Physics 403

Parameter Estimation

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Table of Contents

Principle of Indifference

Uniform and Jeffreys Priors

- **Principle of Indifference**: given $n > 1$ mutually exclusive and exhaustive possibilities, each should be assigned a probability equal to $1/n$.
- Matches our intuition, and we've been applying it throughout the course. We can also use it to derive PDFs.
- Uniform prior is appropriate for a location parameter:

$$p(X|I) = \text{constant} = \frac{1}{x_{\text{max}} - x_{\text{min}}},$$

- Jeffreys prior is appropriate for a scale parameter:

$$p(X|I) = \frac{1}{x \ln\left(x_{\text{max}}/x_{\text{min}}\right)}$$

  It gives equal probability per decade.

Principle of Maximum Entropy

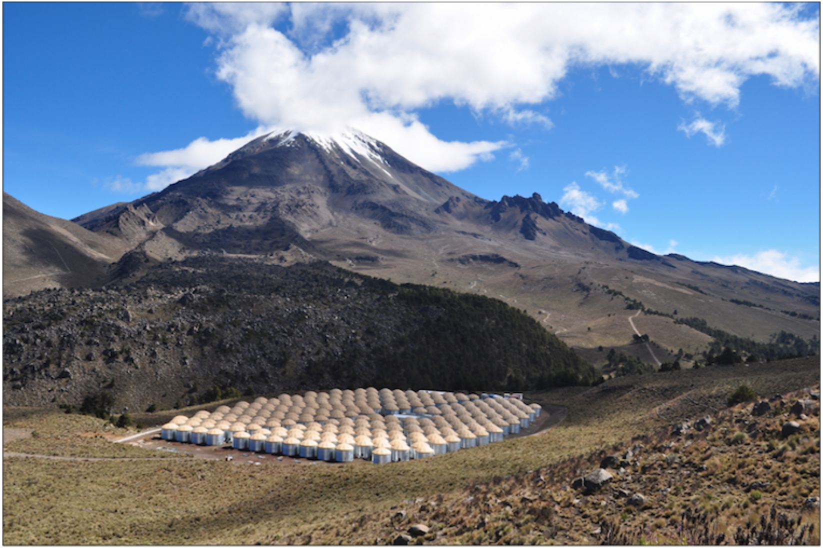
- **Principle of Maximum Entropy**: the least informative prior is the one which maximizes

$$S = -\sum_{i=1}^{N} p_i \ln\left(p_i/m_i\right) \quad \text{or} \quad S = -\int p(x) \ln\left(\frac{p(x)}{m(x)}\right) dx$$

- By maximizing S under different constraints we can derive familiar PDFs using Lagrange multipliers
- Example: given the normalization condition $\sum p_i = 1$, a fixed mean μ, and a fixed variance σ^2, the maximum entropy distribution is a Gaussian
- Important result: a Gaussian model of the uncertainties is a safe choice. Other distributions may give you artificially tight constraints unless you have appropriate prior information

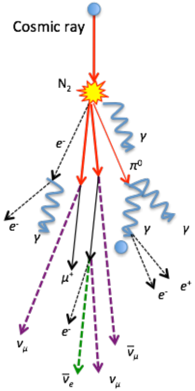
Case Study

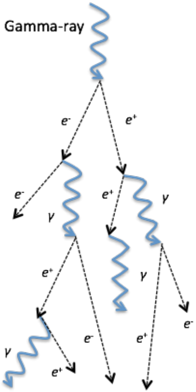
Reconstructing Air Showers with the HAWC Detector



Extensive Air Showers

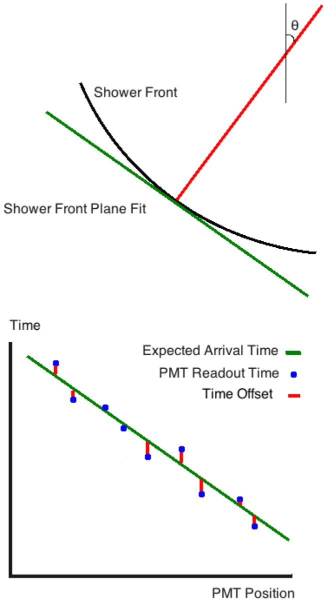
Particle Cascades in the Upper Atmosphere

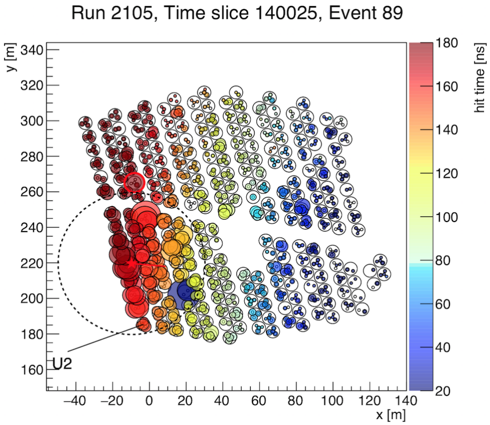




- Gamma rays and nuclear cosmic rays interact in the atmosphere
- A particle cascade, or air shower, of charged particles is produced
- The shower is shaped like a pancake: a few meters thick and $\mathcal{O}(100)$ meters across
- The "pancake" moves at speed $v \approx c$ to the ground, where the charged particles can be detected
- At altitude of Rochester, mostly muons remain at ground level. Flux is $\sim 100\ \text{m}^{-2}\ \text{s}^{-1}\ \text{sr}^{-1}$.

Fitting the Air Shower Plane





Color ∝ timing, circle area ∝ charge.
Two fits: "plane" and "curved" shower.

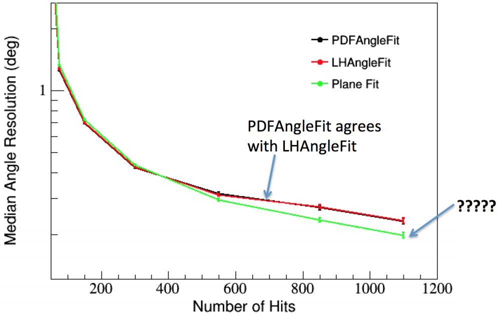
Consequence of Incorrect PDFs

In HAWC we make two fits to the shower front:

1. Planar fit
2. More correct "curved" fit

What happens when we attempt a maximum likelihood fit with simulated time residual PDFs? A worse result than plane fit.

- Why? Timing PDFs are narrow, but wrong
- Naïve parameterization with Gaussian uncertainties is better than correct parameterization with incorrect PDFs.



As $N_{\text{hit}} \to$ large, the likelihood fit with shower curvature should be better than the plane fit. Instead, it gets worse. Solution: try to do better with the PDFs.

Table of Contents

Estimators

- We have seen how the PDF encodes what we want to know about a parameter given data D and relevant background information I.
- An estimator is a summary of this distribution
  - Could be a parameter of the PDF. E.g., p for a binomial distribution
  - Could be a property of the distribution, like the mean
- You have total freedom to make up any estimator you want, but you'll want to report two numbers:
  1. The best estimate itself
  2. A measure of the reliability of the estimate
- Question: what do we mean by "best" estimator?
- Question: what do we mean by the "reliability" of the estimator?

Bayesian Solution to Parameter Estimation

- If the data D are distributed according to a parameter θ, the PDF of θ can be obtained using Bayes' Theorem:

$$p(\theta|D,I) = \frac{p(D|\theta,I)\ p(\theta|I)}{p(D|I)} = \frac{p(D|\theta,I)\ p(\theta|I)}{\int d\theta\ p(D|\theta,I)\ p(\theta|I)}$$

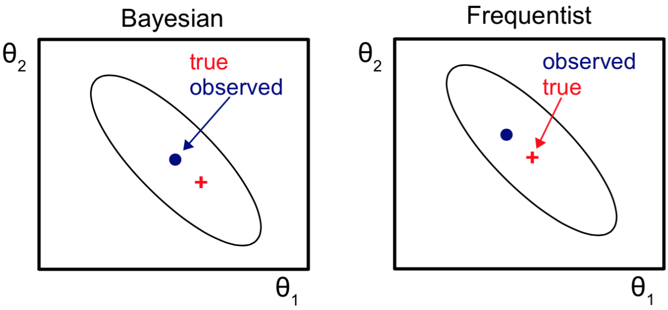
- The posterior $p(\theta|D,I)$ contains all the relevant information about θ.
- You can choose to report the entire distribution or provide a summary of the parameter
- If you're worried about the effect of priors, publish the likelihood $p(D|\theta,I)$ and/or show the effect of different priors on $p(\theta|D,I)$

Frequentist Approach to Parameter Estimation

- Remember that frequentists don't use $p(\theta|D, I)$; only $p(D|\theta, I)$.
- In other words, there is not really a concept of θ varying. Instead, θ has a fixed, "true" value (albeit unknown)
- Consequence: $p(D|\theta, I) =$ "probability of the data given a fixed θ"
- So the frequentist answers the question, "How probable is it that we observed this data D given some value of θ?"
- Most of frequentist statistics involves calculating p-values, or tail probabilities of $p(D|\theta, I)$.
- Because they assume a value for θ, p-values are a little dangerous when used to make decisions about the likelihood of a parameter or a model. They can overstate the evidence against your hypothesis about θ.
- This is one of the reasons that physicists use the 5σ rule of overwhelming evidence when using p-values

Bayesian vs. Frequentist Interpretations

- **Bayesian**: given D, the uncertainties tell us that the true value of the parameter lies within the ellipse centered on the observation with some probability
- **Frequentist**: given the true value of the parameters, the observation lies within an error ellipse centered on the true value with some probability



What is a Best Estimator?

- Let's answer the question of what defines a best estimator.
- Intuitive: it should be where the posterior PDF $p(x|D, I)$ is a maximum, meaning

$$\left.\frac{dp}{dx}\right|_{\hat{x}} = 0$$

  For this to be a maximum, we also require that

$$\left.\frac{d^2p}{dx^2}\right|_{\hat{x}} < 0$$

- If $\hat{x}$ gives the best estimator, then how do we define the reliability of the estimator?
- Look at the behavior of the PDF in a small region around the peak.

Reliability of an Estimator?

- Let's look at the Taylor expansion of p about $\hat{x}$, or better yet, $\ln p$:

$$L = \ln p = \ln p(x|D, I)$$

- We use the logarithm because p will often be a "peaky" function of x near $\hat{x}$. L varies more slowly and is a monotonic function of p.
- Taylor expanding L about $\hat{x}$, we get

$$L = L(\hat{x}) + \frac{1}{2} \frac{d^2 L}{dx^2}\bigg|_{\hat{x}} (x - \hat{x})^2 + \ldots$$

- The first term is a constant. The linear term vanishes (we're at the maximum). So the quadratic term dominates, and

$$p(x|D, I) \approx A\ \exp\left[\frac{1}{2} \frac{d^2 L}{dx^2}\bigg|_{\hat{x}} (x - \hat{x})^2\right]$$

Reliability of an Estimator?

- Compare the Taylor-expanded posterior PDF

$$p(x|D,I) \approx A \ \exp\left[\frac{1}{2}\frac{d^2L}{dx^2}\bigg|_{\hat{x}}(x-\hat{x})^2\right]$$

to the Gaussian

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- We can identify the width of the Gaussian as

$$\sigma = \left(-\frac{d^2L}{dx^2}\bigg|_{\hat{x}}\right)^{-1/2}$$

with $d^2L/dx^2 < 0$ (we're at the maximum). Hence, we express the parameter as

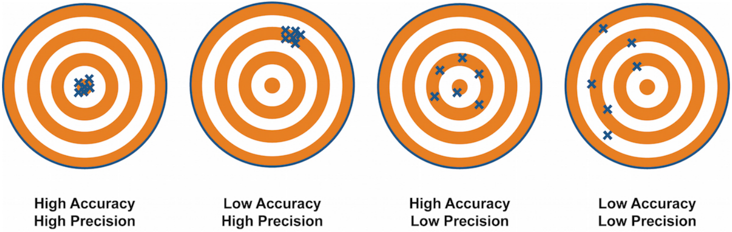
$$x = \hat{x} \pm \sigma,$$

where $\hat{x}$ is the best estimate and σ is its reliability.

Accuracy and Precision

Frequentist Aside

- It is useful to think of an estimator in terms of accuracy and precision
- **Accuracy**: how close is the estimator to true value? (Systematics)
- **Precision**: how clustered is the estimator about a central value? (Variance/Statistics)



Consistency and Bias

Caution: Frequentist Concept

- In the context of a sample of N measurements, we say that an estimator of θ, called $\hat{\theta}$, is consistent if

$$\lim_{N\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0, \quad \forall\ \epsilon > 0$$

  I.e., $\hat{\theta}$ converges to θ in the large N limit.

- We call an estimator unbiased if the bias b

$$b(\theta) = \mathrm{E}\,(\hat{\theta}) - \theta$$

  is zero.

- An estimator can be biased even if it is consistent. If $\hat{\theta} \to \theta$ for an infinite set of measurements in one experiment, it is not necessarily true that $\hat{\theta} \to \theta$ in an infinite set of experiments with a finite number of measurements.

Mean Squared Error (or Deviation)

- It is helpful to think of bias as a systematic error which does not improve with more data
- Another popular measure of the quality of an estimator is the mean squared error, defined as

$$\begin{aligned} d = \mathrm{MSE} &= \mathrm{E}\left((\hat{\theta} - \theta)^2\right) \\ &= \mathrm{E}\left((\hat{\theta} - \mathrm{E}(\hat{\theta}))^2\right) + (\mathrm{E}(\hat{\theta}) - \theta)^2 \\ &= \mathrm{var}(\hat{\theta}) + b^2 \end{aligned}$$

- I.e., the mean squared error (MSE) is the sum of the variance and the square of the bias.
- Classical interpretation: since the variance is the square of the uncertainty in the estimator, the MSE is the quadrature sum of statistical and systematic uncertainties.
- Root mean square (RMS) is defined as $\sqrt{\mathrm{MSE}}$.

What Makes a Good Estimator

Frequentist Aside

Let's define the three properties we expect from a good estimator.

1. **Consistent**: a consistent estimator will tend to the true value as the amount of data approaches infinity:

$$\lim_{N\to\infty} \hat{\theta} = \theta$$

2. **Unbiased**: the expectation value of the estimator is equal to the true value, so its bias b vanishes:

$$b = \langle\hat{\theta}\rangle - \theta = \int d\boldsymbol{x}\ p(\boldsymbol{x}|\theta)\ \hat{\theta}(\boldsymbol{x}) - \theta = 0$$

3. **Efficient**: the variance of the estimator is as small as possible (we'll see how small when we discuss the method of maximum likelihood):

$$\text{var}\,(\hat{\theta}) = \int d\boldsymbol{x}\ p(\boldsymbol{x}|\theta)\ (\hat{\theta}(\boldsymbol{x}) - \hat{\theta})^2$$

$$\text{MSE} = \langle(\hat{\theta} - \theta)^2\rangle = \text{var}\,(\hat{\theta}) + b^2$$

As you have seen, it is not always possible to satisfy all three requirements.

Case Study: Efficiency Uncertainty

Example

Suppose you use simulation to determine a selection efficiency: n out of N events pass some cuts. What is the selection efficiency ϵ and its uncertainty?

This is a binomial process: fixed trials N, fixed successes n, probability of success ϵ. Therefore,

$$p(n|N,\epsilon) \propto \epsilon^n (1-\epsilon)^{N-n}$$

and

$$\begin{aligned} L = \ln p &= \text{constant} + n \ln \epsilon + (N-n) \ln (1-\epsilon) \\ \frac{dL}{d\epsilon} &= \frac{n}{\epsilon} - \frac{N-n}{1-\epsilon} \\ \frac{d^2L}{d\epsilon^2} &= -\frac{n}{\epsilon^2} - \frac{N-n}{(1-\epsilon)^2} \end{aligned}$$

Case Study: Efficiency Uncertainty

Example

For the optimal value of ϵ, $dL/d\epsilon = 0$:

$$\left.\frac{dL}{d\epsilon}\right|_{\hat{\epsilon}} = \frac{n}{\hat{\epsilon}} - \frac{N-n}{1-\hat{\epsilon}}$$

$$\therefore \hat{\epsilon} = \frac{n}{N}$$

This is a pretty intuitive result: the best estimate of the efficiency is just n/N. Mixing in a frequentist concept: is it biased?

$$b = \mathrm{E}\left(\hat{\epsilon}\right) - \epsilon = \frac{\mathrm{E}\left(n\right)}{N} - \epsilon = \frac{N\epsilon}{N} - \epsilon = 0$$

So $\hat{\epsilon}$ is an unbiased estimator.
What about its uncertainty?

Case Study: Efficiency Uncertainty

Example

The estimated variance is given by

$$\hat{\sigma}^2 = -\left(\left. \frac{d^2 L}{d\epsilon^2} \right|_{\hat{\epsilon}} \right)^{-1}$$

After substituting $\hat{\epsilon} = n/N$ and combining terms, this reduces to

$$\left. \frac{d^2 L}{d\epsilon^2} \right|_{\hat{\epsilon}} = -\frac{N}{\hat{\epsilon}(1-\hat{\epsilon})}$$

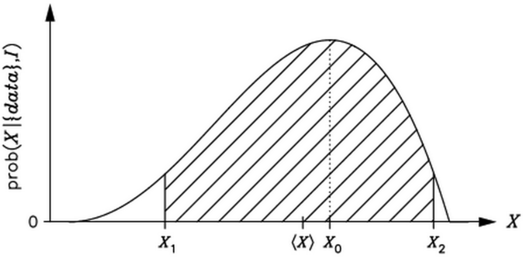
$$\therefore \hat{\sigma}^2 = \frac{\hat{\epsilon}(1-\hat{\epsilon})}{N} = \frac{n(N-n)}{N^3}$$

The expectation of $\hat{\sigma}^2$ is, after some more algebra,

$$\mathrm{E}\left(\hat{\sigma}^2\right) = \frac{N+1}{N}\sigma^2 \qquad \text{(slight bias)}$$

Asymmetric PDFs

- What happens when we have a very asymmetric PDF? In this case the expansion about the maximum may not be so reasonable.



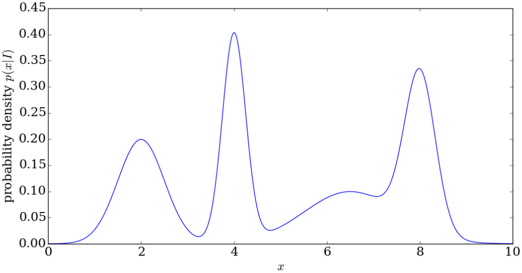
- This is where the concept of confidence intervals (or "credible regions" for a Bayesian) come in. We define

$$p(x_1 \leq x < x_2 | D, I) = \int_{x_1}^{x_2} p(x|D, I)\ dx \approx \alpha,$$

where $\alpha = 0.68$ (for example), and identify x_1 and x_2.

Multimodal PDFs

- What happens when we the PDF is multimodal? Can we even describe a "best parameter" and its uncertainty properly?



- You could try to summarize the posterior using ≥ 2 best estimates and their error bars, or some kind of disjoint confidence interval.
- Alternatively: cut your losses and just report the full posterior PDF.

Gaussian Uncertainties

- Suppose we are measuring values $\boldsymbol{x} = \{x_i\}$ drawn from a Gaussian distribution of mean μ and variance σ^2.
- For today, assume σ^2 is known but μ is not. How do we estimate μ given the data?
- Starting from Bayes' Theorem,

$$p(\mu|\boldsymbol{x}, \sigma^2, I) \propto p(\boldsymbol{x}|\mu, \sigma^2, I)\ p(\mu|\sigma^2, I)$$

- **Likelihood**: If the measurements x_i are independent, then

$$p(\boldsymbol{x}|\mu, \sigma^2, I) = \prod_{i=1}^{N} p(x_i|\mu, \sigma^2, I) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\sum_i \frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

- **Prior**: μ is a location parameter, so we'll use a uniform prior

$$p(\mu|\sigma^2, I) = \frac{1}{\mu_{\text{max}} - \mu_{\text{min}}}$$

which vanishes outside $x \in [\mu_{\text{min}}, \mu_{\text{max}}]$.

Gaussian Uncertainties

Estimate of the Mean

- As in the earlier examples, let's maximize the logarithm of the posterior PDF to get the best estimate for μ:

$$L = \ln p(\mu|\boldsymbol{x}, \sigma^2, I = \text{constant} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

- Differentiating, we have

$$\left.\frac{dL}{d\mu}\right|_{\hat{\mu}} = \sum_{i=1}^{N} \frac{x_i - \mu}{\sigma^2} = 0$$

$$\therefore \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

So the best estimate of μ is the arithmetic mean of the measurements, independent of the spread given by σ.

Gaussian Uncertainties

Uncertainty of the Mean

- We estimate uncertainty of the mean using the second derivative, as before:

$$\left.\frac{d^2L}{d\mu^2}\right|_{\hat{\mu}} = -\sum_{i=1}^{N}\frac{1}{\sigma^2} = -\frac{N}{\sigma^2}$$

- Therefore, our best estimate and uncertainty on the mean is summarized by

$$\mu = \hat{\mu} \pm \frac{\sigma}{\sqrt{N}}$$

- We have recovered the familiar expression often referred to as the "error on the mean," and derived the familiar rule that uncertainties decrease with measurement as $1/\sqrt{N}$.
- The only requirement is the validity of the quadratic expansion of the posterior PDF, which is exactly true for the Gaussian.
- This rule applies often in the lab thanks to the tendency of additive sources of noise to look Gaussian (Central Limit Theorem)

Different-Sized Error Bars

- What happens if the uncertainties in each x_i differ? As long as the source of uncertainties is Gaussian, then

$$p(\boldsymbol{x}|\mu,\sigma_i^2,I) = \prod_{i=1}^{N} p(x_i|\mu,\sigma_i^2,I) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}}\exp\left(-\sum_i \frac{(x_i-\mu)^2}{2\sigma_i^2}\right)$$

  where $\boldsymbol{\Sigma}$ is the diagonal covariance matrix of the $\{x_i\}$.
- Taking the logarithm and differentiating gives

$$L = \ln p = \text{constant} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma_i^2}$$

$$\left.\frac{dL}{d\mu}\right|_{\hat{\mu}} = \sum_{i=0}^{N} \frac{x_i-\mu}{\sigma_i^2} = 0$$

$$\therefore \hat{\mu} = \sum_{i=1}^{N} x_i/\sigma_i^2 \bigg/ \sum_{i=1}^{N} 1/\sigma_i^2 = \sum_{i=1}^{N} x_i w_i \bigg/ \sum_{i=1}^{N} w_i$$

Different-Sized Error Bars

- For the uncertainty on the mean, we have

$$\left.\frac{d^2L}{d\mu^2}\right|_{\hat{\mu}} = -\sum_{i=0}^{N}\frac{1}{\sigma_i^2}$$

$$\therefore \mu = \hat{\mu} \pm \left(\sum_{i=1}^{N} w_i\right)^{-1/2}, \qquad w_i = 1/\sigma_i^2$$

- So for the case of different uncertainties on each measurement x_i, the best estimator of the mean is the arithmetic sum of the data inversely weighted by the uncertainties.
- This makes a lot of sense; we want the data points with the biggest uncertainties to contribute the least to the sum

Table of Contents

Summary

- We can identify the best estimator of a PDF by maximizing it, so that

$$\left.\frac{dp}{dx}\right|_{\hat{x}} = 0$$

- We assessed the reliability of the estimator by Taylor expanding $L = \ln p$ about the best value:

$$\hat{\sigma}^2 = \left(-\left.\frac{d^2L}{dx^2}\right|_{\hat{x}}\right)^{-1}$$

- This only works when the quadratic approximation is reasonable. It may not be:
  1. **Asymmetric PDF**: better to use a confidence interval
  2. **Multimodal PDF**: no clear best estimate; report full PDF
- Frequentists: desire efficient, unbiased, and consistent estimators.

Next Time

- Extension of this technique to the multi-dimensional Gaussian and generalization of the quadratic approximation
- Introduction to the method of maximum likelihood
- Definition of the minimum variance bound
- Method of least squares
- Uncertainty propagation, or changes of variables in a PDF