Data Analytics for Materials Science

A.D. (Tony) Rollett

Dept. Materials Sci. Eng., Carnegie Mellon University

Univariate Statistics

Lecture 3A

Revised: 8th Feb., 2021

Objectives for the lecture

- The *main objective* is to reinforce what you should already know about statistics.
- In more detail, given a set of values (datapoints, measurements) you are expected to be able to evaluate standard statistical quantities such as the mean, median, standard deviation.
- In further detail, you are expected to be able to fit standard distributions to your data as a way of quantifying it and having a model for it. For example, is your dataset normally distributed, or exponential, or gamma …?
- The *secondary objective* is for you to understand that many materials problems of practical engineering interest are *extreme value problems*, which motivates us to quantify the *tails of distributions*. For example, what is the probability of finding a pore > 50 μm in the high stress region of a given sample?

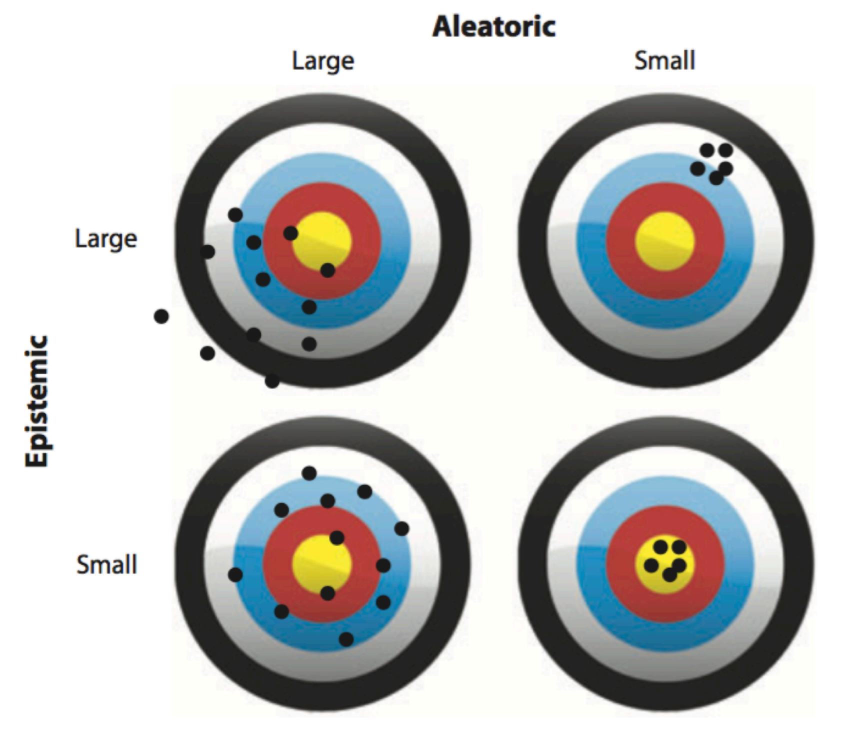
Motivation for lecture

- Developing skills in data analysis is essential to getting answers to hypotheses.
- A hypothesis should always contain a quantitative statement of how success/confirmation can be determined.
- Univariate statistics are sometimes sufficient to provide an answer.

Different Types of Uncertainty

Materials Variations
- Aleatory Uncertainty because each sample has a different microstructure
Characterization:
grain size
grain shape …
- Epistemic Uncertainty because we lack knowledge as to which measures are significant *and* we lack knowledge of how to generate microstructures that match by eye

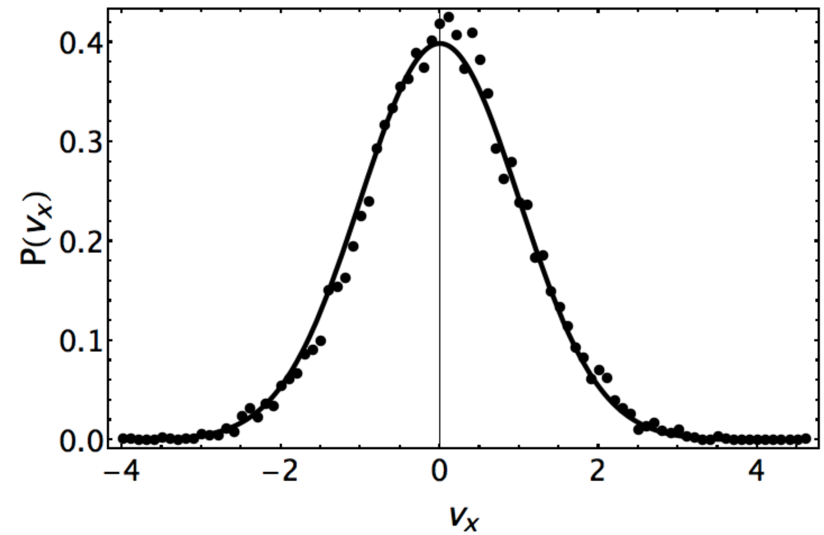
Response Variations
- Aleatory Uncertainty because of microstructure variability (see left box)
- Epistemic Uncertainty because we lack knowledge as to deterministic model allows us to predict where fatigue cracks will start in a given microstructure
- *If we can eliminate the above uncertainty, then we may be able to address the epistemic uncertainty in microstructure characterization*



Aleatoric uncertainty describes inherent randomness and is akin to precision.
Epistemic uncertainty is similar to accuracy in that it describes the bias in the model.

In-class discussion: could univariate statistics help us distinguish these 2 types of uncertainty?

Statistics



mean $$\langle X_i \rangle = \frac{1}{N}\sum_{k=1}^{N} X_{ki}$$

bias-corrected variance $$S_i = \frac{1}{N-1}\sum_{k=1}^{N}\left(X_{ki} - \langle X_i \rangle\right)^2$$

bias-corrected standard deviation $$\sigma_i = \sqrt{S_i}$$

$$\rightarrow$$

bias-corrected covariance between types of data

$$C_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}\left(\vec{X}_{ki} - \langle X_i \rangle\right)\left(\vec{X}_{kj} - \langle X_j \rangle\right)$$

$$C_{ii} = \frac{1}{N-1}\sum_{k=1}^{N}\left(\vec{X}_{ki} - \langle X_i \rangle\right)^2 = S_i$$

Samples from Populations

- To explain about "bias correction" …
- If your dataset is the entire *population* that is relevant to the problem at hand then you can use standard measures of, e.g., variance.
- In-class discussion: differences between a *census* and a *poll*? Should you use $\mathrm{var}(x)$ on census results?!
- Go to https://statisticsglobe.com/variance-in-r-var-function and let's get an idea of the effect of taking a small sample from a large population.
- The "N-1" in the denominator for the sample variance calculation is the *bias correction.*
- Very important: R computes the sample variance (not the population variance).

Distributions

- Normal; “bell curve” – standard assumption in statistical analysis, many mathematical advantages.
- Several tests* of normality such as the Kolmogorov-Smirnov ("KS") and Wilk-Shapiro ("W") tests.
- Log-normal – more commonly found in materials science, e.g., grain or particle sizes. Important to note that this is obtained via a *transformation* of the data.

- And many others … that we will not examine here because the focus of the course is on multi-variate methods (and machine learning).

* section 2.4 in Jobson

Large 3D Data Sets – partial list

Serial Sectioning:

- IN 100 – powder metallurgy Ni alloy (Groeber):
  Groeber-big3D-Grains-Bulk-Edge.txt
- Beta-21S* – single phase Ti alloy (Rowenhorst)
  Rowenhorst-Grains-Bulk-Edge.txt
- Monte Carlo simulation by Seth Wilson, ADR:
  Seth_Wilson-MonteCarlo-volumes-4399.dat
  MonteCarlo-particles-ppU101-GrainList-13000000.txt
- Pure Ni – used for GBCD and GB energy (Rohrer)
- ZrO_2 – used for GBCD and GB energy (Rohrer)
- Y_2O_3 – used for GBCD and GB energy (Rohrer)

Thin Metallic Films (courtesy of Prof. K. Barmak):

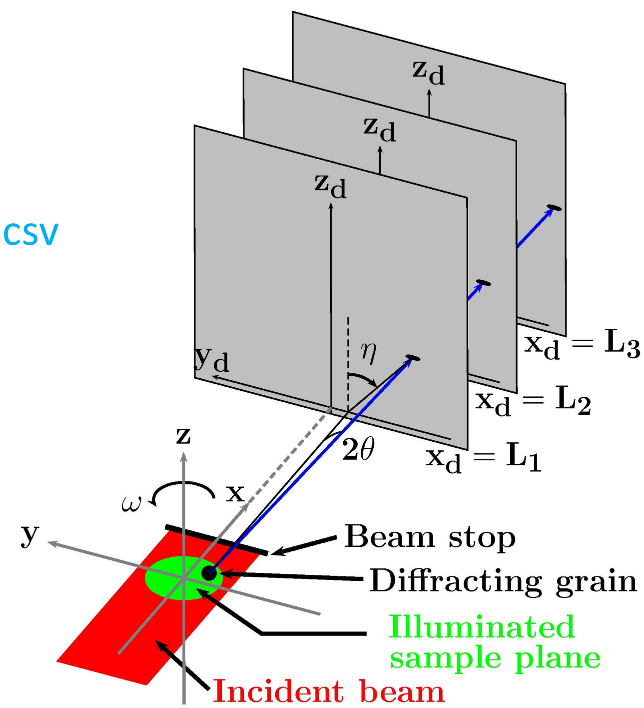
- Mostly Al, some Cu: ReducedDiameter_StagDerrJih_Cu_A-O.csv

Synchrotron Microscopy (Risø, ORNL, Suter):

- Aluminum
- Pure Ni – multiple anneal steps
- Ni doped with Bi
- Cu – multiple strain steps
- **LSHR – fatigue experiment**
- Rene88DT – fatigue experiment

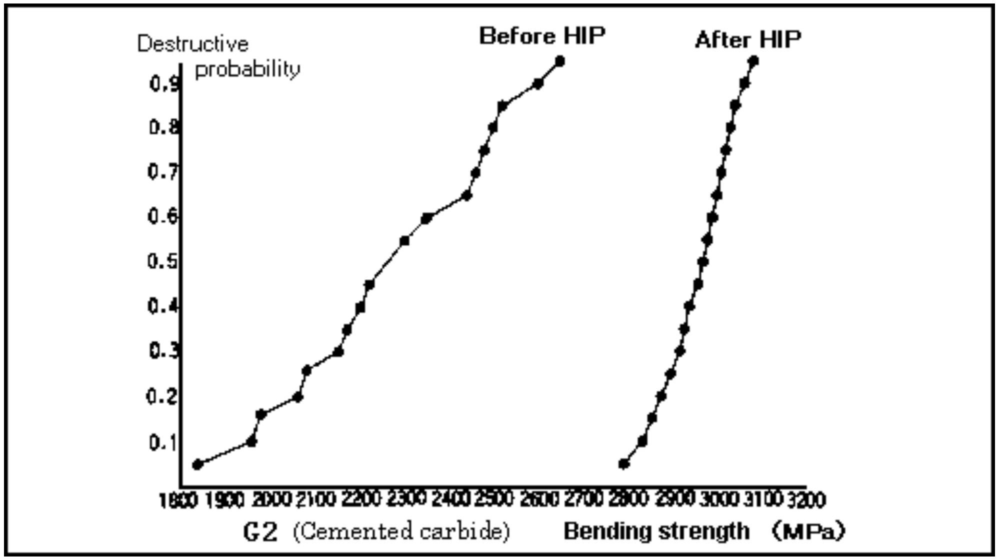
* used in Hwk 1

Files in blue are available in “datasets” on Canvas



What about Extreme Values?

Application of HIP Technology | NIPPON TUNGSTEN CO., LTD.
nittan.co.jp



- If you have ever tested the "strength" of a ceramic, you will have discovered a substantial variation in the measured value.
- The classical approach to this in materials science is to describe the brittle fracture in terms of the "weakest link" , i.e., the largest flaw in the high stress region (of the bend bar).
- This means that the average defect/flaw size is of little consequence because it is the *largest flaw* that will lead to fracture.
- A similar approach is the basis for the Griffith theory of brittle fracture.
- Therefore, we have to consider the extreme values of the population of defects.

Scatter in Fatigue Life

- Measurement of fatigue life generally exhibits substantial scatter, which has been the subject of much research. The motivation is to determine safety factors.
- There has long been a suspicion that large grain sizes play a role in initiating fatigue cracks. This motivates examination of grain size.
- One difficulty has been separating the variability in crack nucleation from that of crack growth.
- Crack growth generally obeys the Paris law with minimal scatter from microstructural variations.
- Crack nucleation, however, often occupies a substantial fraction of total life (depicted in a S-N plot).

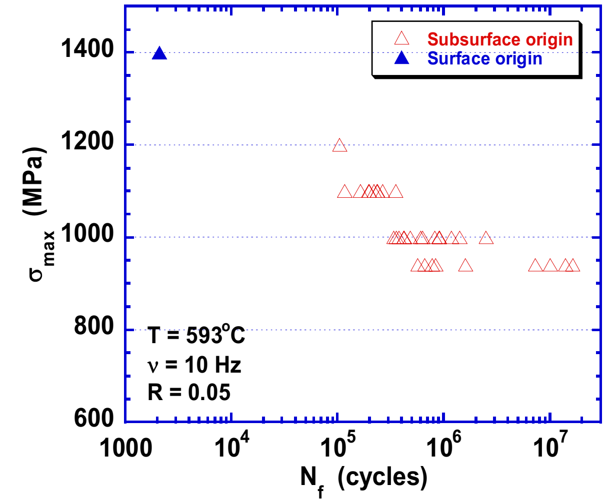


Fig 6: Fatigue life behavior of Rene` 88 DT.

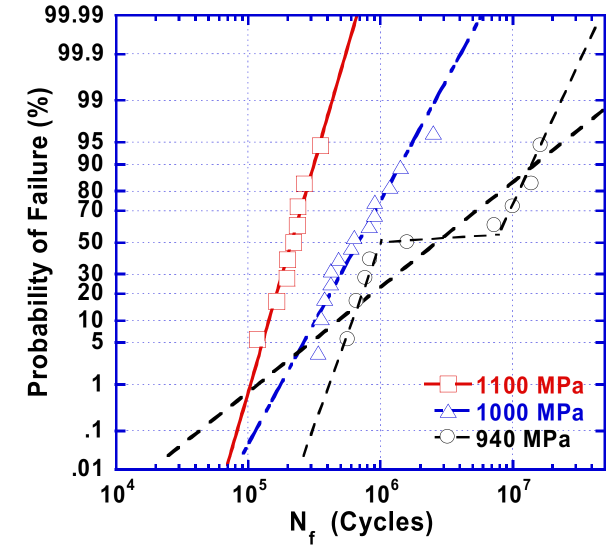


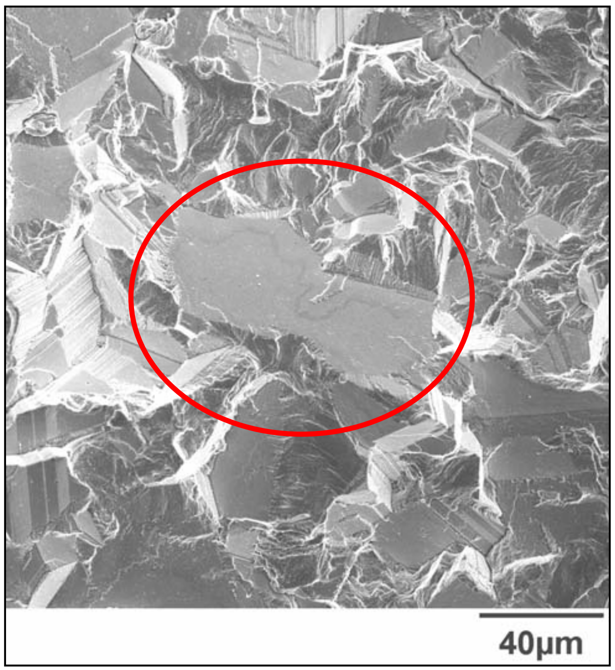
Fig 7: Plot of cumulative distribution functions (CDF) at selected stress levels

Superalloys 2004, Caton et al., "Divergence of mechanisms and the effect on the fatigue life variability of Rene88DT"

Motivation

Motivation to Incorporate Extreme Values

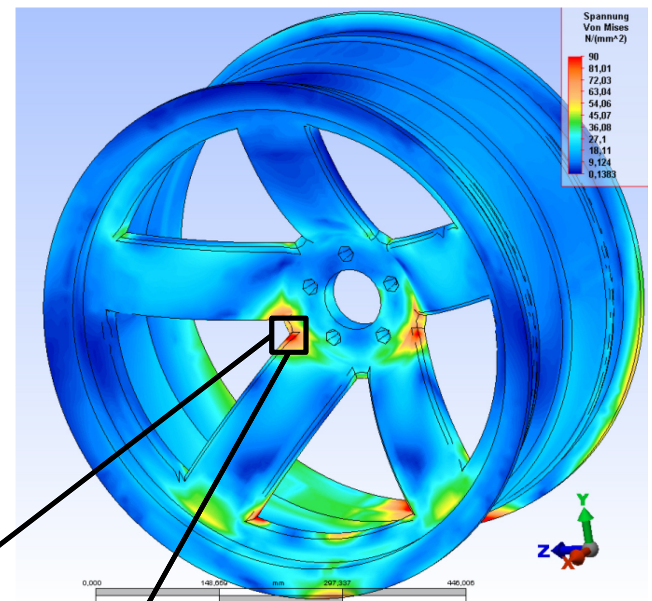
'Forget the Representative Volume Element, show me the ***Weakest*** *Volume Element' – paraphrased from Jim Williams*

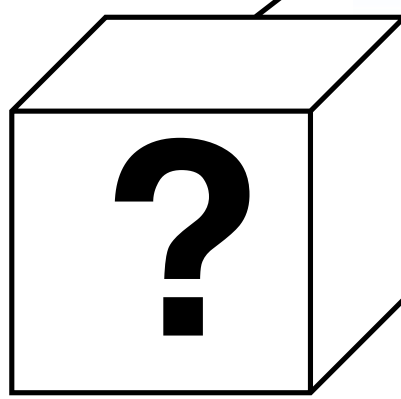


Ni-base superalloys

Fatigue crack initiation was observed in large grains oriented for slip

(Caton, Jha, et al., Superalloys 2004)





Findley et al., IMR **56**;
Larger grain sizes in superalloys lead to surface nucleation via slip

Historical: pre-Industrial Revolution

- Leonardo da Vinci

- da Vinci wrote in the 1500s that “Among cords of equal thickness, the longest is the least strong”.

- Galileo
  - Rejected da Vinci’s implication that cutting a cord or rope could make it stronger, thereby clearly thinking of it in deterministic terms (1638).

- Mariotte
  - Investigated the strength of ropes, paper, tin and described the results in statistical terms (Traité du mouvement des eaux, 1686).

Historical Note

APPLICATION OF THE THEORY OF EXTREME VALUES IN FRACTURE PROBLEMS*

BENJAMIN EPSTEIN

Coal Research Laboratory, Carnegie Institute of Technology

In this paper it is shown that the theory of extreme values is pertinent to the treatment of certain aspects of the fracture or break-down of materials used in modern technology. An attempt is made to integrate some of the results scattered through the technical literature.

J. Amer. Statistical Assoc., **43**:243, 403-412 (1948)

[2] F. T. Peirce, "Tensile Tests for Cotton Yarns V. 'The Weakest Link'—Theorems on the Strength of Long and of Composite Specimens," *Journal of the Textile Institute*, Transactions, *17*, 355 (1926).

[3] W. Weibull, "A Statistical Theory of the Strength of Materials," *Ing. Vetenskaps Akad. Handl.*, No. 151 (1939); "The Phenomenon of Rupture in Solids," *Ibid.*, No. 153 (1939). See also John Tucker, "Statistical Theory of the Effect of the Dimensions and Method of Loading upon the Modulus of Rupture of Beams," *Proceedings of the American Society of Testing Materials*, *41*, 1072 (1941).

[4] N. Davidenkow, E. Shevandin and F. Wittman, "The Influence of Size on the Brittle Strength of Steel," *Journal of Applied Mechanics*, *14*, No. 1, A63–67 (1947).

[5] T. A. Kontorova, *J. Tech. Phys. U.S.S.R.*, *10*, 886 (1940); J. I. Frenkel and T. A. Kontorova, "A Statistical Theory of the Brittle Strength of Real Crystals," *J. Phys. U.S.S.R.*, *7*, 108 (1943).

Analysis

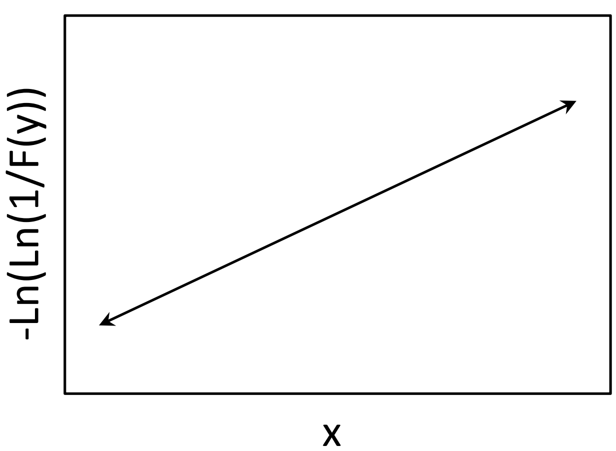
Extreme Value Analysis Methodology

There can exist generally only 3 types of asymptotic distributions for extreme values:

Type I (Gumbel)

Exponentially Distributed Tails
(e.g., Gaussian)

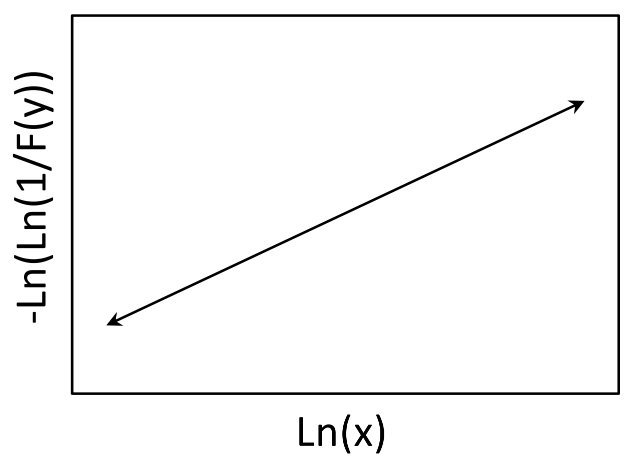
$$F_{Y_n}\left(y_n\right) = e^{-e^{-\alpha_n\left(y_n-u_n\right)}}$$



Type II (Frechet)

Polynomially Distributed Tails
(e.g., Power Law, Lognormal)

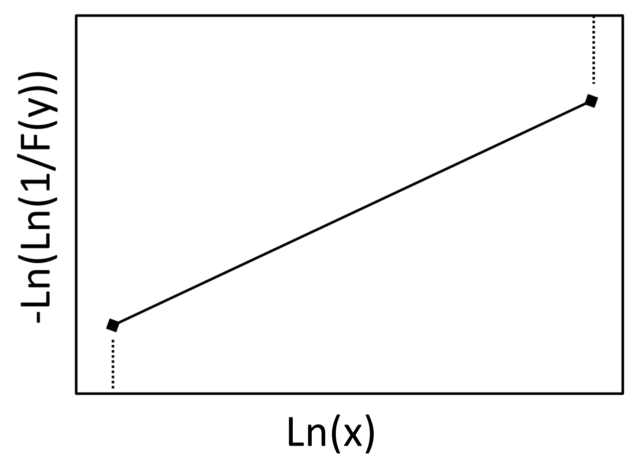
$$F_{Y_n}\left(y_n\right) = e^{-\left(\frac{v_n}{y_n}\right)^k}$$



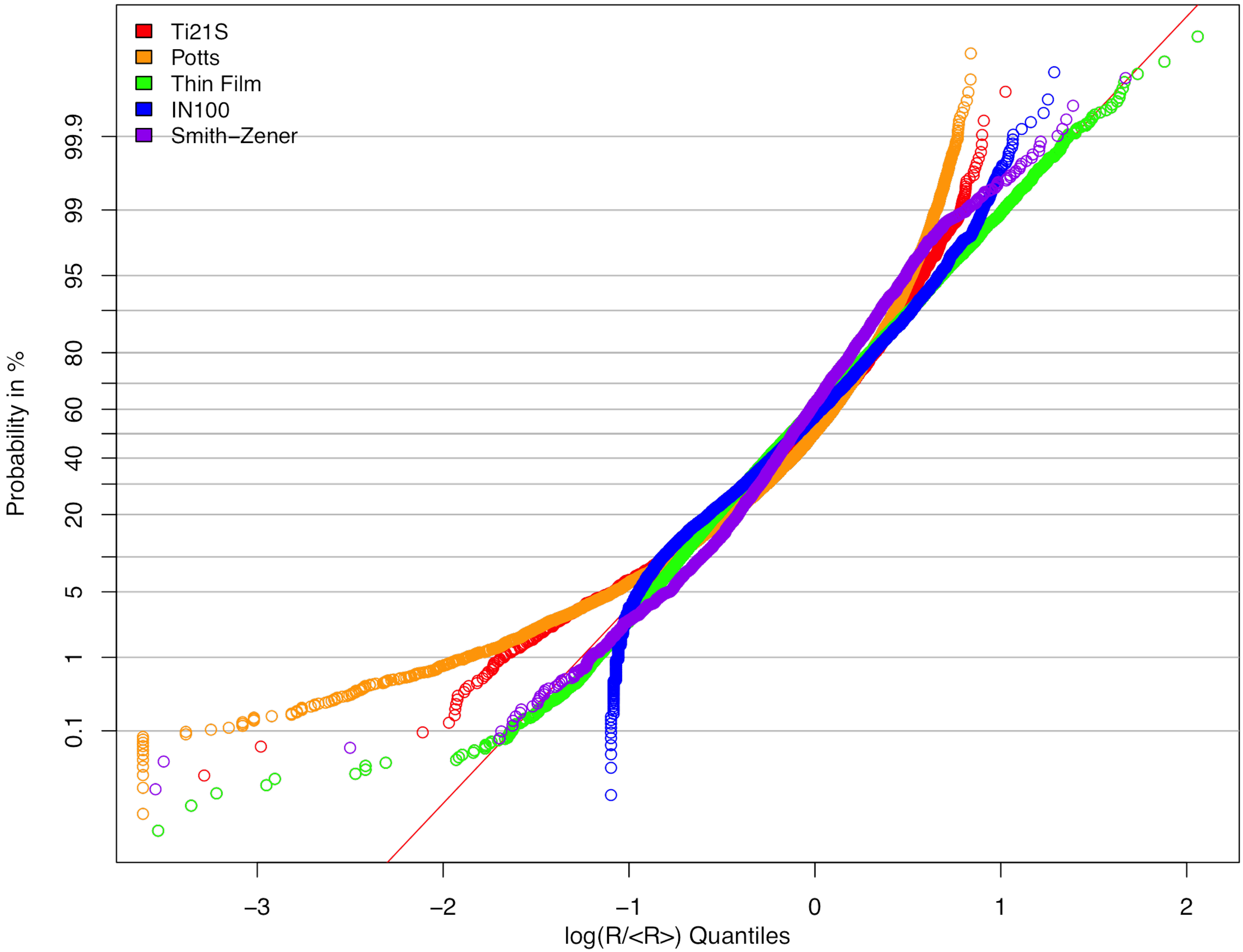
Type III (Weibull)

Polynomial Tails with Cut-Off
(e.g., LSW, Weibull, Hillert)

$$F_{Y_n}\left(y_n\right) = e^{-\left(\frac{\omega-y_n}{\omega-w_n}\right)^k}$$



Combined Probability Plot



http://www.r-project.org/

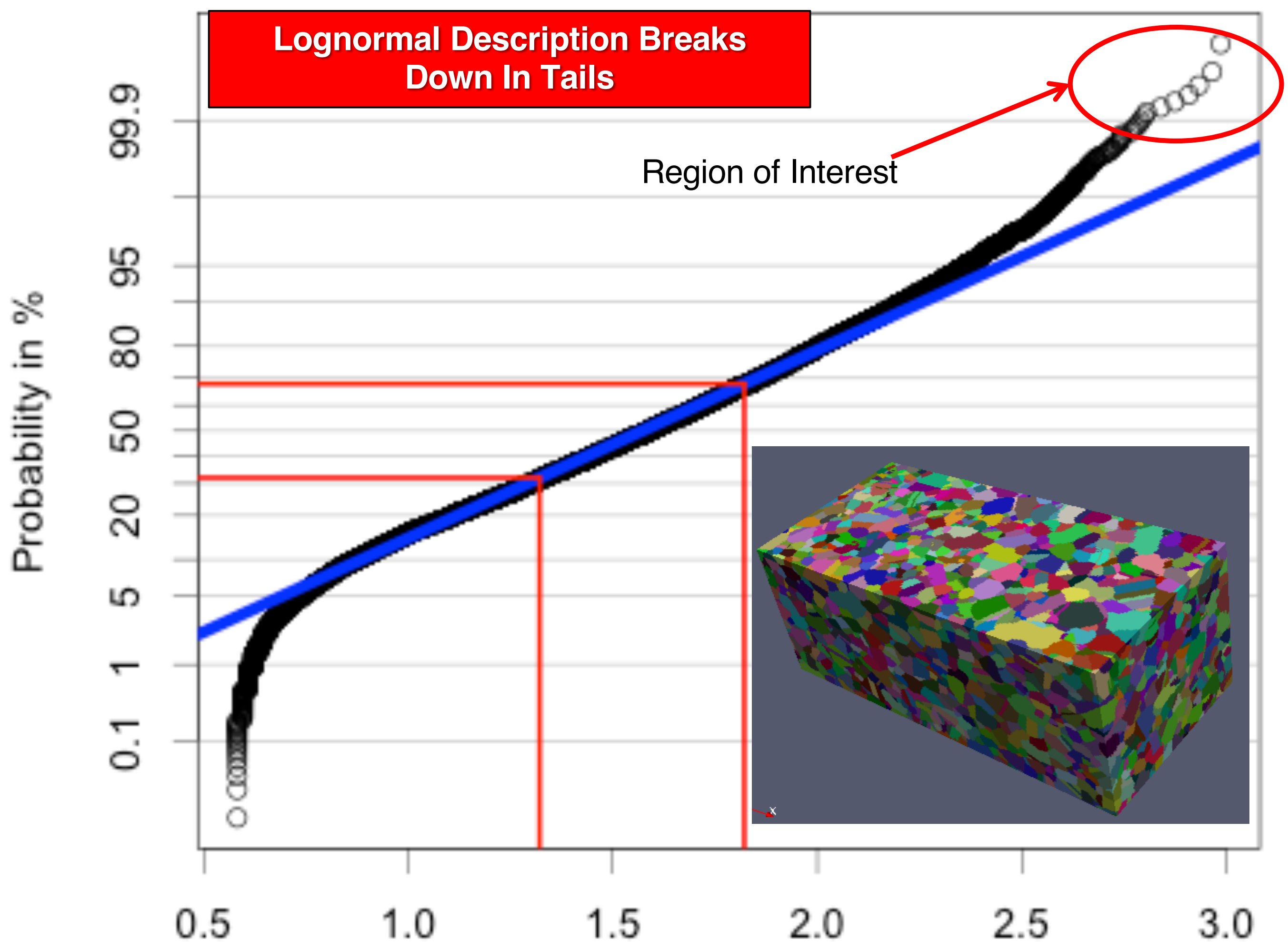
Background Motivation

Assessment of Lognormal 'Nature' of IN100 GSD

Ln(R) should be normal if R is lognormally distributed

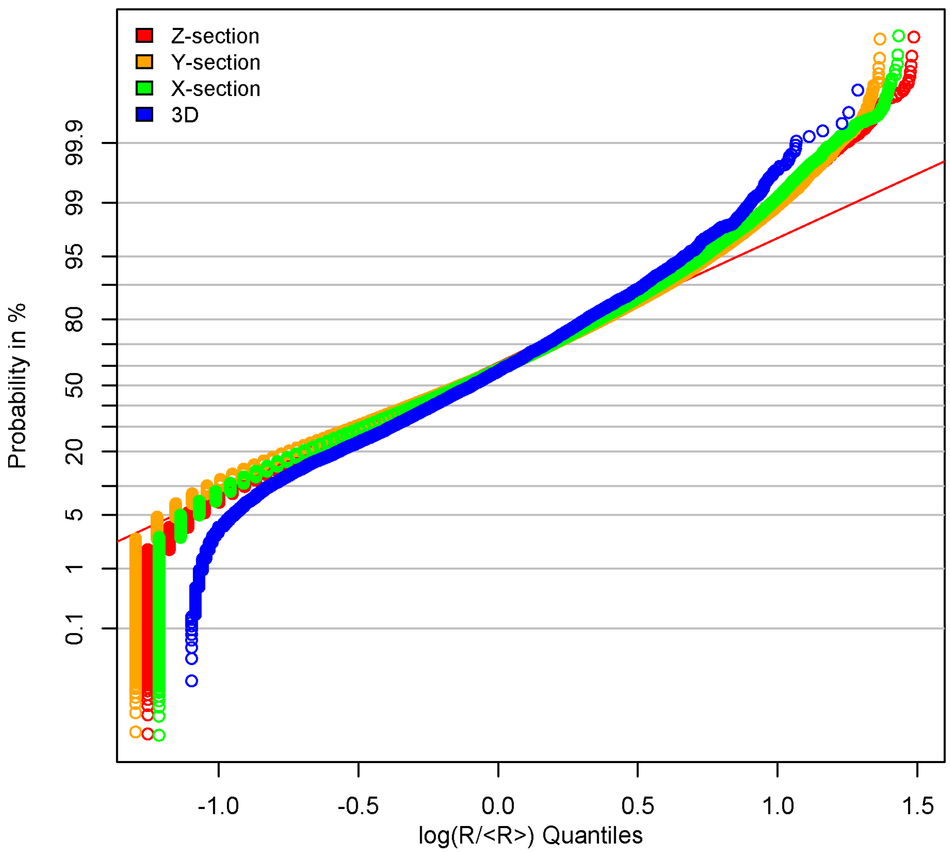
Line is normal distribution with μ and σ equal to that of Ln(R)

Values +/- 2σ from mean follow lognormal

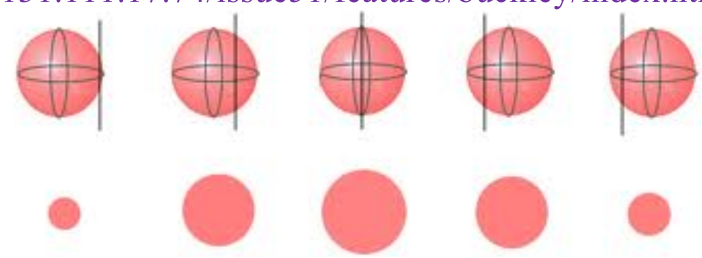


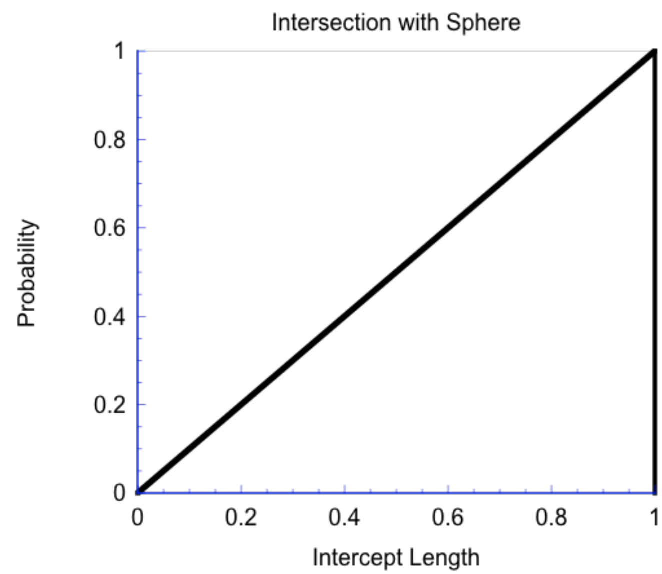
Practical 2D Grain Size Measurement

- WARNING: the following information is qualitative in nature. Significant further work required to reduce the ideas to engineering practice.
- Measurement of true 3D microstructures is necessary for validation of technique but is impractical for everyday use.
- What can we learn from standard 2D cross-sections taken from a full 3D image?



http://131.111.17.74/issue51/features/buckley/index.html





Regeneration of True Grain Size

- For grain size, stereological reconstruction of the 3D distribution is a familiar procedure, as described by Saltykov, Cahn, Fisher, others.
- What is not yet known is how reliably the upper tails can be reproduced.
- The next step is to reconstruct the 3D size distribution from the sections of a known 3D distribution and apply statistical tests for similarity against the original 3D grain size distribution.
- This was done by Tucker *et al*. (2012), *Scripta materialia*, **66**, 554-557; they showed that the upper tails could be deduced from 2D data.

Linear Regression

- We have N pairs of associated quantities, i.e., datapoints.
- One variable is taken to be the independent (explanatory, predictor) variable; the other is taken to be the dependent (response) variable. In software, the name of the first is “x” and the second is “y”.
- More generally, there may be multiple independent variables, in which case we will apply *multiple linear regression*.
- It is always a good idea to check the distribution of each variable: is it normal? Are there outliers? Boxplots or violin plots are useful here.

https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture9.pdf

https://ocw.mit.edu/courses/mathematics/18-s096-topics-in-mathematics-with-applications-in-finance-fall-2013/lecture-notes/MIT18_S096F13_lecnote6.pdf

Questions?