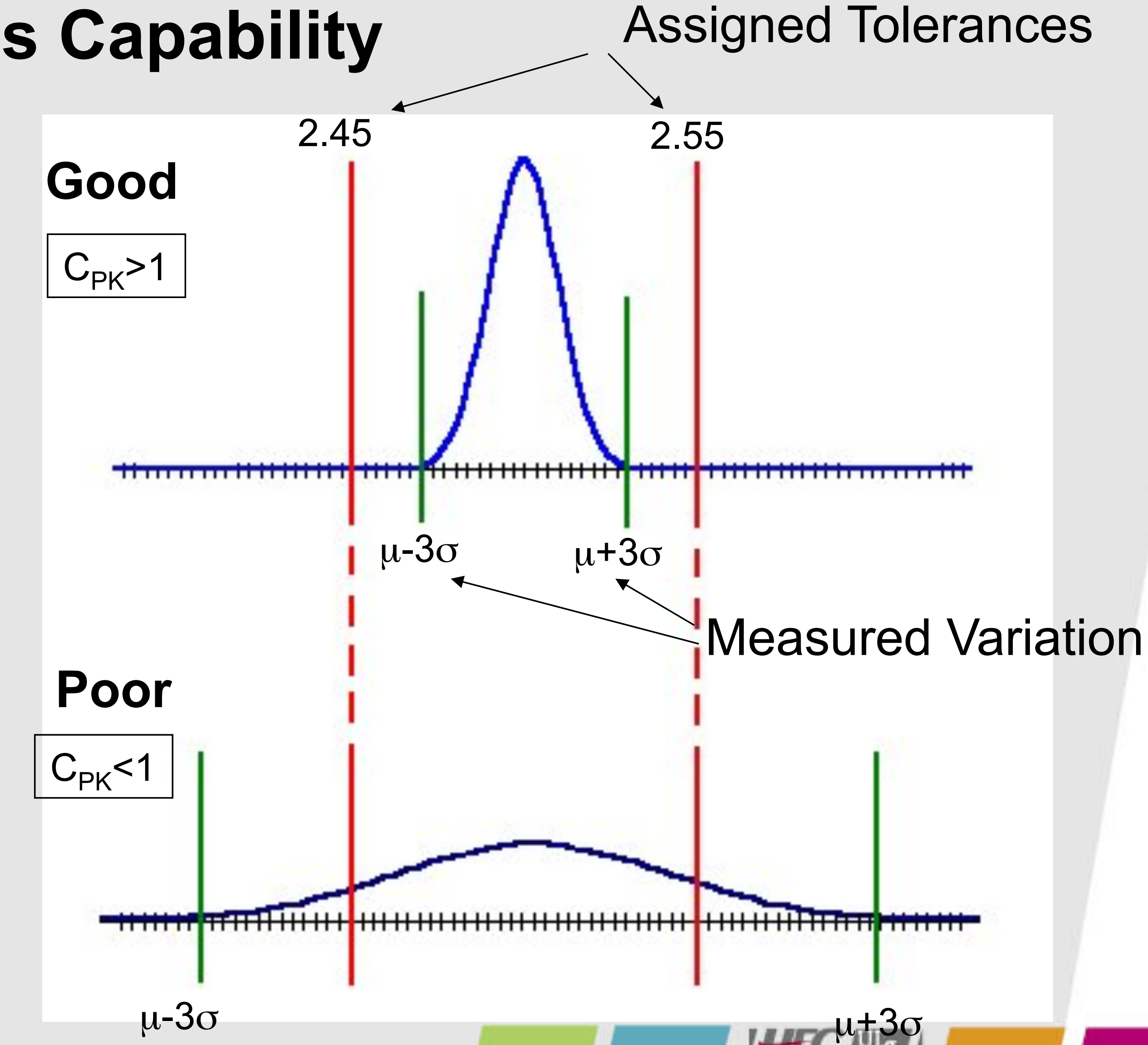
The process performance index takes into account the mean (m) and is defined as:

$$Cpk = \min[\ (USL - m)/3s,\ (m - LSL)/3s\ ]$$

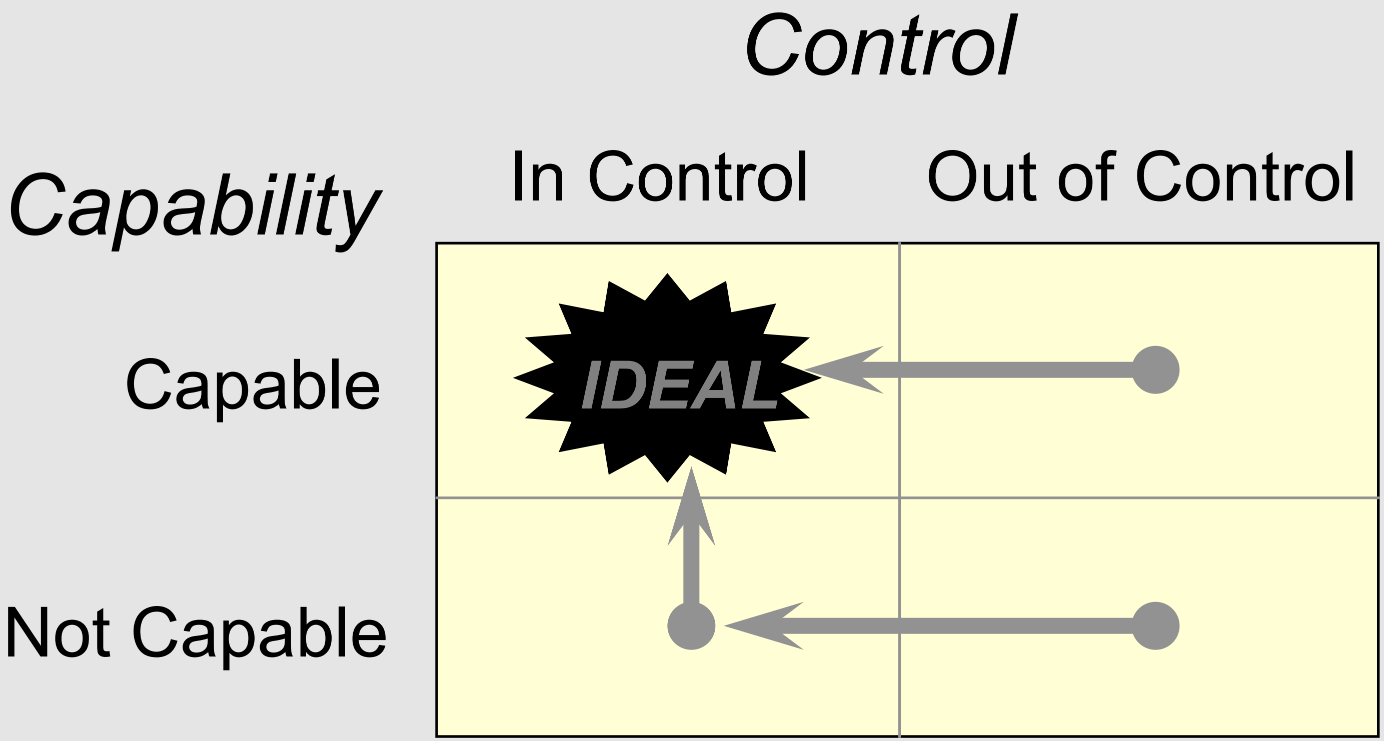


http://lorien.ncl.ac.uk/ming/spc/spc9.htm

Process Capability



Capability Versus Control



Control

In Control

Out of Control

Capability

Capable

Not Capable

IDEAL

Types of control charts

- Variables control charts
  - continuous data are measured. For example: time, weight, distance or temperature with continuous distribution
- Attributes control charts
  - Attribute data: presence or absence, success or failure, accept or reject, correct or false, following discrete distribution
  - Example: the number of errors (a nonnegative integer) of a report has a discrete distribution.

- **Variables control charts**
  - X-bar and R chart
  - X-bar and s chart
  - CUSUM (cumulative sum chart)
  - EWMA (exponentially weighted moving average chart)
  - multivariate chart

- **Attributes control charts**
  - p chart (proportion chart)
  - c chart (count chart)
  - u chart

B) Some problems met and proposed solutions

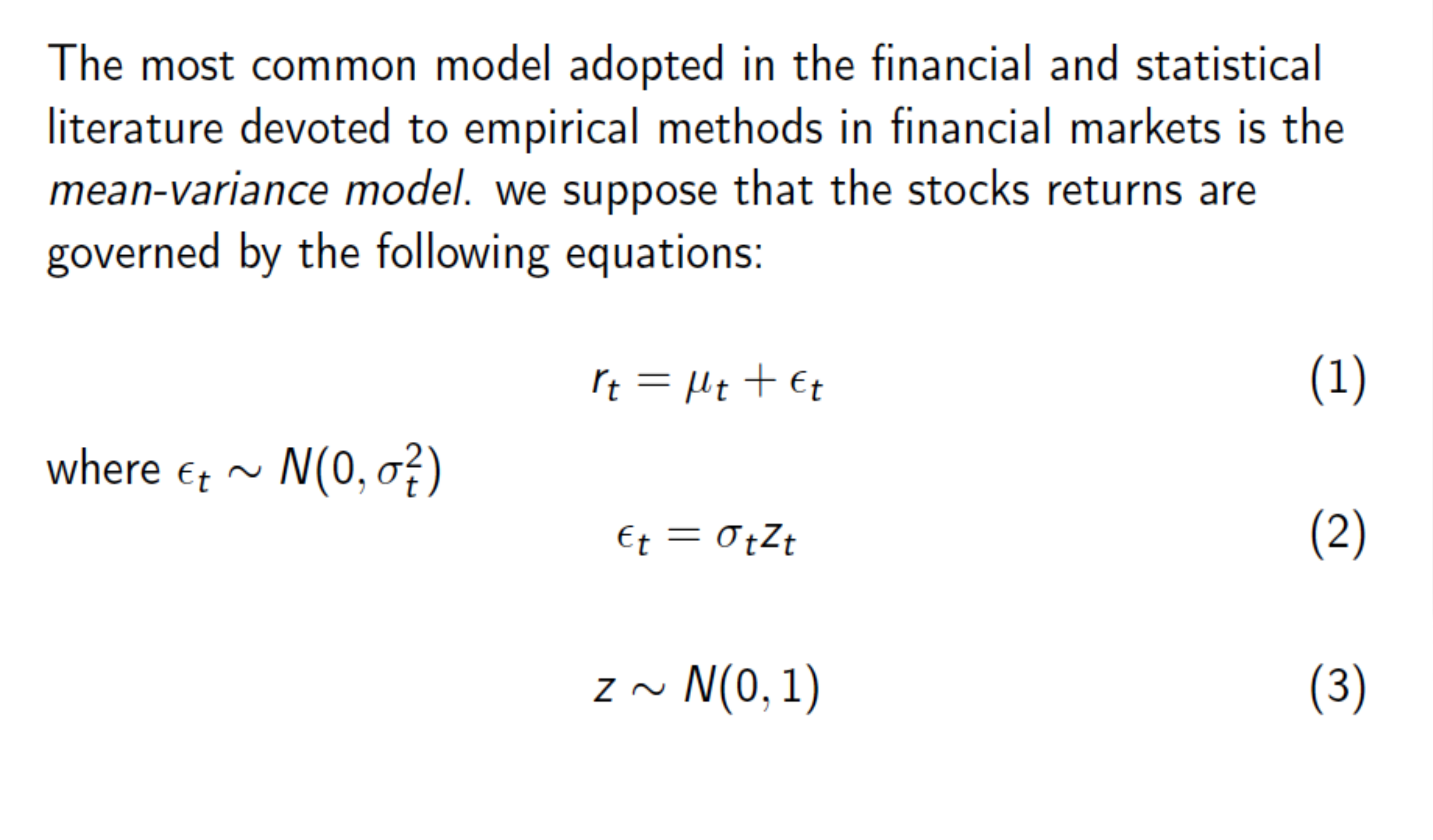
- Delivery chains: supply chains management is a popular discipline which was up to now not monitored globally. A multivariate control chart has therefore been designed to jointly control each delivery path of a supply chain.
- The problem of designing control charts can be very complex. A new idea has been developed in order to provide the parameters of the charts (sample size, time between sampling, control limits,…) that minimize the costs related to monitoring and disfunctioning of the process (find assignable causes, repairing the system, cost of sampling…). This named « economic statistical design » has been applied to many usual control charts.

Flexible Modelling in Financial Risk Management

A) Context and basic ingredients

B) Problems met and proposed solutions

A) Context and basic ingredients



The most common model adopted in the financial and statistical literature devoted to empirical methods in financial markets is the *mean-variance model*. we suppose that the stocks returns are governed by the following equations:

$$r_t = \mu_t + \epsilon_t \tag{1}$$

where $\epsilon_t \sim N(0, \sigma_t^2)$

$$\epsilon_t = \sigma_t z_t \tag{2}$$

$$z \sim N(0, 1) \tag{3}$$

Moreover, the structures of μ_t and σ_t are often specified by ARIMA (p,q) and GARCH (1,1) models. The equations are the following :

$$\mu_t = \sum_{k=1}^{p} \phi_k r_{t-k} + \sum_{k=1}^{q} \theta_k \epsilon_{t-k} \tag{4}$$

$$\sigma_t^2 = \omega + \alpha_{t-1}\sigma_{t-1}^2 + \beta_{t-1}\epsilon_{t-1}^2 \tag{5}$$

Empirical facts

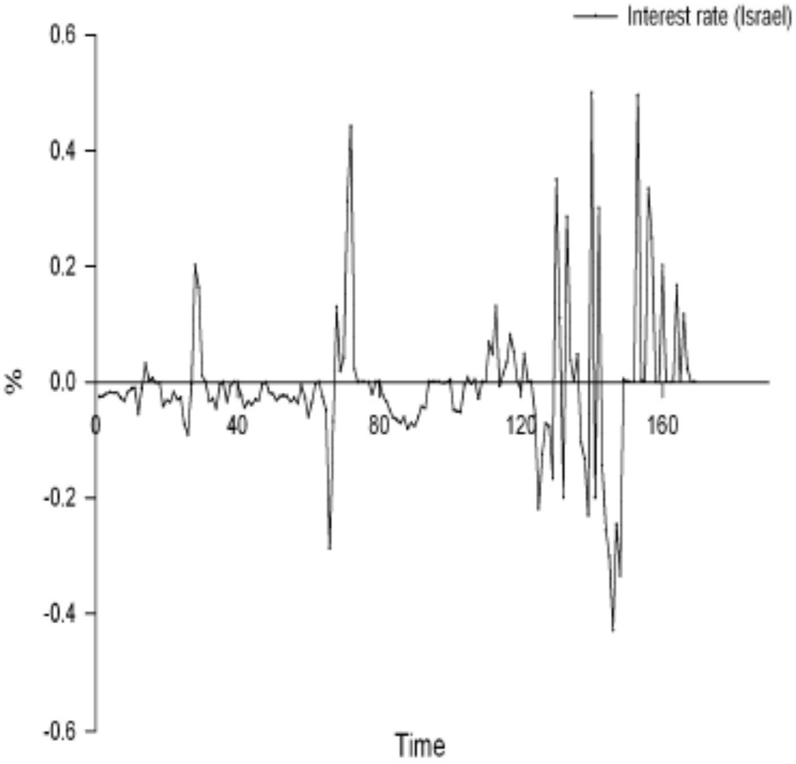


Figure: Israelian interest rates from 1996 to 2011

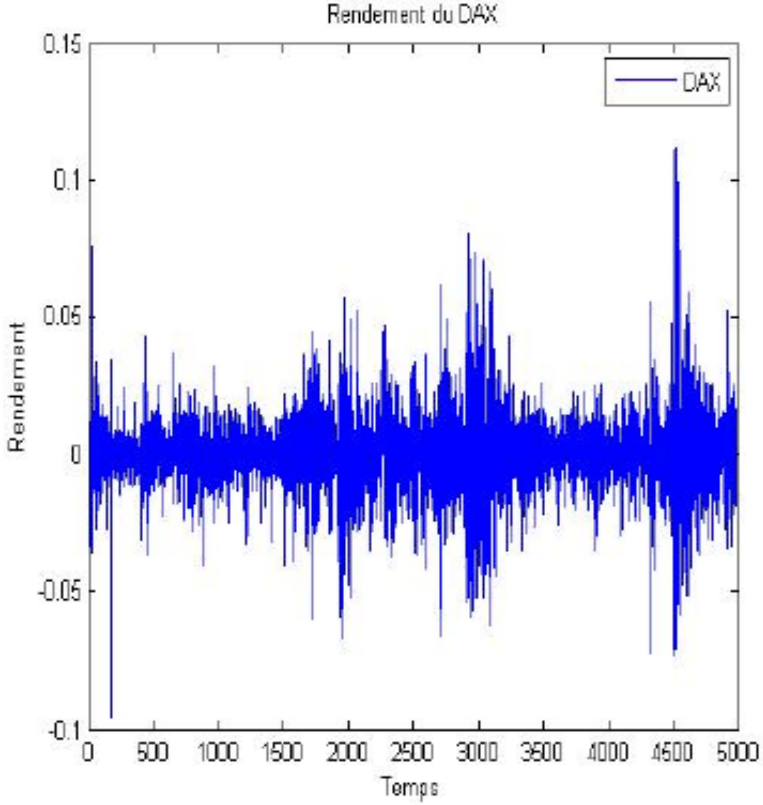


Figure: DAX daily returns from 11/26 1990 to 03/23 2012

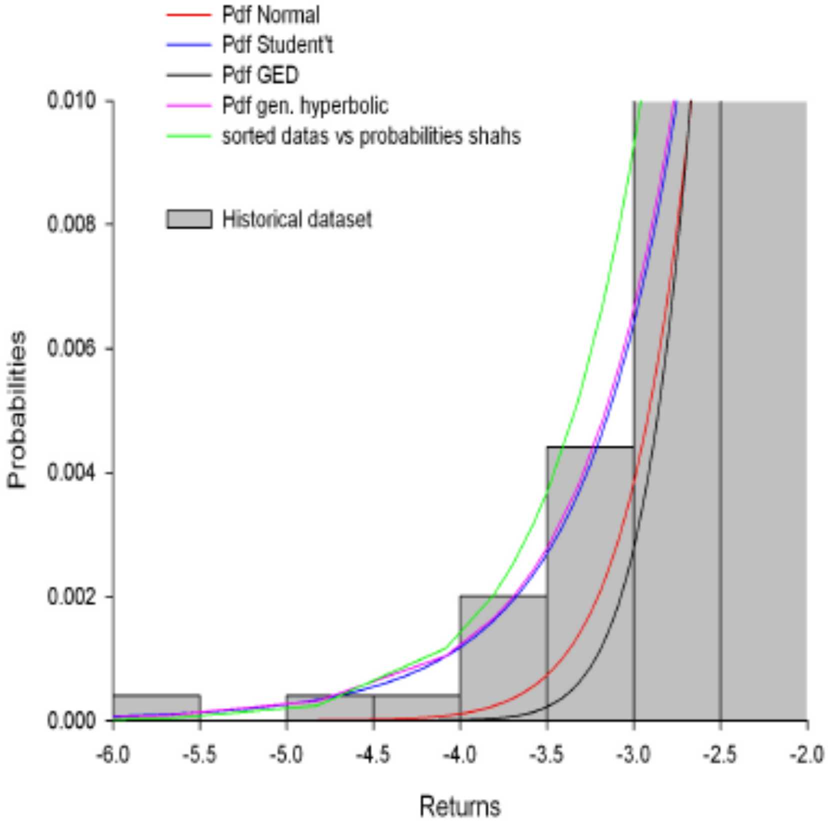


Figure: Adjustments of functions for the residuals of the DJA returns

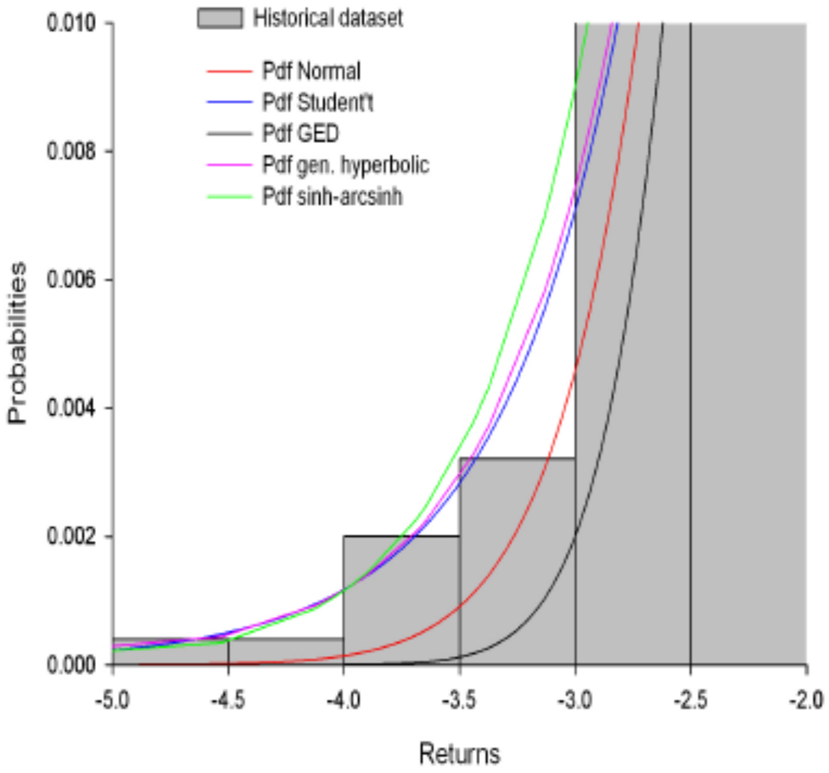


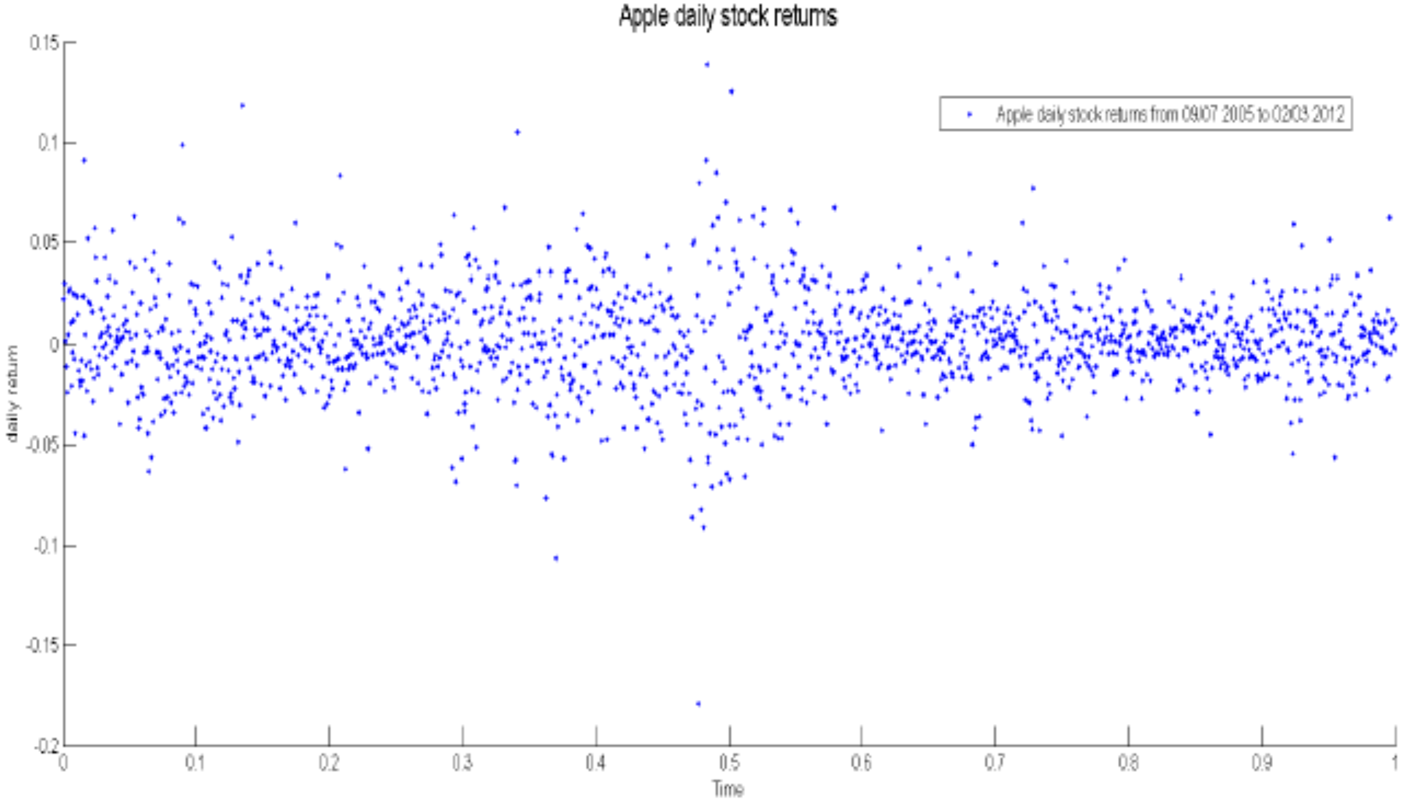
Figure: Adjustments of functions for the residuals of the NIKKEI 225 returns

B) Some problems met and proposed solutions

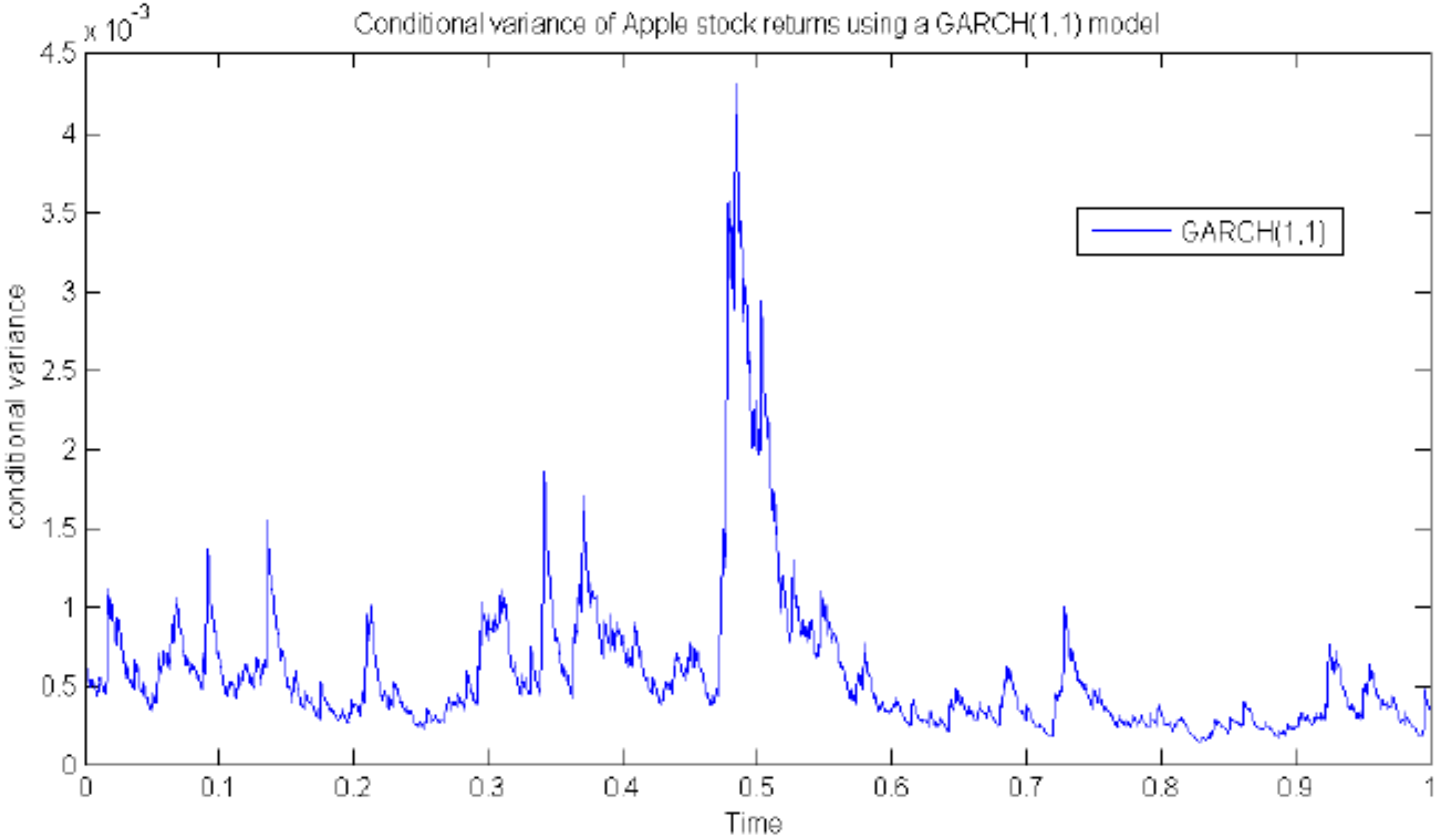
1. Normal and other distributions fail to capture the structure of the tails of the residuals distribution. As a consequence, resulting models fails to predict for example exceptional Values at Risk (VaR).

2. Estimation of the parameters of GARCH models are based on likelihood methods using the parametric distribution of residuals.

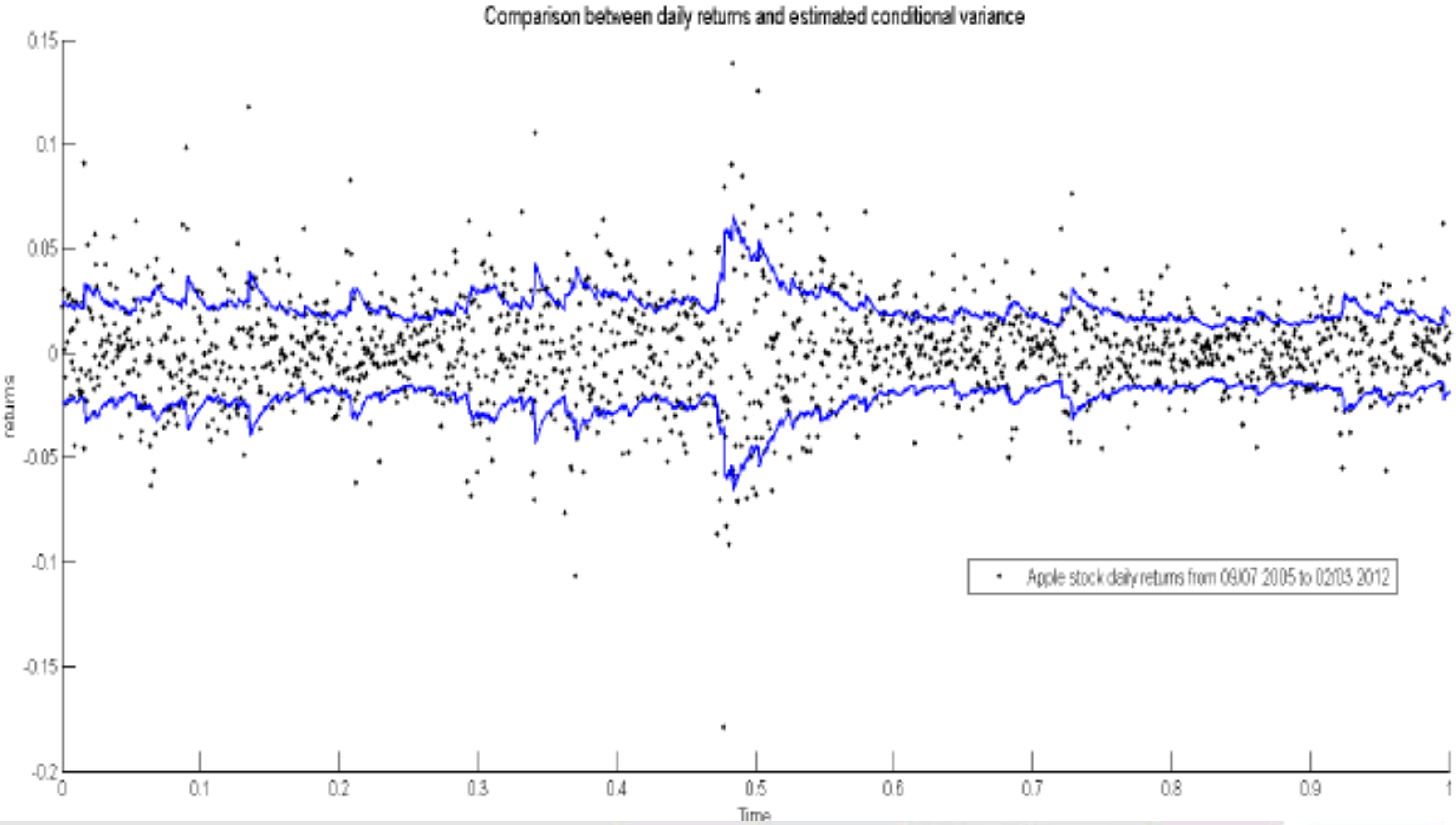
Example : Apple daily stock returns from 09/07 2005 to 02/03 2012



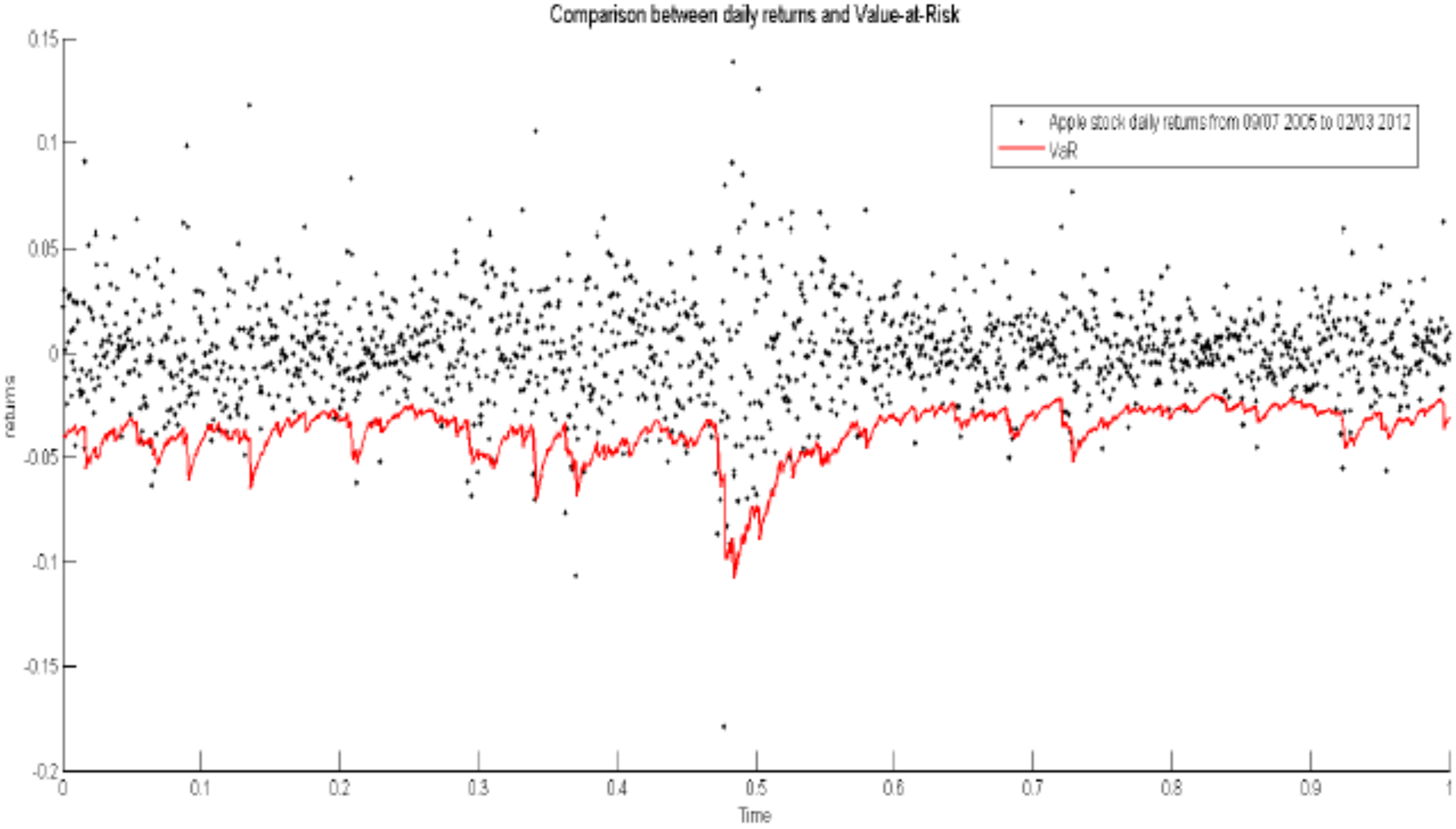
GARCH(1,1) estimated variance



Comparison between estimated GARCH(1,1) variance and daily returns



Comparison between daily returns and estimated VaR

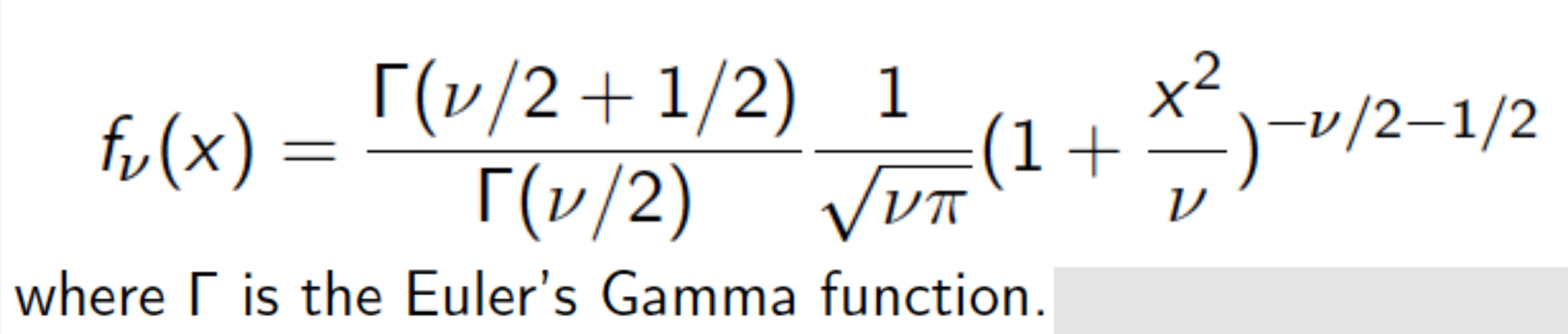


Alternatives to these assumptions:

- use of probability density functions more flexible and closer to the empirical distribution of the returns (like Student-t, generalized hyperbolic or sinh-arcsinh functions) or Extreme Value Theory (EVT)
- use of non-parametric methods to avoid the choice of a bad distribution or a bad structure of variance.
- use of other time-dependent structures in the equations of the variance (alternative GARCH models)

Examples of residuals distributions:

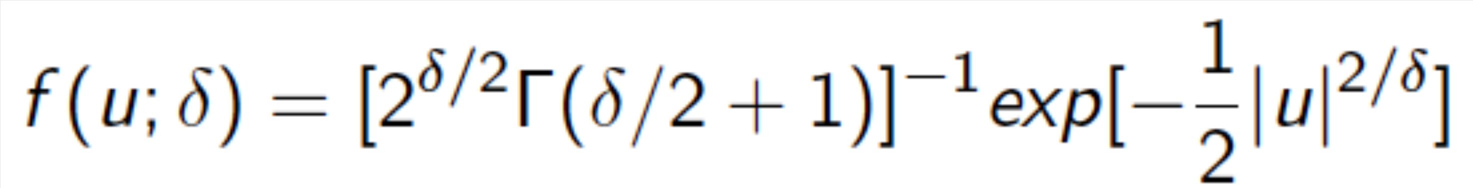
1. t-law



$$f_\nu(x) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}} (1 + \frac{x^2}{\nu})^{-\nu/2-1/2}$$

where Γ is the Euler's Gamma function.

2. General error distribution



$$f(u;\delta) = [2^{\delta/2}\Gamma(\delta/2+1)]^{-1} exp[-\frac{1}{2}|u|^{2/\delta}]$$

3. Generalized hyperbolic distribution
4. …

Basic idea to obtain preliminary nonparametric estimator of the conditional variance at time t:

Use a weighted sum of squares of returns in a neighborhood of t (nonparametric kernel estimators).

Drawback: this method depends on the length of the neighborhood

➔ Possible oversmoothing and need to obtain a good way to automatically compute the length of this neighborhood.

Example with Apple

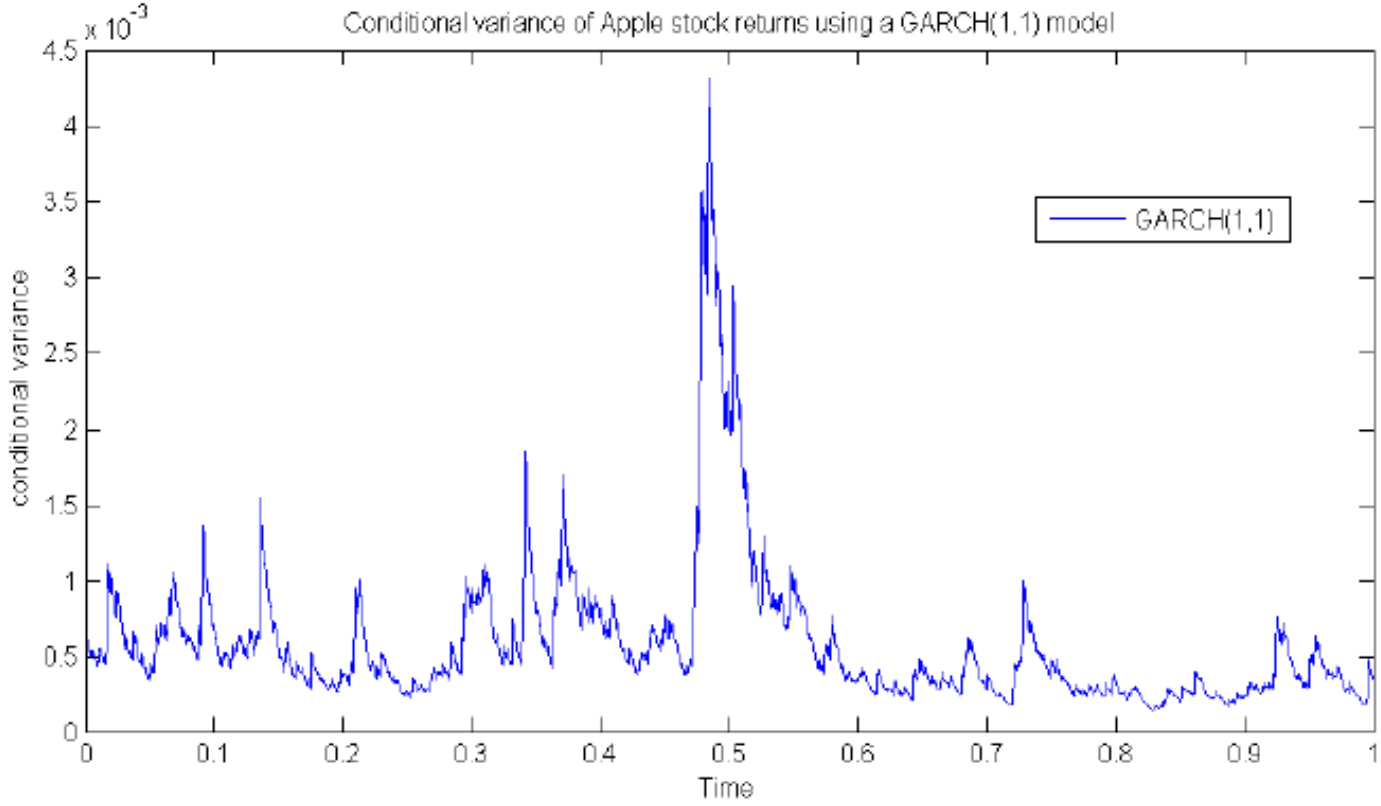
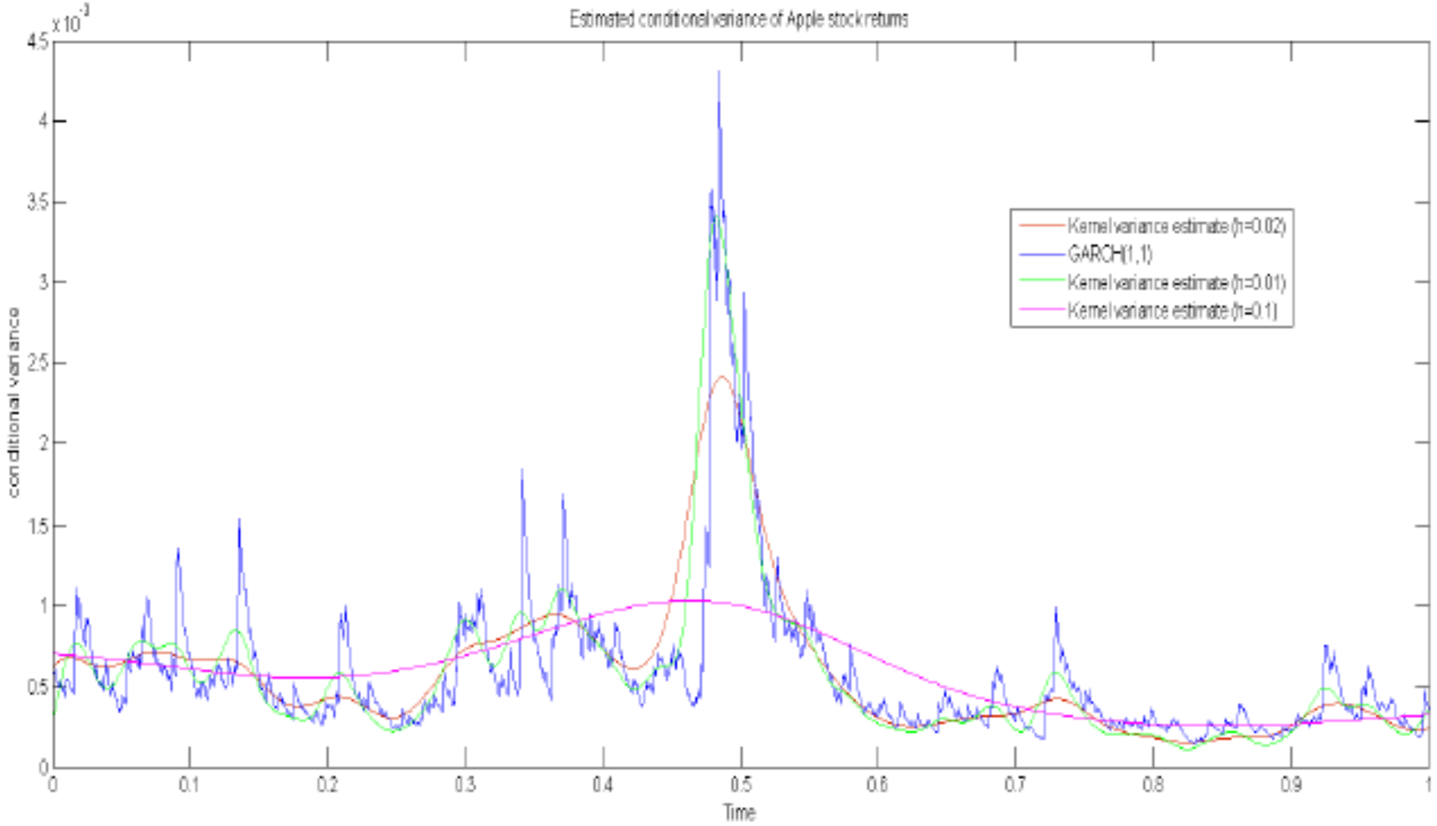


Figure: Conditional variance of Apple stock returns using a GARCH(1,1) model

Comparison with Kernel variance estimate

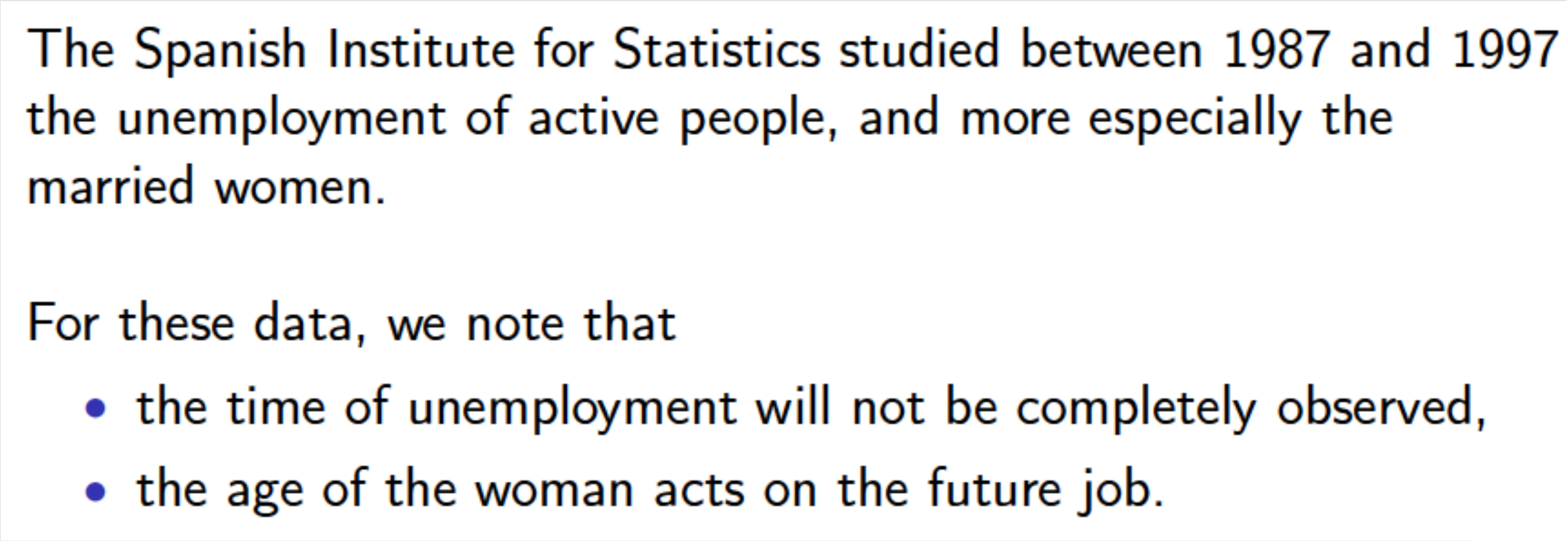


III) Survival Analysis for General Duration Data

A) Example of problems met

B) Idea of proposed solutions

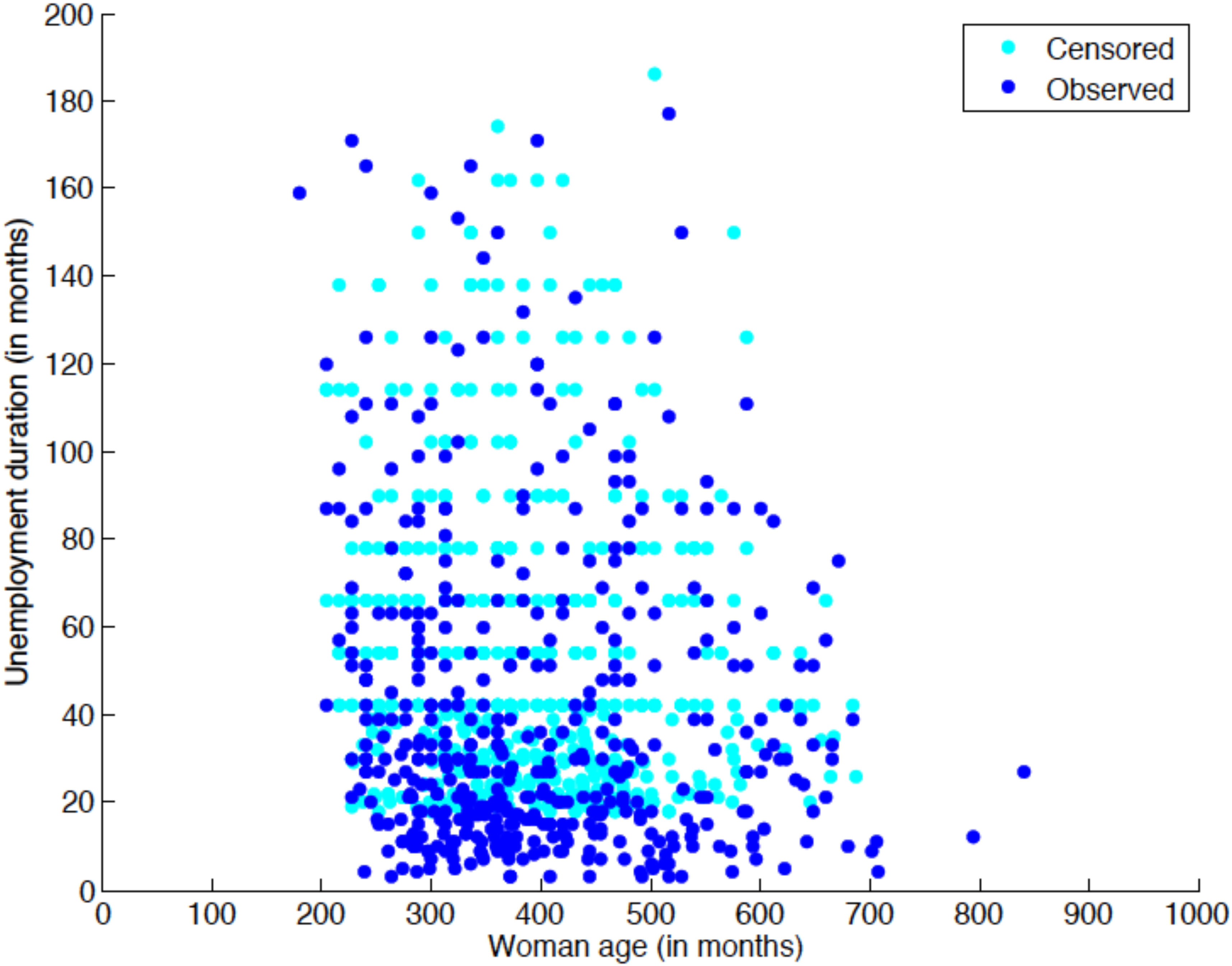
A) Example of problems met



The Spanish Institute for Statistics studied between 1987 and 1997 the unemployment of active people, and more especially the married women.

For these data, we note that

- the time of unemployment will not be completely observed,
- the age of the woman acts on the future job.


Censored
Observed
Unemployment duration (in months)
0
20
40
60
80
100
120
140
160
180
200
Woman age (in months)
0
100
200
300
400
500
600
700
800
900
1000

We consider the nonparametric regression model

$$Y = m(X) + \sigma(X)\varepsilon$$

where

- Y is the response variable
- X is the covariate
- $m(\cdot) = E[Y|\cdot]$ and $\sigma^2(\cdot) = Var[Y|\cdot]$ are unknown smooth functions
- ε is independent of X, with $E[\varepsilon] = 0$ and $Var[\varepsilon] = 1$

Particularity of (X, Y)

- (X, Y) is obtained from cross-sectional sampling
- Y is subject to right censoring.

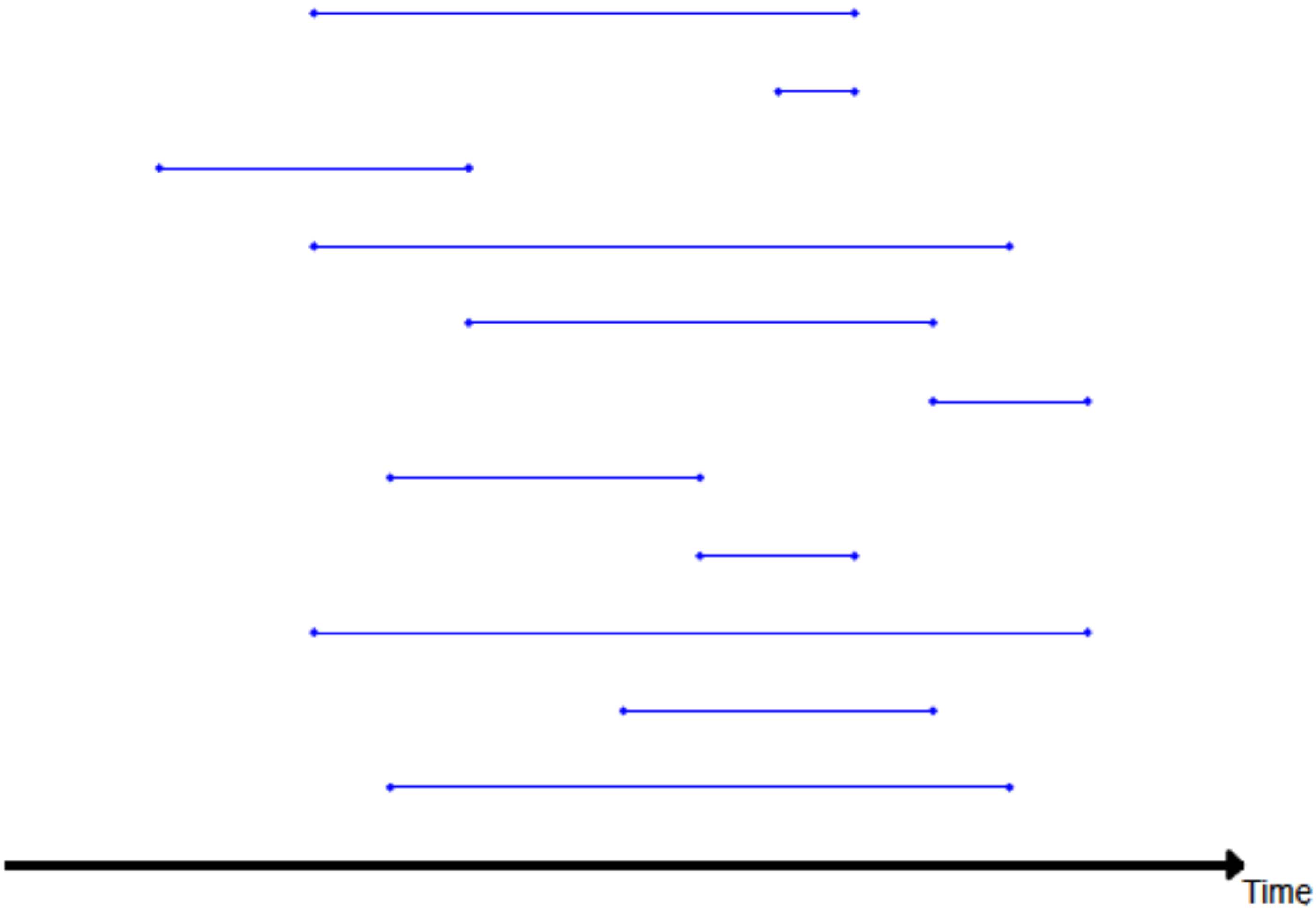
We study the variable Y delimited by

$$T \leq Y \leq C$$

where

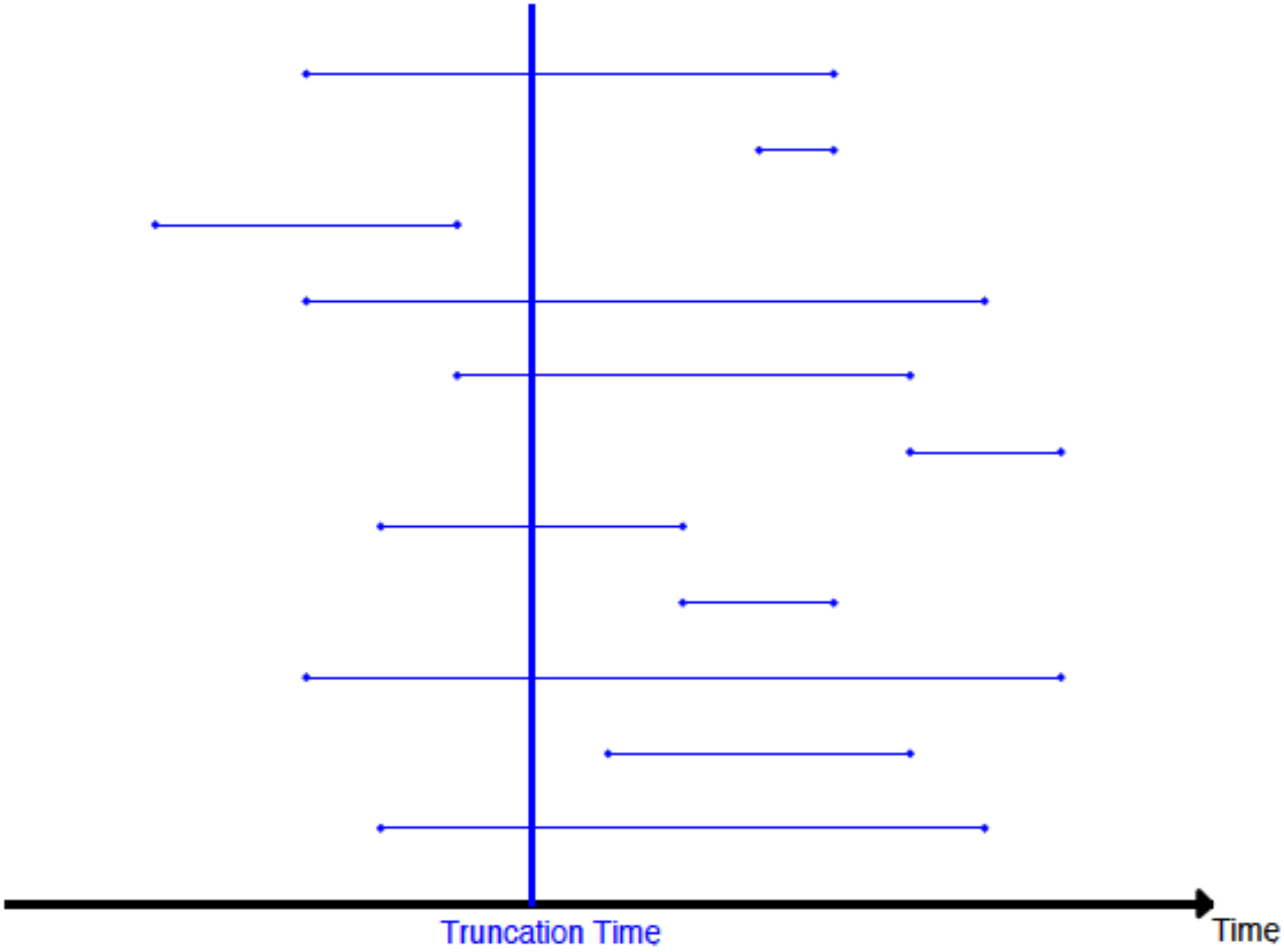
- T is the truncation variable
- C is the censoring variable.

Real World



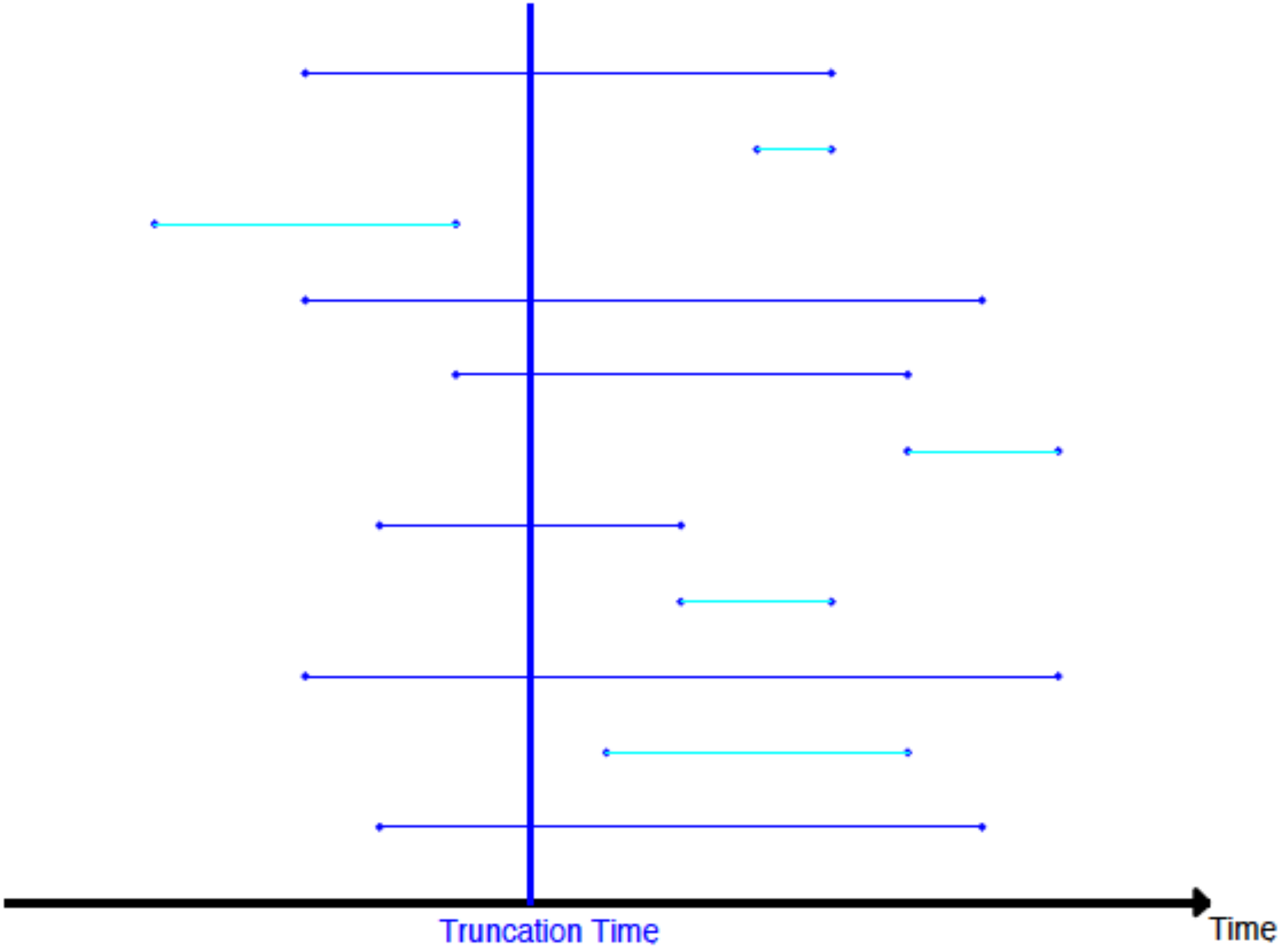
We use as notation F for cdf

Real World



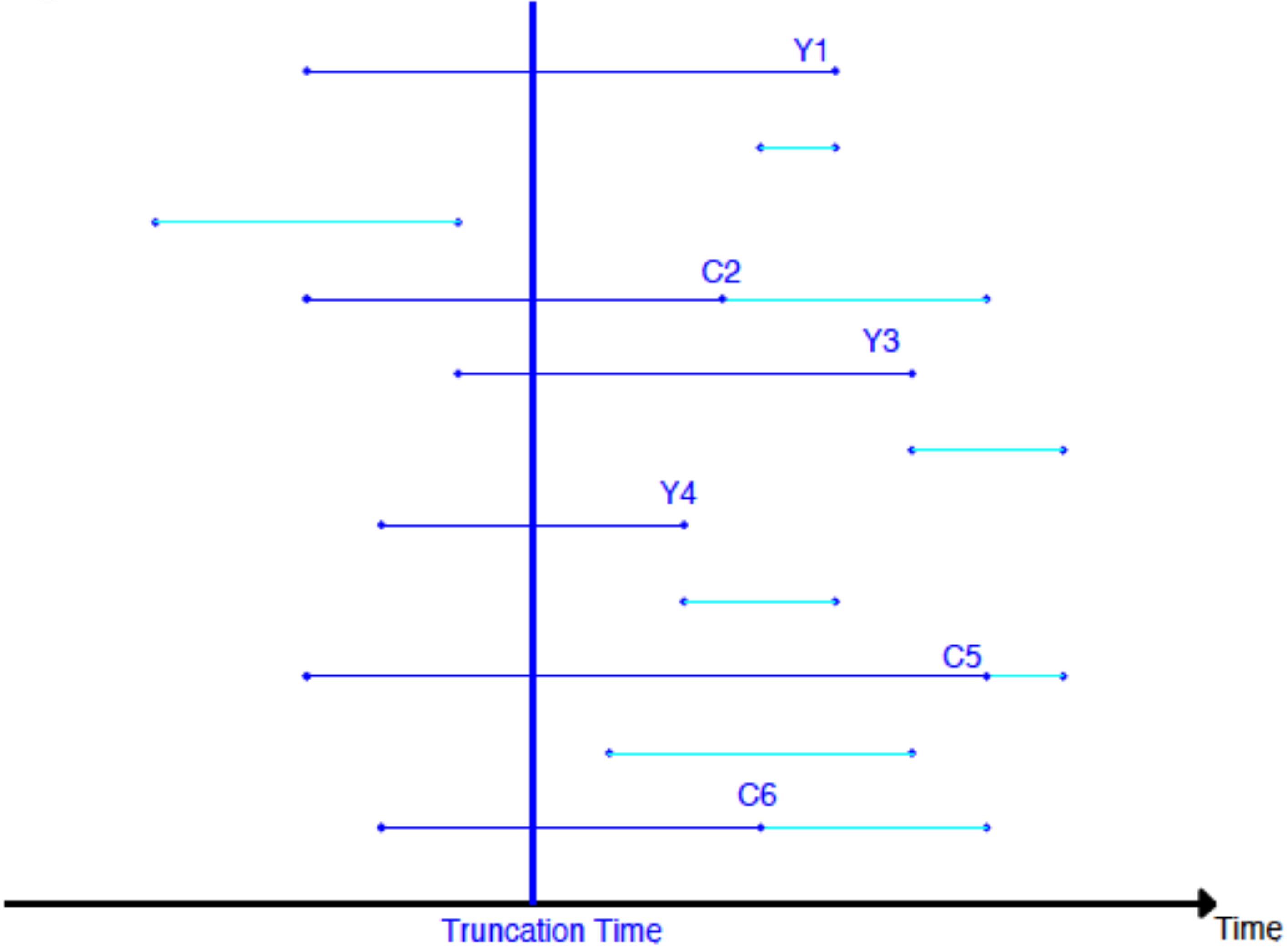
We use as notation F for cdf

Real World



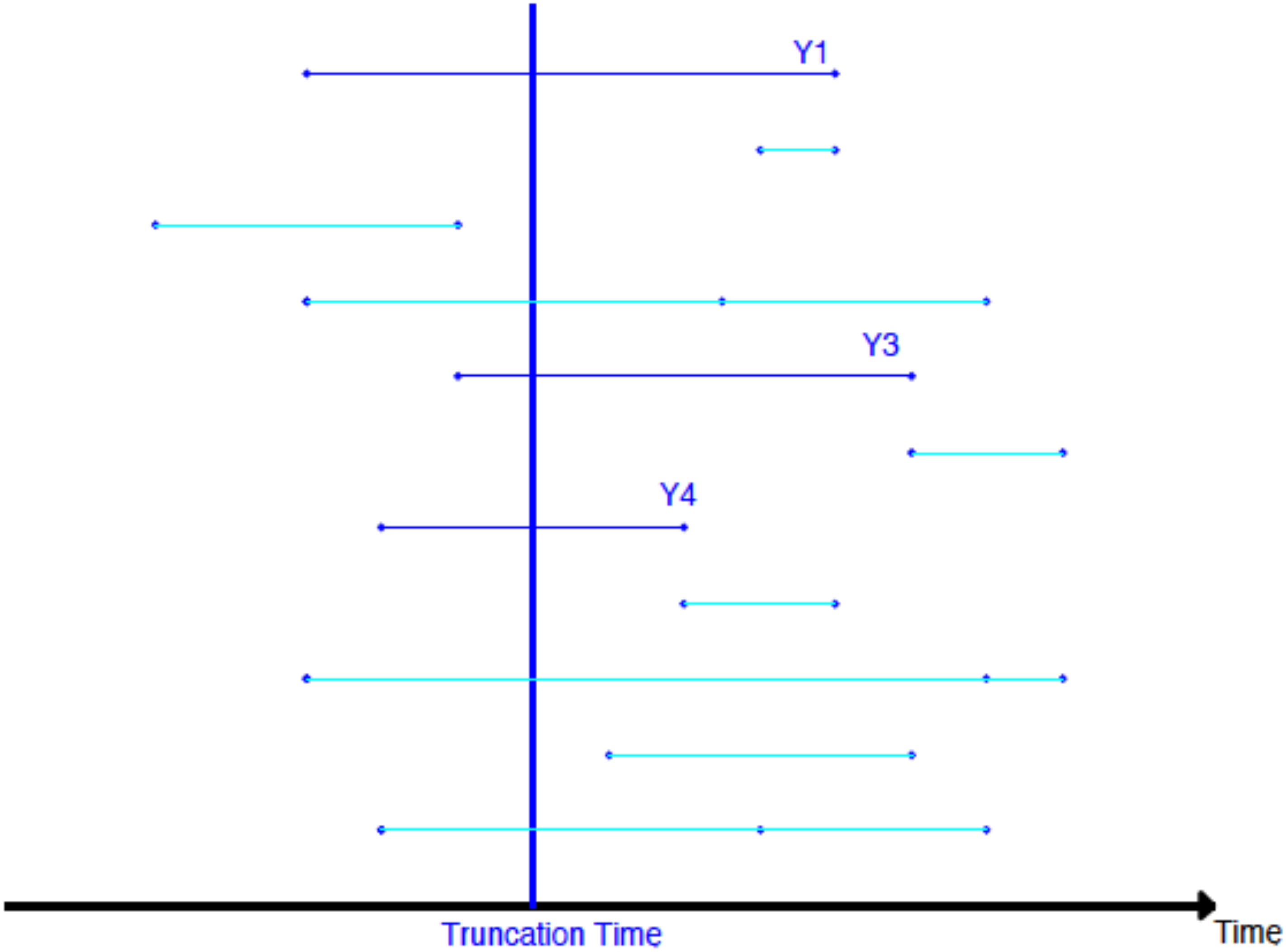
We use as notation F for cdf

Intermediate Observed World



We use as notation $\mathcal{H}$ for cdf, n the sample size

Observed World



We use as notation H for cdf

Aim : Estimation of the error distribution

$$F_{\varepsilon}(e) = \mathbf{P}(\varepsilon \leq e)$$

with

$$(X, Y) \text{ where } T \leq Y \leq C$$

where

- the distribution $F_{T|X}$ is a parametric distribution
- the distribution $F_{C-T|X}$ is completely unknown

Assumptions:

- the variables Y and T are independent, conditionally on X
- for each value x, the support of $F_{Y|X}(\cdot|x)$ is included into the support of $F_{T|X}(\cdot|x)$
- the lower bound of the T support is zero
- the variables (T, Y) and $C - T$ are independent, conditionally on $T \leq Y, X$

We have

$$\begin{aligned} H_{X,Y}(x,y) &= \mathbf{P}(X \leq x, Y \leq y | T \leq Y \leq C) \\ &= (E[w(X,Y)])^{-1} \int_{r \leq x} \int_{s \leq y} w(r,s) dF_{X,Y}(r,s), \end{aligned}$$

the weight function $w(x,y)$ is defined by

$$w(x,y) = \int_{t \leq y} \{1 - \mathcal{G}(y - t|x)\} \, dF_{T|X}(t|x)$$

where $\mathcal{G}(z|x) = \mathbf{P}(C - T \leq z | X = x, T \leq Y)$.

We obtain

$$F_{X,Y}(x,y) = \int_{r \le x} \int_{s \le y} \frac{E[w(X,Y)]}{w(r,s)} dH_{X,Y}(r,s)$$

Therefore,

$$\begin{aligned} F_{\varepsilon}(e) &= \mathbf{P}\left(\frac{Y - m(X)}{\sigma(X)} \le e\right) \\ &= \iint_{\left\{(x,y): \frac{y - m(x)}{\sigma(x)} \le e\right\}} dF_{X,Y}(x,y) \\ &= \iint_{\left\{(x,y): \frac{y - m(x)}{\sigma(x)} \le e\right\}} \frac{E[w(X,Y)]}{w(x,y)} dH_{X,Y}(x,y) \end{aligned}$$

Thus, the estimator is

$$\hat{F}_{\varepsilon}(e) = \frac{1}{M} \sum_{i=1}^{n} \frac{\hat{E}[w(X,Y)]}{\hat{w}(X_i, Y_i)} I\{\hat{\varepsilon}_i \leq e, \Delta_i = 1\}$$

with

$$\hat{\varepsilon}_i = \frac{Y_i - \hat{m}(X_i)}{\hat{\sigma}(X_i)}, \qquad M = \sum_{i=1}^{n} \Delta_i,$$

$$\hat{E}[w(X,Y)] = \left(\frac{1}{M} \sum_{i=1}^{n} \frac{\Delta_i}{\hat{w}(X_i, Y_i)} \right)^{-1}$$

where the functions $\hat{m}(\cdot)$, $\hat{\sigma}(\cdot)$ and $\hat{w}(\cdot,\cdot)$ are nonparametric estimators.

For $\mathcal{G}(t|x)$, we use the Beran (1981) estimator defined by

$$\hat{\mathcal{G}}(t|x) = 1 - \prod_{Z_i \leq t, \Delta_i = 0} \left(1 - \frac{W_i(x, h_n)}{\sum_{j=1}^{n} W_j(x, h_n) I \{Z_j \geq Z_i\}} \right)$$

where

- $Z_i = \min(C_i - T_i, Y_i - T_i)$ and $\Delta_i = I\{Y_i \leq C_i\}$
- $W_i(x, h_n) = \frac{K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)}$ are the Nadaraya-Watson weights
- K is a kernel function
- h_n is a bandwidth sequence tending to 0 when $n \to \infty$

$$=> \hat{w}(x, y) = \int_{t \leq y} \left\{ 1 - \hat{\mathcal{G}}(y - t|x) \right\} dF_{T|X}(t|x)$$

The estimators of $m(\cdot)$ and $\sigma(\cdot)$ are given by

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} \frac{W_i(x,h_n) Y_i \Delta_i}{\hat{w}(x,Y_i)}}{\sum_{i=1}^{n} \frac{W_i(x,h_n) \Delta_i}{\hat{w}(x,Y_i)}},$$

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} \frac{W_i(x,h_n) \Delta_i (Y_i - \hat{m}(x))^2}{\hat{w}(x,Y_i)}}{\sum_{i=1}^{n} \frac{W_i(x,h_n) \Delta_i}{\hat{w}(x,Y_i)}},$$

extension of the estimators in de Uña-Alvarez and Iglesias-Pérez (2008).

- BERAN, R. (1981): *Nonparametric regression with randomly censored survival data*. Technical Report, University of California, Berkeley.
- de UNA-ALVAREZ, J., IGLESIAS-PEREZ, M.C. (2008): Nonparametric estimation of a conditional distribution from length-biased data. *Annals of the Institute of Statistical Mathematics, in press. doi: 10.1007/s10463-008-0178-0.*

R-Chart

- Always look at the Range chart first. The control limits on the X-bar chart are derived from the average range, so if the Range chart is out of control, then the control limits on the X-bar chart are meaningless.
- Look for out of control points. If there are any, then the special causes must be eliminated.
- There should be more than five distinct values plotted.
- If there are values repeated too often, then you have inadequate resolution of your measurements, which will adversely affect your control limit calculations. In this case, you'll have to look at how you measure the variable, and try to measure it more precisely.
- Once the effect of the out of control points from the Range chart is removed, look at the X-bar Chart.

Example: R Control Chart

In the manufacturing of a certain machine part, the percentage of aluminum in the finished part is especially critical. For each production day, the aluminum percentage of five parts is measured. The table below consists of the average aluminum percentage of ten consecutive production days, along with the minimum and maximum sample values (aluminum percentage) for each day. The sum of the 10 samples means (below) is 258.8.

Day	1	2	3	4	5	6	7	8	9	10
Sample Mean	25.2	26.0	25.2	25.2	26.0	25.6	26.0	26.0	24.6	29.0
Maximum Value	26.6	27.6	27.7	27.4	27.6	27.4	27.5	27.9	26.8	31.6
Minimum Value	23.5	24.4	24.6	23.2	23.3	23.3	24.1	23.8	23.5	27.4

$$R_i = \max - \min, i = 1, 2, 3, \ldots, m$$

$$\bar{R} = \frac{\sum R_i}{m}$$

$$UCL = \bar{R}D_4$$

$$LCL = \bar{R}D_3$$

S Chart

- The sample standard deviations are plotted in order to control the variability of a variable.
- For sample size (n>10), the S-chart is more efficient than R-chart.

$$s_i = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$\bar{s} = \frac{\sum s_i}{k}$$

$$UCL = \bar{s}B_4$$

$$LCL = \bar{s}B_3$$

(http://www.statsoft.com/textbook/stquacon.html)

Exponentially-weighted Moving Average (EWMA) Chart

- EWMA Charts are generally used for detecting small shifts in the process mean. They will detect shifts of .5 sigma to 2 sigma much faster.

$$EWMA_{(t+1)} = \lambda Y_t + (1-\lambda)EWMA_t$$

$$UCL = EWMA_1 + ks\sqrt{\frac{\lambda}{2-\lambda}}$$

$$LCL = EWMA_1 - ks\sqrt{\frac{\lambda}{2-\lambda}}$$

where λ is the weighting factor. The factor k is chosen generally to be 2 or 3.

(http://www.statsoft.com/textbook/stquacon.html)

P-Chart

- It is used when the sample size varies: the total number of circuit boards, meals, or bills delivered varies from one sampling period to the next.

Repeated samples of 150 coffee cans are inspected to determine whether a can is out of round or whether it contains leaks due to improper construction. Such a can is said to be nonconforming. Following is the data.

$$UCL = \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n_j}}$$

Sample	1	2	3	4	5	6	7	8	9	10
Nonconforming#	19	10	4	6	8	9	3	1	0	4

$$LCL = \max[0, \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n_j}}]$$

C-Chart

- There are more than one defect per unit.
- Examples might include: the number of defective elements on a circuit board, the number of defects in a dining experience--order wrong, food too cold, check wrong, or the number of defects in bank statement, invoice, or bill.

- The c chart is useful when it's easy to count the number of defects and the sample size is always the same.

An automobile assembly worker is interested in monitoring and controlling the # of minor paint blemishes appearing on the outside door panel on the driver's side of a certain make of automobile. The following data were obtained, using a sample of 25 door panel.

Sample	1	2	3	4	5	6	7	-----	-----	25
# of Paint Blemishes	19	10	4	6	8	9	3	------	-----	4

$$UCL = \bar{c} + 3\sqrt{\bar{c}} \qquad LCL = MAX\left[0, \bar{c} - 3\sqrt{\bar{c}}\right]$$

U-Chart

- The u chart will help evaluate process stability when there can be more than one defect per unit.

- It is used when the sample size varies: the number of circuit boards, meals, or bills delivered each day varies.

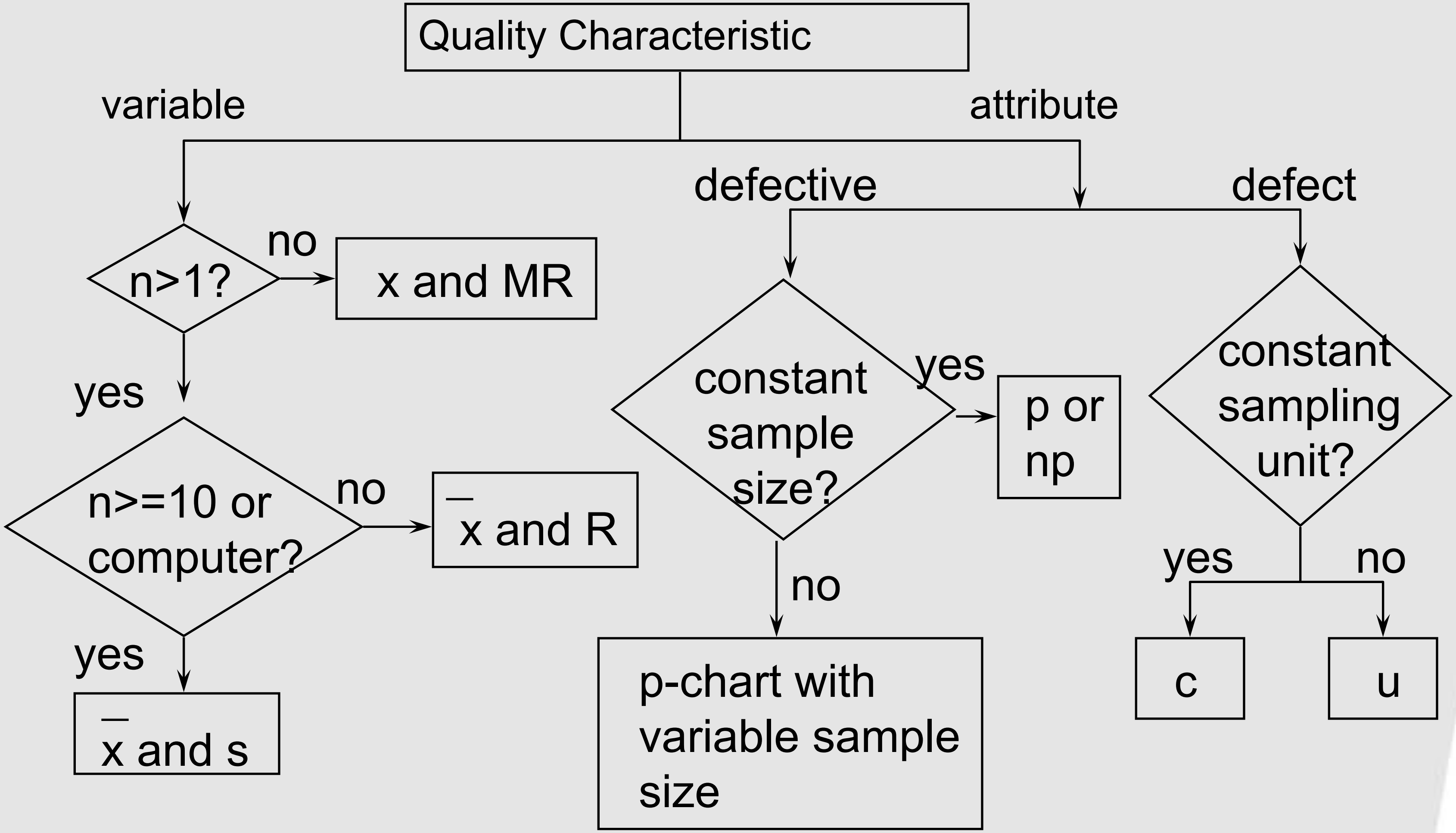
$$u_j = \frac{(count)_j}{n_j}$$

$$UCL = \bar{u} + 3\sqrt{\frac{\bar{u}}{n_j}}$$

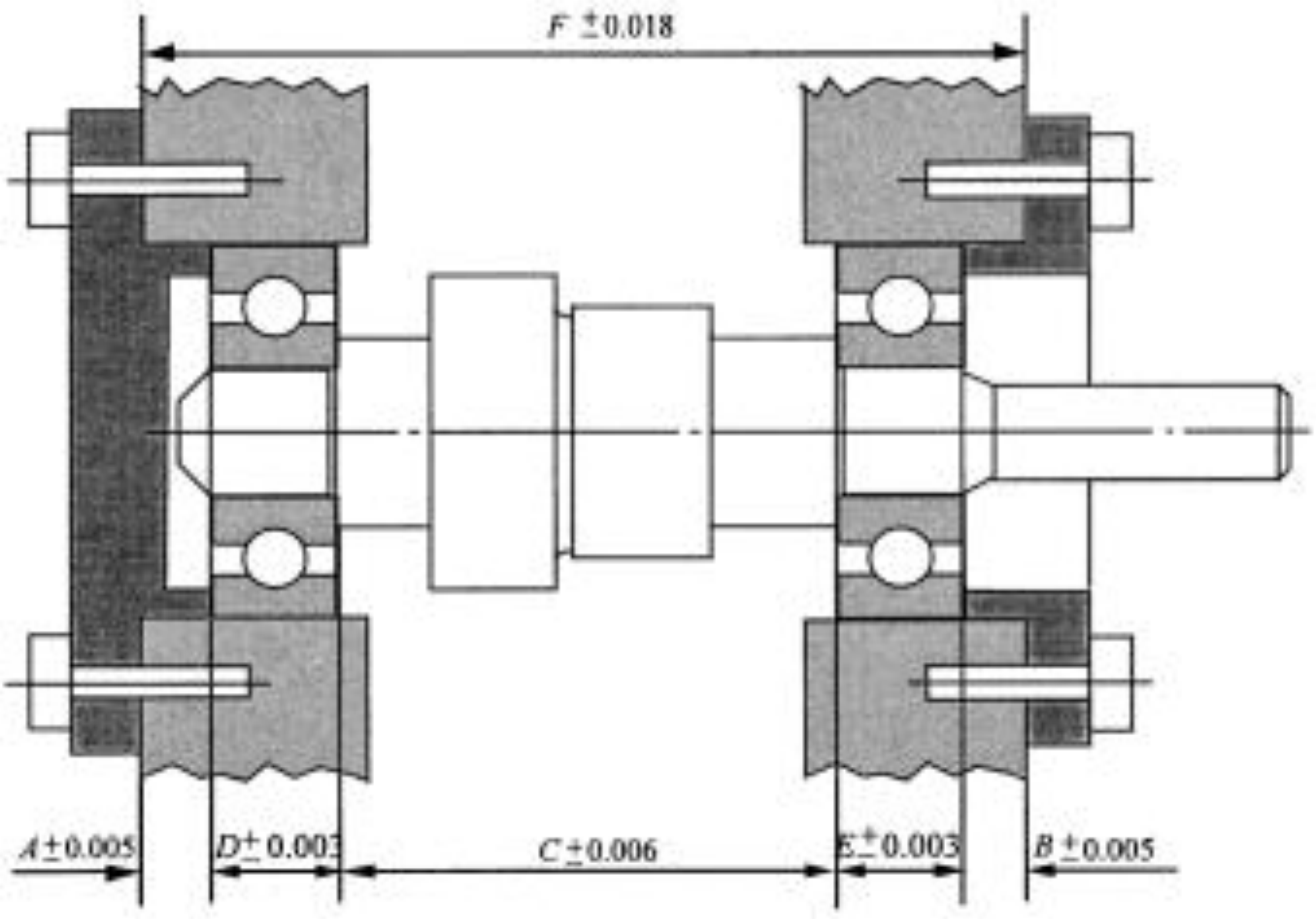
$$\bar{u} = \frac{\sum_{j=1}^{m}(count)_j}{m}$$

$$LCL = MAX\left[0\,,\bar{u} - 3\sqrt{\frac{\bar{u}}{n_j}}\right]$$

Control Chart Selection



Multivariate SPC



Ex- Multi Quality Dimensions of an Education System:

1 Funding

2 student/:staff ratios

3 quality of teaching staff

4 quality of students

5 Classes and Campus size

6 quality of teaching

  1 experience and training

  2 research record

  3 perceived quality, judged by student

7 research environment

8 level of intellectual challenge

9 level of the curriculum

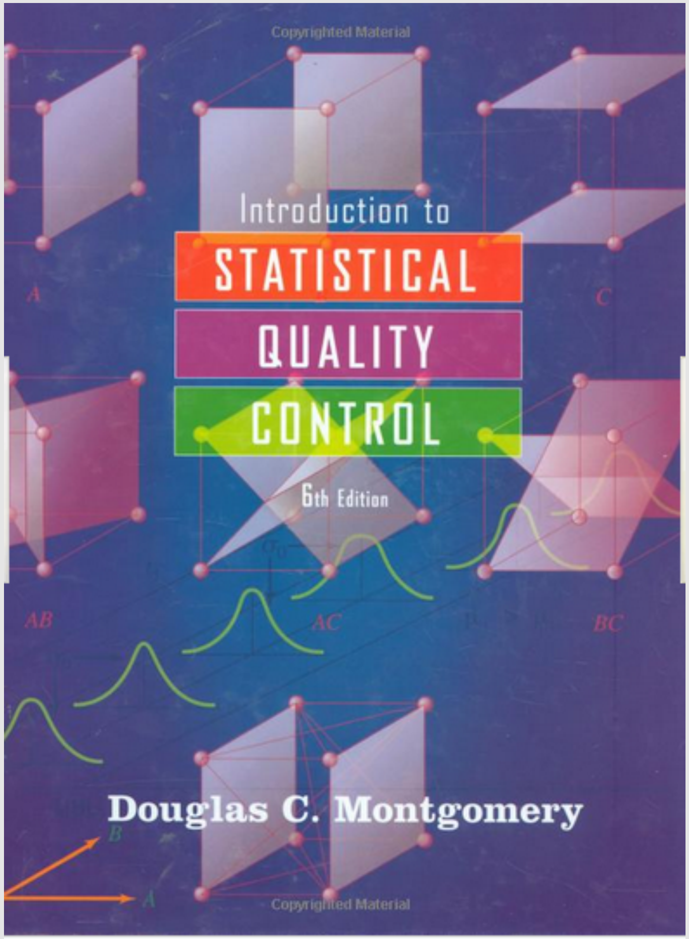
10 student engagement

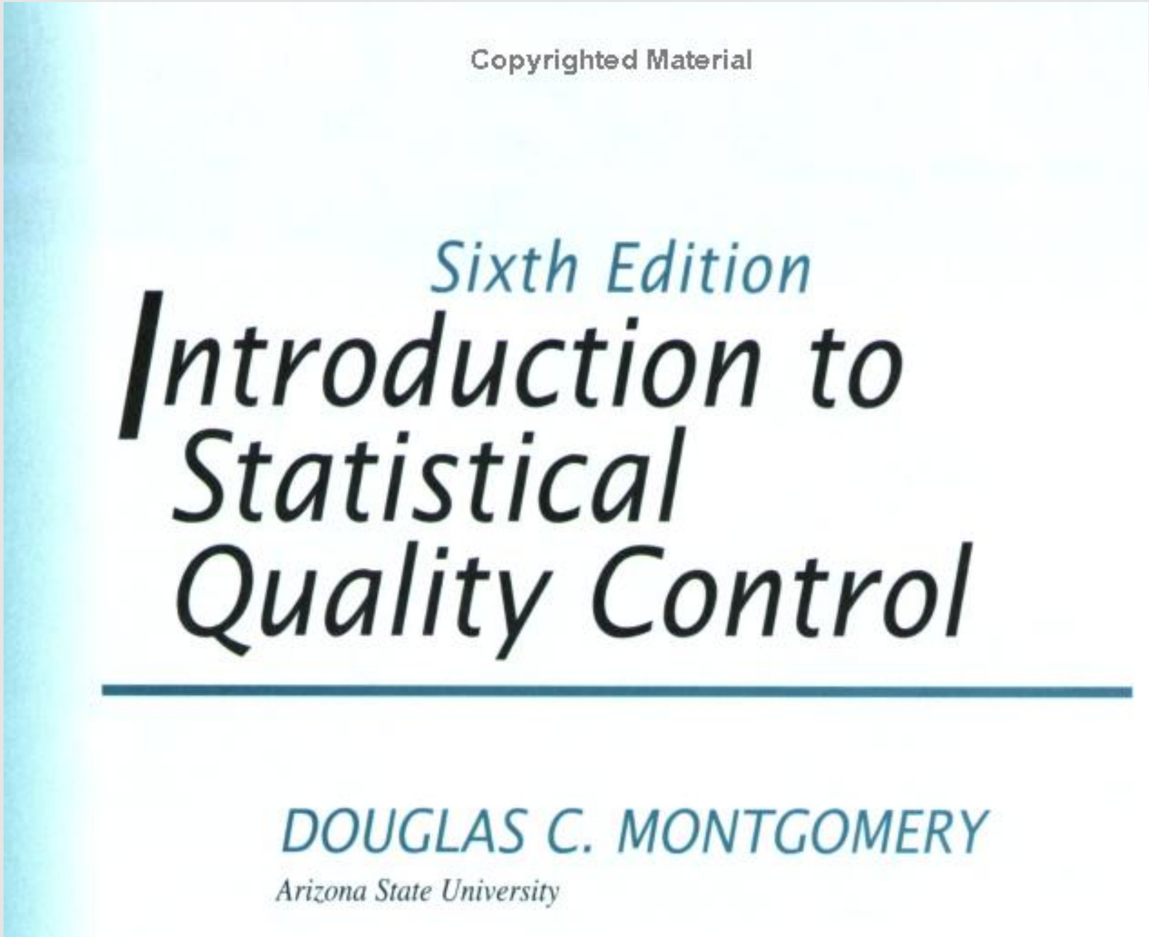
11 Scholar performance and degree classifications

12 student retention and persistence

13 employability and graduate destinations

Reference:





Sixth Edition

Introduction to Statistical Quality Control

DOUGLAS C. MONTGOMERY

Arizona State University