2.2. Reliabilities: best estimates, error-bars + confidence intervals

The posterior pdf encodes our inference about a parameter given the data and background information. Often, however, we would like to summarize the posterior in a few numbers.

First, our best estimate of a parameter X is given by

$$\left.\frac{d\,\text{prob}(X|\{\text{data}\},I)}{dX}\right|_{X_0} = 0 \quad \text{, i.e. maximum}$$

If the above condition holds for more than one value of X, take the one with largest posterior.

To obtain a measure of the reliability of our estimate X_0, we want to quantify the width of the posterior around X_0.
As the posterior is peaked, ~~expand~~ it's nicer to expand the <u>logarithm</u> of the posterior:

$$L \equiv \ln \text{prob}(X|\{\text{data}\},I)$$

$$\Rightarrow L = L(X_0) + \underbrace{\text{linear}}_{=0!} + \frac{1}{2}\left.\frac{d^2L}{dX^2}\right|_{X_0}(X-X_0)^2 + \dots$$

exponentiate this again:

$$\text{prob} = \exp[L] = \exp[\ln \text{prob}(x \mid \{\text{data}\}, I)]$$

$$= \exp\left[L(x_0) + \frac{1}{2} \frac{d^2 L}{dx^2}\Big|_{x_0} (x - x_0)^2 \right]$$

$$= \exp[L(x_0)] \exp\left[\frac{1}{2} \frac{d^2 L}{dx^2}\Big|_{x_0} (x - x_0)^2 \right]$$

Now compare this to the famous normal distribution:

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

Fig. 2.3.

So what we in fact did by taylor expanding around x_0 was approximating the posterior pdf by a normal distribution! Comparing the two exponentials, we find that

$$\sigma = \left[- \frac{d^2 L}{dx^2}\Big|_{x_0} \right]^{-1/2}$$

and one writes

$$x = x_0 \pm \sigma \quad \text{"error bar"}$$

The normal distribution has the property, that 67 % of all probability lies within $\pm\sigma$ and 95 % lies within 2σ:

$$\int_{x_0-\sigma}^{x_0+\sigma} \mathrm{prob}(x \mid \{\mathrm{data}\}, I)\, dx = 0.67$$

$$\int_{x_0-2\sigma}^{x_0+2\sigma} \mathrm{prob}(x \mid \{\mathrm{data}\}, I)\, dx = 0.95$$

Application: The coin example

Fig. 2.4. Remember: $\mathrm{prob}(H \mid \{\mathrm{data}\}, I) \propto H^R (1-H)^{N-R}$

$$L = \ln \mathrm{prob} = R \ln H + (N-R) \ln (1-H) + \mathrm{const}$$

$$\left.\frac{dL}{dH}\right|_{H_0} = \frac{R}{H_0} - \frac{N-R}{1-H_0} \overset{!}{=} 0$$

$$\Rightarrow \quad R(1-H_0) = H_0 (N-R)$$

$$\Rightarrow \quad R - \cancel{R H_0} = H_0 N - \cancel{H_0 R}$$

$$\Rightarrow \quad \boxed{H_0 = \frac{R}{N}}$$

i.e. $\frac{\#\text{ Heads}}{\#\text{ Tosses}}$ ... is best estimate of H.

error bar:

$$\left.\frac{d^2L}{dH^2}\right|_{H_0} = -\frac{R}{H_0^2} - \frac{N-R}{(1-H_0)^2}$$

$$= -\frac{R(1-H_0)^2 + (N-R)H_0^2}{H_0^2 (1-H_0)^2} = \frac{R(1-2H_0 \cancel{+H_0^2}) + H_0^2 N - \cancel{H_0^2 R}}{H_0^2 (1-H_0)^2}$$

[use $R = H_0 N$!] $$= -\frac{H_0 N (1-2H_0) + H_0^2 N}{H_0^2 (1-H_0)^2} = -\frac{H_0 N - H_0^2 N}{H_0^2 (1-H_0)^2}$$

$$= - \frac{H_0 N (1-H_0)}{H_0^2 (1-H_0)^2} = - \frac{N}{H_0(1-H_0)}$$

$$\Rightarrow \quad \sigma = \sqrt{\frac{H_0(1-H_0)}{N}}$$

Please note that after some data has been analyzed, H_0 is quite well known and $\sigma \propto \frac{1}{\sqrt{N}}$

Asymmetric posteriors:

Fig 2.5. Gaussian approximation is usually quite good. Yet, sometimes we get asymmetric posteriors. In this case, it is best to quote x_1, x_2 such that

$$\int_{x_1}^{x_2} \text{prob}(x \mid \{\text{data}\}, I)\, dx = 0.95 \quad \text{e.g.}$$

where the interval $[x_1, x_2]$ is short as possible.

For asymmetric posteriors, peak of posterior might not be representative. Instead use "mean" or "expectation".

$$\langle x \rangle = \int x \underbrace{\text{prob}(x \mid \{\text{data}\}, I)}_{\text{normalized, of course}} dx$$

2.3. Example: Gaussian Noise + Averages

The normal distribution is characterized by the fact that the probability to observe as the k-th datum a value x_k is given by:

$$\mathrm{prob}(x_k \mid \underset{\text{mean}}{\mu}, \underset{\text{error}}{\sigma}) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Let's suppose that σ is known (later unknown...)
Given some measurements $\{x_k\}$, we would like to ~~know~~ infer the mean μ:

$$\mathrm{prob}(\mu \mid \{x_k\}, \sigma, I)$$

From Bayes theorem:

$$\mathrm{prob}(\mu \mid \{x_k\}, \sigma, I) \propto \mathrm{prob}(\{x_k\} \mid \mu, \sigma, I)\, \mathrm{prob}(\mu \mid \sigma, I)$$

If the data are independent, then

$$\mathrm{prob}(\{x_k\} \mid \mu, \sigma, I) = \prod_k \mathrm{prob}(x_k \mid \mu, \sigma, I)$$

Let's use a flat prior $\mathrm{prob}(\mu \mid \sigma, I) = \begin{cases} A\,; & \mu_{min} < \mu < \mu_{max} \\ 0\,; & \text{otherwise} \end{cases}$

$$L = \ln \mathrm{prob}(\mu \mid \{x_k\}, \sigma, I)$$

$$= \ln \mathrm{prob}(\{x_k\} \mid \mu, \sigma, I) + \ln \mathrm{prob}(\mu \mid \sigma, I)$$

$$= \text{const} - \sum_k \frac{-(x_k-\mu)^2}{2\sigma^2}$$

To find our estimate of μ:

$$\frac{dL}{d\mu}\bigg|_{\mu_0} = \sum_{k=1}^{N} \frac{(x_k-\mu)}{\sigma^2} \stackrel{!}{=} 0$$

$$\Rightarrow \sum_{K=1}^{N} (x_K - \mu_0) = 0 \quad \Rightarrow N\mu_0 = \sum_{k=1}^{N} x_k$$

$$\Rightarrow \boxed{\mu_0 = \frac{1}{N}\sum_{k=1}^{N} x_k}$$

So our best estimate of μ is given by average of data.

$$\frac{d^2L}{d\mu^2}\bigg|_{\mu_0} = -\sum_{k=1}^{N}\frac{1}{\sigma^2} = -\frac{N}{\sigma^2}$$

So the error bar for our estimate of μ is given by $\text{error} = \frac{\sigma}{\sqrt{N}}$

and we get: $\mu = \mu_0 \pm \frac{\sigma}{\sqrt{N}}$

<u>Data with different sized errors:</u>

Suppose $\text{prob}(x_k \mid \mu, \sigma_k) = \frac{1}{\sqrt{2\pi}\,\sigma_k} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma_k^2}\right]$

i.e. different experiments.

then

$$L = \ln \text{prob}(\mu \mid \{x_k\}, \{\sigma_k\}, I)$$

$$= \text{constant} - \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma_k^2}$$

$$\left.\frac{dL}{d\mu}\right|_{\mu_0} = 0 \Rightarrow \quad \sum \frac{(x_k - \mu_0)}{\sigma_k^2} = 0$$

$$\Rightarrow \quad \sum \frac{x_k}{\sigma_k^2} = \mu_0 \sum \frac{1}{\sigma_k^2}$$

define "weight" $\quad w_k \equiv \dfrac{1}{\sigma_k^2}$

$$\Rightarrow \quad \sum x_k w_k = \mu_0 \sum w_k$$

$$\Rightarrow \quad \mu_0 = \frac{\sum x_k w_k}{\sum w_k}$$

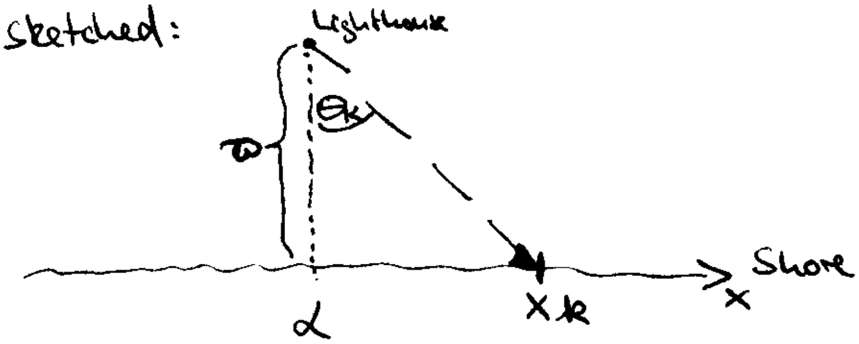
So the better the experiment, the more weight it gets. Repeating

$\dfrac{d^2L}{d\mu^2} = 0$, ~~so~~ you'll find

$$\mu = \mu_0 \pm \left[\sum w_k \right]^{-1/2}$$

2.4. Change of Variables example: the light house

See slide for the setup of our problem, roughly sketched:



The light house is at a distance 'β' from the shore at a position 'α' (which we want to infer!) and emits light pulses that are detected by photomultipliers on the coast. The data is hence a set of positions $\{x_k\}$ where flashes have been detected.

All θ_k are equally likely (and $\theta_k \in [-\frac{\pi}{2}, +\frac{\pi}{2}]$)

$$\text{prob}(\theta_k \mid \alpha, \beta, I) = \frac{1}{\pi}$$

Trigonometry says:

$$\beta \tan \theta_k = x_k - \alpha$$

We need to change variables, because we need in Bayes theorem prob$(x_k \mid \alpha, \beta, I)$ (we only measure x_k's..).

Suppose $\quad y = f(x) \quad$ monotonic

The probability in an interval dx around dy

should be conserved by this transformation:

$$\text{prob}(x|I)\,\Delta x \overset{!}{=} \text{prob}(y|I)\,\Delta y$$

$$\Rightarrow \text{prob}(x|I) = \text{prob}(y|I)\left|\frac{dy}{dx}\right|$$

$$= \text{prob}(y|I)\underbrace{\left|\frac{df}{dx}\right|}_{\text{Jacobian}}$$

For the light house, this means:

$$\beta \tan\theta_k = x_k - \alpha$$

$$\Rightarrow \frac{\beta}{\cos^2\theta_k}\,d\theta_k = dx_k \quad \Rightarrow \quad \frac{d\theta_k}{dx_k} = \frac{\cos^2\theta_k}{\beta}$$

use identity "$\left[\sin^2 + \cos^2 = 1 \Rightarrow \tan^2 + 1 = \frac{1}{\cos^2}\right]$"

$$\Rightarrow \frac{d\theta_k}{dx_k} = \left[\beta(\tan^2\theta_k + 1)\right]^{-1}$$

$$= \left[\beta\left\{1 + \frac{(x_k-\alpha)^2}{\beta^2}\right\}\right]^{-1} = \left[\frac{\beta^2 + (x_k-\alpha)^2}{\beta}\right]^{-1}$$

$$\Rightarrow \text{prob}(x_k|\alpha,\beta,I) = \text{prob}(\theta_k|\alpha,\beta,I)\left|\frac{d\theta}{dx_k}\right|$$

Fig. 2.8. $$= \frac{\beta}{\pi[\beta^2 + (x_k-\alpha)^2]}$$

This functional form is called <u>Lorentzian</u>

and the pdf is the Cauchy distribution.

Suppose we knew position at sea β. Use Bayes theorem:

$$\text{prob}(\alpha \mid \{x_k\}, \beta, I) \propto \text{prob}(\{x_k\} \mid \alpha, \beta, I)\ \underbrace{\text{prob}(\alpha \mid \beta, I)}_{\text{use flat prior for this!}}$$

As the flashes are independent:

$$\text{prob}(\{x_k\} \mid \alpha, \beta, I) = \prod_{k=1}^{N} \text{prob}(x_k \mid \alpha, \beta, I)$$

$$\Rightarrow L = \ln \text{prob}(\alpha \mid \{x_k\}, \beta, I)$$

$$= \text{const} + \sum \ln \frac{\beta}{\pi[\beta^2 + (x_k - \alpha)^2]}$$

$$= \text{const}' - \sum \ln [\beta^2 + (x_k - \alpha)^2]$$

So best estimate of α is:

$$\left.\frac{dL}{d\alpha}\right|_{\alpha_0} = 2 \sum \frac{x_k - \alpha_0}{\beta^2 + (x_k - \alpha_0)^2} = 0$$

Please note that the average of x_k's is not the solution, i.e. $\alpha_0' = \frac{1}{N}\sum x_k$ is not a solution here!

We have to use a "numerical solution" to obtain α_0, simply plot the function above... $\text{prob}(\alpha \mid \{x_k\}, \beta, I)$!

Fig 2.9.

The central limit theorem

In our light house example, something counter-intuitive happened: the average of all data $\{x_k\}$ did not give a good estimate of the position α. In fact, it was as wrong after millions of flashes as it was with only a few. In fact, the Cauchy distribution is one for which the central limit theorem does not apply.

Central-limit-theorem:

If samples are drawn from (almost) any pdf with mean μ, then in the limit of many data N, the average $\frac{1}{N}\sum x_k \rightarrow \mu$

and the error bar for the difference between the average and μ will go down $\propto \frac{1}{\sqrt{N}}$

The Cauchy distribution has a very wide tail, which means that it is not covered by the central-limit-theorem. In fact, all distributions with finite variance $\langle (x-\mu)^2 \rangle$ are covered by the theorem.