Data Analysis

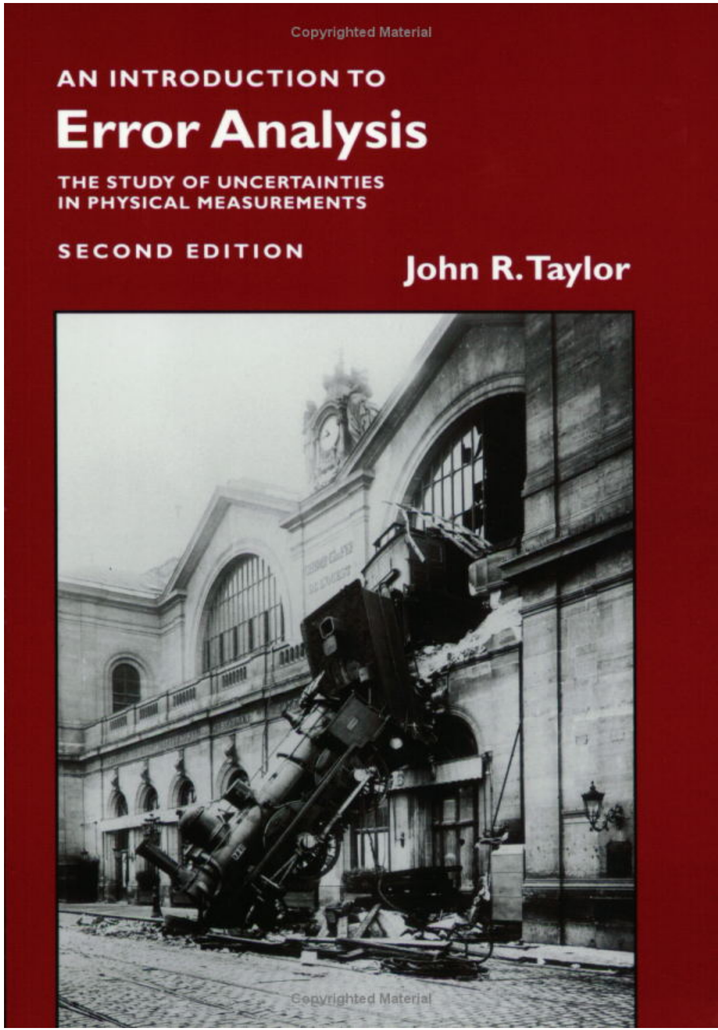
PHY451

September 10, 2014

MICHIGAN STATE UNIVERSITY

References

- "An Introduction to Error Analysis, The Study of Uncertainties in physical measurements", 2nd edition, John. R. Taylor, 1997.
- http://www.lon-capa.org/~mmp/labs/error
- Class Website
  - http://www.pa.msu.edu/courses/PHY451/
  - Some materials pass word protected due to copyright issues
  - Pass word wuli451!



MICHIGAN STATE UNIVERSITY

Probability Distribution

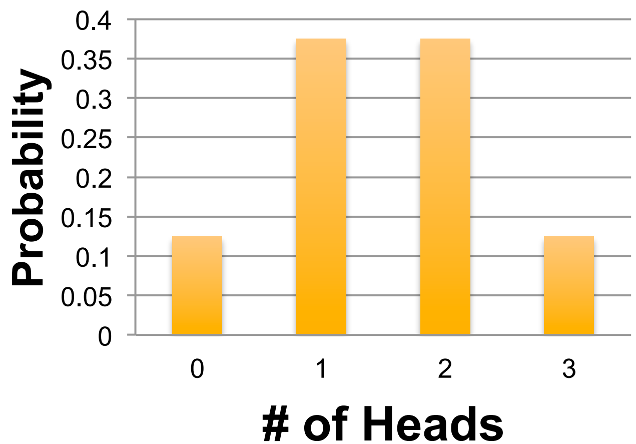
If obtain multiple measurements of a variable and plot frequency (probability) of variable value vs. variable value → obtain probability distribution of variable

Distributions

- **Binomial Distribution – describes distribution of binary (2 outcome) data from finite sample. Give gives probability of getting outcome p from n trials.**
- **Poisson Distribution – describes distribution of binary (2 outcome) data from large sample. (n trials) Gives probability of getting outcome p in n trials.**
- **Poisson Approximation to Binomial Distribution – Poisson good approximation to Binomial as n trials increase**
- **Gaussian or normal distribution – describes continuous data with symmetric distribution – limiting form of Poisson Distribution as n trials becomes infinite**

Binomial Distribution

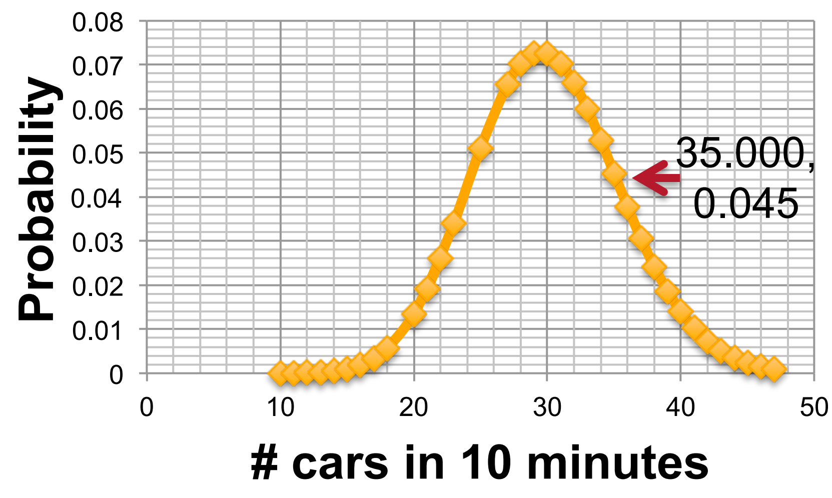
- Two outcomes –
  - Probability of outcome 1 = p
  - Probability of outcome 2 = q
  - Total probability p+q=1 or q=1-p or p=1-q
- Probability distribution $P(n,N)=N!\bullet p^{n}\bullet(1-p)^{N-n}/[n!\bullet(N-n)!]$
  - N = # trials/measurements, n= number with outcome p ➔ (N-n)= number with outcome q
  - Mean value of distribution = $N\bullet p$
- Example – toss coin 3 times for each data set, what is distribution?
  - p=probability of head =1/2, q= probability of tail =1/2
  - Probability of 0 heads ➔$(3!/0!(3-0)!)\bullet(1-0.5)^{3}=1/8$
  - Probability of 1 head ➔$(3!/1!(3-1)!)\bullet(1-0.5)^{2}\bullet 0.5=3/8$
  - Probability of 2 head ➔$(3!/2!(2-1)!)\bullet(1-0.5)^{1}\bullet 0.5^{2}=3/8$
  - Probability of 3 head ➔$(3!/3!(3-0)!)\bullet(1-0.5)^{0}\bullet 0.5^{3}=1/8$



MICHIGAN STATE UNIVERSITY

Poisson Distribution

- Certain number of outcomes within bin (e.g. time period, energy, etc.)
  - # cosmic rays per minute or # babies born per month or # car accidents per week
- Probability distribution $P(x)=(\lambda \bullet t)^{x} \bullet e^{-\lambda \bullet t}/x!$
  - x # of events of specific type, t interval, λ = average value of events in interval t, e= constant=2.718
- Example – if on average, 3 cars cross bridge per minute, what is probability of 35 cars crossing in 10 minutes?
  - λ= 3 car crossings/1 minute = 3 car crossing / minute
  - t= 10 minutes, x= 35
  - $P(x=35)=(3 \text{ crossing/minute} \bullet 10 \text{ minutes})^{35} \bullet e^{-(3 \bullet 10)}/35!=0.045$ (4.5%)
- Probability as function of # cars/10 min
  - Highest probability from λ
  - 3 cars/minute or 30 cars/10 minutes



MICHIGAN STATE UNIVERSITY

Poisson Approximation to Binomial - 1

- IF
  - Sample size n is large and
  - Probability p is small (q is then large)
- THEN Poisson distribution approximates Binomial
  - Larger the n and the smaller the p the better the approximation
- Binomial $\cong P(x) \cong (n \bullet p)^x \bullet e^{-n \bullet p}/x!$
  - n = sample size, p = true probability of success, x = # of successes, e=2.718
  - Mean value = n•p
  - Standard deviation = $(n \bullet p)^{0.5}$

Poisson Approximation to Binomial - 2

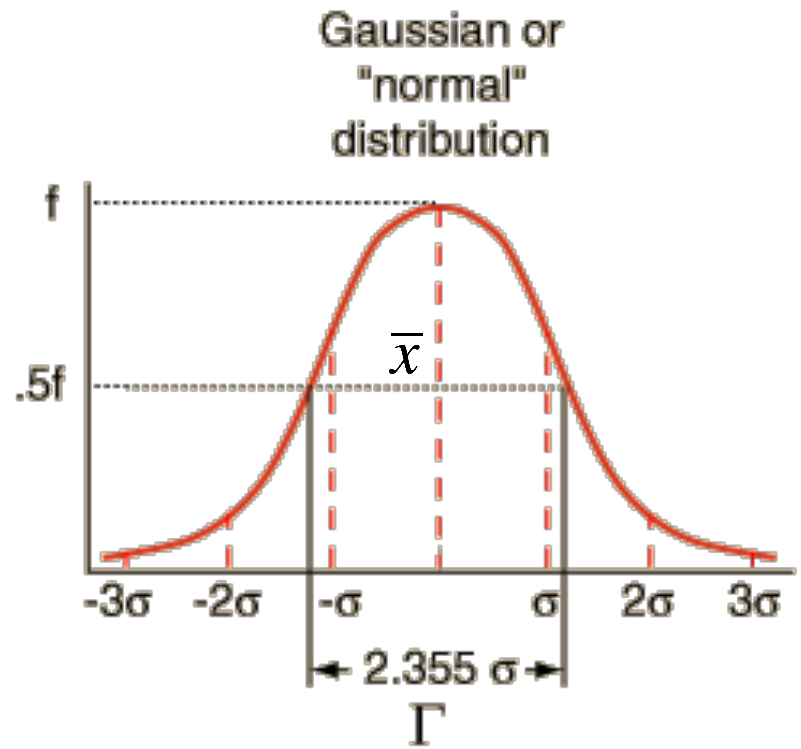
- Example
  - 8% of tires manufactured in plant are defective (= p)
  - Probability of 1 (=x) defective tire from sample of 20 (=n)
- Poisson Approximation (within ~1.6% of Binomial for example)
  - $P(x=1) \cong [e^{-(20)(0.08)} \bullet [(20)(0.08)]^1 / 1! = 0.3230$
  - Larger the n and the smaller the p the better the approximation
- Binomial
  - $P(n,N)=N! \bullet pn \bullet (1-p)N-n/[n! \bullet (N-n)!]=20! \bullet 0.92^{19} \bullet 0.08^1/19!=0.3282$

MICHIGAN STATE UNIVERSITY

Gaussian (normal) distribution

- **Gaussian (normal) distribution**
  - Derived from Poisson distribution but for <u>infinite number</u> of trials (n)
  - Standard deviation = σ_x
  - Average (mean) value $\bar{x}$
  - Uncertainty
    - ±1σ = 68.3% of measurements
    - ±2σ = 95.5% of measurements
    - ±3σ = 99.7% of measurements
    - Full width at half-maximum $\Gamma = 2.355\,\sigma_x$

$$f(x) = \frac{e^{-(x-\bar{x})^2/(2\sigma_x^2)}}{\sigma_x\sqrt{2\pi}}$$



- **Figure From:**

http://hyperphysics.phy-astr.gsu.edu/hbase/math/gaufcn2.html

MICHIGAN STATE UNIVERSITY

Models - 1

Data fit to <u>model</u> – to represent data or as aid to data analysis

- Represent data by fitting to some mathematical function
  - E.G. Gaussian or polynomial or sine/cosine function etc.
  - Provides
    - “Short hand” version of data – one math function representing many data points, function fitting can be method to “average” all data
    - Better data analysis e.g. better way to
      - to find maximum (peak) or minimum (valley) or intercept of data or
      - to interpolate (determine value between data points)

Models - 2

Data fit to model – to check physics model

- Physics model based on theory of variable relationship – very powerful if model correct since then can predict result without measurement
- Check if model is "correct" → agrees with experimental data
  - Can lead to model refinements (to get better agreement with data)
  - Can allow extraction of parameter that is element of another model to test broader consistency
- Experimental data used to confirm - or not – validity of model

Models - 3

Data fit to model – to check physics model

- E.g. Force (f) = mass (m) x acceleration (a) or f=ma
  - For experiments with bodies traveling with velocities much less than speed of light – e.g. 100 miles per hour f=ma works very well (mass is ~constant) – therefore verifies validity of formula for lower velocity bodies
  - For experiments with bodies with velocity substantial fraction of speed of light f=ma would not agree with experimental results (since mass has velocity dependence and therefore not constant)
    - Experiment would not confirm f=ma model with constant m
    - Need to use different model – that has mass dependent on velocity – new velocity dependent mass model can be experimentally tested

Models - 4

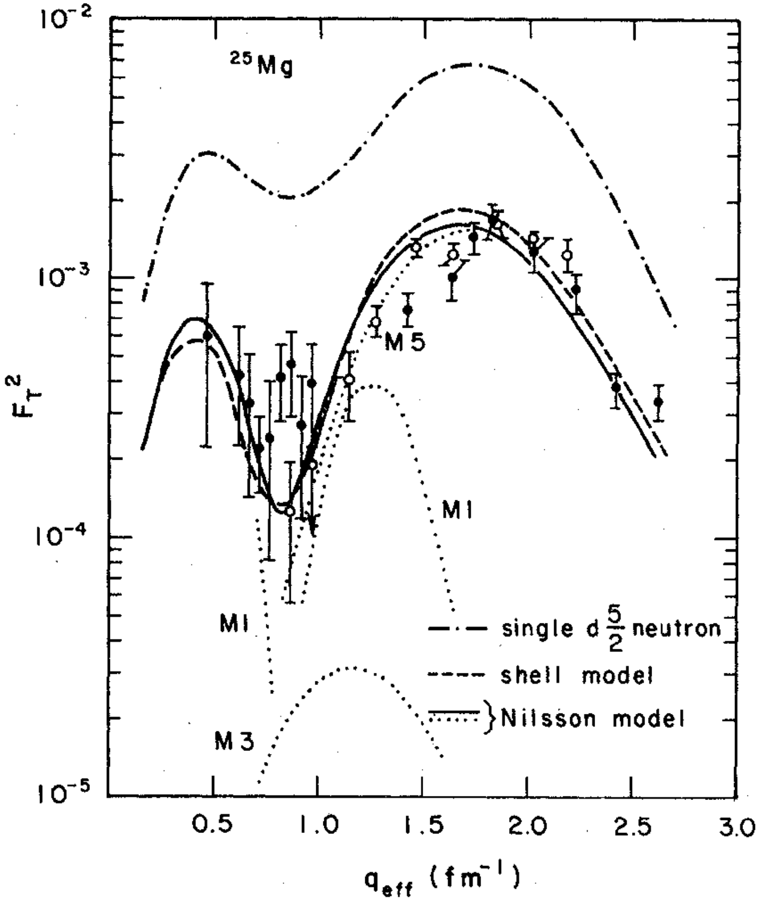


FIG. 1. The square of the transverse elastic form factors of ^{25}Mg. Only statistical errors are shown. The open circle data are the results of this work, and the closed circle data those of Euteneuer *et al.* (Ref. 2). The dotted and solid curves are from a Nilsson model calculation of Moya de Guerra and Dieperink (Ref. 14), and indicate the contributions from the individual multipoles and the sum of the multipoles, respectively. The dash-dotted curve is the result of an ESPM calculation of the sum of the multipole contributions. The dashed curve is the result of a shell model calculation of W. Chung (Ref. 15) of the sum of the multipole contributions.

From: http://journals.aps.org/prc/pdf/10.1103/PhysRevC.19.574

Data Fitting - 1

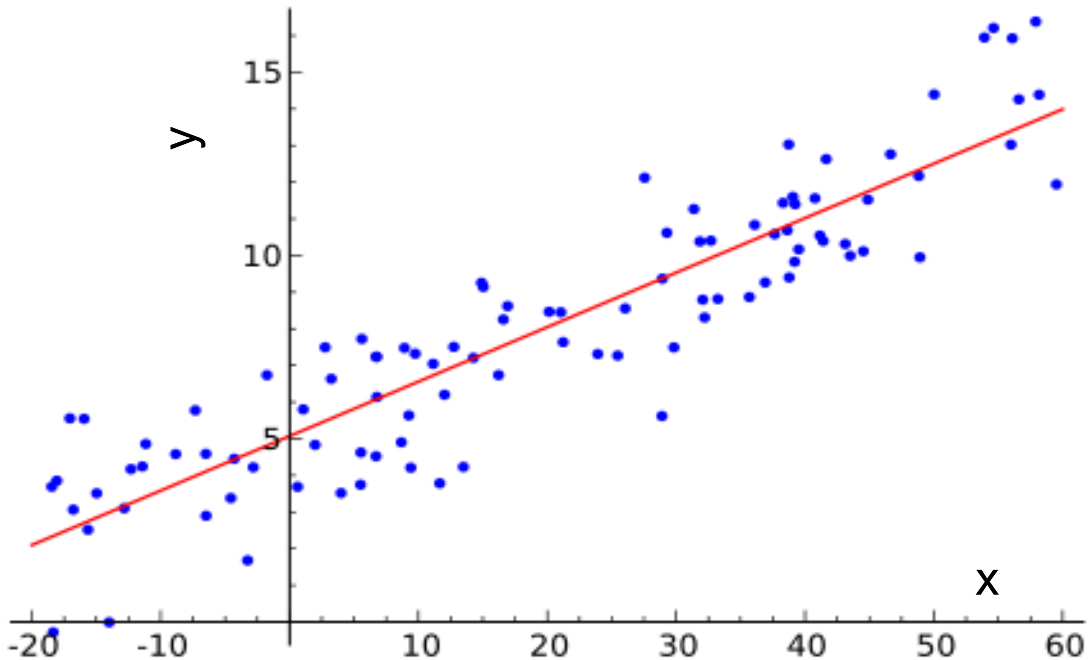
Resources - Scientific Graphing & Data Analysis

- See http://www.pa.msu.edu/courses/PHY451/tools.html
- Also Microsoft EXCEL
- Search web – e.g. “online data fitting”
  - Free online resources

Data Fitting - 1

- Typically have a model (functional dependence) of variables
- Simple model could be f(x) = y = a + b•x
  - See figure (from http://en.wikipedia.org/wiki/Goodness_of_fit)
  - Data value = y_i
  - function value at data point → $yf_i = f(x_i)$
  - Data uncertainty = σ_i
- Fit Quality – "chi squared"

$$X^2 = \frac{\sum_{i=1}^{n} (y_i - yf_i)^2}{\sigma_i^2}$$



MICHIGAN STATE UNIVERSITY

Data Fitting - 2

- **Average, mean, 'typical value'** $\bar{y}$
  - n measurements y_i

$$\bar{y} = \frac{\sum_i y_i}{n}$$

- **Standard deviation, variance,** $\sigma_y = \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{n-1}}$
- **Fit Quality ➔ $R^2 = 1 - X^2/\sigma_y^2$**
- **Best fit for R^2 ➔ 1**