Primer on Analysis of Experimental Data and Design of Experiments

Lecture 3. Physical and Empirical Distributions

Muhammad A. Alam
alam@purdue.edu

Alam
CEED
Group

Outline

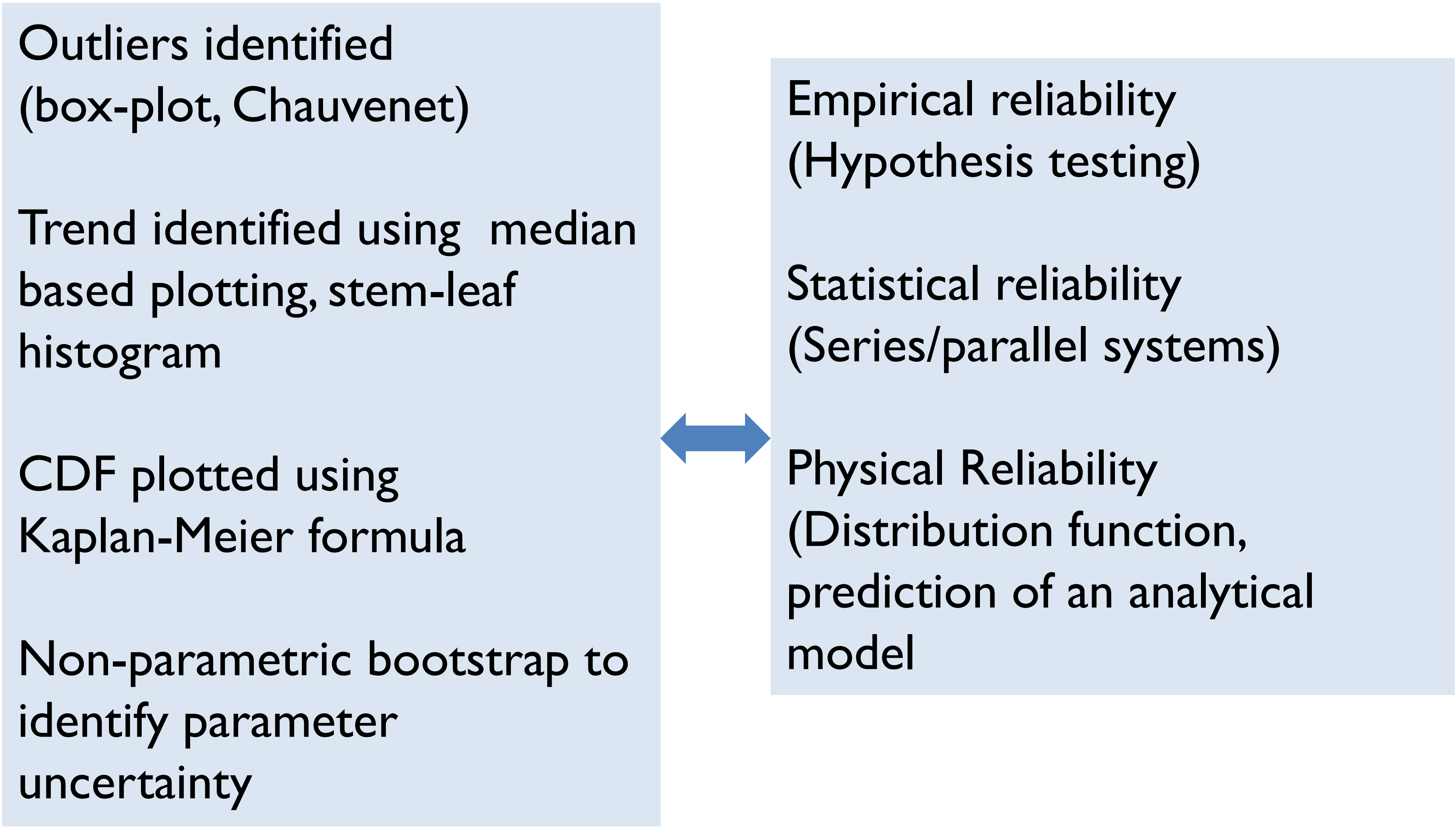
1. Physical Vs. empirical distribution

2. Properties of classical distribution function

3. Moment-based fitting of data

4. Conclusions

Data vs. Hypothesis



Outliers identified
(box-plot, Chauvenet)

Trend identified using median based plotting, stem-leaf histogram

CDF plotted using
Kaplan-Meier formula

Non-parametric bootstrap to identify parameter uncertainty

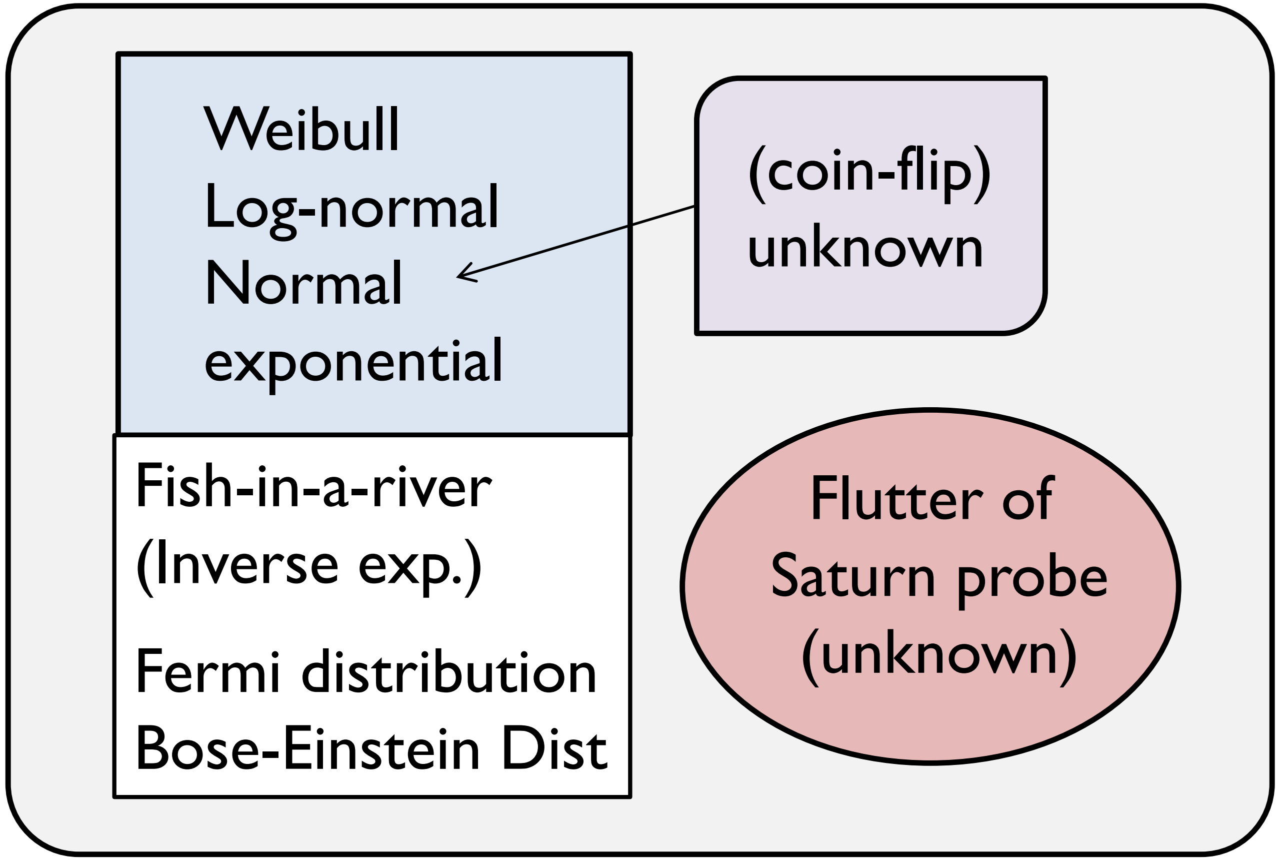
Empirical reliability
(Hypothesis testing)

Statistical reliability
(Series/parallel systems)

Physical Reliability
(Distribution function, prediction of an analytical model

Statistical Distribution is Physical

Experiments


Weibull
Log-normal
Normal
exponential

(coin-flip)
unknown

Fish-in-a-river
(Inverse exp.)

Fermi distribution
Bose-Einstein Dist

Flutter of
Saturn probe
(unknown)

If a problem can be mapped into one of the well known family, large number of results are available.

Outline

1. Physical Vs. empirical distribution
2. Parametric Vs. non-parametric fits
3. Estimating various distribution functions
4. Conclusions

Choosing distribution function

People choose functions that describe wide range of phenomena

❑ **Normal**: After all, everything eventually becomes normal (not really!) Distribution of last resort.

❑ **Log-Normal**: A variant of normal distribution that seems to describe many reliability problems phenomenologically (correlated processes, such as electromigration in interconnects, shunt distribution in solar cells)

❑ **Weibull**: Many physical systems are described by it.

In the limiting case, it becomes Exponential distribution (extreme value problems such as thin oxide breakdown)

Two parameter family: Normal distribution

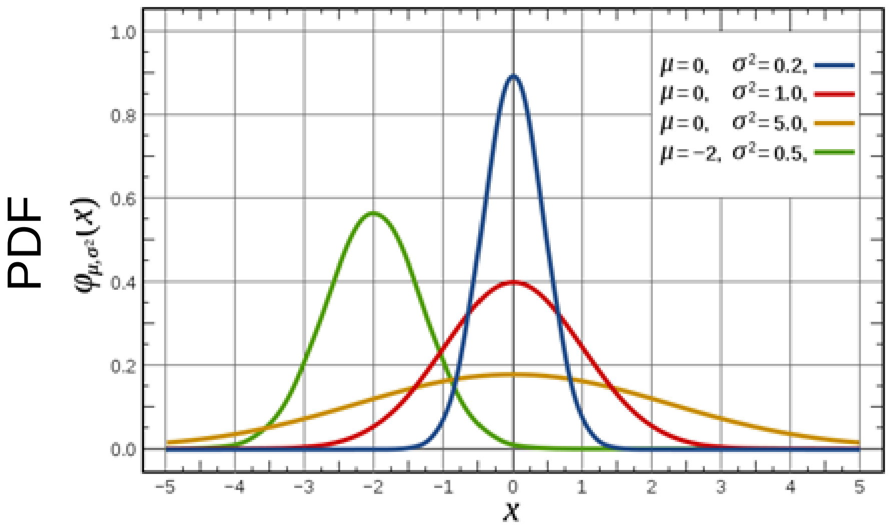
$$f(t;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\cdot\exp\left[-\frac{\{t-\mu\}^{2}}{2\sigma^{2}}\right]$$

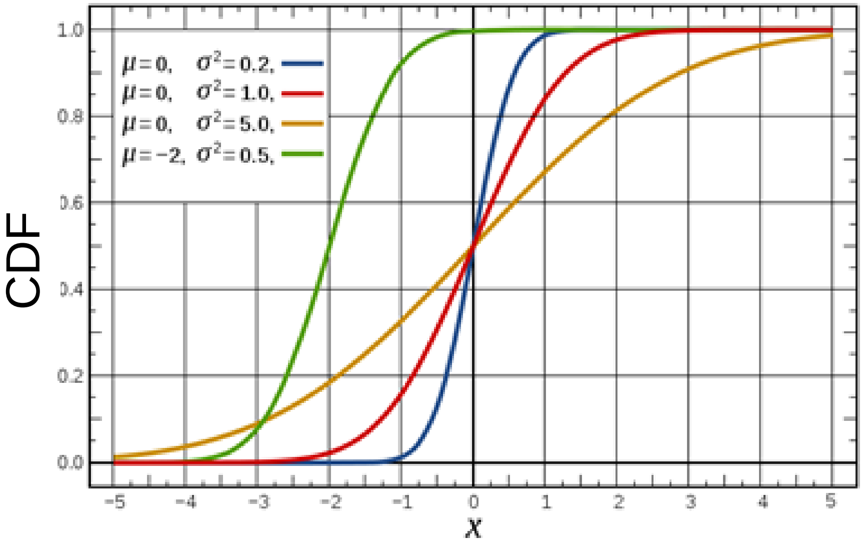
$$F(t)=\Phi\left(\sigma^{-1}(t-\mu)\right)\quad \Phi(z)=\left[1+erf\left(z/\sqrt{2}\right)\right]/2$$

μ=average, σ=standard deviation

Binomial distribution, Poisson distribution, chi-square, student-t distribution …

MATpd = fitdist(data,'Normal')





http://en.wikipedia.org/wiki/Normal_distribution

Two parameter family: log-normal distribution

$$\text{(PDF)}\ f(t;\mu,\sigma)=\frac{1}{t\times\sigma\sqrt{2\pi}}\cdot\exp\left[-\frac{\{\ln(t)-\ln(\mu)\}^2}{2\sigma^2}\right]$$

μ=average, σ=standard deviation

$$\text{(CDF)}\ F(t)=\Phi\left(\sigma^{-1}\ln\frac{t}{\mu}\right)\qquad \Phi(z)=\left[1+erf\,(z/\sqrt{2})\right]/\,2$$

$$\sigma=\ln(t_2/t_1)/\left[\Phi^{-1}(F(t_2))-\Phi^{-1}(F(t_1)\right]$$

$$=\ln(t_2\,@\,F=0.5\,/\,t_1\,@\,F=0.159)$$

$$\lambda(t)=\sqrt{\frac{2}{\pi}}\,\frac{1}{t\sigma}\,\frac{\exp\left[-\sigma^2\{\ln(t/\mu)\}^2/2\right]}{erf\left\{\sqrt{0.5}\ln(t/\mu)/\sigma\right\}}$$

http://en.wikipedia.org/wiki/Log-normal_distribution

Two parameter family: Weibull distribution

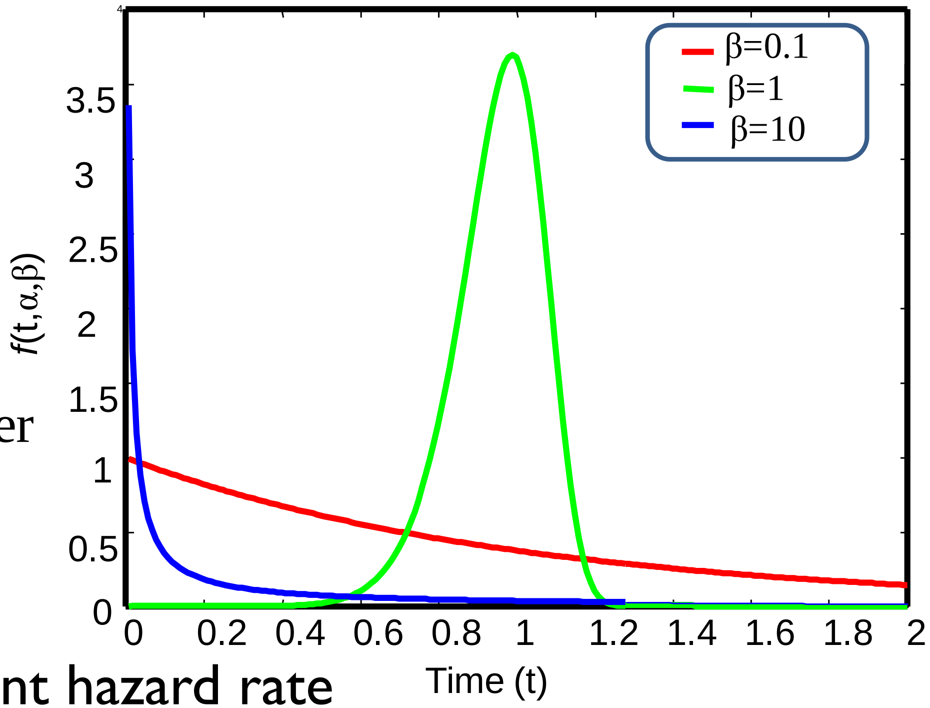
$$f(t;\alpha,\beta)=\frac{\beta}{\alpha^{\beta}}\cdot t^{\beta-1}\cdot e^{-\left(t/\alpha\right)^{\beta}}\quad(\alpha,\beta>0)$$

$$F(t)=1-\exp\left(-(t/\alpha)^{\beta}\right)$$

$$\lambda(t)=\frac{\beta t^{\beta-1}}{\alpha^{\beta}}$$

β=shape parameter, α=scale parameter

Abrahmi recystallization (1905)



β=1 …. Exponential distribution

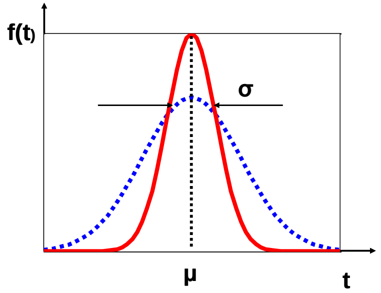
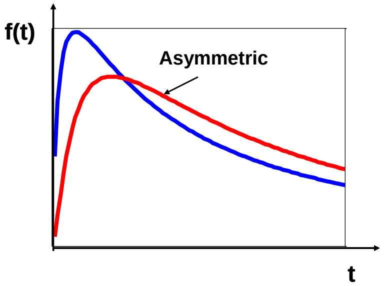
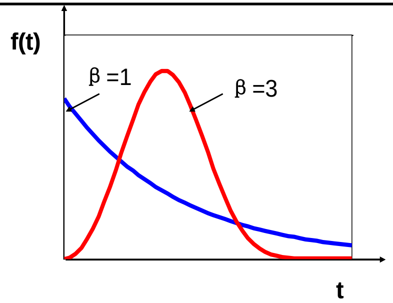
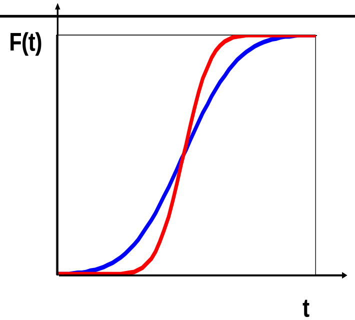
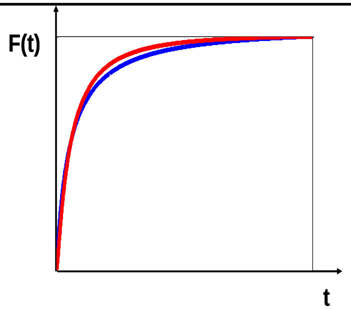
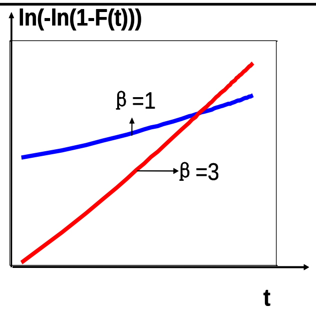
Memory-less distribution, constant hazard rate

β=2 …. Rayleigh distribution

Light scattering, Corrosion in contacts, Failure rate increases with time

http://en.wikipedia.org/wiki/Weibull_distribution

Empirical statistical distributions

	Normal	Log Normal	Weibull
PDF	$\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(t-\mu)^2}{2\sigma^2}}$	$\frac{1}{t\sigma\sqrt{2\pi}}e^{-\frac{(\ln t-\ln\mu)^2}{2\sigma^2}}$	$\frac{\beta}{\alpha}\cdot\left(\frac{t}{\alpha}\right)^{\beta-1}\cdot e^{-\left({}^{t}\!/_{\alpha}\right)^{\beta}}$
PDF			
CDF			
Moment 1st	μ	$\mu\cdot e^{-\frac{\sigma^2}{2}}$	$\alpha\sqrt{{}^{1}\!/_{\beta}}$
Moment 2nd	σ^2	$2\mu\cdot\left(e^{\sigma^2}-1\right)\cdot e^{\sigma^2}$	$\sqrt{\alpha^2\sqrt{1+\frac{2}{\beta}}-\alpha^2\Gamma^2(1+\frac{1}{\beta})}$

Definitions of distribution functions

Name	Symbol	Expression
Prob. distribution	$f(t;\alpha,\beta,\ldots)$	$f(t;\alpha,\beta,\ldots)$
cumulative PDF	$F(t)$	$\int_{-\infty}^{t} f(t^{'})dt^{'}$
survival function	$R(t)$	$1-F(t)$
hazard rate	$\lambda(t)$	$\frac{f(t)}{1-F(t)}$
cum. hazard rate	$H(t)$	$\int_{o}^{t}\lambda(t^{'})dt^{'}$
average hazard	$\lambda_c(t)$	$1/t\int_{o}^{t}\lambda(t^{'})dt^{'}$

These functions are used in difference fields in different ways …

Example: Derivation of hazard function

H(t) … Probability that having survived till time t (event A), it fails within time (t+dt) (event B)

B includes A ….(failed before)

$$P(B|A) = \frac{P(A*B)}{P(A)} = \frac{P(B)}{P(A)} = \frac{f(t)dt}{R(t)} \Rightarrow h(t) = \frac{f(t)}{R(t)}$$

Integrated Hazard till time t ….

$$\boldsymbol{H(t)} = \int_0^{\boldsymbol{t}} \boldsymbol{h(u)du} = -\ln \boldsymbol{R(t)}$$

Discrete Transform: because data is discrete

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

$$f_i = \frac{dF_i}{dt} = \frac{F_{i+1}-F_i}{t_{i+1}-t_i} = \frac{1}{\left(n-2\alpha+1\right)\left(t_{i+1}-t_i\right)}$$

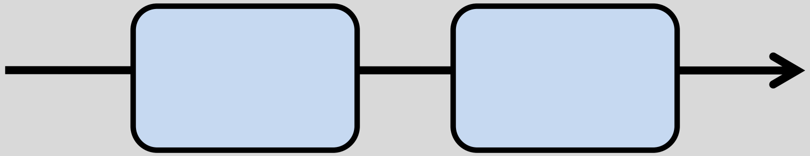
$$\lambda_i = \frac{f_i}{1-F_i} = \frac{1}{\left(n-i-\alpha+1\right)\left(t_{i+1}-t_i\right)}$$

Transformation among reliability functions

	$f(t)$	$F(t)$	$R(t)$	$\lambda(t)$	$H(t)$	$\lambda_c(t)$
$f(t)$	$f(t)$	$\frac{dF(t)}{dt}$	$-\frac{dR(t)}{dt}$	$\lambda(t)e^{-\int_0^t \lambda(t')dt'}$	$e^{-H(t)}\frac{dH(t)}{dt}$	$\left(\lambda_c + t\frac{d\lambda_c(t)}{dt}\right)e^{-\lambda_c(t)t}$
$F(t)$	$\int_o^t f(t')dt'$	$F(t)$	$1-R(t)$	$1-e^{-\int_0^t \lambda(t')dt'}$	$1-e^{-H(t)}$	$1-e^{-\lambda_c(t)t}$
$R(t)$	$\int_t^\infty f(t')dt'$	$1-F(t)$	$R(t)$	$e^{-\int_0^t \lambda(t')dt'}$	$e^{-H(t)}$	$e^{-\lambda_c(t)t}$
$\lambda(t)$	$\frac{f(t)}{\int_t^\infty f(t')dt'}$	$-\frac{d\ln(1-F(t))}{dt}$	$-\frac{d\ln R(t)}{dt}$	$\lambda(t)$	$\frac{dH(t)}{dt}$	$\lambda_c + t\frac{d\lambda_c(t)}{dt}$
$H(t)$	$-\ln\left(\int_t^\infty f(t')dt'\right)$	$-\ln(1-F(t))$	$-\ln R(t)$	$\int_o^t \lambda(t')dt'$	$H(t)$	$t\lambda_c(t)$
$\lambda_c(t)$	$-\frac{1}{t}\ln\left(\int_t^\infty f(t')dt'\right)$	$\frac{-\ln(1-F(t))}{t}$	$\frac{-\ln R(t)}{t}$	$1/t\int_o^t \lambda(t')dt'$	$\frac{H(t)}{t}$	$\lambda_c(t)$

HW: Derive a few reliability functions yourself ...

Series and parallel systems: how to use the distribution functions

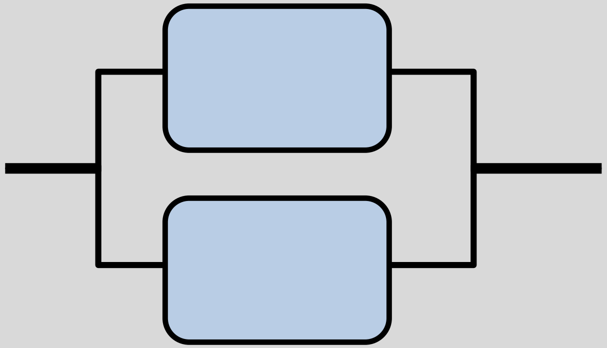


$$1-F_s(t)=\left[1-F_1(t)\right]\times\left[1-F_2(t)\right]\times\ldots$$

$$R_s(t)=R_1(t)\times R_2(t)\times\ldots$$

$$\lambda=\frac{dF/dt}{1-F(t)}=\frac{-dR/dt}{R(t)}=-\frac{d}{dt}\ln R$$

$$\boldsymbol{\lambda_s(t)}=\lambda_1(\boldsymbol{t})+\lambda_2(\boldsymbol{t})+\ldots$$



$$F_s(t)=F_1(t)\times F_2(t)\times\ldots$$

$$1-R_s(t)=\left[1-R_1(t)\right]\times\left[1-R_2(t)\right]\times\ldots$$

Advantage of redundant system ..

$$\frac{\boldsymbol{\lambda_i(t)}}{\boldsymbol{\lambda_s(t)}}=\frac{1+\boldsymbol{F}+\boldsymbol{F}^2+\ldots\boldsymbol{F}^{n-1}}{\boldsymbol{nF}^{n-1}}$$

R1, R2, … may have different distributions

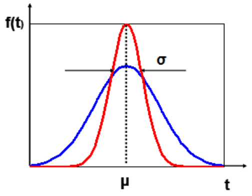
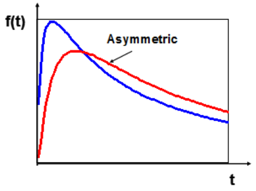
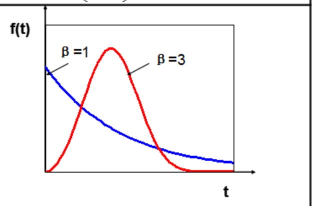
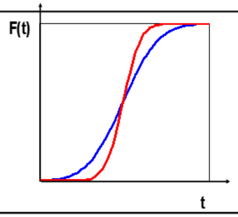
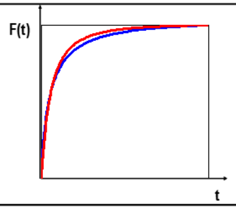
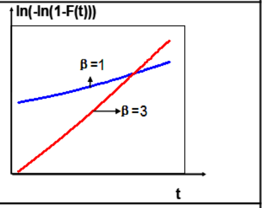
Outline

1. Physical Vs. empirical distribution
2. Properties of classical distribution function
3. Moment-based fitting of data
4. Conclusions

Moment-based fitting

Of 60 oxides, 7 failed in 1000 hrs

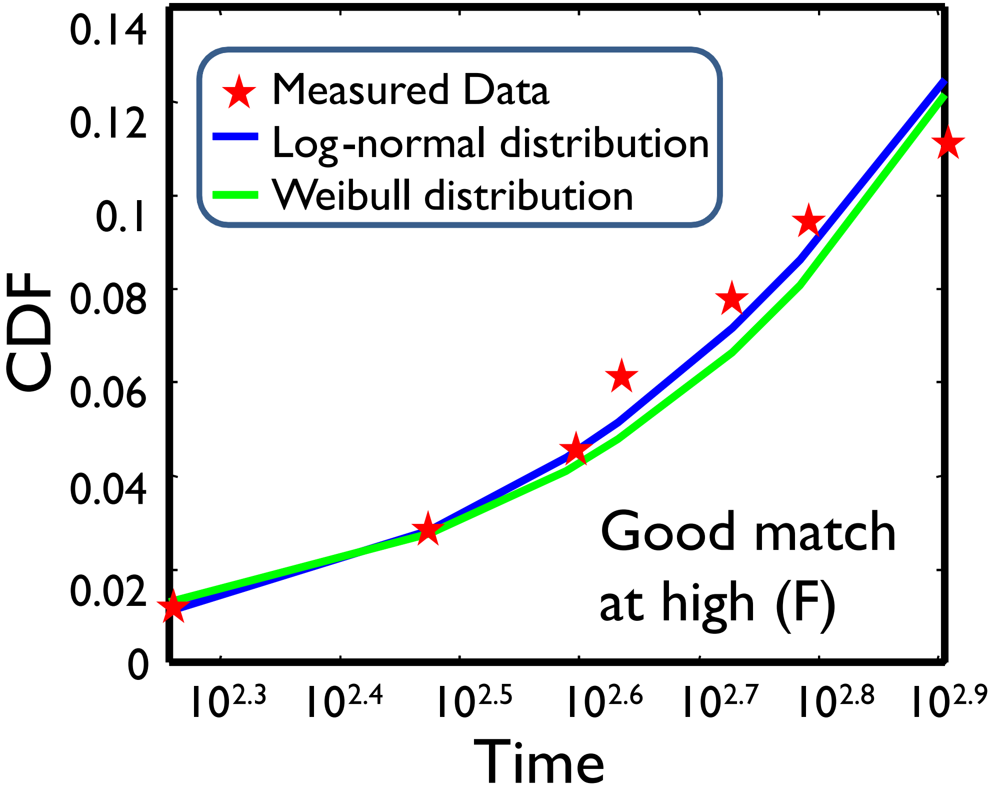
Rank	Lifetime	$F_i =$ $(i - 0.3)/(n + 0.4)$
1	181	0.012
2	299	0.028
3	389	0.045
4	430	0.061
5	535	0.078
6	610	0.094
7	805	0.111

	Normal	Log Normal	Weibull
PDF	$\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(t-\mu)^2}{2\sigma^2}}$	$\frac{1}{t\sigma\sqrt{2\pi}}e^{-\frac{(\ln t-\ln\mu)^2}{2\sigma^2}}$	$\frac{\beta}{\alpha}\cdot\left(\frac{t}{\alpha}\right)^{\beta-1}\cdot e^{-\left(t/\alpha\right)^{\beta}}$
PDF			
CDF			
Moment 1st	μ	$\mu\cdot e^{-\frac{\sigma^2}{2}}$	$\alpha\sqrt{1/\beta}$
Moment 2nd	σ^2	$2\mu\cdot\left(e^{\sigma^2}-1\right)\cdot e^{\sigma^2}$	$\sqrt{\alpha^2\sqrt{1+\frac{2}{\beta}}-\alpha^2\Gamma^2(1+\frac{1}{\beta})}$

Matching moments to distributions

Of 60 oxides, 7 failed in 1000 hrs

Rank	Lifetime	$F_i = (i - 0.3)/(n + 0.4)$
1	181	0.012
2	299	0.028
3	389	0.045
4	430	0.061
5	535	0.078
6	610	0.094
7	805	0.111

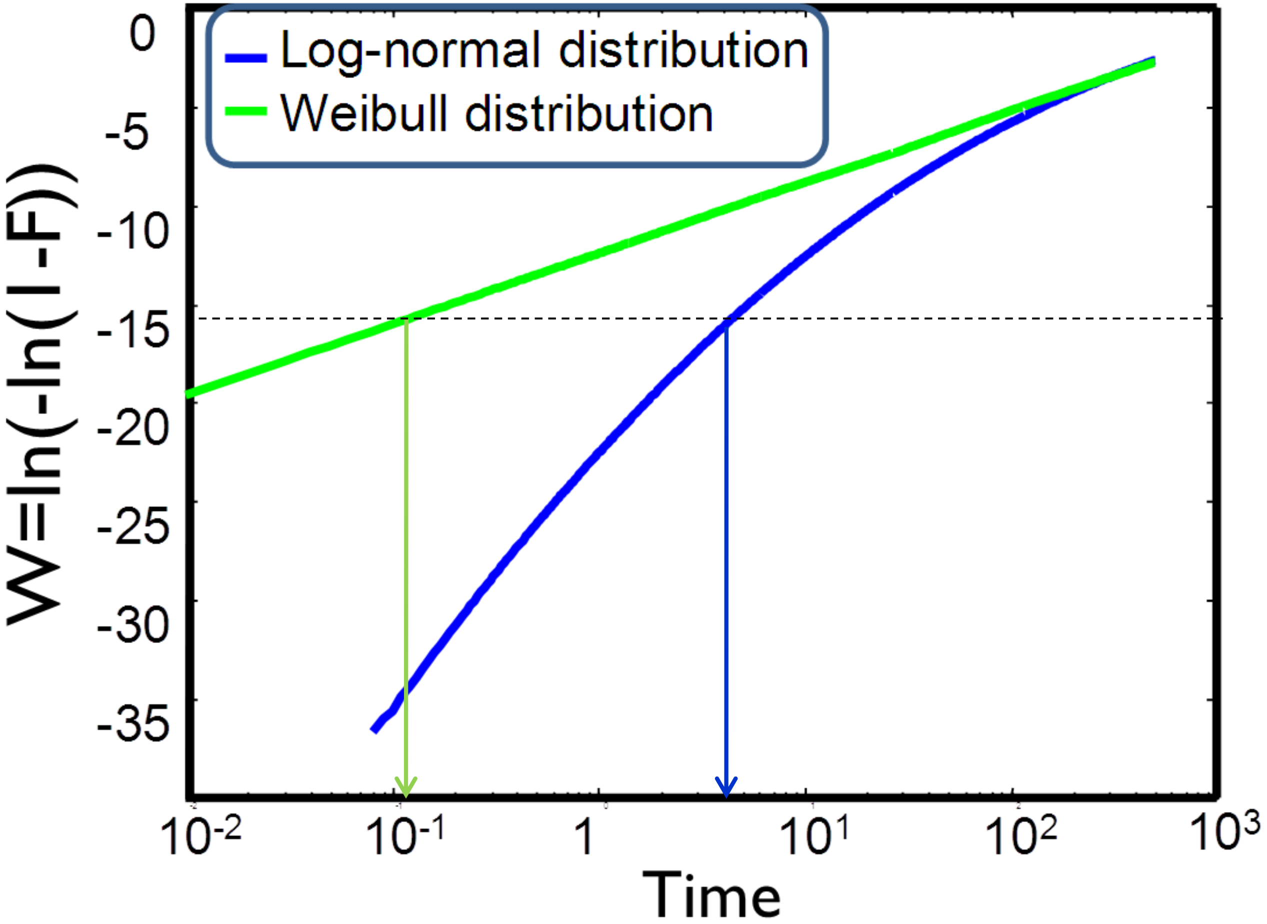


Good match at high (F)

Weibull Distribution Parameters
When $t=\alpha$, $\ln(1-F(t))=-1$, $F(t)=0.632$, a=2990
β estimated using parameter fitting as 1.56

Log-Normal Distribution Parameters
$s=\ln(T_{50\%}/T_{15.9\%})$, $\sigma=\ln(3600/980)=1.30$
$\mu=\ln(T_{50\%})=\ln(3600)=8.19$

Problem of matching the moments

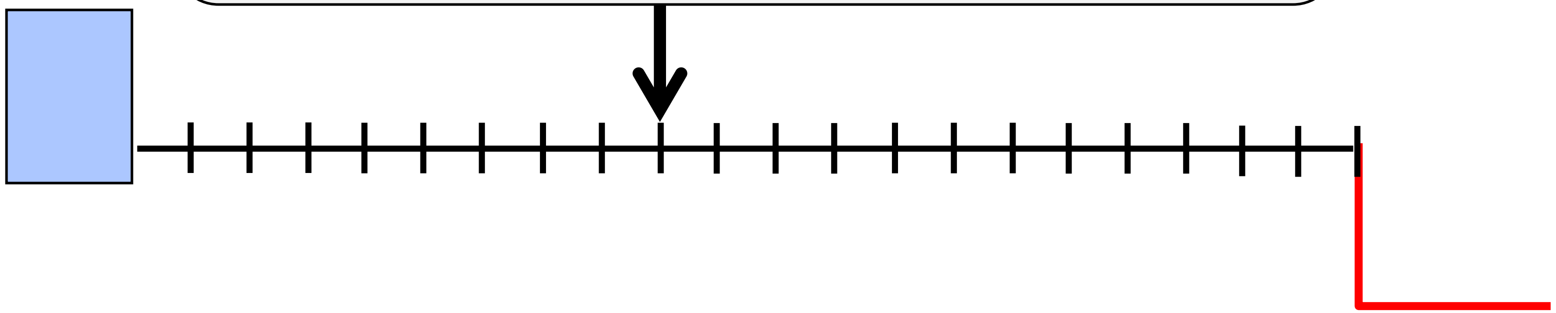


Log-normal distribution is considerably optimistic

Problem of matching distribution: BFRW

$$\frac{\partial P}{\partial t} = D\frac{\partial^2 P}{\partial x^2} \qquad P(x,t=0) = \delta(x - x_0)$$

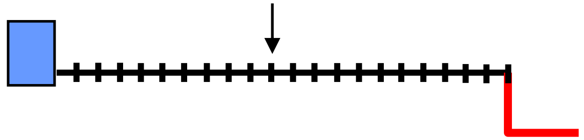
$$P(x=0,t) = 0$$



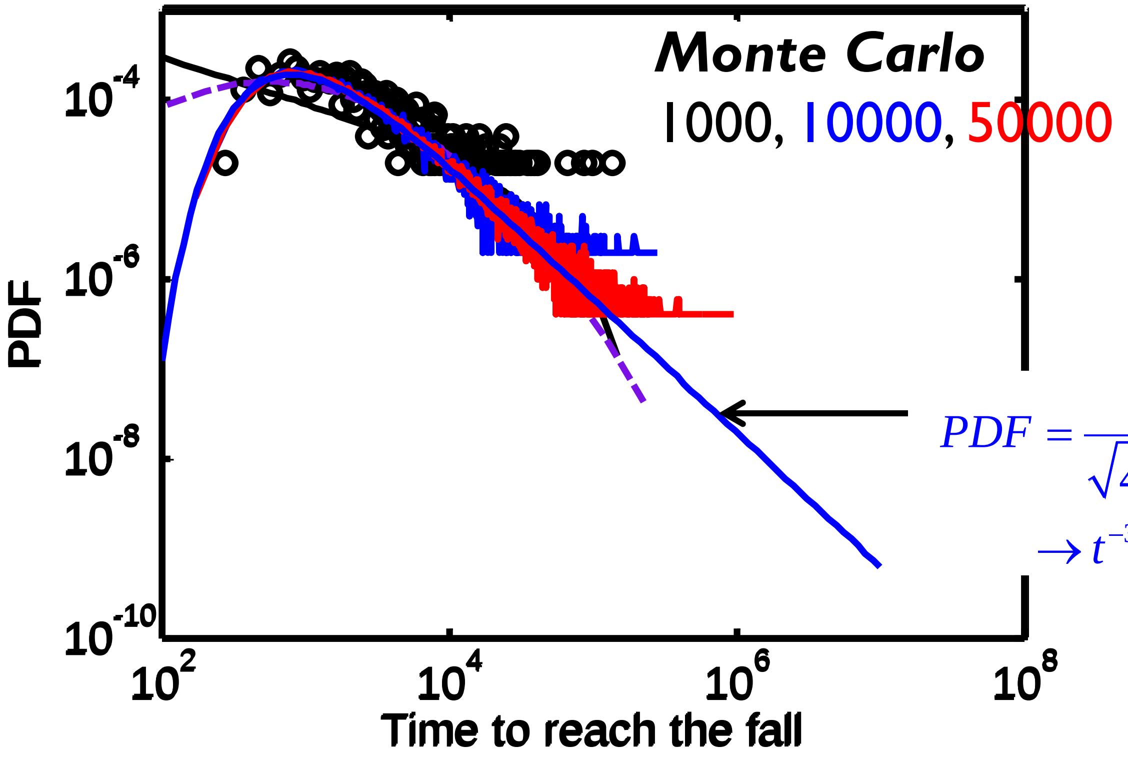
$$P(x,t) = (4\pi Dt)^{-1/2}\left[\mathrm{e}^{-(x\ -x_0)^2/4Dt} - \mathrm{e}^{-(x\ +x_0)^2/4Dt}\right]$$

$$\int_0^t f(\tau)d\tau + \int_0^L P(x,t)dx = 1 \Rightarrow \quad f(t) = \frac{x_0}{\sqrt{4\pi Dt^3}}\mathrm{e}^{-x_0^{\,2}/4Dt}$$

Match apparently reasonable, but wrong



$$f_G(t) = \frac{t^{k-1}e^{-t/\theta}}{\Gamma(k)\theta^k} \qquad T_{avg} = k\theta$$



$$PDF = \frac{x_0}{\sqrt{4\pi D t^3}}\mathrm{e}^{-x_0{}^2/4Dt}$$

$$\rightarrow t^{-3/2}$$

Conclusions

1. Once the data is plotted using the principles discussed in the previous lectures, the phenomenon can be described by a statistical model.

2. If unsuccessful, one should choose functions with least number of variables that described the system. Many applications are described by 2-parameter distributions (e.g. log-normal, Weibull).

3. For an extreme value problem, one should pay particular attention to the tail of the distribution and choose sample size accordingly.

4. Moment-based methods are popular, but cannot distinguish between the tails of the distribution (associated with high moments)

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.

AT&T, “Statistical Quality Control Handbook”.

Joan Fisher Box, “R. A. Fisher and the Design of Experiments, 1922-1926”, *The American Statistician,* vol. 34, no. 1, pp. 1-7, Feb. 1980.

F. Yates, “Sir Ronald Fisher and the Design of Experiments”, *Biometrics,* vol. 20, no. 2, In Memoriam: Ronald Aylmer Fisher, 1890-1962., pp. 307-321, (Jun. 1964.

Ranjith Roy, “A primer on the Taguchi Method”, Van Nostrand Reinhold International Co. Ltd., 1990.

Lloyd W Condra, “Reliability Improvement with design of experiments”, Marcel Dekker Inc., 1993.

Review questions

G1: Why do people use Normal, log-normal, Weibull distributions when they do not know the exact physical distribution?

G2: What is the problem of using empirical distributions? What are the advantages?

G3: If you must choose an empirical distribution, what should be your criteria? (Nos. of parameters, physical principles, etc.)

G4: Why does everyone suggest the use of CDF for empirical data-fitting, rather than PDF? (Obviously one can go from one function to the other)

G5: There are all sorts of distribution functions (e.g. survivability function) ? If everything is related to everything else, why do we need so many?

G6: How would you determine the BFRW failure rates? Mean Hazard rate?

Excellent resource at ….

1. Statistics Online Computational Resource

   http://www.socr.ucla.edu/SOCR.html

2. Excellent toolset within Excel

3. S and S-Plus software set

4. MATLAB has nearly everything!

Parametric vs. non-parametric Bootstrap

0.2 -0.1 0.5 0.3 -0.6

Fit the distribution of your choice by Maximum likelihood estimators (MLE) (obtain parameters, i.e. η_0, β_0)

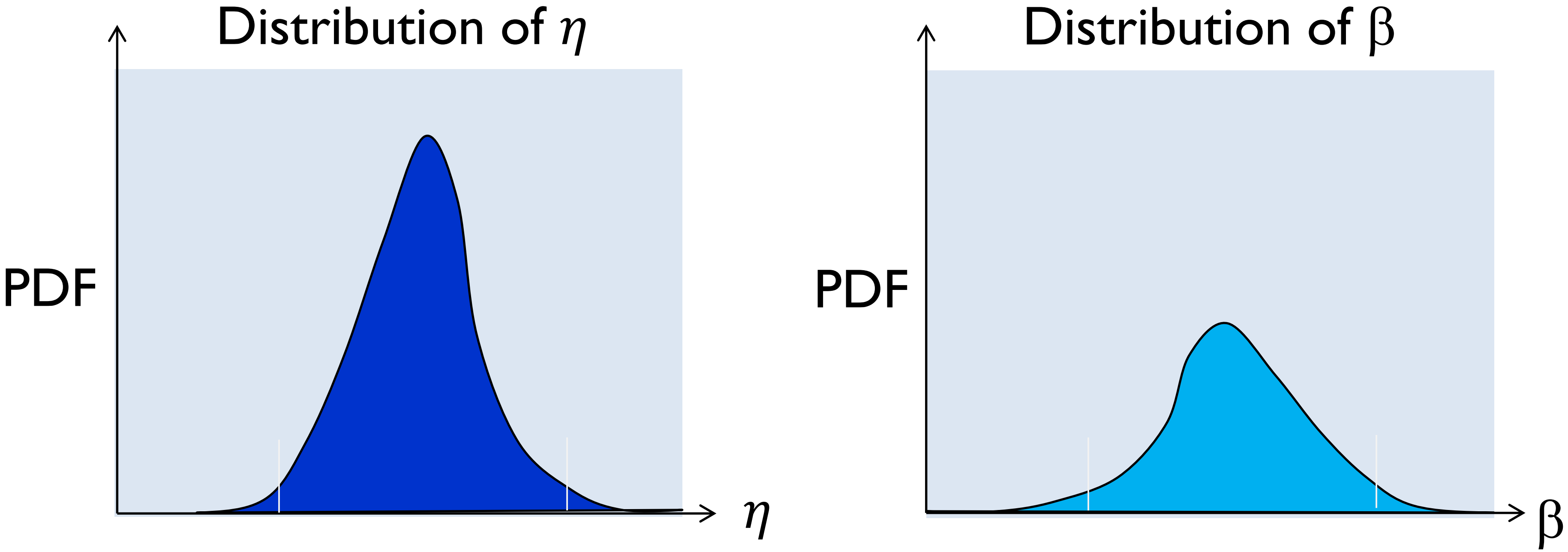
Generate synthetic samples based on the parametric distribution

0.12 -0.17 -0.44 -0.71 0.52 Synthetic sample 1 (new η_1, β_1)

0.32 0.21 -0.69 0.23 0.58 Synthetic sample 2 (new η_2, β_2)

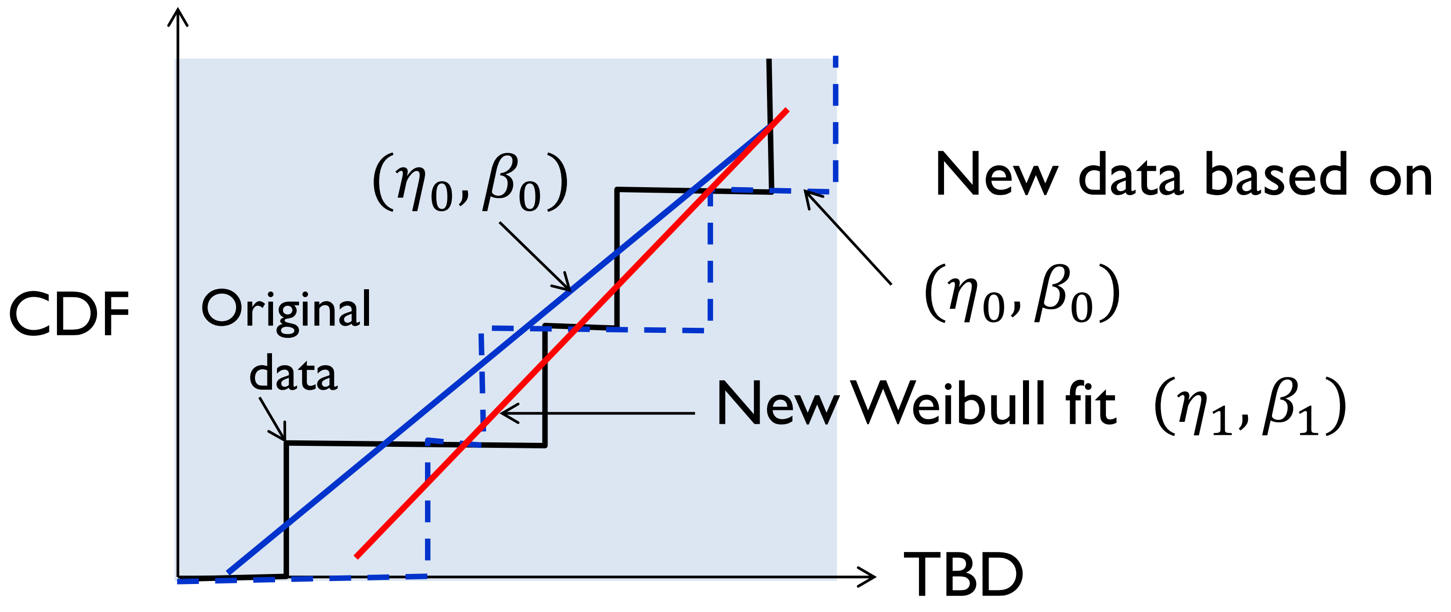
Plot distribution of statics η_i, β_i

Distribution of α and β



Same technique for polling and tenure rate of faculty!

Why resampling from the same distribution generates new fit parameters



Samples taken from the same distribution (η_0, β_0) generates datapoints that are fitted with new (η_i, β_i)

References

1. “Detecting Novel Associations for large scale dataset”, D. Reshef et al. , Science 334, p. 1418, 2011.
2. “Survival Analysis of Faculty Retention in Science and Engineering”, D. Kaminski et al., Science, 335, 864, 2012.
3. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-Law Distributions in Empirical Data,” SIAM Review, vol. 51, no. 4, p. 661, Nov. 2009.