Tools for Physicists: Statistics

Parameter estimation and confidence intervals

Wolfgang Gradl

Institut für Kernphysik

Summer semester 2024

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

The scientific method: how we create ‘knowledge’

Theory / model

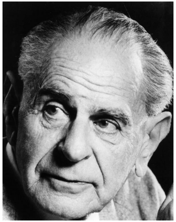
- usually mathematical
- self-consistent
- simple explanations, few (arbitrary) parameters
- testable predictions / hypotheses

Experiment

- modify or even reject theory in case of disagreement with data
- if theory requires too many adjustments it becomes unattractive
- generate surprises

Advance of scientific knowledge is *evolutionary* process with occasional revolutions

Statistical methods are important part of this process
in particular in quantitative sciences like physics



Karl Popper
(1902–1994)

Statistics in science

Statistics is needed to:

- characterise and summarise experimental results (impractical to always deal with raw data)
- quantify uncertainty of a measurement
- assess whether two measurements of the same quantity are compatible,
  combine measurements
- estimate parameters of an underlying model or theory
- test hypotheses:
  determine whether a model is compatible with data
- …

Aims of this mini-series

- Understand statistical concepts
  - Ability to understand physics papers
  - Know some methods / standard statistical toolbox
- Statistical inference: from data to knowledge
  - Should we believe a physics claim?
  - Develop intuition
  - Know (some) pitfalls: avoid making mistakes others have already made
- Use tools
  - Hands-on part with Python / Jupyter
  - Application to your own work? You decide!

JG|U

Practical information

Two sessions:

1. Basics, introduction, statistical distributions
2. Parameter estimation, confidence intervals, hypothesis testing

About 60–90 minutes of lecture, hands-on tutorial in your own time

I hope this will be useful for you,
but keep in mind that there is much more
to statistics than can be covered
in a few brief hours.

Useful reading material

Books:

- G. Cowan, Statistical Data Analysis
- R. Barlow, Statistics: A guide to the use of statistical methods in the physical sciences
- L. Lyons, Statistics for Nuclear and Particle Physicists
- A. J. Bevan, Statistical data analysis for the physical sciences
- G. Bohm, G. Zech, Introduction to Statistics and Data Analysis for Physicists (available online)

Lectures on the web:

- G. Cowan, Royal Holloway University London: Statistical Data Analysis
- K. Reygers, U Heidelberg, Stat. Methods in Particle Physics

Dealing with uncertainty

- Underlying theory is probabilistic (quantum mechanics / QFT)
  source of **true** randomness
- Limited knowledge about measurement process
  even without QM
  random measurement errors
- Things we could know in principle, but don't
  e.g. from limitations of cost, time, …

Quantify uncertainty using tools and concepts from **probability**

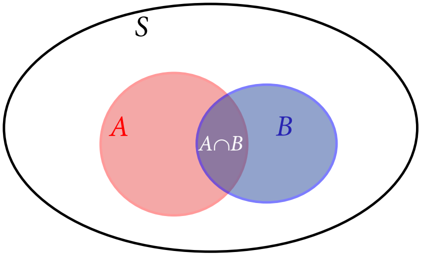
Mathematical definition of probability

Kolmogorov axioms:

Consider a set S (the **sample space**) with subsets A, B, …(**events**).

Define a function on the power set of S, $P : \mathfrak{P}(S) \mapsto [0, 1]$ with

1. $P(A) \geq 0$ for all $A \subset S$
2. $P(S) = 1$
3. $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$, i.e. when A and B are exclusive



From these we can derive further properties:

- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup \bar{A}) = 1$
- $P(\emptyset) = 0$
- If $A \subset B$, then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

for the mathematically inclined: proper treatment will use *measure theory*

Interpretation — intuition about probability

- Classical definition
  - Assign equal probabilities based on symmetry of problem, e.g. rolling ideal dice: $P(6) = 1/6$
  - difficult to generalise, sounds somewhat circular

- Frequentist: relative frequency, proportion of outcomes
  - $A, B, \ldots$ outcomes of a repeatable experiment

$$P(A) = \lim_{n\to\infty} \frac{\text{times outcome is } A \text{ in } n \text{ repetitions}}{n}$$

- Bayesian: subjective probability, degree of belief
  - $A, B, \ldots$ are hypotheses (statements that are either true or false)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

…all three definitions consistent with Kolmogorov's axioms

Conditional probability, independent events

Conditional probability for two events A and B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

"probability of A given B"

Example: rolling dice

$$P(n < 3|n \text{ even}) = \frac{P((n < 3) \cap (n \text{ even}))}{P(n \text{ even})} = \frac{1/6}{1/2} = 1/3$$

Events A and B independent $\iff P(A \cap B) = P(A) \cdot P(B)$

A is independent of B if $P(A|B) = P(A)$

Bayes' theorem

Use definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{and} \qquad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

But obviously $P(A \cap B) = P(B \cap A)$, so:

Theorem

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Allows to 'invert' statements about probability:
of great interest to us. Want to infer $P(\text{theory}|\text{data})$ from $P(\text{data}|\text{theory})$

Often these two are confused, knowingly or unknowingly
(advertising, political campaigns, …)

Bayes' theorem: degree of belief in a theory

likelihood

prior (before seeing the data, subjective)

$$P(\text{theory}|\text{data}) = \frac{P(\text{data}|\text{theory})P(\text{theory})}{P(\text{data})}$$

posterior probability, i.e., after seeing the data

normalization

Example for Bayes' theorem: Rare disease

Base probability (for anyone) to have a disease D:

$$P(D) = 0.0001$$
$$P(\text{no } D) = 0.9999$$

Example for Bayes' theorem: Rare disease

Base probability (for anyone) to have a disease D:

$$P(D) = 0.0001$$
$$P(\text{no } D) = 0.9999$$

Consider a test for D: result is positive or negative (+ or –):

$$P(+|D) = 0.98 \qquad P(+|\text{no } D) = 0.03$$
$$P(-|D) = 0.02 \qquad P(-|\text{no } D) = 0.97$$

JG|U

Example for Bayes' theorem: Rare disease

Base probability (for anyone) to have a disease D:

$$P(D) = 0.0001$$
$$P(\text{no } D) = 0.9999$$

Consider a test for D: result is positive or negative (+ or –):

$$P(+|D) = 0.98 \qquad P(+|\text{no } D) = 0.03$$
$$P(-|D) = 0.02 \qquad P(-|\text{no } D) = 0.97$$

Suppose your result is +; should you be worried?

Example for Bayes' theorem: Rare disease

Base probability (for anyone) to have a disease D:

$$P(D) = 0.0001$$
$$P(\text{no } D) = 0.9999$$

Consider a test for D: result is positive or negative (+ or –):

$$P(+|D) = 0.98 \qquad P(+|\text{no } D) = 0.03$$
$$P(-|D) = 0.02 \qquad P(-|\text{no } D) = 0.97$$

Suppose your result is +; should you be worried?

$$\begin{aligned} P(D|+) &= \frac{P(+|D)\,P(D)}{P(+|D)\,P(D) + P(+|\text{no } D)\,P(\text{no } D)} \\ &= \frac{0.98 \times 0.0001}{0.98 \times 0.0001 + 0.03 \times 0.9999} = 0.0033 \end{aligned}$$

Probability that you have disease is 0.32%, i.e. you're probably ok

Digression: what if prevalence is (much) higher?

Assume $100\times$ higher prevalence in population:

$$P(D) = 0.01$$
$$P(\text{no } D) = 0.99$$

Then,

$$\begin{aligned} P(D|+) &= \frac{P(+|D)\,P(D)}{P(+|D)P(D) + P(+|\text{no } D)P(\text{no } D)} \\ &= \frac{0.98 \times 0.01}{0.98 \times 0.01 + 0.03 \times 0.99} = 0.248 \end{aligned}$$

should you be worried? This can't be answered by statistics, of course …

At least take another (independent) test …

Classification

Population P that either carries (P) or does not carry (N) a specific marker

D or no D, signal candidate or background event, …

Classifier ("test"): predict positive (PP) or negative (PN) outcome

$+$ or $-$

Confusion matrix

		predicted	
		predicted pos.	**predicted neg.**
actual	**positive**	true positive	false negative
	negative	false positive	true negative

Type I error: false positive

Type II error: false negative

Classification

$$\text{sensitivity} = P(+|D) = \frac{\text{true positives}}{\text{actual positives}} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$\text{specificity} = P(-|\text{no } D) = \frac{\text{true negatives}}{\text{actual negatives}} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

Higher sensitivity: lower type II error rate

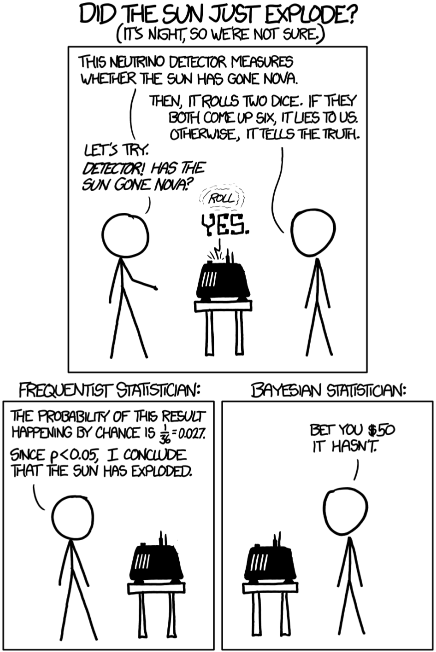
Higher specificity: lower type I error rate

Given a concrete classifier, how can we pick the 'best' threshold?

Criticisms — Frequentists vs. Bayesians

- Criticisms of the frequentist interpretation
  - $n \to \infty$ can never be achieved in practice. When is n large enough?
  - Want to talk about probabilities of events that are not repeatable
    - $P(\text{rain tomorrow})$ — but there's only one tomorrow
    - $P(\text{Universe started with a big bang})$ — only one universe available
  - P is not an intrinsic property of A, but depends on how the ensemble of possible outcomes was constructed
    - $P(\text{person I talk to is a physicist})$ strongly depends on whether I am at a conference or at the beach
- Criticisms of the subjective interpretation
  - 'Subjective' estimate has no place in science
  - How can quantify the prior state of our knowledge?

'Bayesians address the questions everyone is interested in by using assumptions that no one believes, while Frequentists use impeccable logic to deal with an issue that is of no interest to anyone'
— Louis Lyons



https://xkcd.com/1132/

Describing data

Random variables and probability density functions

Random variable:

- Variable whose possible values are numerical outcomes of a random phenomenon

Probability density function (pdf) of a continuous variable:

$$P(X \text{ found in } [x, x + \mathrm{d}x]) = p(x)\mathrm{d}x$$

Normalisation:

$$\int_{-\infty}^{+\infty} p(x)\mathrm{d}x = 1 \qquad x \text{ must be somewhere}$$

Visualisation: Histograms

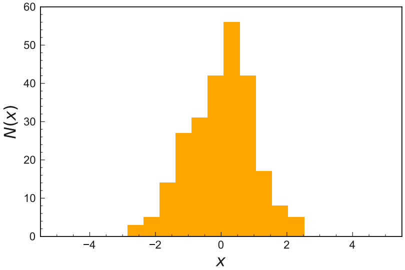
Histogram

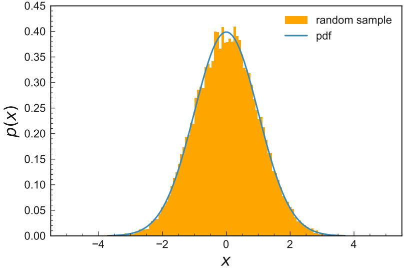
- representation of the frequencies of numerical outcome of a random phenomenon

pdf $\simeq$ histogram for

- infinite data sample
- zero bin width
- normalised to unit area

$$p(x) = \lim_{\Delta x \to 0} \frac{N(x)}{N \Delta x}$$





Median, mean, and mode

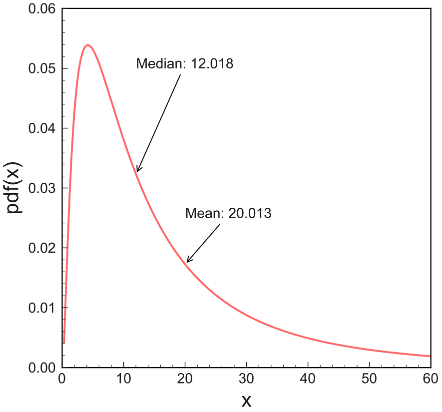
Arithmetic **mean** of a data sample ('sample mean'):

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Mean of a pdf:

$$\mu \equiv \langle x \rangle \equiv \int x\, p(x) \mathrm{d}x$$

$$\equiv \text{ expectation value } E[x]$$



Median:
point with 50% probability above and 50% prob. below

Mode:
most likely value

not necessarily the same, for skewed distributions

Variance, standard deviation

Variance of a **distribution** (pdf):

$$V(x) = \int \mathrm{d}x\, p(x)\, (x-\mu)^2 = E[(x-\mu)^2]$$

Variance of a **data sample**

$$V(x) = \frac{1}{N}\sum_i (x_i - \mu)^2 = \overline{x^2} - \mu^2$$

Requires knowledge of *true* mean μ.

Replacing μ by sample mean $\bar{x}$ results in underestimated variance!

Instead, use this:

$$\hat{V}(x) = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2$$

Standard deviation:

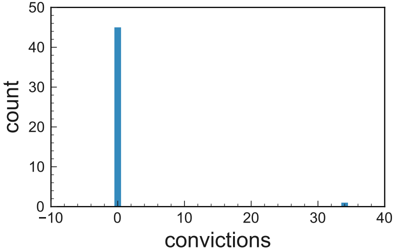
$$\sigma = \sqrt{V(x)}$$

Robustness?

Beware of distributions with large outliers:

Sample mean and variance as defined above not very good ('robust') estimators for the shape of the bulk of the distribution, can be grossly misleading!

Robust statistics deals with methods how to handle this — for a short writeup and pointers to literature, see *e.g.* `https://www.stats.ox.ac.uk/~ripley/StatMethods/Robust.pdf`



As of 31^{st} May 2024, the average US president has been convicted of 0.74 felonies

Multivariate distributions

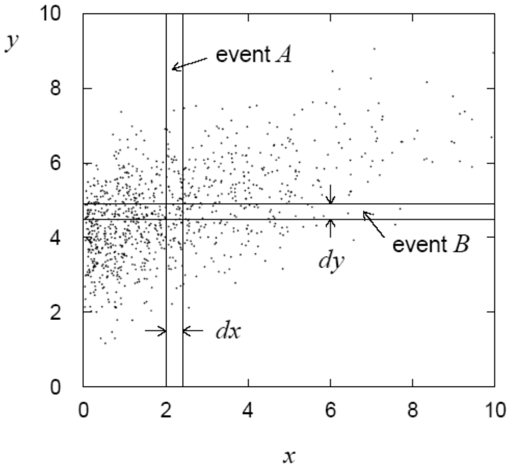
Outcome of an experiment
characterised by tuple $(x_1, \ldots, x_n)$

$$P(A \cap B) = f(x,y)\mathrm{d}x\,\mathrm{d}y$$

with $f(x,y)$ the 'joint pdf'

Normalisation

$$\int \cdots \int f(x_1, \ldots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n = 1$$



Sometimes, only the pdf of one component is wanted:

$$f_1(x_1) = \int \cdots \int f(x_1, \ldots, x_n)\mathrm{d}x_2 \cdots \mathrm{d}x_n$$

≈ projection of joint pdf onto individual axis: marginalised pdf

Covariance and correlation

Covariance:

$$\mathrm{cov}[x,y] = E[(x-\mu_x)(y-\mu_y)]$$

Correlation coefficient:

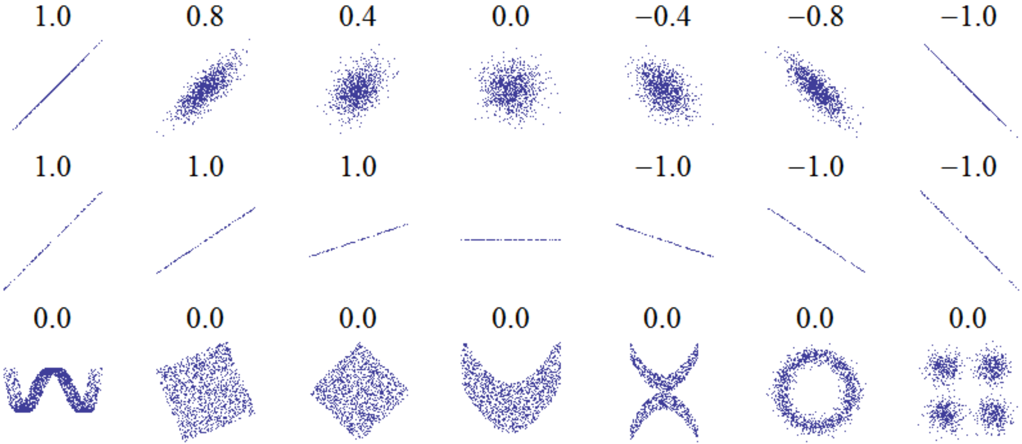
$$\rho_{xy} = \frac{\mathrm{cov}[x,y]}{\sigma_x\,\sigma_y}$$

If x, y independent:
pdf factorises, *i.e.* $f(x,y) = f_x(x)\, f_y(y)$,
and covariance becomes

$$E[(x-\mu_x)(y-\mu_y)] = \int (x-\mu_x) f_x(x)\mathrm{d}x \int (y-\mu_y) f_y(y)\mathrm{d}y = 0$$

Note: converse not necessarily true

Covariance and correlation



Same (linear) correlation coefficient, but very different 2D shapes!

Always visualise your data!

```python
import pandas as pd ...
```

```python
dataset = pd.read_csv('./ds.csv', header=None, names=['x', 'y'])
dataset
```

[2] ✓ 0.7s Python

	x	y
0	55.3846	97.1795
1	51.5385	96.0256
2	46.1538	94.4872
3	42.8205	91.4103
4	40.7692	88.3333
...	...	...
137	39.4872	25.3846
138	91.2821	41.5385
139	50.0000	95.7692
140	47.9487	95.0000
141	44.1026	92.6923

142 rows × 2 columns

Always visualise your data!

```python
dataset.describe()
```

[12] ✓ 0.7s Python

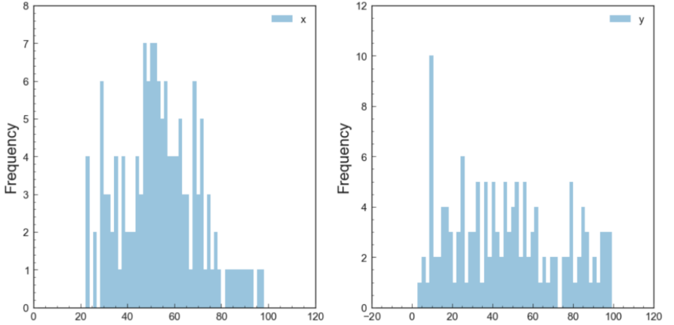
...

	x	y
count	142.000000	142.000000
mean	54.263273	47.832253
std	16.765142	26.935403
min	22.307700	2.948700
25%	44.102600	25.288450
50%	53.333300	46.025600
75%	64.743600	68.525675
max	98.205100	99.487200

Always visualise your data!

```python
fig,axs = plt.subplots(1,2,figsize=(16,8))
dataset.plot('x', kind='hist', bins=50, alpha=0.5, ax=axs[1])
dataset.plot('y', kind='hist', bins=50, alpha=0.5, ax=axs[0]);
```

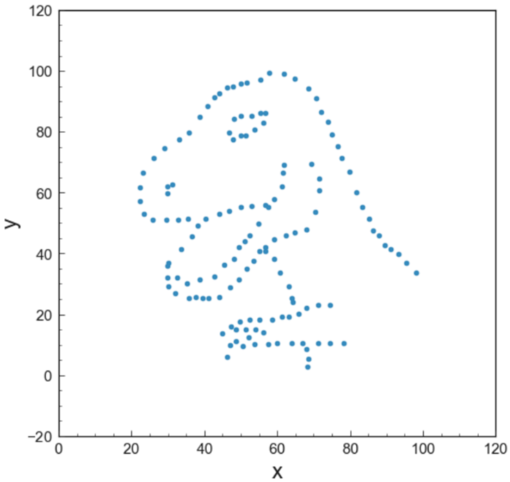
[13] ✓ 1.1s Python



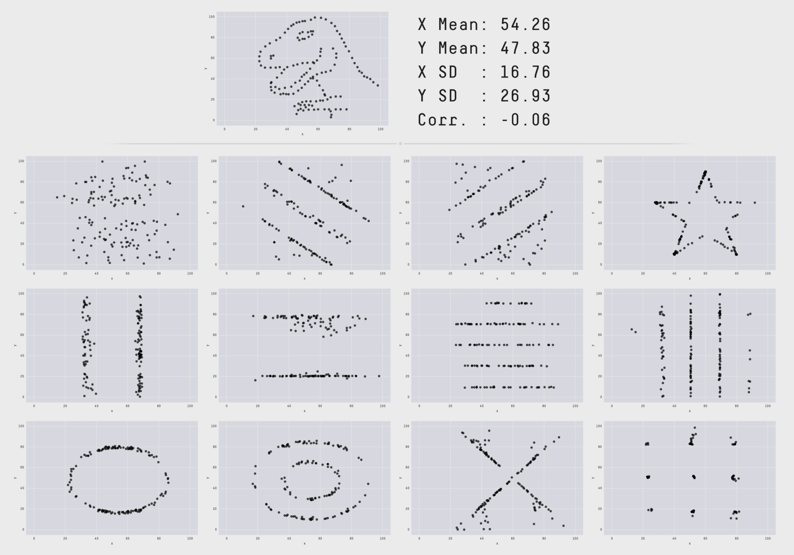
Always visualise your data!

```python
fig, ax = plt.subplots(1,1,figsize=(8,8))
dataset.plot.scatter(x='x', y='y', ax=ax);
```

[16] 0.3s Python



Always visualise your data!



https://www.autodesk.com/research/publications/same-stats-different-graphs

Linear combinations of random variables

Consider two random variables x and y with known covariance $\mathbf{cov}[x, y]$

$$\langle x + y \rangle = \langle x \rangle + \langle y \rangle$$
$$\langle ax \rangle = a \langle x \rangle$$
$$V[ax] = a^2 V[x]$$
$$V[x + y] = V[x] + V[y] + 2\,\mathbf{cov}[x, y]$$

For uncorrelated variables, simply add variances.
How about combination of N independent measurements (estimates) of a quantity, $x_i \pm \sigma$, all drawn from the same underlying distribution?

$$\bar{x} = \frac{1}{N} \sum x_i \quad \text{best estimate}$$
$$V[N\bar{x}] = N^2 \sigma$$
$$\sigma_{\bar{x}} = \frac{1}{\sqrt{N}} \sigma$$

Combination of measurements: weighted mean

Suppose we have N independent measurements of the same quantity, but each with a different uncertainty: $x_i \pm \delta_i$

Weighted sum:

$$x = w_1 x_1 + w_2 x_2$$
$$\delta^2 = w_1^2 \delta_1^2 + w_2^2 \delta_2^2$$

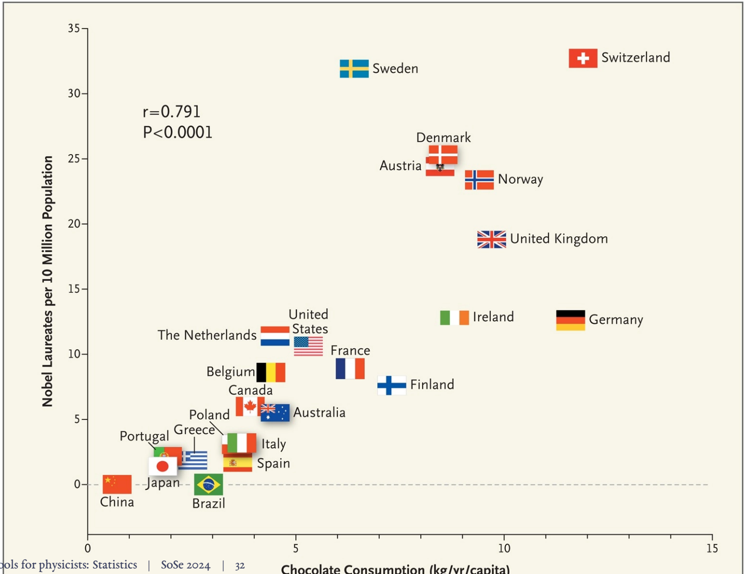
Determine weights w_1, w_2 under constraint $w_1 + w_2 = 1$ such that δ^2 is minimised:

$$w_i = \frac{1/\delta_i^2}{1/\delta_1^2 + 1/\delta_2^2}$$

If original raw data of the two measurements are available, can improve this estimate by combining raw data

alternatively, use log-likelihood curves to combine measurements

Correlation ≠ causation



Correlation coefficient: 0.791

significant correlation
($p < 0.0001$)

0.4 kg/year/capita to produce one additional Nobel laureate

improved cognitive function associated with regular intake of dietary flavonoids?

Some important distributions

Binomial distribution

N independent experiments

- Outcome of each is either ‘success’ or ’failure’
- Probability for success is p

$$f(k;N,p) = \binom{N}{k} p^k (1-p)^{N-k} \qquad E[k] = Np \qquad V[k] = Np(1-p)$$

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

binomial coefficient: number of permutations to have k successes in N tries

Use binomial distribution to model processes with two outcomes
Example: detection efficiency = #(particles seen by detector) / #(all particles passing detector)

In the limit $N \to \infty, p \to 0, Np = \nu =$ const, binomial distribution can be approximated by a Poisson distribution

Poisson distribution

$$p(k;\nu) = \frac{\nu^k}{k!}e^{-\nu}$$

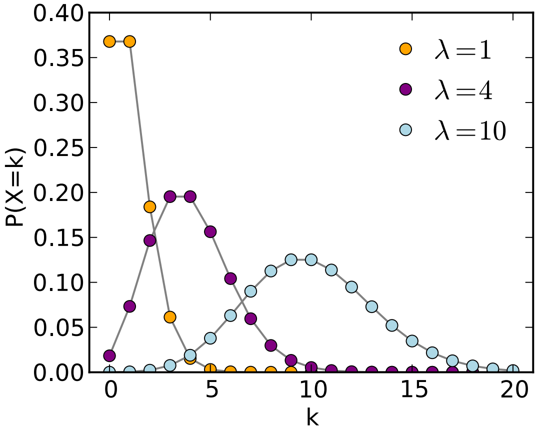
$$E[k] = \nu; \quad V[k] = \nu$$

Properties:

- If n_1, n_2 follow Poisson distribution, then also $n_1 + n_2$
- Can be approximated by Gaussian for large ν

Examples:

- Clicks of a Geiger counter in a given time interval
- Cars arriving at a traffic light in one minute



Poisson distribution

$$p(k;\nu) = \frac{\nu^k}{k!}e^{-\nu}$$

$$E[k] = \nu; \quad V[k] = \nu$$

Properties:

- If n_1, n_2 follow Poisson distribution, then also $n_1 + n_2$
- Can be approximated by Gaussian for large ν

Examples:

- Clicks of a Geiger counter in a given time interval
- Cars arriving at a traffic light in one minute

probability of k events occurring in fixed interval of time if events …

... occur with constant rate

... independently of time since last event

Poisson distribution

$$p(k;\nu) = \frac{\nu^k}{k!}e^{-\nu}$$

$$E[k] = \nu; \quad V[k] = \nu$$

Properties:

- If n_1, n_2 follow Poisson distribution, then also $n_1 + n_2$
- Can be approximated by Gaussian for large ν

Examples:

- Clicks of a Geiger counter in a given time interval
- Cars arriving at a traffic light in one minute

Rare events:

- Number of Prussian cavalrymen killed by horse-kicks

  Observe 10 army corps over 20 years:
  122 deaths due to horse kicks,
  therefore on average 0.61 deaths / (corps $\times$ year)

Number of deaths in 1 corps in 1 year	Actual number of such cases	Poisson prediction
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6

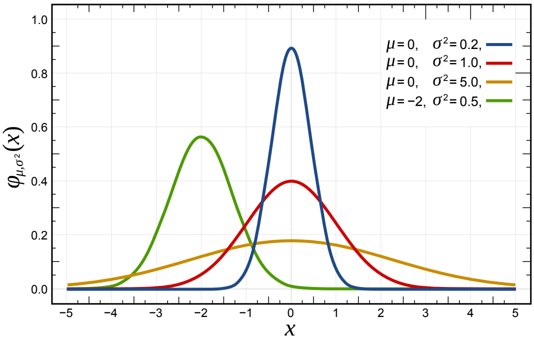
JG|U

Gaussian

A.k.a. normal distribution

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

Mean: $E[x] = \mu$
Variance: $V[x] = \sigma^2$



Standard normal distribution: $\mu = 0, \sigma = 1$

Cumulative distribution related to error function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz = \frac{1}{2}\left[\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) + 1\right]$$

In Python: `scipy.stats.norm(loc, scale)`

Why are Gaussians so useful?

Central limit theorem: sum of n random variables approaches Gaussian distribution, for large n

True, if fluctuation of sum is not dominated by the fluctuation of one (or a few) terms

- Good example: velocity component v_x of air molecules
- So-so example: total deflection due to multiple Coulomb scattering.
  Rare large angle deflections give non-Gaussian tail
- Bad example: energy loss of charged particles traversing thin gas layer.
  Rare collisions make up large fraction of energy loss ➡ Landau PDF

p-value

Probability for a Gaussian distribution corresponding to $[\mu - Z\sigma, \mu + Z\sigma]$:

$$P(Z\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-Z}^{+Z} e^{-\frac{x^2}{2}} = \Phi(Z) - \Phi(-Z) = \mathrm{erf}\left(\frac{Z}{\sqrt{2}}\right)$$

68.27% of area within $\pm 1\sigma$
95.45% of area within $\pm 2\sigma$
99.73% of area within $\pm 3\sigma$

90% of area within $\pm 1.645\sigma$
95% of area within $\pm 1.960\sigma$
99% of area within $\pm 2.576\sigma$

p-value:
probability that random process (fluctuation) produces a measurement at least this far from the true mean

$$p\text{-value} := 1 - P(Z\sigma)$$

Available in ROOT: `TMath::Prob(Z*Z)`
and Python: `2*stats.norm.sf(Z)`

Deviation	p-value (%)
1σ	31.73
2σ	4.55
3σ	0.270
4σ	0.006 33
5σ	0.000 057 3

χ^2 distribution

$x_1, \ldots, x_n$ be n independent standard normal ($\mu = 0, \sigma = 1$) random variables. Then the sum of their squares

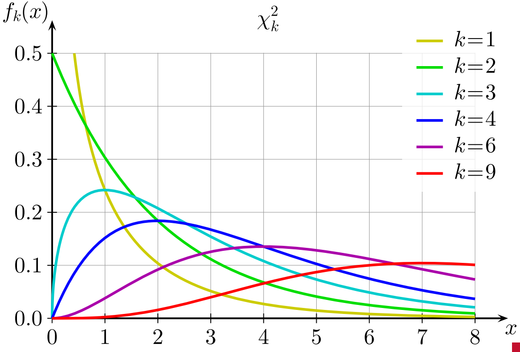
$$z = \sum_{i=1}^{n} x_i^2 = \sum_i \frac{(x' - \mu')^2}{\sigma'^2}$$

follows a χ^2 distribution with n degrees of freedom.

$$f(z; n) = \frac{z^{n/2-1}}{2^{n/2}\Gamma(\frac{n}{2})} e^{-z/2}, \quad z \geq 0$$

$$E[z] = n, \quad V[z] = 2n$$

Quantify goodness of fit, compatibility of measurements, …



Student's t distribution

Let $x_1, \ldots, x_n$ be distributed as $N(\mu, \sigma)$.

Sample mean and estimate of variance:

$$\bar{x} = \frac{1}{n}\sum_i x_i, \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

Don't know true μ, therefore have to estimate variance by $\hat{\sigma}$.

$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ follows $N(0,1)$

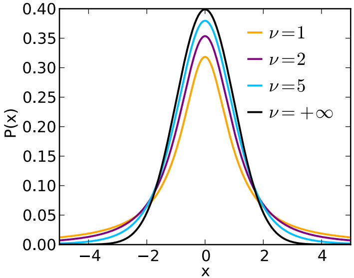
$\frac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{n}}$ not Gaussian:
Student's t-distribution with $n-1$ d.o.f.

$$f(t;n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

For $n \to \infty$, $f(t;n) \to N(t;0,1)$

Applications:

- Hypothesis tests: assess statistical significance between two sample means
- Set confidence intervals (more of that later)



Tools

Usable and useful tools (e.g. for your analysis) depend on environment / external constraints and other factors external constraints and other factors

- within working group
- international collaboration
- personal preferences
- …

Don't underestimate the cost of choosing a different approach than everyone else around you!

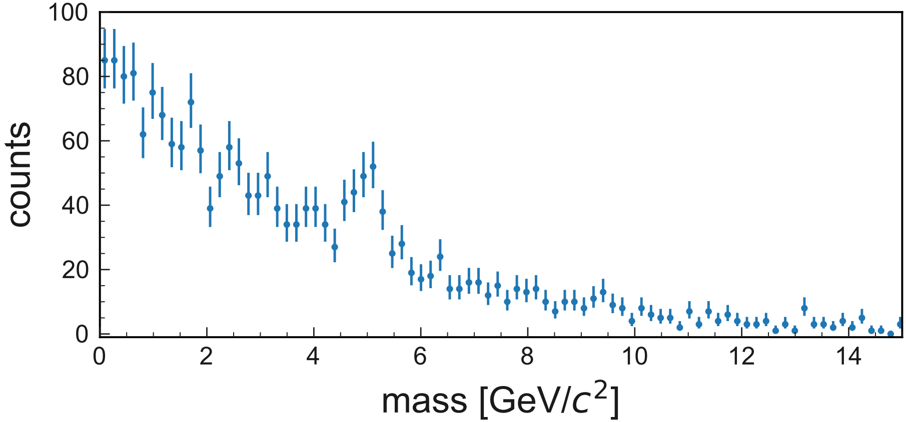
external constraints and other factors It may be worth it, though; just be aware of the implications!
For example: R vs python vs ROOT? Well-maintained or niche packages in python?

Tools

From my own experience with data analysis in HEP experiments:

- To paraphrase Willem van der Poel's 'Zero One Infinity' rule:
  The only numbers you should care about are Zero, One, and Infinity
  If you have to do something more than once, automate!
- Corollary: interactive tools are nice, but scripts are much better 'in production', especially to produce plots
  By all means explore your data using JupyterLab or other interactive tools, but then export the result as executable script
- Use a version control system, such as `git`, to keep track of changes in your code
- Make use of well-maintained libraries, toolkits &c for common tasks
  Yes, you can write your own algorithms to perform function minimisation or matrix inversion, and it is very instructive to do so
  — but should you use this 'in production'?

Motivation for this lecture



Count “events”, signal + background

Q: is there a signal at all? significance? where is it? how wide is it?

Parameter estimation

- Underlying assumption: data points that we measure sample an underlying, true, distribution
- examples:
  - decay of radioactive isotope: decay rate follows exponential distribution
  - mass and line width of a broad resonance: Breit-Wigner (Lorentzian) shape
  - …
- True shape may not be exactly known, but maybe can approximate with analytic function with a few parameters
- detector resolution may ‘smear out’ measured values from true value
- Our task:
  determine the parameters defining the underlying distribution
- would like to have an objective measure of how well model describes data: goodness of fit

Parameter estimation: uncertainties

In addition to point estimate ('what is the lifetime τ of this isotope?', 'how large is the signal strength?'): uncertainty (a.k.a. 'error') on this quantity, confidence interval

Some very well known 'rules of thumb':

- Counts of random events: if Poissonian is a good assumption, $N \pm \sqrt{N}$ for large-ish N
- 'Gaussian error propagation'
  helpful tool: python package `uncertainties`
  `https://pythonhosted.org/uncertainties/user_guide.html`
  ➡ live demo

In 2006: $M_{\mathrm{top}} = 174.3 \pm 5.1\,\mathrm{GeV}/c^2$

What does this mean?

Assuming that the authors quote 68% ("1 σ") uncertainties

- 68% of top quarks have masses between 169.2 and 179.4 GeV/c^2
  WRONG: all top quarks have same mass!

In 2006: $M_{\text{top}} = 174.3 \pm 5.1\ \text{GeV}/c^2$

What does this mean?

Assuming that the authors quote 68% ("1 σ") uncertainties

- 68% of top quarks have masses between 169.2 and 179.4 GeV/c^2
  WRONG: all top quarks have same mass!

- The probability of M_{top} being in the range 169.2 – 179.4 GeV/c^2 is 68%
  WRONG: M_{top} is what it is, it is either in or outside this range. P is 0 or 1.

In 2006: $M_{\text{top}} = 174.3 \pm 5.1\ \text{GeV}/c^2$

What does this mean?

Assuming that the authors quote 68% (“1 σ”) uncertainties

- 68% of top quarks have masses between 169.2 and 179.4 GeV/c^2
  WRONG: all top quarks have same mass!
- The probability of M_{top} being in the range $169.2 - 179.4\ \text{GeV}/c^2$ is 68%
  WRONG: M_{top} is what it is, it is either in or outside this range. P is 0 or 1.
- M_{top} has been measured to be $174.3\ \text{GeV}/c^2$ using a technique which has a 68% probability of being within $5.1\ \text{GeV}/c^2$ of the true result
  RIGHT

In 2006: $M_{\text{top}} = 174.3 \pm 5.1\ \text{GeV}/c^2$

What does this mean?

Assuming that the authors quote 68% ("1 σ") uncertainties

- 68% of top quarks have masses between 169.2 and 179.4 GeV/c^2
  WRONG: all top quarks have same mass!

- The probability of M_{top} being in the range 169.2 – 179.4 GeV/c^2 is 68%
  WRONG: M_{top} is what it is, it is either in or outside this range. P is 0 or 1.

- M_{top} has been measured to be 174.3 GeV/c^2 using a technique which has a 68% probability of being within 5.1 GeV/c^2 of the true result
  RIGHT
  if we repeated the measurement many times, we would obtain many different intervals; they would bracket the true M_{top} in 68% of all cases

Point estimates, limits

Often reported: point estimate and its standard deviation, $\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}$.

In some situations, an interval is reported instead, e.g. when
p.d.f. of the estimator is non-Gaussian, or
there are physical boundaries on the possible values of the parameter

Goals:

- communicate as objectively as possible the result of the experiment
- provide an interval that is constructed to cover the true value of the parameter with a specified probability
- provide information needed to draw conclusions about the parameter or to make a particular decision
- draw conclusions about parameter that incorporate stated prior beliefs

With sufficiently large data sample, point estimate and standard deviation essentially satisfy all these goals.

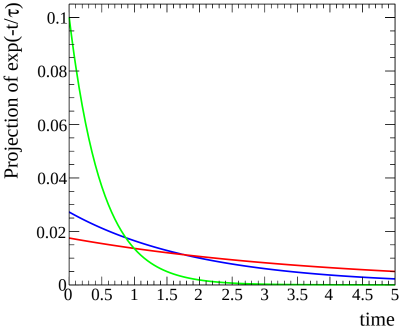
Parameter estimation

Parameters of a pdf are constants that characterise its shape, e.g.

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

x: random variable

θ: shape parameter, here: lifetime τ



Suppose we have a sample of observed values,
$\vec{x} = (x_1, \ldots, x_n)$,
independent, identically distributed (i.i.d.).

Want to find some function of the data to estimate the parameters

$$\hat{\theta}(\vec{x}) \qquad \text{Estimator for } \theta$$

Often, more than one parameter: $\vec{\theta}$

Properties of estimators

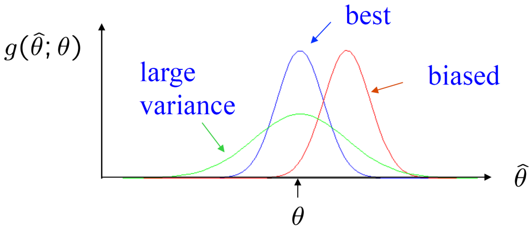
Consistency Estimator is consistent if it converges to the true value

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

Bias Difference between expectation value of estimator and true value

$$b \equiv E[\hat{\theta}] - \theta$$

Efficiency Estimator is efficient if its variance $V[\hat{\theta}]$ is small



Example: estimators for lifetime of a particle

Estimator	Consistent?	Unbiased?	Efficient?
$\hat{\tau} = \frac{t_1+t_2+\ldots+t_n}{n}$	yes	yes	yes
$\hat{\tau} = \frac{t_1+t_2+\ldots+t_n}{n-1}$	yes	no	no
$\hat{\tau} = t_1$	no	yes	no

Unbiased estimators for mean and variance of a distribution

Estimator for the mean:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$b = E[\hat{\mu}] - \mu = 0;\ V[\hat{\mu}] = \frac{\sigma^2}{n}$, i.e. $\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$

Estimator for the variance:

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$b = E[s^2] - \sigma^2 = 0$$

$$V[s^2] = \frac{\sigma^4}{n}\left((\kappa - 1) + \frac{2}{n-1}\right)$$

$\kappa = \mu_4/\sigma^4$: kurtosis.

Note: even though s^2 is unbiased estimator for variance σ^2,
s is a biased estimator for s.d. σ (have to apply non-linear function to get s from s^2)

Likelihood function for i.i.d. data

Suppose we have a measurement of n independent values (i.i.d.)

$$\vec{x} = (x_1, \ldots, x_n)$$

drawn from the same distribution

$$f(x; \vec{\theta}), \quad \vec{\theta} = (\theta_1, \ldots, \theta_m)$$

The joint pdf for the observed values $\vec{x}$ is given by

$$\mathcal{L}(\vec{x}; \vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad \text{likelihood function}$$

Likelihood function for i.i.d. data

Suppose we have a measurement of n independent values (i.i.d.)

$$\vec{x} = (x_1, \ldots, x_n)$$

drawn from the same distribution

$$f(x; \vec{\theta}), \quad \vec{\theta} = (\theta_1, \ldots, \theta_m)$$

The joint pdf for the observed values $\vec{x}$ is given by

$$\mathcal{L}(\vec{x}; \vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad \text{likelihood function}$$

Note

Likelihood $\mathcal{L}(\vec{\theta})$ is not a pdf: not normalized (unclear whether $\int \mathrm{d}\theta \mathcal{L}(\theta)$ exists at all)

Can be normalized using

$$\int \mathrm{d}\theta \mathcal{L}(\theta) p(\theta)$$

but $p(\theta)$ not uniquely determined! (used in Bayesian reasoning: prior)

Likelihood function for i.i.d. data

$$\mathcal{L}(\vec{x};\vec{\theta}) = \prod_{i=1}^{n} f(x_i;\vec{\theta})$$

Consider $\vec{x}$ as constant, so $\mathcal{L}(\vec{x};\vec{\theta})$ is a function of the parameters $\vec{\theta}$ only.

The maximum likelihood estimate (MLE) of the parameters are the values $\vec{\theta}$ for which $\mathcal{L}(\vec{x};\vec{\theta})$ has a global maximum.

For practical reasons, usually use

$$\log \mathcal{L}(\vec{x};\vec{\theta}) = \sum_{i=1}^{n} \log f(x_i;\vec{\theta})$$

(computers can cope with sum of small numbers much better than with product of small numbers)

ML Example: Exponential decay

Consider exponential pdf: $f(t;\tau) = \frac{1}{\tau}e^{-t/\tau}$

Independent measurements drawn from this distribution: $t_1, t_2, \ldots, t_n$

Likelihood function:

$$\mathcal{L}(\tau) = \prod_i \frac{1}{\tau} e^{-t_i/\tau}$$

$\mathcal{L}(\tau)$ is maximal where $\log \mathcal{L}(\tau)$ is maximal:

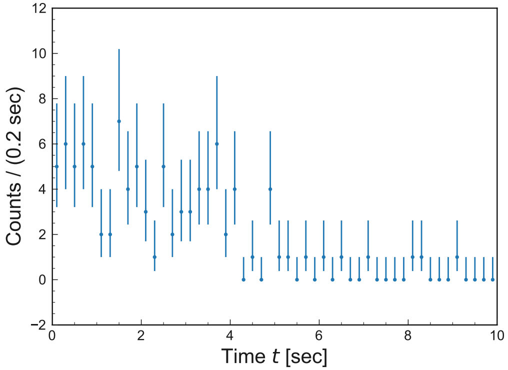
$$\log \mathcal{L}(\tau) = \sum_{i=1}^{n} \log f(t_i;\tau) = \sum_{i=1}^{n} \left(\log \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

Find maximum:

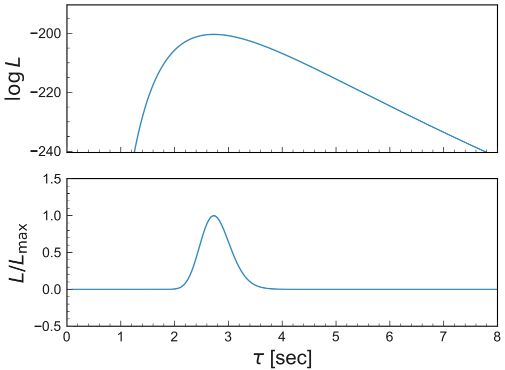
$$\frac{\partial \log \mathcal{L}(\tau)}{\partial \tau} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \quad \Rightarrow \quad \hat{\tau} = \frac{1}{n} \sum_i t_i$$

ML Example: Exponential decay

Raw data (100 ‘measurements’)



Scan of likelihood function $\mathcal{L}(\vec{x};\tau)$



ML Example: Gaussian

Consider $x_1, \ldots, x_n$ drawn from Gaussian(μ, σ^2)

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Log-likelihood function:

$$\log \mathcal{L}(\mu,\sigma^2) = \sum_i \log f(x_i;\mu,\sigma^2) = \sum_i \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

Derivatives w.r.t μ and σ^2:

$$\frac{\partial \log \mathcal{L}(\mu,\sigma^2)}{\partial \mu} = \sum_i \frac{x_i - \mu}{\sigma^2}; \qquad \frac{\partial \log \mathcal{L}(\mu,\sigma^2)}{\partial \sigma^2} = \sum_i \left(\frac{(x_i-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right)$$

ML Example: Gaussian

Setting derivatives w.r.t. μ and σ^2 to zero, and solving the equations:

$$\hat{\mu} = \frac{1}{n}\sum_i x_i; \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_i (x_i - \hat{\mu})^2$$

- Find that the ML estimator for σ^2 is biased!
- For Gaussian distribution, μ and σ can be estimated simply from histogram mean and RMS!

Properties of the ML estimator

- ML estimator is consistent, i.e. it approaches the true value asymptotically
- In general, ML estimator is biased for finite n
  (need to check size of bias)
- ML estimator is invariant under parameter transformation

$$\psi = g(\theta) \quad \Rightarrow \quad \hat{\psi} = g(\hat{\theta})$$

Averaging measurements with Gaussian uncertainties

Assume n measurements, same mean μ, but different resolutions σ

$$f(x;\mu,\sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu)^2}{2\sigma_i^2}}$$

log-likelihood, similar to before:

$$\log \mathcal{L}(\mu) = \sum_i \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma_i - \frac{(x_i-\mu)^2}{2\sigma_i^2} \right)$$

We obtain formula for weighted average, as before:

$$\left.\frac{\partial \log \mathcal{L}(\mu)}{\partial \mu}\right|_{\mu=\hat{\mu}} \stackrel{!}{=} 0 \qquad \Rightarrow \qquad \hat{\mu} = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}$$

Averaging measurements with Gaussian uncertainties

Uncertainty? Taylor expansion exact, because $\log \mathcal{L}(\mu)$ is parabola:

$$\log \mathcal{L}(\mu) = \log \mathcal{L}(\hat{\mu}) + \underbrace{\left[\frac{\partial \log \mathcal{L}}{\partial \mu}\right]_{\mu=\hat{\mu}} (\mu - \hat{\mu})}_{=0} - \frac{h}{2}(\mu - \hat{\mu})^2, \qquad h = - \left.\frac{\partial^2 \log \mathcal{L}(\mu)}{\partial \mu^2}\right|_{\mu=\hat{\mu}}$$

This means that likelihood function is a Gaussian:

$$\mathcal{L}(\mu) \propto \exp\left(-\frac{h}{2}(\mu - \hat{\mu})^2\right)$$

with a standard deviation

$$\sigma_{\hat{\mu}} = 1/\sqrt{h} = \left(\left.\frac{\partial^2 \log \mathcal{L}(\mu)}{\partial \mu^2}\right|_{\mu=\hat{\mu}}\right)^{-1}$$

$$h = \sum_i \frac{1}{\sigma_i^2} \qquad \Rightarrow \qquad \sigma_{\hat{\mu}} = \left(\sum_i \frac{1}{\sigma_i^2}\right)^{-1/2}$$

Uncertainty bounds

Likelihood function with only one parameter:

$$\mathcal{L}(\vec{x};\theta) = \mathcal{L}(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

and $\hat{\theta}$ an estimator of the parameter θ

Without proof: it can be shown that the variance of a (biased, with bias b) estimator satisfies

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

Cramér-Rao minimum variance bound (MVB)

Uncertainty of the MLE: Approach I

Approximation

$$E\left[-\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right] \approx -\left.\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right|_{\theta=\hat{\theta}}$$

good for large n (and away from any explicit boundaries on θ)

In this approximation, variance of ML estimator is given by

$$V[\hat{\theta}] = -\left(\left.\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right|_{\theta=\hat{\theta}}\right)^{-1}$$

so we only need to evaluate the second derivative of $\log \mathcal{L}$ at its maximum.

Uncertainty of the MLE: Approach II (‘graphical method’)

Taylor expansion of $\log \mathcal{L}$ around maximum:

$$\log \mathcal{L}(\theta) = \log \mathcal{L}(\hat{\theta}) + \underbrace{\left[\frac{\partial \log \mathcal{L}}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})}_{=0} + \frac{1}{2}\left[\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \cdots$$

If $\mathcal{L}$ approximately Gaussian ($\log \mathcal{L}$ approx. a parabola):

$$\log \mathcal{L}(\theta) \approx \log \mathcal{L}_{\text{max}} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2_{\hat{\theta}}}}$$

Estimate uncertainties from the points where $\log \mathcal{L}$ has dropped by 1/2 from its maximum:

$$\log \mathcal{L}(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \log \mathcal{L}_{\text{max}} - \frac{1}{2}$$

This can be used even if $\mathcal{L}(\theta)$ is not Gaussian

If $\mathcal{L}(\theta)$ is Gaussian: results of approach I & II identical

Example:
uncertainty of the decay time for an exponential decay

Variance of the estimated decay time:

$$\frac{\partial^2 \log \mathcal{L}(\tau)}{\partial \tau^2} = \sum_i \left(\frac{1}{\tau^2} - 2\frac{t_i}{\tau^3} \right) = \frac{n}{\tau^2} - \frac{2}{\tau^3} \sum_i t_i = \frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right)$$

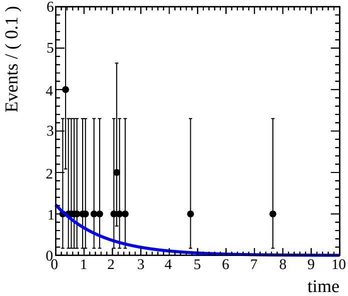
Thus,

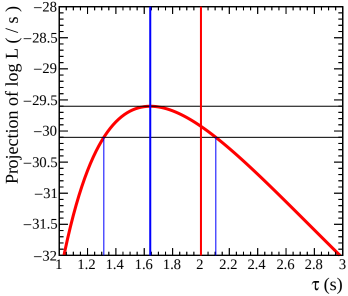
$$V[\hat{\tau}] = - \left(\frac{\partial^2 \log \mathcal{L}(\tau)}{\partial \tau^2} \right)^{-1}_{\tau=\hat{\tau}} = \frac{\hat{\tau}^2}{n}$$

$$\Rightarrow \quad \hat{\sigma}_{\hat{\tau}} = \frac{\hat{\tau}}{\sqrt{n}}$$

Exponential decay: illustration

20 data points sampled from $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$ with $\tau = 2$





ML estimate:

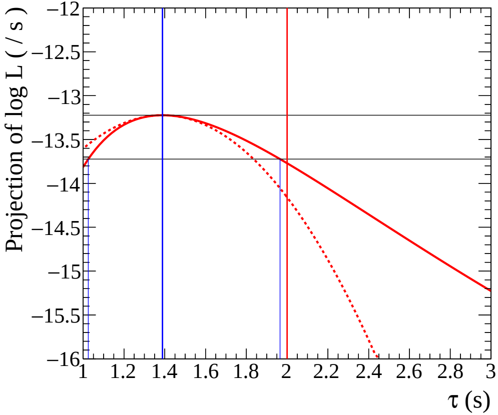
$$\hat{\tau} = 1.65$$

$$\hat{\sigma} = 1.65/\sqrt{20} = 0.37$$ using quadratic approximation of $\mathcal{L}(\tau)$

or $$\hat{\sigma} = {}^{+0.47}_{-0.34}$$ using shape of $-\log \mathcal{L}$ curve

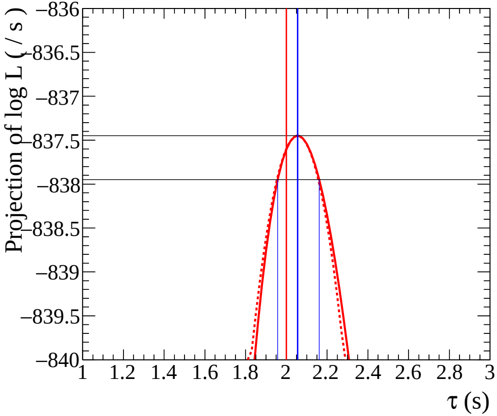
Exponential decay: $\log \mathcal{L}$ for different sample sizes

10 data points



quadratic approximation for $\log \mathcal{L}$ not very good

500 data points



quadratic approximation for $\log \mathcal{L}$ excellent

Variance of the ML estimator for m parameters

In limit of large sample size, $\mathcal{L}$ approaches multivariate Gaussian distribution for any probability density :

$$\mathcal{L}(\vec{\theta}) \propto \exp\left(-\frac{1}{2}(\vec{\theta}-\hat{\vec{\theta}})^T V^{-1}[\hat{\vec{\theta}}](\vec{\theta}-\hat{\vec{\theta}})\right)$$

Variance of ML estimator reaches MVB (minimum variance bound), related to the Fisher information matrix:

$$V[\hat{\vec{\theta}}] \to I(\theta)^{-1}, \qquad I_{jk}[\vec{\theta}] = -E\left[\frac{\partial^2 \log \mathcal{L}(\vec{\theta})}{\partial\theta_j \partial\theta_k}\right]$$

Covariance matrix of the estimated parameters:

$$V[\hat{\vec{\theta}}] \approx \left[-\left.\frac{\partial^2 \log \mathcal{L}(\vec{x};\vec{\theta})}{\partial\vec{\theta}^2}\right|_{\vec{\theta}=\hat{\vec{\theta}}}\right]^{-1}$$

Standard deviation of a single parameter:

$$\hat{\sigma}_{\hat{\theta}_j} = \sqrt{(V[\hat{\vec{\theta}}])_{jj}}$$

MLE in practice: numeric minimisation

Analytic expression for $\mathcal{L}(\theta)$ and its derivatives often not easily known

Use a generic minimiser like MINUIT to find (global) minimum of $-\log \mathcal{L}(\theta)$

Typically uses gradient descent method to find minimum and then scans around minimum to obtain $\mathcal{L}_{\text{max}} - 1/2$ contour

make sure you don't get stuck in a local minimum: check likelihood profiles

➡ see today's practical part for a hands-on

MINUIT

MINUIT

generic minimiser

around since the 1970s (Fred James, CERN; first implementation in FORTRAN)

ported to C++ (Minuit2 in ROOT), Python interface (iminuit)

features:

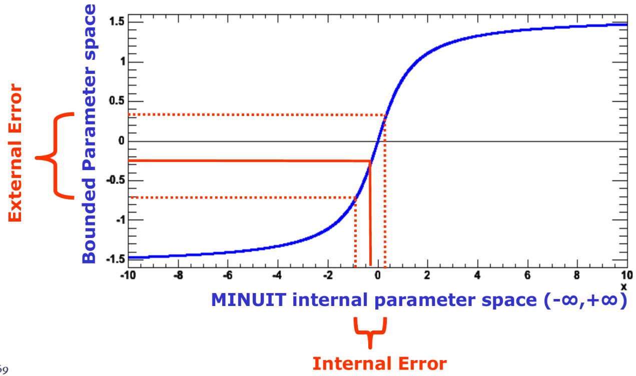
- several algorithms for minimisation
- one of the few minimisers that returns estimates for parameter errors
- compute confidence intervals by scanning likelihood function around minimum
- …

use for generic minimisation only — dedicated fit routines (e.g. for track fits) may have better performance

Bounds on parameters in MINUIT

Sometimes, you may want to bound the allowed range of fit parameters
e.g. to prevent (numerical) instabilities or
avoid unphysical results (‘fraction f should be in $[0, 1]$’, ‘mass ≥ 0’)

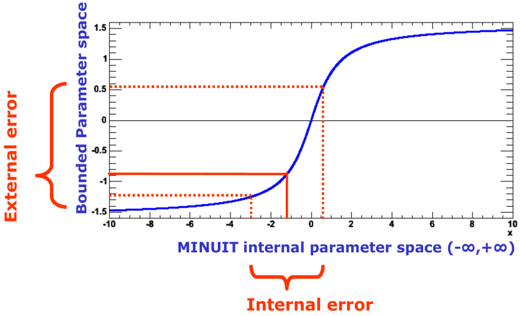
MINUIT internally transforms parameter y with two-sided bounds with an $\arcsin(y)$ function to an unbounded parameter x:



Bounds on parameters in MINUIT

If fitted parameter value is close to boundary, errors will become asymmetric and maybe even incorrect

Placing very large limits 'just in case' (such as $[0, 10^{10}]$) can lead to total loss of precision for small parameter values



- Try to find alternative parametrisation to avoid region of instability.
  E.g. complex number
  $z = re^{i\phi}$ with bounds $r \geq 0$, $0 \leq \phi < 2\pi$
  $z = x + iy$ may be better behaved
- If bounds were placed to avoid 'unphysical' region, consider not imposing the limits and dealing with the restriction to the physical region after the fit.

Extended ML method

In standard ML method, information about unknown parameters is encoded in shape of the distribution of the data.

Sometimes, the number of observed events also contains information about the parameters (e.g. when measuring a decay rate).

Normal ML method:

$$\int f(x;\vec{\theta})\mathrm{d}x = 1$$

Extended ML method:

$$\int q(x;\vec{\theta})\mathrm{d}x = \nu(\vec{\theta}) = \text{predicted number of events}$$

Extended ML method (II)

Likelihood function becomes:

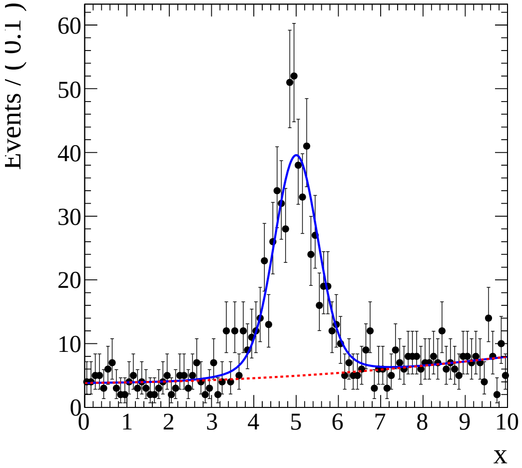
$$\mathcal{L}(\vec{\theta}) = \frac{\nu^n e^{-\nu}}{n!} \prod_i f(x_i; \vec{\theta}) \qquad \text{where } \nu \equiv \nu(\vec{\theta})$$

And log-likelihood function:

$$\log \mathcal{L}(\vec{\theta}) = -\log(n!) - \nu(\vec{\theta}) + \sum_i \log[f(x_i; \vec{\theta})\nu(\vec{\theta})]$$

$\log n!$ does not depend on parameters. Can be omitted in minimisation

Application of Extended ML method



Example:

- Two-component fit (signal + background)
- Unbinned ML fit, histogram for visualisation only
- Want to obtain meaningful estimate of the uncertainties of signal and background yields

Normalised pdf:

$$f(x; r_s, \vec{\theta}) = r_s f_s(x; \vec{\theta}) + (1 - r_s) f_b(x; \vec{\theta})$$

$$r_s = \frac{s}{s+b}, \quad r_b = 1 - r_s = \frac{b}{s+b}$$

$$-\log \tilde{\mathcal{L}}(s, b, \vec{\theta}) = s + b - \sum_i \log[s f_s(x_i; \vec{\theta}) + b f_b(x_i; \vec{\theta})]$$

Application of Extended ML method (II)

Could have just fitted normalised pdf to our n events, with r_S an additional parameter.

Good estimate of the number of signal events: $r_S \times n$

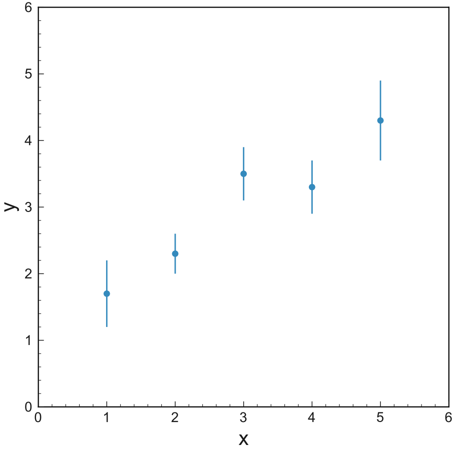
However, $\sigma_{r_S} \times n$ is not a good estimate for the variation of the number of signal events: ignores fluctuations of n.

Using extended ML fixes this.

Least squares from ML

Consider n measured values $y_1(x_1), y_2(x_2), \ldots, y_n(x_n)$, assumed to be independent Gaussian r.v. with known variances, $V[y_i] = \sigma_i^2$.

x	y	σ_y
1	1.7	0.5
2	2.3	0.3
3	3.5	0.4
4	3.3	0.4
5	4.3	0.6



Least squares from ML

Consider n measured values $y_1(x_1), y_2(x_2), \ldots, y_n(x_n)$, assumed to be independent Gaussian r.v. with known variances, $V[y_i] = \sigma_i^2$.

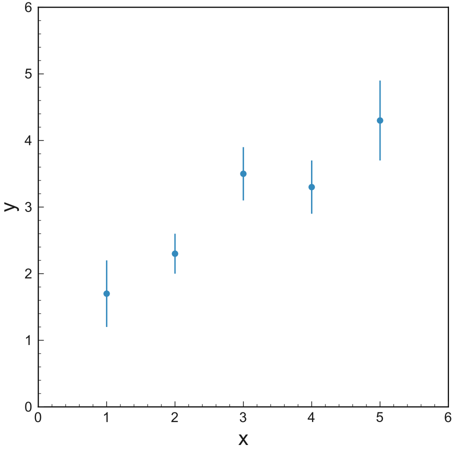
Assume we have a model for the functional dependence of y_i on x_i,

$$E[y_i] = f(x_i; \vec{\theta})$$

Want to estimate $\vec{\theta}$

Likelihood function:

$$\mathcal{L}(\vec{\theta}) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{y_i - f(x_i;\vec{\theta})}{\sigma_i}\right)^2\right]$$



Least squares from ML (II)

Log-likelihood function:

$$\log \mathcal{L}(\vec{\theta}) = -\frac{1}{2}\sum_i \left(\frac{y_i - f(x_i;\vec{\theta})}{\sigma_i}\right)^2 + \text{terms not depending on } \vec{\theta}$$

Maximising this is equivalent to minimising

$$\chi^2(\vec{\theta}) = \sum_i \left(\frac{y_i - f(x_i;\vec{\theta})}{\sigma_i}\right)^2$$

so, for Gaussian uncertainties, method of least squares coincides with maximum likelihood method.

Error definition: points where $\chi^2 = \chi^2_{\text{min}} + Z^2$ for a $Z\sigma$ interval
(compare: $\log \mathcal{L} = \log \mathcal{L}_{\text{max}} - \frac{1}{2}Z^2$ for MLE)

Linear least squares

Important special case: consider function linear in the parameters:

$$f(x;\vec{\theta}) = \sum_j a_j(x)\theta_j \qquad n \text{ data points, } m \text{ parameters}$$

χ^2 in matrix form:

$$\begin{aligned}\chi^2 &= (\vec{y} - A\vec{\theta})^T V^{-1} (\vec{y} - A\vec{\theta}), \qquad A_{i,j} = a_j(x_i) \\ &= \vec{y}^T V^{-1} \vec{y} - 2\vec{y}^T V^{-1} A\vec{\theta} + \vec{\theta}^T A^T V^{-1} A\vec{\theta}\end{aligned}$$

Set derivatives w.r.t. θ_i to zero:

$$\nabla\chi^2 = -2(A^T V^{-1}\vec{y} - A^T V^{-1} A\vec{\theta}) = 0$$

Solution:

$$\widehat{\vec{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \vec{y} \equiv L\vec{y}$$

Linear least squares

Covariance matrix U of the parameters, from error propagation
(exact, because estimated parameter vector is linear function of data points y_i)

$$\begin{aligned} U &= LVL^T \\ &= (A^T V^{-1} A)^{-1} \end{aligned}$$

Equivalently, calculate numerically

$$(U^{-1})_{ij} = \frac{1}{2} \left[\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j} \right]_{\vec{\theta}=\widehat{\vec{\theta}}}$$

Example: straight line fit

$$y = \theta_0 + \theta_1 x$$

Conditions $\partial\chi^2/\partial\theta_0 = 0$ and $\partial\chi^2/\partial\theta_1 = 0$ yield two linear equations with two variables that are easy to solve.

With the shorthand notation

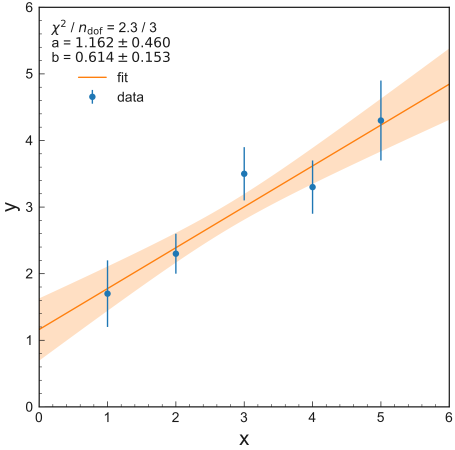
$$[z] := \sum_i \frac{z}{\sigma_i^2}$$

we finally obtain

$$\hat{\theta}_0 = \frac{[x^2][y] - [x][xy]}{[1][x^2] - [x][x]}, \qquad \hat{\theta}_1 = \frac{-[x][y] + [1][xy]}{[1][x^2] - [x][x]}$$

Simple, huh? At least, easy to program and compute, given a set of data
(I'll put the complete calculation for this in the appendix of the slides)

Example: straight line fit



Analytic fit result:

$$\hat{\theta}_0 = \frac{[x^2][y] - [x][xy]}{[1][x^2] - [x][x]} = 1.16207$$

$$\hat{\theta}_1 = \frac{-[x][y] + [1][xy]}{[1][x^2] - [x][x]} = 0.613945$$

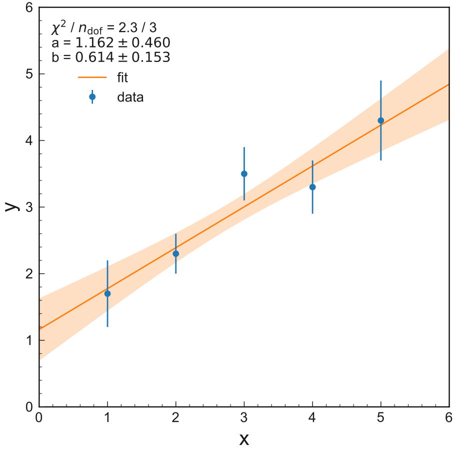
Covariance matrix of (θ_0, θ_1):

$$U = (A^T V^{-1} A)^{-1}$$
$$= \begin{pmatrix} 0.211\,186 & -0.064\,603\,5 \\ -0.064\,603\,5 & 0.023\,410\,5 \end{pmatrix}$$

Error band from

$$e^2(x) = \vec{g}(x)^T U \vec{g}(x) \qquad \text{with } \vec{g} = \nabla f(x; \vec{\theta})$$

Example: straight line fit



Numerical estimate with MINUIT:

```
****************************************
Minimizer is Minuit / Migrad
Chi2                      =      2.29557
NDf                       =            3
Edm                       =  3.23988e-23
NCalls                    =           32
p0                        =      1.16207   +/-   0.45955
p1                        =     0.613945   +/-   0.153005

Covariance Matrix:

                    p0           p1
p0             0.21119    -0.064603
p1           -0.064603      0.02341

Correlation Matrix:

                    p0           p1
p0                   1     -0.91879
p1            -0.91879            1
```

Fitting binned data

Very popular application of least-squares fit: fit a model (curve) to binned data (a histogram)

Number of events occurring in each bin j is assumed to follow Poisson distribution with mean f_j.

$$\chi^2 = \sum_{j=1}^{m} \frac{(n_j - f_j)^2}{f_j}$$

Further common simplification: ‘modified least-squares method’, assuming that $\sigma_{n_j}^2 = n_j$:

$$\chi^2 \approx \sum_{j=1}^{m} \frac{(n_j - f_j)^2}{n_j}$$

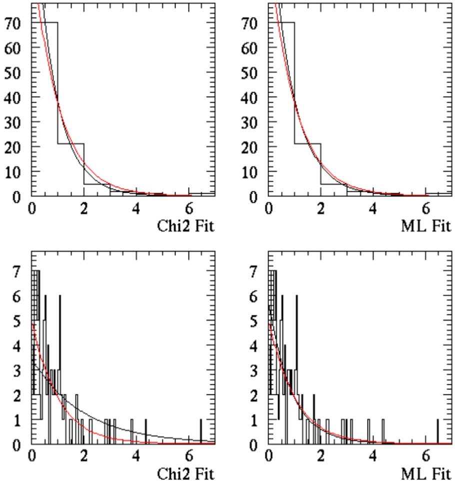
Can get away with this when all n_j are sufficiently large, but what about bins with small contents, or even zero events?

➡ Frequently, bins with $n_j = 0$ are simply excluded.

This throws away information, and will lead to biased results of your fit!

Fitting binned data

Example: exponential distribution, 100 events



red: true distribution
black: fit

The more bins you have with small statistics, the worse the MLS fit becomes.

ML method gives more reliable results in this case.

If you must use MLS, then at least rebin your data, at the loss of information.

Oser, https://www.phas.ubc.ca/~oser/p509/Lec_09.pdf

Practical estimation — verifying the validity of your fits

Want to demonstrate that

- your fit procedure gives, at least on average, the correct answer: no bias
- uncertainty quoted by your fit is an accurate measure for the statistical spread in your measurement: correct error

Validation is particularly important for low-statistics fits
intrinsic ML bias proportional $1/n$

Also important for problems with multi-dimensional observables:
mis-modelled correlations between observables can lead to bias

Basic validation strategy

Simulation study

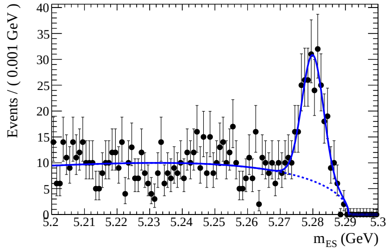
1. Obtain (very) large sample of simulated events
2. Divide simulated events in $O(100 - 1000)$ independent samples with the same size as the problem under study
3. Repeat fit procedure for each data-sized simulated sample
4. Compare average value of fitted parameter values with generated value
   ➠ demonstrate (absence of) bias
5. Compare spread in fitted parameter values with quoted parameter error
   ➠ demonstrate (in)correctness of error

Practical example — validation study

Example fit model in 1D (B mass)

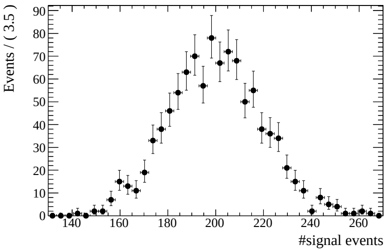
- signal component is Gaussian centred at B mass
- background component is ARGUS function (models phase space near kinematic limit)

$$q(m; n_{\text{sig}}, n_{\text{bkg}}, \vec{p}_{\text{sig}}, \vec{p}_{\text{bkg}}) = n_{\text{sig}} G(m; \vec{p}_{\text{sig}}) + n_{\text{bkg}} A(m; \vec{p}_{\text{bkg}})$$



Fit parameter under study: n_{sig}

- result of simulation study: 1000 experiments with $\left\langle n_{\text{sig}}^{\text{gen}} \right\rangle = 200$, $\left\langle n_{\text{bkg}}^{\text{gen}} \right\rangle = 800$
- distribution of $n_{\text{sig}}^{\text{fit}}$
- …looks good



Validation study — pull distribution

What about validity of the error estimate?

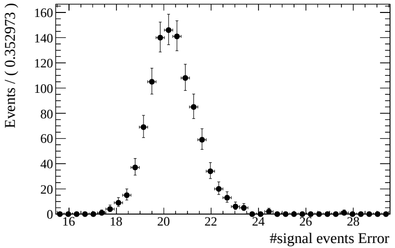
- distribution of error from simulated experiments is difficult to interpret …
- don't have equivalent of $n_{\text{sig}}^{\text{gen}}$ for the error

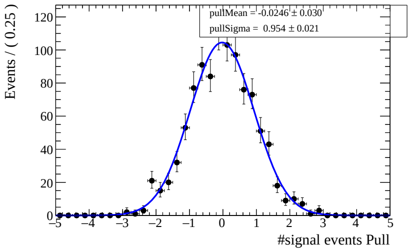
Solution: look at pull distribution

- Definition:

$$\text{pull}(n_{\text{sig}}) \equiv \frac{n_{\text{sig}}^{\text{fit}} - n_{\text{sig}}^{\text{gen}}}{\sigma_n^{\text{fit}}}$$

- Properties of pull:
  - follows Gaussian distribution if parameter and error 'sensible'
  - Mean is 0 if no bias
  - Width is 1 if error is correct

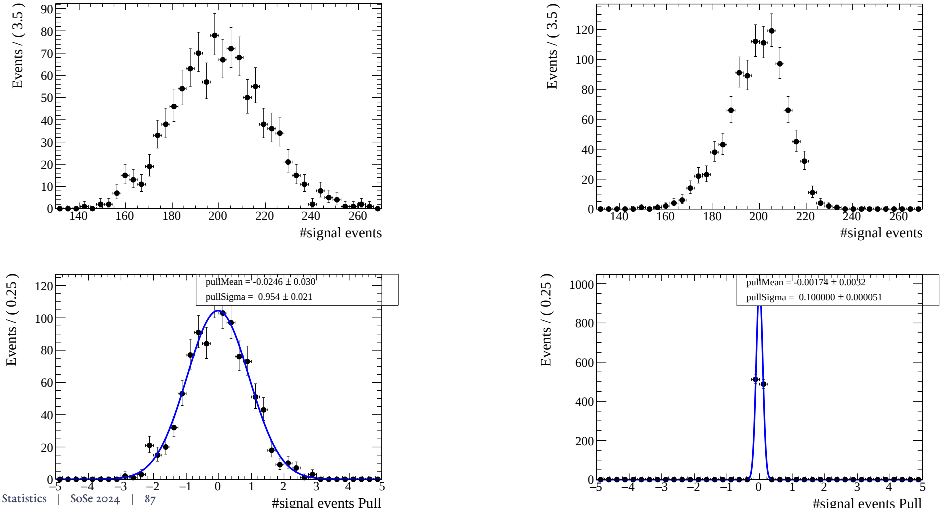




Validation study — extended ML!

As an aside, ran this toy study also with standard (not extended) ML method:

Extended | Standard



Validation study — low statistics example

Special care needs to be taken when fitting small data samples,
also if fitting small signal component in large sample

Possible causes of trouble

- χ^2 estimators become approximate as Gaussian approximation of Poisson statistics becomes inaccurate
- ML estimators may no longer be efficient
  error estimate from 2nd derivative inaccurate
- Bias term $\propto 1/n$ may no longer be small compared to $1/\sqrt{n}$

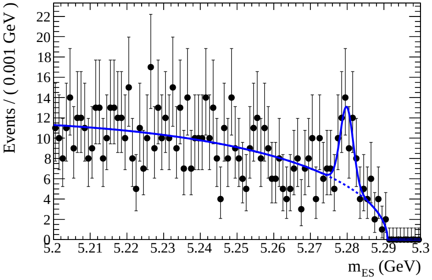
In general, absence of bias, correctness of error cannot be assumed.

- Use unbinned ML fits wherever possible — more robust
- **explicitly verify the validity of your fit**

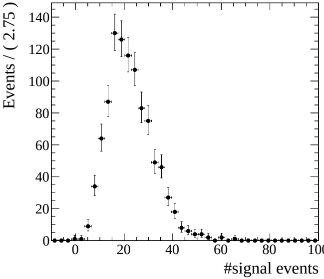
Fit bias at low n

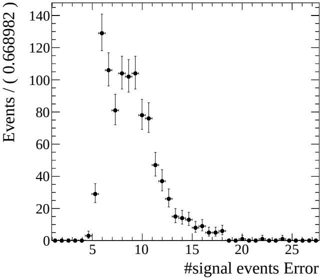
Low statistics example:

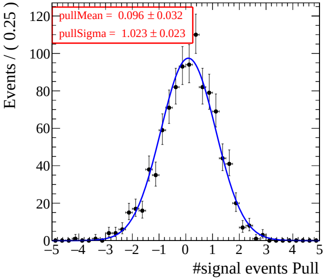
- model as before, but with $\left\langle n_{\text{sig}}^{\text{gen}} \right\rangle = 20$



Result of simulation study:







Place limit on n_{sig}?

Very tempting to limit signal yield to be ≥ 0

After all, negative signal yield is unphysical!

But: remember shape of n_{sig} in our toy experiments. Removing small values of n_{sig} will introduce (additional) positive bias

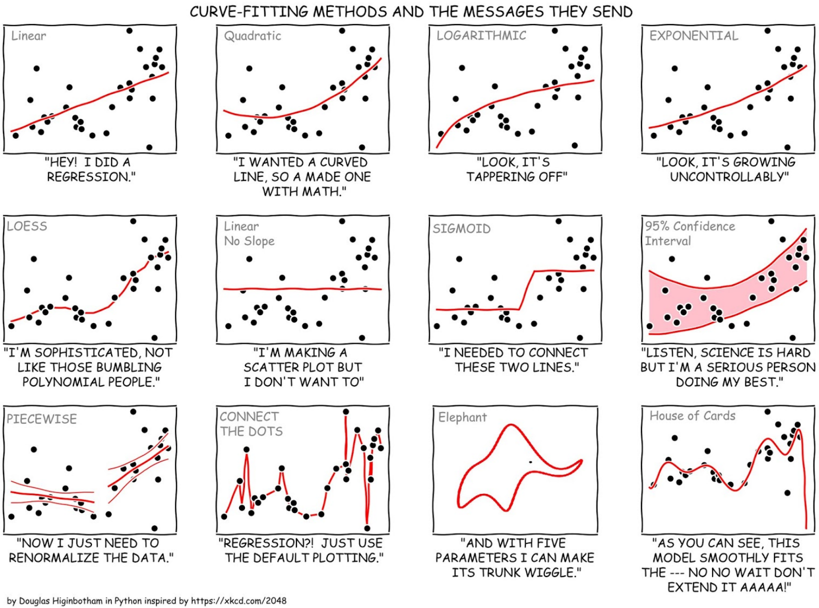
Validation study — how to obtain 10^7 simulated events?

Practical issue: usually need very large amounts of simulated events for a fit validation study

- Of order 1000x (number of events in data), easily $> 10^6$ events
- Using data generated through full (GEANT-based) detector simulation can be prohibitively expensive

Solution: sample events directly from fit function

- Technique called toy Monte Carlo sampling
- Advantage: easy to do, very fast
- Good to determine fit bias due to low statistics, choice of parametrisation, bounds on parameters, …
- Cannot test assumptions built in to fit model:
  absence of correlations between observables, …
  still need full simulation for this


CURVE-FITTING METHODS AND THE MESSAGES THEY SEND
Linear
"HEY! I DID A REGRESSION."
Quadratic
"I WANTED A CURVED LINE, SO A MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPPERING OFF"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
Linear No Slope
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO"
SIGMOID
"I NEEDED TO CONNECT THESE TWO LINES."
95% Confidence Interval
"LISTEN, SCIENCE IS HARD BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"NOW I JUST NEED TO RENORMALIZE THE DATA."
CONNECT THE DOTS
"REGRESSION?! JUST USE THE DEFAULT PLOTTING."
Elephant
"AND WITH FIVE PARAMETERS I CAN MAKE ITS TRUNK WIGGLE."
House of Cards
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE --- NO NO WAIT DON'T EXTEND IT AAAAA!"
by Douglas Higinbotham in Python inspired by https://xkcd.com/2048

Confidence intervals

Confidence intervals: Choices, choices!

We can choose:

- The confidence level
  two-sided confidence intervals: typically 68%, corresponding to $\pm 1\sigma$
  upper (or lower) limits: frequently 90%, but 95% not uncommon …
- Whether to quote an upper limit or a two-sided confidence interval
- What sort of two-sided limit
  central (i.e. symmetric), shortest, …

Important: document what you are doing!

Estimation of confidence intervals

Typically, use **fit** to determine event yields or parameters of a distribution

Least square fit (for binned datasets) or maximum likelihood fits (can also deal with unbinned data)

Error definition, for one degree of freedom:

LSQ : 1σ confidence interval from $S = S_{\text{min}} + 1$

ML : 1σ confidence interval from $\log \mathcal{L} = \log \mathcal{L}_{\text{max}} - \frac{1}{2}$

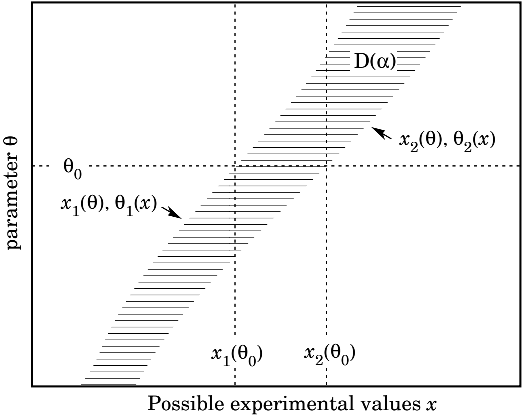
$n\sigma$ conf. intervals from $2\Delta \log \mathcal{L} = n^2$

See today's practical part what happens for joint confidence region for ν parameters

Construction of frequentist confidence intervals

Neyman construction of ‘confidence belts’:

for a given value of parameter θ, find interval of possible measured values x such that $[x_1, x_2]$ is a CL confidence interval:



Constrained parameters

Measure a mass

$$M_X = -2 \pm 5\,\mathrm{GeV}$$

or even

$$M_X = -5 \pm 2\,\mathrm{GeV}$$

'M_X lies between -7 and -3' with 68% confidence

???

Counting experiment

Expect 2.8 background events

See 0 events; so, 90% CL upper limit is 2.3 events

so, signal < -0.5 events

???

What's happened?

Two views:

Nothing has gone wrong

(Up to) 10% of our 90% CL statements can be wrong; this is just one of them

Publish this, to avoid bias!

Everything wrong!

There are physical constraints (masses are non-negative, so are cross sections!)

No way to input this into the statistical apparatus

We will not publish results that are manifestly wrong

This is broken and needs fixing

What should be done with ‘unphysical’ results?

Best, but mostly not possible: publish full likelihood (or log-likelihood) function. This allows optimal combination of results, but is rarely done.

Preferred solution: publish both solutions,
i.e. the ‘raw’, maybe nonsensical two-sided confidence interval,
and one-sided C.I. taking extra constraints into account

May have to fight against (internal and external) referees who insist that publishing a two-sided confidence interval is equivalent to claiming “observation”

Bayesian credible intervals

After a fit of our model to data: have likelihood function

$$\mathcal{L}(\vec{x}|\vec{\theta})$$

(reminder: this makes a statement on the data given a set of parameters. In general, not normalised, i.e. not a p.d.f.)

Want to turn this into a statement about the model parameters $\vec{\theta}$ given our data $\vec{x}$: use Bayes' theorem

$$b(\vec{\theta}|\vec{x}) = \frac{\mathcal{L}(\vec{x}|\vec{\theta}) \times P(\vec{\theta})}{\int \mathcal{L}(\vec{x}|\vec{\theta}) \times P(\theta)\mathrm{d}\vec{\theta}}$$

with a suitable prior $P(\vec{\theta})$

$b(\vec{\theta}|\vec{x})$: Bayes' distribution

if it exists, call it the posterior p.d.f. for the parameters

$b(\vec{\theta}|\vec{x})$ updates our prior knowledge of θ with the new measurement

Bayesian credible intervals

Bayesian approach: report full posterior p.d.f. (i.e. the Bayes' distribution)

If a range is desired: integrate posterior p.d.f. $b(\theta|x)$

$$1 - \alpha = \int_{\theta_{\text{lo}}}^{\theta_{\text{up}}} b(\theta|x) \mathrm{d}\theta$$

e.g. $1 - \alpha = 0.9$: "90% credible interval"

Several choices possible to construct $[\theta_{\text{lo}}, \theta_{\text{up}}]$:

- $[-\infty; \theta_{\text{lo}}]$ and $[\theta_{\text{up}}; \infty]$ both correspond to probability $\alpha/2$
- Symmetric interval around maximum value of b, corresponding to probability $1 - \alpha$
- $b(\theta|x)$ higher than any θ not belonging to the set
- …

Choice of prior

Some remarks on the prior $P(\theta)$:

- How to parametrise ‘complete ignorance’?
  Flat prior: hope that $\mathcal{L}$ is sufficiently peaked that we can ‘cut off’ large values
  e.g. use $P = 1/(\Sigma^+ - \Sigma^-)$ around maximum of $\mathcal{L}$ and let $\Sigma^\pm \to \pm\infty$
- But: can easily implement ‘physical limits’ such as
  ‘masses are non-negative’: $P(m) = 0$ for $m < 0$
- Non-linear parameter transformations do not leave prior invariant:
  check whether this makes a large difference!

Non-informative prior

Fisher information matrix for likelihood $\mathcal{L}(x;\theta)$

$$\mathcal{I}(\theta)_{i,j} \equiv \mathsf{E}\left[\left(\frac{\partial}{\partial\theta_i}\log\mathcal{L}(x;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log\mathcal{L}(x;\theta)\right)\middle|\,\theta\right] = -\,\mathsf{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log\mathcal{L}(x;\theta)\right]$$

can be numerically estimated as the Hessian matrix of the log-likelihood function near maximum

Jeffreys prior: non-informative, invariant under reparametrisation

$$p(\theta) \propto \sqrt{\det\mathcal{I}(\theta)}$$

if θ, ϕ two possible parametrisations of our problem, and $\theta(\phi)$ is continuously differentiable, we want to have

$$p_\theta(\theta) = p_\phi(\phi)\left|\frac{\partial\theta}{\partial\phi}\right|$$

Example for Jeffreys prior (I)

Gaussian pdf with fixed σ, parameter of interest is the scale parameter μ:

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p(\mu) \propto \sqrt{\mathcal{I}(\mu)} = \sqrt{\mathrm{E}\left[\left(\frac{d}{d\mu}\log f(x;\mu)\right)^2\right]} = \sqrt{\mathrm{E}\left[\left(\frac{x-\mu}{\sigma^2}\right)^2\right]}$$
$$= \sqrt{\int_{-\infty}^{+\infty} f(x;\mu)\left(\frac{x-\mu}{\sigma^2}\right)^2 dx} = \sqrt{\sigma^2/\sigma^4} \propto 1$$

i.e. translation-invariant measure on the real numbers: all mean values equally likely

Example for Jeffreys prior (II)

Gaussian pdf with fixed μ, parameter of interest is the standard deviation parameter σ:

$$f(x;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p(\sigma) \propto \sqrt{\mathcal{I}(\sigma)} = \sqrt{\mathrm{E}\left[\left(\frac{d}{d\sigma}\log f(x;\sigma)\right)^2\right]} = \sqrt{\mathrm{E}\left[\left(\frac{(x-\mu)^2-\sigma^2}{\sigma^3}\right)^2\right]}$$

$$= \sqrt{2/\sigma^2} \propto \frac{1}{\sigma}$$

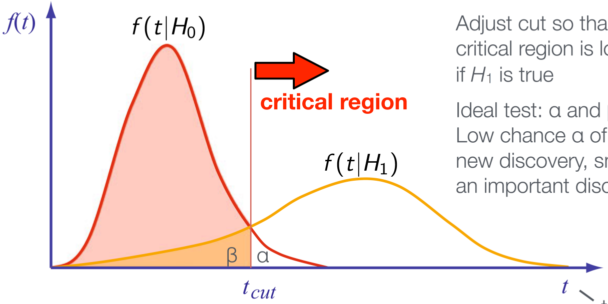
Hypothesis tests

Hypotheses and tests

- Hypothesis test
  - Goal: draw conclusions from the data
  - Statement about validity of a model
  - Decide which of two competing models is more consistent with data
- Simple hypothesis: no free parameters
  - Examples: particle is a π; data follow Poissonian with mean 5
- Composite hypothesis: contains free parameters
- Null hypothesis H_0 and alternative hypothesis H_1
  - H_0 often the background-only hypothesis
    (e.g. Standard Model only; no additional resonance; …)
  - H_1 often signal or signal+background hypothesis
- Question: can H_0 be rejected by data?
- Test statistic t: (scalar) variable that is a function of the data alone, that can be used to test hypothesis

Critical region

Reject null hypothesis if value of t lies in critical region: $t > t_{\text{cut}}$



test statistic

Adjust cut so that probability to be in critical region is low if H_0 is true and high if H_1 is true

Ideal test: α and β small:
Low chance α of incorrectly claiming a new discovery, small chance β of missing an important discovery

Probability for H_0 to be rejected while H_0 is true:

$$\int_{t_{\text{cut}}}^{\infty} f(t|H_0)\mathrm{d}t = \alpha$$

α: "size" or significance level of test

Probability for H_1 to be rejected even though it is true:

$$\int_{-\infty}^{t_{\text{cut}}} f(t|H_1)\mathrm{d}t = \beta$$

$1-\beta$: power of the test

Type I and Type II errors

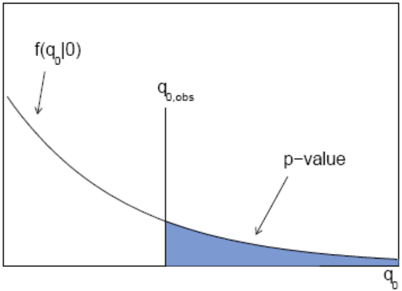
Statistics jargon, getting more and more common also in HEP

Type I error: Probability of rejecting null hypothesis H_0 when it is actually true
also known as false discovery rate

Type II error: Probability to fail to reject null hypothesis H_0 while it is actually false
also known as false exclusion rate

p-value

p-value: probability to observe data set that is as consistent or worse with null hypothesis as the actual observation

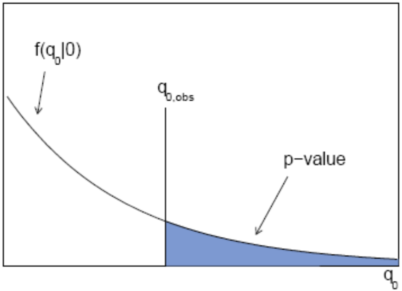


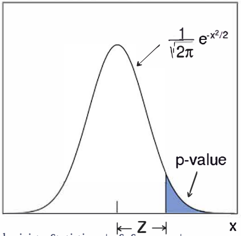
test statistic: q_0
pdf for q_0 under H_0: $f(q_0|0)$
critical region: large values of q_0
$q_{0,\text{obs}}$: observed value in data

$$p_0 = \int_{q_{0,\text{obs}}}^{\infty} f(q_0|0)\mathrm{d}q_0$$

pdf for q_0 under H_0 frequently needs to be estimated with simulation
p-value is a random variable (contrast: significance level α fixed before measurement).
if $p_0 < \alpha$: reject H_0
$1 - p_0$: confidence level of test

p-value and significance





(a)

if $p_0 < \alpha$, then reject null hypothesis

Frequent convention in HEP:

for discovery, require $p < 2.87 \times 10^{-7}$

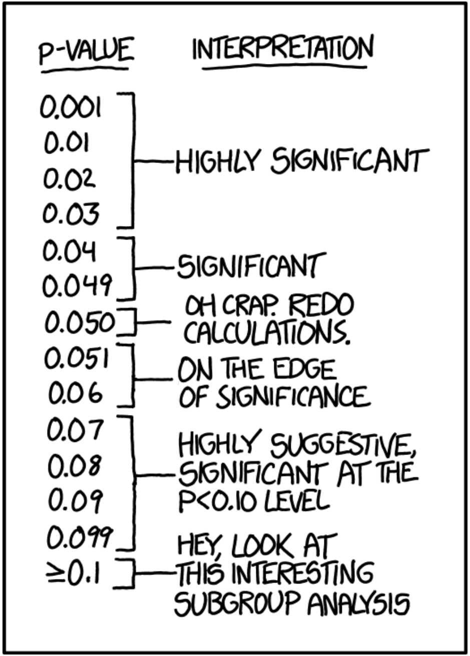
for exclusion, require $p < 0.05$

translate p-value to significance Z via Standard Normal pdf

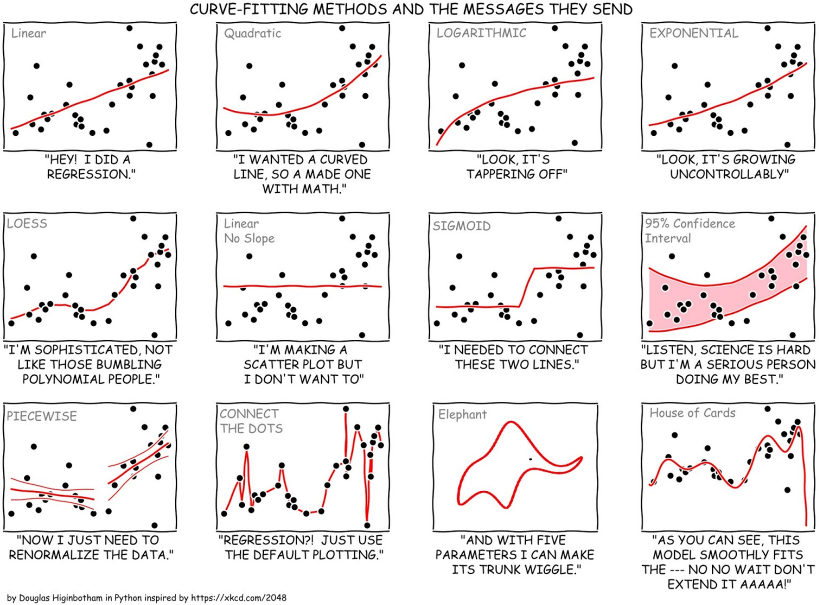
$$p_0 = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z)$$

$$Z = \Phi^{-1}(1 - p_0)$$

Significance of 5 (1.64) s.d. corresponds to $p = 2.87 \times 10^{-7} (0.05)$


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

how can we objectively tell which model fits better?



Least squares: Goodness-of-fit

Minimum value of S in the least squares method is a measure of agreement between model and data:

$$S_{\min} = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i; \widehat{\vec{\theta}})}{\sigma_i} \right)^2$$

Large value of $S_{\min}$: can reject model.

If model is correct, then $S_{\min}$ for repeated experiments follows a χ^2 distribution with n_{df} degrees of freedom:

$$f(t; n_{\mathrm{df}}) = \frac{t^{n_{\mathrm{df}}/2-1}}{2^{n_{\mathrm{df}}/2}\Gamma\left(\frac{n_{\mathrm{df}}}{2}\right)} e^{-t/2}, \quad t = \chi^2_{\min}$$

with $n_{\mathrm{df}} = n - m =$ number of data points $-$ number of fit parameters

Least squares: Goodness-of-fit

Expectation value of χ^2 distribution is n_{df}
➡ $\chi^2 \approx n_{\text{df}}$ indicates good fit

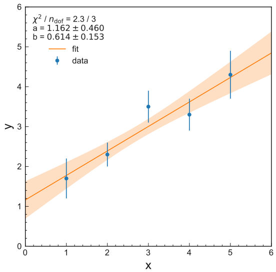
Consistency of a model with data is quantified with the p-value:

$$p = \int_{S_{\text{min}}}^{+\infty} f(t; n_{\text{df}}) \mathrm{d}t$$

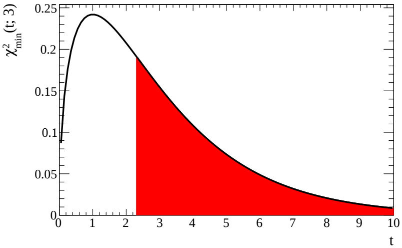
p-value: probability to get a χ^2_{min} at least as high as the observed one, if the model is correct.

p-value is **not** the probability that the model is correct!

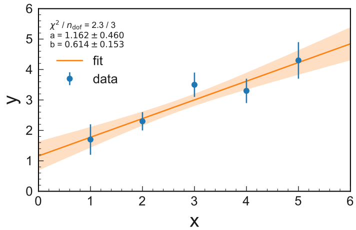
p-value for the straight line fit example



$$S_{\text{min}} = 2.29557,\, n_{\text{df}} = 3$$
$$p\text{-value: prob}(S_{\text{min}}, n_{\text{df}}) = 0.51337011$$



p-value for the straight line fit example

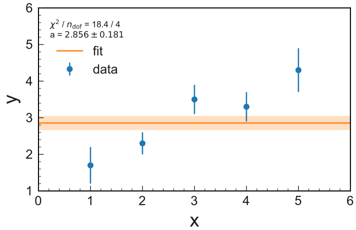


$$S_{\min} = 2.29557, \quad n_{\mathrm{df}} = 3$$

$$p\text{-value} = 0.5134$$

$$\hat{\theta}_0 = 1.16 \pm 0.46$$

$$\hat{\theta}_1 = 0.614 \pm 0.153$$



$$S_{\min} = 18.3964, \quad n_{\mathrm{df}} = 4$$

$$p\text{-value} = 0.00103$$

$$\hat{\theta}_0 = 2.856 \pm 0.181$$

Stat. uncertainty on fit parameter does not tell us whether model is correct

Side remark: quoting χ^2 and ndf

Always remember to quote χ^2 and n_{df} separately,
instead of just the ‘reduced χ^2/n_{df} — there *is* a difference!

$$\mathrm{prob}(15, 10) = 0.132$$
$$\mathrm{prob}(1500, 1000) = 1.05 \times 10^{-22}$$

Goodness of fit for unbinned ML fits

In the case of unbinned ML fit, can bin data and model prediction into histogram and then perform χ^2 test

Consider the likelihood ratio

$$\lambda = \frac{\mathcal{L}(\vec{n}|\vec{\nu})}{\mathcal{L}(\vec{n}|\vec{n})}, \qquad \vec{\nu} = \vec{\nu}(\vec{\theta})$$

For multinomially (“M”, n_{tot} fixed) and Poisson distributed data (“P”), one obtains for k bins

$$\lambda_M = \prod_i^k \left(\frac{\nu_i}{n_i}\right)^{n_i}, \qquad \lambda_P = e^{n_{\text{tot}} - \nu_{\text{tot}}} \prod_i^k \left(\frac{\nu_i}{n_i}\right)^{n_i}$$

Now consider test statistic

$$t \equiv -2 \log \lambda$$

Goodness of fit for unbinned ML fits

For multinomially distributed data, in the large sample limit

$$t_M = -2\log\lambda_M = 2\sum_{i=1}^{k} n_i \log\frac{n_i}{\hat{\nu}_i}$$

follows χ^2 distribution for $k - m - 1$ degrees of freedom.

For Poisson distributed data,

$$t_P = -2\log\lambda_P = 2\sum_{i=1}^{k}\left(n_i \log\frac{n_i}{\hat{\nu}_i} + \hat{\nu}_i - n_i\right)$$

follows χ^2 distribution for $k - m$ degrees of freedom.

Profile likelihood ratio: hypothesis tests with nuisance parameters

Base significance test on the profile likelihood

$$\lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\hat{\theta}})}{\mathcal{L}(\hat{\mu}, \hat{\theta})} = \frac{\text{maximised } \mathcal{L} \text{ for specified } \mu}{\text{globally maximised } \mathcal{L}}$$

Likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma).

Composite hypothesis: parameter μ is only fixed under H_0, but not under H_1.

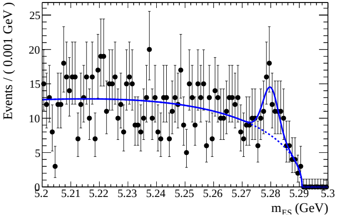
Wilks' theorem:

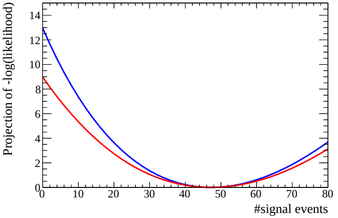
$$q_0 = -2 \log \lambda$$

asymptotically approaches chi-square distribution for k degrees of freedom, where k is the difference in dimensionality of H_1 and H_0

Profile likelihood ratio

Example: B mass fit from last time; 40 signal events, 1000 background events





3 parameters in the fit: signal and background yields, shape parameter for background

$$\hat{n}_{\text{sig}} = 47 \pm 12$$

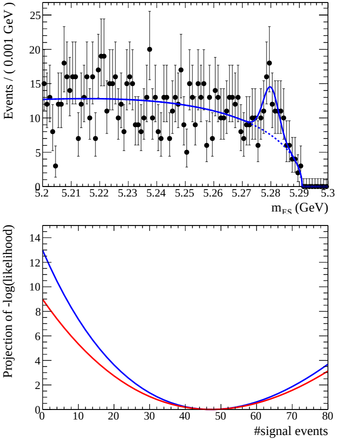
$$\hat{n}_{\text{bkg}} = 992 \pm 33$$

scan of $\mathcal{L}(n_{\text{sig}}, \hat{\theta})$ with nuisance parameters fixed to values from global minimum

profile likelihood: $\mathcal{L}(n_{\text{sig}}; \hat{\hat{\theta}})$

Profile likelihood ratio

Example: B mass fit from last time; 40 signal events, 1000 background events



3 parameters in the fit: signal and background yields, shape parameter for background

$$\hat{n}_{\mathrm{sig}} = 47 \pm 12$$
$$\hat{n}_{\mathrm{bkg}} = 992 \pm 33$$

From scan of profile likelihood:

$$2\Delta \log \mathcal{L} = 17.94$$

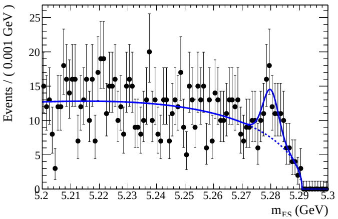
And therefore p-value for H_0:
$1.139\,27 \times 10^{-5}$, or significance for $n_{\mathrm{sig}} \neq 0$

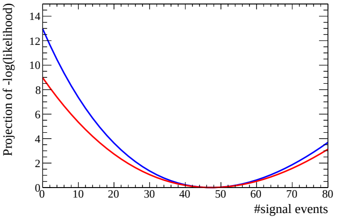
$$Z = \sqrt{2\Delta \log \mathcal{L}} = 4.2\sigma$$

(one degree of freedom!)

Profile likelihood ratio

Example: B mass fit from last time; 40 signal events, 1000 background events





3 parameters in the fit: signal and background yields, shape parameter for background

$$\hat{n}_{\text{sig}} = 47 \pm 12$$
$$\hat{n}_{\text{bkg}} = 992 \pm 33$$

now leave also mean and width of signal peak free in fit: two additional nuisance parameters (that cannot really be determined when $n_{\text{sig}} = 0$).

p-value = 0.0697557

$Z = 1.48\ \sigma$

Look-elsewhere effect

A Swedish study in 1992 tried to determine whether or not power lines caused some kind of poor health effects. The researchers surveyed everyone living within 300 meters of high-voltage power lines over a 25-year period and looked for statistically significant increases in rates of over 800 ailments. The study found that the incidence of childhood leukemia was four times higher among those that lived closest to the power lines, and it spurred calls to action by the Swedish government. The problem with the conclusion, however, was that they failed to compensate for the look-elsewhere effect; in any collection of 800 random samples, it is likely that at least one will be at least 3 standard deviations above the expected value, by chance alone. Subsequent studies failed to show any links between power lines and childhood leukemia, neither in causation nor even in correlation.

`https://en.wikipedia.org/wiki/Look-elsewhere_effect`

Look-elsewhere effect

In general, a p-value of $1/n$ is likely to occur after n tests.

Solution: apply ‘trials penalty’, or ‘trials factor’, *i.e.* make threshold more stringent for large n.

Not entirely trivial to choose trials factor: need to count effective number of ‘independent’ regions. Suppose you look at a range of invariant masses large compared to the mass resolution, then $N \sim \Delta M/\sigma_M$.

See e.g. Gross & Vitells, arXiv:1005.1891 [physics.data-an] for a recipe

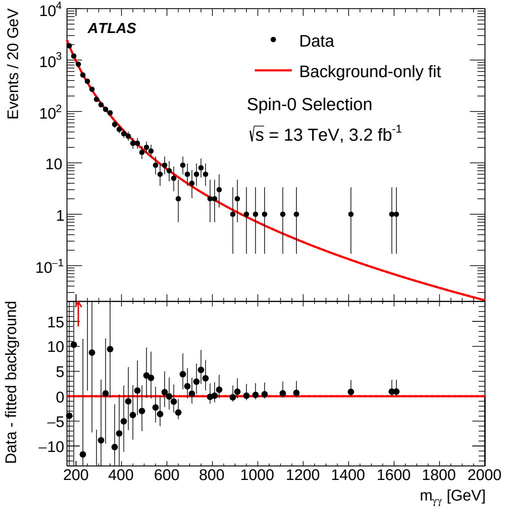
Look-elsewhere effect

Can make substantial change to claimed significance:

for example ATLAS observation of an enhancement around 750 GeV in $\gamma\gamma$ invariant mass:

Local significance 3.9σ, corresponding to a p-value of $p = 9.6 \times 10^{-5}$, i.e. roughly 1:10000

Global significance only 2.1σ, corresponding to a p-value of $p = 0.0357$, i.e. roughly 1:28



ATLAS, JHEP 09 (2016) 001

(Final) digression: *p*-value debate

In many fields (esp. social sciences, psychology, etc.), significant means $p < 0.05$

Relatively weak statistical standard, but often not realised as such!

We've seen that getting $p < 0.05$ isn't that rare, especially if you run many experiments!

May be a contributing factor to the 'reproducibility crisis'
and may be exacerbated by p-value hacking

5σ for discovery in particle physics?

5σ corresponds to p-value of 2.87×10^{-7} (one-sided test)

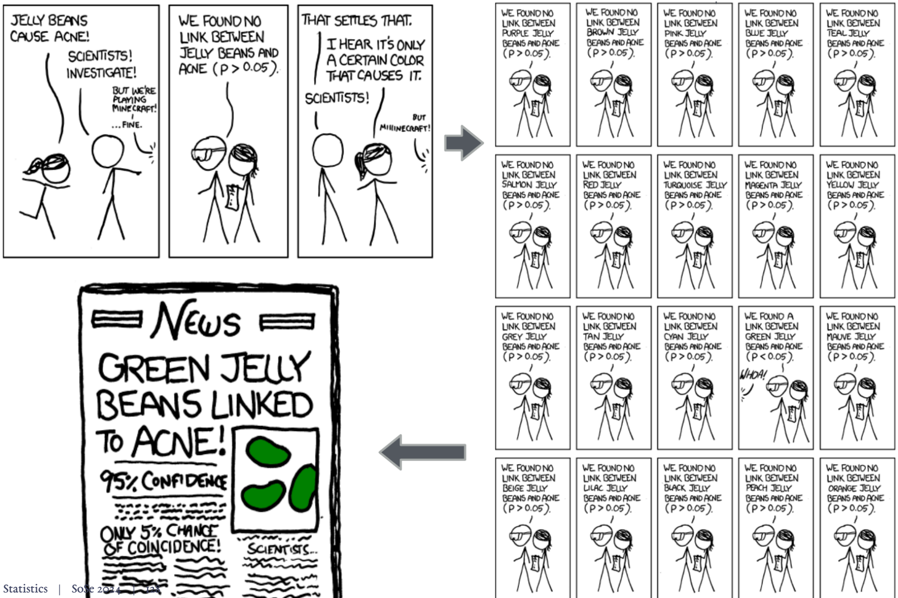
- History: many cases where 3σ and 4σ effects have disappeared with more data
- Look-elsewhere effect
- Systematics: often difficult to quantify / estimate
- Subconscious Bayes factor:
  - physicists tend to (subconsciously) assess Bayesian probabilities $p(H_1|\text{data})$ and $p(H_0|\text{data})$
  - If H_1 involves something very unexpected (e.g. superluminal neutrinos), then prior probability for H_0 is much larger than for H_1
  - Extraordinary claims require extraordinary evidence

May be unreasonable to have single criterion for all experiments

Louis Lyons, Statistical issues in searches for new physics, arXiv:1409.1903

p-value hacking

http://xkcd.com/822



Appendix

Best Linear Unbiased Estimate (BLUE)

Have seen how to combine uncorrelated measurements.

Now consider n data points y_i, $\vec{y} = (y_1, \ldots, y_n)$ with covariance matrix V.

Calculate weighted average λ by minimising

$$\chi^2(\lambda) = (\vec{y} - \vec{\lambda})^T V^{-1} (\vec{y} - \vec{\lambda}) \qquad \vec{\lambda} = (\lambda, \ldots, \lambda)$$

Result:

$$\hat{\lambda} = \sum_i w_i y_i, \qquad \text{with} \quad w_i = \frac{\sum_k (V^{-1})_{ik}}{\sum_{k,l} (V^{-1})_{kl}}$$

Variance:

$$\sigma^2_{\hat{\lambda}} = \vec{w}^T V \vec{w} = \sum_{i,j} w_i V_{ij} w_j$$

This is the best linear unbiased estimator, i.e. the linar unbiased estimator with the lowest variance

BLUE

Special case: two correlated measurements

Consider two measurements y_1, y_2, with covariance matrix (ρ is correlation coefficient)

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Applying formulas from above:

$$V^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}; \qquad \hat{\lambda} = wy_1 + (1-w)y_2$$

$$w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}; \qquad V[\hat{\lambda}] = \sigma^2 = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

Weighted average of correlated measurements: interesting example

adapted from Cowan's book and Scott Oser's lecture:

Measure length of an object with two rulers. Both are calibrated to be accurate at temperature $T = T_0$, but otherwise have a temperature dependency: true length y is related to measured length L by

$$y_i = L_i + c_i(T - T_0)$$

Assume that we know c_i and the (Gaussian) uncertainties. We measure L_1, L_2, and T, and want to combine the measurements to get the best estimate of the true length.

Weighted average of correlated measurements

Start by forming covariance matrix of the two measurements:

$$y_i = L_i + c_i(T - T_0); \qquad \sigma_i^2 = \sigma_L^2 + c_i^2\sigma_T^2$$

$$\mathrm{cov}[y_1, y_2] = c_1 c_2 \sigma_T^2$$

Use the following parameter values, just for concreteness:

$$c_1 = 0.1 \qquad L_1 = 2.0 \pm 0.1 \qquad y_1 = 1.80 \pm 0.22 \qquad T_0 = 25$$

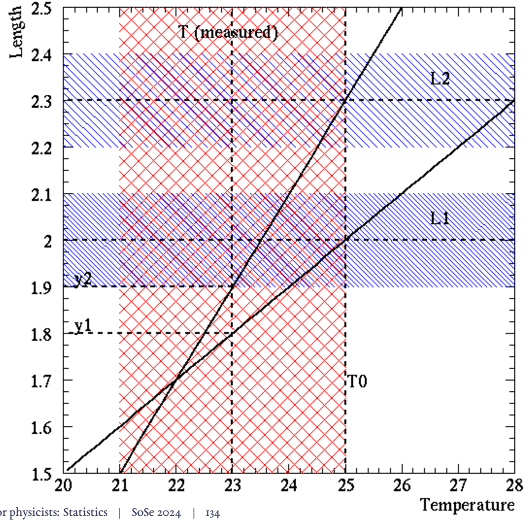
$$c_2 = 0.2 \qquad L_2 = 2.3 \pm 0.1 \qquad y_2 = 1.90 \pm 0.41 \qquad T = 23 \pm 2$$

With the formulas above, we obtain the following weighted average

$$y = 1.75 \pm 0.19$$

Why doesn't y lie between y_1 and y_2? Weird!

Weighted average of correlated measurements



y_1 and y_2 were calculated assuming $T = 23$

Fit adjusts temperature and finds best agreement at $\hat{T} = 22$

Temperature is a nuisance parameter in this case

Here, data themselves provide information about nuisance parameter

Addendum: Linear least squares (I)

Fit model: $y = \theta_1 x + \theta_0$

Apply general solution developed for linear least squares fit:

$A_{i,j} = a_j(x_i)$

$L = (A^T V^{-1} A)^{-1} A^T V^{-1}, \qquad \hat{\vec{\theta}} = L\vec{y}$

$$A^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}; \quad V^{-1} = \begin{pmatrix} 1/\sigma_1^2 & & & \\ & 1/\sigma_2^2 & & \\ & & \ddots & \\ & & & 1/\sigma_n^2 \end{pmatrix}$$

$$A^T V^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 1/\sigma_2^2 & \cdots & 1/\sigma_n^2 \\ x_1/\sigma_1^2 & x_2/\sigma_2^2 & \cdots & x_n/\sigma_n^2 \end{pmatrix}$$

$$A^T V^{-1} A = \begin{pmatrix} 1/\sigma_1^2 & 1/\sigma_2^2 & \cdots & 1/\sigma_n^2 \\ x_1/\sigma_1^2 & x_2/\sigma_2^2 & \cdots & x_n/\sigma_n^2 \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} \sum_i 1/\sigma_i^2 & \sum_i x_i/\sigma_i^2 \\ \sum_i x_i/\sigma_i^2 & \sum_i x_i^2/\sigma_i^2 \end{pmatrix}$$

Addendum: Linear least squares (II)

2×2 matrix easy to invert. Using shorthand notation $[z] = \sum_i z/\sigma_i^2$:

$$(A^T V^{-1} A)^{-1} = \frac{1}{[1][x^2] - [x][x]} \begin{pmatrix} [x^2] & -[x] \\ -[x] & [1] \end{pmatrix}$$

And therefore

$$\begin{aligned} L &= (A^T V^{-1} A)^{-1} A^T V^{-1} \\ &= \frac{1}{[1][x^2] - [x][x]} \begin{pmatrix} [x^2] & -[x] \\ -[x] & [1] \end{pmatrix} \cdot \begin{pmatrix} 1/\sigma_1^2 & 1/\sigma_2^2 & \cdots & 1/\sigma_n^2 \\ x_1/\sigma_1^2 & x_2/\sigma_2^2 & \cdots & x_n/\sigma_n^2 \end{pmatrix} \\ &= \frac{1}{[1][x^2] - [x][x]} \begin{pmatrix} \frac{[x^2]}{\sigma_1^2} - \frac{[x]x_1}{\sigma_1^2} & \cdots & \frac{[x^2]}{\sigma_n^2} - \frac{[x]x_n}{\sigma_n^2} \\ \frac{-[x]}{\sigma_1^2} + \frac{[1]x_1}{\sigma_1^2} & \cdots & \frac{-[x]}{\sigma_n^2} + \frac{[1]x_n}{\sigma_n^2} \end{pmatrix} \end{aligned}$$

And finally:

$$\hat{\theta}_0 = \frac{[x^2][y] - [x][xy]}{[1][x^2] - [x][x]}, \qquad \hat{\theta}_1 = \frac{-[x][y] + [1][xy]}{[1][x^2] - [x][x]}$$