# Central Limit Theorem

Alan Heavens

ICIC
Imperial Centre
for Inference & Cosmology

# Better answer to Day 1 number count problem: Likelihood for α

- Normalised distribution is $p(S)dS = (\alpha - 1)\left(\frac{S}{S_0}\right)^{-\alpha} \frac{dS}{S_0}$
- Expected number of sources in *(S,S+ΔS) is λ=Np(S)ΔS*
- Choose ΔS so small that the observed number n in any bin of width ΔS is 0 or 1. p(n=0) = exp(-λ); p(n=1) = λexp(-λ)
- Independent, so prob of entire set of source values is

$$\prod_{\text{empty cells}} e^{-\lambda} \prod_{\text{filled cells}} \lambda e^{-\lambda}$$

$$\lambda \to 0, \qquad p(\{S_i\}) = \prod_{\text{filled cells}} \lambda_i \propto \prod_{\text{sources}} p(S_i)$$

ICIC

- Likelihood (and hence posterior, if we assume a uniform prior for α) is therefore

$$L(\alpha) \propto \prod_{i=1}^{n} (\alpha - 1) S_0^{\alpha-1} S_i^{-\alpha}$$

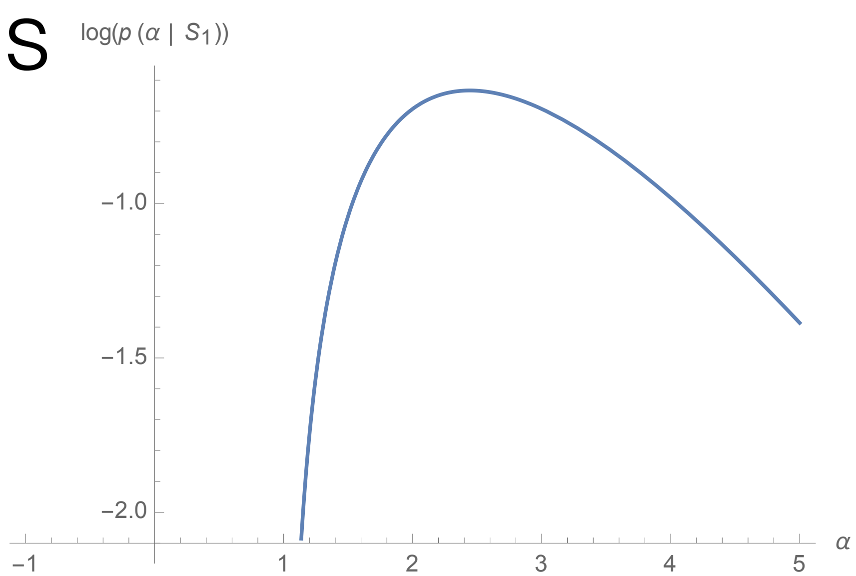
$$\ln L = \sum_{i=1}^{n} [\ln(\alpha - 1) + (\alpha - 1) \ln S_0 - \alpha \ln S_i] + \text{constant}$$

- Maximising lnL w.r.t. α gives

$$\frac{\partial}{\partial \alpha} \ln L = \sum_{i=1}^{n} \left( \frac{1}{\alpha - 1} + \ln S_0 - \ln S_i \right) = 0$$

- i.e. The maximum likelihood value is

$$\alpha = 1 + \frac{n}{\sum_{i=1}^{n} \ln \frac{S_i}{S_0}}$$




- For n=1 and $S_1 = 2S_0$, $\alpha_{ML} = 2.44$

ICIC

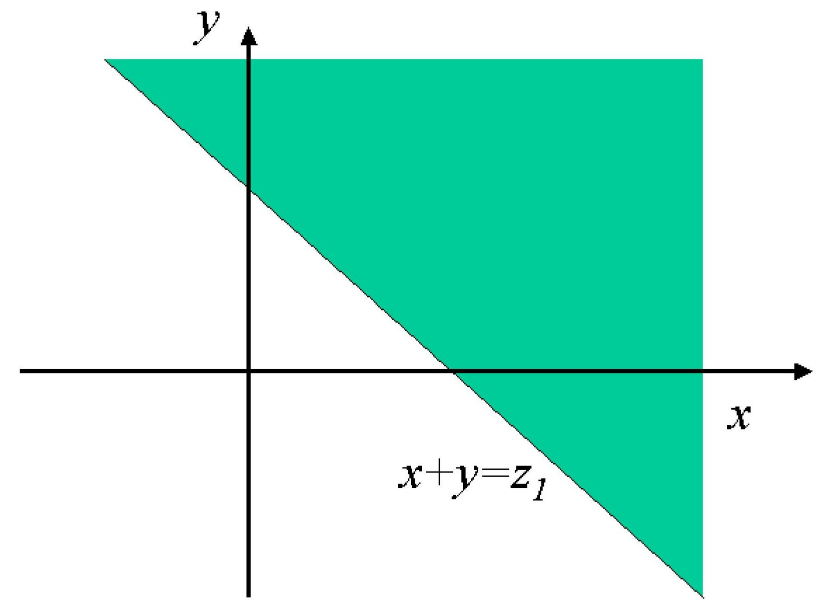
# Central Limit Theorem

## Preamble: adding two random variables

- Probability of a random variable $Z$ being larger than $z_1$:

$$p(Z \geq z_1) = \int_{z_1}^{\infty} dz\, p(z)$$

- Let $Z = x + y$.




$$p(Z \geq z_1) = \int_{-\infty}^{\infty} dy \int_{z_1 - y}^{\infty} dx\, p(x, y)$$

- If $x$ and $y$ are independent,

$$p(Z \geq z_1) = \int_{-\infty}^{\infty} dy \int_{z_1 - y}^{\infty} dx\, p_x(x)\, p_y(y)$$

ICIC

$$p(Z \geq z_1) = \int_{-\infty}^{\infty} dy \int_{z_1 - y}^{\infty} dx\, p_x(x)\, p_y(y)$$

- Change from *x* to *z=x+y; x=z-y (Note: it's more obvious perhaps to leave y as the outer integral first, but then notice that the limit on z does not depend on y, so the order can be reversed)*

$$p(Z \geq z_1) = \int_{z_1}^{\infty} dz \int_{-\infty}^{\infty} dy\, p_x(z-y)\, p_y(y)$$

- Since

$$p(Z \geq z_1) = \int_{z_1}^{\infty} dz\, p(z)$$

- we can read off

$$p(z) = \int_{-\infty}^{\infty} dy\, p_x(z-y)\, p_y(y)$$

ICIC

- i.e. *p(z)* is a *convolution*

# Convolution=product in Fourier space

- 'Characteristic function' (or generating function):

$$\phi(k) = \int_{-\infty}^{\infty} dx\, p(x)\, e^{ikx}$$

- Characteristic function for *z* is

$$\phi_z(k) = \phi_x(k)\phi_y(k)$$

- For n independent observations from the same *p(x)*

$$\phi_z(k) = \phi^n(k)$$

ICIC

# Central limit theorem

$$\phi_x(k) = \int_{-\infty}^{\infty} dx\, p(x) e^{ikx} = \int_{-\infty}^{\infty} dx\, p(x) \left[1 + ikx + \frac{(ikx)^2}{2!} + \ldots\right]$$

- i.e. $$\phi_x(k) = 1 + i\langle x\rangle k - \frac{1}{2}\langle x^2\rangle k^2 + \ldots,$$
- Consider $$X = \frac{1}{\sqrt{n}}(x_1 + x_2 + \cdots + x_n)$$
- Its characteristic function is $$\Phi_X(k) = [\phi_x(k/\sqrt{n})]^n$$
- If $<x>=0$ and $<x^2>=\sigma^2$, then, truncating the expansion at second order:

$$\Phi_x(k) = \left[1 - \frac{\sigma_x^2 k^2}{2n}\right]^n \to e^{-\sigma_x^2 k^2/2}$$

ICIC

# Central limit theorem

- Invert the Fourier transform: $X = \frac{1}{\sqrt{n}}(x_1 + x_2 + \cdots + x_n)$

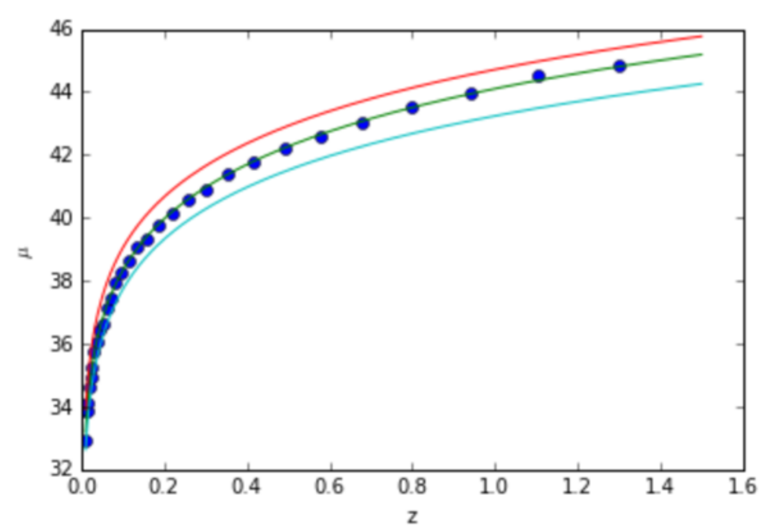
$$p(X) = \frac{e^{-X^2/(2\sigma_x^2)}}{\sqrt{2\pi}\sigma_x}.$$

- So we find the pdf for the *average* $\bar{X} = X/\sqrt{n}$

$$p(\bar{X}) = p(X)\sqrt{n} = \frac{e^{-X^2/(2\sigma_x^2/n)}}{\sqrt{2\pi\sigma_x^2/n}}.$$

- The average of n (many) identically-distributed random variables tends to a gaussian, with a variance given by the individual variances divided by n.

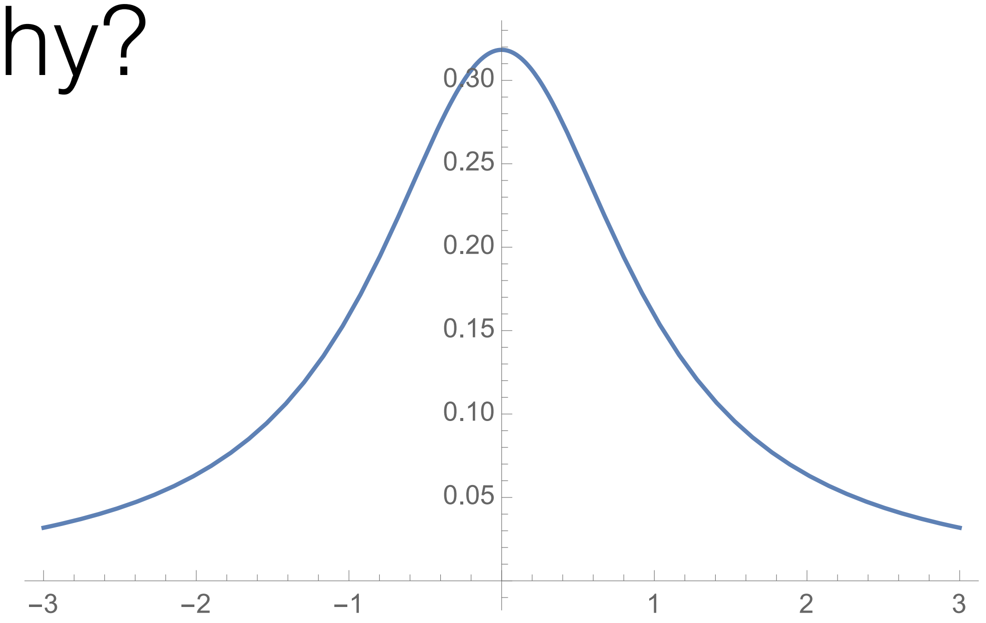
Amazing theorem - we don't need to know $p_x$, and the errors go down with more observations.



# Pathological pdfs

- Cauchy distribution $$p(x) = \frac{1}{\pi\left(1+x^2\right)}$$

  does not obey CLT. Why?



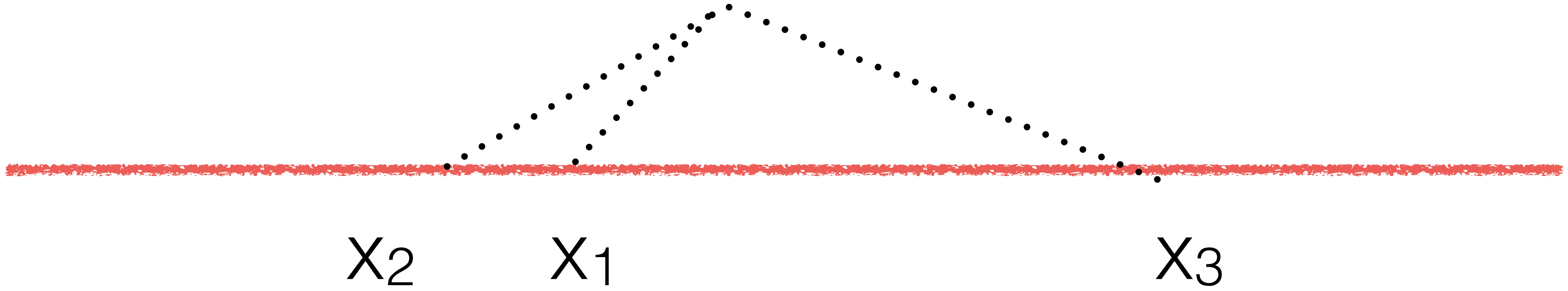

- Variance is infinite

ICIC

# The Lighthouse: Bayes vs estimator-based Frequentist

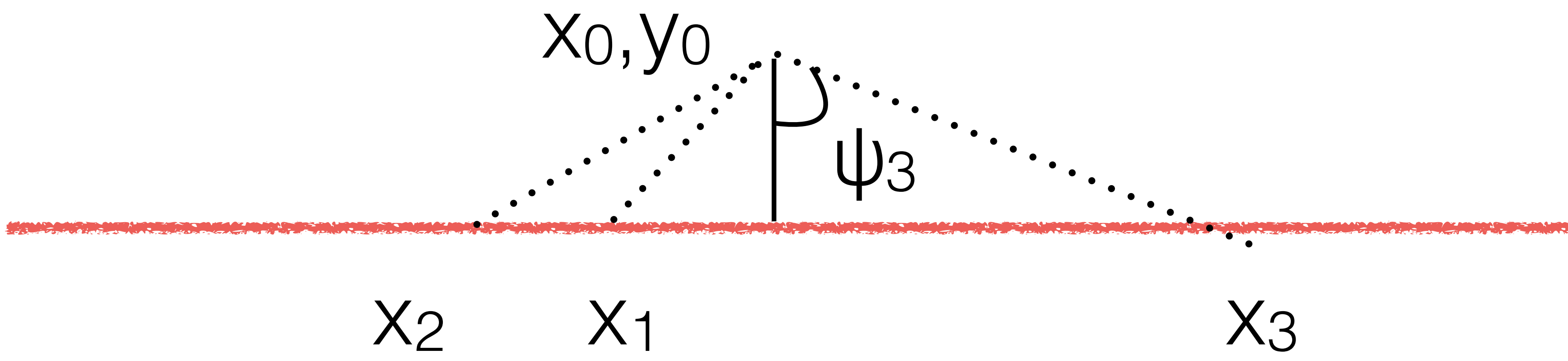
Steve Gull, 1st year University of Cambridge tutorial problem, 1988

A lighthouse is situated at unknown coordinates $x_0,y_0$ with respect to a straight coastline $y=0$. It sends a series of N flashes in random directions, and these are recorded on the coastline at positions $x_i$.




Using a Bayesian approach, find the posterior distribution of $x_0,y_0$, given the positions $x_i$.

# Solution




- Rule 1: we want $p(x_0, y_0|\{x_i\})$
- Use Bayes, assuming a uniform prior on $x_0,y_0$

$$p(x_0, y_0|\{x_i\}) \propto p(\{x_i\}|x_0, y_0)p(x_0, y_0) \propto \prod_i p(x_i|x_0, y_0)$$

- Let the angles wrt the vertical be $\psi_i$. Geometry gives

$$\frac{x_i - x_0}{y_0} = \tan\psi_i.$$

$$p(x_i|x_0, y_0) = p(\psi_i|x_0, y_0)\left|\frac{d\psi_i}{dx_i}\right|$$

$$p(\psi_i|x_0, y_0) = 1/\pi \;\; (-\pi/2 < \psi < \pi/2)$$

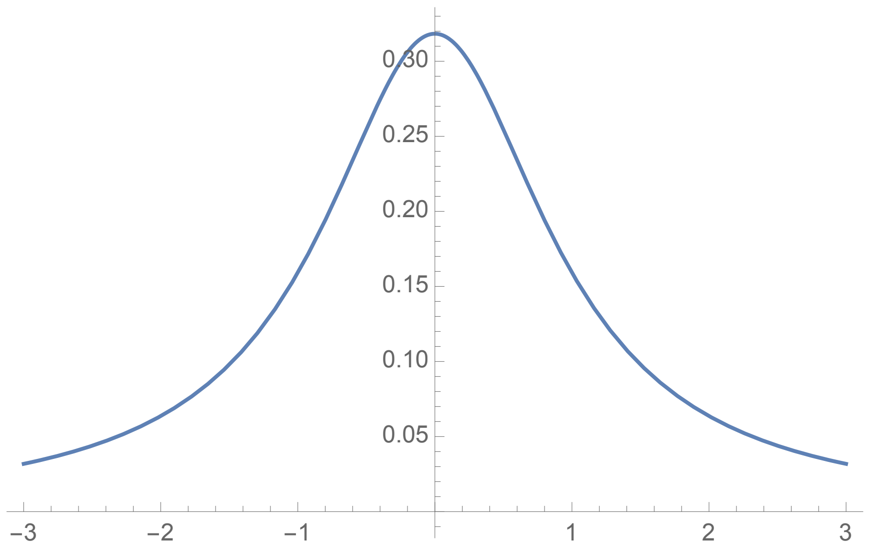
$$\sec^2\psi_i \frac{d\psi_i}{dx_i} = \frac{1}{y_0} \Rightarrow \left[1 + \frac{(x_i - x_0)^2}{y_0^2}\right]\frac{d\psi_i}{dx_i} = \frac{1}{y_0}$$

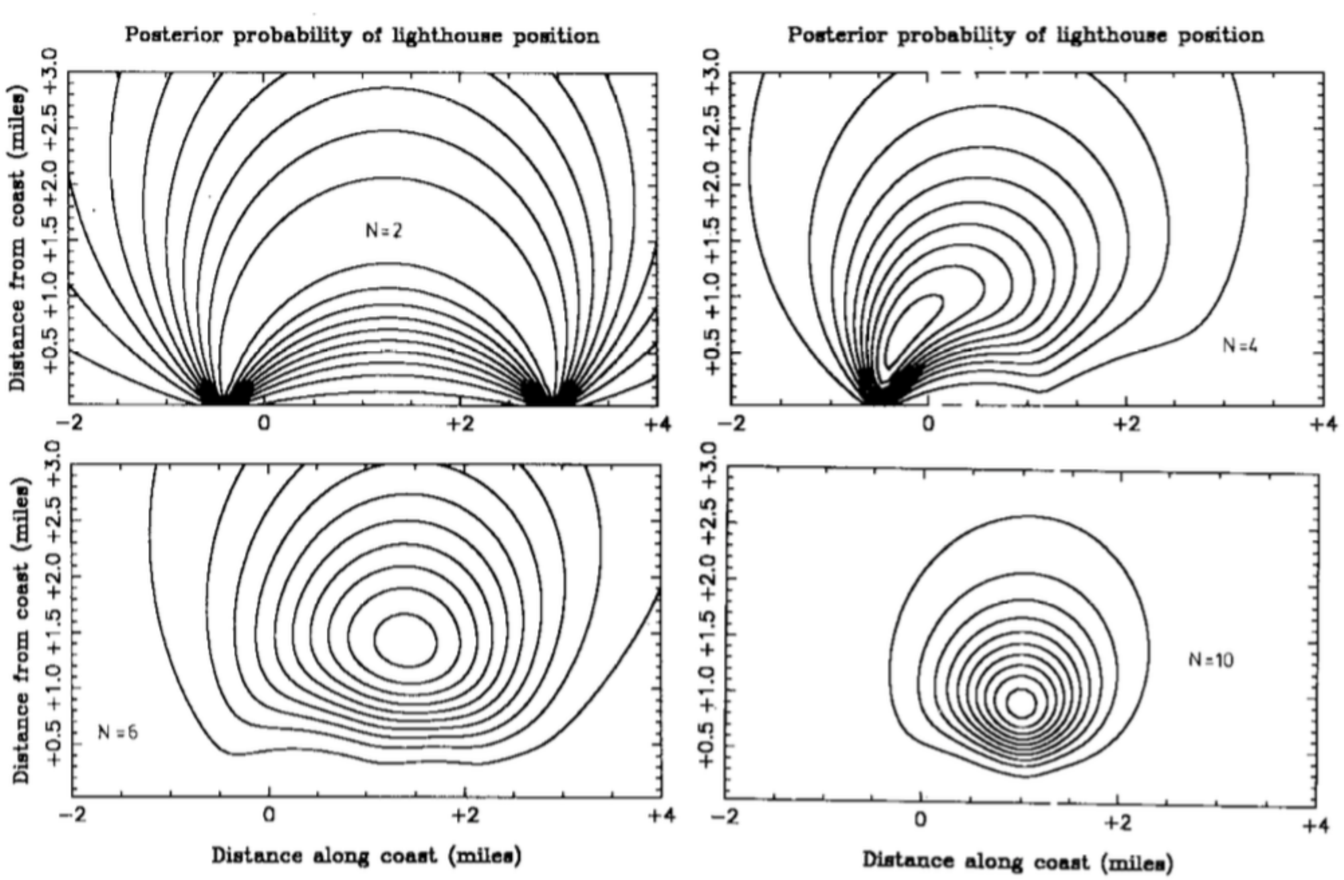
ICIC

- Hence the posterior is

$$p(x_0, y_0|\{x_i\}) \propto \prod_i \frac{1}{\pi y_0 \left[1 + \frac{(x_i - x_0)^2}{y_0^2}\right]}$$

- Product of Cauchy distributions



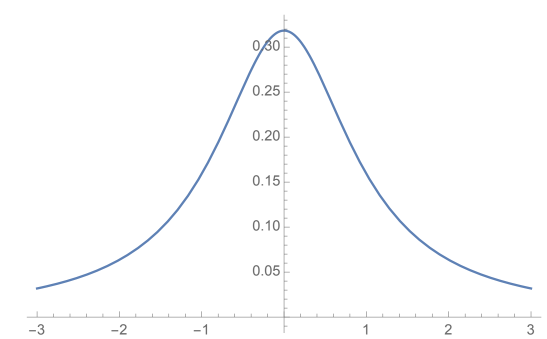



# Estimator-based Frequentist analysis for $x_0$

- Define an estimator. What would you choose?
- The average of the $x_i$: $$\hat{x}_0 = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{Z}{N}$$
- Z has a characteristic function $\Phi(k) = \phi^N(k)$

$$\phi(k) = \int_{-\infty}^{\infty} e^{ikx} \frac{1}{\left[1 + \frac{(x-x_0)^2}{y_0^2}\right]} dx = e^{ikx_0 - |k|y_0}.$$

$$\Phi(k) = e^{iNkx_0 - N|k|y_0}$$




- Invert to get *p(Z)*: $$p(N\hat{x}_0) = \frac{1}{\pi N y_0 \left[1 + \frac{(N\hat{x}_0 - Nx_0)^2}{N^2 y_0^2}\right]}$$

$$p(\hat{x}_0) = \frac{1}{\pi y_0 \left[1 + \frac{(\hat{x}_0 - x_0)^2}{y_0^2}\right]}$$

Having 1000 measurements is no better than having 1!