POLITECNICO
MILANO 1863

Quality Data Analysis

Data modeling- part 2

Bianca Maria Colosimo – biancamaria.colosimo@polimi.it

How can we decide whether the standard model is appropriate?

Assumptions	Hypothesis test (to check the assumption)	Remedy in case of violation
“independence” (random pattern)	- Runs test - Bartlett’s test - LBQ’s test	-gapping -batching -(Linear) regression -Time series (ARIMA)
Normal distribution	Normality test	Transform data

Exploring the data- data snooping

Qualitative: Histogram, Boxplot, Run chart

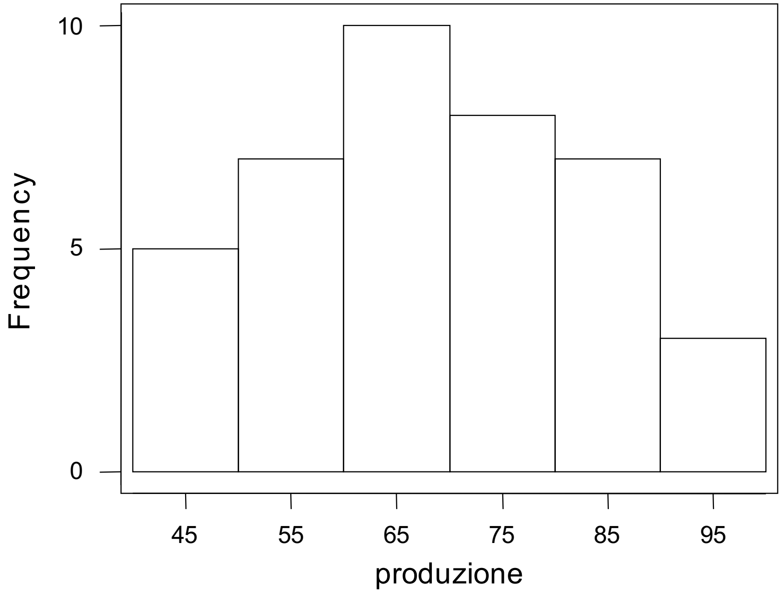
Weekly production of a semiconductor production

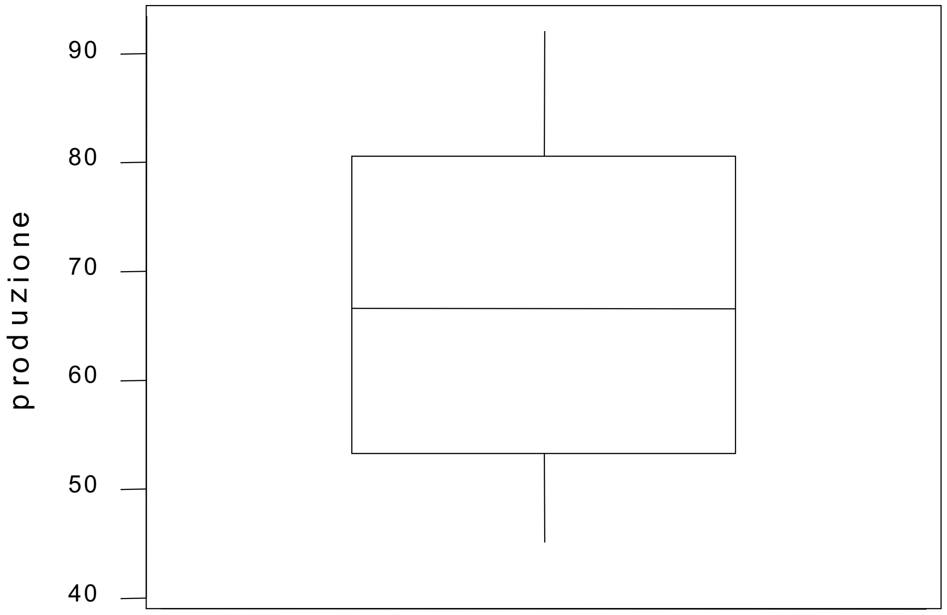
week	products	week	products	week	products	week	products
1	48	11	59	21	68	31	75
2	53	12	54	22	65	32	85
3	49	13	47	23	73	33	81
4	52	14	49	24	88	34	77
5	51	15	45	25	69	35	82
6	52	16	64	26	83	36	76
7	63	17	79	27	78	37	75
8	60	18	65	28	81	38	91
9	53	19	62	29	86	39	73
10	64	20	60	30	92	40	92

40 data: min 45; max 92

Table 2.1 Montgomery

Exploring data: histogram and boxplot





Goodness-of-fit test

The goodness of fit (GOF) tests measure the agreement of a random sample with a theoretical probability distribution function.

We'll see:

- Chi-squared tests
- Shapiro-Wilk

The procedure consists of defining a test statistic (i.e., a random variable that is calculated from sample data to determine whether to reject the null hypothesis. The test statistic compares your data with what is expected under the null hypothesis).

Chi-squared test

H_0: data follow a given F distribution.

H_a: data do not follow this specified distribution

This test is applied to **binned data**. Although there is no optimal choice for the number of bins (k), there are several formulas which can be used to calculate this number based on the sample size (N). For example, $k=1+\log_2 N$

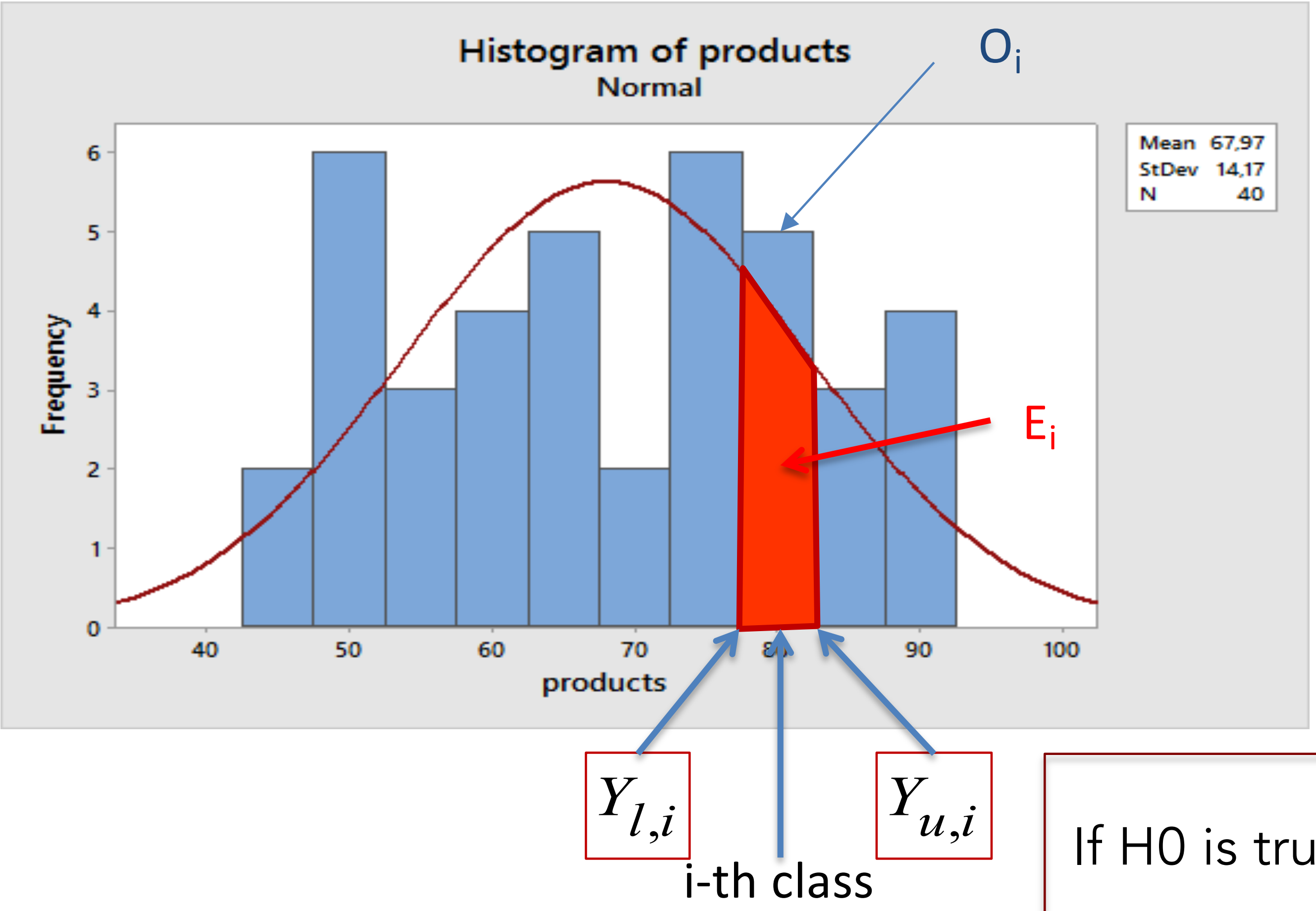
Let i represent the i-th class (i=1,..,*K*);

O_i=observed frequency in class *i*; O_i

E_i=expected frequency in class *i* $E_i = N(F(Y_{u,i}) - F(Y_{l,i}))$

Where $Y_{u,i}$ e $Y_{l,i}$ represent the upper e lower limits for the *i-th class*

https://www.itl.nist.gov/div898/handbook/



O_i

E_i

$Y_{l,i}$

$Y_{u,i}$

i-th class

$$\chi^2 = \sum_{i=1}^{K} (O_i - E_i)^2 / E_i$$

If H0 is true: $\chi^2 \sim \chi^2_{K-c}$

where *c=number of estimated parameters* (ex: mean, variance)

Normality test: Shapiro-Wilk

The Shapiro-Wilk test, calculates a *W* statistic that tests whether a random sample, $x_1, x_2, \dots, x_n$ comes from (specifically) a normal distribution .

Small values of W are evidence of departure from normality and percentage points for the W statistic, obtained via Monte Carlo simulations, were reproduced by Pearson and Hartley (1972, Table 16).

The W statistic is calculated as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{[i]}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where the x(i) are the ordered sample values (x(1) is the smallest) and the ai are constants generated from the means, variances and covariances of the order statistics of a sample of size n from a normal distribution (see Pearson and Hartley (1972, Table 15).

https://www.itl.nist.gov/div898/handbook/prc/section2/prc213.htm

Normality test: Anderson-Darling

The Anderson-Darling Test was developed in 1952 by Theodore Anderson and Donald Darling. It is a statistical test whether or not a dataset comes from a certain probability distribution, e.g., the normal distribution.

H0: The data follows the normal distribution
H1: The data do not follow the normal distribution

The test involves calculating the Anderson-Darling statistic.

$$A^2 = -\frac{\sum_{i=1,n}(2i-1)[\ln F(x_{[i]}) + \ln(1-F(x_{[n+1-i]}))]}{n} - n$$

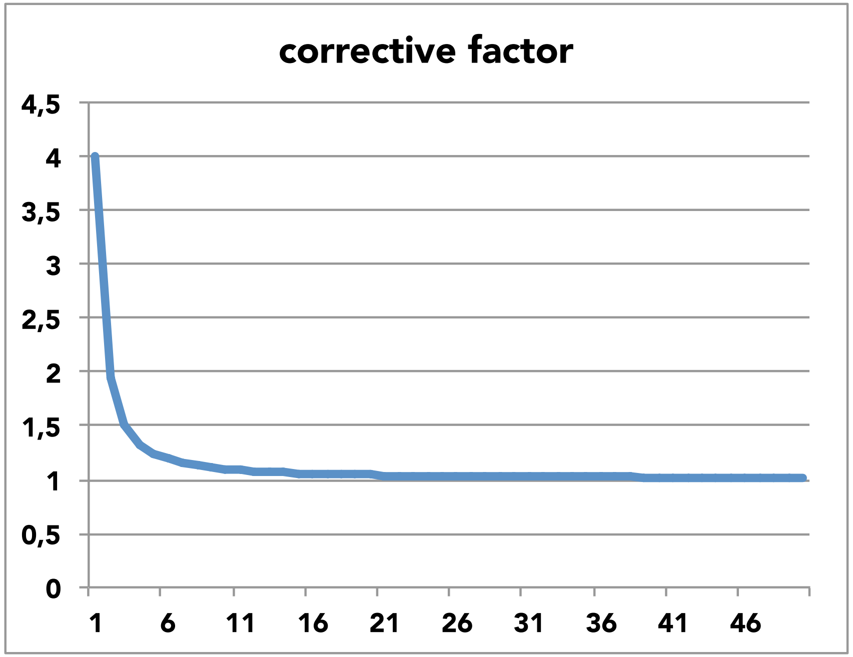
Where:

- n = sample size,
- F(x) = cumulative distribution function for the specified distribution and
- i = the i-th sample when the data is sorted in ascending order.

This statistic is often called A2.

$x_{[1]}, x_{[2]}, \ldots, x_{[i]}, \ldots, x_{[n]}$ Data sorted in ascending order

$F(x_{[1]}), F(x_{[2]}), \ldots, F(x_{[i]}), \ldots, F(x_{[n]})$



The value of A^2 needs to be adjusted for small sample sizes. The adjusted A^2 value is given by:

$$A^2* = A^2 \underbrace{(1 + \frac{0.75}{n} + \frac{2.25}{n^2})}_{\text{Corrective factor}}$$

The calculation of the p value is not straightforward*.

There are different equations depending on the value of A^{2*}. These are given by:

If A^{2*}=>0.6, then p-value = exp(1.2937 - 5.709(A^{2*})+ 0.0186(A^{2*})2

If 0.34 < A^{2*} < .6, then p-value = exp(0.9177 - 4.279(A^{2*}) - 1.38(A^{2*})2

If 0.2 < A^{2*} < 0.34, then p-value = 1 - exp(-8.318 + 42.796(A^{2*})- 59.938(A^{2*})2)

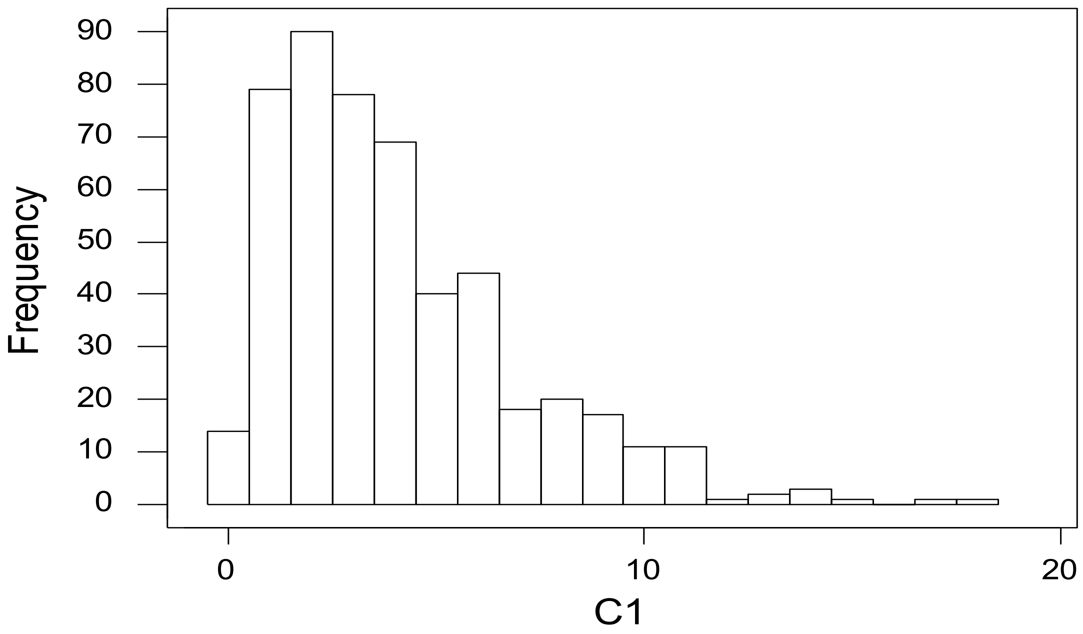
If A^{2*} <= 0.2, then p-value = 1 - exp(-13.436 + 101.14(A^{2*})- 223.73(A^{2*})2)

*The reference most people use is R.B. D'Augostino and M.A. Stephens, Eds., 1986, Goodness-of-Fit Techniques, Marcel Dekker.

Normality test – A-D

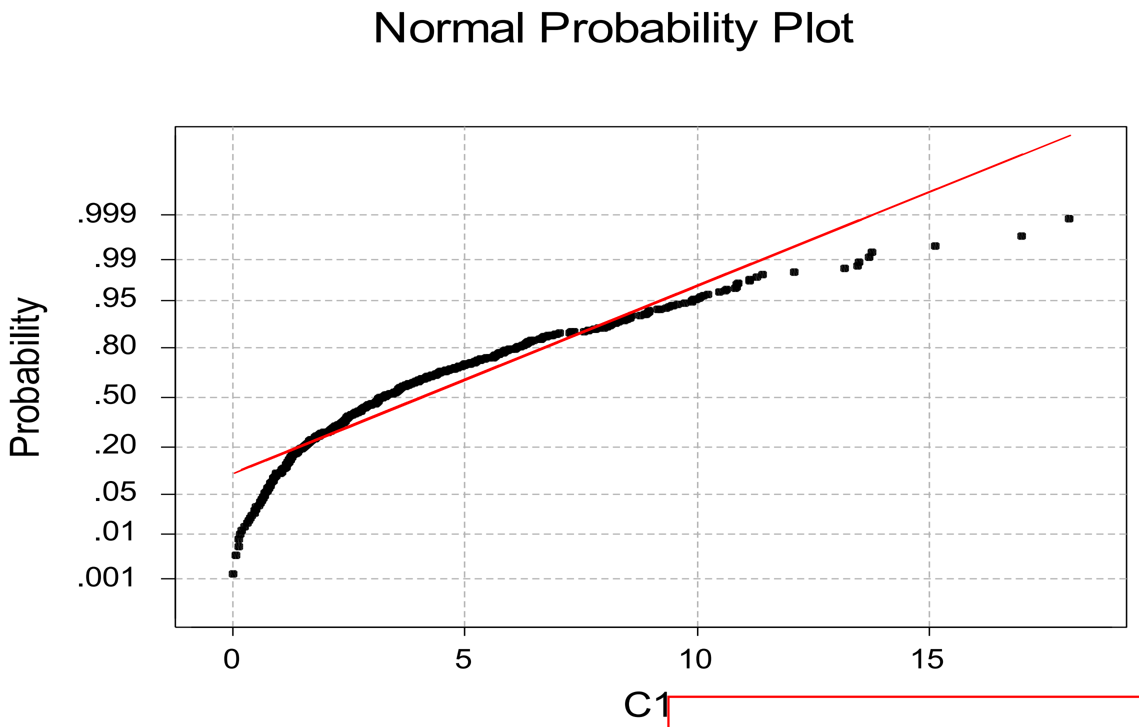
Ex: 500 data

From a $\chi^2(4)$



H_0: data follow a normal distribution.

If the p-value of the test is less than your α level, reject H_0.



Average: 4.07499
StDev: 2.96533
N: 500

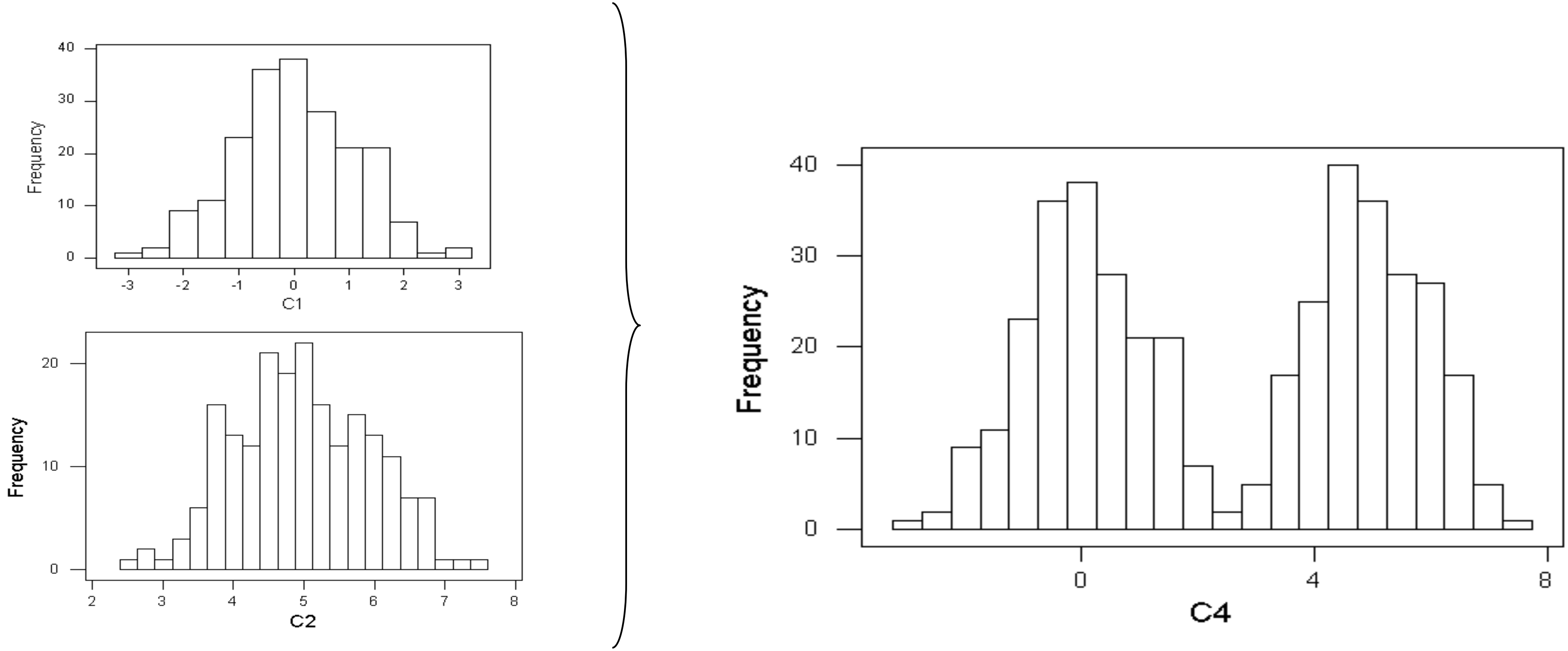
Anderson-Darling Normality Test
A-Squared: 13.292
P-Value: 0.000

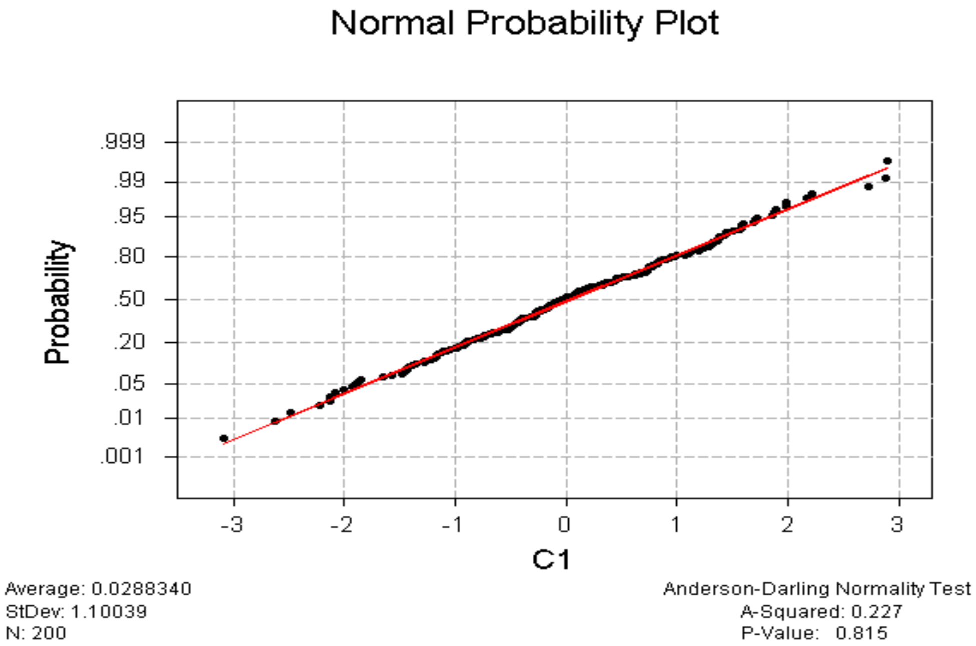
Additional info on normal probability plot:

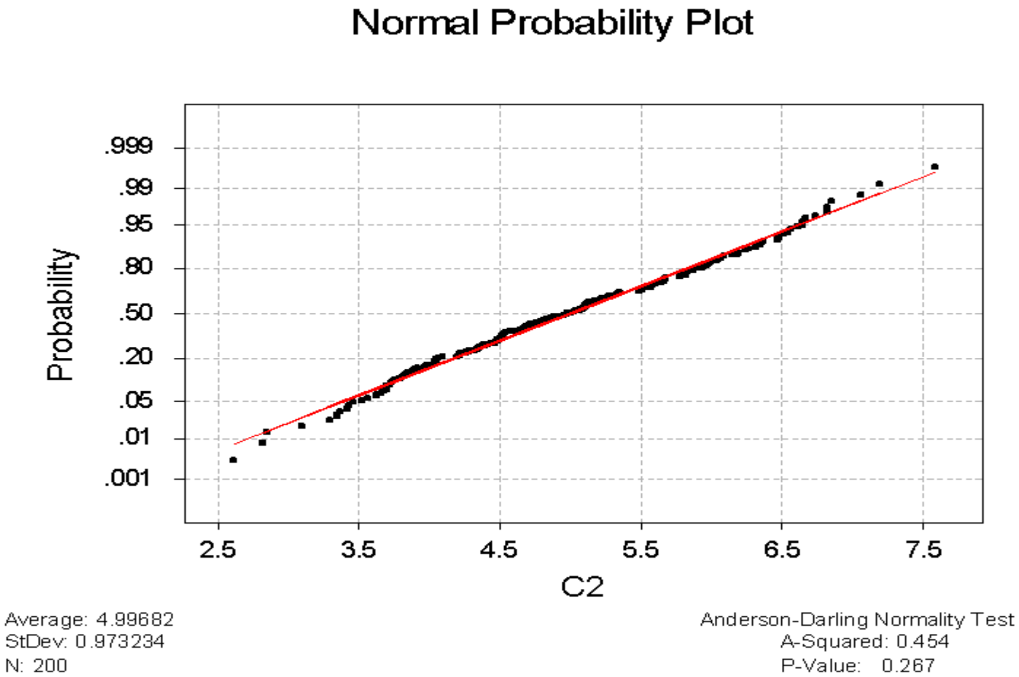
https://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm

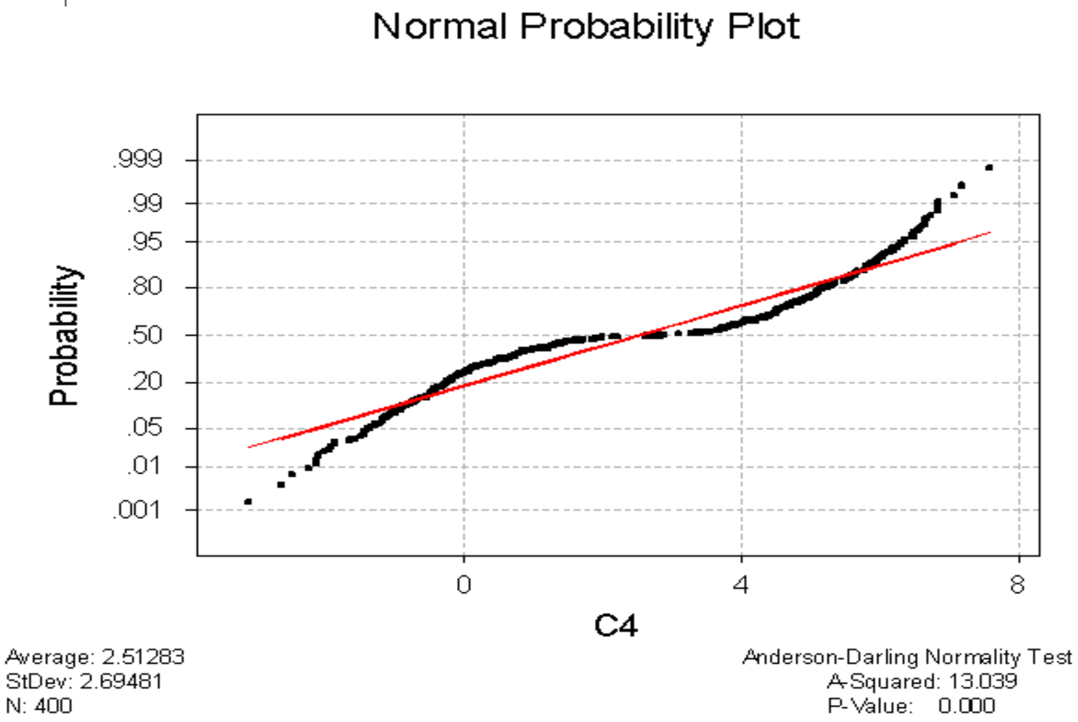
What can we do if data are not normal?

- Mixture: example of output from two different processes that should be (in principle) characterized by the same distribution.




Normal Probability Plot
Probability
.999
.99
.95
.80
.50
.20
.05
.01
.001
-3
-2
-1
0
1
2
3
C1
Average: 0.0288340
StDev: 1.10039
N: 200
Anderson-Darling Normality Test
A-Squared: 0.227
P-Value: 0.815


Normal Probability Plot
Probability
.999
.99
.95
.80
.50
.20
.05
.01
.001
2.5
3.5
4.5
5.5
6.5
7.5
C2
Average: 4.99682
StDev: 0.973234
N: 200
Anderson-Darling Normality Test
A-Squared: 0.454
P-Value: 0.267


Normal Probability Plot
Probability
.999
.99
.95
.80
.50
.20
.05
.01
.001
0
4
8
C4
Average: 2.51283
StDev: 2.69481
N: 400
Anderson-Darling Normality Test
A-Squared: 13.039
P-Value: 0.000

What we can do if data are not normal

What can we do if data are not normal?

- It can be intrinsically due to the process
  - Example:
    - Physical processes:
      - Distorsion measurement (eccentricity, roughness)
      - Electrical phenomena: capacitance, insulation resistance
      - Small levels of substances in the material: porosity, contaminants;
      - Other physical properties (ultimate tensile stress , time to failure)
    - Other processes:
      - Waiting time
      - Km/day for a sale representative
      - Time to repair.

Non normality

Solutions:

1. Use the real distribution

   Ex: esponenziale, Weibull, Gamma for time to failure

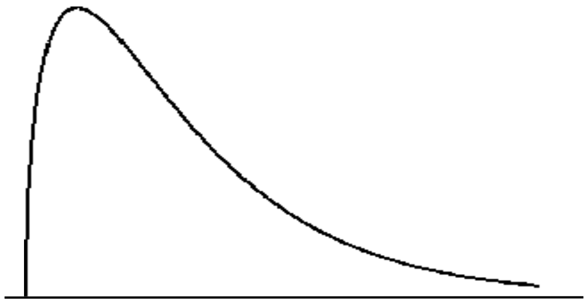
2. Manage the data sampling to deal with the sample average instead of dealing with single data (Central Limit Theorem)
3. Nonparametric methods (ex. runs test): limited use in SPC because they are unsenstitive (robust) to extreme points as outliers. They are not useful when we want to detect outliers
4. Transform data

Transform data for non-normality

Power transformation (Box-Cox transformation):

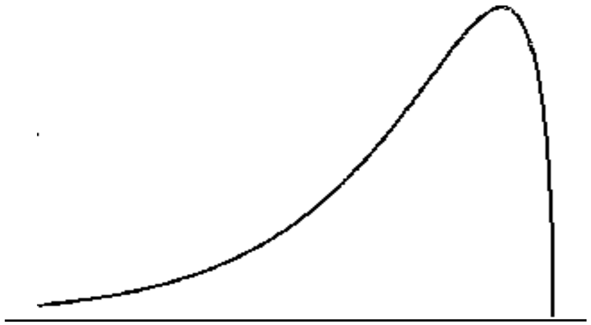
$$f(x) = \begin{cases} x^{\lambda} & \text{se } \lambda \neq 0 \\ \ln x & \text{se } \lambda = 0 \end{cases} \qquad \longrightarrow \qquad \sqrt{x}, {}^{1}\!/_{x}, x^{2}, \ln x$$

- Trial and error:



$\lambda<1$

Positive skewness



$\lambda>1$

Negative skewness

Box-Cox transformation

Problems:

data must be greater than zero (to apply ln and square root):

- If data are negative: we can add a constant
- Trasformation does not work if data that has higher frequence is the smallest data in the set
- … other tranformations (e.g., Johnson transformation)

A useful theorem to draw random data from any CFD

Test based on the empirical cumulative distribution function (CDF)

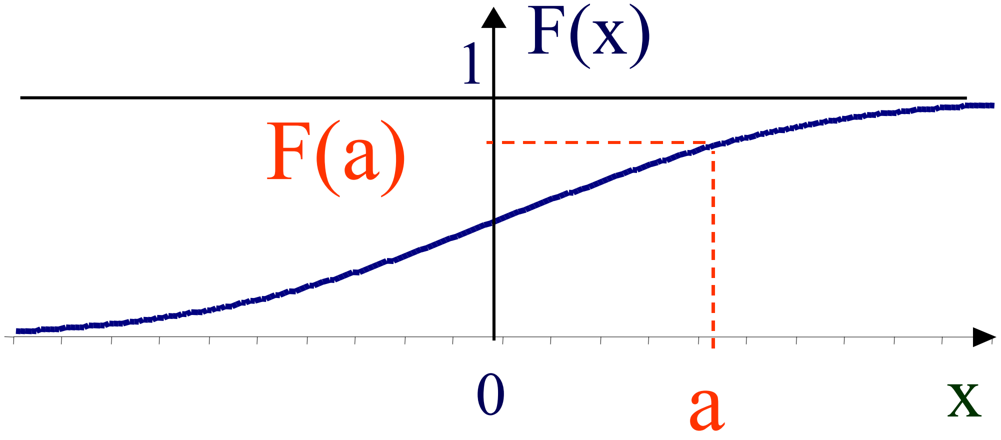
Basic idea:

f(x) : probability density function

$F(x) = \int_{-\infty}^{x} f(t)\,dt$: CDF

Regardless of the specific f(x)

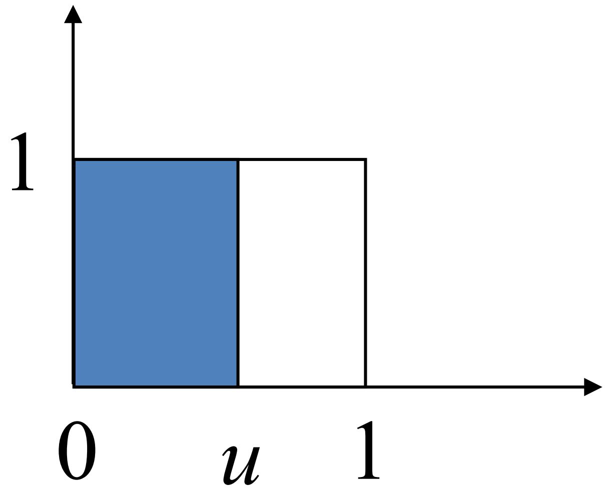
$$\boxed{F \sim \mathrm{Unif}(0,1)}$$

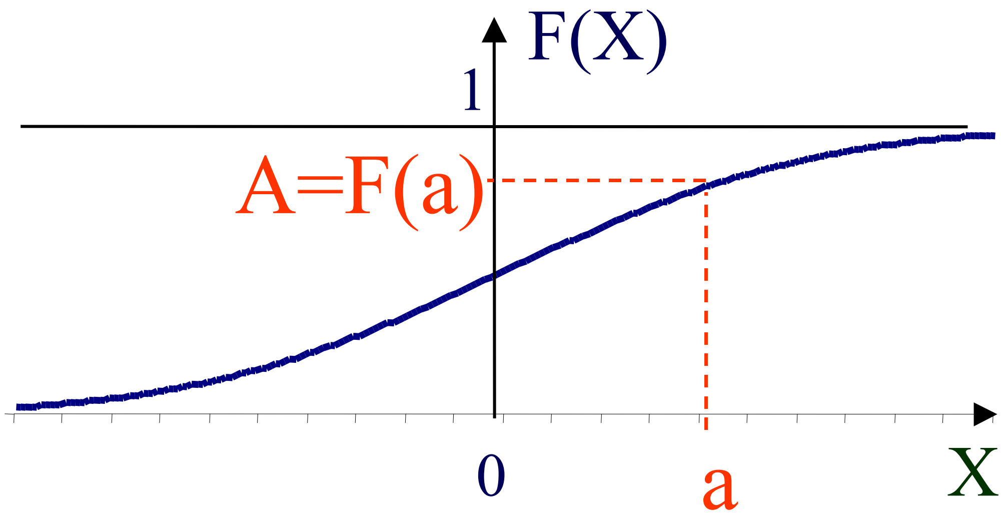


Proof:

U~Unif(0,1) if and only if

$$\Pr(U \le u) = u * 1 = u$$





X=g(F) is a strictly increasing function

$$\Pr(F \le A) = \Pr(g(F) \le g(A)) =$$

$$= \Pr(X \le a) = F(a) = A$$

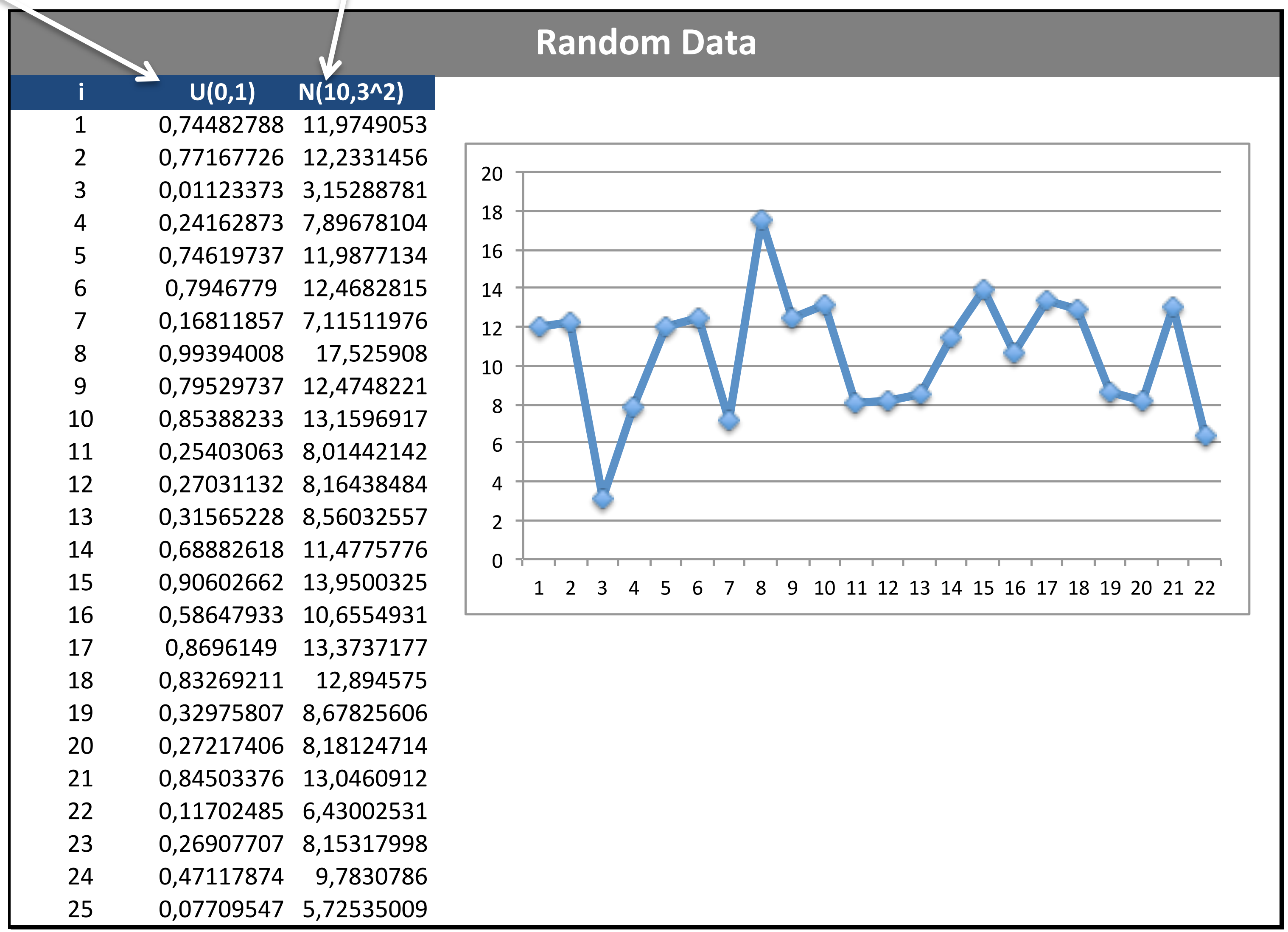
By definition

$\Rightarrow$ F ~Unif(0, 1)

How to draw random data from (any) random distribution

CASUALE()
RANDOM()

INV.NORM.(prob;mean; std dev)



Random Data

i	U(0,1)	N(10,3^2)
1	0,74482788	11,9749053
2	0,77167726	12,2331456
3	0,01123373	3,15288781
4	0,24162873	7,89678104
5	0,74619737	11,9877134
6	0,7946779	12,4682815
7	0,16811857	7,11511976
8	0,99394008	17,525908
9	0,79529737	12,4748221
10	0,85388233	13,1596917
11	0,25403063	8,01442142
12	0,27031132	8,16438484
13	0,31565228	8,56032557
14	0,68882618	11,4775776
15	0,90602662	13,9500325
16	0,58647933	10,6554931
17	0,8696149	13,3737177
18	0,83269211	12,894575
19	0,32975807	8,67825606
20	0,27217406	8,18124714
21	0,84503376	13,0460912
22	0,11702485	6,43002531
23	0,26907707	8,15317998
24	0,47117874	9,7830786
25	0,07709547	5,72535009