Statistical Analysis of Materials Data

Morning Session of Day One

Modules One and Two:

Random Variables

and their

Distributions

Session Overview

- Introduction: why this course?
- Random Variables and Distributions
- Parameters and Estimations
- Continuous and Discrete Distributions
- Measures of Central Tendency/Dispersion
- Probabilities, events, rare events & outliers
- Data presentation and display

Course Contents Overview

- Random variables, distributions, parameters
- Confidence intervals and tolerance limits
- Hypothesis Testing and Goodness of Fit
- Multivariate (bivariate) statistics:
  - Two or more data from each specimen (subject)
- Covariance, correlation and regression
- ANOVA (analysis of variance) model
- Case studies in materials data analysis.

Materials Engineering Motivation

- Why should I take this statistics course?

Several Positive Reasons

- Materials are inherently variable
- This induces performance variability
- and also materials properties variability
- Hence, need to obtain property allowables
- To extrapolate results to the population
- And to assess the variability of such allowables (estimations).

Several Statistical Concerns

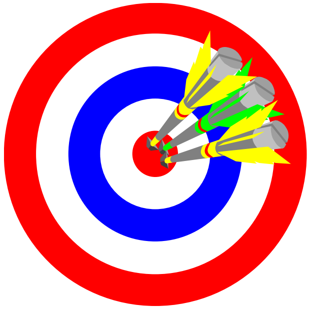
- Poor allowables produce bad designs
- Bad statistics produces poor allowables
- Lack of understanding induces bad statistics
- Poor ***statistical thinking*** hinders the understanding of statistical procedures
- Manuals and books provide how-to's
- This course stresses ***statistical thinking.***

Data and Statistical Analysis

- Data is at the center of statistical analysis
- Good data is required (or GIGO Model)
    - Garbage-in-Garbage-out
- Collecting Good Data costs money
- Collecting Good Data takes time
- Good Data also requires statistical planning of your collection effort.

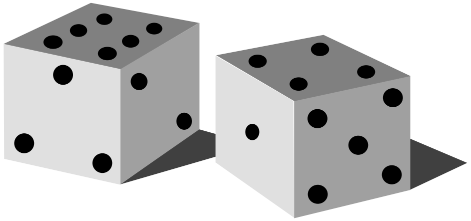
Statistical Characteristics of Data

- *Random*
  - *not gathered by convenience*
- *Representative*
  - *of the situation under study*
- *Samples*
  - *corresponding statistical procedure.*



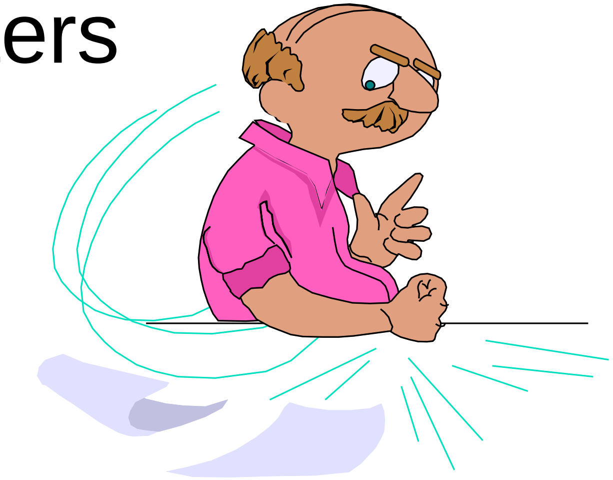
Statistical Sampling Procedures

- **Random**: obtained by chance
- **With Replacement**: returning items
- **Without Replacement**: “destroying”
- **Representative**: extrapolation
- **Homogeneous**: stratification
- **Large** Samples: CLT.
  - – Central Limit Theorem



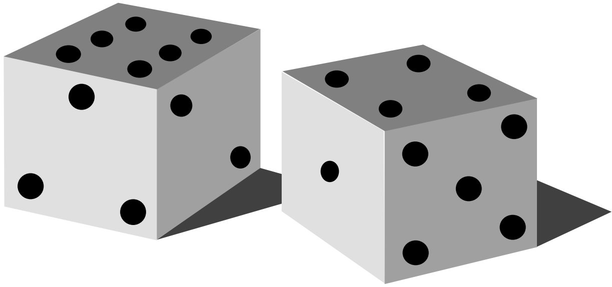
Issues on Sampling Results

- Estimators not parameters
- Variability inherent
- Sampling error
- Design of experiments
- Extrapolation to population
- Pilot studies.



Random Variables

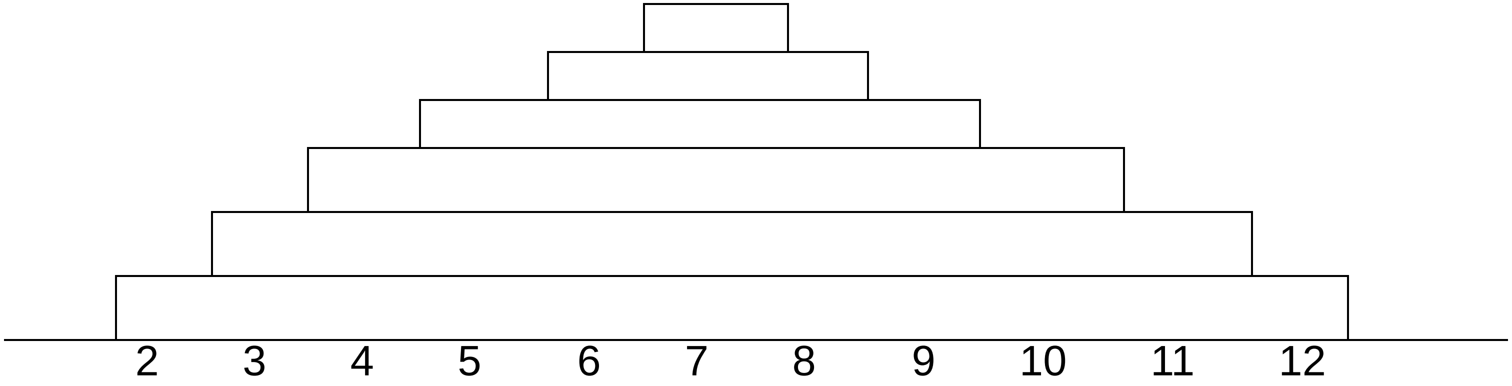
- Results of a random phenomenon
- Two or more random outcomes
- Sampling space and “events”
- Distributions as “patterns”
- Parameters of a distribution
- CDF and PDF of a distribution
- Examples: stress, strength, dice.



Example:The Sum of Two Dice

- Die one and Die two: equiprobable
- Dice Sum outcomes: 2 through 12
- Results no longer equiprobable
- Parameters of the Distribution
- Probability Mass (Density) Function
- Cumulative Mass (Distribution) Function
- Probability of Event: Rolling a Sum of ...

The Graphical Pattern



DICE	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

X is the Sum of Two Honest Dice

f(X) is the probability of two honest dice adding up to a particular value

x	f(x)
2	0.028
3	0.056
4	0.083
5	0.111
6	0.139
7	0.167
8	0.139
9	0.111
10	0.083
11	0.056
12	0.028

Examples of Probabilities

- For the Sum of Two Dice:
  - – Prob {Sum of Five or Six}
  - – Prob {Sum is Greater than Seven}
- For the Length of a Crack:
  - – Prob {Length will be less than 3 units}
  - – Prob {Length is between 1 and 3 units}
- Strength and Stress of a Material.

Measurement Scale Levels

- Qualitative
  - – Nominal (only classificatory)
  - – Ordinal (there is an order)
- Quantitative:
  - – Interval (there is a distance)
  - – Ratio (there is an absolute zero)
- Statistical methods depends on scale.

Examples of Scale Levels

- Nominal: defective or non-defective
- Ordinal: bad, poor, medium, acceptable and excellent (quality of a material)
- Interval: (Farenheit, Celsius degrees) temperatures, where the zeros are arbitrarily set
- Ratio: weight, height, mass (zero is set and meaningful).

Discrete Distributions

- Discrete sampling space
- Mass functions are “cylinders”
- Simple events have nonzero probability
- Counting rules are very important
- Permutations, combinations, etc.
- Probability as the Ratio of Favorable Cases to Total Cases.

Continuous Distributions

- Continuous sampling space
- Density functions (continuous)
- Areas under the (density) curve
- Events characterized by ranges
- Integration is important now
- Can always be "discretized"
  - – translate ratings to quality categories.

Some Distribution Characteristics

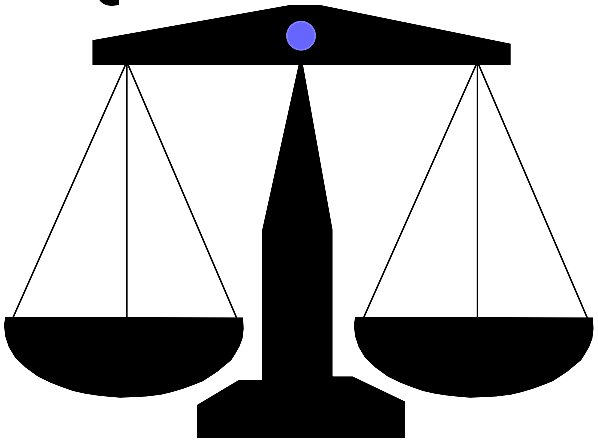
- Measures of Central Tendency
  - – characterize concentration
- Measures of Dispersion
  - – characterize variation
- Symmetry (and skewness)
- Modality (one or more modes)
- Kurtosis (flat or peaked).

Measures of Location

- Respond to the question “Where”
- Maximum, minimum values
- Mean (outcome at the center of gravity)
- Median (splits the population in halves)
- Mode (outcome with largest frequency)
- Quartiles (Q1,Q3), percentiles
  - Percent of the population left behind.

Measures of Dispersion

- Are associated with variability or “risk”
- Range = Max - Min
- Interquartile Range = Q3 - Q1
- Variance
- Standard Deviation
- Coefficient of Variation.



Other Distribution Parameters

- Shape (e.g. in Weibull)
- Scale (e.g. in the Exponential)
- Thresh-hold (minimum feasible value)
- Parameters Main Usefulness:
  - *Help describe the outcome pattern*
  - *a.k.a. the random variable distribution.*

Outliers or Rare Events

- Assume a particular distribution
- Specify all distribution parameters
- Then, under such outcome pattern
- Observe “Main Stream” events
- Detect any “Rare” event (or outlier)
- Both depend on the assumptions made
- Analyze with care! Do not stereotype!

Well-Known Distributions

- There are theoretically an infinite number of distributions
- For practical purposes: having a good approximation is enough
- Several approximations, well studied
- They exhibit good properties and fit well
- These are the “well-known distributions”
- They have been tabulated, graphed.

Population Parameters

- The mean or expected value
  - – meaning: center of gravity of pattern
- The variance
  - – meaning: dispersion about the mean
- The coefficient of variation:
  - – a “standardized variance”
- Uses of mean, variance, etc.
  - – display and comparison of patterns.

Estimations

- Obtained (reduced) from data (samples)
- Point toward their respective parameter
- Are random variables themselves
- Hence, they have a distribution
- They also have (related) parameters
- Hence, are used in statistical analyses
  - – to "recuperate" the unknown parameters.

Data Presentation

- Data provide a wealth of information
- But first, samples must be processed
  - Tabulation of the data
  - Graphical data displays
- Interpretation of these results
- Preliminary ideas (hypotheses)
- *EDA* (Exploratory Data Analysis).

Tabulation of the Data

- Organize and sort your data, then:
- Frequency/cumulative frequency tables
- The mean and standard deviation
- Five-Number descriptors:
  - – Median, quartiles (Q1,Q3), max, min
- Interpretation and comparisons to some
  - – Well-known distributions.

Example of Frequency Table

	Freq	CumFreq	Perc	CumPer
<300	3	3	0.085	0.085
300-305	5	8	0.142	0.228
305-310	6	14	0.171	0.4
310-315	8	22	0.228	0.628
315-320	6	28	0.171	0.8
320-325	4	32	0.114	0.91
325-330	3	35	0.085	1

Examples of Tabulation

Data Sample and its Descriptive Statistics

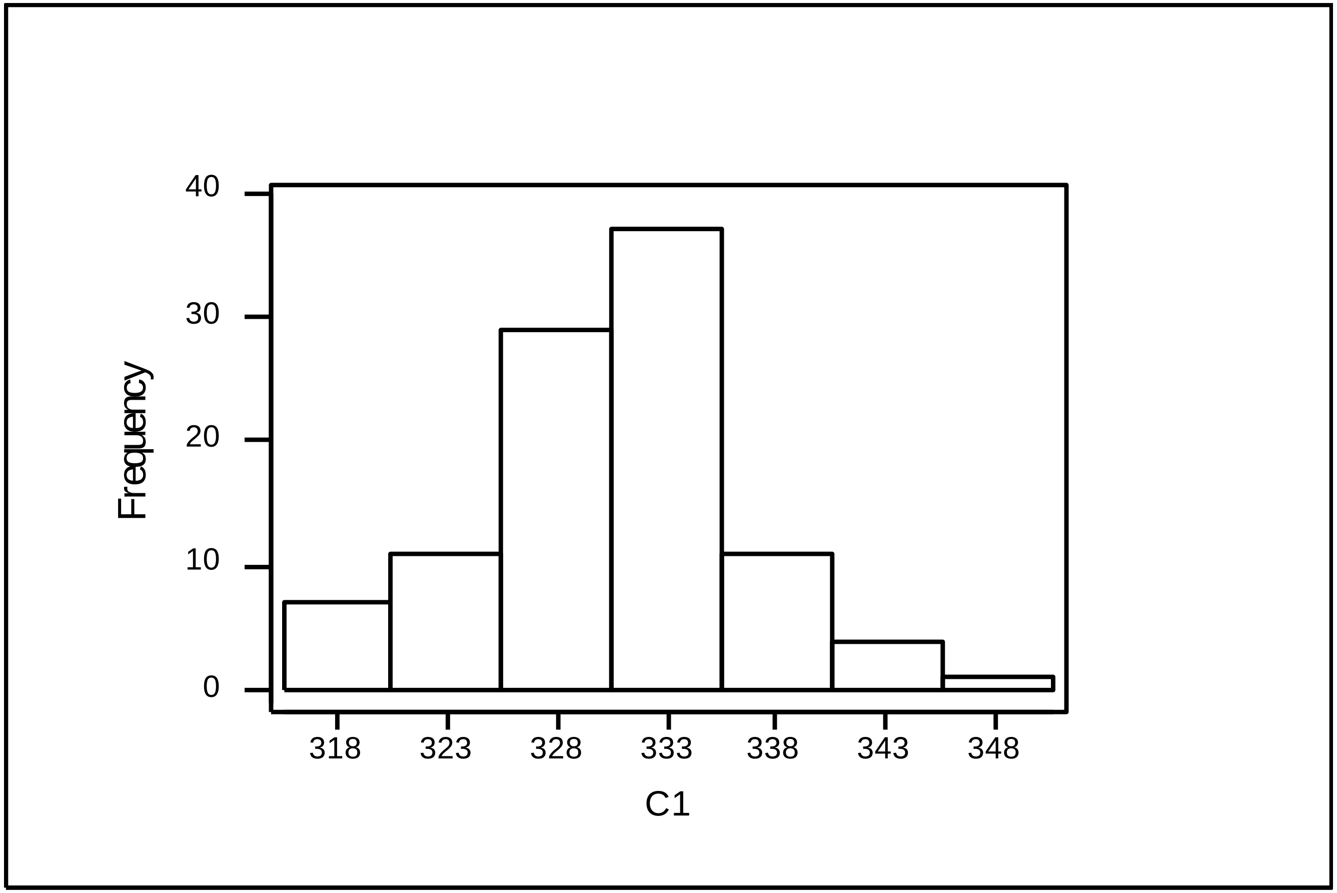
```
324.067   339.668   326.397   328.389   327.287   329.930   335.063
331.985   339.055   331.064   329.806   320.234   327.432   328.192
322.861   339.008   332.106   331.224   338.830   330.281
```

```
               N     MEAN   STDEV      MIN      MAX      Q1      Q3
uniform       20   330.64    5.51   320.23   339.67   327.32  334.32
```

Graphical Presentation

- Histograms, stem-and-leaf displays
- Box and whiskers plots and outliers
- Cumulative frequency plots
- Graphical parameter estimation
- Reference Distributions and Outliers
- Interpretation and Outlier Detection.

Histogram Graphical Display



Example of Data Display

```
 1    32 0
 2    32 2
 3    32 4
 6    32 677
10    32 8899
10    33 0111
 6    33 2
 5    33 5
 4    33
 4    33 8999
```

Stem-and-Leaf Tabular and Graphical Representation (compare to the histograms)

The Box-and-Whiskers Plot

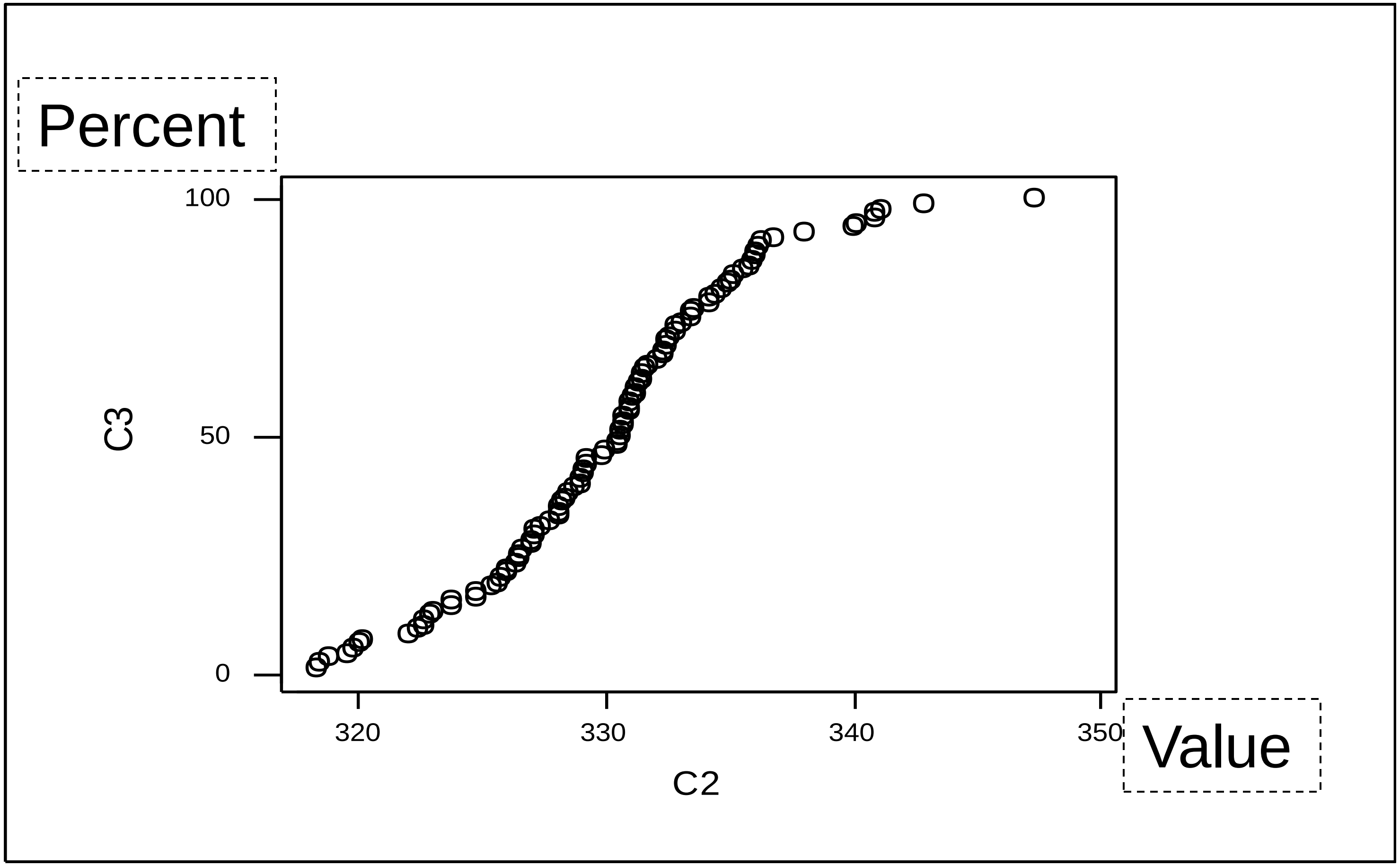
Min, Lower Quart, Median, Upper Quart, Max

```
                      -----------------

     -----------------I      +        I---------------

                      -----------------

----+---------+---------+---------+---------+---------uniform

320.0     324.0     328.0     332.0     336.0     340.0
```

The Dot-Plot Graph

```
Each dot represents 5 points

                                       .
                                   . ..:::
                                   :.:::::.
                                 . ::::::::
                                 ::::::::::.
                               ..:::::::::::.
                              .::::::::::::::
                          .::::::::::::::::::.
          ..  . ......::::::::::::::::::::::::::.
    +---------+---------+---------+---------+---------+----
  288       300       312       324       336       348
```

The Cumulative Plot



Summary

- Random Variables:
  - – the outcome/result of random phenomena
- Statistical Distributions:
  - – describe their outcome patterns
- Location, Dispersion, etc. Parameters:
  - – characterize the distribution patterns
- Statistics is about identifying them!

Questions