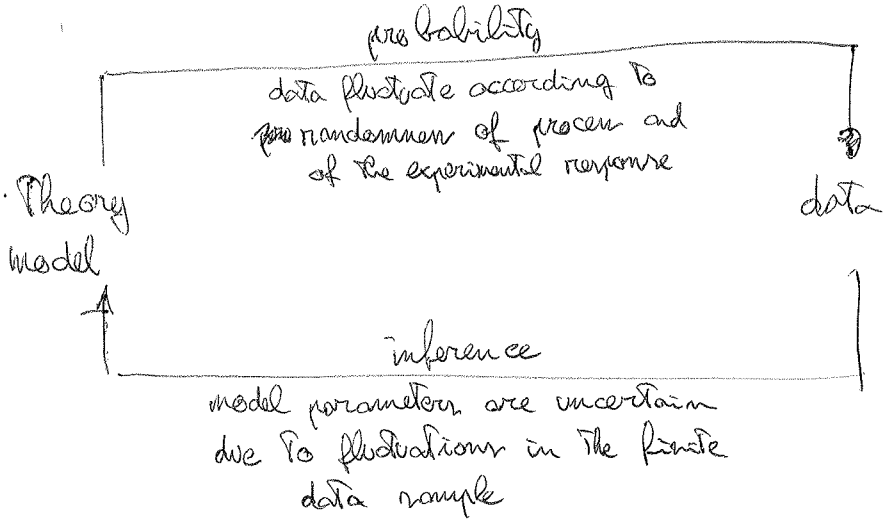
Point Estimation

~~Part of this applies to both the frequentist and Bayesian approach~~
Part of this in 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.



- Estimator and estimate
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

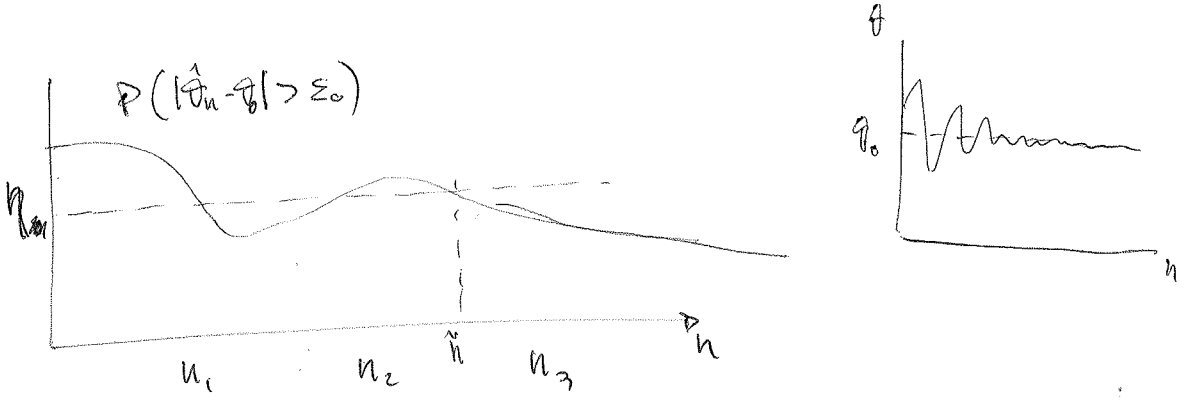
- Consistency

An estimator is consistent if it converges in probability to the true value of the unknown parameter θ_0.

$$\forall \varepsilon > 0 \quad : \quad \lim_{n \to \infty} P\left(|\hat{\theta}_n - \theta_0| < \varepsilon\right) = 1$$

In other words, if $\forall \varepsilon > 0$ and $\eta > 0$, there is an $\tilde{n}$ such that

$$P\left(|\hat{\theta}_n - \theta_0| > \varepsilon\right) < \eta \qquad \forall n > \tilde{n}$$



Notice: → The law of large numbers states that the sample mean is a consistent estimator of the parent mean.

→ Consistency is an asymptotic property. It does not imply that precision is a monotonic function of n.

- Bias

The bias of an estimator is the expected value of the deviation of the parameter estimate from the corresponding true value θ_0

Assuming n observations:

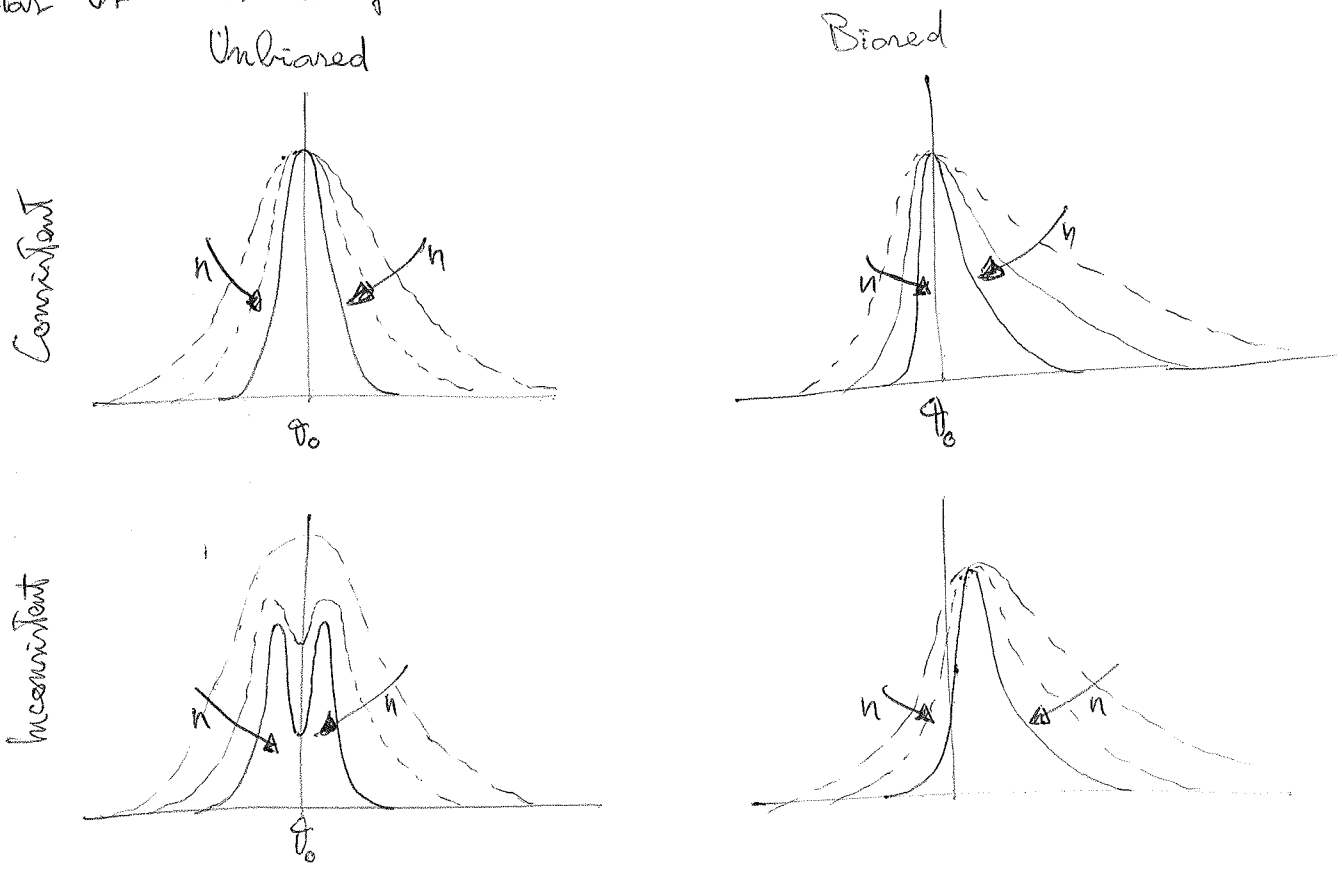
$$b_n(\hat{\theta}) = E(\hat{\theta}) - \theta_0 = E(\hat{\theta} - \theta_0)$$

An estimator is unbiased if $\forall n$ and $\forall \theta_0$: $\quad b_n(\hat{\theta}) = 0$

$$E(\hat{\theta}) = \theta_0$$

Point 2

- 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:

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

→ bias of estimator

→ Fisher information

$$\Downarrow$$

$$V_{CR}(\hat{\theta})$$

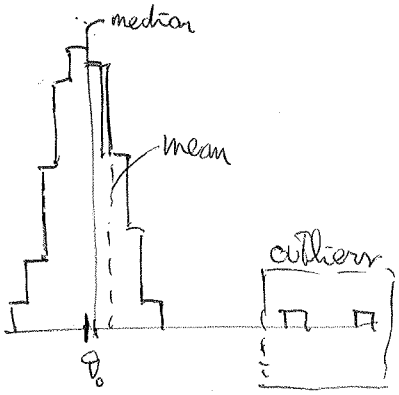
We define the efficiency of the estimator as: $\mathcal{E}(\hat{\theta}) = \frac{V_{CR}(\hat{\theta})}{V(\hat{\theta})}$

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

• Robustness

A good estimator should be not ~~very~~ too sensitive if the real distribution has some small deviation from the assumed PDF model.

Entries in the data that introduce visible deviations from the assumed PDF are called outliers.



Example: Suppose we take n measurements of some quantity x, distributed as in the figure.
The median ~~estimator~~ estimator is more robust than the mean estimator.

Trick: Remove outliers and compute the average on the "clean" sample.
This is called "Trimmed average".
It works, as long as we are confident that we are removing only outliers, and that the outlier ~~population~~ sample does not overlap too much with the good sample.

• Maximum likelihood estimator

The maximum likelihood estimate of the parameter θ is that value $\hat{\theta}$ for which $\mathcal{L}(\vec{x}|\theta)$ is maximal, given the observed data $\vec{x}$.

Finding the maximum likelihood is also called finding the "best fit", because it determines the value of the parameter that best fits the data.

Suppose we have n measurements consisting of m values of the data, so that the total sample is $\vec{x} = \{(x_1^1, \dots, x_m^1), \dots, (x_1^n, \dots, x_m^n)\}$

The likelihood is: $$\mathcal{L}(\vec{x}|\vec{\theta}) = \prod_{i=1}^{n} f(x_1^i, \dots, x_m^i | \vec{\theta})$$

Since the logarithm is monotonic, finding the $\max(\mathcal{L})$ is equivalent to finding $\max(\ln \mathcal{L})$. The advantage is that we can turn the product into a sum:

$$\ln \mathcal{L} = \sum_{i=1}^{n} f(\vec{x}_i | \vec{\theta})$$

So we need an algorithm that maximizes $\ln \mathcal{L}$, or minimizes $-\ln \mathcal{L}$.

Point 4

- 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~~ 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 where average λ depends on the parameter(s) θ:

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

→ A common case in when the events can be induced by a signal or background,

where: signal = new physics that we're interested in

background = known physics that we wish wouldn't be there to disturb our measurement

The signal and background will have their PDFs f_s and f_b.

The overall PDF will be,

$$f(x|\vec{\theta}) = \frac{s}{s+b} f_s(x|\vec{\theta}) + \frac{b}{s+b} f_b(x|\vec{\theta})$$

where s and b are the expected number of ~~events~~ signal and background events.

Therefore, the expectation value for n will be $\lambda = s+b$.

In the end we have:

$$\mathcal{L} = \frac{(s+b)^n e^{-(s+b)}}{n!} \prod_{i=1}^{n} \frac{s f_s(x|\vec{\theta}) + b f_b(x|\vec{\theta})}{s+b}$$

$$= \frac{e^{-(s+b)}}{n!} \prod_{i=1}^{n} \left[s f_s(x|\vec{\theta}) + b f_b(x|\vec{\theta}) \right]$$

Taking the logarithm:

$$\ln \mathcal{L} = -s - b - \ln(n!) + \sum_{i=1}^{n} \ln \left[s f_s(x|\vec{\theta}) + b f_b(x|\vec{\theta}) \right]$$

↳ Notice that s and b are parameters of the model, in addition to the parameters of f_s and f_b (if any).

↳ Notice that the term $\ln(n!)$ does not depend on any parameter, so it can be omitted from the maximization.

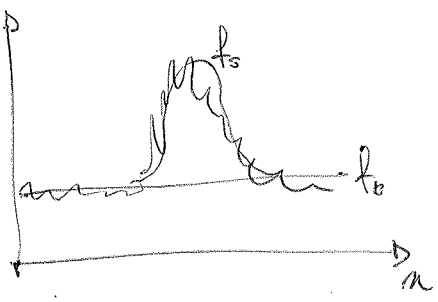
↳ Further simplifications might be possible, depending on the actual form of f_s and f_b.

Example: Extended likelihood fit of Gaussian peak over flat background

$$f_s = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

$$f_b = \frac{1}{\Delta x} \qquad \Delta x = \text{fit range (fixed)}$$

Parameters: $\vec{\theta} = \{s, b, \mu, \sigma\}$



Point 6

- 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 x in $m << n$ bins.

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

The likelihood will be a multinomial:

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

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
$\hookrightarrow p_i = \int_{\Delta x_i} f(x|\theta)\,dx$ Δx_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
$\sum k_i = n$

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

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

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

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

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

$\Rightarrow$ The binned version of an extended $\mathcal{L}$ with a Poisson ~~PDF~~ 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

• ~~Minimizer algorithms~~

→ ~~Minimizer algorithms available in almost pretty much any programming language~~

• Least - squares method

Sometimes this is also referred as χ^2 method, although ~~the~~ the squared residuals behave as a χ^2 only in specific circumstances.

Consider a set of measurements with ~~symmetric~~ Gaussian uncertainties $y_i \pm \sigma_i$, $i = 1, ..., n$.
Assume each y_i corresponds to a perfectly known x_i.

This would be for example the case of a calibration function, where x is the literature energy value of γ lines, and y is the amplitude measured by our detector.

Assume the relation between y and x is: $y = y(x, \vec{\theta})$

Notice: ~~...~~ $y(x, \vec{\theta})$ is just a function, not a PDF.

If the measurements y_i are distributed around the curve $y(x, \vec{\theta})$ according to a Gaussian distribution with STD σ_i, the likelihood will be a product of Gaussian PDFs:

$$\mathcal{L}(\vec{y}\,|\,\vec{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(\frac{-\left(y_i - y(x_i, \vec{\theta})\right)^2}{2\sigma_i^2}\right)$$

Maximizing $\mathcal{L}$ is equivalent to minimize $-2 \ln \mathcal{L}$:

$$-2 \ln \mathcal{L} = \sum_{i=1}^{n} \frac{\left(y_i - y(x_i, \vec{\theta})\right)^2}{\sigma_i^2} + 2 \sum_{i=1}^{n} \ln\left(2\pi\sigma_i^2\right)$$

First term: This is a sum of standard normal variables, so it is a χ^2 distro!

Second term: → does not depend on $\vec{\theta}$, so we can drop it

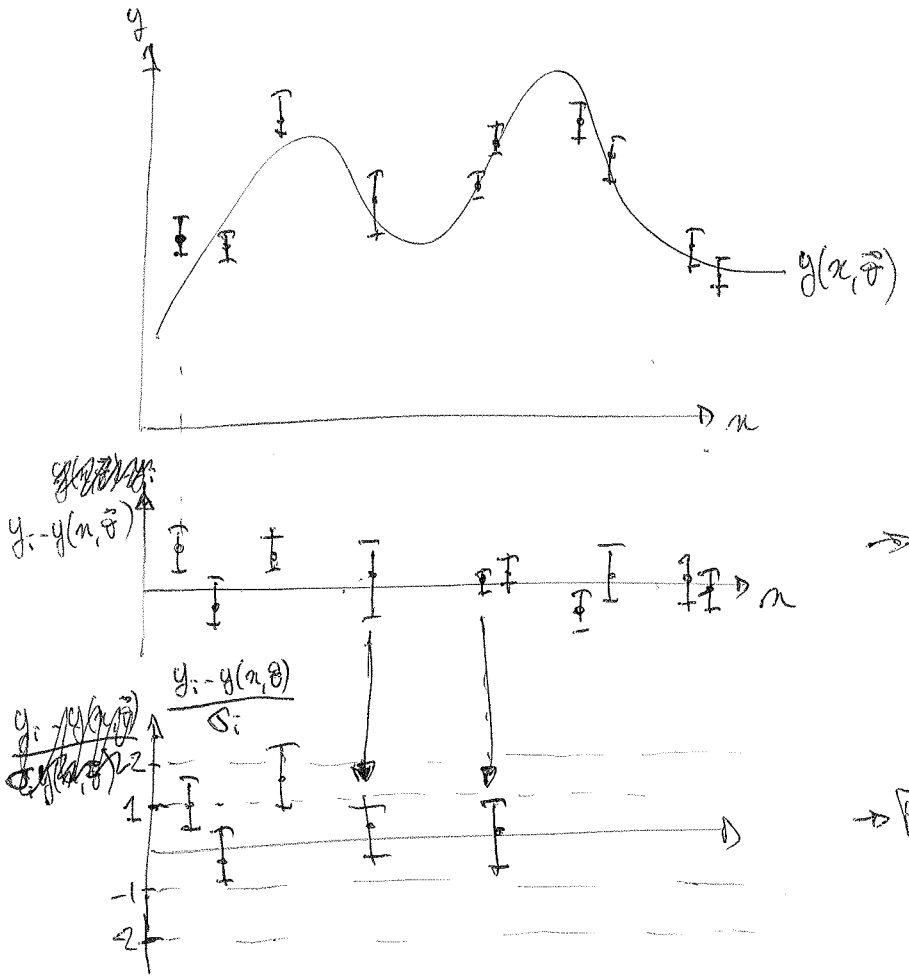
$$\chi^2(\vec{\theta}) = \sum_{i=1}^{n} \frac{\left(y_i - y(x_i, \vec{\theta})\right)^2}{\sigma_i^2}$$

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

→ If $\vec{\theta}$ is not known, that combination of $\vec{\theta}$ value 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

~~• Least squares for binned histograms~~

• Linear regression

Suppose you have a set of measurements y_i with Gaussian uncertainties σ_i.

Suppose $y(x, \vec{\theta})$ is of the form: $y = ax + b$ with $\vec{\theta} = (a, b)$.

The χ^2 is: $$\chi^2(a,b) = \sum_{i=1}^{n} \frac{(y_i - a x_i - b)^2}{\sigma_i^2}$$

For convenience, let's introduce weights: $$w_i = \frac{\frac{1}{\sigma_i^2}}{\frac{1}{\sigma^2}} \qquad \frac{1}{\sigma^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}$$

So that: $\sum w_i = 1$

We can minimize χ^2 analitically by imposing: $$\begin{cases} \frac{\partial \chi^2}{\partial a} = 0 \\ \frac{\partial \chi^2}{\partial b} = 0 \end{cases}$$

This corresponds to: $$\begin{cases} 2 \sum w_i (y_i - a x_i - b) x_i = 0 \\ 2 \sum w_i (y_i - a x_i - b) = 0 \end{cases}$$

~~$\sum w_i y_i = \sum$~~

$$\begin{cases} \sum w_i x_i y_i = a \sum w_i x_i^2 + b \sum w_i x_i^2 \\ \sum w_i y_i = a \sum w_i x_i + b \end{cases}$$

Which can be easily solved for a and b, obtaining:

$$\hat{a} = \bar{y} - b \bar{x}$$

$$\hat{b} = \frac{\mathrm{cov}(x, y)}{V[x]}$$

→ A similar approach can be used (sometimes) ~~for solve~~ for nonlinear functions

→ Otherwise we need numerical methods.

- Numerical and historical 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 k_i, 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: ~~xxxx~~
→ 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 $\sum \ln(2\pi\sigma_i^2)$.

Point 11

At this point we have: $\chi^2 = \sum_{i=1}^{n} \frac{(k_i - \lambda_i(\vec{\theta}))^2}{\sigma_i^2}$

Now we can make two approximations/assumptions:

→ Neyman's χ^2: $\sigma_i^2 = k_i$ = measured number of counts in bin i

$$\chi^2_N = \sum_{i=1}^{m} \frac{(k_i - \lambda_i(\vec{\theta}))^2}{k_i}$$

↳ Standard for many pre-defined fit methods, e.g. in ROOT / TMinuit

↳ Fails miserably if one bin is empty

→ Pearson's χ^2: $\sigma_i^2 = \lambda_i$ = expected number of counts in bin i

$$\chi^2_P = \sum_{i=1}^{m} \frac{(k_i - \lambda_i(\vec{\theta}))^2}{\lambda_i(\vec{\theta})}$$

↳ Available/optional in some pre-defined fit methods

↳ Works also for empty bins

↳ Also called "modified least squares method"

- Limitations of least-squares fits for binned histograms.

→ For bins with few events, k_i are not approximately normally distributed.

→ In presence of empty bins, the χ^2_N is usually modified so that empty bins are ignored. This leads to a significant bias.

→ The χ^2_P also has some (smaller) bias in the presence of empty bins.

⇒ Do not use χ^2 fits on histograms!

Point 12

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

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

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

- Exercises / examples:
  - → Unbinned vs Binned fit (extended vs binned likelihood)
  - → Scatter Plot
  - → Scatter plot with overestimated uncertainty
  - → Scatter plot with unaccounted systematic
  - → Likelihood vs χ^2 for Poisson distributed counts
  - → Efficiency curve fit with Binomial, Poisson and Gamma distribution
  - → ~~Radioactive decay fit (only N, N, $T_{1/2}$)~~

- Detection efficiency and weights

Whenever we decide a detector, we have an efficiency $\varepsilon < 1$ of detecting the events due to several reasons:

a) Geometry: not all parts of the detector are sensitive / active
b) Particle type: we might be sensitive only to some particles
c) Timing: we might have dead times, pile-up issues, ...
d) Noise: the signal induced by a particle might be happening in a period affected by high noise
e) Energy / amplitude: the efficiency has some roll-up curve as a function of the signal amplitude / energy.

We have two options: 1) Correct the data. This ~~has the~~ introduces the problem that we need to account for the implemented correction ~~in the~~ by assigning a corresponding uncertainty to the data points
2) Include the efficiency in the fit model. This is the exact method, and is preferable whenever possible.

Point 14

• 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 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 $p = 0.9$, $t = 2$ keV, $\sigma_t = 1$ keV

Suppose we have an exponential background:

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

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

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

Our likelihood will be:

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

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

(the integral in the denominator does not depend on E_i)

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

Point 15

• Simultaneous Fit

Suppose we have two sets of data $\vec{x}$ and $\vec{y}$ that ~~depend~~ have some parameters in common:

$$f(\vec{x}|\vec{\theta},\vec{\nu}) \qquad f(\vec{y}|\vec{\theta},\vec{\eta})$$

We can run a combined simultaneous fit simply by multiplying the two likelihoods:

$$\mathcal{L}(\vec{x},\vec{y}|\vec{\theta},\vec{\nu},\vec{\eta}) = \mathcal{L}(\vec{x}|\vec{\theta},\vec{\nu}) \cdot \mathcal{L}(\vec{y}|\vec{\theta},\vec{\eta})$$

Example: Combined efficiency and signal + bkg fit

Suppose we have a set of data $\vec{k}$ corresponding to the number of triggered events from pulser injected events.
We inject 1000 events with a pulse generator into a detector for 20 energy values between ~~200~~ and 1 and 20 keV, and count how many survive.
Assume the efficiency follows the curve:

$$\varepsilon(E) = \frac{p}{2}\left[1 + \text{erf}\left(\frac{E-\mu_\varepsilon}{\sigma_\varepsilon}\right)\right] \quad \text{with} \quad \begin{matrix} p = 0.9 \\ \mu_\varepsilon = 3 \text{ keV} \\ \sigma = 1 \text{ keV} \end{matrix} \Big\} \text{ Fit parameters}$$

Suppose then we measure some data $\vec{E}$ distributed as

a) A Gaussian signal: $f_s(\vec{E})_{\vec{\mu}} = \frac{1}{\sqrt{2\pi}\sigma}\text{e}^{\left(\frac{(E-\mu)^2}{2\sigma^2}\right)}$ with $\begin{matrix} \mu = 5 \text{ keV} \\ \sigma = 1.2 \text{ keV} \end{matrix}$ Fixed

b) A flat background: $f_b(\vec{E}) = \frac{1}{\Delta E}$ with $\Delta E = E_{max} - E_{min}$

$E_{min} = 1$ keV

$E_{max} = 20$ keV

~~Suppose the true~~

The likelihood for the efficiency measurement is:

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

$$\mathcal{L}(\vec{k_E} | p, \mu_\varepsilon, \sigma_\varepsilon) = \prod_{i=1}^{n_k} \frac{n!}{k_i!(n-k_i)!} \varepsilon(E_i)^{k_i} \left(1 - \varepsilon(E_i)\right)^{n-k_i} \quad \text{with } n = 1000 \text{ injected events}$$

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

$n_k = 20$ energy points

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

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

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

n_E = number of detected events

s = expected value for signal events

b = expected value for background events

The combined likelihood is:

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

Point 17