Statistics, Probability, Distributions, & Error Propagation

James R. Graham

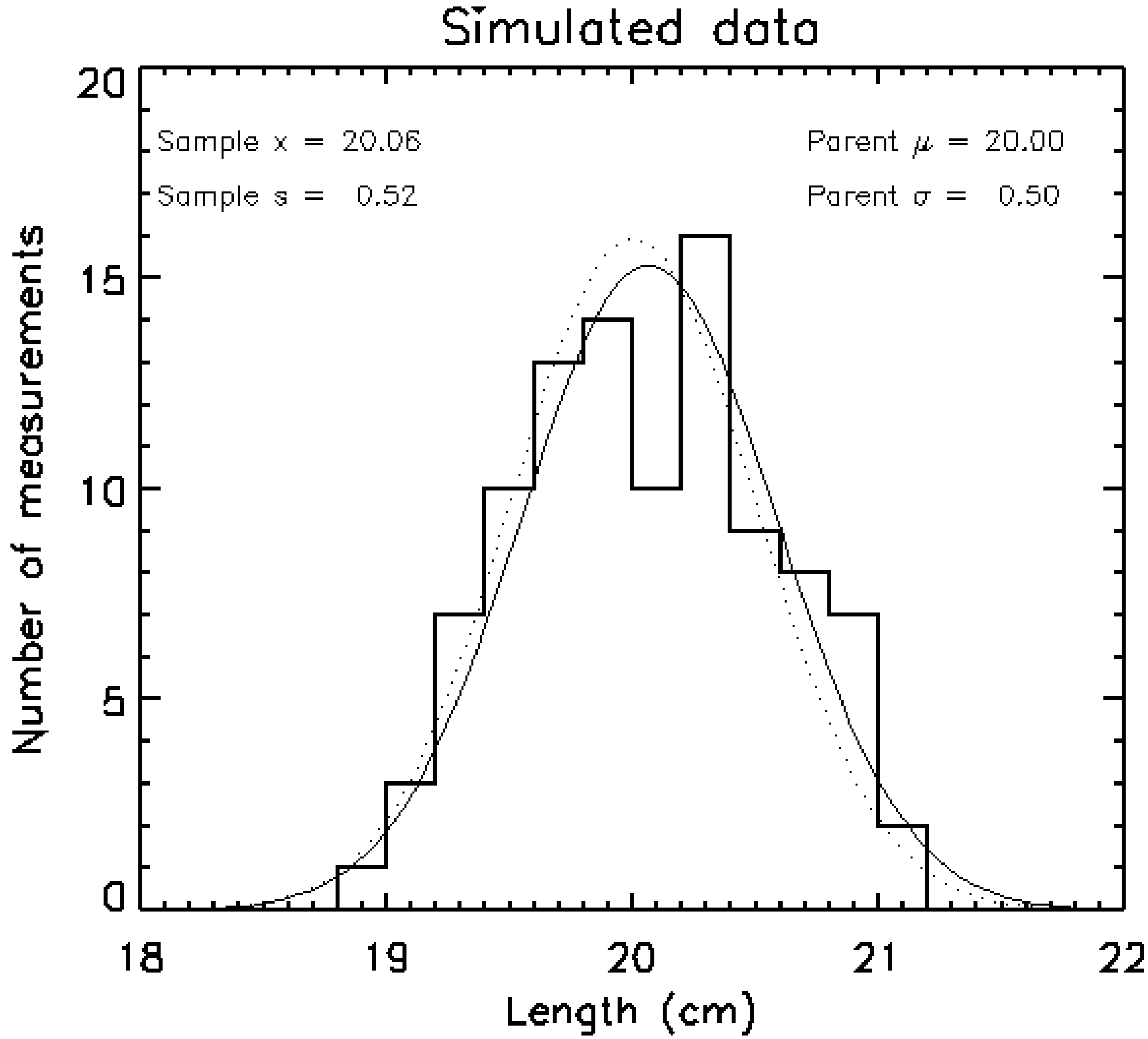
9/2/09

Sample & Parent Populations

- Make measurements
  - x_1
  - x_2
  - In general do not expect $x_1 = x_2$
  - But as you take more and more measurements a pattern emerges in this sample
- With an infinite sample x_i, $i \in \{1\ldots\infty\}$ we can
  - Expect a pattern to emerge with a characteristic value
  - Exactly specify the ***distribution*** of x_i
  - The hypothetical pool of all possible measurements is the ***parent population***
  - Any finite sequence is the ***sample population***

Histograms & Distributions

- Histogram represents the ***occurrence*** or ***frequency*** of discrete measurements
  - Parent population (dotted)
  - Inferred parent distribution (solid)



Notation

- Parent distribution: Greek, e.g., μ
- Sample distribution: Latin, $\bar{x}$
  – To determine properties of the parent distribution assume that the properties of the sample distribution tend to those of the parent as *N* tends to infinity

Summation

- If we make *N* measurements, x_1, x_2, x_3, etc. the sum of these measurements is

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + x_3 + ... + x_N$$

- Typically, we use the shorthand

$$\sum_{i=1}^{N} x_i = \sum x_i$$

Mean

- The mean of an experimental distribution is

$$\bar{x} = \frac{1}{N}\sum x_i$$

- The mean of the parent population is defined as

$$\mu = \lim_{N \to \infty}\left(\frac{1}{N}\sum x_i\right)$$

Median

- The median of the parent population $\mu_{1/2}$ is the value for which half of $x_i < \mu_{1/2}$

$$P(x_i < \mu_{1/2}) = P(x_i \geq \mu_{1/2}) = 1/2$$

- The median cuts the area under the probability distribution in half

Mode

- The mode is the most probable value drawn from the parent distribution
  - The mode is the most likely value to occur in an experiment
  - For a symmetrical distribution the mean, median and mode are all the same

Deviation

- The deviation, d_i , of a measurement, x_i , from the mean is defined as

$$d_i = x_i - \mu$$

- If μ is the true mean value the deviation is the error in x_i

Mean Deviation

- The mean deviation vanishes!
  - Evident from the definition

$$\lim_{N\to\infty} \bar{d} = \lim_{N\to\infty}\left[\frac{1}{N}\sum(x_i - \mu)\right] = \underbrace{\lim_{N\to\infty}\left[\frac{1}{N}\sum x_i\right]}_{\mu} - \mu$$

Mean Square Deviation

- The mean square deviation is easy to use analytically and justified theoretically

$$\sigma^2 = \lim_{N\to\infty}\left[\frac{1}{N}\sum\left(x_i-\mu\right)^2\right] = \lim_{N\to\infty}\left[\frac{1}{N}\sum x_i^2\right]-\mu^2$$

- σ^2 is also known as the ***variance***
  - Derive this expression
  - Computation of σ^2 assumes we know μ

Population Mean Square Deviation

- The ***estimate*** of the standard deviation, *s,* from a sample population is

$$s^2 = \frac{1}{N-1}\sum\left(x_i - \bar{x}\right)^2$$

- The factor (*N*-1) is used instead of *N* to account for the fact that the mean must be derived from the data

Significance

- The mean of the sample is the best estimate of the mean of the parent distribution
  - The standard deviation, *s*, is characteristic of the uncertainties associated with attempts to measure μ
  - But what is the uncertainty in μ?
- To answer these questions we need probability distributions…

μ and σ of Distributions

- Define μ and σ in terms of the parent probability distribution $P(x)$
  - Definition of $P(x)$
    - Limit as $N \rightarrow \infty$
    - The number of observations dN that yield values between x and $x + dx$ is

$$dN/N = P(x)\,dx$$

Expectation Values

- The mean, μ, is the expectation value of some quantity *x*

$$\langle x \rangle$$

- The variance, σ^2, is the expectation value of the deviation squared

$$\langle (x-\mu)^2 \rangle$$

Expectation Values

- For a discrete distribution, *N*, observations and *n* distinct outcomes

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

$$= \lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{n} x_j n_{x_j} \quad \text{each } x_j \text{ is a unique value}$$

$$= \lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{n} x_j N P(x_j)$$

$$= \lim_{N\to\infty} \sum_{j=1}^{n} x_j P(x_j)$$

Expectation Values

- For a discrete distribution, *N*, observations and *n* distinct outcomes

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

$$= \underset{N\to\infty}{Lim}\frac{1}{N}\sum_{j=1}^{n}(x_j-\mu)^2 NP(x_j)$$

$$= \underset{N\to\infty}{Lim}\sum_{j=1}^{n}\left[(x_j-\mu)^2 P(x_j)\right]$$

Expectation values

- The expectation value of any continuous function of *x*

$$\langle f(x) \rangle = \int_{-\infty}^{\infty} f(x) P(x) dx$$

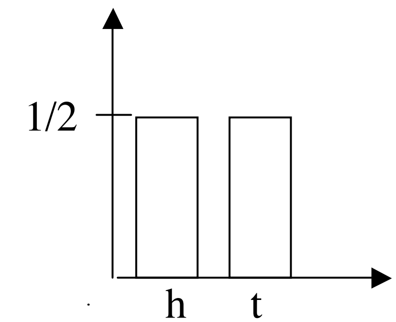
$$\mu = \int_{-\infty}^{\infty} x P(x) dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 P(x) dx$$

where $\int_{-\infty}^{\infty} P(x) dx = 1$

Binomial Distribution

- Suppose we have two possible outcomes with probability p and q = 1-p
  - e.g., a coin toss, p = 1/2, q = 1/2
- If we flip n coins what is the probability of getting x heads?
  - Answer is given by the *Binomial Distribution*



$$P(x;n,p) = C(n,x)\,p^x q^{n-x}$$

- $C(n, x)$ is the number of combinations of n items taken x at a time = $n!/[x!(n-x)!]$

Binomial Distribution

- The expectation value

$$\mu = \sum_{x=0}^{n} x P(x;n,p)$$

$$= \sum_{x=0}^{n} x C(n,x) p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \left[x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \right] = np$$

Poisson Distribution

- The Poisson distribution is the limit of the Binomial distribution when $\mu << n$ because p is small
  - The binomial distribution describes the probability $P(x; n, p)$ of observing x events per unit time out of n possible events
  - Usually we don't know n or p but we do know μ

Poisson Distribution

- Suppose *p* << 1 then *x* << *n*

$$P(x;n,p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$\frac{n!}{(n-x)!} = n(n-1)(n-2)...(n-x-2)(n-x-1)$$

$$\approx n^x \quad \text{when n >> x}$$

$$\frac{n!}{(n-x)!} p^x \approx (np)^x = \mu^x$$

$$(1-p)^{n-x} = (1-p)^{-x}(1-p)^n \approx 1 \times (1-p)^n \text{ since p << 1}$$

$$\underset{p \to 0}{Lim}(1-p)^n = \underset{p \to 0}{Lim}\left[(1-p)^{1/p}\right]^\mu = \left(e^{-1}\right)^\mu = e^{-\mu}$$

$$P(x,\mu) = \frac{\mu^x}{x!} e^{-\mu}$$

Poisson Distribution

- The expectation value of *x* is

$$\langle x \rangle = \sum_{x=0}^{\infty} x P(x,\mu) = \sum_{x=0}^{\infty} x \frac{\mu^x}{x!} e^{-\mu} = \mu$$

- Expectation value of $(x\text{-}\mu)^2$

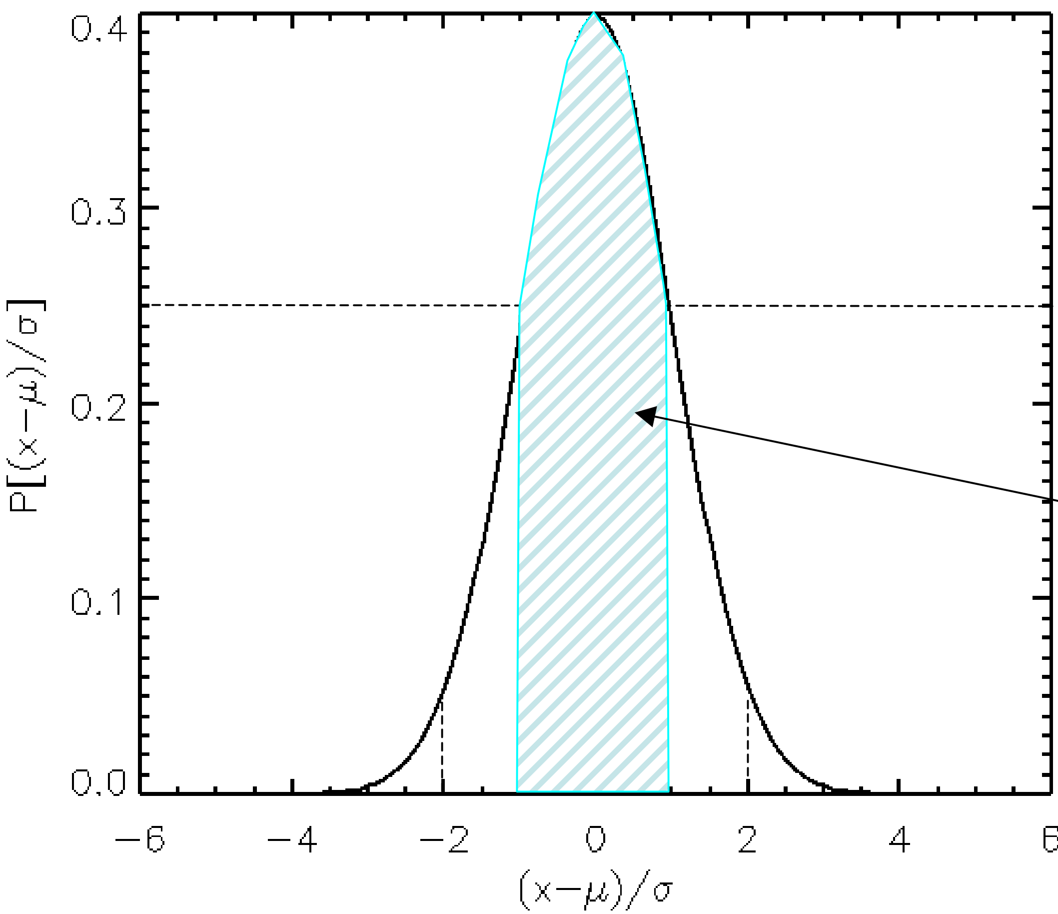
$$\sigma^2 = \left\langle (x-\mu)^2 \right\rangle = \sum_{x=0}^{\infty} (x-\mu)^2 \frac{\mu^x}{x!} e^{-\mu} = \mu$$

Gaussian or Normal Distribution

- The Gaussian distribution is an approximation to the binomial distribution for large *n* and large *np*

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Gaussian or Normal Distribution



$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-\frac{1}{2}x^2} dx = 0.683$$

+/- 1σ: 68.3%

+/- 2σ: 95.5%

+/- 3σ: 99.7%

Combining Two Observations

- Suppose I have two sets of measurements, a_i, and b_i
  - A derived quantity $c_i = a_i + b_i$
  - What is the relation between the means and standard deviations of a_i and b_i and c_i
  - Suppose we have the same number of observations N of a_i and b_i

Combining Two Observations

$$N = N_a = N_b$$

$$\bar{a} = \frac{1}{N}\sum a_i \qquad \bar{b} = \frac{1}{N}\sum b_i$$

$$\bar{c} = \frac{1}{N}\sum c_i \qquad s_c^2 = \frac{1}{N-1}\sum \left(c_i - \bar{c}\right)^2$$

$$c_i = a_i + b_i$$

$$\bar{c} = \frac{1}{N}\sum (a_i + b_i) = \frac{1}{N}\sum a_i + \frac{1}{N}\sum b_i$$

$$\boxed{= \bar{a} + \bar{b}}$$

Combining Two Observations

$$s_c^2 = \tfrac{1}{N-1}\sum\left(c_i - \overline{c}\right)^2, \quad \overline{c} = \overline{a} + \overline{b}$$

$$s_c^2 = \tfrac{1}{N-1}\sum\left[a_i + b_i - \left(\overline{a} + \overline{b}\right)\right]^2$$

$$= \tfrac{1}{N-1}\sum\left[\left(a_i + b_i\right)^2 - 2\left(a_i + b_i\right)\left(\overline{a} + \overline{b}\right) + \left(\overline{a} + \overline{b}\right)^2\right]$$

$$= \tfrac{1}{N-1}\sum\left[a_i^2 + b_i^2 + 2a_i b_i - 2\left(a_i\overline{a} + a_i\overline{b} + b_i\overline{a} + b_i\overline{b}\right) + (\overline{a})^2 + 2\overline{a}\overline{b} + \left(\overline{b}\right)^2\right]$$

$$= \tfrac{N}{N-1}\overline{a^2} + \tfrac{N}{N-1}\overline{b^2} + \tfrac{2}{N-1}\sum a_i b_i - \tfrac{N}{N-1}(\overline{a})^2 - \tfrac{2N}{N-1}\overline{a}\overline{b} - \tfrac{N}{N-1}\left(\overline{b}\right)^2$$

Combining Two Observations

$$s_c^2 = \tfrac{1}{N-1}\sum\left(c_i - \bar{c}\right)^2, \quad \bar{c} = \bar{a} + \bar{b}$$

$$= \tfrac{N}{N-1}\overline{a^2} + \tfrac{N}{N-1}\overline{b^2} + \tfrac{2}{N-1}\sum a_i b_i - \tfrac{N}{N-1}(\bar{a})^2 - \tfrac{2N}{N-1}\bar{a}\bar{b} - \tfrac{N}{N-1}\left(\bar{b}\right)^2$$

$$= \underbrace{\tfrac{N}{N-1}\left[\overline{a^2} - (\bar{a})^2\right]}_{s_a^2} + \underbrace{\tfrac{N}{N-1}\left[\overline{b^2} - \left(\bar{b}\right)^2\right]}_{s_b^2} + \underbrace{\tfrac{2N}{N-1}\left(\overline{ab} - \bar{a}\bar{b}\right)}_{2s_{ab}^2}$$

$$\boxed{s_c^2 = s_a^2 + s_b^2 + 2s_{ab}^2}$$

- The term s^2_{ab} is the covariance
  - Murphy's law factor
  - s_{ab} can be negative, zero or positive

Combining Two Uncorrelated Observations

- When *a* and *b* are uncorrelated the covariance is zero

$$s_{ab}^2 = \tfrac{1}{N-1}\sum\left(a_i - \overline{a}\right)\left(b_i - \overline{b}\right) = 0$$

$$s_c^2 = s_a^2 + s_b^2$$

  - **The variance of *c* is the sum of the variances of *a* and *b***
- This demonstrates the fundamentals of error propagation

Propagation of Errors

- Suppose we want to determine *x* which is a function of measured quantities, *u*, *v*, etc.

$$x = f(u, v, \ldots)$$

- Assume that

$$\bar{x} = f(\bar{u}, \bar{v}, \ldots)$$

Propagation of Errors

- The uncertainty in *x* can be found by considering the spread of the values of *x* resulting from individual measurements, u_i, v_i , etc.,

$$x_i = f(u_i, v_i, \ldots)$$

- In the limit of $N \to \infty$ the variance of *x*

$$\sigma_x^2 = \lim_{N \to \infty} \tfrac{1}{N} \sum_i \left(x_i - \bar{x}\right)^2$$

Propagation of Errors

- Taylor expand the deviation ($N \to \infty$ assumed

$$x_i - \overline{x} = \left(u_i - \overline{u}\right)\frac{\partial f}{\partial u}\bigg|_{\overline{u}} + \left(v_i - \overline{v}\right)\frac{\partial f}{\partial v}\bigg|_{\overline{v}} + \ldots$$

$$\sigma_x^2 = \tfrac{1}{N}\sum_i\left[\left(u_i - \overline{u}\right)\frac{\partial f}{\partial u}\bigg|_{\overline{u}} + \left(v_i - \overline{v}\right)\frac{\partial f}{\partial v}\bigg|_{\overline{v}} + \ldots\right]^2$$

$$= \tfrac{1}{N}\sum_i\left[\left(u_i - \overline{u}\right)^2\left(\frac{\partial f}{\partial u}\right)_{\overline{u}}^2 + \left(v_i - \overline{v}\right)^2\left(\frac{\partial f}{\partial v}\right)_{\overline{v}}^2 + 2\left(u_i - \overline{u}\right)\left(v_i - \overline{v}\right)\frac{\partial f}{\partial u}\bigg|_{\overline{u}}\frac{\partial f}{\partial v}\bigg|_{\overline{v}} \ldots\right]$$

Propagation of Errors

$$\sigma_x^2 = \tfrac{1}{N}\sum_i \left[(u_i - \bar{u})^2 \left(\frac{\partial f}{\partial u}\right)_{\bar{u}}^2 + (v_i - \bar{v})^2 \left(\frac{\partial f}{\partial v}\right)_{\bar{v}}^2 + 2(u_i - \bar{u})(v_i - \bar{v}) \left.\frac{\partial f}{\partial u}\right|_{\bar{u}} \left.\frac{\partial f}{\partial v}\right|_{\bar{v}} \ldots \right]$$

$$= \tfrac{1}{N}\sum_i (u_i - \bar{u})^2 \left(\frac{\partial f}{\partial u}\right)_{\bar{u}}^2 +$$

$$\tfrac{1}{N}\sum_i (v_i - \bar{v})^2 \left(\frac{\partial f}{\partial v}\right)_{\bar{v}}^2 +$$

$$\tfrac{2}{N}\sum_i (u_i - \bar{u})(v_i - \bar{v}) \left.\frac{\partial f}{\partial u}\right|_{\bar{u}} \left.\frac{\partial f}{\partial v}\right|_{\bar{v}} + \ldots$$

$$\boxed{\sigma_x^2 = \sigma_u^2 \left(\frac{\partial f}{\partial u}\right)_{\bar{u}}^2 + \sigma_v^2 \left(\frac{\partial f}{\partial v}\right)_{\bar{v}}^2 + 2\sigma_{uv}^2 \left.\frac{\partial f}{\partial u}\right|_{\bar{u}} \left.\frac{\partial f}{\partial v}\right|_{\bar{v}} + \ldots}$$

Examples of Error Propagation

- Suppose $a = b + c$
  - We know that

$$\bar{a} = \bar{b} + \bar{c}$$

$$\sigma_a^2 = \sigma_b^2 + \sigma_c^2$$

assuming that the covariance is 0

- What about $a = b/c$?

Examples of Error Propagation

- Suppose *a* = *b*/*c*?

$$\bar{a} = \bar{b}/\bar{c}$$

and

$$\sigma_a^2 = \sigma_b^2 \left(\frac{\partial a}{\partial b}\right)_{\bar{b}}^2 + \sigma_c^2 \left(\frac{\partial a}{\partial c}\right)_{\bar{c}}^2 + 2\sigma_{bc}^2 \left.\frac{\partial a}{\partial b}\right|_{\bar{b}} \left.\frac{\partial a}{\partial c}\right|_{\bar{c}} + \ldots$$

$$\sigma_a^2 = \sigma_b^2 \frac{1}{c^2} + \sigma_c^2 \left(\frac{b}{c^2}\right)^2$$

or

$$\frac{\sigma_a^2}{a^2} = \frac{\sigma_b^2}{b^2} + \frac{\sigma_c^2}{c^2}$$

assuming that the covariance is 0

Error of the Mean

- Suppose we have *N* measurements, x_i with uncertainties characterized by s_i

$$\bar{x} = \tfrac{1}{N}\left(x_1 + x_2 + x_3 + \ldots + x_N\right) = \tfrac{1}{N}\sum_i x_i$$

$$s_{\bar{x}}^2 = s_1^2\left(\frac{\partial \bar{x}}{\partial x_1}\right)_{\bar{x}}^2 + s_2^2\left(\frac{\partial \bar{x}}{\partial x_2}\right)_{\bar{x}}^2 + s_3^2\left(\frac{\partial \bar{x}}{\partial x_3}\right)_{\bar{x}}^2 + \ldots + s_N^2\left(\frac{\partial \bar{x}}{\partial x_N}\right)_{\bar{x}}^2$$

$$= \sum_i s_i^2\left(\frac{\partial \bar{x}}{\partial x_i}\right)_{\bar{x}}^2$$

assuming that the covariance is 0

Error of the Mean

- Suppose the errors on all points are equal so that $s_i = s$

$$s_{\bar{x}}^2 = \sum_i s_i^2 \left(\frac{\partial \bar{x}}{\partial x_i} \right)_{\bar{x}}^2$$

$$\frac{\partial \bar{x}}{\partial x_i} = \frac{\partial}{\partial x_i} \left(\frac{1}{N} \sum_j x_j \right) = \frac{1}{N} \qquad\qquad \frac{\partial x_j}{\partial x_i} = \delta_{ij}$$

$$\boxed{\begin{aligned} s_{\bar{x}}^2 &= \sum_i s^2 \left(\frac{1}{N} \right)^2 \\ &= \frac{s^2}{N} \end{aligned}}$$

Examples of Error Propagation

- What happens when m = $-2.5 \log_{10}(F/F_0)$?
  - What is the error in m?

$$m = -2.5\log_{10}(F/F_0)$$

and

$$\sigma_m^2 = \sigma_F^2 \left(\frac{\partial m}{\partial F} \right)_{\bar{F}}^2$$

$$\sigma_m^2 = \sigma_F^2 \left(\frac{2.5}{F \log(10)} \right)^2$$

$$\sigma_m^2 = (1.087)^2 \left(\frac{\sigma_F}{F} \right)^2$$