POINT ESTIMATION

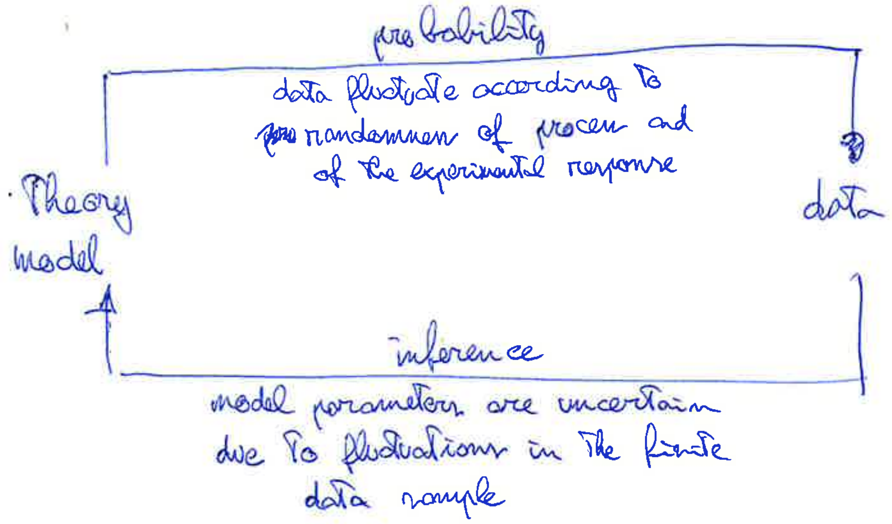
~~Most of this applies to both the frequentist and Bayesian approach~~

Most of this is frequentist exclusively, even if some methods are easily applicable to a Bayesian approach.

- ~~Estimates and estimators~~
  ~~An estimate of a parameter is a statistic, i.e. it is a function of the data.~~

- Inference

The inference is the process of determining an estimated value $\hat{\theta}$ and the corresponding uncertainty of some parameter θ from experimental data.



- Estimators and estimates

The estimate of an unknown parameter is a mathematical procedure to determine the central value of the parameter as a function of the observed data sample.

The function of the data sample that returns the estimate is called "estimator".

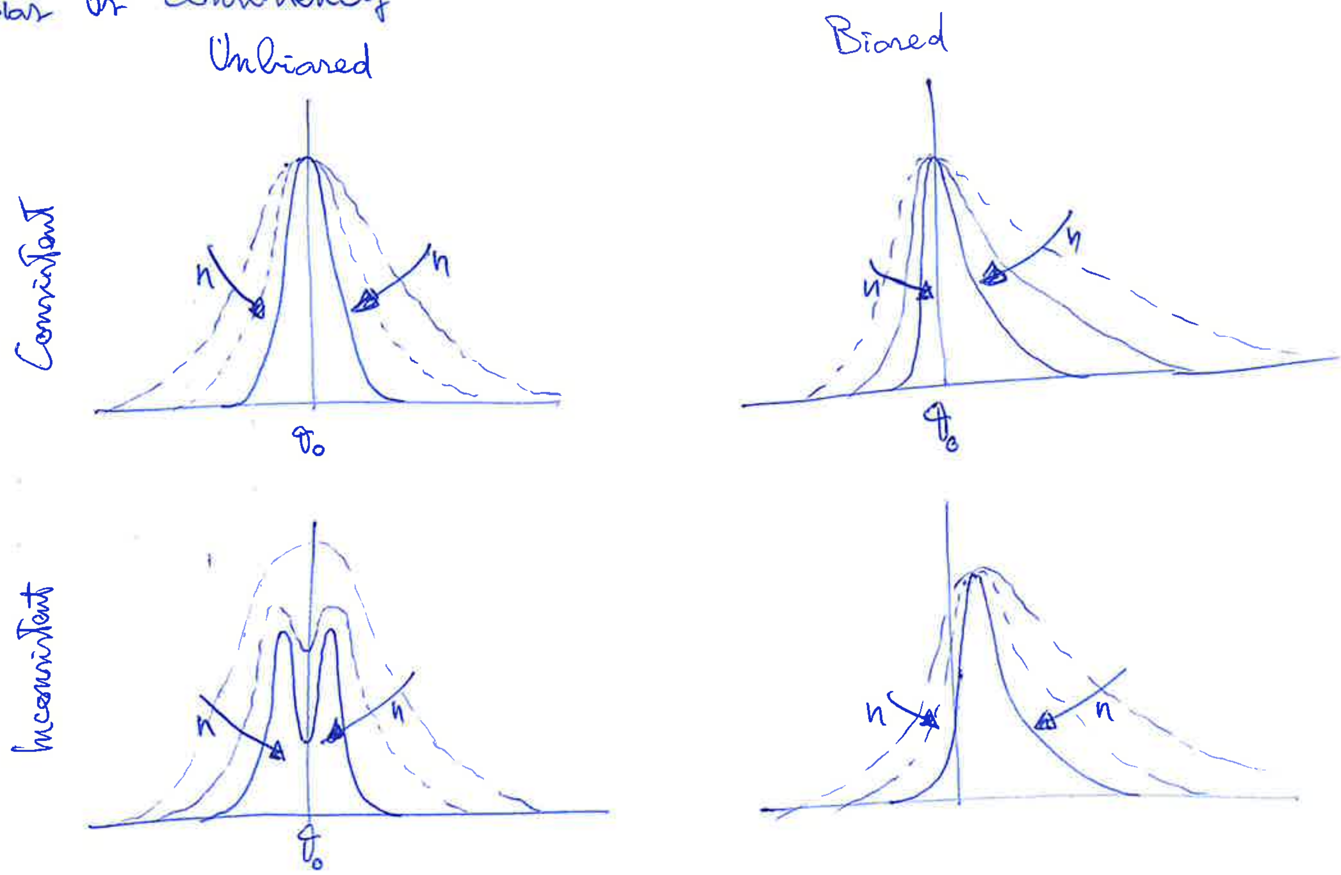
Example: If I measure the mass of an object once, the measured value is an estimate of the object mass:

$$\hat{m}(x) = x$$

Properties of estimators:
→ Consistency
→ Unbiasedness
→ Information content or efficiency
→ Robustness

Point 1

- Bias vs consistency



- Efficiency: minimum variance and Cramer-Rao inequality

Neglecting the bias, the smaller the variance of the estimator, the more certain we are that the estimate is near the true value of the parameter.

One can prove that the variance $V(\hat{\theta})$ of any consistent estimation is subject to a lower bound given by:

$$\boxed{V(\hat{\theta}) \geq \underbrace{\frac{\left(1+\frac{\partial b(\hat{\theta})}{\partial \theta}\right)^2}{E\left[\left(\frac{\partial \ln \mathcal{L}(\vec{x}|\theta)}{\partial \theta}\right)^2\right]}}_{V_{CR}(\hat{\theta})}}$$

→ bias of estimator

→ Fisher information

We define the efficiency of the estimator as: $\varepsilon(\hat{\theta}) = \dfrac{V_{CR}(\hat{\theta})}{V(\hat{\theta})}$

Any consistent estimator $\hat{\theta}$ has an efficiency which is at most equal to 1.

Print 3

- Properties of maximum likelihood:

We need to distinguish between → asymptotic properties → hold for sufficiently large n
↳ finite sample properties → hold for any n

→ max $\mathcal{L}$ estimators are consistent: asymptotically, one of the maxima will go arbitrarily close to the true value

→ max $\mathcal{L}$ estimators might be biased, but $\lim_{n\to\infty} b(\mathcal{L}) = 0$

→ max $\mathcal{L}$ estimators are asymptotically Normally distributed with minimum variance, and their variance is given by the Cramer-Rao lower bound.
↳ Therefore, the efficiency of max $\mathcal{L}$ estimators tends asymptotically to 1.
↳ max $\mathcal{L}$ estimators have asymptotically the lowest variance of any consistent estimator.

→ max $\mathcal{L}$ estimators are invariant under reparameterization.
If we reparametrize $\mathcal{L}$, ~~the transf~~ with $\tau(\theta)$, ~~we~~ we find $\hat{\tau} = \tau(\hat{\theta})$

→ For finite n, $\mathcal{L}$ might have multiple maxima, but we wouldn't know which is the closest to the true one.
↳ ~~However~~, in my experience this is a rare case that happens only when we have highly (anti)correlated parameters.

- Extended likelihood

Suppose we ~~measure~~ perform n measurements of a random variable x with PDF $f(x|\theta)$.
The number of observations n is itself a random variable, with distribution $P(n|\theta)$.
The likelihood needs to be extended to include $P(n|\theta)$:

$$\boxed{\mathcal{L} = P(n|\theta) \prod_{i=1}^{n} f(x_i|\theta)}$$

→ In most cases, $P(n|\theta)$ is a Poisson distribution whose average λ depends on the parameter(s) θ:

$$\mathcal{L} = \frac{e^{-\lambda}\lambda^n}{n!} \prod_{i=1}^{n} f(x_i|\theta) \quad \text{with } \lambda = \lambda(\theta)$$

Point 5

- Binned likelihood

Suppose the number of measurement n is so large that computing $\sum_{i=1}^{n} \ln f(x_i|\theta)$ would take too long.

We can simplify the problem (from the computational point of view) by binning the data n in $m \ll n$ bins.

If the data are ~~multi~~ vectors $\vec{x}$, we can do multidimensional bins.

The ~~likelihood will be a multinomial:~~

$$\cancel{\mathcal{L} = n! \prod_{\text{bin } i=1}^{m} \frac{p_i^{k_i}}{k_i!}}$$

~~where: i is the bin index (not the event index!)~~
~~k_i is the number of events in bin i~~
~~p_i is the expectation value for the number of counts in bin i~~
~~$\hookrightarrow p_i = \int_{\Delta x_i} f(x|\theta)\,dx$ $\quad \Delta x_i$ = bin width~~

The likelihood will be a multinomial, times an extended term for the total number of measurements n:

$$\mathcal{L} = P(n|\theta)\left[\prod_{i=1}^{m} \frac{p_i^{k_i}}{k_i!}\right] n!$$

where: i is the bin index (not the event index!)
k_i is the number of events in bin i
p_i is the probability associated to bin i
$\sum k_i = n$

$\rightarrow$ If $P(n|\theta)$ is a Poisson distribution with expectation value λ:

$$\lambda = \sum_{i=1}^{m} \lambda_i \qquad \lambda_i = \text{expected number of events in bin } i$$

$$p_i = \frac{\lambda_i}{\lambda} \qquad \lambda_i \stackrel{!}{=} \lambda \cdot \int_{\Delta x_i} f(x|\theta)\,dx \qquad \Delta x_i = \text{width of bin } i$$

$$\mathcal{L} = \frac{e^{-\lambda}\lambda^n}{n!}\, n! \cdot \prod_{i=1}^{m} \left(\frac{\lambda_i}{\lambda}\right)^{k_i} \frac{1}{k_i!} = e^{-\lambda} \cancel{\lambda^n} \prod_{i=1}^{m} \frac{\lambda_i^{k_i}}{\cancel{\lambda^{k_i}} \cdot k_i!}$$

$$\boxed{\mathcal{L} = \prod_{i=1}^{m} \frac{e^{-\lambda_i}\lambda_i^{k_i}}{k_i!}}$$

$\Rightarrow$ The binned version of an extended $\mathcal{L}$ with a Poisson ~~distr~~ distribution for n, is the product of a Poisson term for each bin!

Notice: This derivation is different than the one reported in literature, e.g. in the Cowan.

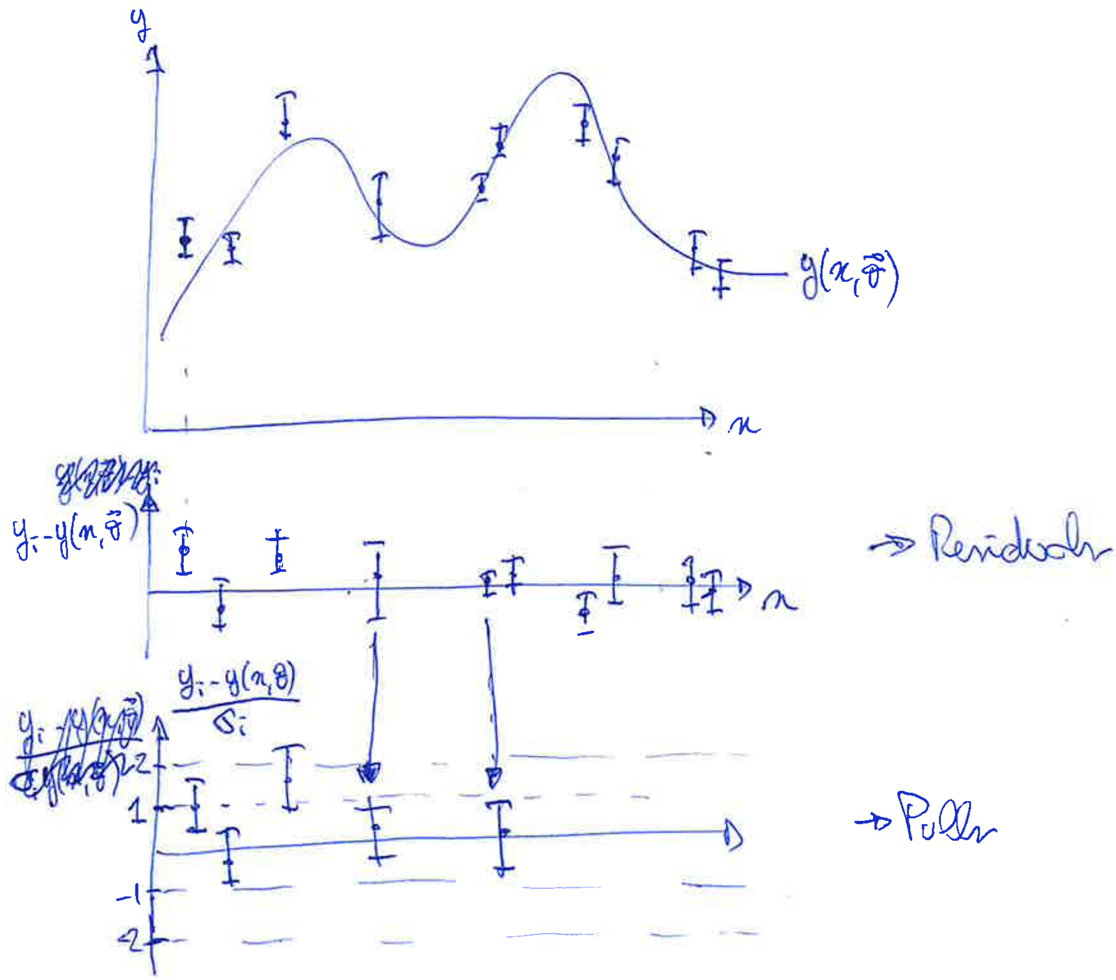
Point 7

Notice that: → If the parameters $\vec{\theta}$ are known, this is truly a χ^2 distr.

→ If $\vec{\theta}$ is not known, that combination of $\vec{\theta}$ values that minimizes $\sum \frac{(y_i - y(\,))^2}{\sigma^2}$ ~~...~~ is a χ^2 distribution only ~~...~~ if all measurements y_i are uncorrelated.

→ If the individual measurements y_i are not normally distributed, or if they are correlated, or if $y(x, \vec{\theta})$ does not perfectly describe the data, the least-square method can still be used ~~...~~ for estimating parameters, but it will not behave as a χ^2.

↳ This can actually be used to test the goodness of fit!



Point 9

• Numerical and hystorical considerations

Often times, we deal with histograms rather than scatter plots.
In the past, the computational power was not enough to numerically maximize $\mathcal{L}$,
~~On the other hand,~~ especially when dealing with factorials.
On the other hand, it is much easier to minimize the sum of squares.
This has lead to the ~~use of the~~ application of the least-squares method
to the fit of histograms.

• Least squares method for fitting histograms

Assume we have a set of measurements x_i following a PDF $f(x_i|\vec{\theta})$.
~~Assume we bin the spectrum~~ Assume we bin the data, and that every bin has a number of entries k_i large enough so that the corresponding Poisson distribution can be approximated by a Gaussian ~~with variance given by the ... expected number of counts in that bin.~~
with mean given by the expected number of counts in that bin:

$$\lambda = \text{expected number of total counts}$$

$$\lambda_i = \int_{\delta x_i} \lambda f(x_i|\vec{\theta})\, dx$$

The likelihood reads: $$\mathcal{L} = \prod_{\text{bin } i=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(\frac{-(k_i - \lambda_i(\vec{\theta}))^2}{2\sigma_i^2}\right)$$

Taking the logarithm: $$-2\ln\mathcal{L} = \sum_{i=1}^{m} \ln(2\pi\sigma_i^2) + \sum_{i=1}^{m} \frac{(k_i - \lambda_i(\vec{\theta}))^2}{\sigma_i^2}$$

The problem is, we don't know σ_i.

We can make some considerations:
→ σ_i^2 is the variance of the number of counts (measured or expected) in bin i. If this is large, it is equal to the number of counts.
But the logarithm of a large number is small and varies very slowly, so we can neglect the term $\sum \ln(2\pi\sigma_i^2)$.

$\boxed{\text{Point 11}}$

• Point estimation in practice

So, how do we minimize the χ^2 or maximize $\mathcal{L}$?

Every programming language has a tool (or many tools) for performing χ^2 fits or ~~Poisson~~ $\mathcal{L}$ fits with Poisson ~~statist~~ models.

The issue ~~ma~~ arises when we want to tweak the function that we want to minimize for precision or economy (speed) reasons.

Possible tools are:

→ ROOT (C++ and Python) ~~→ Default fit~~
root.cern.ch
→ Available χ^2 fit for scatter plots (called TGraphErrors)
→ Default χ^2_N for histograms
→ χ^2_P fit with option "P"
→ $\mathcal{L}$ fit with option "L"
→ Custom $\mathcal{L}$ possible, but complicated

→ BAT (C++ and ROOT) → It's actually a framework for Bayesian fits, but it can be easily forced to perform frequentist ones.
github.com/bat/bat
→ Nice feature: the user must implement the $\mathcal{L}$. Then one can use model→FindMode() to call the underlying ROOT fitter (Minuit) using the custom $\mathcal{L}$.

→ numpy + scipy.optimize (Python) → easy to define custom $\mathcal{L}$ which is ~~calle~~ passed to "minimize"

→ Julia → minimizer available e.g. from "Optim" ~~~~ package
→ can minimize custom $\mathcal{L}$

- Efficiency-corrected likelihood

Suppose each "true" event has an efficiency $\varepsilon(\vec{x})$ to be detected.
ε depends on the parameters $\vec{\theta}$ directly, if we fit some efficiency curve, or indirectly, i.e. through $\vec{x}$.

The ~~ev~~ PDF of detected events $\vec{x}$ will be: $\varepsilon(\vec{x}|\vec{\theta})\, f(\vec{x}|\vec{\theta})$

The likelihood will be: $\mathcal{L} = \frac{e^{-\lambda}\lambda^n}{n!} \prod_i \varepsilon(\vec{x}|\vec{\theta})\, f(\vec{x}|\vec{\theta})$

→ Example: Looking for peak near threshold

Suppose we have an X-ray detector with the following properties:

Energy resolution $\sigma = 1$ keV

Trigger Efficiency: $\varepsilon(E, t, \sigma_t) = \frac{p}{2}\left[1 + \mathrm{erf}\left(\frac{E-t}{\sigma_t}\right)\right]$ with $p = 0.9$, $t = 3$ keV, $\sigma_t = 1$ keV

Suppose we have an exponential background:

$$f_b(E|\lambda) = \lambda \exp(-\lambda E) \qquad \text{with } \lambda = 3 \text{ keV}$$

Suppose we look for an X-ray signal peak at 5 keV:

$$f_s(E|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(E-\mu)^2}{2\sigma^2}\right] \qquad \text{with } \mu = 5 \text{ keV}, \ \sigma = 1 \text{ keV}$$

Our likelihood will be:

$$\cancel{\mathcal{L} = \frac{e^{-\nu}\nu^n}{n!} \prod_{i=1}^{n} \left[\frac{s}{s+b}\varepsilon(E_i|t,\sigma_t) f_s(E_i|\mu,\sigma) + \frac{b}{s+b} f_b(E_i|\lambda)\right]} \quad \text{with } \nu = s+b$$

$$\mathcal{L} = \frac{e^{-\nu}\nu^n}{n!} \prod_{i=1}^{n} \frac{\frac{s}{s+b}\varepsilon(E_i|t,\sigma_t) f_s(E_i|\mu,\sigma) + \frac{b}{s+b} f_b(E_i|\lambda)\,\varepsilon(E_i|t,\sigma_t)}{\underbrace{\int\left[\frac{s}{s+b}\varepsilon f_s + \frac{b}{s+b}\varepsilon f_b\right] dE}_{\text{does not depend on } E_i}} \qquad \text{with } \nu = s+b$$

$$\mathcal{L} \propto \frac{e^{-\nu}\nu^n}{n!} \prod \left[\frac{s}{s+b}\varepsilon f_s + \frac{b}{s+b}\varepsilon f_b\right]$$

$\boxed{\text{Point 15}}$

The likelihood for the efficiency measurement is:

$$\cancel{\mathcal{L}(\vec{k} \mid \mu_\varepsilon, \sigma_\varepsilon, p) = \prod_{i=1}^{n_k} \frac{n!}{k_i!(n-k_i)!} p^{k_i}(1-p)^{n-k_i} \text{ with } n=}$$

$$\mathcal{L}(\vec{k}(E) \mid p, \mu_\varepsilon, \sigma_\varepsilon) = \prod_{i=1}^{n_k} \frac{n!}{k_i!(n-k_i)!} \varepsilon(E_i)^{k_i}\left(1-\varepsilon(E_i)\right)^{n-k_i}$$

with $n=1000$ injected events

$E_i = 1, 2, \ldots, 20$ keV

$n_k = 20$ energy points

The likelihood for the physics data $\vec{E}$ is:

$$\mathcal{L}(\vec{E} \mid s, b, p) = \frac{e^{-\lambda}\lambda^{n_E}}{n_E!} \prod_{i=1}^{n_E} \left[\frac{\varepsilon(E)\, s}{\lambda} f_s(E_i) + \frac{\varepsilon(E)\, b}{\lambda} f_b(E_i) \right]$$

with $\lambda = \cancel{p(s+b)} \sum_{i=1}^{n_E} \varepsilon(E_i)(s+b)$

n_E = number of detected events

s = expected value for signal events

b = expected value for background events

The combined likelihood is:

$$\mathcal{L}(\vec{k}, \vec{E} \mid p, \mu_\varepsilon, \sigma_\varepsilon, s, b) = \mathcal{L}(\vec{k} \mid p, \mu_\varepsilon, \sigma_\varepsilon) \cdot \mathcal{L}(\vec{E} \mid s, b, p)$$

$\boxed{\text{Point 17}}$