EXAMPLES / EXERCISES

- Religious belief / Dogma

Take a set of possible events $\{A_i\}$ constituting a non-intersecting partition of Ω.
Imagine we have the following prior probability for each A_i:

$$\pi(A_i) = \begin{cases} 1 & \text{if } i = 0 \\ \phi & \text{if } i \neq 0 \end{cases}$$

which means we believe A_1 is absolutely true, and $A_{i \neq 1}$ are absolutely false.
Assume we acquire more information through the observation of some event B.

From Bayes theorem: $$P(A_i | B) = \frac{P(B | A_i)\, \pi(A_i)}{P(B)}$$

If $i \neq 0$: $P(A_i | B) = 0 = P(A_i)$

If $i = 0$: $$P(A_0 | B) = \frac{P(B|A_0)\,\pi(A_0)}{\sum_i P(B|A_i)\,\pi(A_i)} = \frac{P(B|A_0)\cdot 1}{P(B|A_0)\cdot 1} = 1 = P(A_0)$$

This is what we call a dogma or religious belief, i.e. a belief that we cannot change, no matter how much new knowledge we accumulate.

The scientific method has allowed to improve our knowledge of Nature by progressingly accumulating new experimental evidences.
On the other hand, scientific progress is not possible in the presence of religious beliefs about observable facts.

- Change of variables – Quadratic dependence

Take the variables x and $y = x^2$.

Assume x is Gaussian-distributed with $\mu = 2$ and $\sigma = 1$.

Populate a histogram with 10^6 random values of x, and a second histogram with the corresponding values of y.

Compare the random-generated distribution of y with the theoretical curve.

Solution: $$f(y) = \sum \frac{f(x)}{|Y'(x)|} = \sum \frac{f(x)}{\left|\frac{dy}{dx}\right|}$$

$\downarrow$ sum over segments for $x<0$ and $x \geq 0$

$$= \frac{f(-x) + f(x)}{2|x|}$$

Do the same assuming x is Poisson-distributed with $\lambda = 3$.

- Change of variables – Ratio

Take the variables x and y, Gaussian distributed with $\mu = 0$ and $\sigma = 1$.
Compute (analytically and numerically) the distribution of their ratio $\frac{x}{y}$.

Solution: Define a change of variable: $u = \frac{x}{y}$

$v = y$

The joint-PDF is: $$h(u,v)\,du\,dv = f(x)\,g(y)\,dx\,dy = f(uv)\,g(v)\left\|\frac{\partial(x,y)}{\partial(u,v)}\right\| du\,dv = f(uv)\,g(v)\,v\,du\,dv$$

The density of u is: $$p(u) = \int_0^\infty f(uv)\,g(v)\,v\,dv \underbrace{- \int_{-\infty}^0 f(uv)\,g(v)\,v\,dv}_{\text{minus sign for } v<0}$$

$$p(u) = \int_0^\infty \frac{1}{2\pi} e^{-\frac{u^2v^2}{2}} e^{-\frac{v^2}{2}}\,dv - \int_{-\infty}^0 \frac{1}{2\pi} e^{-\frac{u^2v^2}{2}} e^{-\frac{v^2}{2}}\,dv$$

$$= \frac{1}{2\pi}\int_0^\infty e^{-\frac{u^2+1}{2}v^2}\,dv - \int_{-\infty}^0 \cdots$$

$$= \frac{1}{2\pi}\frac{6}{u^2+1}x^3\Big|_0^\infty + \frac{1}{2\pi}\frac{6}{u^2+1}x^3\Big|_{-\infty}^0$$

$$p(u) = \frac{1}{\pi(1+u^2)} \rightarrow \text{Cauchy distro (infinite variance)}$$

$\boxed{\text{Ex 2}}$

• Binomial

~~Plot, [illegible] on the same~~

~~Compute the~~ Compute the distribution of Binomial variables with $n = 150$ and

$p = 0.1,\ 0.3,\ 0.5,\ 0.7,\ 0.9$

Plot them all together (overlay them). Do not use any predefined method for the binomial or factorial!

Repeat for $n = 500$.

Tricks: 1) Binomial coefficient with exp(log)

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n\cdot(n-1)\cdot(n-2)\cdots 3\cdot 2\cdot 1}{k(k-1)\cdots 3\cdot 2\cdot (n-k)(n-k-1)\cdots 3\cdot 2}$$

$$= \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 3\cdot 2}$$ ↳ works numerically up to $k \approx 170$

$$= \exp\left[\ln\left(n(n-1)\cdots(n-k+1)\right) - \ln\left(k(k-1)\cdots 3\cdot 2\right)\right]$$

$$= \exp\left[\sum_{i=n-k+1}^{n} \ln(i) - \sum_{i=2}^{k} \ln(i)\right]$$

2) Bernoully term with exp(log) to solve situation with small p and large (1-p)

$$p^k (1-p)^{n-k} = \exp\left[\ln\left(p^k (1-p)^{n-k}\right)\right]$$

$$= \exp\left[k \ln(p) + (n-k)\ln(1-p)\right]$$

• ~~Poisson~~

~~Populate histograms with 10^5 random generated values~~

Ex 3

• ~~Poisson~~

~~Populate histograms with random generated events~~

~~1) Generate 10^5 uniformly distributed values of t between 0 and $\Delta t = 10^3$~~

~~2) Test wether t happens within a window $\delta t = p \cdot \Delta t$ with $p = 0.01$~~

• Poisson

1) Generate $n = 10$ uniformly distributed values of t in interval $[0, \Delta t = 10^3]$
2) Compute number of accepted events k that happen in interval $[0, \delta t]$ with $\delta t = p \cdot \Delta t$ and $p = 0.01$
3) Repeat 10^5 times, and populate histogram of k
4) Plot histogram of k and compare with theoretical Binomial and Poisson distribution. Do not use any predefined method!
5) Repeat for all combinations of: $n = 10, 100, 1000$
   $p = 0.01, 0.05, 0.1, 0.5$

• Exponential from uniform

1) Generate 10^6 values of variable t, uniformly distributed in $[0, \Delta t = 10^3]$
2) Sort the list of generated values
3) Compute the interval between subsequent events
   $$\Delta t_i = t_i - t_{i-1}$$
4) Plot the distribution of Δt_i (on a histogram)
5) Fit it with an exponential. Does it work? Do you get back the input parameters?

$$\cancel{f(x, n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)}$$

- χ^2 distribution

  ~~1) Populate a histogram with 10^5 randomly generated~~

  1) Generate 10^5 sets of n standard-normal random numbers x, and populate a histogram with $\chi^2 = \sum_{i=1}^{n} x^2$
  2) Repeat point 1 for $n = 1, 2, 3, 5, 10, 20$
  3) Overlap the histograms with the theoretical χ^2 distribution with the corresponding number of DOF, properly scaled to the number of generated values.

     Do not use any predefined χ^2 method.

- Central limit theorem for Gaussian random number generator

  1) Generate 10^8 sets of n uniformly distributed random numbers in $[0, 1[$, with $n = 1, 2, \dots, 10$

  ~~2) Populate For every n, populate a histogram~~

  2) For every n, populate a histogram with the 10^8 sums of the n generated random numbers
  3) Plot the histogram in linear and log-scale (on the y axis)

     Do you notice anything?

- Acceptance-rejection method

  1) Generate 10^6 points uniformly distributed in $x \in [-1, 1]$ and $y \in [-1, 1]$
  2) Compute the fraction of points falling inside a circle of radius 1
  3) Repeat for $\dim = 2, 3, 4, \dots, 15$
  4) Plot fraction of accepted points as a function of dimension
  5) Assign uncertainty to your estimate of accepted points
  6) Compare (overlap plots) with theoretical expectation of number of accepted points.

$\boxed{\text{Ex 5}}$

- Random number correlation

1) Generate $(10^5 + 1)$ random numbers distributed as a Gaussian with $\mu = 3$ and $\sigma = 2$

2) Plot the distribution of x_i and of $(x_i - x_{i-1})$

3) Do the same, but ~~populate the~~ this time generate the numbers x_i using Metropolis-Hastings, until you accept $(10^5 + 1)$ points
As a proposal function, use a Gaussian with $\mu = x_i$ and $\sigma = 0.3$

How do the distributions look like?
Can you explain the STD of the $(x_i - x_{i-1})$ distributions in the two cases?

- Extended vs binned likelihood

1) Generate data distributed with the following signal and background PDFs:

$$f_S(E)\big|_{\mu} = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(E-\mu)^2}{2\sigma^2}\right) \quad \text{with} \quad \mu = 2039 \text{ keV}, \ \sigma = 1.5 \text{ keV}$$

$$f_B(E) = \frac{1}{E_{max} - E_{min}} \quad \text{with} \quad E_{min} = 2000 \text{ keV}, \ E_{max} = 2080 \text{ keV} \ \to \text{this is our fit range}$$

Generate ~~data~~ 10^4 Toy-MC ~~spectra~~ data using all combinations of the following number of ~~co~~ signal and background counts:

$S = \{5, 20, 50, 100\}$

$B = \{10, 30, 100, 200\}$

2) Fit each Toy-MC ~~spectra~~ data with the extended likelihood (Poisson)

3) Bin the Toy-MC data with 1 keV binning, and fit the spectra with the binned $\mathcal{L}$ (Poisson term for each bin)

4) Extract the ~~average~~ CPU time required for each ~~combinati~~ fit.

5) Plot the average of $\frac{t_{extended}}{t_{binned}}$ vs $\frac{N_{events}}{N_{bins}}$

$\boxed{\text{Ex 6}}$

- Scatter plot (χ^2 fit)

1) Generate $N = 11$ points with the following coordinates:

$$x_i = 0, 100, 200, \ldots, 1000$$

$$y_i = m\, x_i + q + \delta y_i \qquad \text{where} \qquad m = 1.5$$

$$q = 3$$

$$\delta y_i = \text{random Gaus}(\mu = 0, \sigma_i)$$

$$\sigma_i = \text{random uniform in } [10, 100[$$

To each point, assign the uncertainties:

$$\sigma_{x_i} = 0$$

$$\sigma_{y_i} = \sigma_i$$

2) Fit the scatterplot with the χ^2 fit, and retrieve the minimum χ^2 from the fitter.

↳ In ROOT:

```
TFitResultPtr resPtr = scatterplot -> Fit(...);
TFitResult* res = resPtr.Get();
double chi2 = res->Chi2();
```

3) Repeat for 10^4 Toy-MC data, and plot the histogram of min-χ^2.

4) Overlap with the theoretical χ^2 distribution with the same number of DOF

↳ $nDOF = n - 2$ (because we have 2 parameters)

- Scatter plot χ^2 fit with overestimated uncertainty

Repeat the previous exercise, artificially enlarging the uncertainty assigned to y_i with respect to the one used to generate δy_i.

Use the following: $\sigma_{y_i} = k\, \sigma_i$ with $k = \{1, 1.1, 1.2, 1.3, 1.4\}$

- Scatter plot χ^2 fit with unaccounted systematic

Repeat the scatter plot exercise but generate the data with a 2-nd order polinomial:

$$y_i = q + m x_i + c x_i^2 \qquad c = \{0, 10^{-4}, 2\cdot 10^{-4}, 3\cdot 10^{-4}, 4\cdot 10^{-4}, 5\cdot 10^{-4}\}$$

However, keep fitting the data with the linear function

$$y = mx + q$$

752

- Likelihood vs χ^2 for binned Poisson data

1) Generate data distributed with the following signal and background PDFs:

$$f_s(E) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(E-\mu)^2}{2\sigma^2}\right) \qquad \text{with } \mu = 2039 \text{ keV}, \ \sigma = 1.5 \text{ keV}$$

$$f_b(E) = \frac{1}{E_{max} - E_{min}} \qquad \text{with } E_{min} = 2000 \text{ keV}, \ E_{max} = 2080 \text{ keV}$$

Generate 300 ToyMC data using all combinations of the following number of signal and background counts:

$$s = \{10, 100, 1000, 10\,000\}$$
$$b = \{10, 100, 1000, 10\,000\}$$

Bin the data with 0.1 keV binning

2) Fit each histogram with $\mathcal{L}$, χ^2_N, and χ^2_P

3) Produce the following plots:

a) Distribution of best fit of s for all combinations of s and b, ~~b) Distrib~~ with the distributions obtained from the 3 methods overlapped together

b) Same, but for b

c) 2-dim distribution of $\hat{s}$ vs $\hat{b}$ for all combinations of s and b. This time, do 3 separate plots for $\mathcal{L}$, χ^2_N, χ^2_P otherwise it's a mess.

Ex 8

- Efficiency curve fit with ~~binomial, Poisson and a~~ binomial $\mathcal{L}$, Poisson $\mathcal{L}$ and Gaussian χ^2

1) Generate 20 points with coordinates (E_i, K_i)

$$E_i = 1, 2, 3, \ldots, 20 \text{ keV}$$

$$\varepsilon_i = \frac{p}{2}\left[1 + \mathrm{erf}\left(\frac{E-\mu}{\sigma}\right)\right] \qquad \text{with } \begin{array}{l} p = 0.9, 0.95, 0.99 \\ \mu = 3 \text{ keV} \\ \sigma = 1 \text{ keV} \end{array}$$

$\forall \varepsilon_i$, generate n random values uniformly distributed in $[0, 1[$ and ~~count how set $K_i = \#$,~~ set K_i to the number of values $< \varepsilon_i$.

Consider the following values for n: 100, 1000, 10 000

2) Fit the data ~~assumi~~ assuming the following distributions for K_i:

   a) Binomial

   b) Poisson

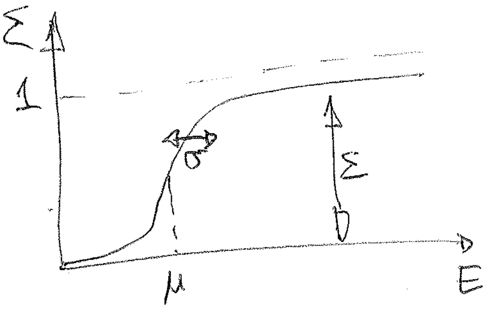
   c) χ^2 (Neyman or Pearson, your choice)

~~3) Plot the chart Repeat for all~~

3) Repeat for 10^3 toys for all combinations of n and p

4) Plot the best fit distribution of $\hat{\varepsilon}$ obtained with the 3 methods, for all combinations of p and n.

This ~~example~~ is the case of a pulse generator injecting n ~~pu~~ artificial pulses into a detector, with 20 different energy values between 1 and 20 keV.



ε is the overall trigger efficiency

μ is the trigger threshold

σ describes the rise of the efficiency curve.

$\boxed{\text{Ex. 9}}$

- Ordering rules

Suppose to have a variable λ with PDF $f(\lambda) = \text{Poisson}(n, \lambda) = \frac{e^{-\lambda}\lambda^n}{n!}$

with $n = 3$.

Notice that we are looking at the Poisson as a function of the expectation value, not of the number of measured counts!

1) Compute (and plot) the 68% intervals with the following methods:
   a) central interval
   b) equal areas
   c) shortest interval

2) As a cross check, print the coverage of each method.

- Neyman belt for Gaussian

Assume $t(x) = x$ and $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$ with θ unknown, $\sigma = 1$

1) Create a 2-dim histogram of (x, θ) with $0 \leq x \leq 10$ and $0 \leq \theta \leq 20$.
2) For each value of θ, populate the corresponding x with 10^5 values of x randomly sampled from $f(x|\theta)$.
3) Invert the belt for $x = 5$, and plot ~~f(θ|x=5)~~ the corresponding $f(x)$
4) Fit ~~the~~ $f(\theta)$ with a Gaussian

- Neyman belt for Binomial

Assume ~~again~~ we have a detector injected with n pulser events, and that we want to measure the trigger efficiency p.

Take: n = # of injected events = $\{0, 1, 2, \ldots, 10\}$

p = input trigger probability $= 0, 10^{-3}, 2\cdot 10^{-3}, \ldots, 1$

1) Create a 2-dim histogram for (k, p).
   Notice that $0 \leq k \leq n$, so k must go from 0 to 10.
2) For each value of p, populate the corresponding distribution of k with a Binomial$(k|n,p)$
3) For each value of p, extract the shortest 68% interval
4) Plot the Neyman confidence belt
5) Compute the coverage of p.

- Flip-flopping

Take a variable x with PDF: $f(x|\mu) = \frac{1}{2\pi} \exp\left(-\frac{(x-\mu)^2}{2}\right)$

Assume x can be negative, but $\mu \geq 0$.

Assume we decide to quote a central 90% interval if $x \geq 3$, or an upper 90% limit if $x < 3$.

1) Prepare a 2-dim histogram for (x, μ)
2) Populate the Neyman belt
3) Construct the 90% central confidence belt for $x \geq 3$
4) Construct the 90% lower confidence belt for $x < 3$
5) Join the two belts and plot them.
6) Compute and plot the coverage of μ.

Ex 11

- Feldman - Cousins for Gaussian case

Take a variable x with range $[-3, 11]$ and PDF $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$

1) Prepare a 2-dim histogram for (x, μ) with $\mu \in [0, 7]$

2) Populate the Neyman belt

3) Find $\max(f(x|\mu))$ for each value of x and ~~store~~ plot it as a function of x. Compare it with the analytical calculation.

4) Populate a 2-dim histogram (same ranges as before) with the likelihood ratio

5) Construct and plot the FC confidence belt.

6) Compute the coverage as a function of μ and plot it using the values of the Neyman belt.

7) Compute the coverage using Toy-MC:

   7.a) Generate 10^4 values of x for a fixed value of μ

   7.b) Compute the fraction of x that fall inside the FC belt for that specific value of μ

   7.c) ~~Iterate~~ Repeat for all values of μ

Ex 12

• Feldman - Cousins for electron neutrino mass

In this example, we try to reproduce the ~~mass~~ limit on electron ν mass reported by ~~KATRIN~~ KATRIN in Phys. Rev. Lett. 123 (2019) 221802.

Their limit is: $m_\nu < 1.1\ eV$

obtained from: $m_\nu^2 = -1.0^{+0.9}_{-1.1}\ eV^2$

~~1) Prepare a bank~~ Our variables are m_ν (the parameter of the model) and m^2 (the observable). Notice that m_ν can't be negative, but the observable m^2 can.

Let's approximate the PDF of m^2 with: $f(m^2 | m_\nu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(m^2 - m_\nu^2)^2}{2\sigma^2}\right)$

Let's assume $\sigma = 1$, which is close to the asymmetric uncertainties quoted above.

1) Populate the Neyman belt, using $m^2 \in [-5, 20]$ and $m_\nu \in [0, 4]$

2) Extract $\max f(m^2 | m_\nu)$ for each value of m_ν

3) Populate the likelihood ratio 2-dim histogram

4) Extract the FC confidence belt

5) Compute the coverage (using the Neyman belt)

6) Extract the central interval or limit ~~correspondi~~ on m_ν corresponding to a measured value $m^2 = -1.0$

How does it compare with the published one?

$\boxed{\text{Ex 13}}$

• Example: Poisson posterior for rate

Suppose we have a random variable k with Poisson PDF

$$f(k|\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!} \qquad \text{with } \lambda = 5 \rightarrow \text{This is the \underline{true} value of } \lambda.$$

~~Suppose we measure k once, obtaining k = 5.~~

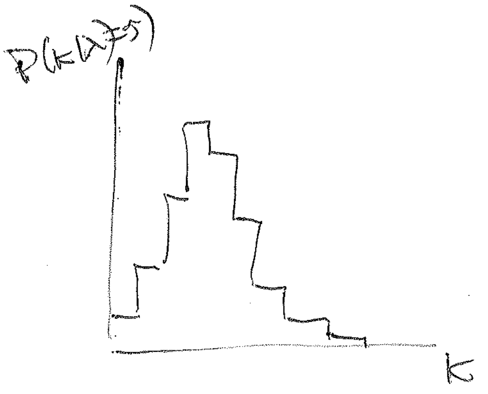
Suppose our prior knowledge on k is flat between 0 and 20.

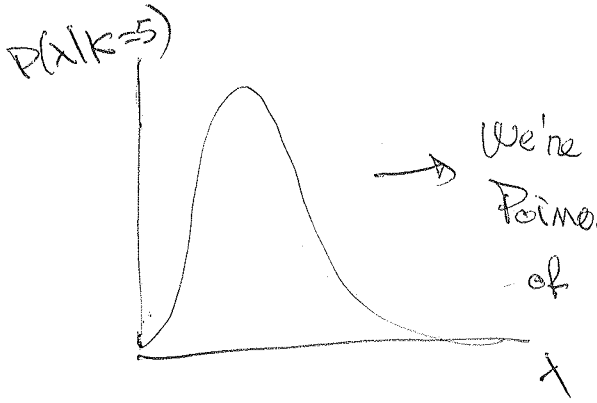
Suppose we measure k once, obtaining $k = 5$.

~~What is our~~

1) Compute the posterior PDF for λ
2) What's the ~~best fit~~ most probable value for λ?
3) What's the 68% shortest C.I.?
4) Why is the best fit $\hat{\lambda} \neq 5$ (as we would expect)?

You expect something like





→ We're looking at the Poisson as a function of λ now!!!

- Bayesian efficiency fit

1) Generate 20 points with coordinates (E_i, k_i)

$$E_i = 1, 2, 3, \ldots, 20 \text{ keV}$$

$$\mathcal{E}_i = \frac{p}{2}\left[1 + \mathrm{erf}\left(\frac{E-\mu}{\sigma}\right)\right] \quad \text{with} \quad \begin{array}{l} p = 0.99 \\ \mu = 3 \text{ keV} \\ \sigma = 1 \text{ keV} \end{array}$$

$\forall\, \mathcal{E}_i$, generate 1000 random values uniformly distributed in $[0, 1[$ and set k_i to the number of values $< \mathcal{E}_i$.

2) Run a Bayesian fit with Metropolis-Hastings using the following distributions for k_i:

a) Binomial

b) Poisson

c) χ^2

3) Plot the posterior of p for the three fits.

- ~~Radioactive decay fit~~

- Simultaneous fit

Perform a combined fit of 2 data sets:

a) Pulser-injected data produced in the same way as in the "efficiency" example.

b) Physics data, recorded between 10 an 20 keV, and featuring:

→ a Gaussian signal with $\mu = 15$ keV and $\sigma = 1.2$ keV

→ a flat background

1) Generate pulser data as in the previous example

2) Generate 20 signal and 20 background events, and apply the efficiency curve to them

3) Perform a simultaneous fit, with E_a as a common parameter.

- Evidence

1) Generate the usual data with a flat bkg and Gaussian signal.

Use: $s = 5$

$b = 25$

Unbinned PDF. Just one toy dataset.

2) Run ~~the fit twi~~ a Bayesian fit twice:

a) Using H_0 = bkg only

b) Using H_1 = signal + bkg

Use flat priors for both s and b

3) Compute the evidence of H_0 and H_1 using the acceptance-rejection method

4) Compute the Bayes factor, $P(H_0|\vec{E})$ and $P(H_1|\vec{E})$