Data analysis and interpretation

Lecture 2

Based on Sheldon M Ross

Chebyshev inequality

- Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data set
- Assume $s > 0$
- Chebyeshev inequality states that for any $k \geq 1$
- greater than $100\left(1 - 1/k^2\right)$ percent of the data lies within the interval $\bar{x} - ks$ to $\bar{x} + ks$

Chebyshev inequality

- Note the word greater than
- This is the lower limit
- It can be sharpened based on the specification of the data set
- Meaning of the statement?

Proof

$$(n-1)s^2 = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)$$

$$(n-1)s^2 = \sum_{i \in S_k} \left(x_i - \bar{x}\right)^2 + \sum_{i \notin S_k} \left(x_i - \bar{x}\right)^2$$

Proof ...

- Both the terms are positive (since they are squared)
- Hence, if you consider only one term, then, it should be less than the sum; the worst case is where there is equality; in that case one of the terms should be zero.

Proof ...

$$(n-1)s^2 \geq \sum_{i \notin S_k} (x_i - \bar{x})^2$$

Note, by definition, if i does not belong to the set S_k, then,

$$(x_i - \bar{x})^2 \geq k^2 s^2$$

There are n-$|S_k|$ terms in the summation.

Proof ...

$$(n-1)s^2 \geq k^2 s^2 (n-|S_k|)$$

Divide both sides by *N* k^2 s^2, to obtain

$$\frac{(n-1)}{nk^2} \geq \frac{(n-|S_k|)}{n}$$

Proof ...

$$\frac{(n-1)}{nk^2} \geq \frac{(n-|S_k|)}{n}$$

$$\frac{(n-1)}{nk^2} \geq 1-\frac{|S_k|}{n}$$

Proof ...

$$\frac{(n-1)}{nk^2} \geq 1 - \frac{|S_k|}{n}$$

$$\frac{|S_k|}{n} \geq 1 - \frac{(n-1)}{nk^2}$$

Proof ...

$$\frac{|S_k|}{n} \geq 1 - \frac{(n-1)}{nk^2}$$

$$\frac{|S_k|}{n} > 1 - \frac{1}{k^2}$$

I think! Check!!

Chebyshev inequality

- Note: all we assumed is that we know the mean and standard deviation for the data
- Very powerful
- Also, only the limit -- can be sharpened if you know more information

Chebyshev inequality ...

- Suppose we are interested in the fraction of data values that exceed the sample mean by at least k sample standard deviations, where k is positive
- By Chebyshev: $$\frac{N(k)}{n} \leqslant \frac{1}{k^2}$$

Stronger statement: one-sided Chebyshev inequality

- Let $\bar{x}$ and S be the sample mean and sample standard deviation of a data set
- Assume $s > 0$
- If N(k) are the number of data points which are outside ks from the mean, then, for any k > 0,

$$\frac{N(k)}{n} \leqslant \frac{1}{k^2}$$

Proof

$$y_i = x_i - \bar{x}, i = 1, 2, \ldots n$$

For any b > 0,

$$\sum_{i=1}^{n} (y_i + b)^2 \geq \sum_{i: y_i \geq ks} (y_i + b)^2$$

Proof ...

$$\sum_{i=1}^{n} \left(y_i + b \right)^2 \geq \sum_{i:\, y_i \geq ks} \left(y_i + b \right)^2$$

$$\sum_{i=1}^{n} \left(y_i + b \right)^2 \geq \sum_{i:\, y_i \geq ks} \left(ks + b \right)^2$$

Proof ...

$$\sum_{i=1}^{n} \left(y_i + b \right)^2 \geq \sum_{i:\, y_i \geq ks} \left(ks + b \right)^2$$

$$\sum_{i=1}^{n} \left(y_i + b \right)^2 \geq N\left(k \right)\left(ks + b \right)^2$$

Proof ...

$$\sum_{i=1}^{n} (y_i + b)^2 = (n-1)s^2 + nb^2$$

Why? Home work. Hint: expand and do the sums individually. And, sum y_i is zero.

Proof ...

$$\sum_{i=1}^{n} (y_i + b)^2 \geq N(k)(ks + b)^2$$

$$(n-1)s^2 + nb^2 \geq N(k)(ks + b)^2$$

Proof ...

$$(n-1)s^2+nb^2 \geq N(k)(ks+b)^2$$

$$N(k) \leqslant \frac{(n-1)s^2+nb^2}{(ks+b)^2}$$

Proof ...

$$N(k) \leqslant \frac{(n-1)s^2 + n b^2}{(ks+b)^2}$$

$$N(k) < \frac{ns^2 + n b^2}{(ks+b)^2}$$

Why? Check!!

Proof ...

$$N(k) < \frac{ns^2 + nb^2}{(ks+b)^2}$$

$$\frac{N(k)}{n} < \frac{s^2 + b^2}{(ks+b)^2}$$

Proof ...

$$\frac{N(k)}{n} < \frac{s^2 + b^2}{(ks + b)^2}$$

This is valid for all b. Choose b = s/k

$$\frac{N(k)}{n} < \frac{s^2 + (s/k)^2}{(ks + (s/k))^2}$$

Proof ...

Multiply and divide the RHS by k^2/s^2:

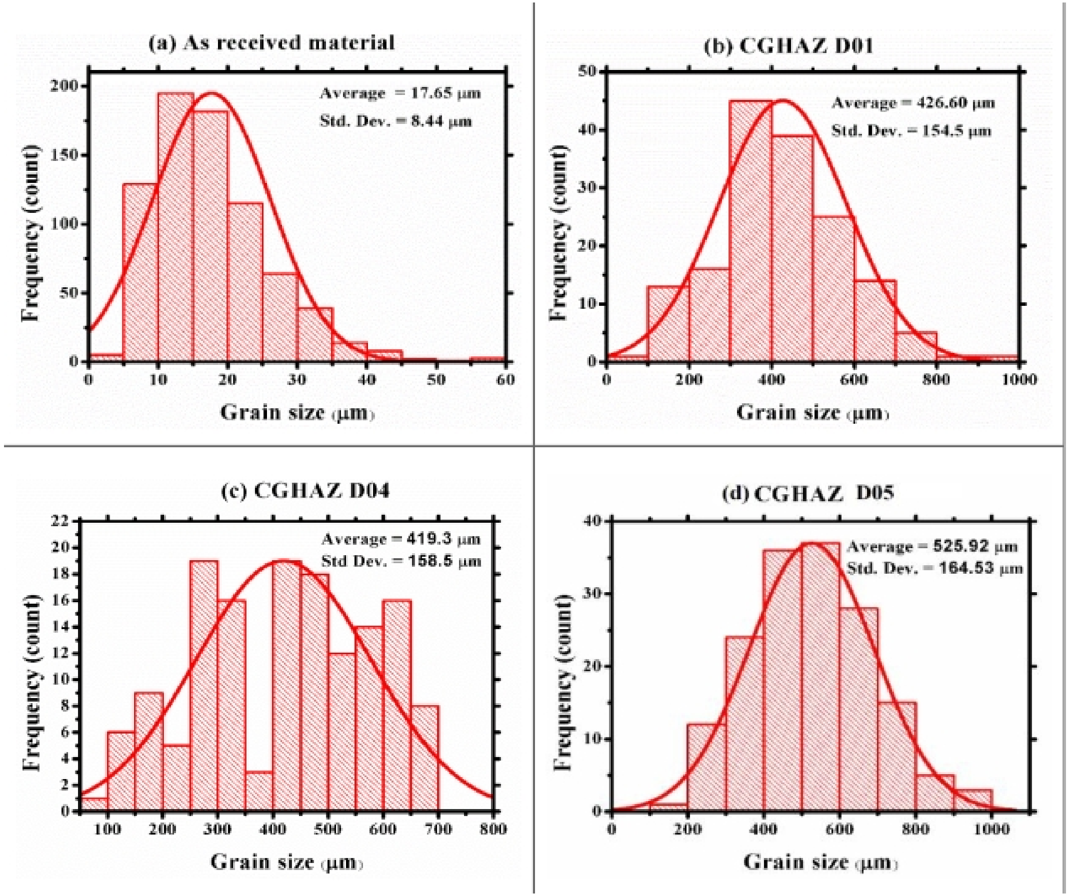
$$\frac{N(k)}{n} < \frac{\left(s^2 + (s/k)^2\right)\left(\dfrac{k^2}{s^2}\right)}{\left(ks + (s/k)\right)^2\left(\dfrac{k^2}{s^2}\right)}$$

Proof ...

$$\frac{N(k)}{n} < \frac{(k^2+1)}{(k^2+1)^2}$$

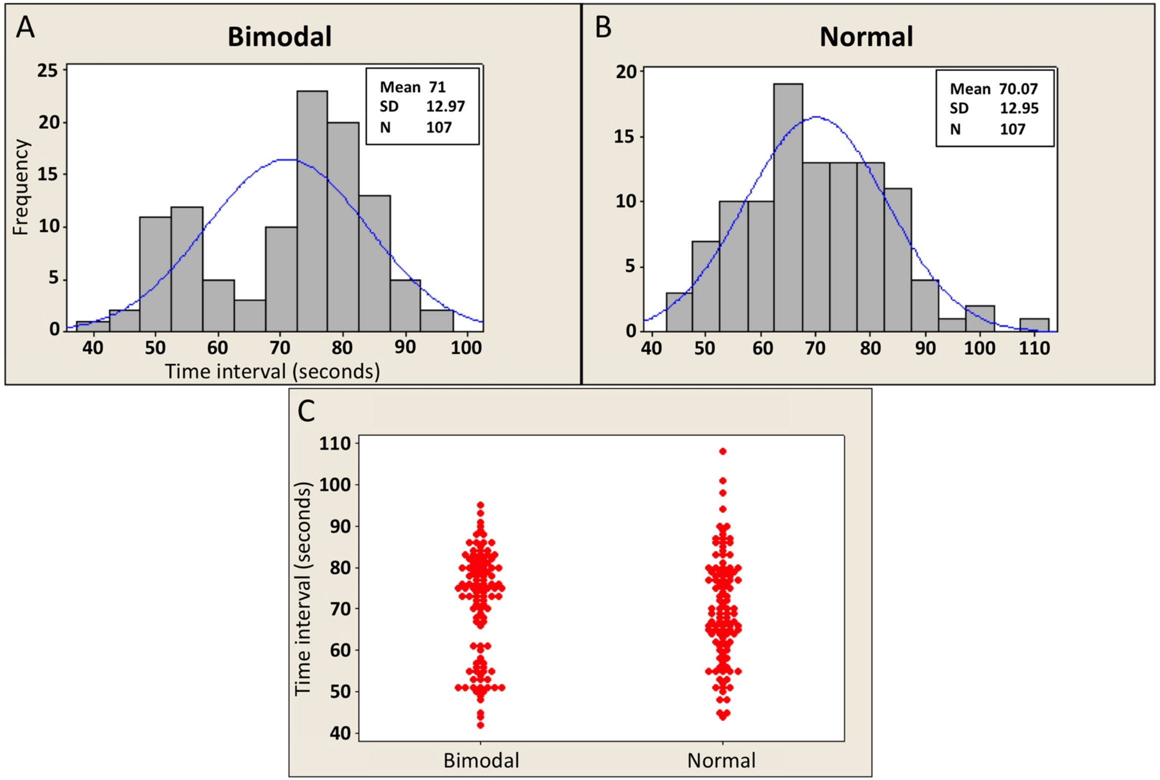
$$\frac{N(k)}{n} < \frac{(1)}{(k^2+1)}$$

Normal distribution



M Akhtar, NIT Warangal, Thesis: grain size distribution during welding (from ResearchGate)

Bi-modal distribution



D S Fay and K Gerow, A biologist's guide to statistical thinking and analysis, *Wormbook*

Normal distribution

- If the data set is nearly normal,
  68, 95 and 99.7 percent of the obversations lie within
  1, 2 and 3 standard deviations from the mean, respectively.

Paired data set

- Suppose I buy some ice-cream: note the flavour
- Note also how much time it takes for me to start my car for every time I buy the ice-cream
- These two data sets form a paired data set
- My car does not like anything other than vanilla flavour!!

Correlation coefficient

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) S_x S_y}$$

Correlation coefficient

$$r = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}}$$

Properties of r

- Lies between -1 and +1
- If $y_i = a + b\ x_i$ for constants a and b >0, then r=1
- If $y_i = a + b\ x_i$ for constants a and b <0, then r=-1
- If r is the correlation coefficient for (xi,yi), it is also the correlation coefficient for $a + b\ x_i$ and $c + d\ y_i$ provided both b and d are both positive or both negative. (Dimensions of measurement does not matter)
- Correlation is association and not causation!