PURDUE UNIVERSITY

NCN

www.nanohub.org

ECE695: Reliability Physics of Nano-Transistors

Lecture 31: Collecting and Plotting Data

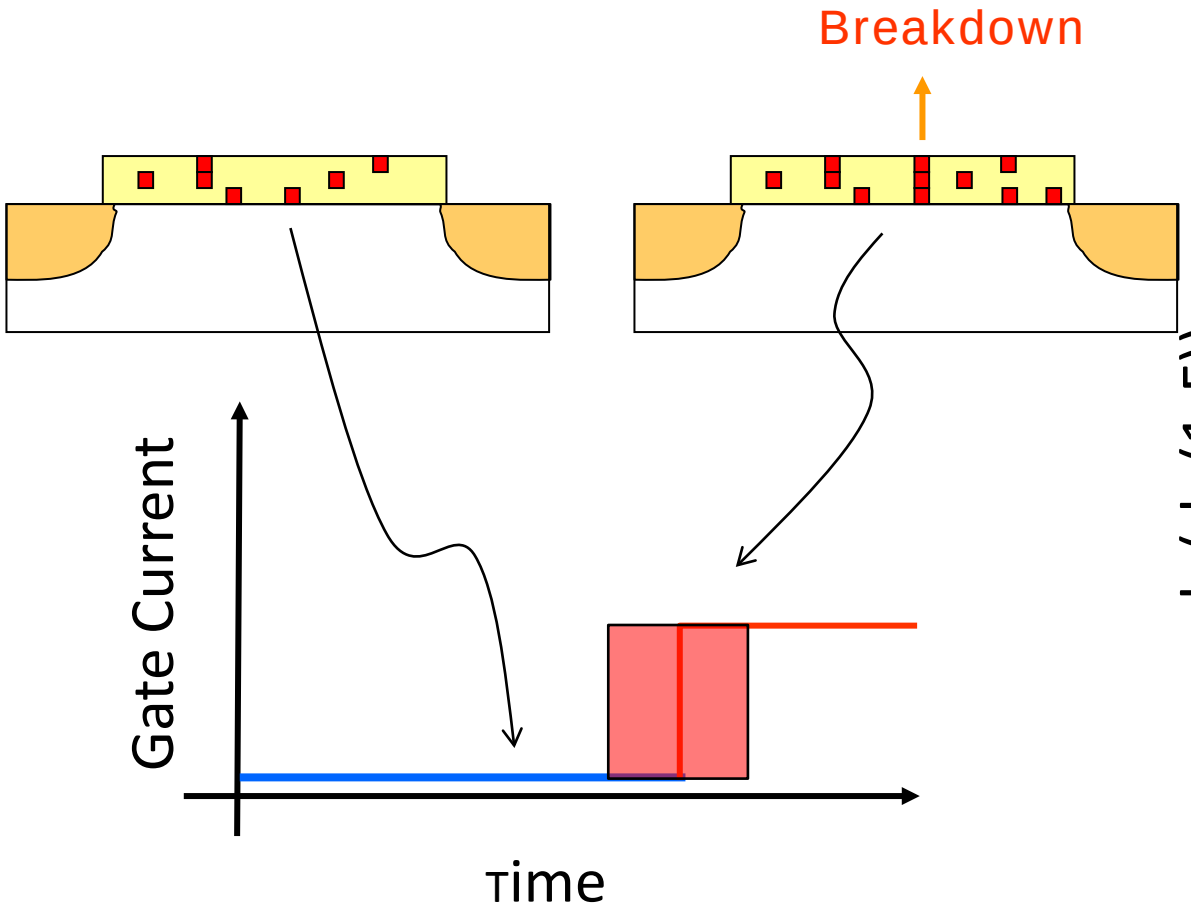
Muhammad Ashraful Alam
alam@purdue.edu

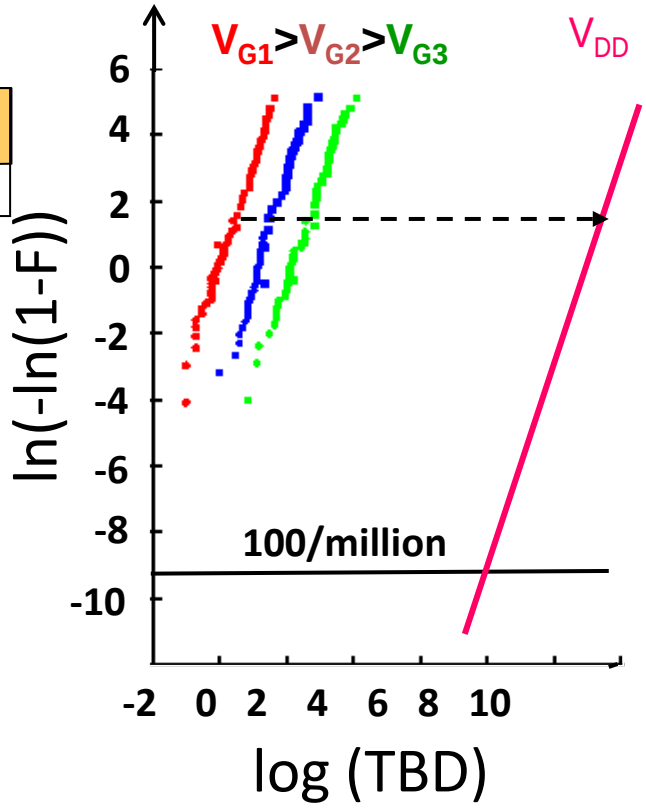
copyright 2011

Outline

1. Origin of data, Field Acceleration vs. Statistical Inference
2. Nonparametric information
3. Preparing data for projection: Hazen formula
4. Preparing data for projection: Kaplan formula
5. Conclusions

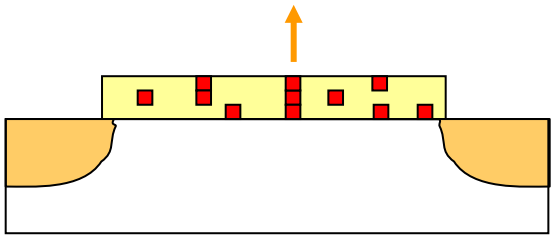
Where do data come from: TDDB Example

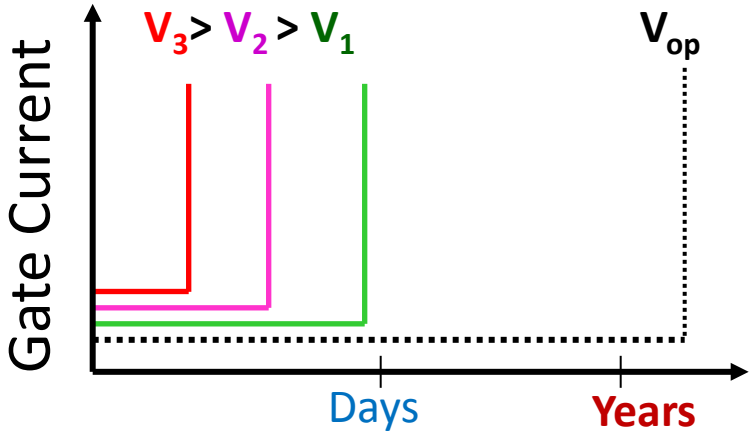


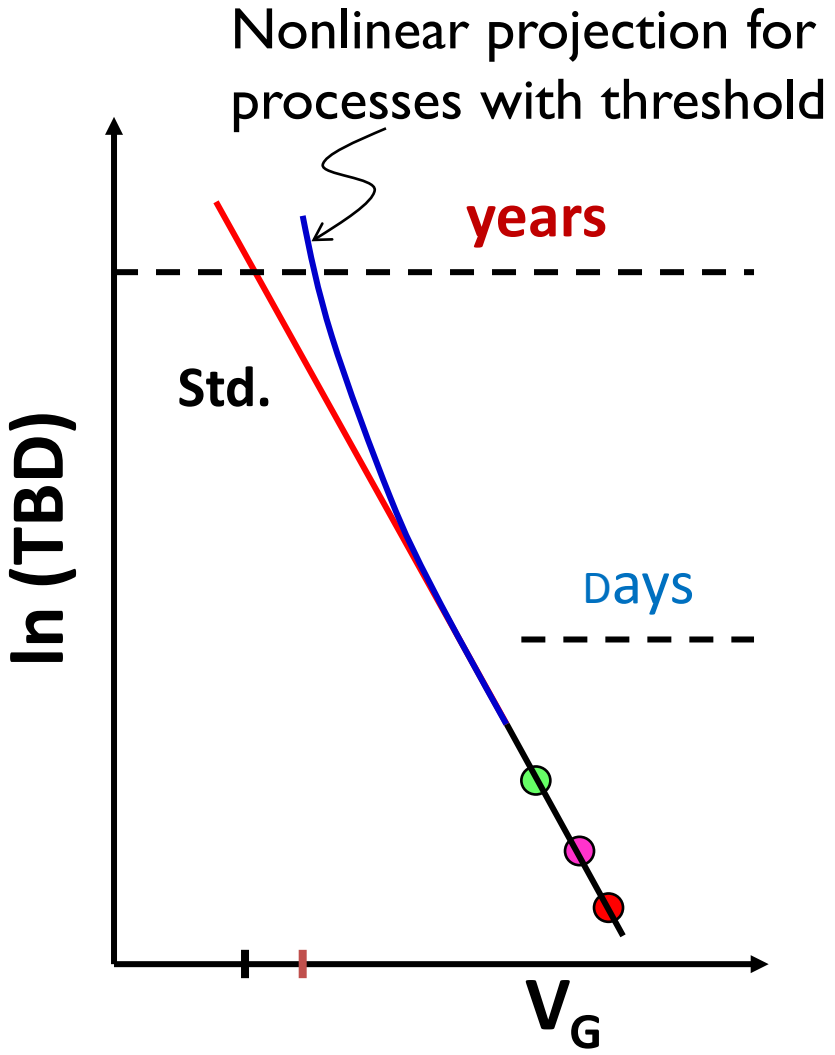


Small changes in the slope can be serious ….

Where do data come from: TDDB Example







Small errors can have serious consequences …
Generation of data is very costly ...

Issues with data

- Small errors can have serious consequences …
- Generation of data is costly in terms of equipment, time, deadlines. Have to maximize information from small dataset.
- Often the dataset may be incomplete, the quality of the data nonuniform, and still we have to make the best decision possible.
- Often there could be competing hypothesis for a given distribution. Have to decide which one fits the data best. Based on the principles of Statistical decision theory.

Outline

1. Origin of data, Field Acceleration vs. Statistical Inference
2. Nonparametric information
3. Preparing data for projection: Hazen formula
4. Preparing data for projection: Kaplan formula
5. Conclusions

Moments of the Experimental Data (or discrete distribution)

Distribution-free statistical measure of data ….

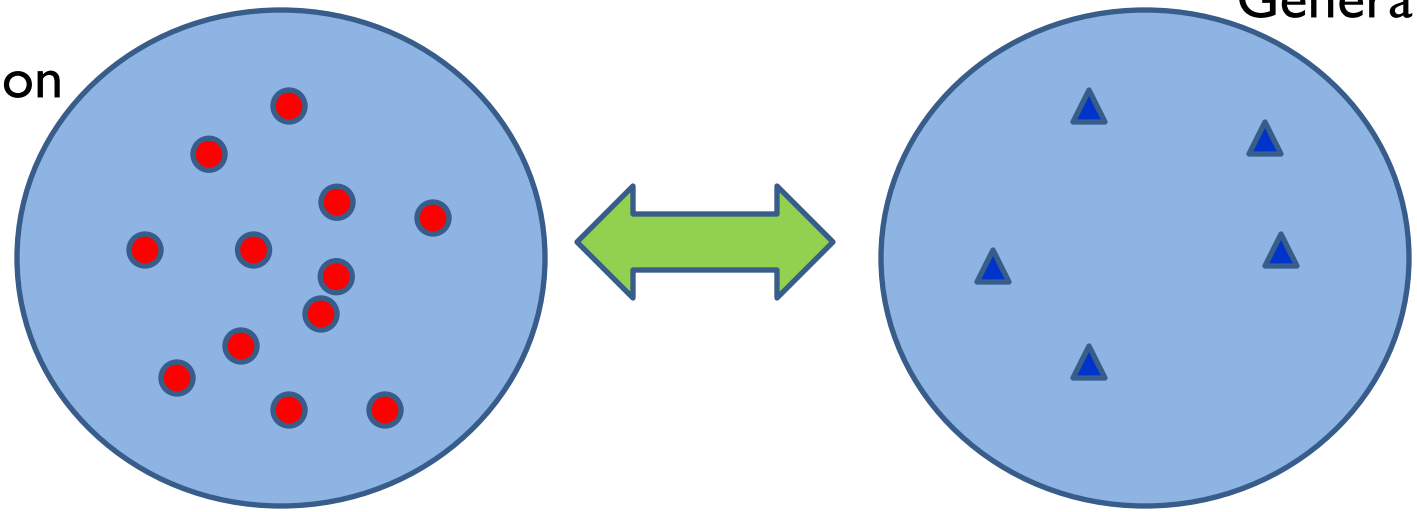
Parameter-space

$$\langle t \rangle = \frac{\sum_{j=1,N} t_j}{N} \qquad s^2 = \frac{\sum_{j=1,N} (t_j - \langle t \rangle)^2}{N-1}$$

$$\delta_{T_K} = \sqrt[k]{\frac{\sum_{j=1}^{N} (t_i - \langle t \rangle)^k}{N-k+1}}$$

General formula

Population



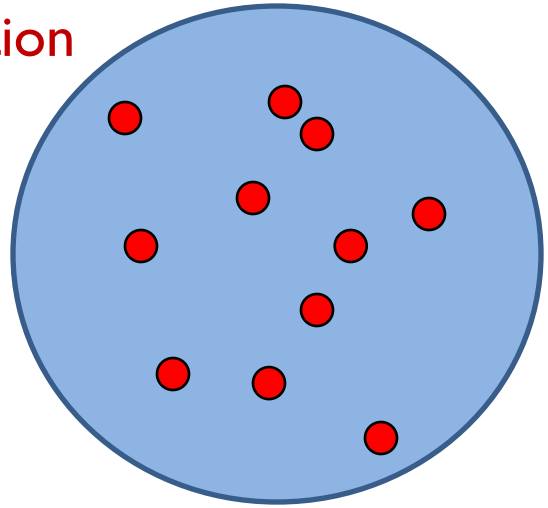
Parameter

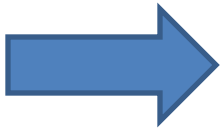
Similar to Fourier Series, First used by Brahe for Alpha Aretis

Good for comparison, but not appropriate for projection

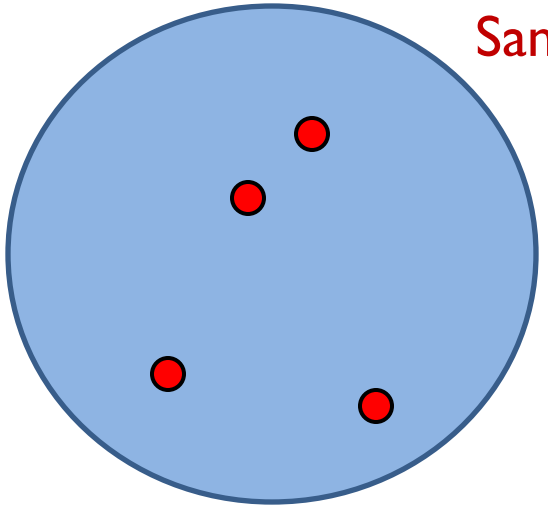
Population vs. Sample Distribution

Population





Sample



$$\langle t \rangle = \frac{\sum_{j=1,N} t_j}{N}$$

$$s^2 = \frac{\sum_{j=1,N} (t_j - \langle t \rangle)^2}{N-1}$$

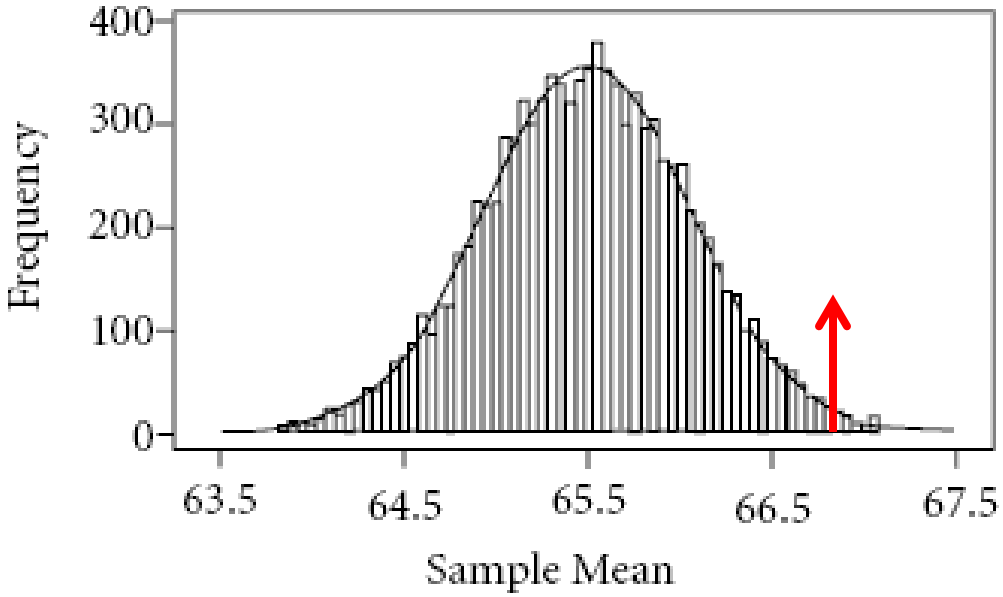
$$\sigma^2 = \frac{\sum_{j=1,N} (t_j - \langle t \rangle)^2}{N}$$

Example Excel routines …

STDEV (2.1, 3.5, 4.5, 5.6) =1.488
STDEVP= (2.1,3.5,4.5,5.6) = 1.2891

Distribution of the Sample Statistic/Moment (e.g. Mean)

Sample Size =20
Number of samples=10k (from population)



Meaning of p-value

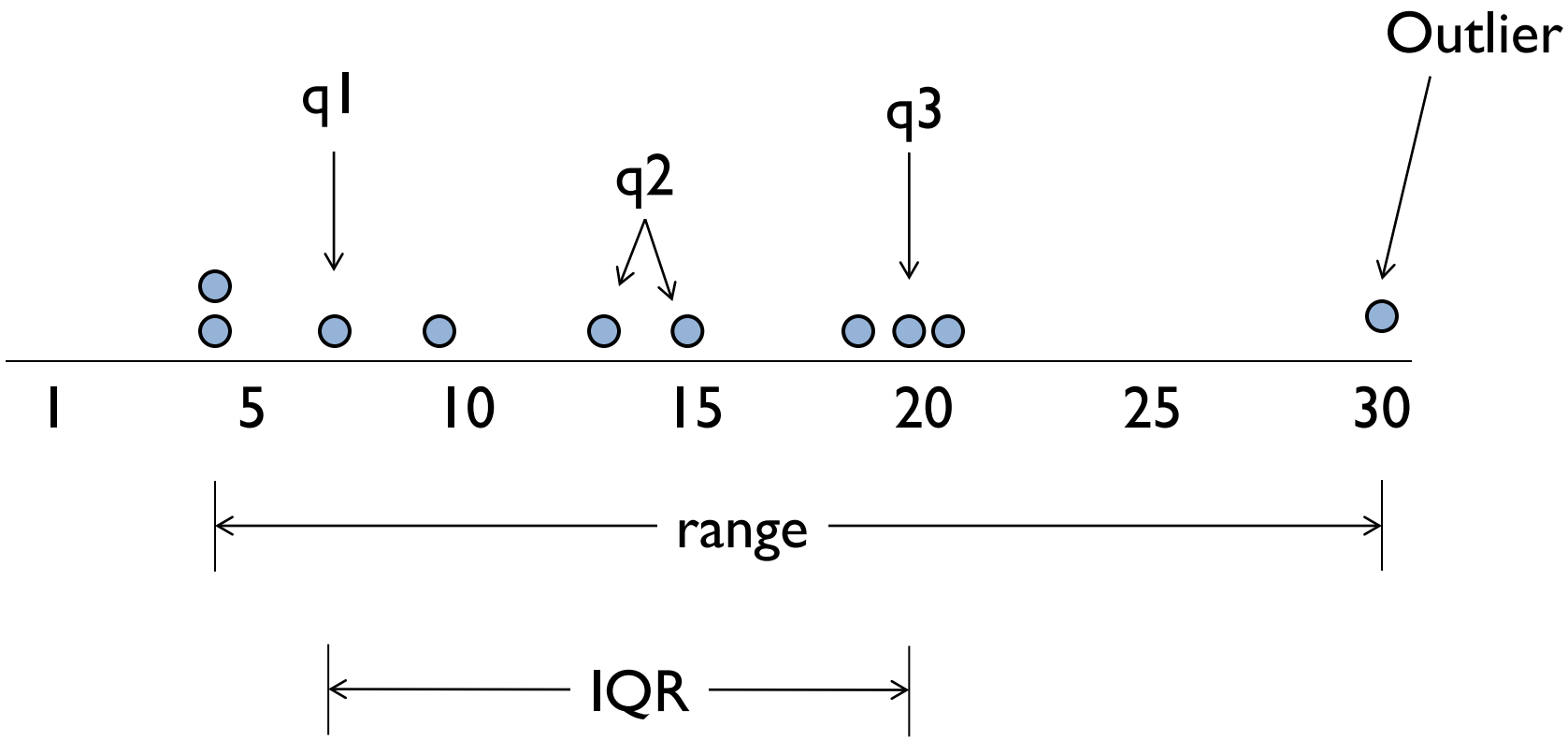
$\mu_x = \mu$

$\sigma_x = \sigma / \sqrt{N}$

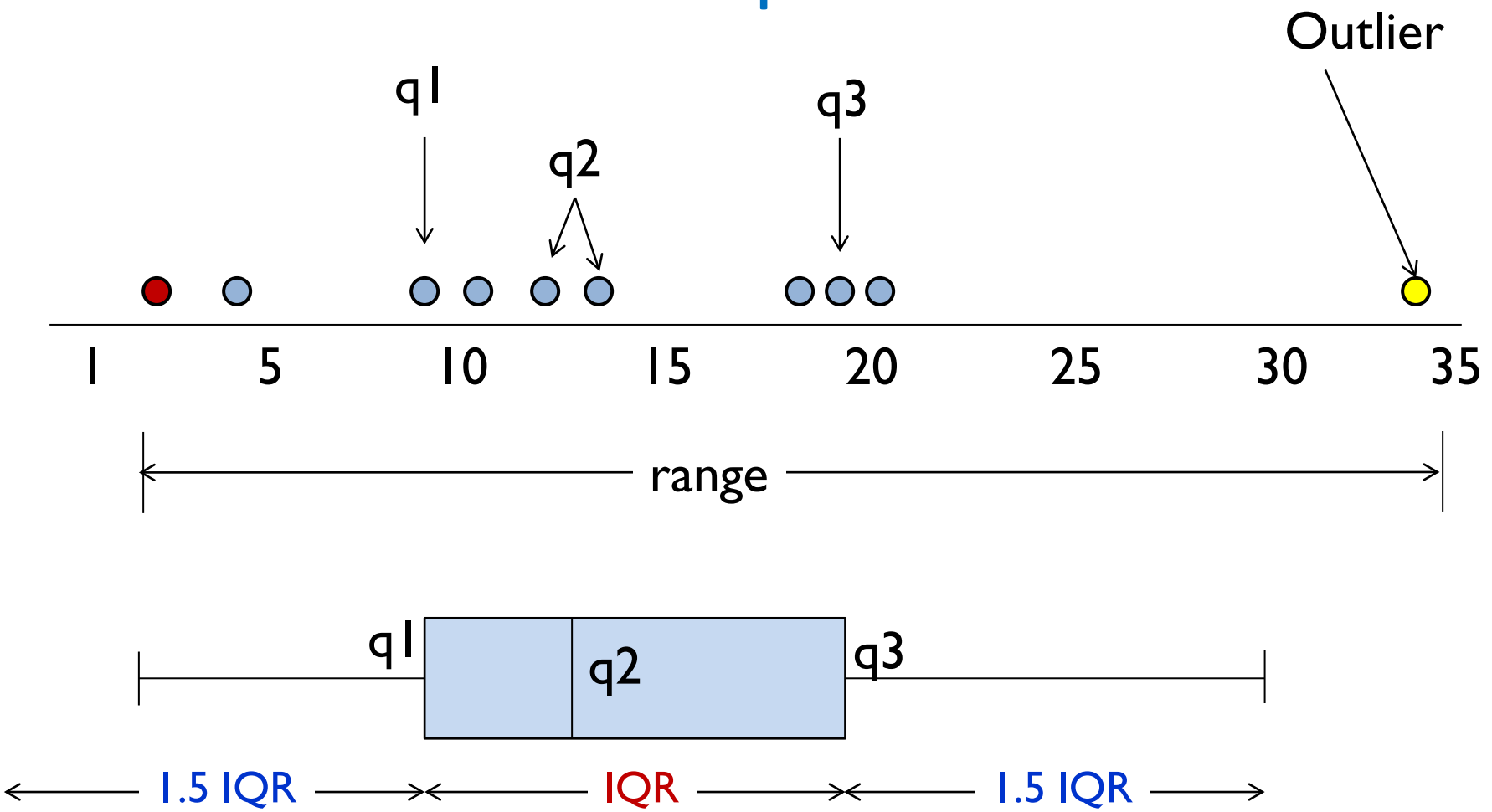
$Z=(X-\mu)/(\sigma/\sqrt{N})$ N> 30

$Z=(X-\mu)/(s/\sqrt{N})$ N < 30

Problem with Sample Moments
Quantiles and robust data description

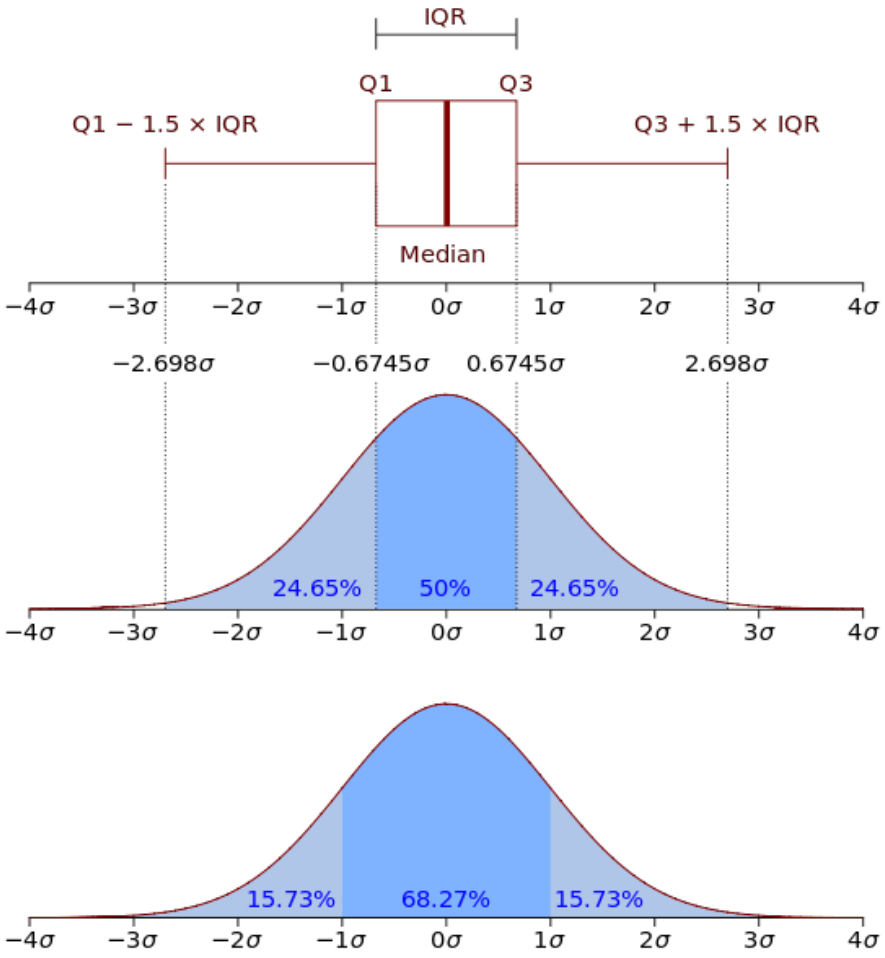


Box plot


Outlier
q1
q3
q2
1
5
10
15
20
25
30
35
range
q1
q2
q3
1.5 IQR
IQR
1.5 IQR

Removing Outliers: The logic of 1.5 IQR in Box plot

Image from Wikipedia



Discrete Data

Continuous distribution

John Tukey, Exploratory Data Analysis, 1977. Adison-Wesley

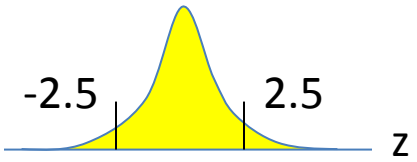
Removing outliers based on Chauvenet's Criteria

TBD (s)
560
540
570
550
560
660 ???
580
570
550

1) Calculate average value of TBD… 571.1 sec.
2) Calculate sample variance s = 35.51 sec.
3) Find normalized variable z= (660-571.1)/35.51=2.5
4) Calculate the tails of Normal distribution with

   p=P(-inf <z < -2.5) + P(inf<z<2.5)= 0.01242
5) Calculate Np = 9*0.01242=0.1
6) If the Np<0.5, throw away the datapoint.

Data Analysis with Excel,
Les Kirkup, Cambridge University Press, p. 185,



Be careful about data rejection – repeat experiment if possible.

Stem and leaf display: Pre-histogram

Order data

44 46 47 49 63 64 66 68 68 72 72 75 76 81 84 88 106

n=17

4 | 4679 ← Leaf
5 |
6 | 34688
7 | 2256
8 | 148
9 |
10 | 6

stem

$L = [10 \times \log_{10} n] \sim 13$

h_n = (Range/L) to power of 10 (i.e. $4.77 \rightarrow 10$)

Therefore, 40, 50, 60 ...90, 100 are stem values

Should use the same approach for histogram
Histogram should not increase precision

Aside: Derivation of Scott's formula for histogram size

Minimize:

$$MSE(x) = \int E[\, f_n(x) - f(x)]^2 \, dx$$

$$h_n = \left\{ \frac{6}{\int_{-\infty}^{\infty} \left[f'(x) \right]^2 dx} \right\}^{1/3} n^{-1/3}$$

$$h_n = 3.49 \times s \times n^{-(1/3)}$$

Freedman/Diaconis-1:

$$h_n = 1.66 \times s \times \left(\frac{\ln(n)}{n} \right)^{1/3}$$

Freedman/Diaconis-2:

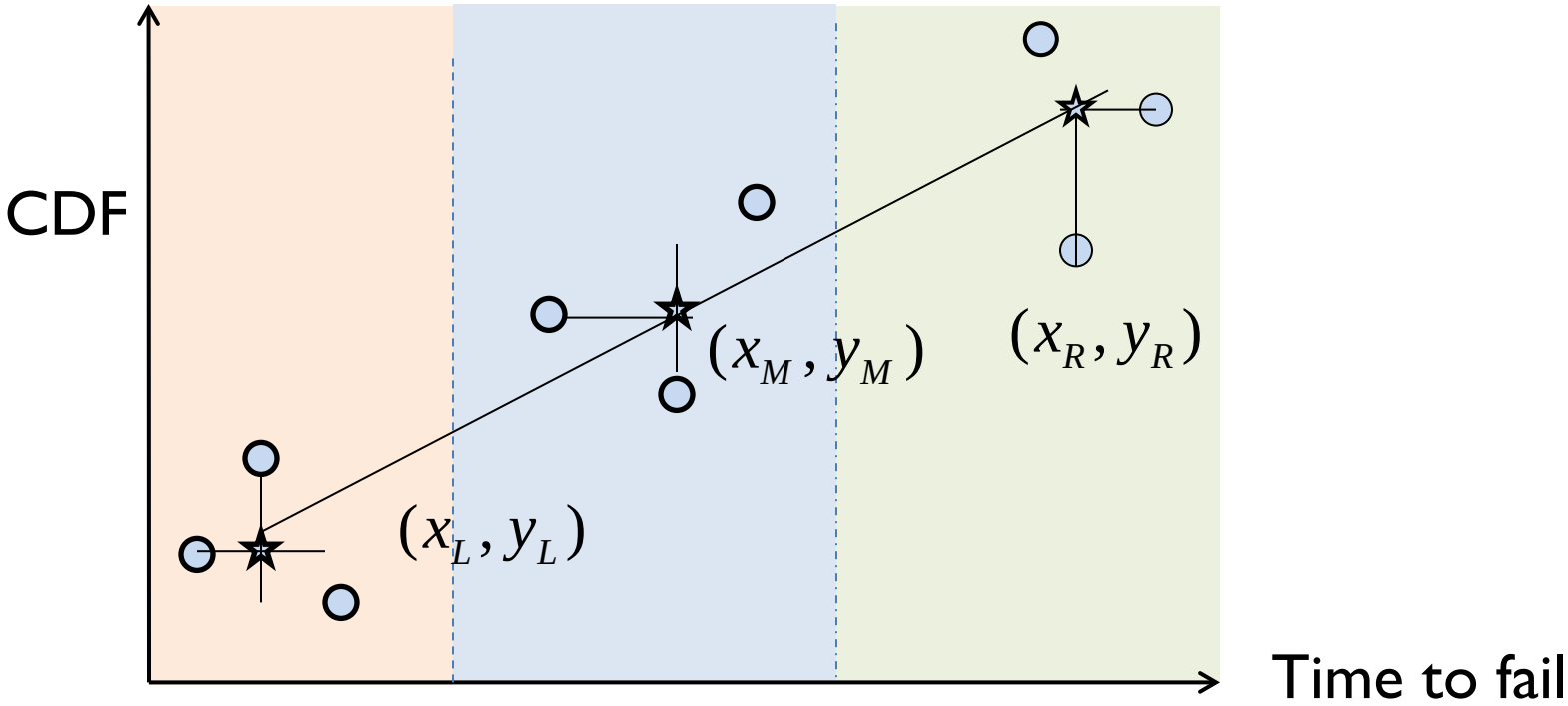
$$h_n = 2(IQR) \left(\frac{1}{n} \right)^{1/3}$$

Scott:

$$h_n = 3.49 \times s \times n^{-(1/3)}$$

Choose any of these formula, but remain consistent

Drawing lines resistant to outliers



Divide the data into three groups, i.e.

For n=3k (k, k, and k)

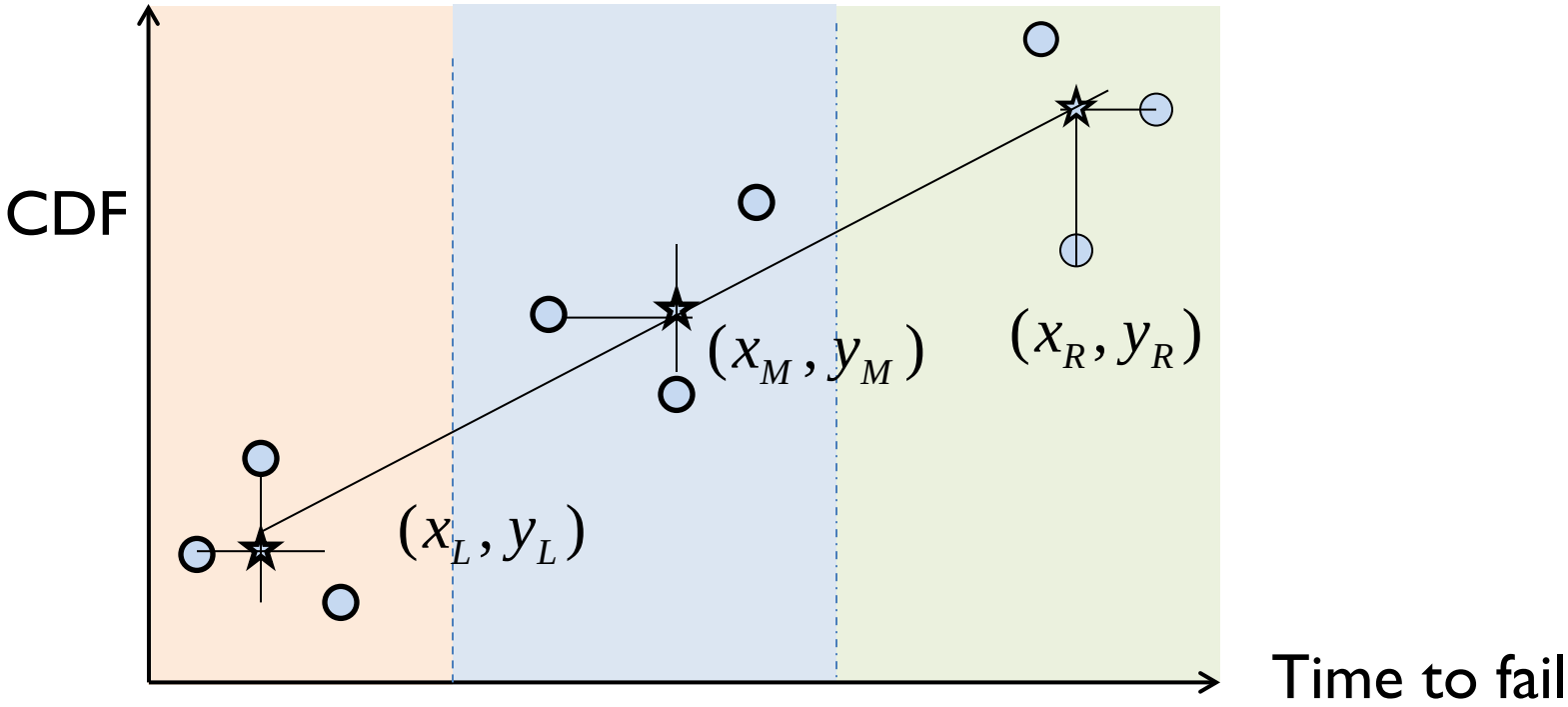
For n=3k+1 (k, k+1, k)

For n=3k+2 (k+1, k, k+1)

Calculate the median (x, y) of each group.

Draw the line.

Drawing lines resistant to outliers



$$y = b(x - x_M) + a$$

$$b_0 = (y_R - y_L)/(x_R - x_L)$$

$$3a_0 = [y_L - b_0(x_L - x_M)] + y_M + [y_R - b_0(x_R - x_M)]$$

$$r_i = y_i - [a_0 + b_0(x_i - x_0)]$$

$$a_1 = a_0 + \gamma_1 \quad b_1 = b_0 + \delta_1$$

Outline

1. Origin of data, Field Acceleration vs. Statistical Inference
2. Nonparametric information
3. Preparing data for projection: Hazen formula
4. Preparing data for projection: Kaplan formula
5. Conclusions

Problem of data plotting and numerical CDF

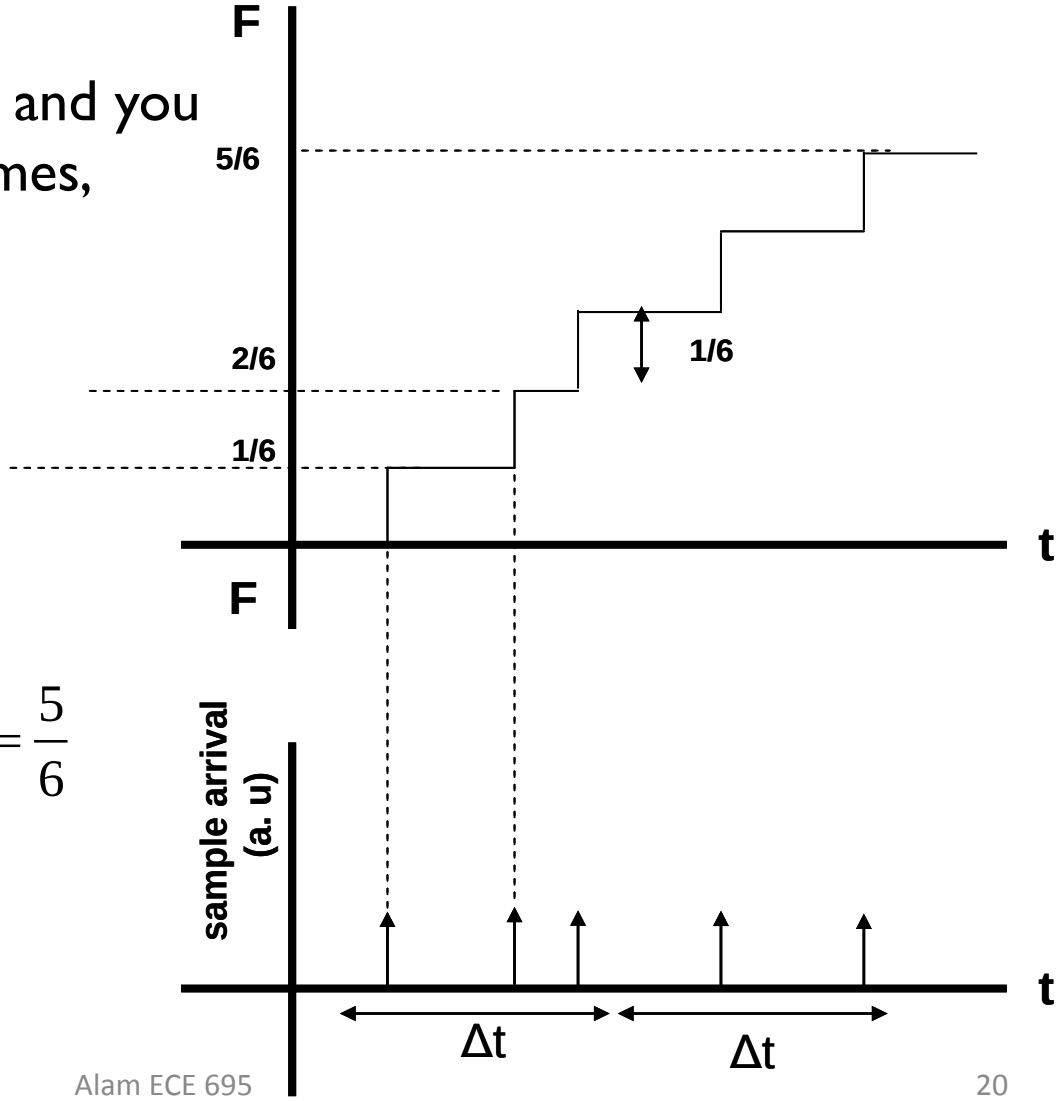
Assume you have 5 transistors and you have collected 5 breakdown times, t_1, t_2, t_3, t_4, t_5

How do we find the CDF?

$$F_i = \frac{i}{n} \text{ or } F_i = \frac{i}{n+1}?$$

$$F_1 = \frac{1}{6} \quad F_2 = \frac{2}{6} \quad F_3 = \frac{3}{6} \quad F_4 = \frac{4}{6} \quad F_5 = \frac{5}{6}$$

$$W = \ln(-\ln(1-F_i))$$



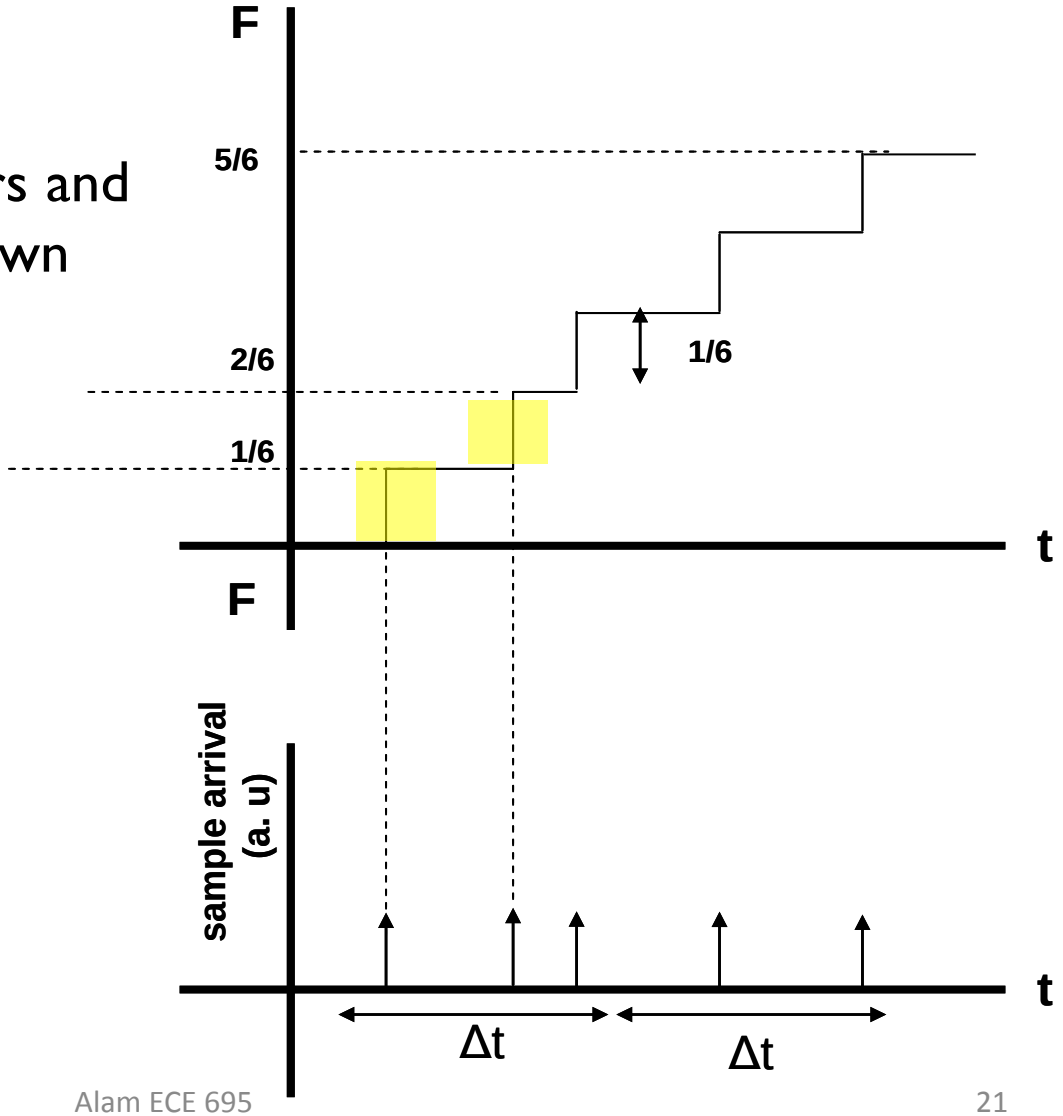
… there is a problem (Failure time is statistical)

Assume you have 5 transistors and you have collected 5 breakdown times, t_1, t_2, t_3, t_4, t_5

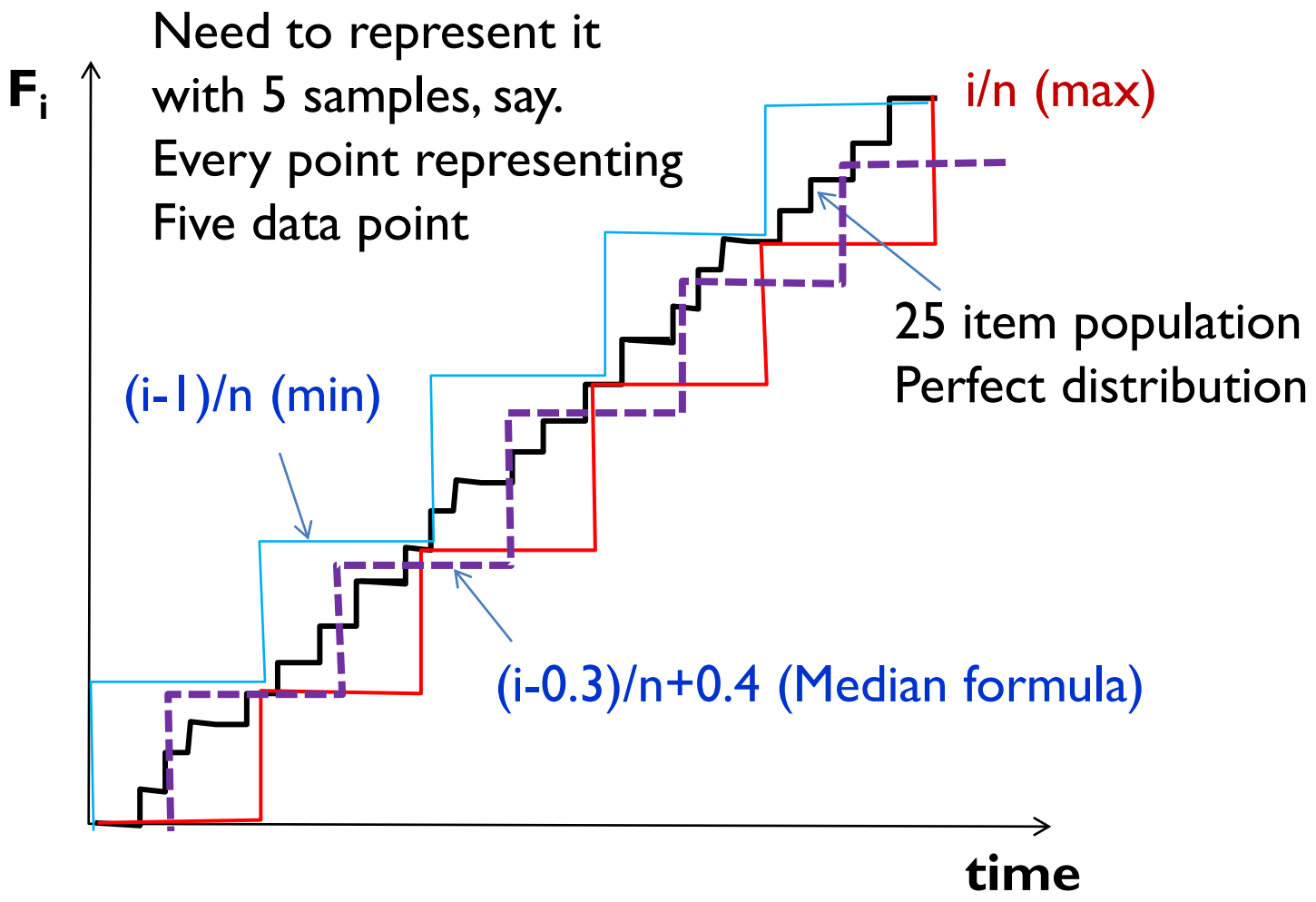
How do we find the CDF?

$$F_i = \frac{i - \alpha}{n - 2\alpha + 1}$$

$$W = \ln(-\ln(1 - F_i))$$



Relationship among various formula



Analogous to a congressman …

Aside: Derivation of Hazen Formula

$$F_i = \frac{i-\alpha}{n-2\alpha+1}$$

p = Probable CDF location,

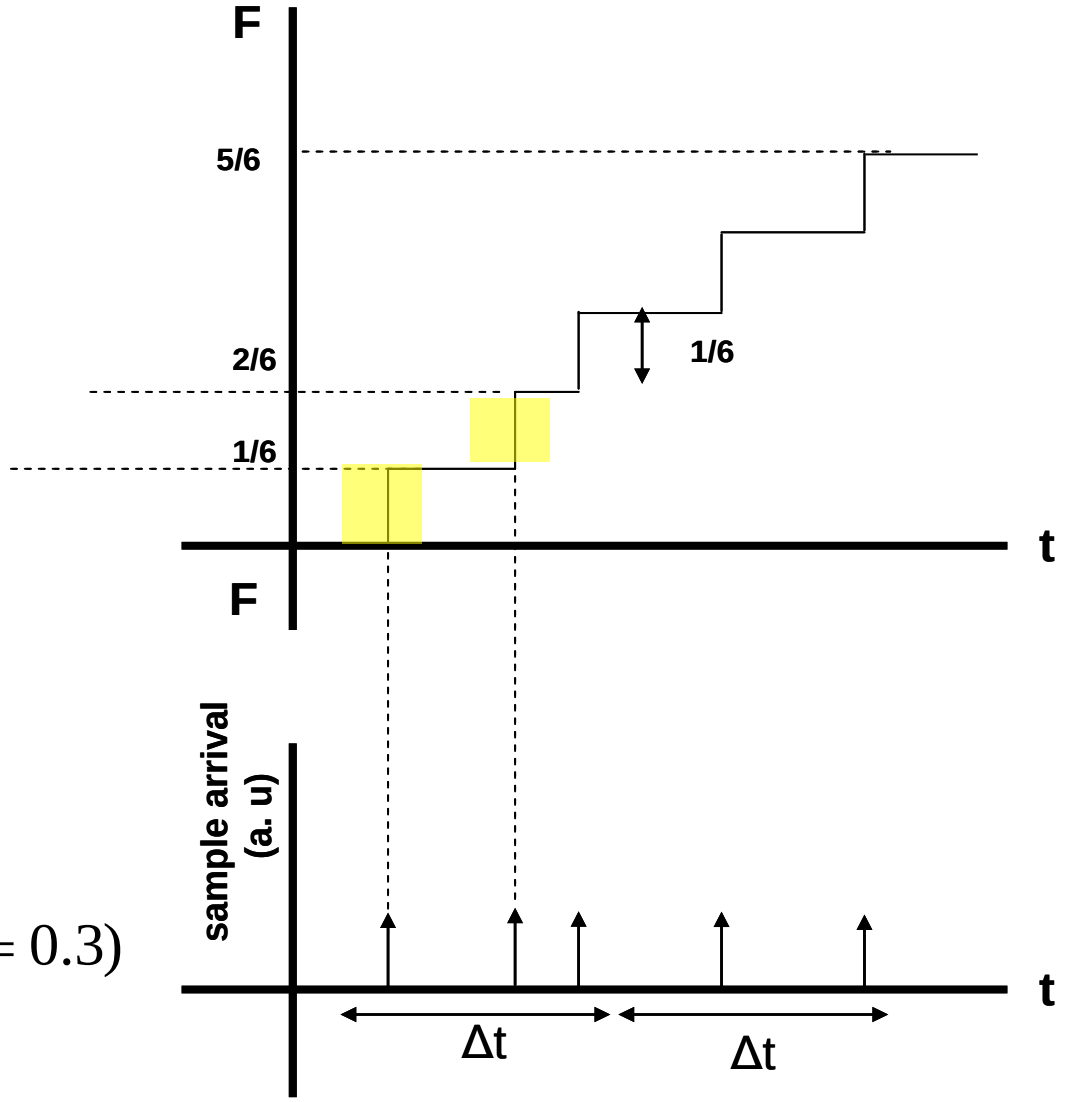
F_i, of the i-th data

$$G = \binom{n}{i} p^i (1-p)^{n-i}$$

$$g = \frac{dG}{dp} = i\binom{n}{i} p^{i-1}(1-p)^{n-i}$$

$$\int_0^{F_{Median,i}} g(p)dp = 1/2$$

$$\Rightarrow F_{Median,i} = \frac{i-\alpha}{n-2\alpha+1} \quad (\alpha = 0.3)$$

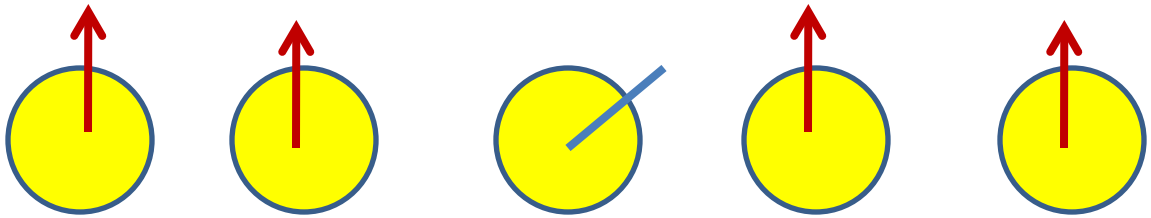


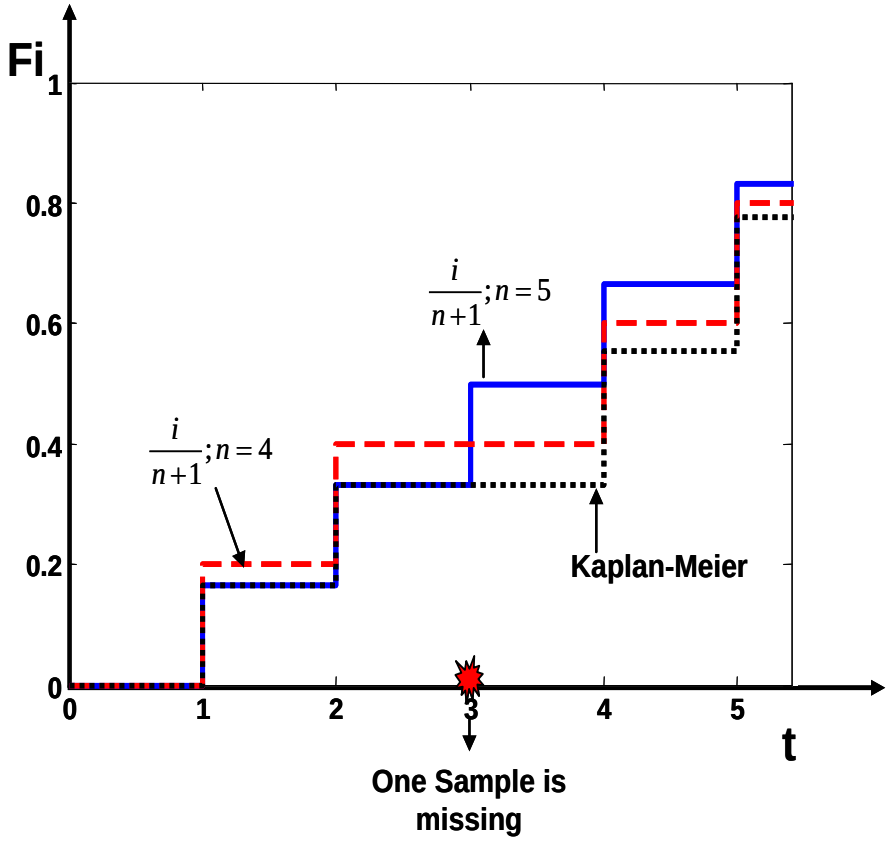
Ref. Ross, 1999 Conf. on Elec. Insulation/Dielectric Phenomena, p. 170; Ohring, p. 223

Outline

1. Origin of data, Field Acceleration vs. Statistical Inference
2. Nonparametric information
3. Preparing data for projection: Hazen formula
4. Preparing data for projection: Kaplan formula
5. Conclusions

Censored data and imperfect sampling





$$F_i = \frac{i-\alpha}{n-2\alpha+1} \qquad F_i = \frac{i}{n+1}$$

$$F_1 = \frac{1}{6} \quad F_2 = \frac{2}{6} \quad F_3 = \frac{3}{6} \quad F_4 = \frac{4}{6} \quad F_5 = \frac{5}{6}$$

With 4 data points now, most people would do …..

$$F_1 = \frac{1}{5} \qquad F_2 = \frac{2}{5} \qquad F_3^* = \frac{3}{5} \qquad F_4* = \frac{4}{5}$$

… but this would be wrong!

Hazen (approximate) formula for censored data

$$F_1 = \frac{1}{5}$$

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

$$F_2 = \frac{2}{5}$$

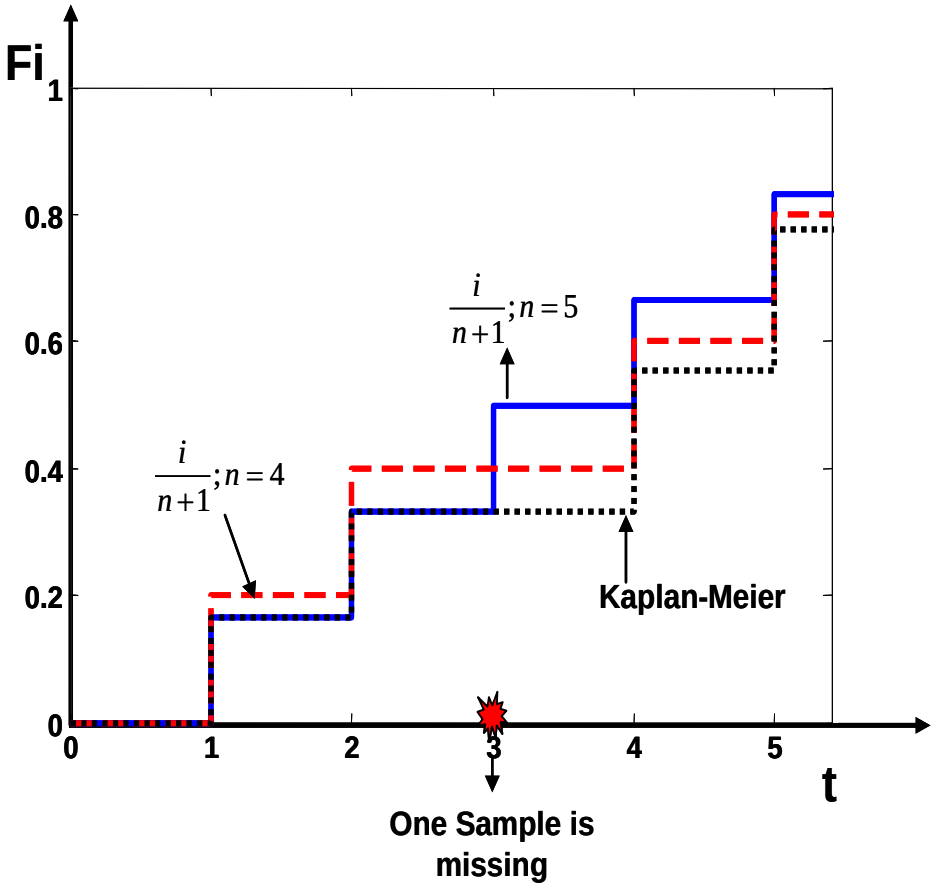
$$F_3 = \frac{2}{5}$$

$$F_4 = \frac{3}{5}$$

$$F_5 = \frac{4}{5}$$

Loss of a sample N=4

but the sample did survive till t2 …



5 data-points, same as before, but with effective reduction in sample size

Past data affected by future problems … does not seem correct

Kaplan-Meier (proper) Formula

$$F_i = 1 - \left(\frac{n - \alpha + 1}{n - 2\alpha + 1} \right) \prod_{i=1}^{f} \left(\frac{n_{si} + 1 - \alpha}{n_{si} + 2 - \alpha} \right)$$

Total number of samples

Number of surviving samples after time t_i

Assume α=0, so that

$$F_i = 1 - \prod_{i=1}^{f} \left(\frac{n_{si} + 1}{n_{si} + 2} \right)$$

For uncensored traditional data …

$$F_i = 1 - \prod_{i=1}^{f} \left(\frac{n_{si} + 1}{n_{si} + 2} \right)$$

$$F_1 = 1 - \frac{5}{6} = \frac{1}{6}$$

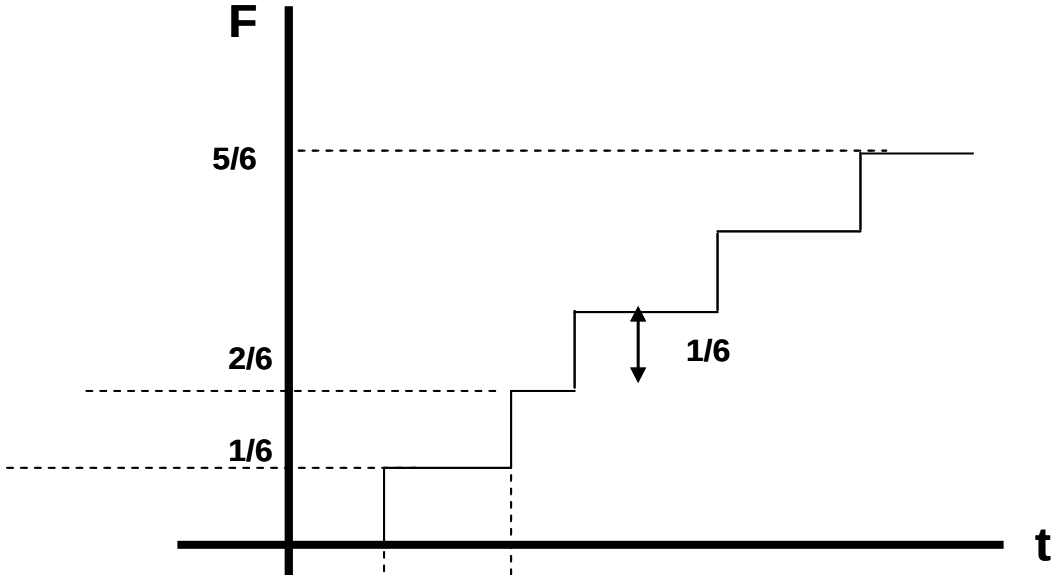
$$F_2 = 1 - \left(\frac{5}{6}\right)\left(\frac{4}{5}\right) = \frac{2}{6}$$

$$F_3 = 1 - \frac{5}{6}\frac{4}{5}\frac{3}{4} = \frac{3}{6}$$

$$F_4 = 1 - \frac{5}{6}\frac{4}{5}\frac{3}{4}\frac{2}{3} = \frac{4}{6}$$

$$F_5 = 1 - \frac{5}{6}\frac{4}{5}\frac{3}{4}\frac{2}{3}\frac{1}{2} = \frac{5}{6}$$

n_{si} before t_i	5	4	3	2	1
n_{si} after t_i	4	3	2	1	0



Same as before …

For censored data

Assume that at time t_3, one sample is taken out of the experiments (censored)

n_{si} before t_i	5	4	3	2	1
n_{si} after t_i	4	3	2	1	0

$$F_1 = 1 - \frac{4+1}{4+2} = \frac{1}{6}$$

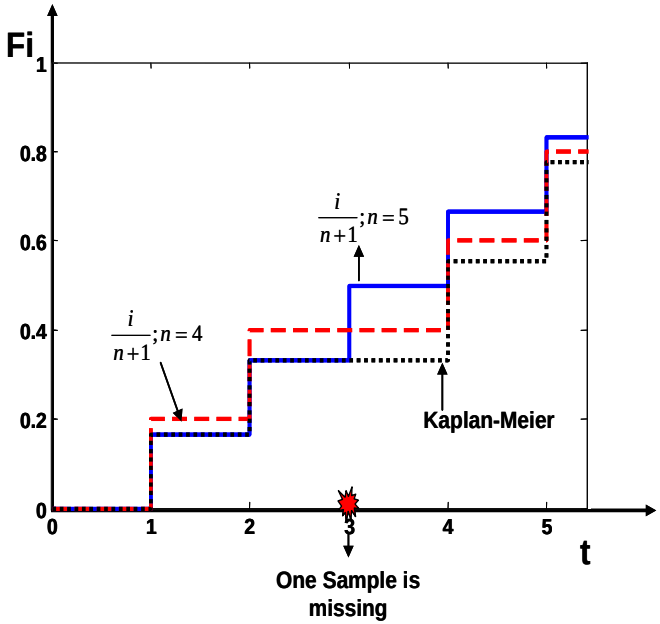
$$F_2 = 1 - \frac{4+1}{4+2}\frac{3+1}{3+2} = \frac{2}{6}$$

$$F_3 = 1 - \frac{4+1}{4+2}\frac{3+1}{3+2} = \frac{2}{6}$$

$$F_4 = 1 - \frac{4+1}{4+2}\frac{3+1}{3+2}\frac{1+1}{1+2} = 1 - \frac{5}{6}\frac{4}{5}\bullet\frac{2}{3} = \frac{5}{9}$$

$$F_5 = 1 - \frac{5}{6}\frac{4}{5}\frac{2}{3}\frac{1}{2} = \frac{7}{9}$$

¾ missing …



Summary

Method	T=1	T=2	T=3	T=4	T=5
Hazen's Formula	1/6	2/6	3/6	4/6	5/6
Hazen's Formula with one sample missing	1/5	2/5	Missing Sample	3/5	4/5
Kaplan-Meier Method	1/6	2/6	Missing Sample	5/9	7/9

Larger sample

Assume time for oxide breakdown… (or test for cancer drug)

6,6,6,6*,7,9*,10,10*,11*,13,16,17*,19*, 20*,22,23,25*,32*,34*,35*
(* data stopped)

$F_1,F_2,F_3,S_1,F_4,S_2,F_5,S_3,S_4,F_6,F_7,S_5,S_6,\ldots.$

Samples before
21 20 19 18 17 ….. 12

Samples after (n_s)
20 19 18 17 16 …. 11

R_1=20/21,
R_2=19/20*(20/21)
R_3=0.857
R_4=R_3*(16/17)=0.8067 ..
R_5=R_4*(14/15)=0.753

$$F_i = 1 - \prod_{i=1}^{f} \left(\frac{n_{si}+1}{n_{si}+2} \right) = 1 - R_i$$

Dramatically different plot ….

Ref. L. Wolstenhome, Reliability Modeling

Conclusions

1. Treat your data with respect! They have stories to tell. A photon on your window may have the memory of a galaxy.

2. Focus on non-parametric data analysis. Simple non-parametric estimates like mean, standard deviation, median are all useful indicators that helps selecting appropriate distribution functions.

3. Non parametric plotting of distribution function is very important. Censored and uncensored data have very different plotting approaches. Outliers distort, therefore, median-based techniques is often useful.

References

- D. C. Hoaglen, F. Mosteller, and J.W. Tukey, "Understanding Robust and Exploratory Data Analysis", Wiley Interscience, 1983. Explains the importance of Median based analysis when the dataset is small and the quality cannot be guaranteed.
- Linda C. Wolsterholme, "Reliability Modeling – A Statistical Approach, Chapman Hall, CRC, 1999. Chapter 1-7 has excellent summary of 'Goodness of Fit" analysis.
- R. H. Myers and D.C. Montgomery, "Response Surface Methodology", Wiley Interscience, 2002. This book discusses design of experiment in great detail.
- An excellent textbook that covers many topics discussed in this Lectures is Applied Statstics and Proability for Engineers, 3rd Edision, D.C. Montgomery and G. C. Runger, Wiley, 2003.
- J. Stuart Hunter had a Television Series on Statistics and some of the lectures are now posted at Nanohub. You can search under his name or get started by following link http://www.youtube.com/watch?v=AVUAt0Qly60

Review Questions

1. What is the difference between parametric estimation vs. non-parametric estimation?
2. What principle did Tacho Brahe's approach assume?
3. What is the difference between population and sample? When we collect data for TDDB or NBTI, what type of data are we collecting?
4. What problem does Hazen formula avoid regarding Fi=i/n? How is this justified?
5. What is the problem of Hazen formula with respect to censored data?
6. Can you think of a situation where data censoring may be necessary for NBTI test?
7. What is the difference between an outlier vs. a censored data?
8. Do I need to know what the physical distribution is before using Hazen or Kaplan formula? Why or why not?