Statistics and Data Analysis (HEP at LHC)

Few problems for your homework

Slides extracted from N. Berger lectures at CERN Summer School 2019

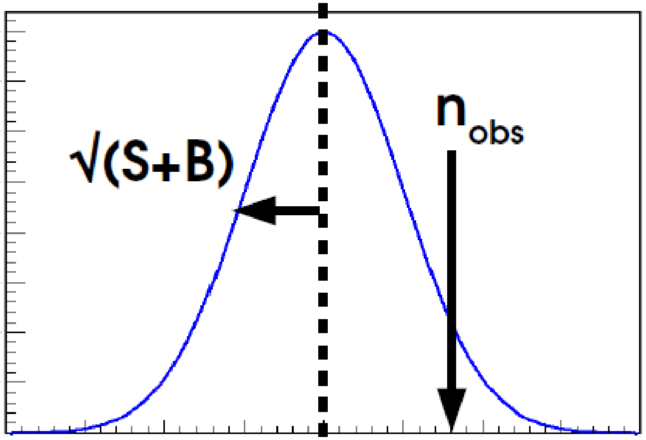
Prof. dr hab. Elżbieta Richter-Wąs

Homework 1: Gaussian Counting

Count number of events n in data

→ assume n large enough so process is Gaussian

→ assume B is known, measure S



Likelihood : $$L(S;n_{obs}) = e^{-\frac{1}{2}\left(\frac{n_{obs}-(S+B)}{\sqrt{S+B}}\right)^2}$$

→ Find the best-fit value (MLE) $\hat{S}$ for the signal (can use λ = -2 log L instead of L for simplicity)

→ Find the expression of q_0 for $\hat{S} > 0$.

→ Find the expression for the significance

$$Z = \frac{\hat{S}}{\sqrt{B}}$$

$\sqrt{B}$ is the uncertainty on S (remember $\sqrt{n}$?) so this gives "how many times its uncertainty" $\hat{S}$ is from 0 ⇒ Natural expression.

→ Only valid in Gaussian regime!

Homework 2: Poisson Counting

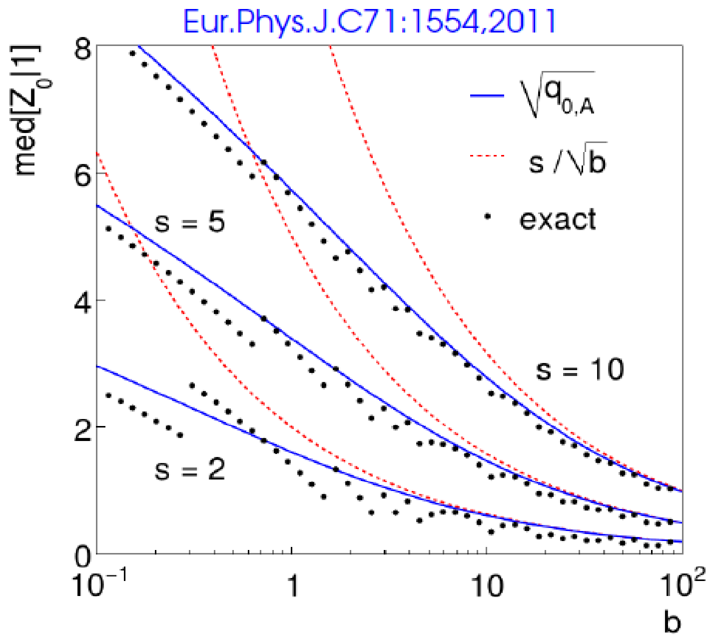
Same problem but now ***not*** assuming Gaussian behavior:

$$L(S;n) = e^{-(S+B)}(S+B)^n$$

(Can remove the n! constant since we're only dealing with L ratios)

→ As before, compute $\hat{S}$, and q_0

→ Compute $Z = \sqrt{q_0}$, assuming asymptotic behavior (weaker form of the Gaussian assumption)



Solution:

$$Z = \sqrt{2\left[(\hat{S}+B)\log\left(1+\frac{\hat{S}}{B}\right) - \hat{S}\right]}$$

Exact result can be obtained using pseudo-experiments → close to $\sqrt{q_0}$ result

Asymptotic formulas justified by Gaussian regime, but remain valid even for small values of S+B (down to 5 events!)

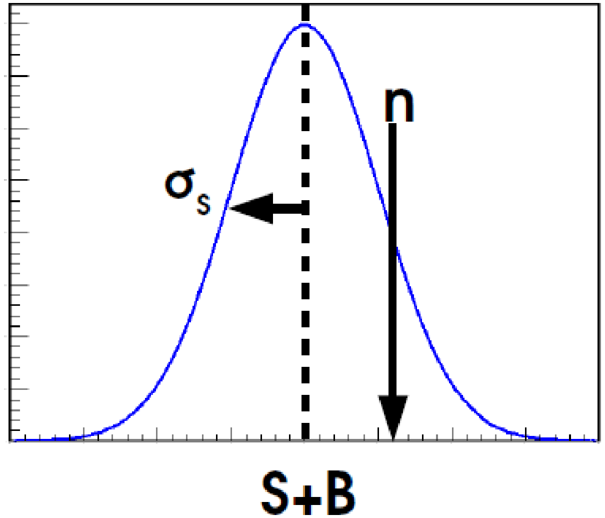
35

Homework 3: Gaussian example

Usual Gaussian counting example with known B:

$$L(S;n) = e^{-\frac{1}{2}\left(\frac{n-(S+B)}{\sigma_S}\right)^2}$$

$\sigma_S \sim \sqrt{B}$ for small S



Reminder: Significance: $Z = \hat{S}/\sigma_S$

→ Compute q_{S0}

→ Compute the 95% CL upper limit on S, S_{up}, by solving $q_{S0} = 2.70$.

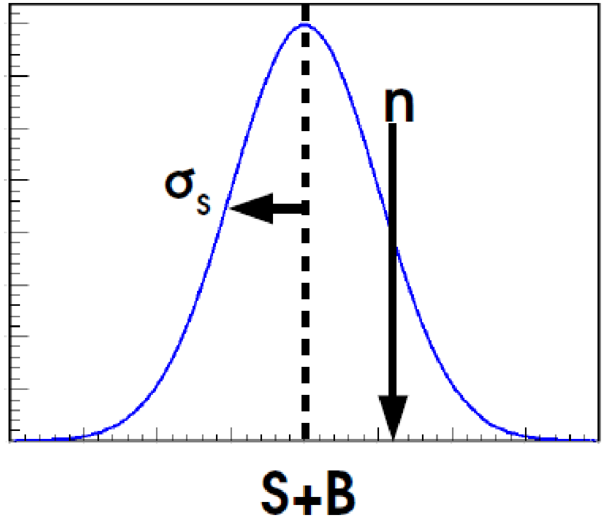
Solution: $S_{up} = \hat{S} + 1.64\,\sigma_S$ at 95 % CL

Homework 4: CL_s Gaussian Case

Usual Gaussian counting example with known B:

$$L(S;n) = e^{-\frac{1}{2}\left(\frac{n-(S+B)}{\sigma_S}\right)^2}$$

$\sigma_S \sim \sqrt{B}$ for small S



Reminder

CL_{s+b} limit: $S_{up} = \hat{S} + 1.64\,\sigma_S$ at 95 % CL

CL_s upper limit :

→ Compute p_{S0} (same as for CLs+b)

→ Compute p_B (hard!)

Solution: $S_{up} = \hat{S} + \left[\Phi^{-1}\left(1 - 0.05\,\Phi\left(\hat{S}/\sigma_S\right)\right)\right]\sigma_S$ at 95% CL

for $\hat{S} \sim 0$, $S_{up} = \hat{S} + 1.96\,\sigma_S$ at 95% CL

Homework 5: CL_S Rule of Thumb for $n_{obs} = 0$

Same exercise, for the Poisson case with $n_{obs} = 0$. Perform an exact computation of the 95% CLs upper limit based on the definition of the p-value:
p-value : *sum probabilities of cases at least as extreme as the data*

Hint: for $n_{obs}=0$, there are no "more extreme" cases (cannot have n<0 !), so
p_{S0} = Poisson(n=0 | S_0+B) and 1 - p_B = Poisson(n=0 | B)

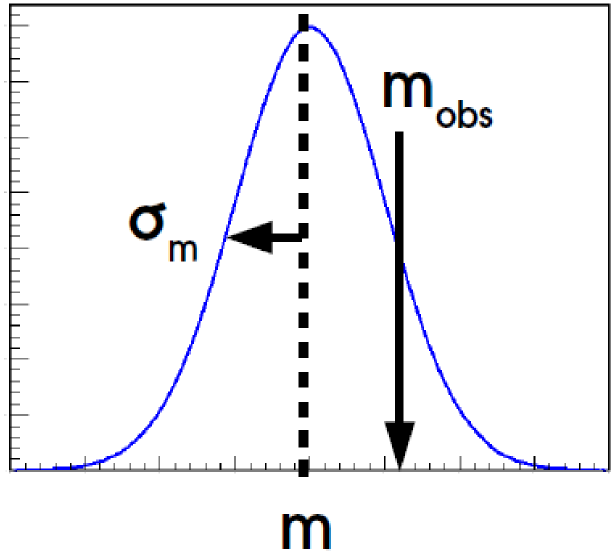
Solution: $S_{up}(n_{obs}=0) = \log(20) = 2.996 \approx 3$

⇒ **Rule of thumb: when $n_{obs} = 0$, the 95% CL_s limit is 3 events (for any B)**

Homework 6: Likelihood Intervals Gaussian case

Consider a parameter m (e.g. Higgs boson mass) whose measurement is Gaussian with known width σ_m, and we measure m_{obs}:

$$L(m; m_{obs}) = e^{-\frac{1}{2}\left(\frac{m - m_{obs}}{\sigma_m}\right)^2}$$



→ Compute the best-fit value (MLE) $\hat{m}$

→ Compute t_m

→ Compute the 1-σ (Z=1, ~68% CL) interval on m

Solution: $m = m_{obs} \pm \sigma_m$

→ Not really a surprise – the method works as expected on this simple case

→ General method can be applied in the same way to more complex cases

Homework 7: Gaussian profiling

Counting experiment with background uncertainty: **n = S + B** :

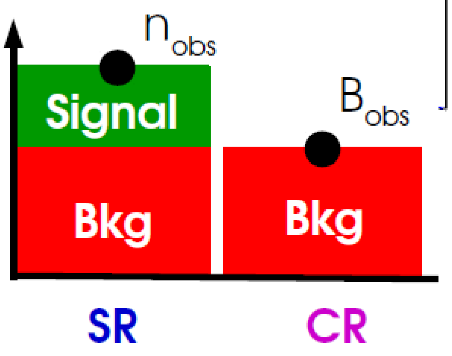
→ **Signal region (SR)**: n_{obs} ~ G(S + B, σ_{stat})

→ **Control region (CR)**: B_{obs} ~ G(B, σ_{bkg})

$$L(S,B) = G(n_{obs}; S+B, \sigma_{stat})\, G(B_{obs}; B, \sigma_{bkg})$$

Recall: Signal region only (fixed B):

$$t_S = \left(\frac{S - n_{obs}}{\sigma_{stat}}\right)^2 \qquad S = (n_{obs} - B) \pm \sigma_{stat}$$



→ Compute the best-fit (MLEs) for S and B

→ Show that the conditional MLE for B is

$$\hat{\hat{B}}(S) = B_{obs} + \frac{\sigma_{bkg}^2}{\sigma_{stat}^2 + \sigma_{bkg}^2}(\hat{S} - S)$$

→ Compute the profile likelihood t_S

→ Compute the 1σ confidence interval on S

$$S = (n_{obs} - B_{obs}) \pm \sqrt{\sigma_{stat}^2 + \sigma_{bkg}^2} \qquad \sigma_S = \sqrt{\sigma_{stat}^2 + \sigma_{bkg}^2}$$

Stat uncertainty (on n) and systematic (on B) add in quadrature

Homework 8: Bayesian methods and CLs

Gaussian counting problem with systematic on background: $n = S + B + \sigma_{syst}\theta$

$$P(n; S, \theta) = G(n; S + B + \sigma_{syst}\theta, \sigma_{stat})\, G(\theta_{obs} = 0; \theta, 1)$$

→ **What is the 95% CL upper limit on S, given a measurement n_{obs} ?**

1. CLs computation:

- Use the result of Homework 7 to compute the PLR for S
- Use the result of Homework 6 to compute the CLs upper limit

2. Bayesian computation:

- Integrate P(n; S, θ) over θ to get the marginalized P(n | S)
- Use Bayes' theorem to compute P(S | n) ∝ P(n | S) P(S), with P(S) a constant prior over S>0.
- Find the 95% CL limit by solving $\int_{S_{up}}^{\infty} P(S|n)\, dS = 5\,\%$

Solution:
In both cases

$$S_{up}^{CL_s} = n - B + \left[\Phi^{-1}\left(1 - 0.05\,\Phi\left(\frac{n-B}{\sqrt{\sigma_{stat}^2 + \sigma_{syst}^2}}\right)\right)\right]\sqrt{\sigma_{stat}^2 + \sigma_{syst}^2}$$