Primer on Analysis of Experimental Data and Design of Experiments

Lecture 2. Collecting and Plotting Data

Muhammad A. Alam
alam@purdue.edu

Alam CEED Group

Outline

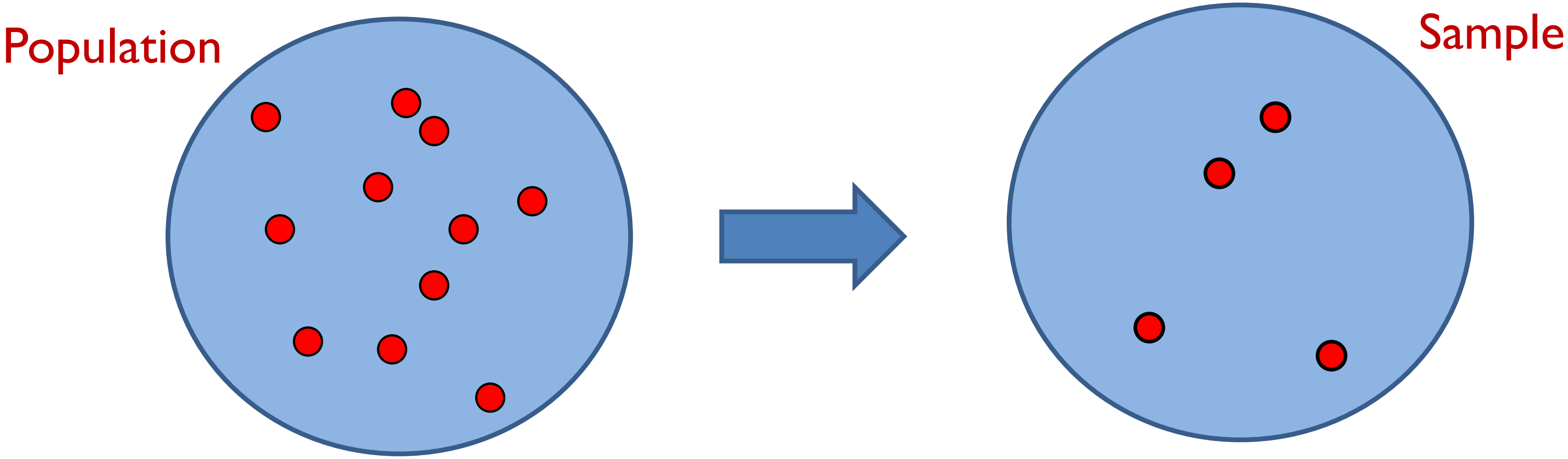
1. Sample vs. population: A Review of traditional statistics

2. Trouble with traditional statistics

   - If the population is not described by Gaussian distribution
   - If some of the datapoints are outliers
   - If some of some experiments ended early

3. Conclusions

Population vs. Sample Distribution



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

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

Moments of the Experimental Data (or discrete distribution)

Distribution-free statistical measure of data ….

Parameter-space

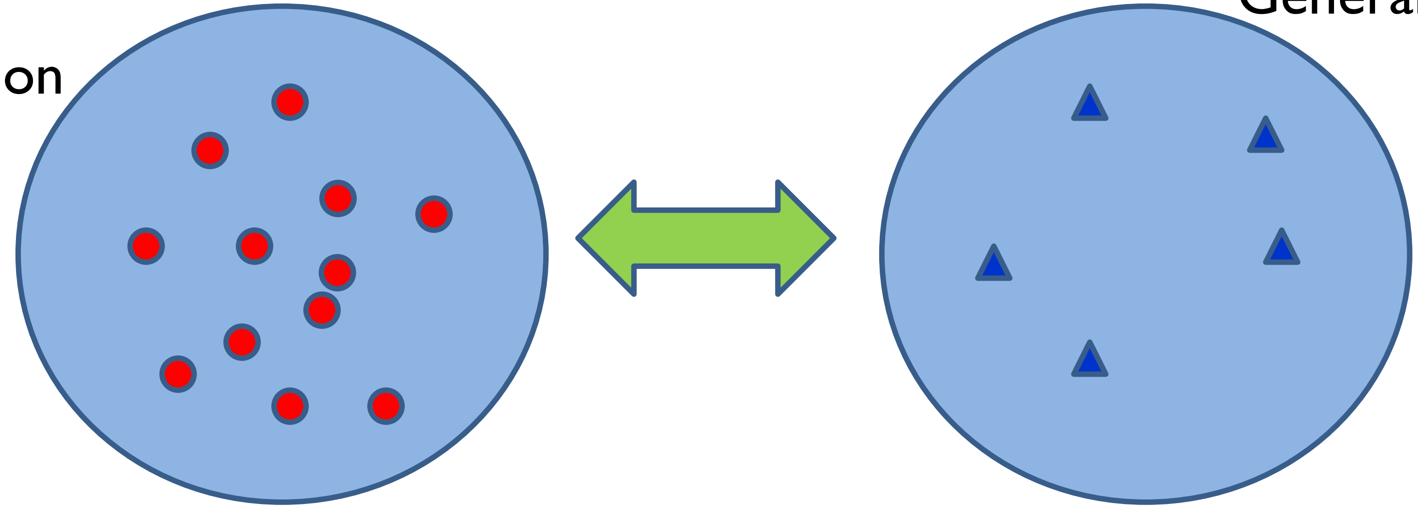
$$\langle \boldsymbol{t} \rangle = \frac{\sum_{j=1,N} \boldsymbol{t}_j}{\boldsymbol{N}}$$

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

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

General formula

Population

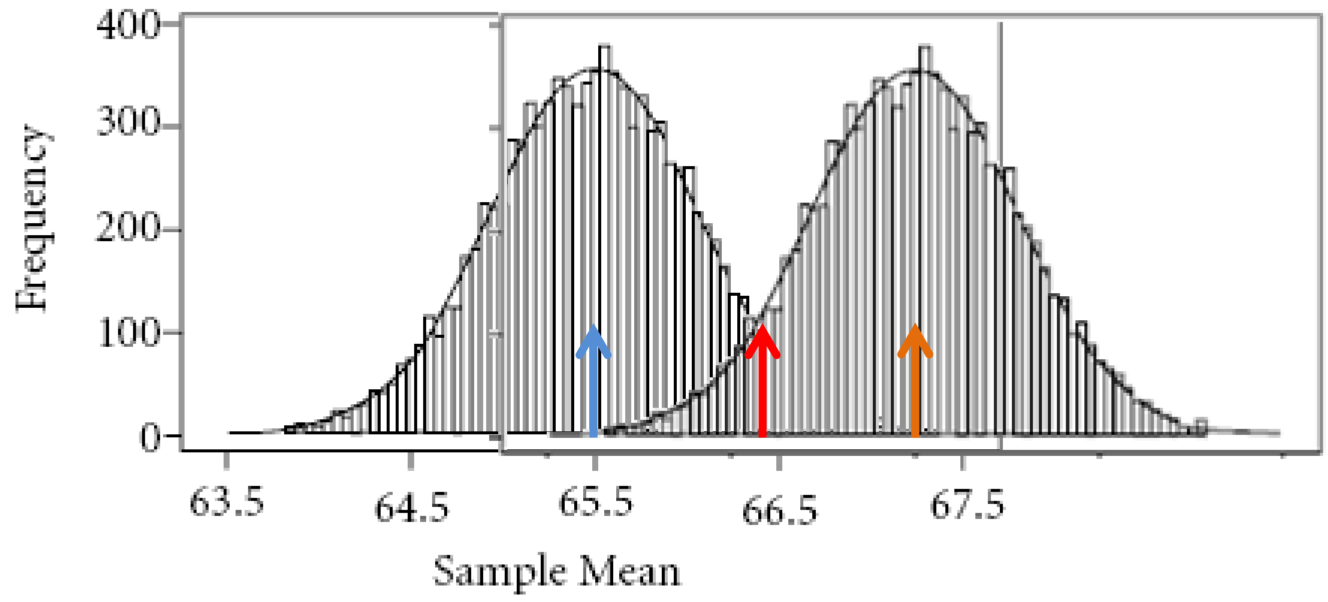


Parameter

Similar to Fourier Series

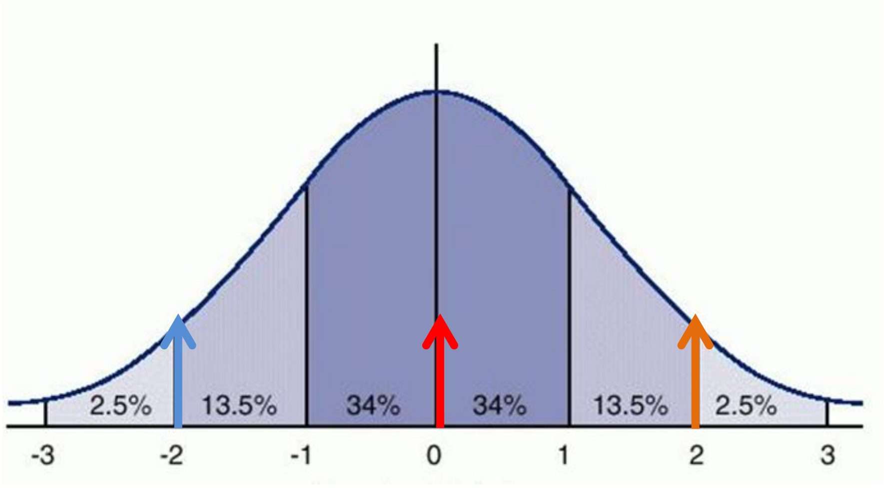
Distribution of the Sample Statistic/Moment (e.g. Mean)

Sample Size =20, Population size=10k



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

Meaning of p-value



Z

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

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

Outline

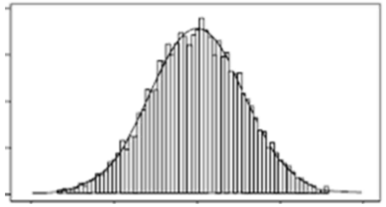
1. Sample vs. population: A Review of traditional statistics

2. Trouble with traditional statistics

   - If the population is not described by Gaussian distribution
   - If some of the datapoints are outliers
   - If some of some experiments ended early

3. Conclusions

Bootstrap: standard deviation if the distribution is unknown



0.2 -0.1 0.5 0.3 -0.6

All you have is a single sample ..

Generate synthetic samples from the original (with replacement)

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

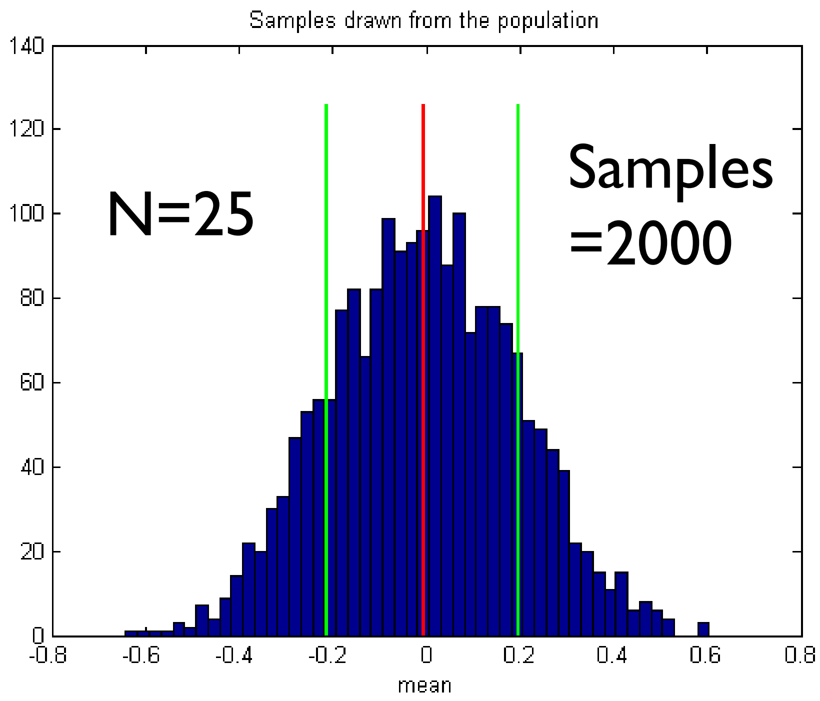
0.2 -0.1 -0.6 -0.1 0.5 Synthetic sample 1

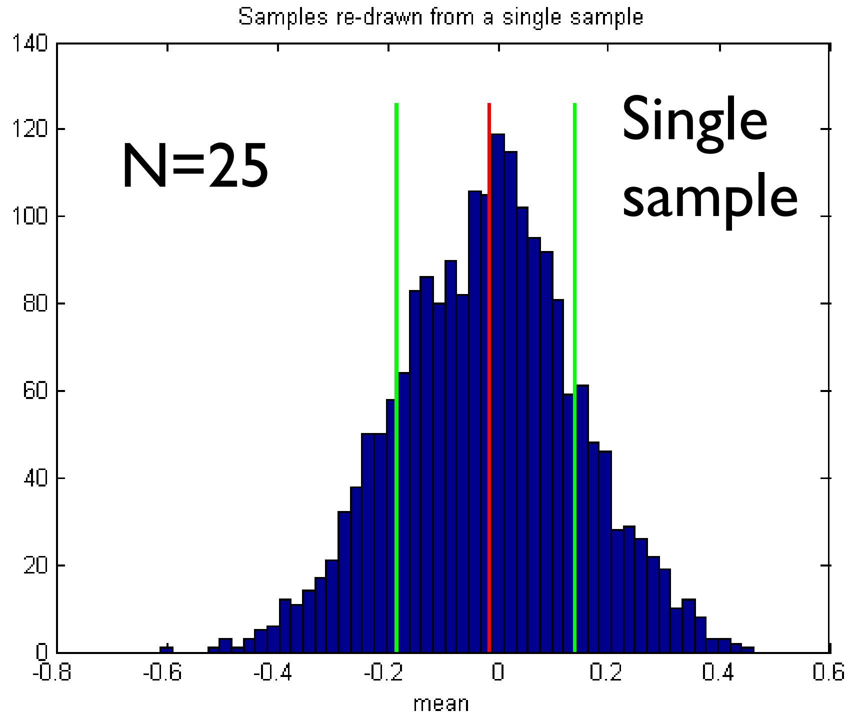
0.3 0.2 -0.6 0.2 0.5 Synthetic sample 2

0.5 -0.1 0.5 0.2 0.3 Synthetic sample 3

```
m = bootstrp(100,@mean,y);
figure; [fi,xi] = ksdensity(m); plot(xi,fi);
```

Multiple sample vs. single sample





The mean-distribution from the population and that from a single samples are essentially identical

Outline

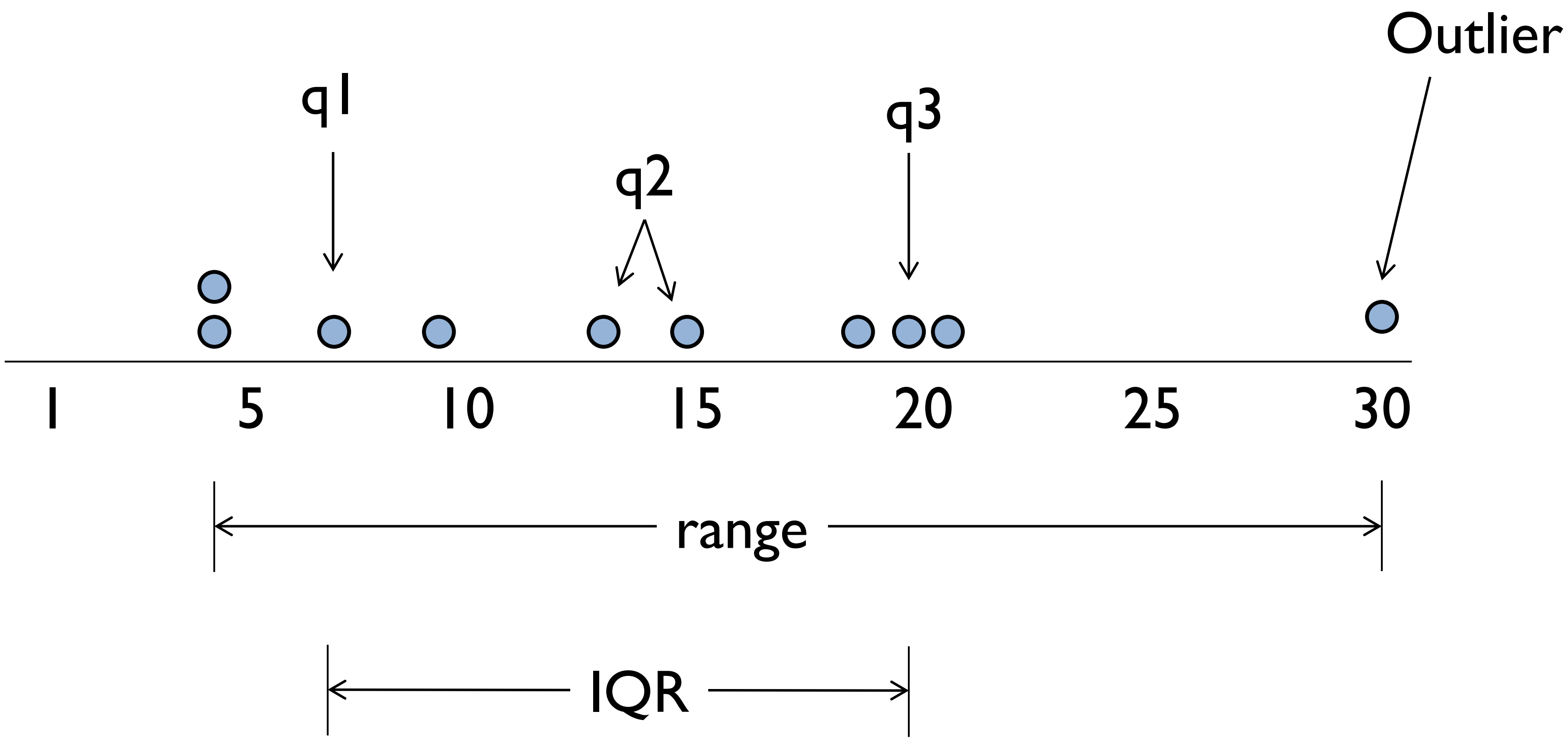
1. Sample vs. population: A Review of traditional statistics

2. Trouble with traditional statistics

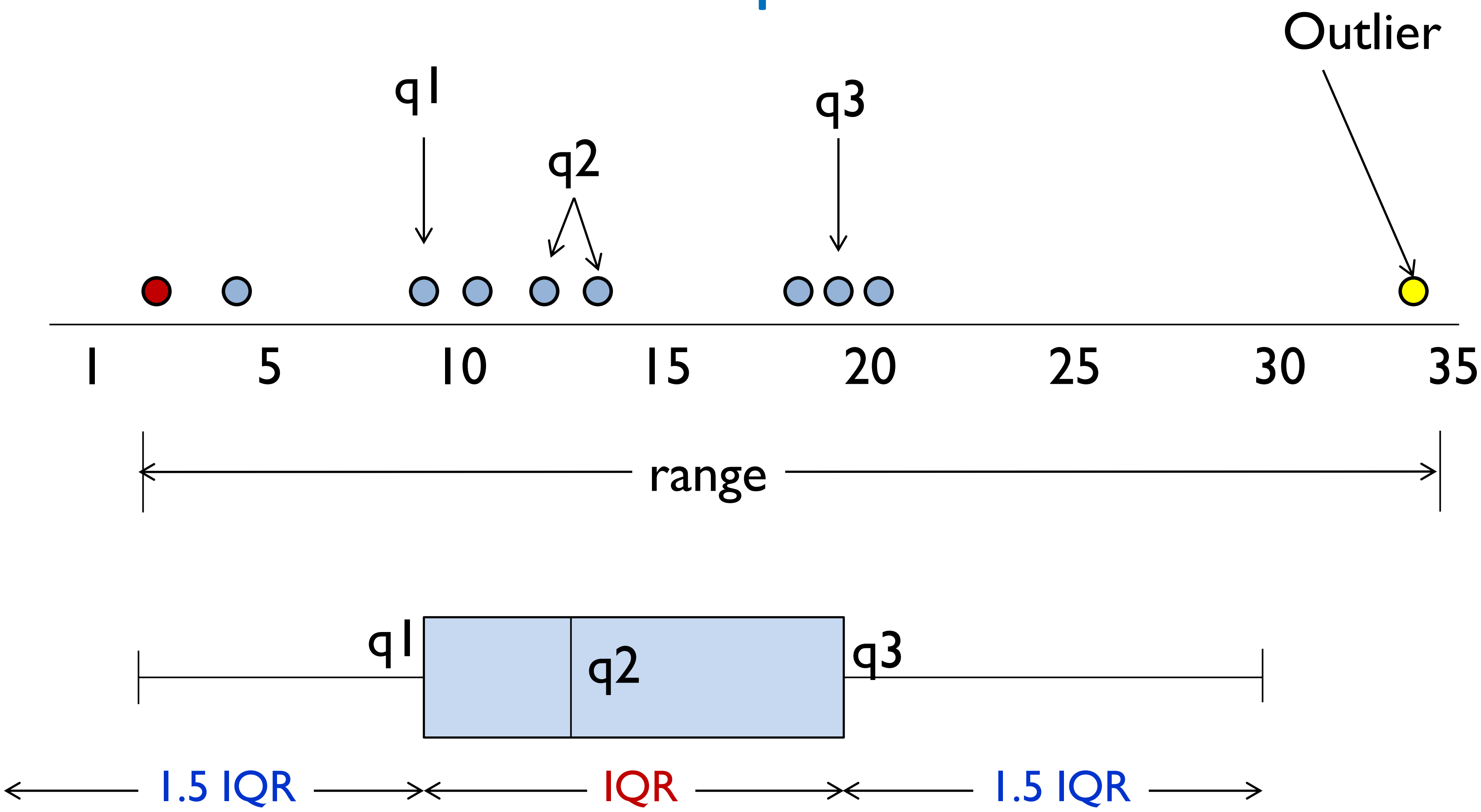
    - If the population is not described by Gaussian distribution
    - If some of the datapoints are outliers
    - If some of some experiments ended early

3. Conclusions

Problem with Sample Moments
Quartiles and robust data description



Box plot



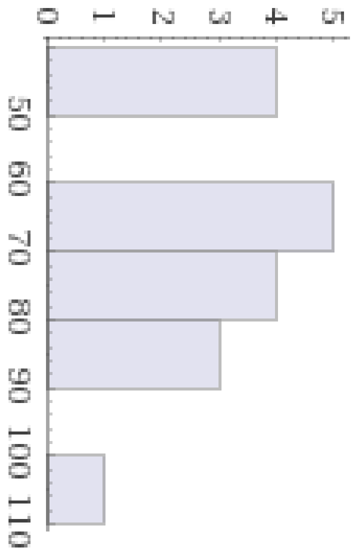
Outlier
q1
q2
q3
1
5
10
15
20
25
30
35
range
q1
q2
q3
1.5 IQR
IQR
1.5 IQR

Stem and leaf display: Pre-histogram

Order data

n=17

44 46 47 49 63 64 66 68 68 72 72 75 76 81 84 88 103



4 | 4679 ← Leaf
5 |
6 | 34688
7 | 2256
8 | 148
9 |
10 | 3
↑
stem

$$L = 10 \times \log_{10} n = 10 \times \log_{10} 17 = 12.3 \sim 13$$

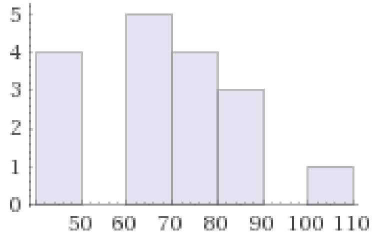
$$h_n = \left(\frac{Range}{L}\right) = \frac{103 - 44}{13} = 4.53$$

~ 10 rounded to 10 power

i.e. 40, 50, … 90, 100

Should use the same approach for histogram

Histogram should not increase precision



side: Derivation of 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:

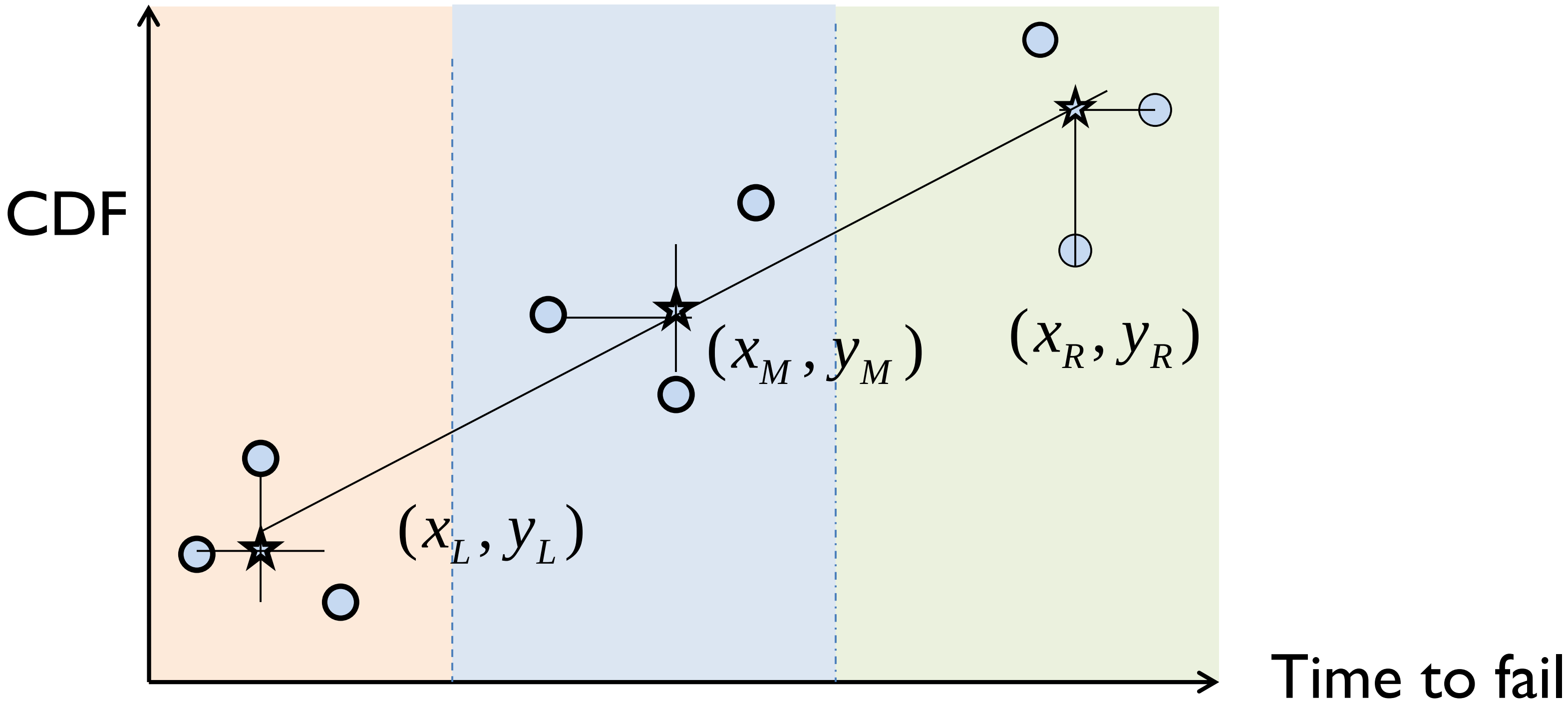
$$h_n = 2(IQR)\left(\frac{1}{n} \right)^{1/3}$$

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)

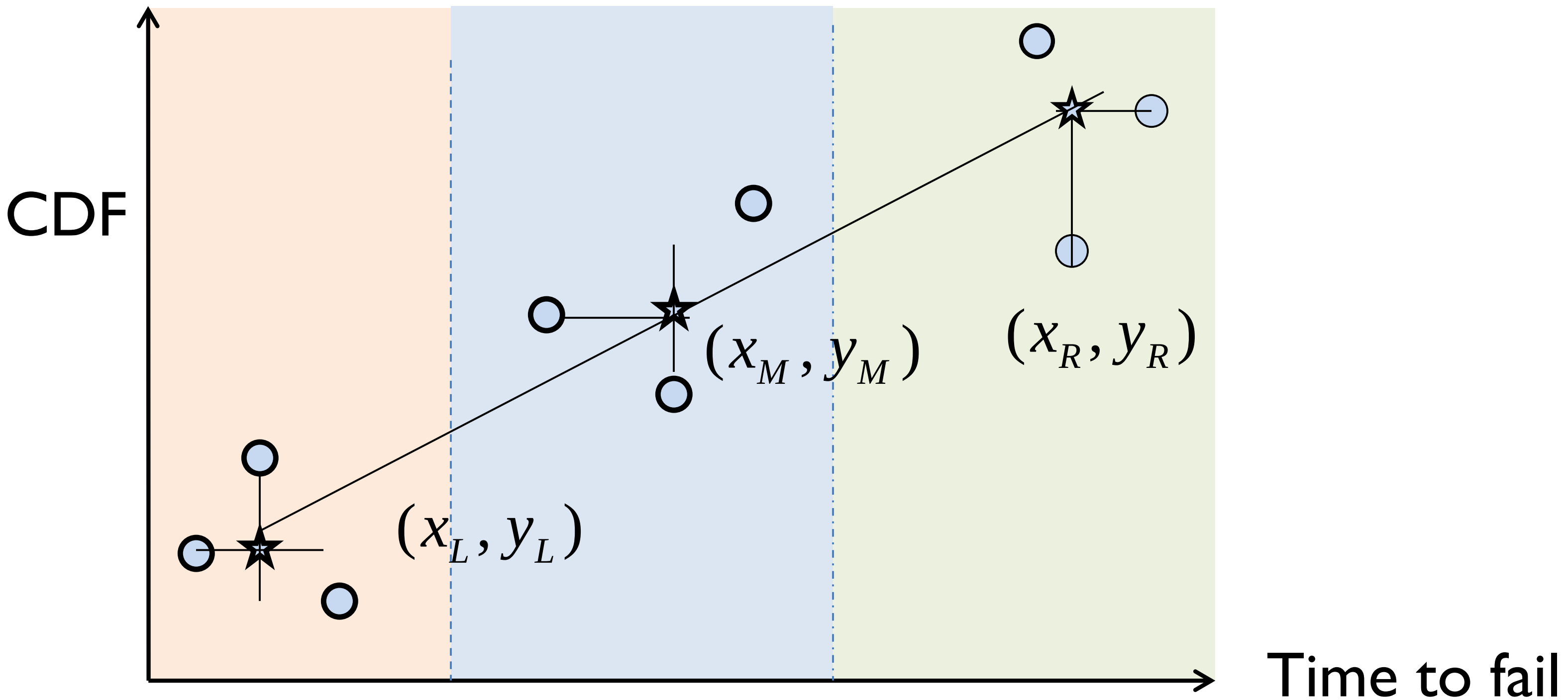
For n=3k+1 (k, k+1, k)

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. Sample vs. population: A Review of traditional statistics

2. Trouble with traditional statistics

   - If the population is not described by Gaussian distribution
   - If some of the datapoints are outliers
   - If some of some experiments ended early

3. Conclusions

Problem of data plotting and numerical CDF

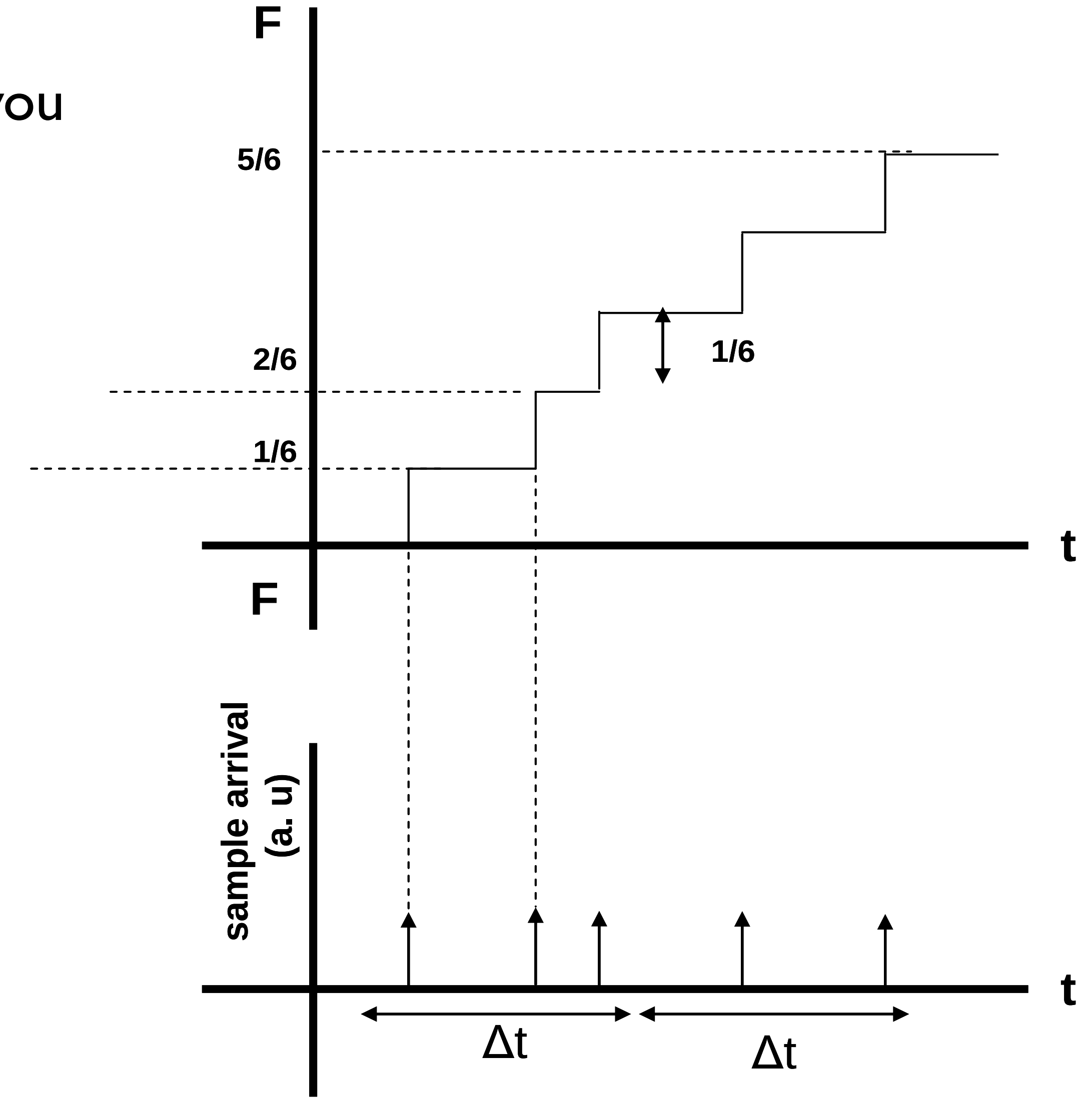
Assume you have 5 transistors and you have collected 5 breakdown times, t_1,t_2,t_3, t_4,t_5

How do we find the CDF?

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

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

$$\boldsymbol{W} = \ln(-\ln(1-\boldsymbol{F}_i))$$



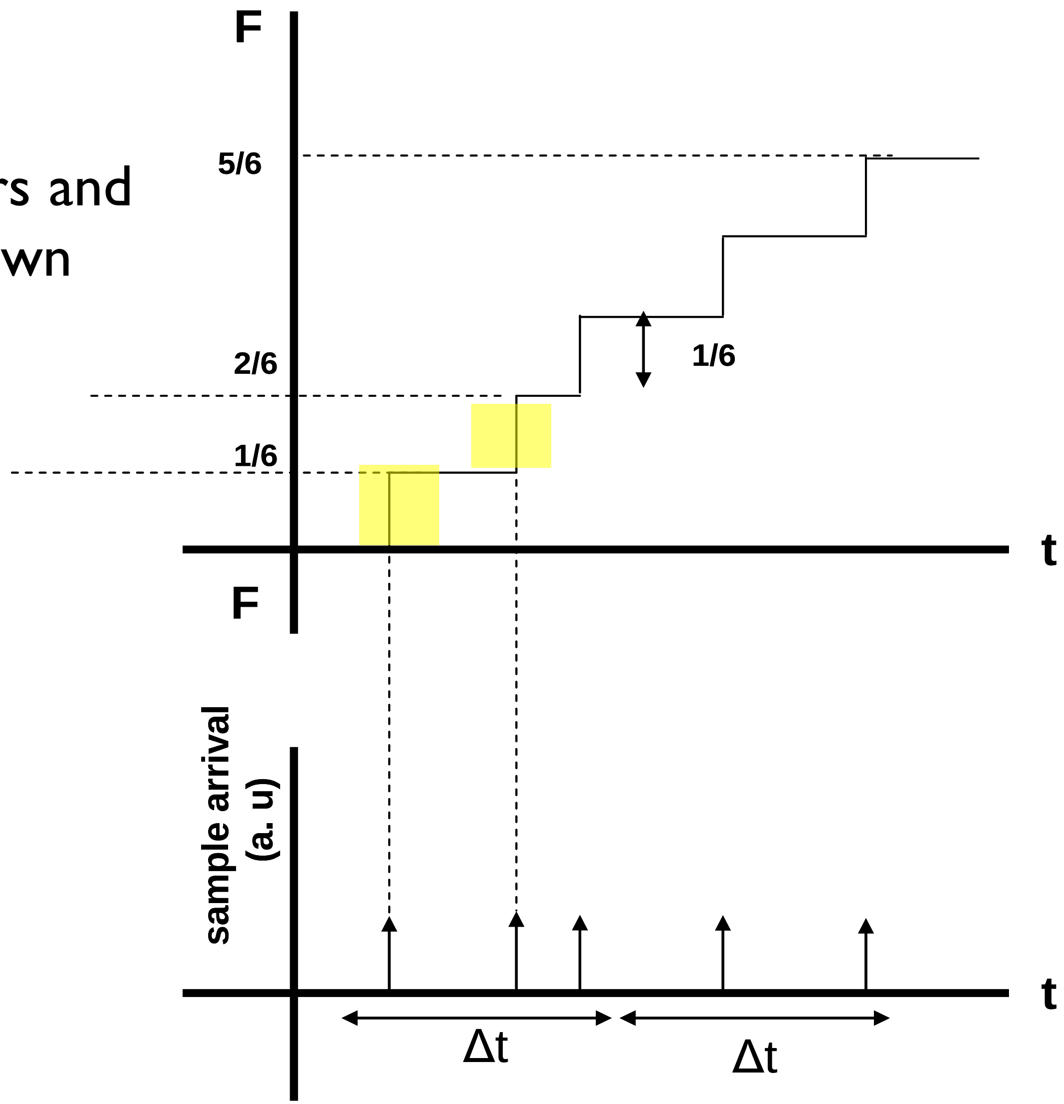
… there is a problem (Failure time is statistical)

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

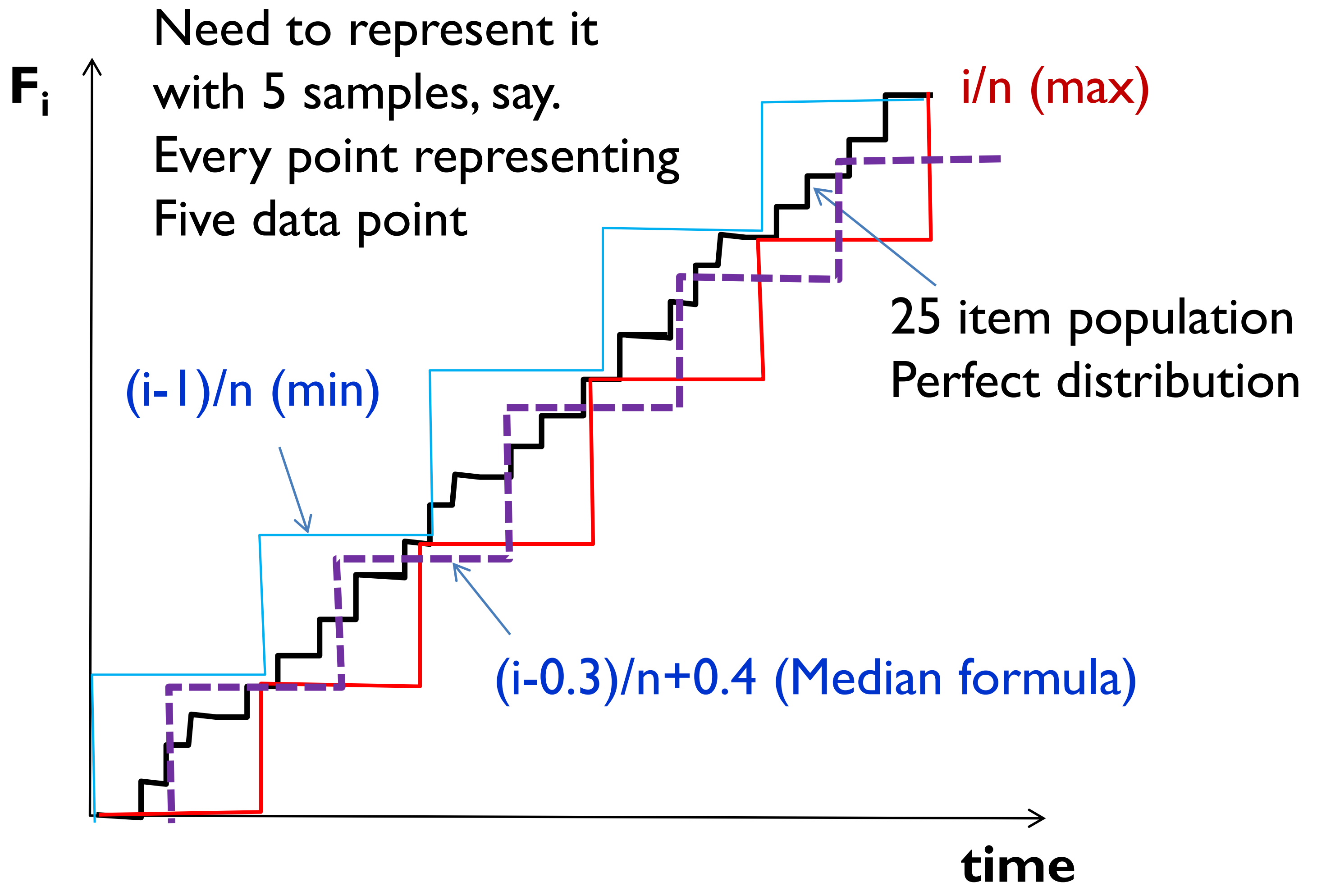
How do we find the CDF?

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

$$\boldsymbol{W} = \ln(-\ln(1 - \boldsymbol{F_i}))$$



Relationship among various formula



Analogous to a congressman …

Aside: Derivation of Hazen Formula

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

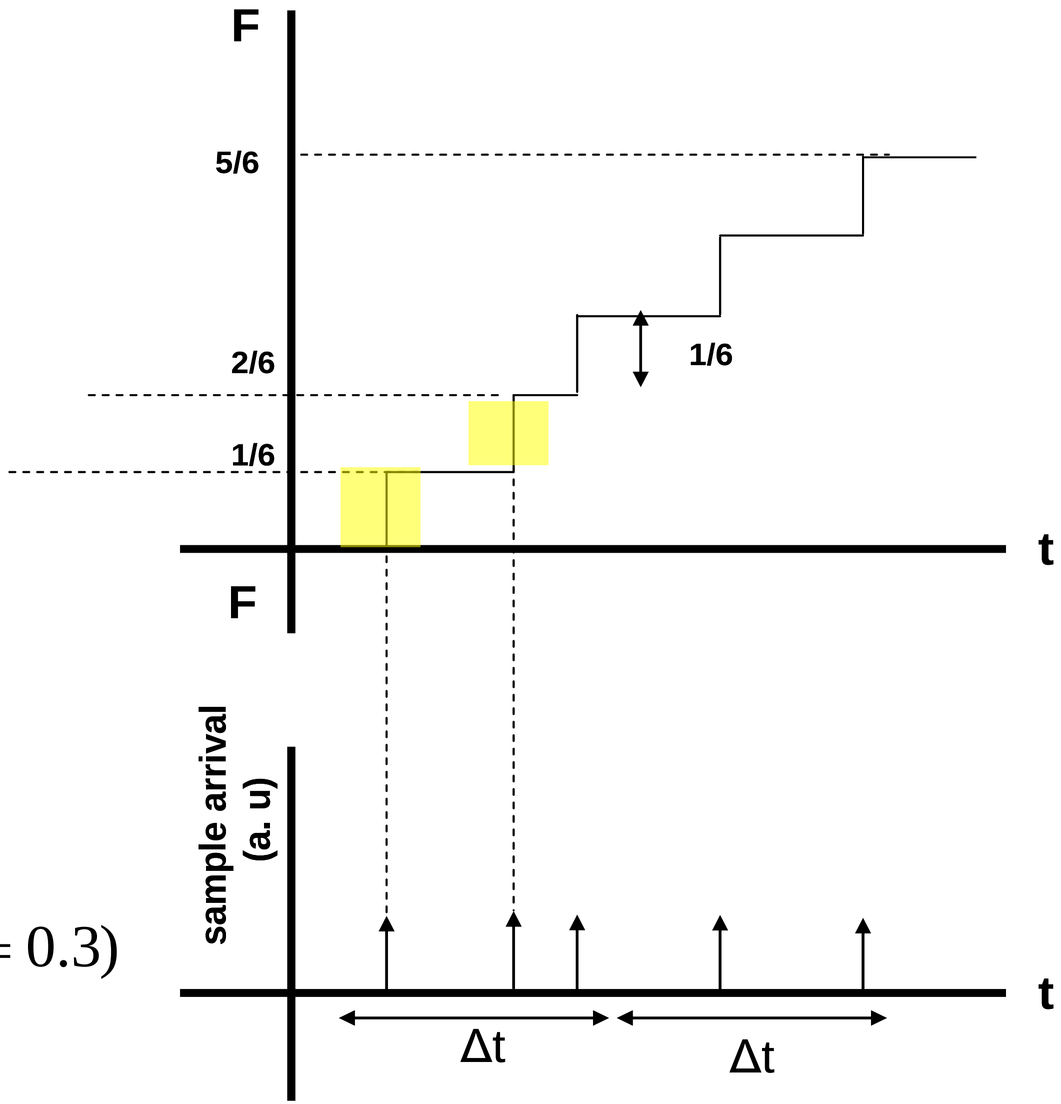
p = Probable CDF location, F_i, of the i-th data

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

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

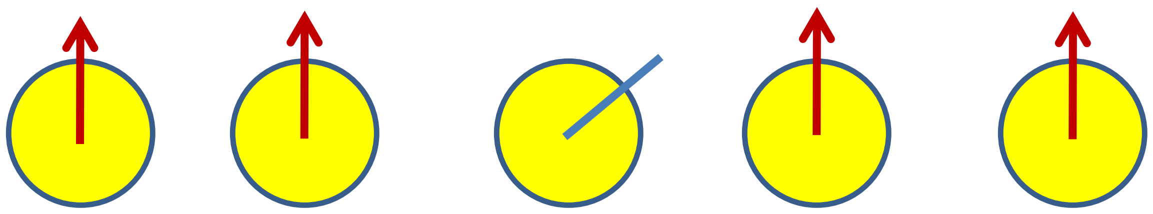
$$\int_0^{F_{Median,i}} g(p)dp \;\; = 1/2$$

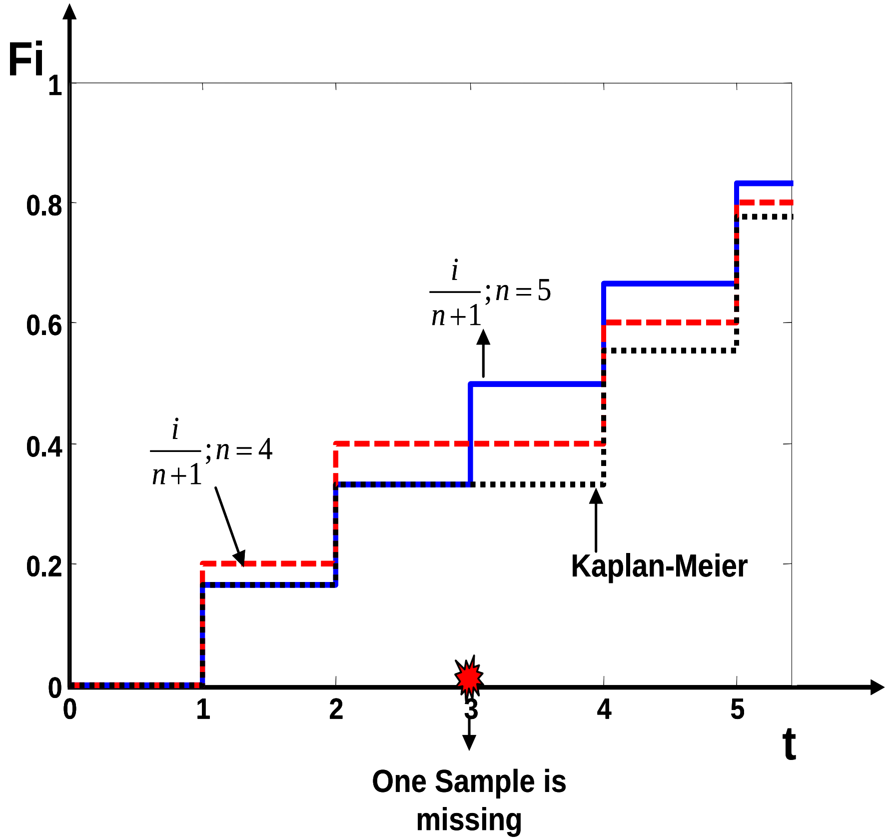
$$\Rightarrow F_{Median,i} = \frac{i-\alpha}{n-2\alpha+1} \quad (\alpha = 0.3)$$



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

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!

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 ...

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

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

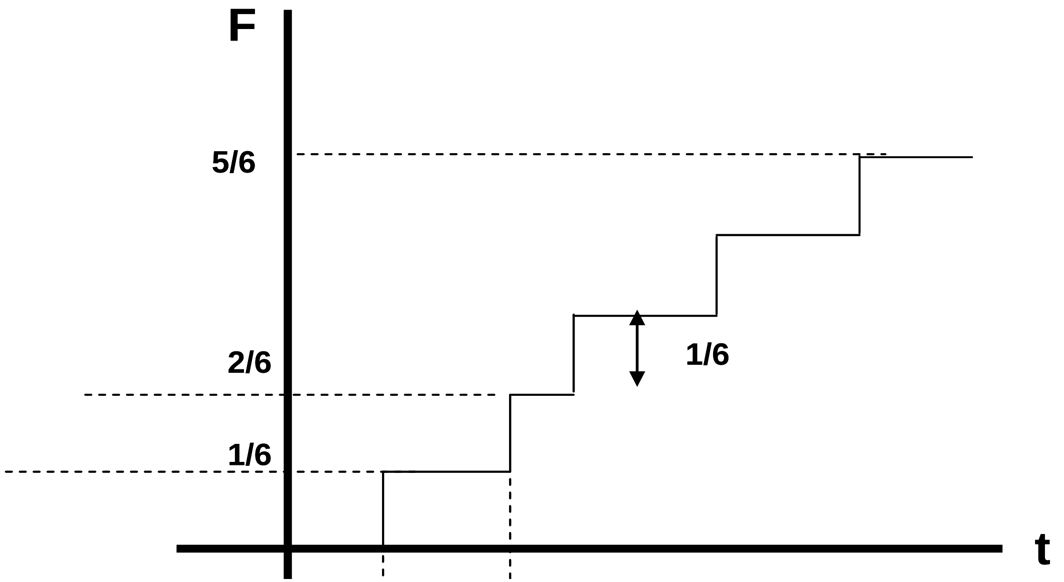
$$F_1 = 1 - \frac{5}{6} = \frac{1}{6}$$

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

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

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

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



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

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

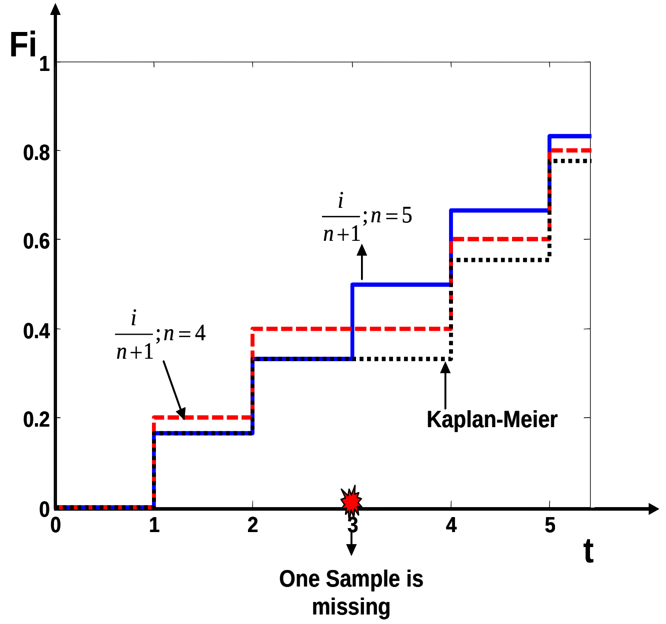
$$F_2 = 1 - \frac{4+1}{4+2} \cdot \frac{3+1}{3+2} = \frac{2}{6}$$

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

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

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

¾ missing …



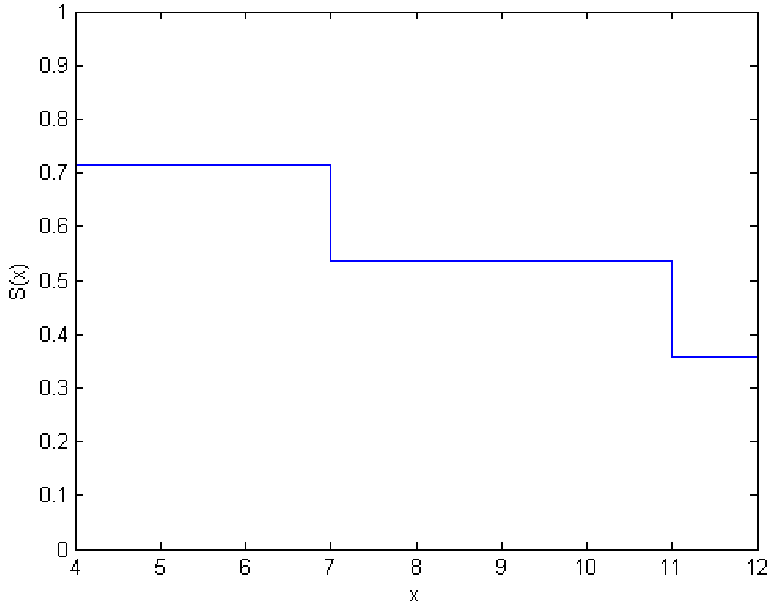
MATLAB Routine for Censored Data

Kaplan-Meier algorithm

```
y =       [4 4 4 7  11  11 12];
cens =  [0 1 0 0    1    0   0];
[f,x] = ecdf(y,'censoring',cens)
```

```
figure()
ecdf(y,'censoring',cens,'function','survivor');
```

Survival function



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.