Data Analysis II

CU- Boulder

CHEM-4181

Instrumental Analysis Laboratory

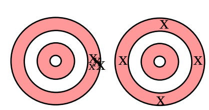
Prof. Jose-Luis Jimenez

Spring 2007

Lecture will be posted on course web page – based on lab manual, Skoog, web links

Summary of Last Lecture

- *Treat data in your lab reports and student choice exp. in a professional way*
- Topics covered in lecture I
  - Significant figures
  - Precision vs. accuracy
  - Errors
    - E_a and RE
    - Gross errors -> outliers
    - Random errors
      - Treat with statistics (Gaussian distribution)
    - Systematic errors
      - Identify and get rid of them
- Today: treatment of random errors & Excel



Population and Sample Mean

- Sample Mean ($\bar{x}$)
  - Average of a finite set of data
  - In general not the same as μ, because of finite error
  - AVERAGE() in Excel

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Population Mean (μ)
  - Also "limiting mean"
  - It is the true value of the quantity being measured

$$\mu = \lim_{N\to\infty} \frac{\sum_{i=1}^{N} x_i}{N}$$

Standard Deviation and Variance I

- Population Standard Deviation (σ)
  - Measure of the precision of a population of data
  - STDEVP() in Excel

$$\sigma = \sqrt{\lim_{N\to\infty} \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

- Variance (σ^2)
  - Std. dev. has same units of x, variance as units of x^2
  - Variance from different effects is often additive
    - $\sigma^2 = \sigma_1^2+\sigma_2^2+\sigma_3^2+\ldots$ (INDEPENDENT effects)
    - Std. Dev. is not!
  - VAR() in Excel

Standard Deviation and Variance II

- Sample Standard Deviation (s)
  - s instead of σ
  - $\bar{x}$ instead of μ
  - (N-1) instead of N
    - "Number of degrees of freedom", $\nu = N-1$
      - Because $\bar{x}$ is used in the calculation, only $N-1$ values are independent, the last one can be calculated from the mean and the other values
  - STDEV() in Excel

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

RSD and CV

- Relative Standard Deviation (RSD)
  - Often more informative than absolute SDs

$$RSD = \frac{s}{\bar{x}} \cdot 10^z$$

  - z = 2 => percent
  - z = 3 => ppth
- Coefficient of Variation (CV)
  - RSD expressed as a percent

$$CV = \frac{s}{\bar{x}} \cdot 10^2$$

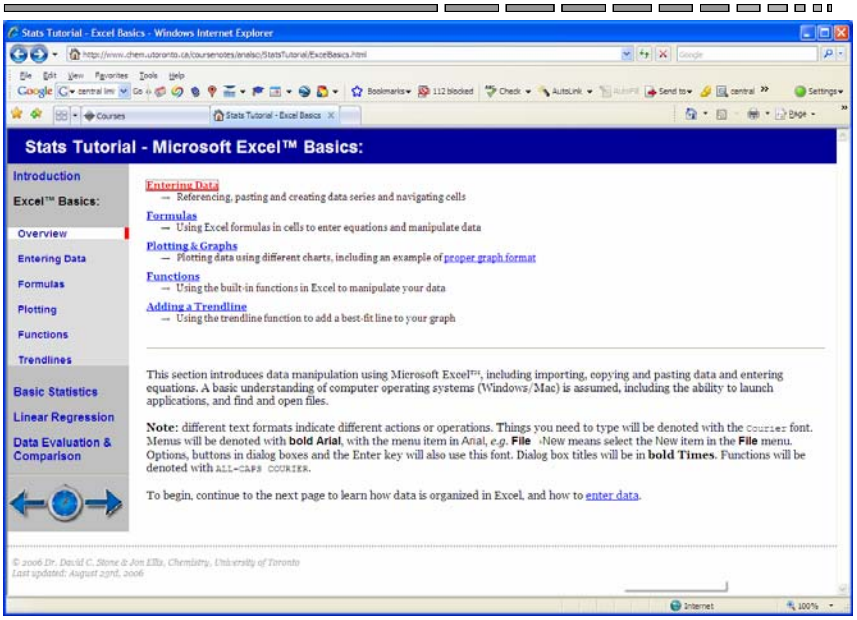
Standard Error of the Mean

- Standard deviation
  - estimate of the probable error of a single measurement
- Standard error of the mean
  - Estimate of the probable error of the mean of N measurements

$$\sigma_m = \frac{\sigma}{\sqrt{N}} \qquad s_m = \frac{s}{\sqrt{N}}$$

- More generally
  - The mean of N measurements has a distribution $N(\mu,\sigma_m^2)$
  - This is true in the limit even if error is NOT Gaussian
    - "Central limit theorem" of probability

Excel Tutorial – Part 1



http://www.chem.utoronto.ca/coursenotes/analsci/StatsTutorial/ExcelBasics.html

Entering Data

- In Excel data are entered in cells
  - Cells can be empty or contain data or formulas
  - Every cell has coordinates
    - A1, B12, etc.
    - Absolute coordinates: $A1 or A$1 or A1
      - Very important distinctions!

CQ: Do you know how to use relative and absolute references in Excel?
A. yes
B. a little
C. no

- Demo: pasting a series of data
  - Useful to create regularly spaced data

Formulas and Equations I

- Numerical operators:

Task	Operator	Example	Result
Multiplication	*	2*3	6
Division	/	4/2	2
Exponent	^	2^3	8
Order of Operations	(..)	2*3+5 or 2*(3+5)	11 or 16
Power of ten	e or E	3.2e+2 or 3.2e-2	320 or 0.032

CQ: 10e4 in computer notation equals:
A. 1,000
B. 10,000
C. 100,000
D. Neither
E. I don’t know

Formulas and Equations II

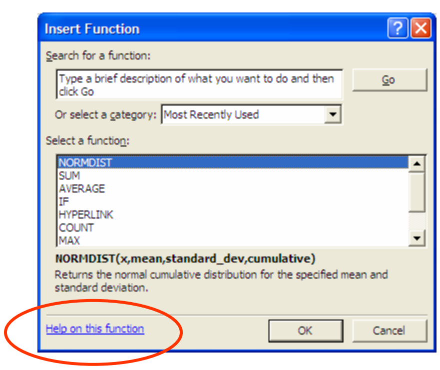
- Enter a formula which is calculated based on other cells
- Drag or Copy / Paste
  - Note that the result is different if you use absolute or relative references

C16 = =B16*2+5

	A	B	C	D
1	1	0	5	
2	2	0.1	5.2	
3	3	0.2	5.4	
4	4	0.3	5.6	
5	5	0.4	5.8	
6	6	0.5	6	
7	7	0.6	6.2	
8	8	0.7	6.4	
9	9	0.8	6.6	
10	10	0.9	6.8	
11		1	7	
12		1.1	7.2	
13		1.2	7.4	
14		1.3	7.6	
15		1.4	7.8	
16		1.5	8	
17				
18				

Pre-Programmed Functions

- Excel has lots of pre-programmed functions
  - E.g. normal distribution
  - Click f_x symbol to get a menu
  - Also look up the help files



Useful Pre-Programmed Functions

- AVERAGE()
- STDDEV()
- STDEVP()
- MEDIAN()
- MAX()
- MIN()
- VAR()
- NORMDIST()
- NORMDISTINV()
- TDIST()
- TTEST()

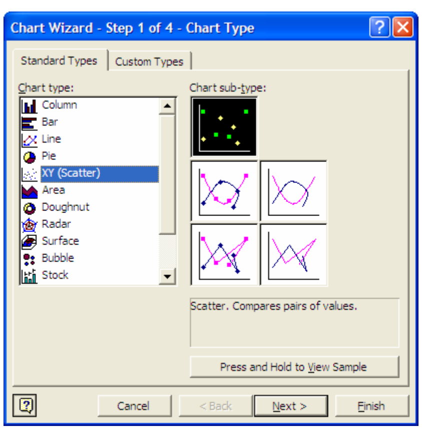
CQ: I have used

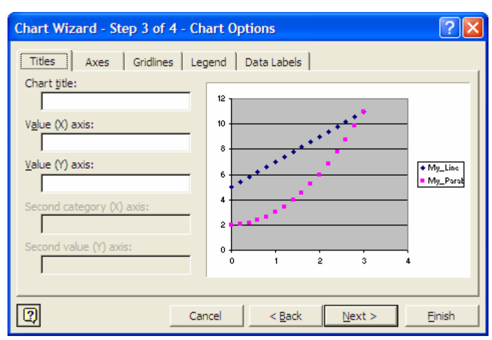
A. all of these
B. most of these
C. a few of these
D. none of these
E. what are these?

These are only a few of the statistical functions, there are lots more!

Plotting in Excel

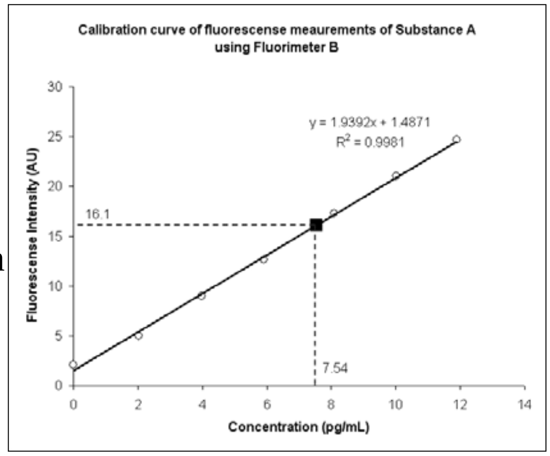
- Select data
- Choose Insert -> Chart





Proper Graph Formatting for Reports

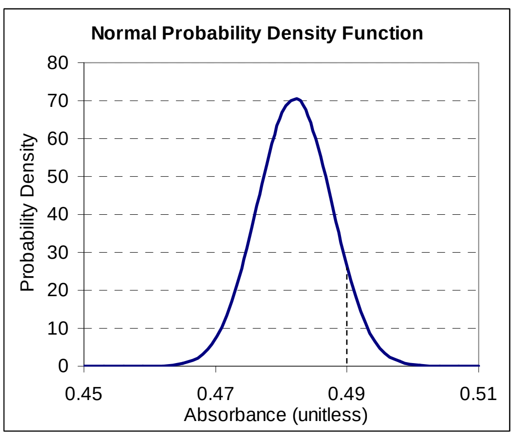
- Label axes w/ Units (AU or arb. Units if needed)
- Independent variable on X-axis
- Dependent variable on Y-axis.
- Scatter (not line) plot
- Add a regression line (if appropriate)
- Descriptive title
- *Remember Sigfigs!*



- Do not plot too many data sets on single graph - make multiple graphs instead

The Normal Error Curve

- Random errors are often distributed according to the normal error law



CQ: the probability that the absorbance is exactly 0.49 is:

A. 0.49
B. 25
C. infinity
D. zero
E. I don't know

The Normal Error Law I

- The fraction of a population of observations whose values are between x and $x+dx$ is:

$$\frac{dN}{N} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx$$

- I.e. the probability that an observation is between x and $x+dx$ *is:*

$$P(x, x+dx) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx$$

- *Probability density:* $$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

The Normal Error Law II

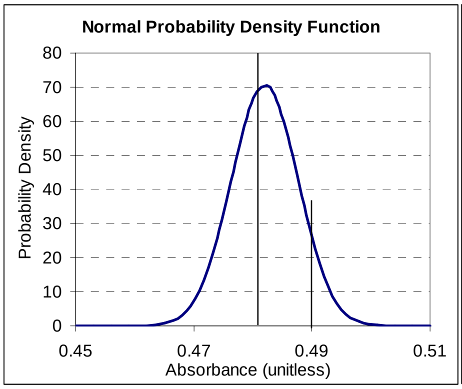
- Cumulative probability
  - The probability that x has a value between x_1 *and* x_2 is:

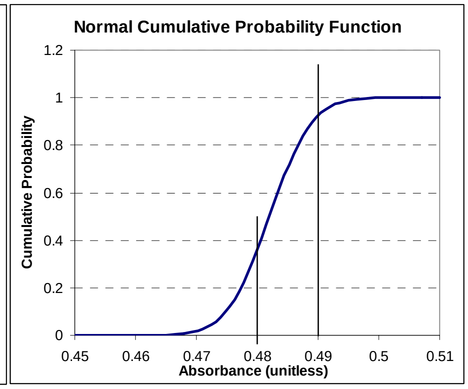
$$P(x_1, x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx$$

- Normalized distribution: $$z = \frac{x-\mu}{\sigma}$$
- In Excel
  - Can type the whole formula (prone to errors)
  - NORMDIST(x,μ,σ,*FALSE) for PDF*
    - *E.g. =NORMDIST(3,1,0.23,FALSE)*
  - NORMDIST(x,μ,σ,*TRUE) for PDF*

PDF vs CDF

- Probability Density vs. Cumulative Probability



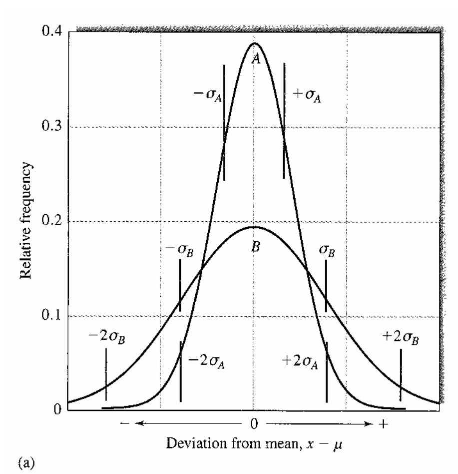


CQ: the probability that the absorbance is between 0.48 and 0.49 is:

A. zero
B. It is not defined
C. 0.37
D. 0.56
E. I don't know

Normal Error Curves

With units:



Normalized:

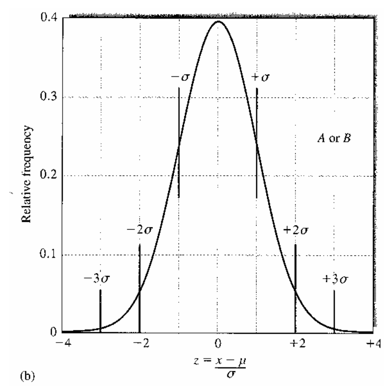
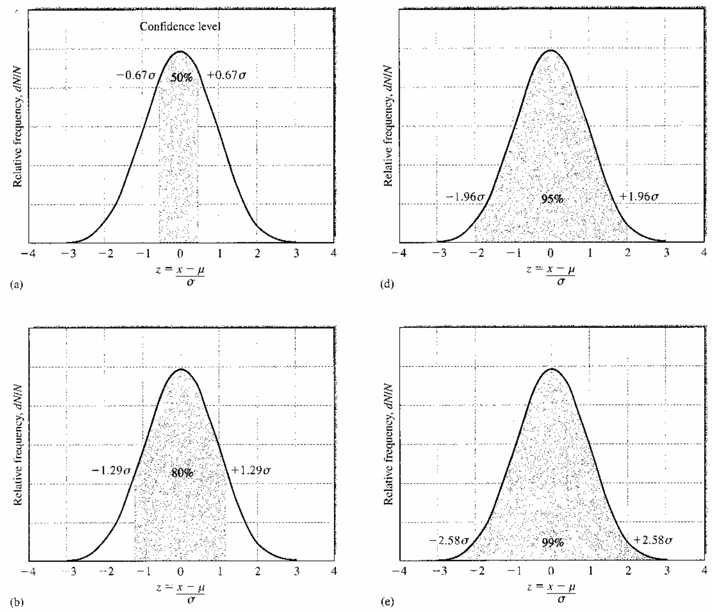


FIGURE a1-3 Normal error curves. The standard deviation of B is twice that of A; that is, $\sigma_B = 2\sigma_A$. (a) The abscissa is the deviation from the mean in the units of the measurement. (b) The abscissa is the deviation from the mean relative to σ. Thus, A and B produce identical curves when the abscissa is $z = (x - \mu)/\sigma$.

Areas under Gaussian Curve



Also know that +/- 1σ has 67%

Q: calculation w/ Excel?

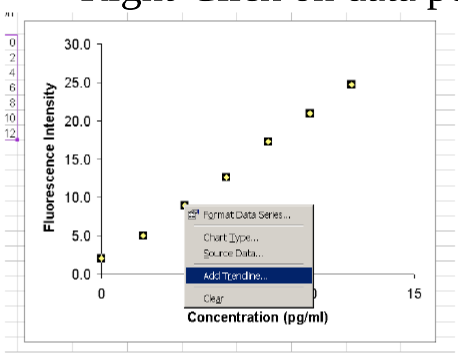
Linear Regression in Excel I

- Easy way (in graph)
- More complex way & more information (Analysis ToolPak)
- Example: calibration curve for fluorescence
- Input data from table:

Fluorescence Intensities	**Concentration (pg/ml)**
2.1	0
5.0	2
9.0	4
12.6	6
17.3	8
21.0	10
24.7	12

Linear Regression in Excel II

- Right-Click on data point:



Steps 2 & 3: