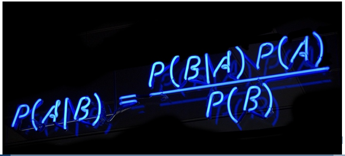
Statistics in Data Analysis

All you ever wanted to know about statistics but never dared to ask

part 6

Pawel Brückman de Renstrom
(`pawel.bruckman@ifj.edu.pl`)

April 17, 2024



Putting it to work...

1D fit of the signal yield

N data events have been collected in an experiment yielding a scalar random variable x:

- The sample consists of a mixture of *signal* and *background* events with known p.d.f.'s.
- Background p.d.f.: $f_B(x) = \frac{1}{\tau} e^{-x/\tau}$, $\tau = 5$,
- Signal p.d.f.: $f_S(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$, $\mu = 10$, $\sigma = 3$,
- The varaible x has been recorded in the range $(0, 30)$.
- No assumption about background yield can be made: $N = N_S + N_B$.

Our task is to:

1. Estimate number of signal events N_S in the observed sample,
2. assess the error of the N_S estimate from the $\log L$ or χ^2 profile.

MIND: This is NOT an extended fit.

Putting it to work...

1D fit of the signal yield

We shall use three strategies:

1. the *Unbinned Maximum Likelihood* fit:
   `Unbinned ML Python notebook template in Colab`
2. the *Binned Maximum Likelihood* fit; 10 bins over (0, 30):
   `Binned ML Python notebook template in Colab`
3. the *Binned Least Squares* (NOT modified) fit; 10 bins over (0, 30):
   `Binned LS Python notebook template in Colab`

and four data samples:

1. `pickled data sample 1 from GitHub`
2. `pickled data sample 2 from GitHub`
3. `pickled data sample 3 from GitHub`
4. `pickled data sample 4 from GitHub`

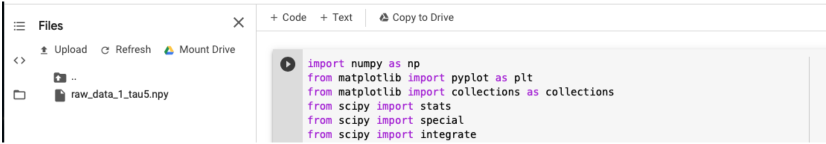
All shall be executed on the *Google Colaboratory* platform.

Detailed instructions

⟶ Click on one of the Python notebook links in order to open it in *Google Colaboratory*.

⟶ Click to download the assigned dataset file from GitHub.

⟶ Click on one of the *Files* icon on the left bar of your *Colab* interface. If you cannot see any datafiles, click on the *Upload* button and select previously downloaded file. As a result you should see:



```
import numpy as np
from matplotlib import pyplot as plt
from matplotlib import collections as collections
from scipy import stats
from scipy import special
from scipy import integrate
```

Hints part 1: fit the N_s, $\log L$ or χ^2 function

$\longrightarrow$ You need p.d.f. normalization factors in the range (0,30). For this purpose calculate `scales` and `scaleb` just as it is done in the main part of the Python script. You will need `xmin`, `xmax`, `tau0`, `mu0` and `sigma0`.

$\longrightarrow$ You need the total number of collected events. For the unbinned ML this is just the length of the `data` vector (`len(data)`). For the binned methods loop over the `hist` array and sum all entries. `nbins=len(hist)` gives you the number of bins.

$\longrightarrow$ For the unbinned ML you need to loop over the `data` array, for each entry calculate the normalized p.d.f.'s (`gauss` & `decay`) and acumulate $\log L$ according to Eq. (25) of lecture 4 and using the combined S+B p.d.f.

$\longrightarrow$ For the binned methods you need to loop over the bins, the `hist` array (e.g. `for k in range(nbins)`). You need to get the prediction for the bin by integrating the normalized p.d.f., see Eq. (11) of lecture 5. Bin k is delimitted by `binsy[k]` and `binsy[k+1]`.

$\longrightarrow$ For the binned ML increment the $\log L$ using Eq. (13) of lecture 5 and the combined S+B p.d.f.

$\longrightarrow$ For the binned LS increment the χ^2 using Eq. (36) of lecture 5 and the combined S+B p.d.f.

$\longrightarrow$ NOTE: We are fitting just one free parameter, N_S (coded as `mus`). Make sure you properly define the combined S+B p.d.f. using N_S and the total number of collected events.

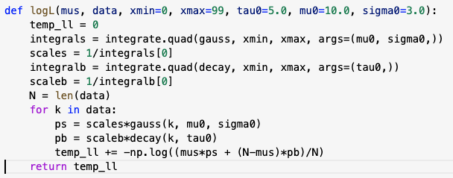
Hints part 2: estimate the error on N_s using $\log L$ or χ^2 profile

$\longrightarrow$ The code provides you with the `pl` & `ll_array` arrays which contain the estimated N_S and the corresponding value of either $\log L$ or χ^2, respectively.
$\longrightarrow$ Your task is to find values of N_S corresponding to $+1\sigma$ and -1σ about the fitted value (coded as `sigma_neg` & `sigma_pos`).
$\longrightarrow$ For the purpose, recall Eq. (7) and Eq. (32) of lecture 5.

Example solution

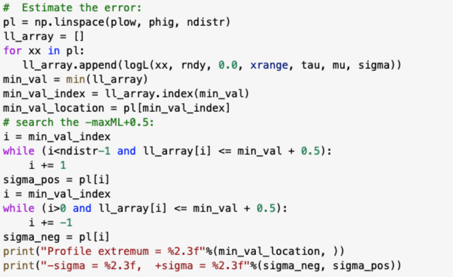
Unbinned Maximum Likelihood

$\log L$ function definition:



```
def logL(mus, data, xmin=0, xmax=99, tau0=5.0, mu0=10.0, sigma0=3.0):
    temp_ll = 0
    integrals = integrate.quad(gauss, xmin, xmax, args=(mu0, sigma0,))
    scales = 1/integrals[0]
    integralb = integrate.quad(decay, xmin, xmax, args=(tau0,))
    scaleb = 1/integralb[0]
    N = len(data)
    for k in data:
        ps = scales*gauss(k, mu0, sigma0)
        pb = scaleb*decay(k, tau0)
        temp_ll += -np.log((mus*ps + (N-mus)*pb)/N)
    return temp_ll
```

Uncertainties from the $\log L$ profile:



```
#  Estimate the error:
pl = np.linspace(plow, phig, ndistr)
ll_array = []
for xx in pl:
    ll_array.append(logL(xx, rndy, 0.0, xrange, tau, mu, sigma))
min_val = min(ll_array)
min_val_index = ll_array.index(min_val)
min_val_location = pl[min_val_index]
# search the -maxML+0.5:
i = min_val_index
while (i<ndistr-1 and ll_array[i] <= min_val + 0.5):
    i += 1
sigma_pos = pl[i]
i = min_val_index
while (i>0 and ll_array[i] <= min_val + 0.5):
    i += -1
sigma_neg = pl[i]
print("Profile extremum = %2.3f"%(min_val_location, ))
print("-sigma = %2.3f,  +sigma = %2.3f"%(sigma_neg, sigma_pos))
```

Results:

DS 1: $N_S = 21.52 + 7.82 - 7.47$

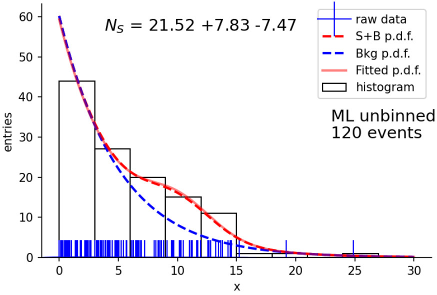
DS 2: $N_S = 29.99 + 7.93 - 7.69$

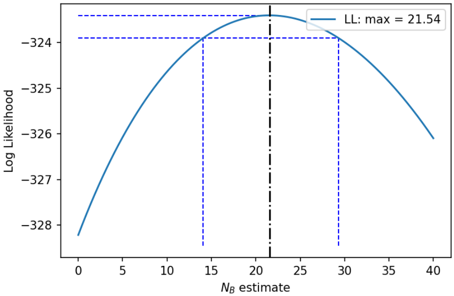
DS 3: $N_S = 18.29 + 7.69 - 7.32$

DS 4: $N_S = 21.09 + 7.94 - 7.55$

Example solution

Unbinned Maximum Likelihood



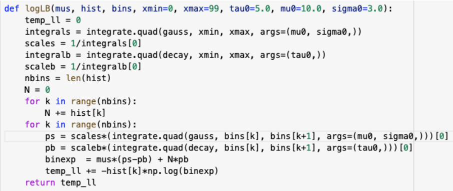


Fit uncertainty obtained from the $\log L$ profile at $\log L(N_S \pm \sigma_{N_S}) = \log L_{\max} - \frac{1}{2}$.

Example solution

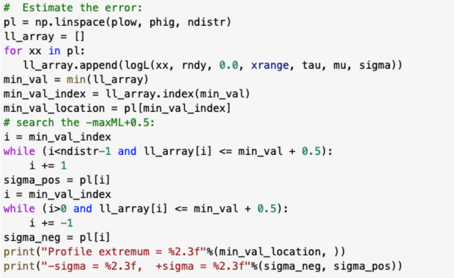
Binned Maximum Likelihood

$\log L$ function definition:



```python
def logLB(mus, hist, bins, xmin=0, xmax=99, tau0=5.0, mu0=10.0, sigma0=3.0):
    temp_ll = 0
    integrals = integrate.quad(gauss, xmin, xmax, args=(mu0, sigma0,))
    scales = 1/integrals[0]
    integralb = integrate.quad(decay, xmin, xmax, args=(tau0,))
    scaleb = 1/integralb[0]
    nbins = len(hist)
    N = 0
    for k in range(nbins):
        N += hist[k]
    for k in range(nbins):
        ps = scales*(integrate.quad(gauss, bins[k], bins[k+1], args=(mu0, sigma0,)))[0]
        pb = scaleb*(integrate.quad(decay, bins[k], bins[k+1], args=(tau0,)))[0]
        binexp  = mus*(ps-pb) + N*pb
        temp_ll += -hist[k]*np.log(binexp)
    return temp_ll
```

Uncertainties from the $\log L$ profile:



```python
#  Estimate the error:
pl = np.linspace(plow, phig, ndistr)
ll_array = []
for xx in pl:
    ll_array.append(logL(xx, rndy, 0.0, xrange, tau, mu, sigma))
min_val = min(ll_array)
min_val_index = ll_array.index(min_val)
min_val_location = pl[min_val_index]
# search the -maxML+0.5:
i = min_val_index
while (i<ndistr-1 and ll_array[i] <= min_val + 0.5):
    i += 1
sigma_pos = pl[i]
i = min_val_index
while (i>0 and ll_array[i] <= min_val + 0.5):
    i += -1
sigma_neg = pl[i]
print("Profile extremum = %2.3f"%(min_val_location, ))
print("-sigma = %2.3f,  +sigma = %2.3f"%(sigma_neg, sigma_pos))
```

Results:

DS 1: $N_S = 23.73 + 8.07 - 7.75$

DS 2: $N_S = 29.90 + 8.05 - 7.80$

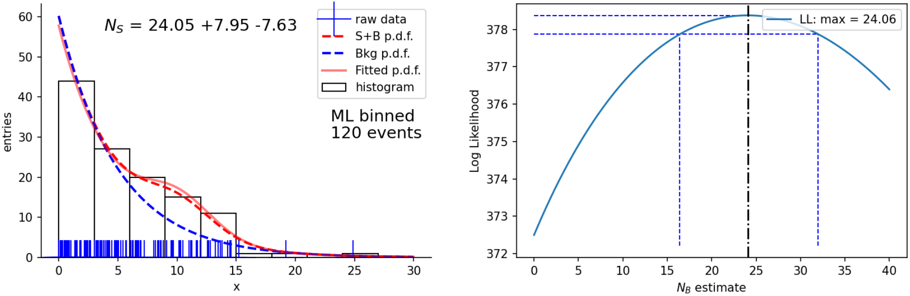
DS 3: $N_S = 17.45 + 8.02 - 7.60$

DS 4: $N_S = 22.76 + 8.19 - 7.86$

Example solution

Binned Maximum Likelihood

10 bins:

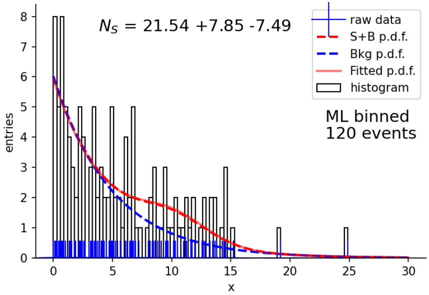


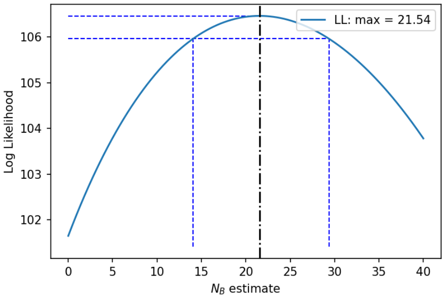
Fit uncertainty obtained from the $\log L$ profile at $\log L(N_S \pm \sigma_{N_S}) = \log L_{\max} - \frac{1}{2}$.

Example solution

Binned Maximum Likelihood

100 bins (for very fine binning result converges to the unbinned one!):



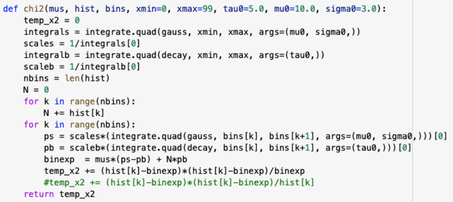


Fit uncertainty obtained from the $\log L$ profile at $\log L(N_S \pm \sigma_{N_S}) = \log L_{\max} - \frac{1}{2}$.

Example solution

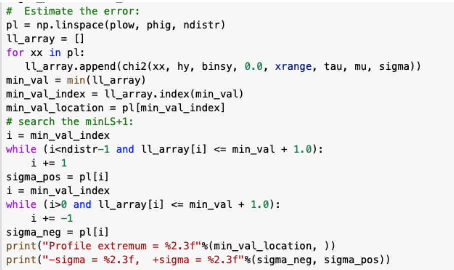
Binned Least Squares

χ^2 function definition:



```python
def chi2(mus, hist, bins, xmin=0, xmax=99, tau0=5.0, mu0=10.0, sigma0=3.0):
    temp_x2 = 0
    integrals = integrate.quad(gauss, xmin, xmax, args=(mu0, sigma0,))
    scales = 1/integrals[0]
    integralb = integrate.quad(decay, xmin, xmax, args=(tau0,))
    scaleb = 1/integralb[0]
    nbins = len(hist)
    N = 0
    for k in range(nbins):
        N += hist[k]
    for k in range(nbins):
        ps = scales*(integrate.quad(gauss, bins[k], bins[k+1], args=(mu0, sigma0,)))[0]
        pb = scaleb*(integrate.quad(decay, bins[k], bins[k+1], args=(tau0,)))[0]
        binexp  = mus*(ps-pb) + N*pb
        temp_x2 += (hist[k]-binexp)*(hist[k]-binexp)/binexp
        #temp_x2 += (hist[k]-binexp)*(hist[k]-binexp)/hist[k]
    return temp_x2
```

Uncertainties from the χ^2 profile:



```python
#  Estimate the error:
pl = np.linspace(plow, phig, ndistr)
ll_array = []
for xx in pl:
   ll_array.append(chi2(xx, hy, binsy, 0.0, xrange, tau, mu, sigma))
min_val = min(ll_array)
min_val_index = ll_array.index(min_val)
min_val_location = pl[min_val_index]
# search the minLS+1:
i = min_val_index
while (i<ndistr-1 and ll_array[i] <= min_val + 1.0):
    i += 1
sigma_pos = pl[i]
i = min_val_index
while (i>0 and ll_array[i] <= min_val + 1.0):
    i += -1
sigma_neg = pl[i]
print("Profile extremum = %2.3f"%(min_val_location, ))
print("-sigma = %2.3f,  +sigma = %2.3f"%(sigma_neg, sigma_pos))
```

Results:

DS 1: $N_S = 23.74 + 7.93 - 7.49$

DS 2: $N_S = 28.32 + 7.40 - 7.05$

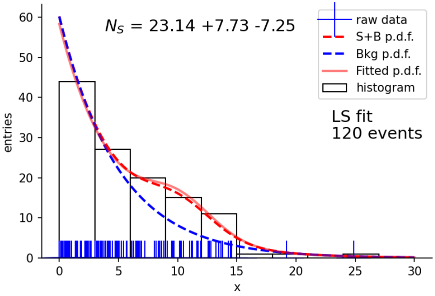
DS 3: $N_S = 17.11 + 7.95 - 7.42$

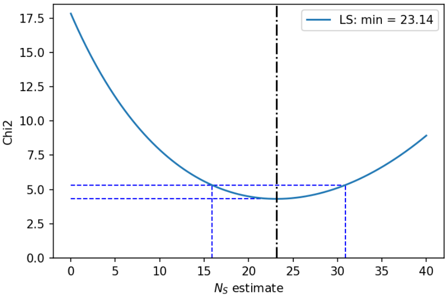
DS 4: $N_S = 23.04 + 8.19 - 7.67$

Example solution

Binned Least Squares

10 bins:



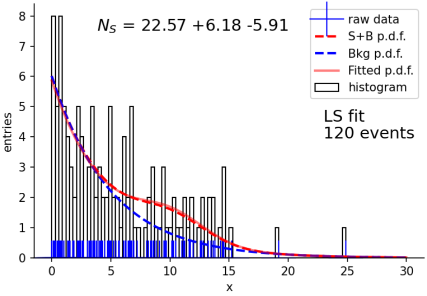


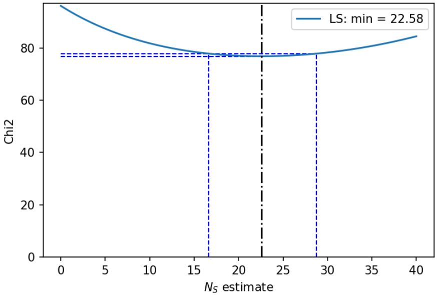
Fit uncertainty obtained from the χ^2 profile at $\chi^2(N_S \pm \sigma_{N_S}) = \chi^2_{\text{min}} + 1$.

Example solution

Binned Least Squares

100 bins (cannot reproduce the ML result exactly):
Small and empty bins make the uncertainty estimation less reliable.





Fit uncertainty obtained from the χ^2 profile at $\chi^2(N_S \pm \sigma_{N_S}) = \chi^2_{\min} + 1$.

Statistical error & confidence interval

Limits & signal significance

- So far, our discussion of uncertainties was mostly limited to (co)variance, or simply standard deviation.
- At some point we discussed quantiles of the Gaussian distribution to be more quantitative about the probability of a statistical outcome (*Lecture 3*),
- We also discussed statistical tests (*Lecture 4*),
- We also talked about goodness-of-fit and significance of the signal (*Lecture 4*),
- Finally we introduced the Pearson's χ^2 test (*Lecture 3*).

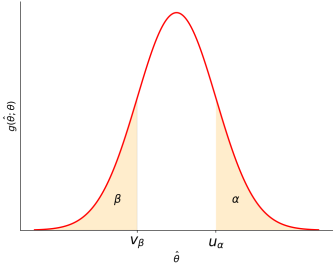
Now we shall be more detailed on confidence interval...
For a Gaussian distributed $g(\hat{\theta})$, the e.g. 68.3% confidence interval is the same as the interval covered by $\hat{\theta}_{\text{obs}} \pm \hat{\sigma}_{\hat{\theta}}$.

Statistical error & confidence interval

Classical confidence intervals

Assumptions:

- An estimator for a parameter θ is based on n observations $\hat{\theta}(x_1, ..., x_n)$.
- The value of the parameter is not known, but the p.d.f. of the estimator under the assumption of the true parameter value, $g(\hat{\theta};\theta)$ is known.
- From $g(\hat{\theta};\theta)$ one can determine u_α and v_β such that:



$$\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta) d\hat{\theta} = 1 - G(u_\alpha(\theta);\theta), \quad (1)$$

$$\beta = P(\hat{\theta} \leq v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta) d\hat{\theta} = G(v_\beta(\theta);\theta), \quad (2)$$

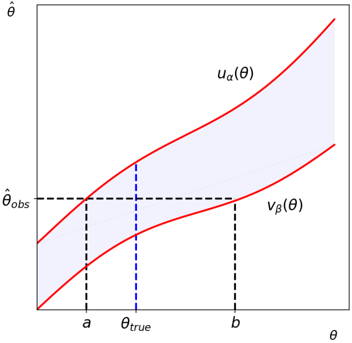
where G is the cumulative distribution of $g(\hat{\theta};\theta)$.

Statistical error & confidence belt

Classical confidence intervals

- A hypothetical shape of $u_\alpha(\theta)$ and $v_\beta(\theta)$ as a function of true value θ. $\longrightarrow$
- The region between the two curves is called the **confidence belt**.
- The probability for the estimator to be inside the belt (regardless of the value of θ) is:

$$P(v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta)) = 1 - \alpha - \beta \qquad (3)$$



It is also useful to define:

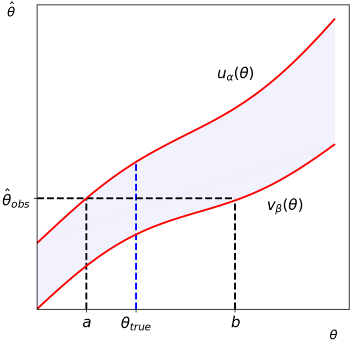
$$\begin{aligned} a(\hat{\theta}) &\equiv u_\alpha^{-1}(\hat{\theta}) \\ b(\hat{\theta}) &\equiv v_\beta^{-1}(\hat{\theta}) \end{aligned} \qquad (4)$$

$$\text{hence :} \qquad P(a(\hat{\theta}) \leq \theta \leq b(\hat{\theta})) = 1 - \alpha - \beta \qquad (5)$$

Confidence level & confidence interval

Classical confidence intervals

- The interval $[a(\hat{\theta}_{\mathrm{obs}}), b(\hat{\theta}_{\mathrm{obs}})]$ is called a **confidence interval** at a **confidence level** or **confidence probability** of $1 - \alpha - \beta$.
- Interpretation: If a similar experiment is repeated multiple times, the interval $[a, b]$ will cover θ_{true} with the probability $1 - \alpha - \beta$.
- One often chooses $\alpha = \beta = \gamma/2$, a so called **central confidence interval** with probability $1 - \gamma$.
- One also defines **one-sided confidence limit** such that a represents a lower limit on the parameter θ ($\theta \geq a$ with probability $1 - \alpha$). Similarly, b represents an upper limit ($\theta \leq b$ with probability $1 - \beta$).



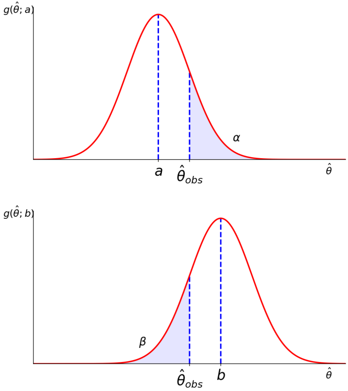
Confidence limits

Classical confidence intervals

Taking $\hat{\theta}_{\mathrm{obs}} = u_\alpha(a) = v_\beta(b)$, equations 1 and 2 become:

$$\alpha = \int_{\hat{\theta}_{\mathrm{obs}}}^{\infty} g(\hat{\theta}; a) d\hat{\theta} = 1 - G(\hat{\theta}_{\mathrm{obs}}; a), \quad (6)$$

$$\beta = \int_{-\infty}^{\hat{\theta}_{\mathrm{obs}}} g(\hat{\theta}; b) d\hat{\theta} = \quad G(\hat{\theta}_{\mathrm{obs}}; b). \quad (7)$$



This is closely connected to goodness-of-fit introduced in *lecture 4*. Here, P-value is set to α and a is a random variable that depends on data.

- The major difficulty of constructing confidence intervals is that the p.d.f. of the estimator $g(\hat{\theta}; \theta)$ (or $G(\hat{\theta}; \theta)$) has to be known.
- In practice, the p.d.f. is often Gaussian or approximately Gaussian which allows for easy construction of the intervals.

Gaussian distributed estimator

Confidence interval

If $\hat{\theta}$ is Gaussian distributed with mean θ and standard deviation $\sigma_{\hat{\theta}}$, we have:

$$\begin{aligned} G(\hat{\theta};\theta,\sigma_{\hat{\theta}}) = & \int_{-\infty}^{\hat{\theta}} \frac{1}{\sqrt{2\pi\sigma_{\hat{\theta}}^2}} \exp\left(\frac{-(\hat{\theta}'-\theta)^2}{2\sigma_{\hat{\theta}}^2}\right) d\hat{\theta}' = \\ & = \Phi\left(\frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}\right). \end{aligned} \tag{8}$$

This gives the solution to Eqs. 6 and 7:

$$\begin{aligned} a &= \hat{\theta}_{\text{obs}} - \sigma_{\hat{\theta}}\Phi^{-1}(1-\alpha), \\ b &= \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}}\Phi^{-1}(1-\beta). \end{aligned} \tag{9}$$

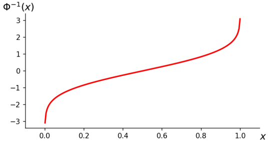
Here, Φ^{-1} is the inverse *error function*, i.e. the quantile of the standard Gaussian (*normal*).

single-sided intervals:

$\Phi^{-1}(1-\alpha)$	$1-\alpha$
1.0	0.8413
1.282	0.90
1.645	0.95
2.0	0.9772
2.326	0.99
3.0	0.9987

central intervals:

$\Phi^{-1}(1-\gamma/2)$	$1-\gamma$
1.0	0.6827
1.645	0.90
1.960	0.95
2.0	0.9544
2.576	0.99
3.0	0.9973



Poisson distribution

Confidence interval

Poisson estimator takes discrete (integer) values while the mean ν is a real positive:

$$f(n;\nu) = \frac{\nu^n}{n!}e^{-\nu}, \quad E[n] = \nu, \quad V[n] = \nu. \tag{10}$$

We can formulate the condition for confidence interval $[a, b]$:

$$\alpha = P(\hat{\nu} \geq \hat{\nu}_{\text{obs}}; a) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{a^n}{n!}e^{-a}, \tag{11}$$

$$\beta = P(\hat{\nu} \leq \hat{\nu}_{\text{obs}}; b) = \sum_{n=0}^{n_{\text{obs}}} \frac{b^n}{n!}e^{-b}. \tag{12}$$

n_{obs}	lower limit a $\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	upper limit b $\beta = 0.1$	$\beta = 0.05$	$\beta = 0.01$
0	–	–	–	2.30	3.00	4.61
1	0.105	0.051	0.010	3.89	4.74	6.64
2	0.532	0.355	0.149	5.32	6.30	8.41
3	1.10	0.818	0.436	6.68	7.75	10.04
4	1.74	1.37	0.832	7.99	9.15	11.60
5	2.43	1.97	1.28	9.27	10.51	13.11

The limits defined this way are conservative:

$$\begin{aligned} P(\nu \geq a) &\geq 1 - \alpha \\ P(\nu \leq b) &\geq 1 - \beta \\ P(a \leq \nu \leq b) &\geq 1 - \alpha - \beta \end{aligned} \tag{13}$$

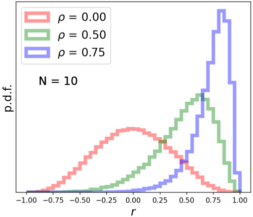
Note: A lower limit a cannot be determined for $n_{\text{obs}} = 0$. The upper one is defined by:

$$\beta = \sum_{n=0}^{0} \frac{b^n e^{-b}}{n!} = e^{-b} \Longrightarrow b = -\ln(\beta)$$

E.g.: $-\ln(0.05) \approx 3$, so if $n_{\text{obs}} = 0$, the 95% upper limit on the mean is 3.

Correlation coefficient

Confidence interval



Let's recall from *lecture 4* estimator for the covariance:

$$\widehat{V}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{n}{n-1}(\overline{xy}-\bar{x}\bar{y}) \qquad (14)$$

Hence, the (asymptotically unbiased) estimator for correlation coefficient:

$$r = \frac{\widehat{V}_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j-\bar{x})^2\sum_{k=1}^{n}(y_k-\bar{y})^2}} = \frac{\overline{xy}-\bar{x}\bar{y}}{\sqrt{(\overline{x^2}-\bar{x}^2)(\overline{y^2}-\bar{y}^2)}} \qquad (15)$$

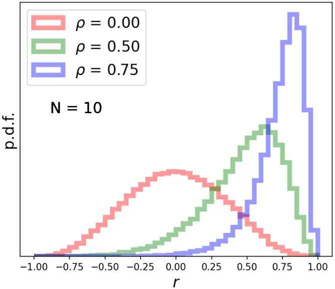
Issues when dealing with small statistics:

- The form of $g(r;\rho,n)$ is generally non-Gaussian and the solution (inversion of cumulative G) cannot be analytically found.
- The standard deviation and asymmetry depends on ρ.
- The estimator r is biased (although both V and s are not):

$$E[r] = \rho - \frac{\rho(1-\rho^2)}{2n} + \mathcal{O}(1/n^2), \qquad V[r] = \frac{1}{n}(1-\rho^2)^2 + \mathcal{O}(1/n^2).$$

Correlation coefficient

Confidence interval



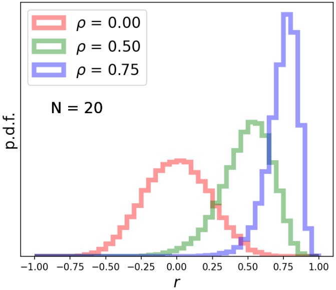
Let us assume an experiment results in $N = 10$ events yielding pairs on random variables. We have estimated the correlation to be $r = 0.75$. What is the significance of this result?

- The naive solution: $\hat{\sigma}_r = \frac{1-r^2}{\sqrt{n}} = 0.138$, so we get the 99% confidence interval of [0.39, 1.11]. "The probability of $\rho = 0$ is 6×10^{-6}. We have confirmed a positive correlation!"
  Really, have we?
- The correct reasoning: Assume the $\rho = 0$ hypothesis. What is the 99% confidence interval? $\hat{\sigma}_0 = 0.32$. The corresponding 99% confidence interval is [-0.81, 0.81]. Actually, the probability of obtaining $r = 0.75$ or larger is 1.8%! We have not demonstrated the correlation at the claimed confidence level !!!

Correlation coefficient

Confidence interval

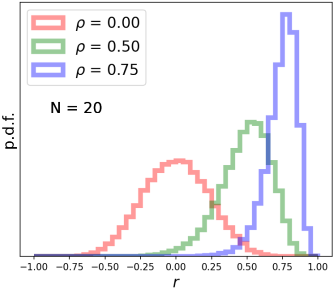
Let us assume an experiment results in $N = 20$ events yielding pairs on random variables. We have estimated the correlation to be $r = 0.5$. What is the significance of this result?



- The naive solution: $\hat{\sigma}_r = \frac{1-r^2}{\sqrt{n}} = 0.168$, so we get the 99% confidence interval of [0.07, 0.93]. “The probability of $\rho = 0$ is less than 0.5% . We have a strong evidence of a positive correlation!”
  Really, have we?
- The correct reasoning: Assume the $\rho = 0$ hypothesis. What is the 99% confidence interval? $\hat{\sigma}_0 = 0.223$. The corresponding 99% confidence interval is [-0.58, 0.58]. Actually, the probability of obtaining $r = 0.5$ or larger is 2.5%!
  We have not demonstrated the correlation at the claimed confidence level !!!

Correlation coefficient

Confidence interval



An approximate solution can be obtained using variable transformation.
It has been shown by Fisher that the p.d.f. of the statistic

$$z = \tanh^{-1} r = \frac{1}{2} \log \frac{1+r}{1-r}$$

approaches the Gaussian limit much more quickly as a function of n.

- The expectation value and variation are approximately:

$$E[z] \simeq \frac{1}{2} \log \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}, \qquad V[z] \simeq \frac{1}{n-3} \quad \text{(no z dependence).}$$

- The approximate solution: $z = 0.549$ and $\hat{\sigma}_z = 0.243$, so we get the 99% confidence interval of [-0.075, 1.174] for z.
- The inverse transformation gives the 99% confidence interval for r of [-0.075, 0.826].
- Or the probability of obtaining $r = 0.5$ or larger is 2.3% (in decent agreement with the exact calculation).

Confidence intervals using log-likelihood or χ^2

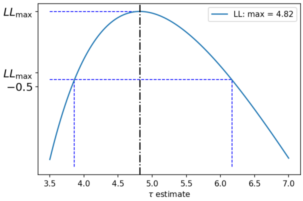
Even for non-Gaussian estimators, the interval can be determined approximately using profile of the $\log L$ or χ^2 functions. As discussed in *Lecture 5*, from Taylor expansion and assuming the RCF bound we get:

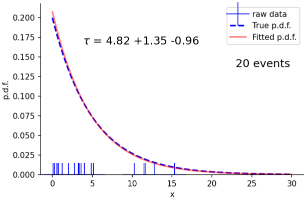
$$\log L(\theta^{+d}_{-c}) = \log L_{\max} - \frac{N^2}{2}, \quad \text{or} \quad \chi^2(\theta^{+d}_{-c}) = \chi^2_{\min} + N^2, \tag{16}$$

where the central confidence interval is given by $[a, b] = [\widehat{\theta} - c, \widehat{\theta} + d]$ and $N = \Phi^{-1}(1 - \gamma/2)$ is the quantile of the standard Gaussian (*normal dist.*) corresponding to the desired confidence level $1 - \gamma$.

Note: With the assumption of Gaussian errors one has $\log L = -\chi^2/2$.

Recall our example of $\hat{\tau} = \frac{1}{n}\sum_i t_i$. The asymmetry justifies to quote $\hat{\tau} = 4.82^{+1.35}_{-0.95}$ 95%C̃L, rather than $\hat{\tau} = 4.82 \pm 1.08$:





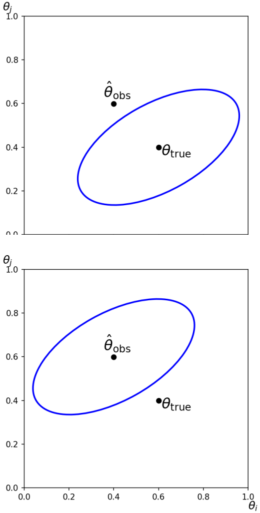
Multidimensional confidence intervals

In case of estimators of more than one parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, the confidence interval is replaced by the **confidence region**. In the large sample limit the joint p.d.f. becomes:

$$g(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left[-\frac{1}{2}Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\right], \qquad (17)$$

$$\text{with} \quad Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T V^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}).$$

The hyperellipsoids in the $\hat{\boldsymbol{\theta}}$-space delimit confidence regions. Due to the symmetry between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$, the confidence region looks the same no matter which of the two is regarded constant.



Multidimensional confidence intervals

- For sufficiently large samples (n-dimensional Gaussian) the quantity $Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ is distributed according to χ^2 with n DoF's. The **confidence region** with CL $1-\gamma$ is given by:

$$Q(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \leq Q_\gamma = F^{-1}(1-\gamma; n), \qquad \int_0^{Q_\gamma} f(z; n) dz = 1-\gamma. \tag{18}$$

  Q_γ is the quantile of order $1-\gamma$ of the χ^2 distribution.

- The confidence region boudaries can be constructed finding values of $\boldsymbol{\theta}$ satisfying:

$$\log L(\boldsymbol{\theta}) = \log L_{\max} - \frac{Q_\gamma}{2}. \tag{19}$$

	Q_γ				
$1-\gamma$	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.693	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Limits near a physical boundary

- It often happens that an estimator can attain values outside a physically allowed region (e.g. negative quantity of a sought for admixture).
- This is typical when the estimator results from subtracting two random variables: $\hat{\theta} = x - y$.
- If both x and y are Gaussian distributed, so is $\hat{\theta}$, with expectation $\theta = \mu_x - \mu_y$ and variance $\sigma^2_{\hat{\theta}} = \sigma^2_x + \sigma^2_y$.
- We can end up with not only the estimated value outside of the physical bound but even the upper limit may be outside.

  Imagine e.g. $m < -3$mg @ 95% CL – not a very useful result ☹).

Limits near a physical boundary

Imagine e.g. our result must not be negative.
There are three most common solutions to this problem:

1. Classical: take as is despite disturbing interpretation (mathematically correct).
2. Shift the observation to the boundary of allowed interval:

$$\theta_{\mathrm{up}} = \max(\hat{\theta}_{\mathrm{obs}}, 0) + \sigma_{\hat{\theta}} \Phi^{-1}(1-\beta). \tag{20}$$

   Overconservative, as $1-\beta$ probability no longer applies. On the other hand, limit is never smaller than the experimental resolution.

3. Use the Bayesian posterior p.d.f.:

$$p(\theta|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\theta)\pi(\theta)}{\int L(\boldsymbol{x}|\theta')\pi(\theta')d\theta'}, \quad 1-\beta = \int_{-\infty}^{\theta_{\mathrm{up}}} p(\theta|\boldsymbol{x})d\theta = \frac{\int_{-\infty}^{\theta_{\mathrm{up}}} L(\boldsymbol{x}|\theta)\pi(\theta)d\theta}{\int_{-\infty}^{\infty} L(\boldsymbol{x}|\theta)\pi(\theta)d\theta} \tag{21}$$

   What remains undefined is the choice of prior $\pi(\theta)$. The simplest choice is a flat prior:

$$\pi(\theta) = \begin{cases} 0 & \theta < 0 \\ 1 & \theta \geq 0 \end{cases}$$

Limit on Poisson signal over background

$n = n_s + n_b$. This is a special example of the previous case.

$$f(n;\nu_s,\nu_b) = \frac{(\nu_s+\nu_b)^n}{n!} e^{-(\nu_s+\nu_b)}, \tag{22}$$

with the unbiased ML estimator for ν_s:

$$\hat{\nu_s} = n - \nu_b, \qquad E[n] = \nu_s + \nu_b. \tag{23}$$

- $\hat{\nu}_s$ and its variance must be reported if results of multiple experiments are to be combined.
- Classical limit is not recommended when ν_b is large compared to ν_s. For seting limits, the Bayesian approach with flat prior is usually used:

$$1-\beta = \frac{\int_0^{\nu_s^{\mathrm{up}}} L(n_{\mathrm{obs}}|\nu_s)d\nu_s}{\int_0^{\infty} L(n_{\mathrm{obs}}|\nu_s)d\nu_s} = \frac{\int_0^{\nu_s^{\mathrm{up}}} (\nu_s+\nu_b)^n_{\mathrm{obs}} e^{-(\nu_s+\nu_b)} d\nu_s}{\int_0^{\infty} (\nu_s+\nu_b)^n_{\mathrm{obs}} e^{-(\nu_s+\nu_b)} d\nu_s} \tag{24}$$

which, for no background is equivalent to Eq. 12 and asymptotically decrease to $-\ln\beta$ for growing ν_b.

Limit from the functional shape

x is a random variable with different p.d.f.'s $f_s(x)$ and $f_b(x)$, respectivly.

$$f(x;\nu_s,\nu_b) = \frac{\nu_s f_s(x) + \nu_b f_b(x)}{\nu_s + \nu_b}. \tag{25}$$

We can formulate the fit in two different ways:

1. The extended ML using:

$$\begin{aligned} L(\nu_s) &= \frac{(\nu_s+\nu_b)^n}{n!} e^{-(\nu_s+\nu_b)} \prod_{i=1}^{n} \frac{\nu_s f_s(x_i) + \nu_b f_b(x_i)}{\nu_s+\nu_b} \\ \Longrightarrow \quad \log L(\nu_s) &= -\nu_s + \sum_{i=1}^{n} \ln\left(\nu_s f_s(x_i) + \nu_b f_b(x_i)\right) \end{aligned} \tag{26}$$

2. The normal ML using:

$$L(\nu_s) = \prod_{i=1}^{n} \frac{\nu_s f_s(x_i) + \nu_b f_b(x_i)}{\nu_s+\nu_b} \quad \Longrightarrow \quad \log L(\nu_s) = \sum_{i=1}^{n} \ln\left(\frac{\nu_s f_s(x_i) + \nu_b f_b(x_i)}{\nu_s+\nu_b}\right) \tag{27}$$

NOTE: The latter was used in our homework exercise!

Confidence intervals with binned data and systematic uncertainties

Consider a typical situation when experiment results in a histogram $\mathbf{n} = (n_1, ..., n_N)$, where contents of a bin depends on existence of the sought for signal (with *signal strength* μ) and additionally on a set of tunable experimental or theoretical *nuisance parameters* $\boldsymbol{\theta}$:

$$E[n_1] = \mu s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}) \tag{28}$$

The Likelihood function is given by:

$$L(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)}. \tag{29}$$

To test a hypothesized value of μ we consider the **profile likelihood ratio**:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}, \quad \text{where} \quad \begin{array}{l} \hat{\hat{\boldsymbol{\theta}}} \text{ maximizes } L \text{ for a given } \mu \\ \hat{\mu}, \hat{\boldsymbol{\theta}} \text{ realise the absolute maximum of } L \end{array} \tag{30}$$

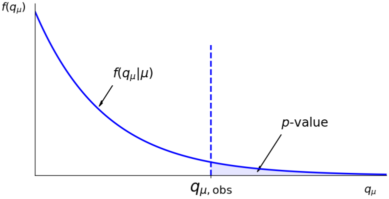
Profile Likelihood Ratio

The maximum of profile likelihood ratio (PLR) should be a near-optimal estimator for μ with nuisance parameters $\boldsymbol{\theta}$.
A monotonic function of PLR provides an equally good test statistic:

$$q_\mu = -2\log\lambda(\mu). \tag{31}$$

Large q_μ means increasing incompatibility between the data and hypothesis (μ), therefore p-value for an observed $q_{\mu,\mathrm{obs}}$ is:

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\mathrm{d}q_\mu. \tag{32}$$



NOTE: Significance: $Z = \Phi^{-1}(1-p)$, where Φ is the cumulative of *normal* dist.

Profile Likelihood Ratio

Wald approximation

In practice, we need a decent approximation of the q_μ p.d.f. in order to calculate quantiles and calculate p-values.
The desired distribution $f(q_\mu \mid \mu')$ can be found using a result due to A. Wald (1943), who showed that for the case of a single parameter of interest:

$$q_\mu = -2\log\lambda(\mu) = \frac{(\hat{\mu}-\mu)^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N}), \qquad (33)$$

with $\hat{\mu} \sim \mathrm{Gaussian}(\mu', \sigma)$, i.e., $E[\hat{\mu}] = \mu'$. Here, σ can be estimated e.g. from the fit Hessian thanks to RCF relation:

$$\left(\widehat{V^{-1}}\right)_{i,j} = -\frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \qquad (34)$$

The p.d.f. is a noncentral χ^2 distribution (noncentrality param.: $\Lambda = \frac{(\mu-\mu')^2}{\sigma^2}$):

$$f(q_\mu|\mu') = \frac{1}{2\sqrt{q_\mu}}\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}(\sqrt{q_\mu}+\sqrt{\Lambda})^2\right) + \exp\left(-\frac{1}{2}(\sqrt{q_\mu}-\sqrt{\Lambda})^2\right)\right]. \qquad (35)$$

N.b.: For $\mu = \mu'$, q_μ approaches a χ^2_1 distribution, as shown by S. Wilks (1938).

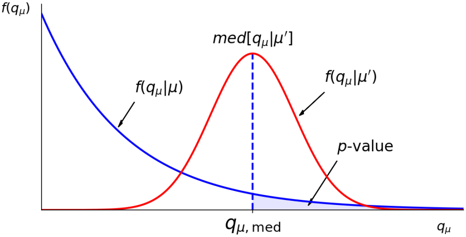
Profile Likelihood Ratio

The sensitivity

If we want to assess sensitivity of our experiment, we need to find value of μ' which gives (on average) required p-value for our null hypothesis μ.

The **Asimov** data set can be used to assess the median value of q_μ statistic. It is defined by all bins content equal to their expectation values:

$$\hat{\mu} = \mu', \qquad \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}. \tag{36}$$



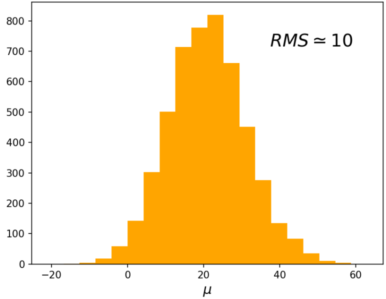
NOTE: What we are trying to estimate here is how incompatible is our assumed signal hypothesis μ' with the *null hypothesis* (μ). We settle on the value witch is equal to the required p-value (or equivalently significance).

Profile Likelihood Ratio

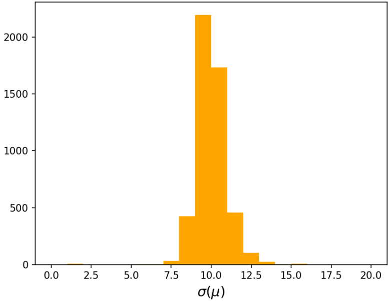
Base example (signal significance)

Let us revisit the fit of Gaussian signal over exponential background (last HW). Let us assume our *parameter of interest* (PoI) is still the signal yield, while the background yield and its shape (parameterised by τ) are *nuisance parameters*. Let us check the asymptotic formula for q_μ statistic p.d.f. and extract the signal significance. We run 5000 toys...

fitted signal strength μ



fitted uncertainty on μ, $\sigma(\mu)$

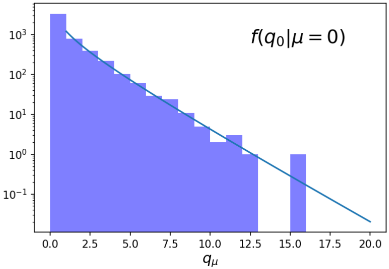


Profile Likelihood Ratio

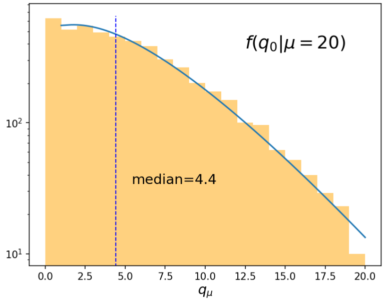
Base example (signal significance)

Let us revisit the fit of Gaussian signal over exponential background (last HW). Let us assume our *parameter of interest* (PoI) is still the signal yield, while the background yield and its shape (parameterised by τ) are *nuisance parameters*. Let us check the asymptotic formula for q_μ statistic p.d.f. and extract the signal significance. We run 5000 toys...

$f(q_0|\mu = 0)$, $\Lambda = 0$



$f(q_0|\mu = 20)$, $\sigma \simeq 10$, $\Lambda \simeq 4$



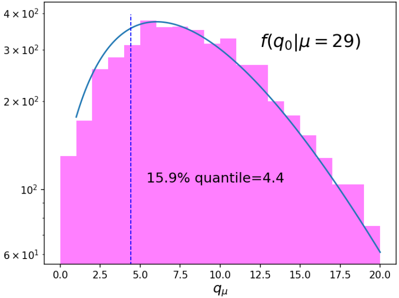
Asimov: q_μ=4.4, p-value=0.036 Data: q_μ=8.1, p-value=0.004

Profile Likelihood Ratio

Discovery example (signal significance)

- We are not only interested in the median. We always want to know how much statistical variation to expect from a real data set. But we have the full $f(q_0|\mu)$. We can get any desired quantiles.
- In particular, values of μ for which median-1σ and median$+1\sigma$ happen at the Asimov (or fitted) q_μ define the $\pm 1\sigma$ uncertainty band.

$f(q_0|\mu = 29)$, $\sigma \simeq 10$, $\Lambda \simeq 9$



- The Profile Likelihood Ratio fit can be extended to several fitting areas, including e.g. control regions constraining certain nuisance parameters, etc. Likelihood is a product of individual ones.
- Gaussian constraints on the nuisance parameters are typically present in the PLR as well.

Profile Likelihood Ratio

Discovery vs Upper Limit

The so far discussed statistic allows for both up and down fluctuations resulting in $\hat{\mu}$ going beyond its physically meaningful boubnds.
This is why modified statistics are commonly used:

DISCOVERY

Try to reject background-only ($\mu = 0$) hypothesis using:

$$q_0 = \begin{cases} -2\log\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases} \tag{37}$$

UPPER LIMIT

For purposes of setting an upper limit on μ use:

$$q_\mu = \begin{cases} -2\log\lambda(0) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} \tag{38}$$

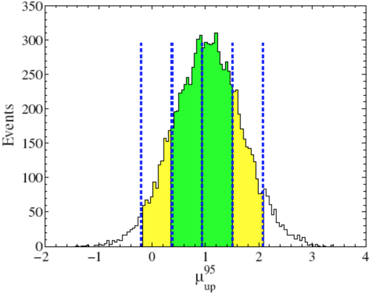
NOTE: Essentially, we are probing only single sided departures from the *null hypothesis*. P.d.f.'s of concerned statistics are slightly modified but still analytically defined and allow for making numerical predictions.

Profile Likelihood Ratio

Limit on the observation - example

Distribution of upper limit on μ

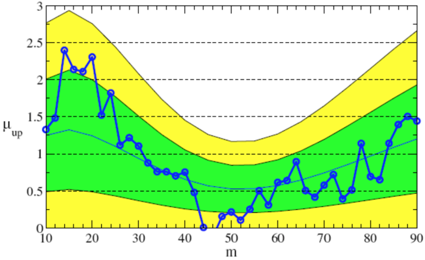
$\pm1\sigma$ (green) and $\pm2\sigma$ (yellow) bands from MC; Vertical lines from asymptotic formulae



credit: G. Cowan

Limit on μ versus peak position

This is the famous “brasilian plot”



This and much, much more on PLR and asymptotic formulae hypothesis tests can be found here.

Thank you

Back-up

Profile Likelihood Ratio

Wald approximation

$$q_\mu = -2\log\lambda(\mu) = \frac{(\hat{\mu}-\mu)^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N}), \tag{39}$$

The p.d.f. is a noncentral χ^2 distribution (noncentrality param.: $\Lambda = \frac{(\mu-\mu')^2}{\sigma^2}$):

$$f(q_\mu|\mu') = \frac{1}{2\sqrt{q_\mu}}\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}(\sqrt{q_\mu}+\sqrt{\Lambda})^2\right) + \exp\left(-\frac{1}{2}(\sqrt{q_\mu}-\sqrt{\Lambda})^2\right)\right]. \tag{40}$$

N.b.: For $\mu = \mu'$, q_μ approaches a χ^2_1 distribution, as shown by S. Wilks (1938).

$$f(q_\mu|\mu) = \frac{1}{\sqrt{q_\mu}}\frac{1}{\sqrt{2\pi}}\exp^{-q_\mu/2} \tag{41}$$

The cumulative distribution of q_μ assuming μ' is:

$$F(q_\mu|\mu') = \Phi\left(\sqrt{q_\mu}+\sqrt{\Lambda}\right) + \Phi\left(\sqrt{q_\mu}-\sqrt{\Lambda}\right) - 1. \tag{42}$$

Profile Likelihood Ratio

Discovery

Try to reject background-only ($\mu = 0$) hypothesis using:

$$q_0 = \begin{cases} -2\log\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases} \tag{43}$$

Assuming the validity of the Wald approximation, one gets:

$$f(q_0|\mu') = \left(1 - \Phi\frac{\mu'}{\sigma}\right)\delta(q_0) + \frac{1}{2\sqrt{q_0}}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]. \tag{44}$$

The corresponding cumulative distribution is found to be:

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right) \tag{45}$$

The p-value of the hypothesis $\mu = 0$, p_0, is obtained from these distributions by using $\mu' = 0$. For the significance one finds the simple formula:

$$Z_0 = \Phi^{-1}(1 - p_0) = \sqrt{q_0} \tag{46}$$

Profile Likelihood Ratio

Upper limit

For purposes of setting an upper limit on μ use:

$$q_\mu = \begin{cases} -2\log\lambda(0) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} \tag{47}$$

Assuming the validity of the Wald approximation, one gets:

$$f(q_\mu|\mu') = \Phi\left(\frac{\mu' - \mu}{\sigma}\right)\delta(q_\mu) + \frac{1}{2\sqrt{q_\mu}}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\sqrt{q_\mu} - \frac{\mu' - \mu}{\sigma}\right)^2\right]. \tag{48}$$

The corresponding cumulative distribution is found to be:

$$F(q_\mu|\mu') = \Phi\left(\sqrt{q_\mu} - \frac{\mu' - \mu}{\sigma}\right) \tag{49}$$

The p-value of the hypothesis μ, p_μ, is obtained from these distributions by using $\mu' = 0$. For the significance one finds the simple formula:

$$Z_\mu = \Phi^{-1}(1 - p_\mu) = \sqrt{q_\mu} \tag{50}$$