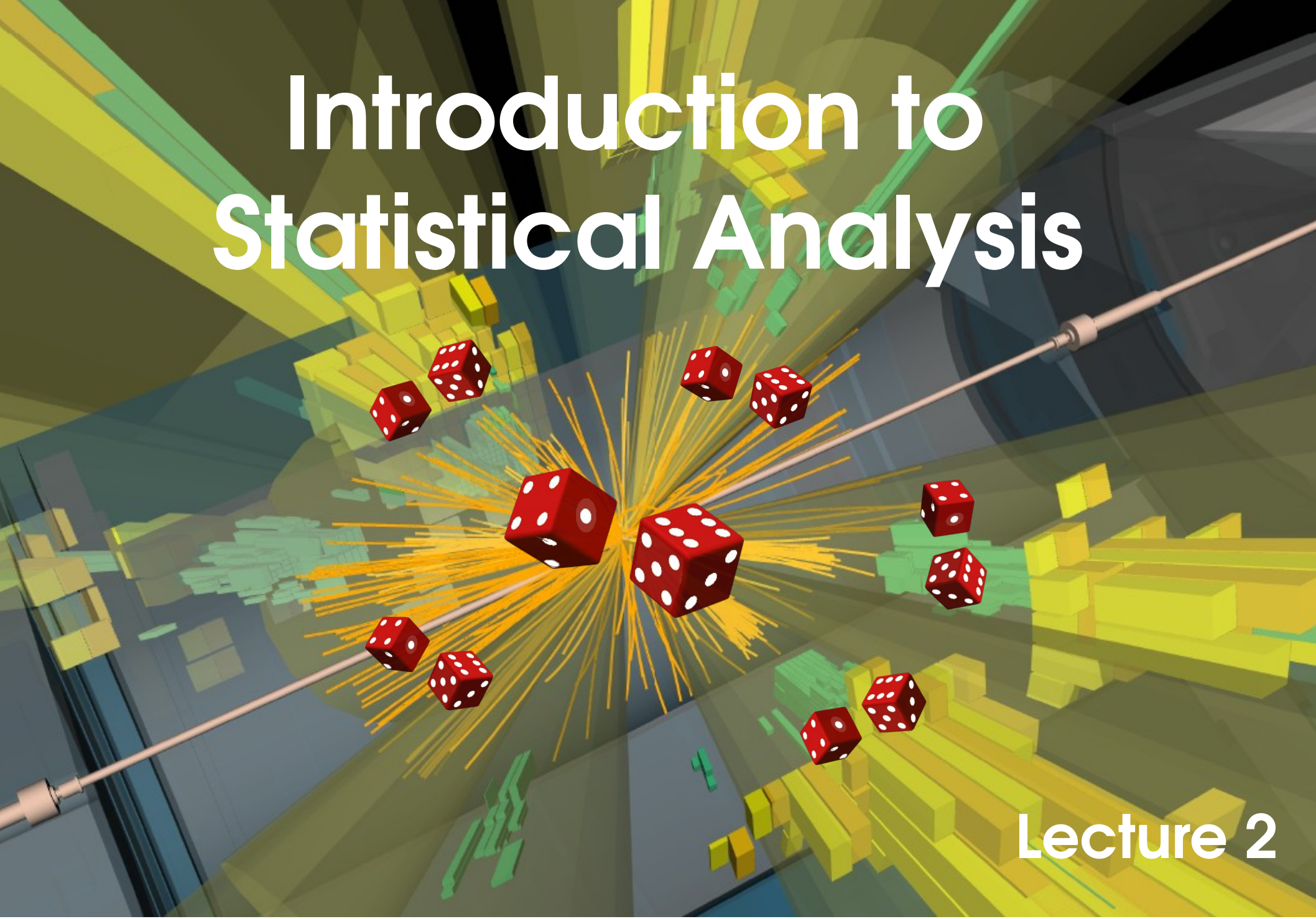
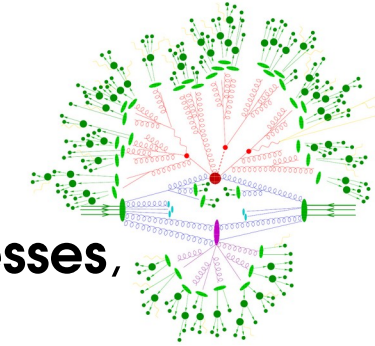
Introduction to Statistical Analysis



Lecture 2

Nicolas Berger (LAPP Annecy)

Physics Modeling : Summary of Lecture 1



Physics measurement data are produced through **random processes**,
Need to be described using a statistical model:

Description	Observable	Likelihood
Counting	n	**Poisson** $P(n;S,B) = e^{-(S+B)}\frac{(S+B)^n}{n!}$
Binned shape analysis	n_i, $i=1..N_{bins}$	**Poisson product** $P(n_i;S,B) = \prod_{i=1}^{N_{bins}} e^{-(S f_i^{sig} + B f_i^{bkg})}\frac{(S f_i^{sig} + B f_i^{bkg})^{n_i}}{n_i!}$
Unbinned shape analysis	m_i, $i=1..n_{evts}$	**Extended Unbinned Likelihood** $P(m_i;S,B) = \frac{e^{-(S+B)}}{n_{evts}!}\prod_{i=1}^{n_{evts}} S\,P_{sig}(m_i)+B\,P_{bkg}(m_i)$

Model can include multiple **categories**, each with a separate description
Includes **parameters of interest** (POIs) but also **nuisance parameters** (NPs)
Next step: use the model to obtain information on the POIs

Model Parameters

Model typically includes:

- **Parameters of interest** (POIs) : what we want to measure

  → **S, σ×B, m_W**, ...

- **Nuisance parameters** (NPs) : other parameters needed to define the model

  → **B**

  → For binned data, f^{sig}_i , f^{bkg}_i

  → For unbinned data, parameters needed

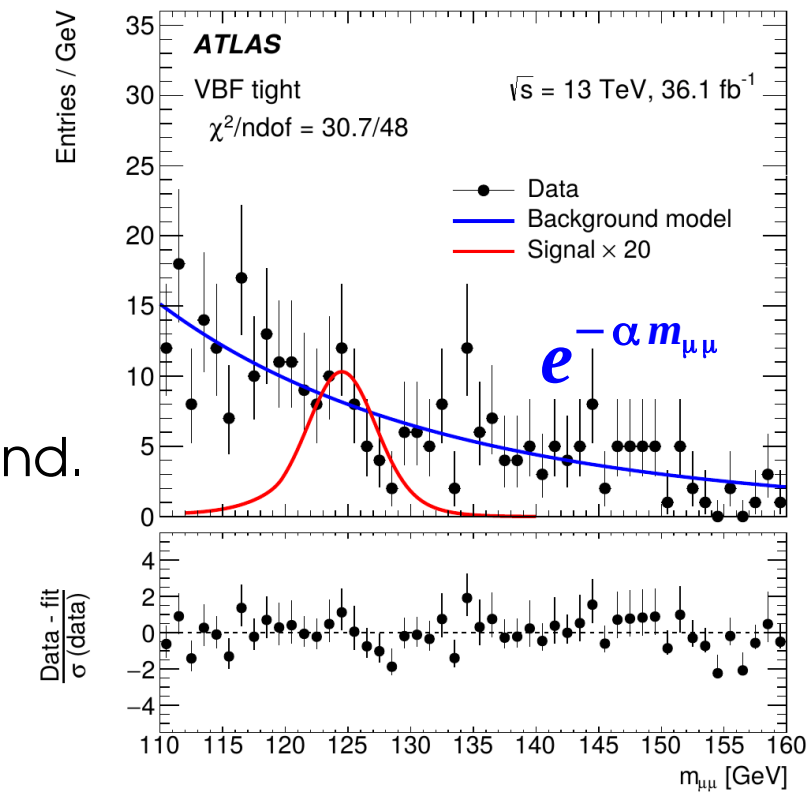
  to define P_{bkg}

  e.g. exponential slope α of H→μμ background.

NPs must be either

→ **known a priori** (possibly within systematics) or

→ **constrained by the data** (e.g. in sidebands)



Phys. Rev. Lett. 119 (2017) 051802

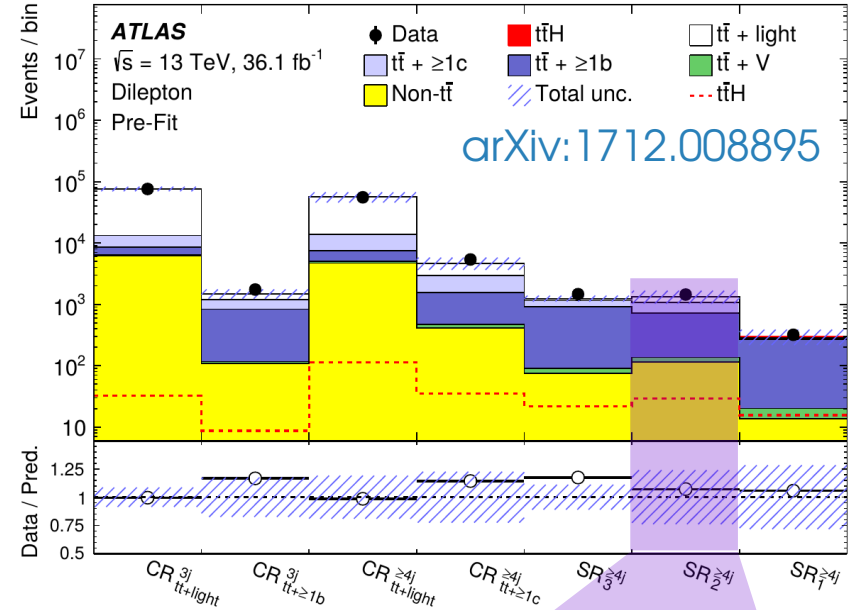
Categories

Multiple analysis regions often used:

e.g. for different decay modes, etc.

→ Useful to model these separately if better sensitivity in some regions (avoids dilution)

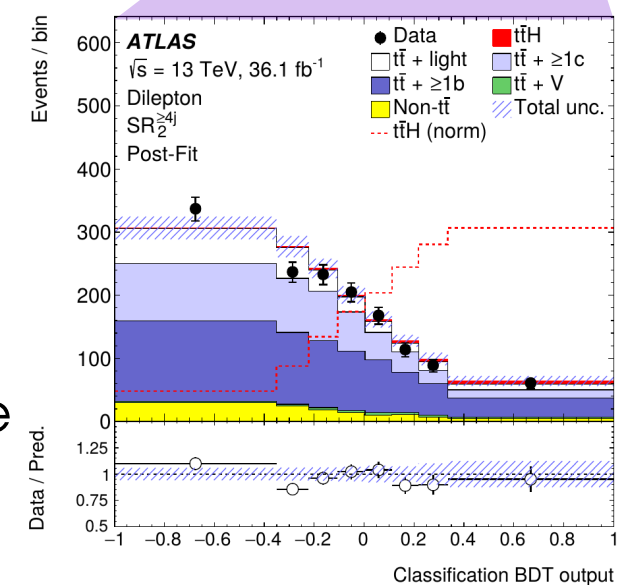
– Also ***Control regions*** for backgrounds



⇒ ***Analysis categories*** :

PDF for category k

$$P\left(S;\{n_i^{(k)}\}_{i=1\ldots n_{\text{evts}}^{(k)}}^{k=1\ldots n_{\text{cats}}}\right) = \prod_{k=1}^{n_{\text{cats}}} P_k\left(S;\{n_i^{(k)}\}_{i=1\ldots n_{\text{evts}}^{(k)}}\right)$$



No overlaps between categories ⇒ Stat. Indepedence

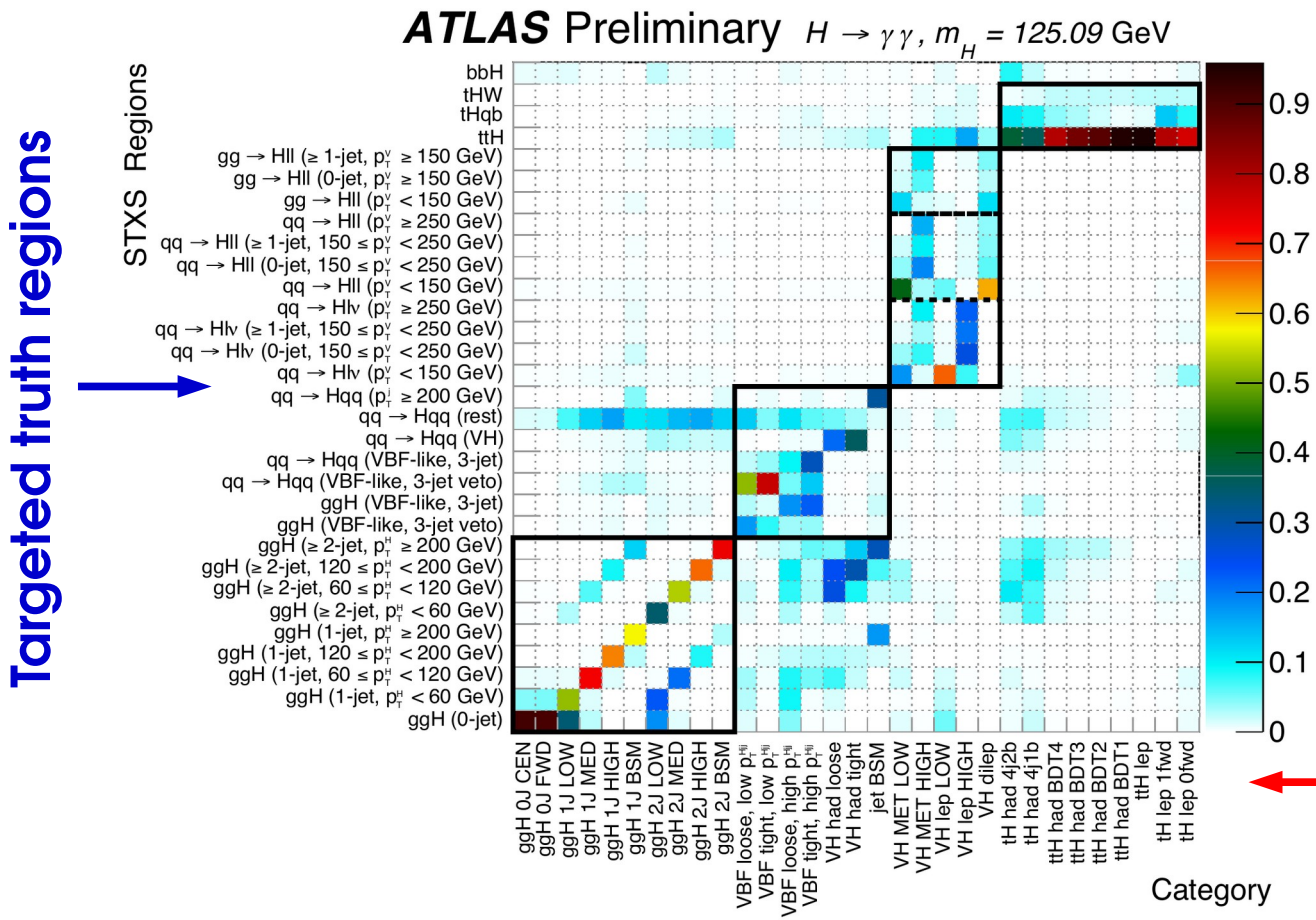
⇒ **can simply take product of PDFs.**

→ Similar to a-posteriori combination of the various regions, but allows proper handling of correlated parameters (e.g. systematics).

Categories for H→γγ Property Measurements

Categories also useful to provide measurements of separate kinematic regions

→ e.g. differential cross-section measurements



Targeted truth regions

Analysis Selections

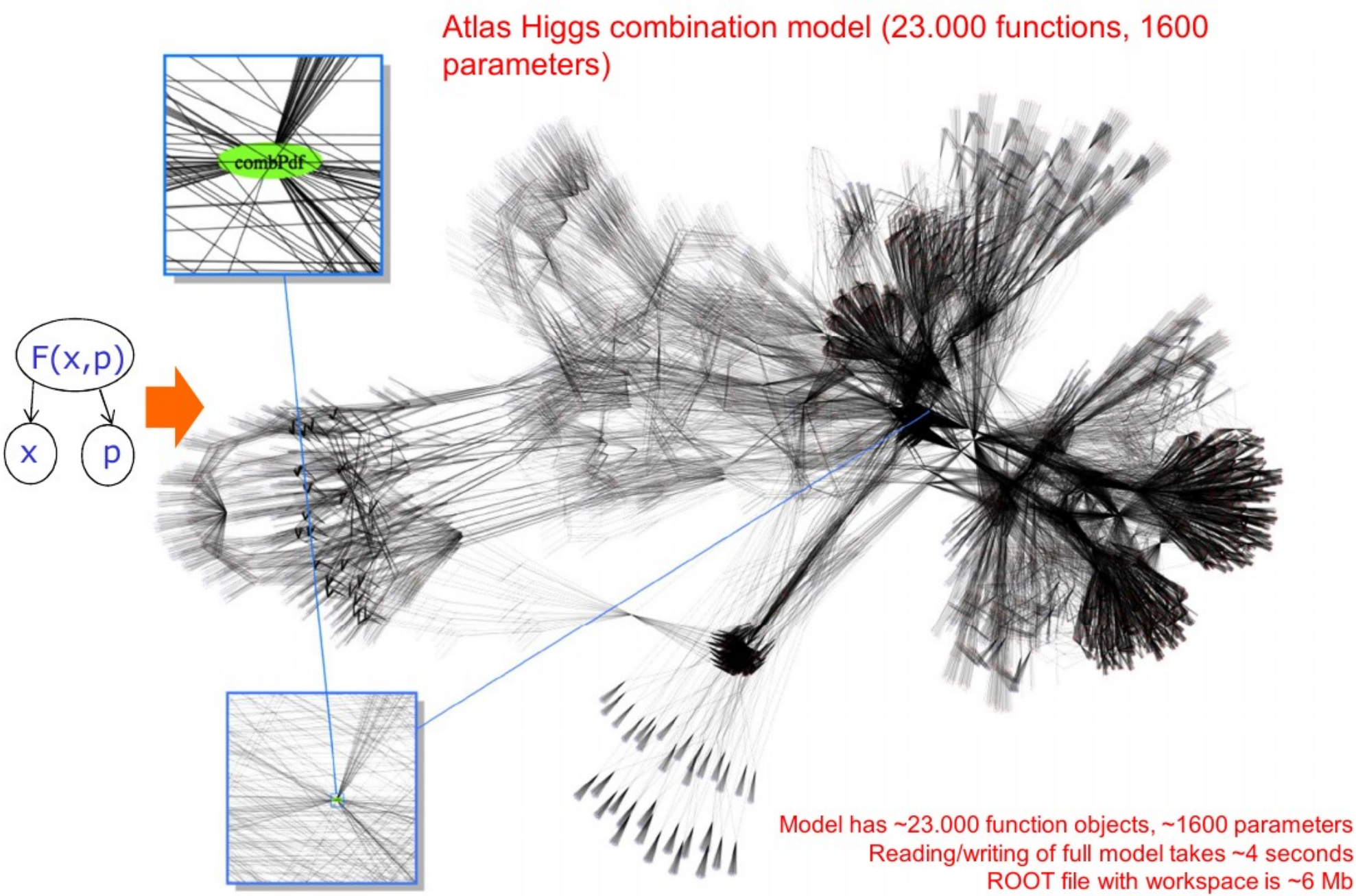
H→ γγ Properties Measurement (ATLAS-CONF-2017-045)

Most **categories** aimed at one particular **truth region**

→ also cross-feed from other regions (detector acceptance, pileup, etc.)

⇒ Combined analysis for optimal use of all information

ATLAS Higgs Combination Model

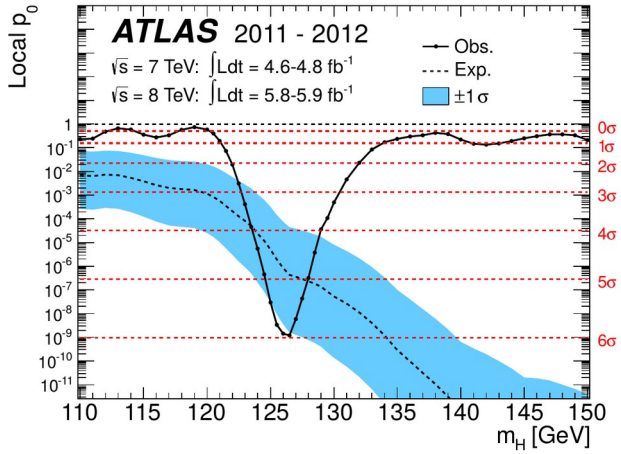


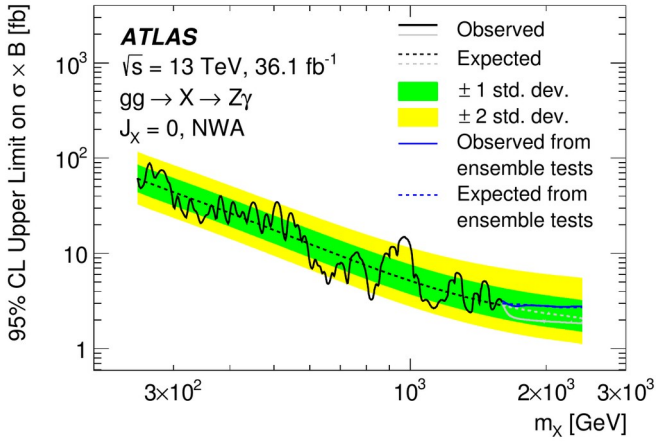
Statistical computations

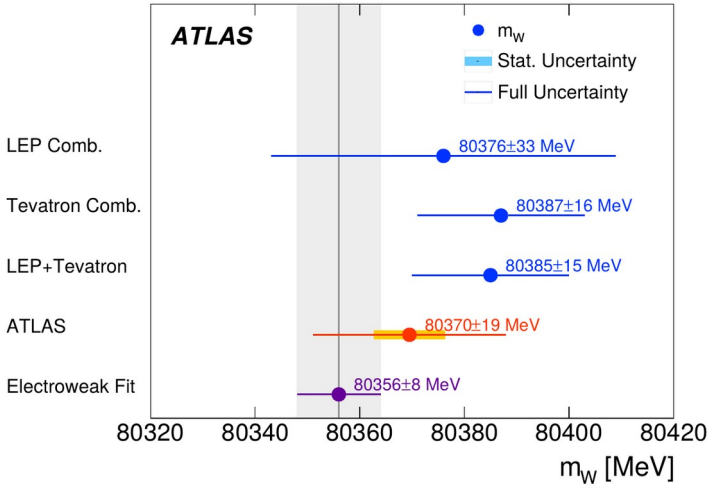
Now that we have a model, can use it to compute analysis results:

- **Discovery significance**: we see an excess – is it a (new) signal, or a background fluctuation ?
- **Upper limit on signal yield**: we don't see an excess – if there is a signal present, how small must it be ?
- **Parameter measurement:** what is the allowed range for a model parameter ? ("confidence interval")

→ The Statistical Model already contains all the needed information – how to use it ?







Plan for today

Lecture 1:

- Statistics basics
- Describing measurements

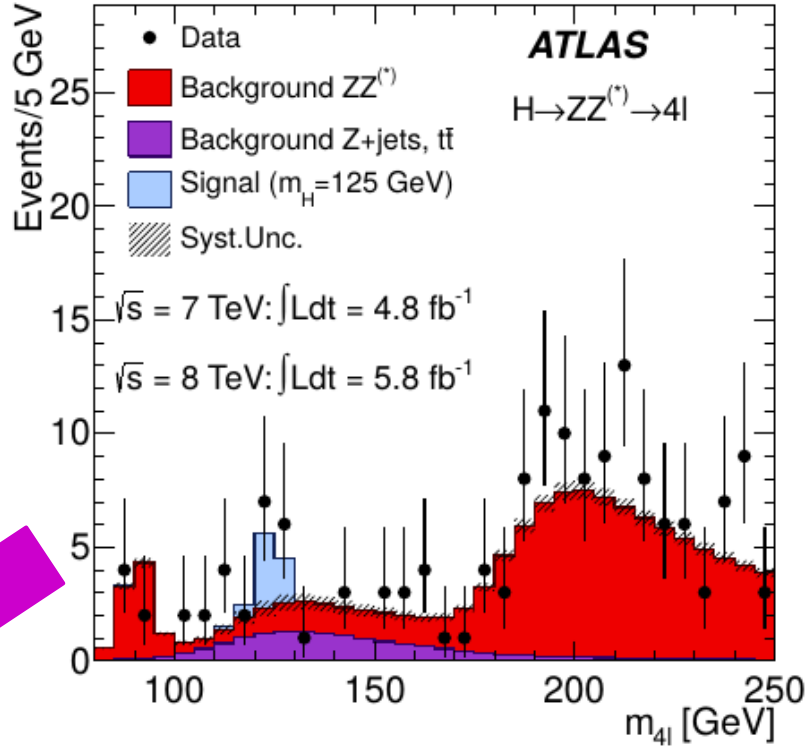
Today:

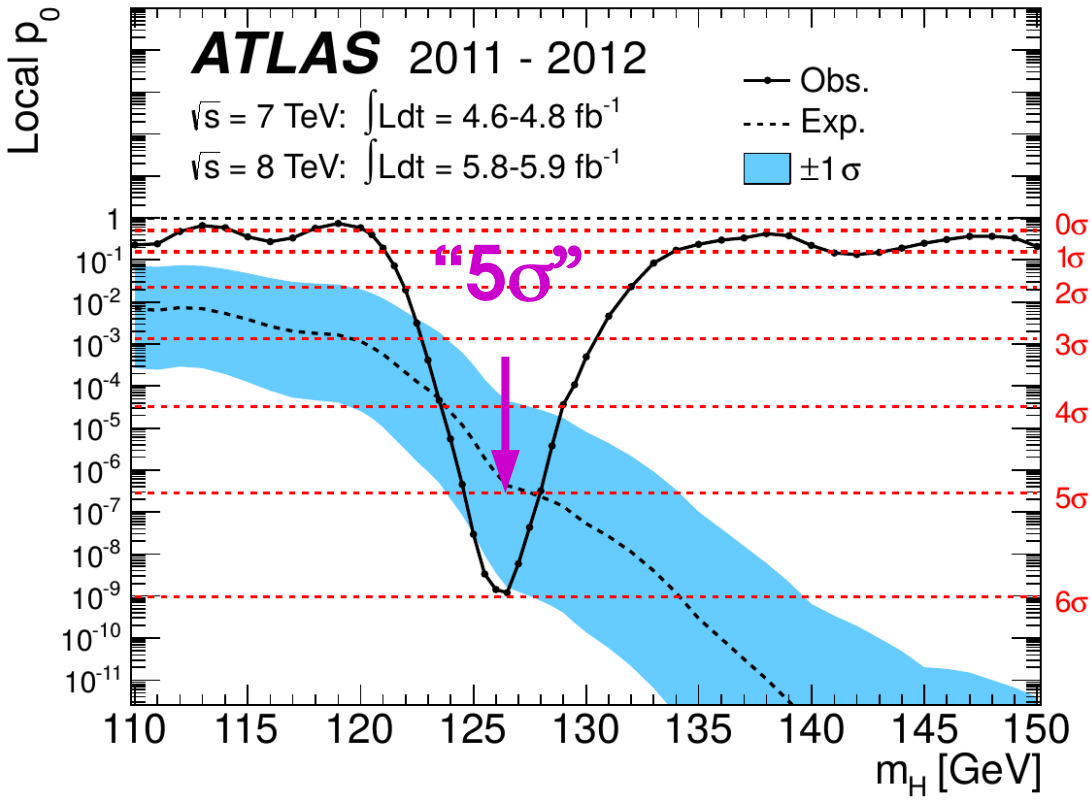
- **Computing statistical results:**
  - Estimating a parameter
  - Testing hypotheses
  - Discovery testing

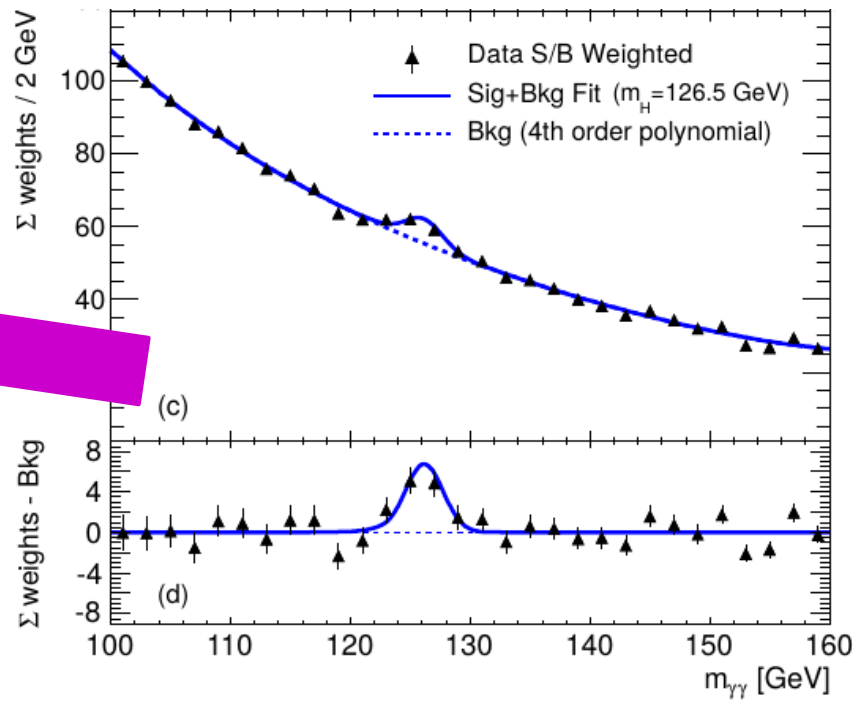
Discovery Testing

We see an unexpected feature in our data, is it a signal for new physics or a fluctuation ?

e.g. Higgs discovery : **"We have 5σ" !**

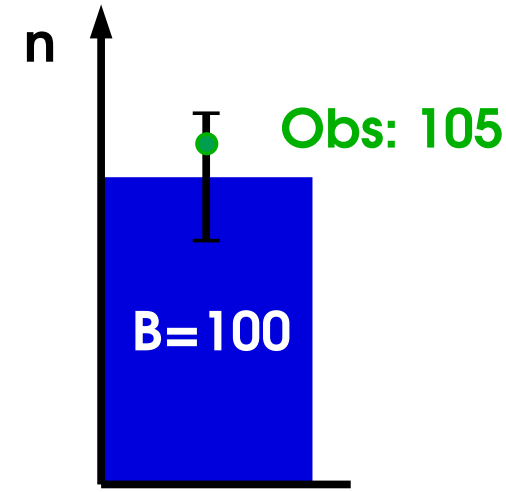






Phys. Lett. B 716 (2012) 1-29

Discovery Testing



Say we have a Gaussian measurement with a background B=100, and we measure n=105

Do we have a signal ?

Probably not :

S = 5, uncertainty on B is $\sqrt{B}$ = 10

⇒ we are only ~ 0.5σ away from S=0.

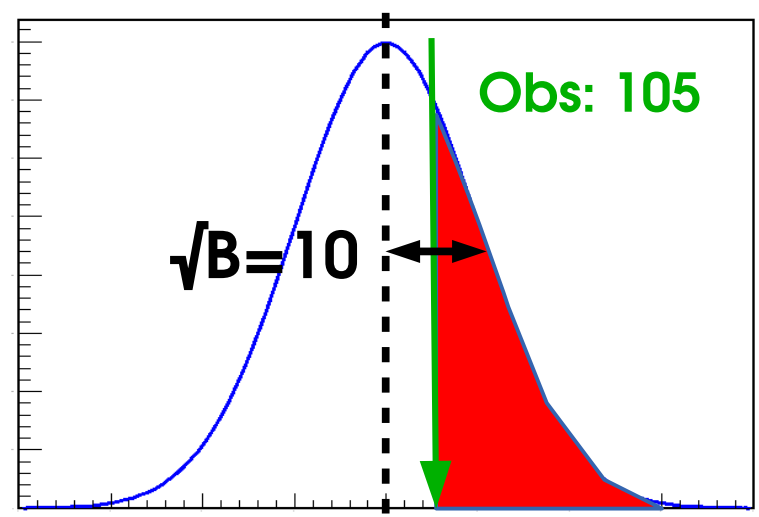
Gaussian quantiles : happens ~31% of the time for S=0, so not exceptional

$S = n_{obs} - B$

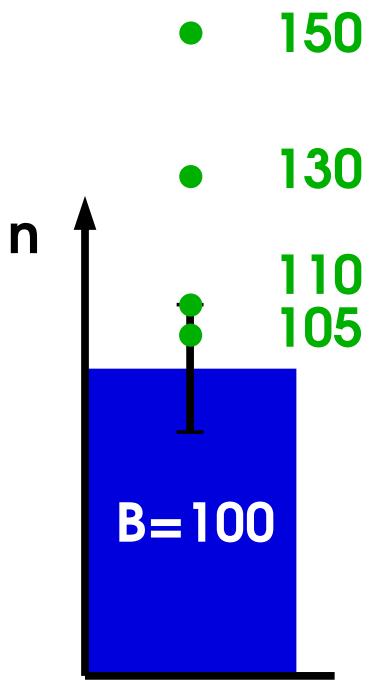
Significance: $$Z = \frac{S}{\sqrt{B}}$$

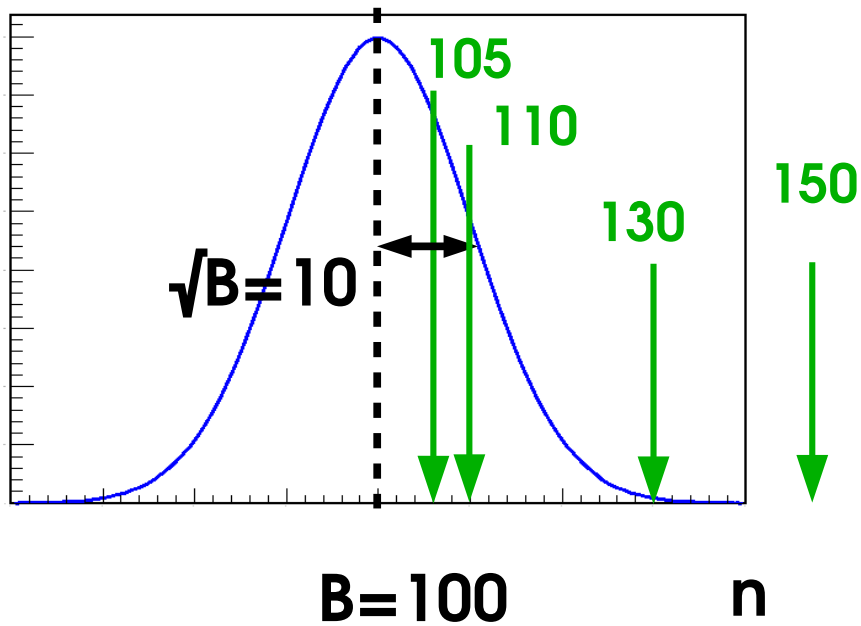
P-value: $$p_0 = 1 - \Phi(Z)$$

$$\Phi(z) = \int_{-\infty}^{z} G(u;0,1)\, du$$



Discovery Testing





n	S	Z	p_0	
105	5	0.5σ	31%	
110	10	1σ	16%	
130	30	3σ	0.1%	Evidence
150	50	5σ	$3\ 10^{-7}$	Discovery

Straightforward in this Gaussian case

In more complex cases, also need to

- **Determine S**
- **Compute Z and p_0**

Outline

Using the PDF

Model describes the distribution of the observable: **P(data; parameters)**

⇒ Possible outcomes of the experiment, for given parameter values

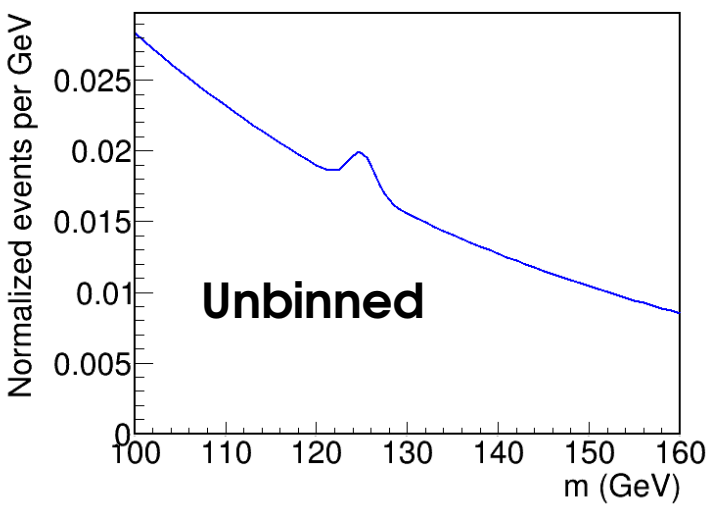
Can draw random events according to PDF : **generate *pseudo-data***

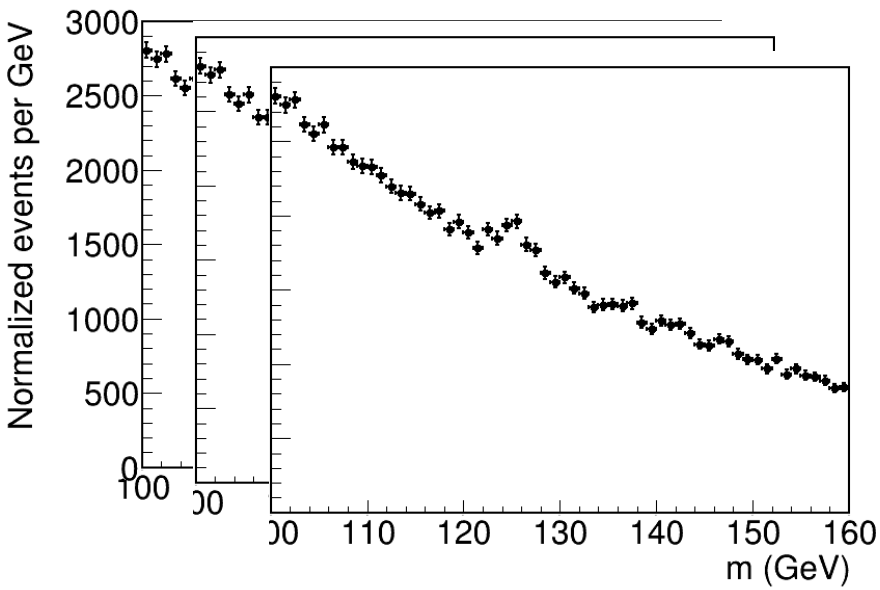
$$P(\lambda = 5)$$

2, 5, 3, 7, 4, 9, ...

Each entry = separate "experiment"

Generate





Likelihood

Model describes the distribution of the observable: **P(data; parameters)**

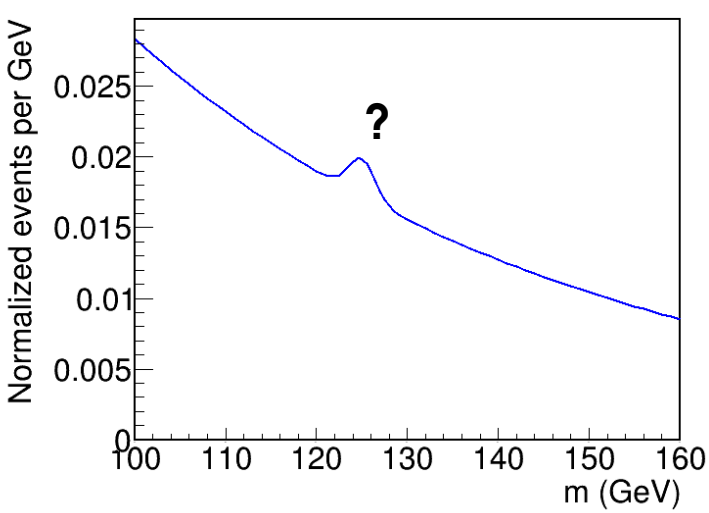
⇒ Possible outcomes of the experiment, for given parameter values

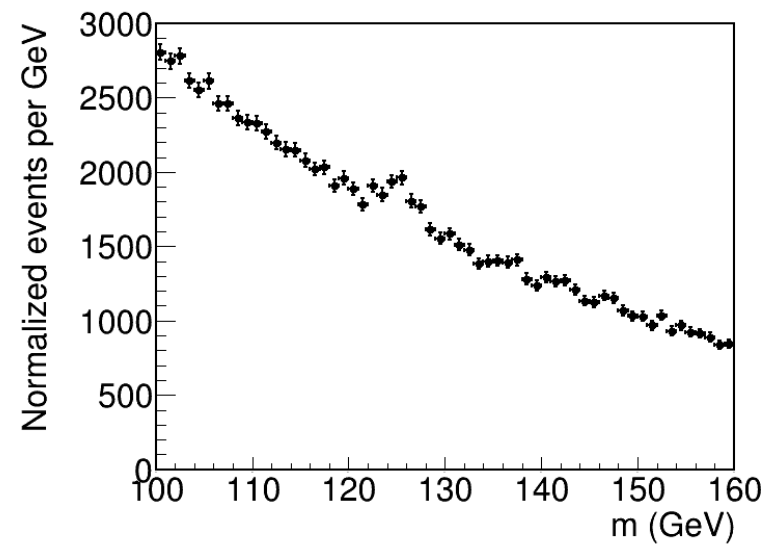
We want the **other** direction: **use data to get information on parameters**

$$P(\lambda = ?)$$

2

Estimate





Likelihood: **L(parameters) = P(data; parameters)**

→ **same as the PDF,** but seen as function of the parameters

Poisson Example

Assume **Poisson distribution** with B = 0 :

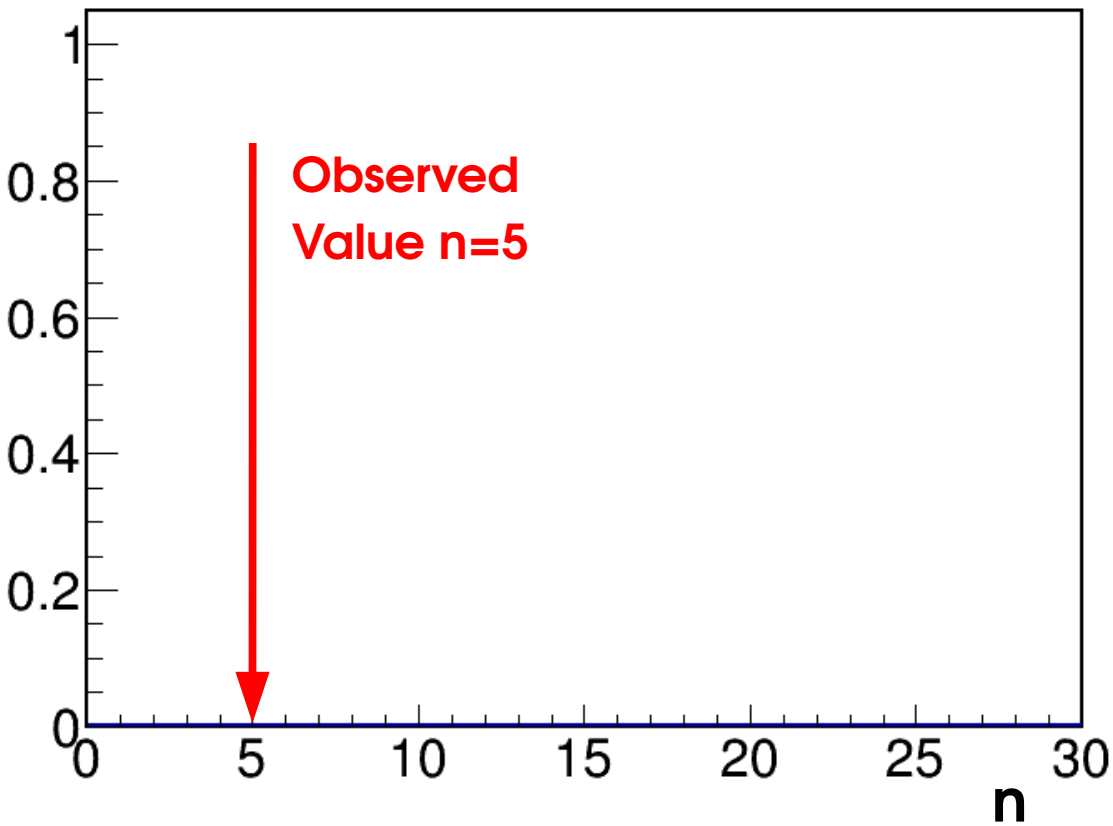
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

Say we **observe n=5**, want to infer information on the parameter **S**

→ Try different values of S for a fixed data value n=5

→ Varying parameter, fixed data: **likelihood**

$$L(S;n=5) = e^{-S}\frac{S^5}{5!}$$



Poisson Example

Assume **Poisson distribution** with B = 0 :

$$P(n;S) = e^{-s}\frac{S^n}{n!}$$

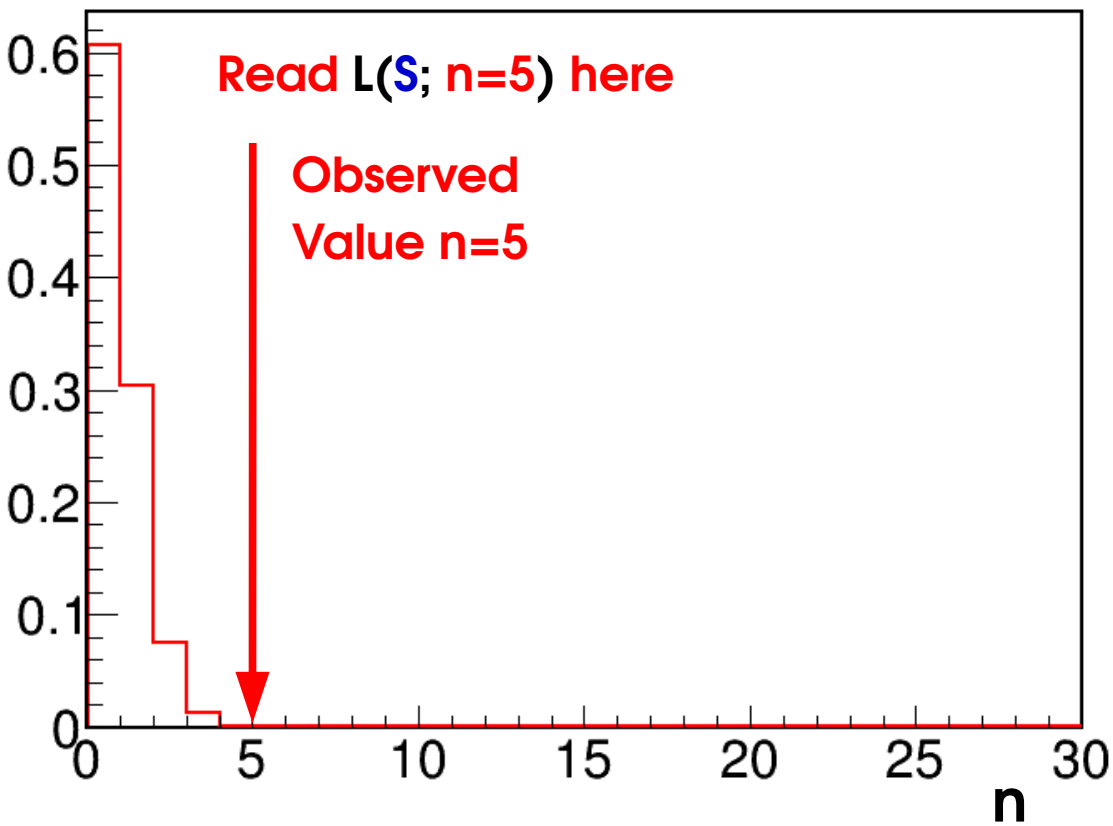
Say we **observe** n=5, want to infer information on the parameter S

→ Try different values of S for a fixed data value n=5

→ Varying parameter, fixed data: **likelihood**

$$L(S;n=5) = e^{-s}\frac{S^5}{5!}$$

P(S = 0.5)
Low
likelihood



Poisson Example

Assume **Poisson distribution** with B = 0 :

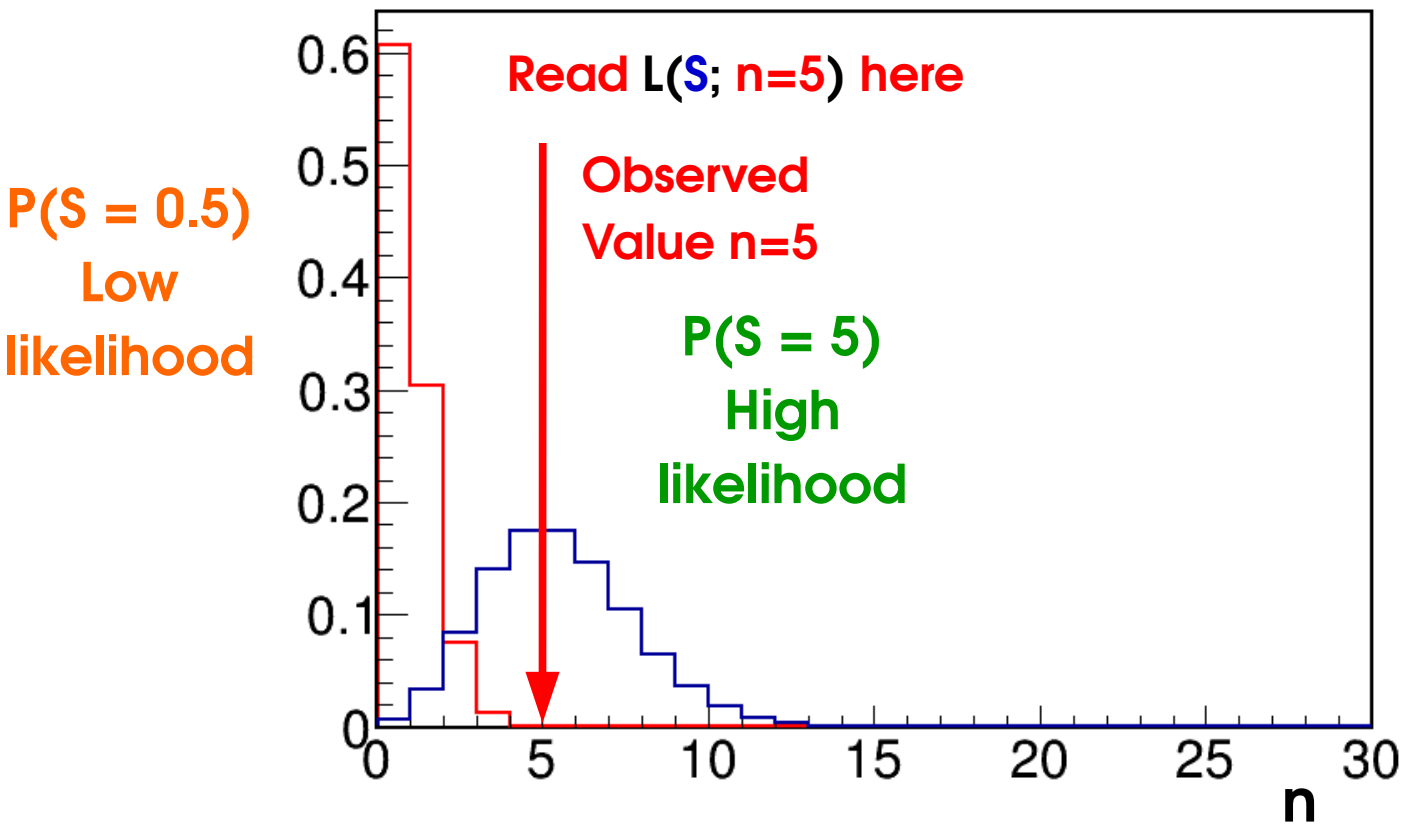
$$P(n;S) = e^{-s}\frac{S^n}{n!}$$

Say we **observe** n=5, want to infer information on the parameter S

→ Try different values of S for a fixed data value n=5

→ Varying parameter, fixed data: **likelihood**

$$L(S;n=5) = e^{-s}\frac{S^5}{5!}$$



Poisson Example

Assume **Poisson distribution** with B = 0 :

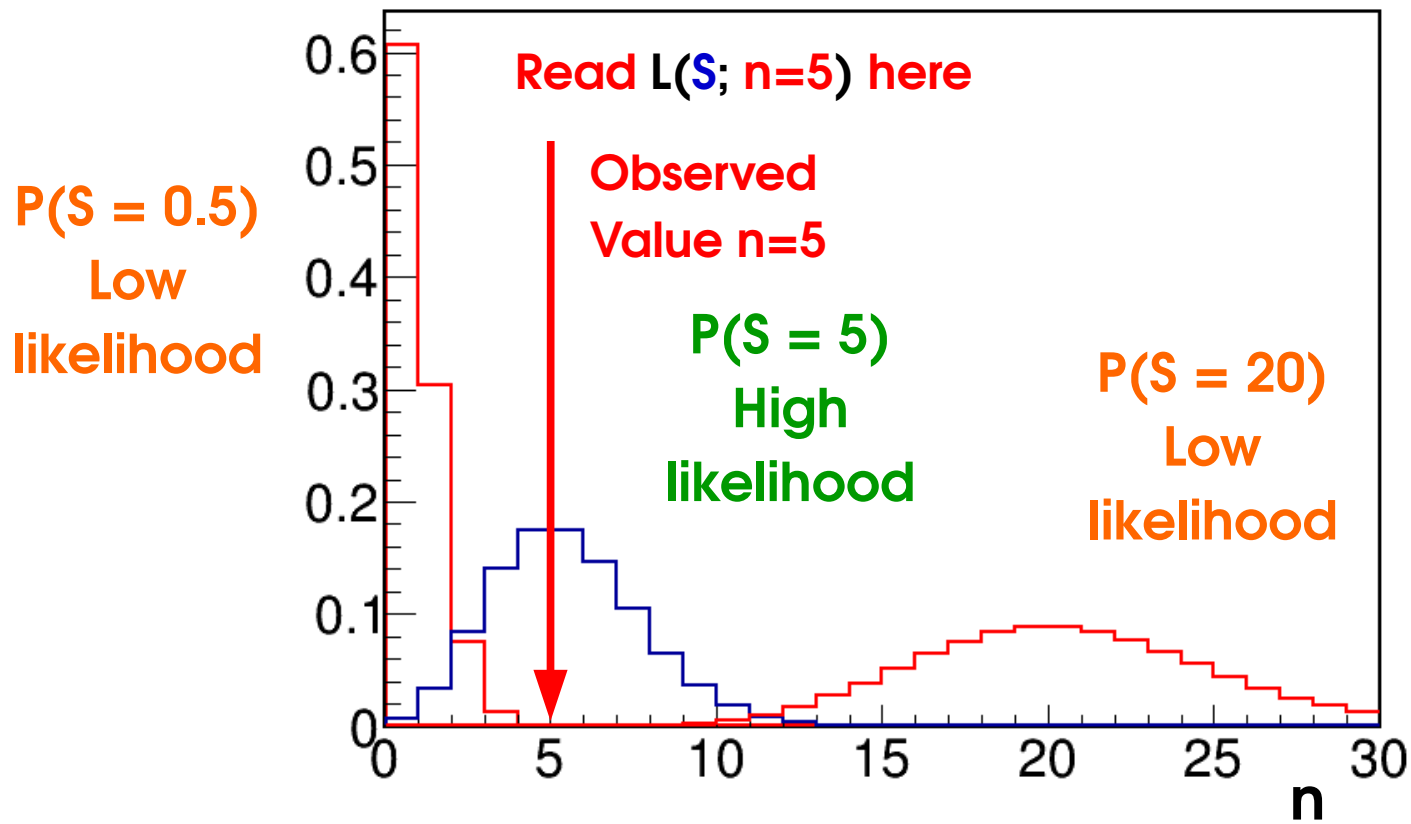
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

Say we **observe** n=5, want to infer information on the parameter S

→ Try different values of S for a fixed data value n=5

→ Varying parameter, fixed data: **likelihood**

$$L(S;n=5) = e^{-S}\frac{S^5}{5!}$$



Poisson Example

Assume **Poisson distribution** with B = 0 :

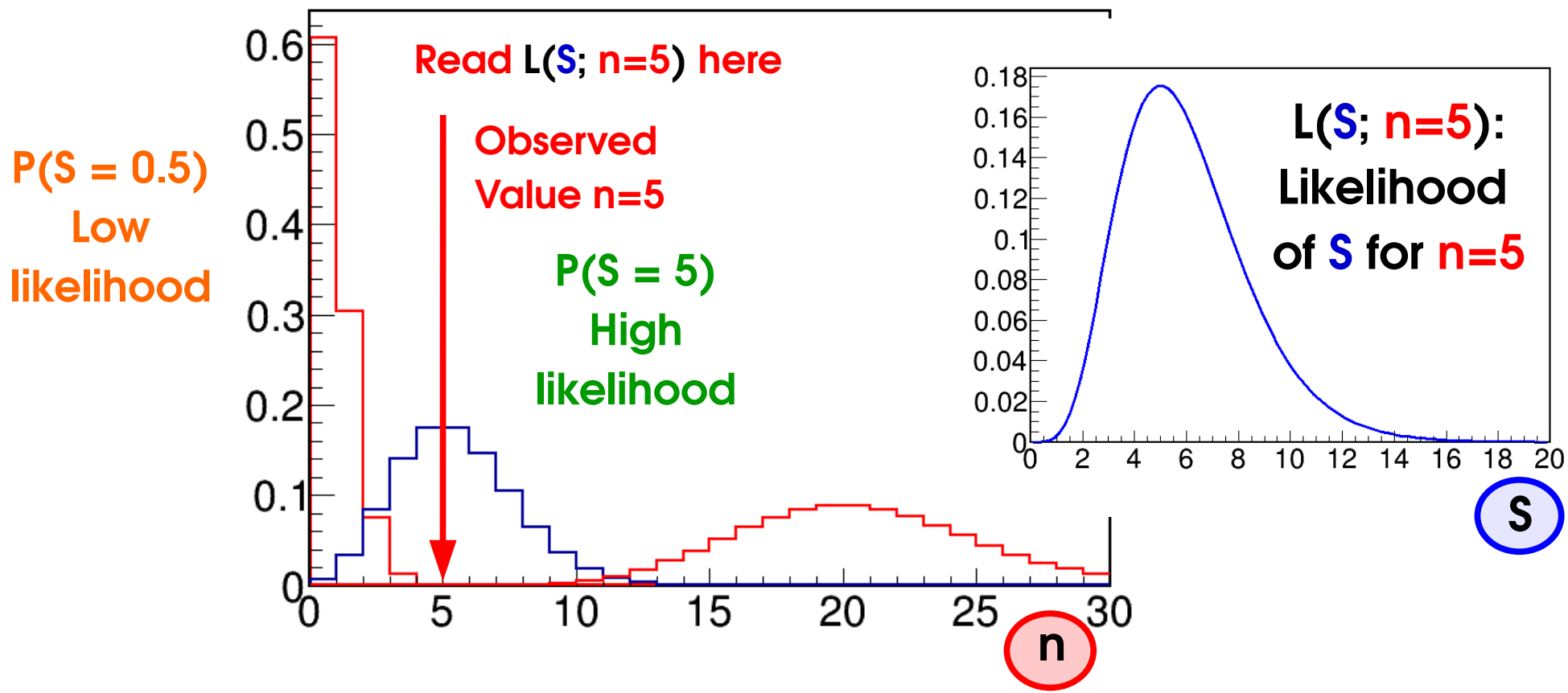
$$P(n;S) = e^{-S}\frac{S^n}{n!}$$

Say we **observe n=5**, want to infer information on the parameter **S**

→ Try different values of S for a fixed data value n=5

→ Varying parameter, fixed data: **likelihood**

$$L(S;n=5) = e^{-S}\frac{S^5}{5!}$$

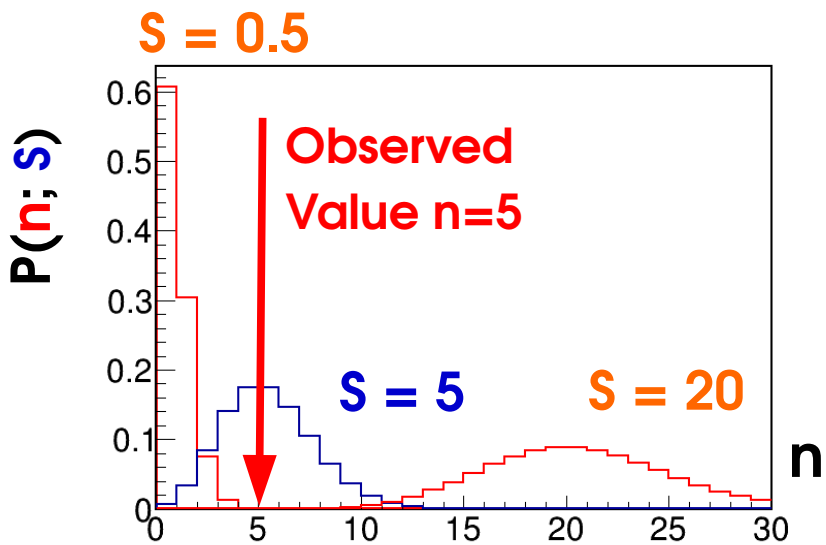


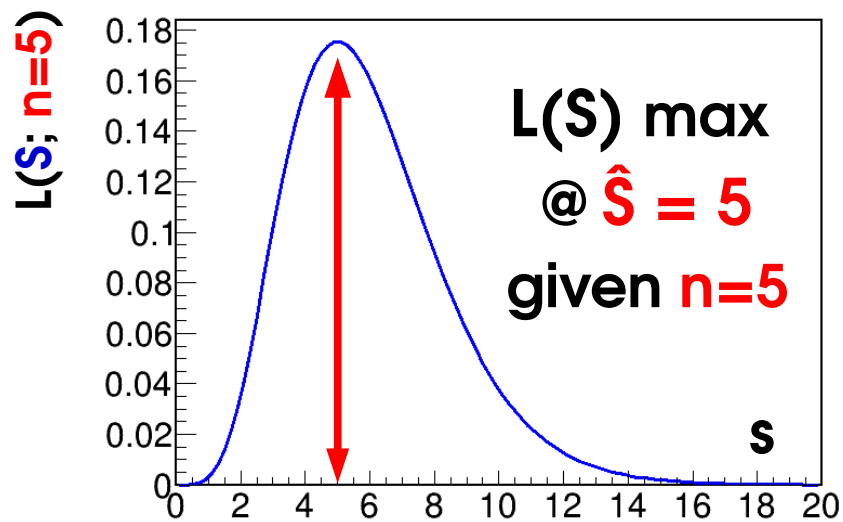
Maximum Likelihood Estimation

To estimate a parameter μ, find the **value $\hat{\mu}$ that maximizes** L(μ)

Maximum Likelihood Estimator (MLE) $\hat{\mu}$:

$$\hat{\mu} = arg\,max\, L(\mu)$$



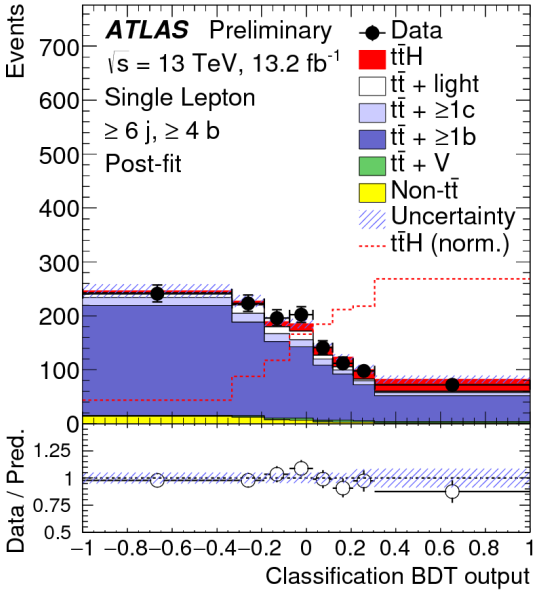


MLE: the value of μ for which **this data** was ***most likely to occur***

The MLE is a function of the data – itself an **observable**

No guarantee it is the true value (data may be "unlikely") but sensible estimate

MLEs in Shape Analyses



Binned shape analysis:

$$L(\boldsymbol{S};\boldsymbol{n}_i) = P(\boldsymbol{n}_i;\boldsymbol{S}) = \prod_{i=1}^{N} \mathrm{Pois}(\boldsymbol{n}_i;\boldsymbol{S}f_i + B_i)$$

Maximize global L(S) (each bin may prefer a different **S**)

In practice easier to minimize

$$\lambda_{\mathrm{Pois}}(\boldsymbol{S}) = -2\log L(\boldsymbol{S}) = -2\sum_{i=1}^{N} \log \mathrm{Pois}(\boldsymbol{n}_i;\boldsymbol{S}f_i + B_i)$$

Needs a computer...

In the Gaussian limit

$$\lambda_{\mathrm{Gaus}}(\boldsymbol{S}) = \sum_{i=1}^{N} -2\log G(\boldsymbol{n}_i;\boldsymbol{S}f_i + B_i,\sigma_i) = \sum_{i=1}^{N}\left(\frac{\boldsymbol{n}_i-(\boldsymbol{S}f_i + B_i)}{\sigma_i}\right)^2$$

χ^2 formula!

→ **Gaussian MLE** (min χ^2 or min λ_{Gaus}) : ***Best fit value*** in a χ^2 (Least-squares) fit

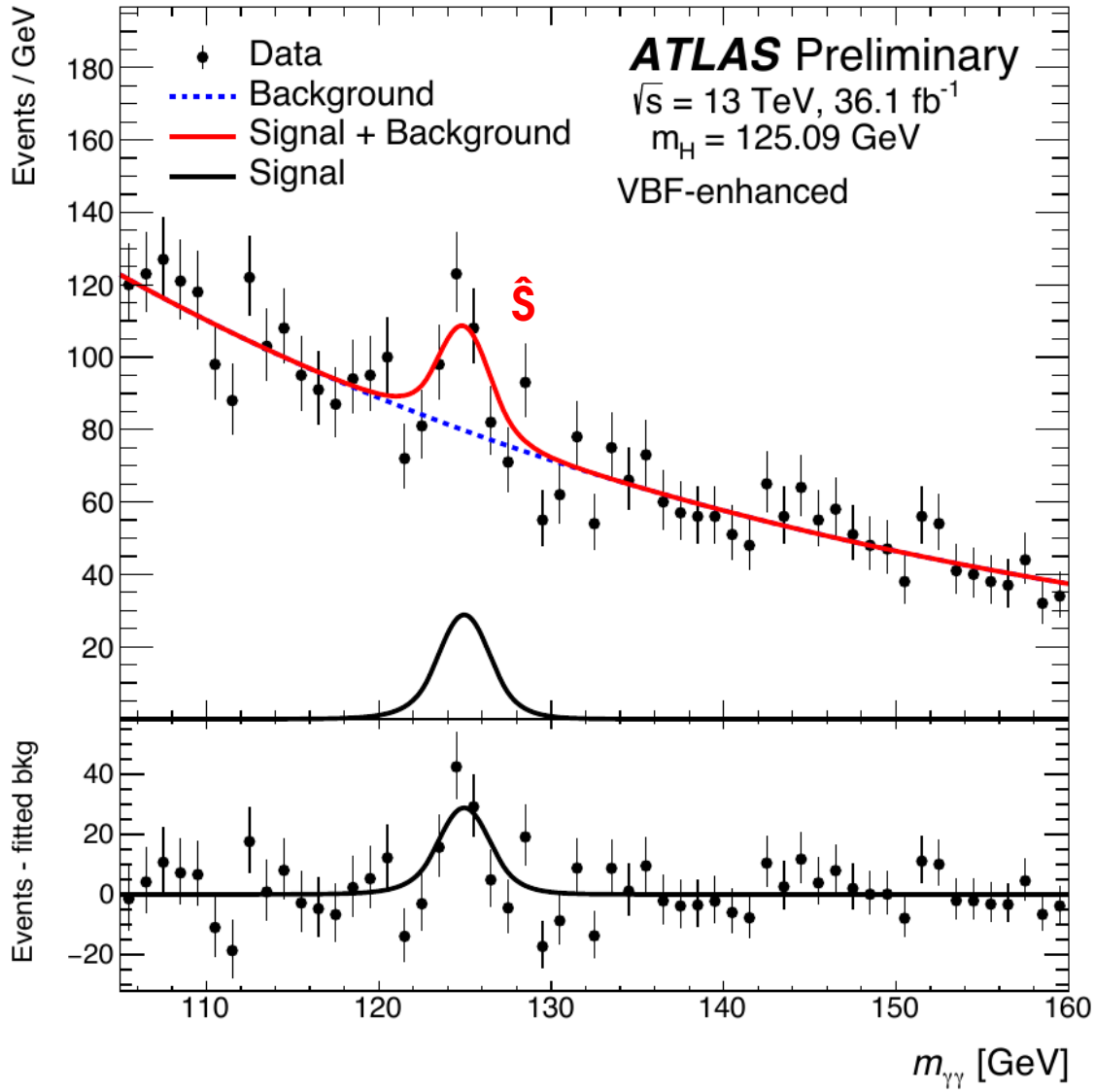
→ **Poisson MLE** (min λ_{Pois}) : ***Best fit value*** in a *likelihood* fit (in `ROOT`, fit option "`L`")

In RooFit, $\boldsymbol{\lambda}_{\mathrm{Pois}}$ ⇒ `RooAbsPdf::fitTo()`, $\boldsymbol{\lambda}_{\mathrm{Gaus}}$ ⇒ `RooAbsPdf::chi2FitTo()`.

In both cases, MLE ⇔ *Best Fit*

H→γγ

$$L(S,B;m_i)=e^{-(S+B)}\prod_{i=1}^{n_{evts}} S\,P_{sig}(m_i)+B\,P_{bkg}(m_i)$$



Estimate the MLE $\hat{S}$ *of* S *?*

→ Perform (likelihood) best-fit of model to data

⇒ fit result for S is the desired $\hat{S}$.

In particle physics, often use the *MINUIT* minimizer within ROOT.

ATLAS-CONF-2017-045

MLE Properties

- **Asymptotically Gaussian** $P(\hat{\mu}) \propto \exp\left(-\frac{(\hat{\mu}-\mu^*)^2}{2\sigma_{\hat{\mu}}^2}\right)$ for $n \to \infty$

  and unbiased $\langle \hat{\mu} \rangle = \mu^*$ for $n \to \infty$

  Standard deviation of the distribution of $\hat{\mu}$

  for large enough datasets

- **Asymptotically Efficient** : $\sigma_{\hat{\mu}}$ is the **lowest possible value** (in the limit $n \to \infty$) among consistent estimators.

  → MLE captures all the available information in the data

- Also **consistent**: $\hat{\mu}$ converges to the true value for large n, $\hat{\mu} \xrightarrow{n \to \infty} \mu^*$

- **Log-likelihood :** Can also **minimize** $\lambda = -2 \log L$

  → Usually more efficient numerically

  → For Gaussian L, λ is parabolic:

- Can **drop multiplicative constants in L** (additive constants in λ)

Extra: Fisher Information

Fisher Information:

$$I(\mu) = \left\langle \left(\frac{\partial}{\partial \mu} \log L(\mu) \right)^2 \right\rangle = -\left\langle \frac{\partial^2}{\partial \mu^2} \log L(\mu) \right\rangle$$

Measures the **amount of information** available in the measurement of μ.

Gaussian likelihood: $$I(\mu) = \frac{1}{\sigma^2_{\text{Gauss}}}$$

→ smaller σ_{Gauss} ⇒ more information.

Gaussian case:

- For a Gaussian estimator $\tilde{\mu}$

$$P(\tilde{\mu}) \propto \exp\left(-\frac{(\tilde{\mu} - \mu^*)^2}{2\sigma^2_{\tilde{\mu}}} \right)$$

- **MLE**: $\text{Var}(\hat{\mu}) = \sigma_{\hat{\mu}}^2$

Cramer-Rao: $\text{Var}(\tilde{\mu}) \geq \sigma_{\text{Gauss}}^2 = \sigma_{\tilde{\mu}}^2$

Cramer-Rao bound: $$\text{Var}(\tilde{\mu}) \geq \frac{1}{I(\mu)}$$

For any estimator $\tilde{\mu}$.

→ cannot be more precise than allowed by information in the measurement.

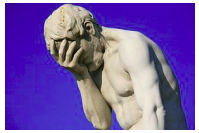
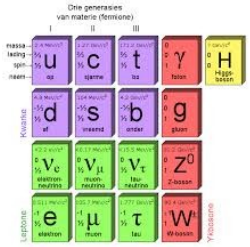
Efficient estimators reach the bound : **e.g. MLE in the large dataset limit.**

Outline

Hypothesis Testing

Hypothesis: assumption on parameters of interest, say value of S (e.g. **H_0 : S=0**)

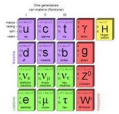
→ **Goal** : decide if H_0 is favored or disfavored using a test based on the data

Possible outcomes:	**Data disfavors H_0 (Discovery claim)**	**Data favors H_0 (Nothing found)**
H_0 is false (New physics!)	**Discovery!**  	**Missed discovery Type-II error (1 - Power)** 
H_0 is true (Nothing new)	**False discovery claim Type-I error (→ p-value, significance)** 	**No new physics, none found** 

"... the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only to give the facts a chance of disproving the null hypothesis." – R. A. Fisher

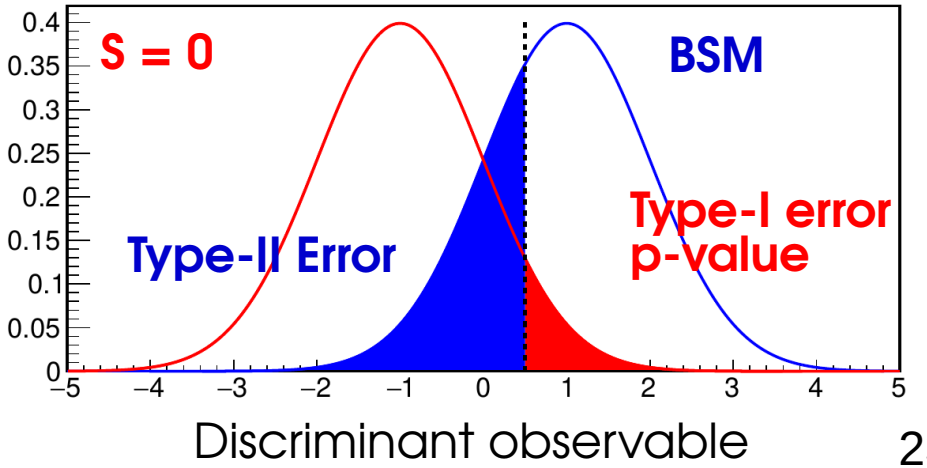
Hypothesis Testing

Hypothesis: assumption on model parameters, say value of S (e.g. **H_0 : S=0**)

	Data disfavors H_0 (Discovery claim)	Data favors H_0 (Nothing found)
H_0 is false (New physics!)	Discovery! 	Type-II error (Missed discovery) 
H_0 is true (Nothing new)	Type-I error (False discovery) 	No new physics, none found 

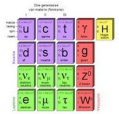
Lower Type-I errors ⇔ **Higher Type-II errors** and vice versa: cannot have everything!

→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.



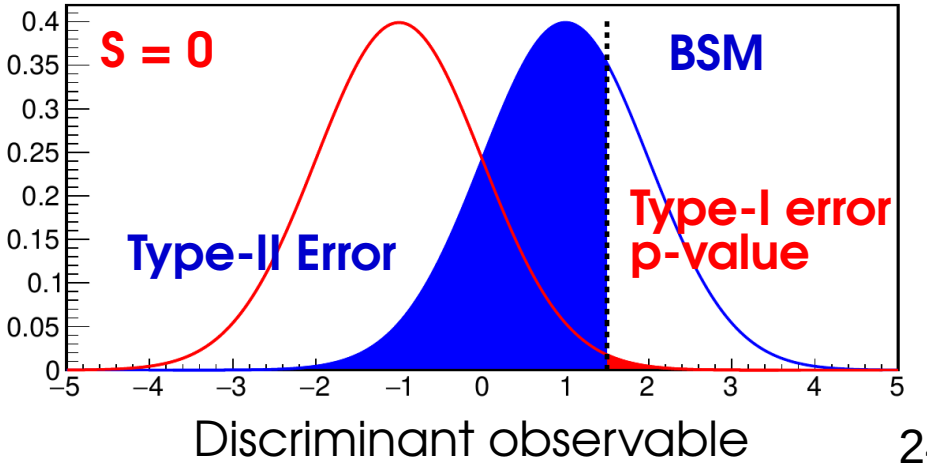
Hypothesis Testing

Hypothesis: assumption on model parameters, say value of S (e.g. **H_0 : S=0**)

	Data disfavors H_0 (Discovery claim)	Data favors H_0 (Nothing found)
H_0 is false (New physics!)	Discovery! 	Type-II error (Missed discovery) 
H_0 is true (Nothing new)	Type-I error (False discovery) 	No new physics, none found 

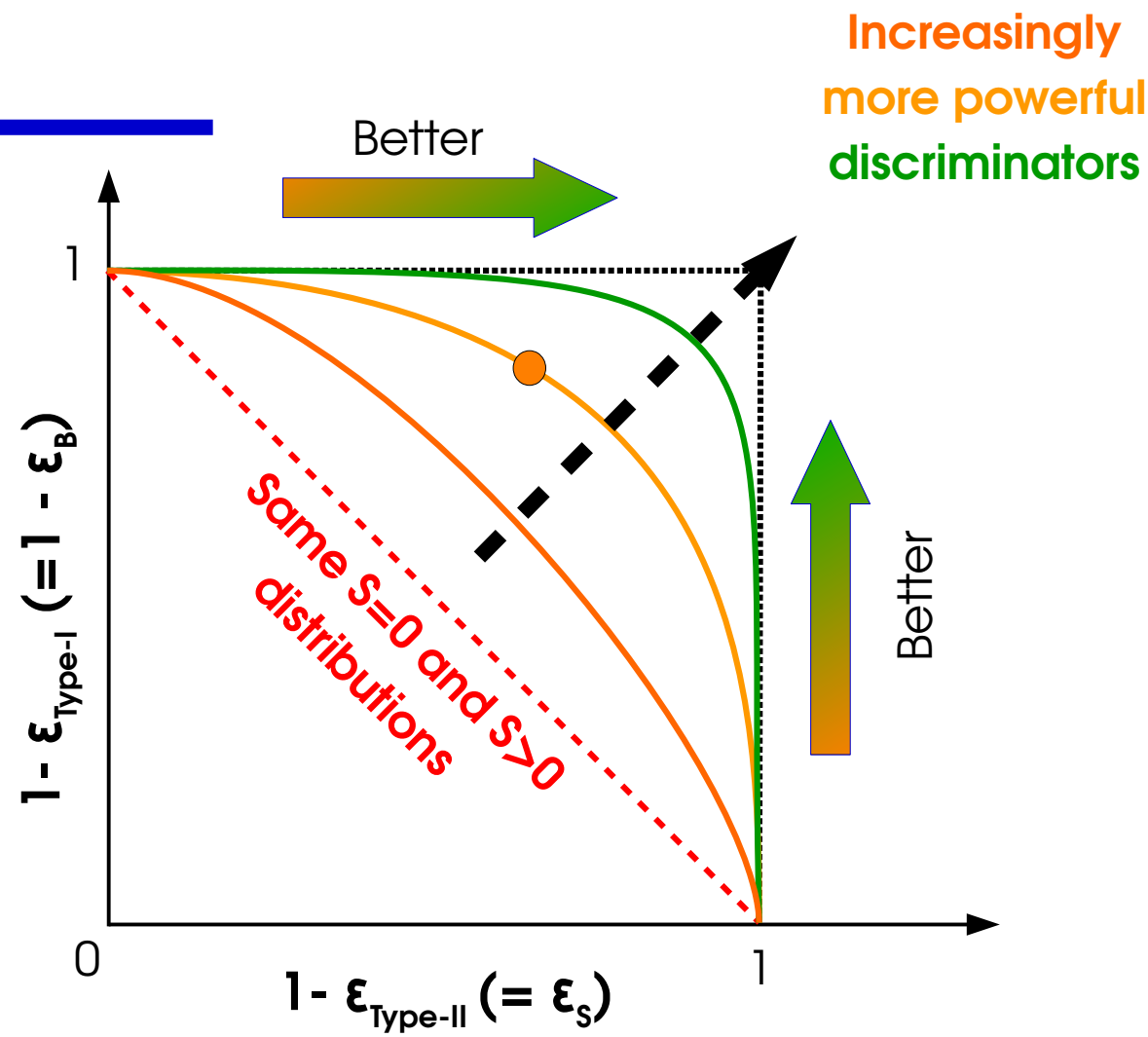
Lower Type-I errors ⇔ **Higher Type-II errors** and vice versa: cannot have everything!

→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.



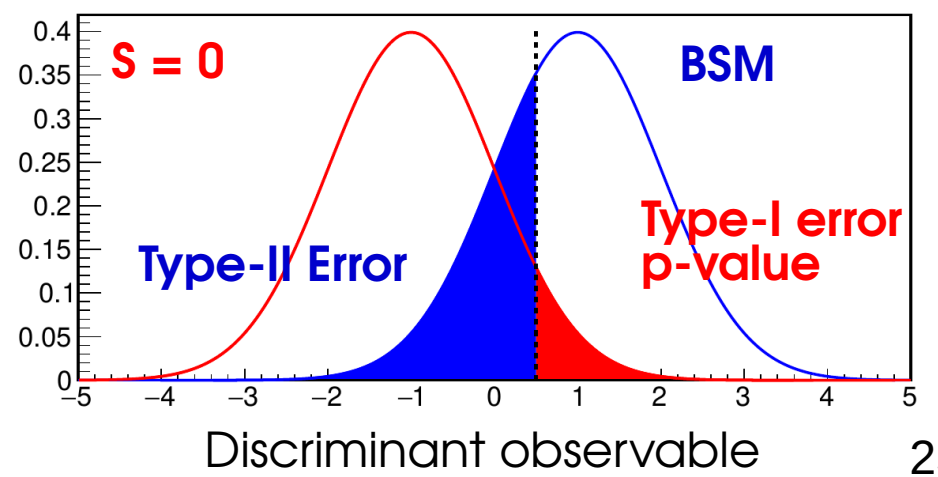
ROC Curves

"Receiver operating characteristic" (ROC) Curve:
→ Plot Type-I vs Type-II rates for different cut values
→ All curves monotonically decrease from (0,1) to (1,0)
→ Better discriminators more bent towards (1,1)



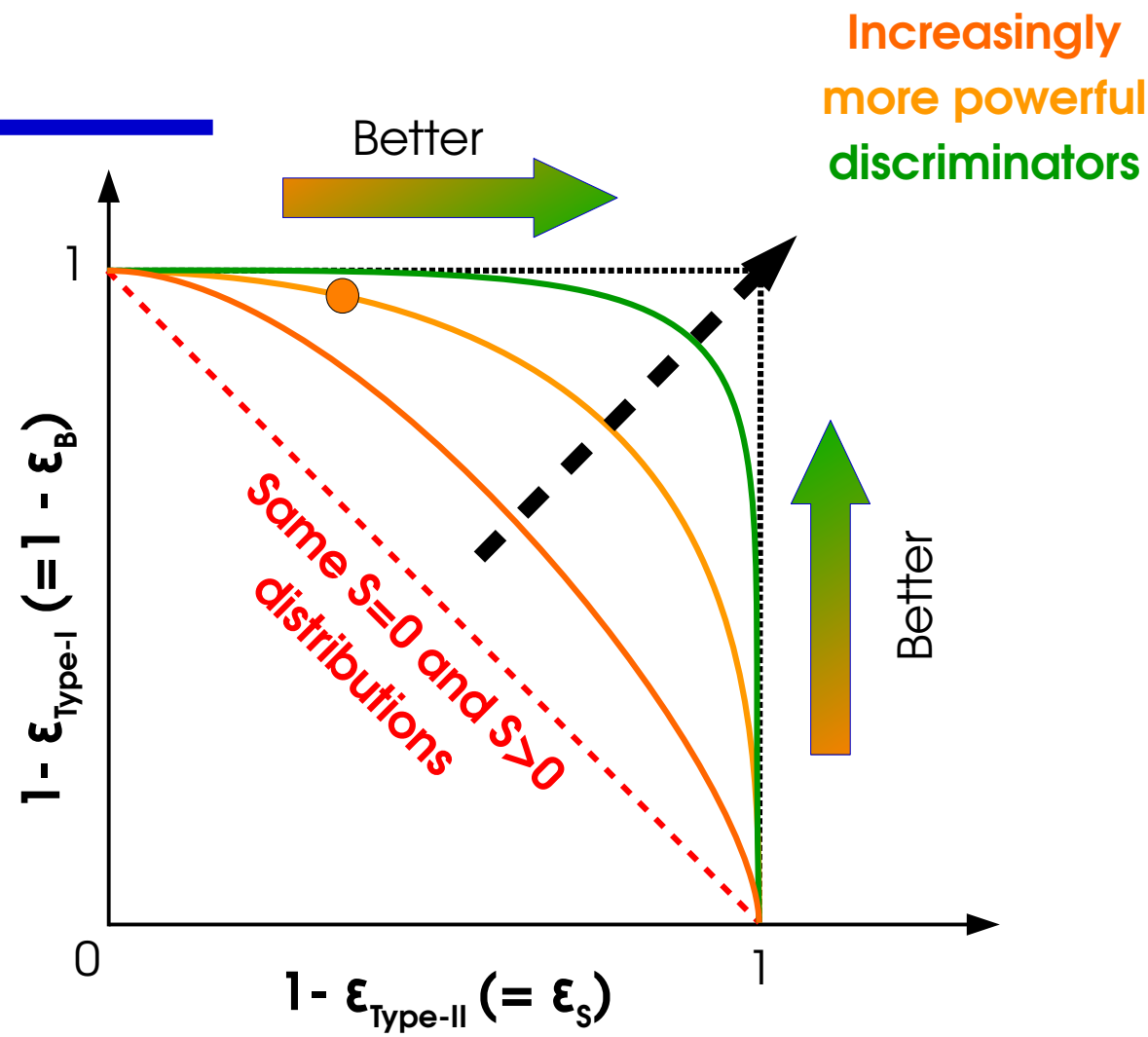
→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.

→ Usually set predefined level of **acceptable Type-I error** (e.g. "5σ")



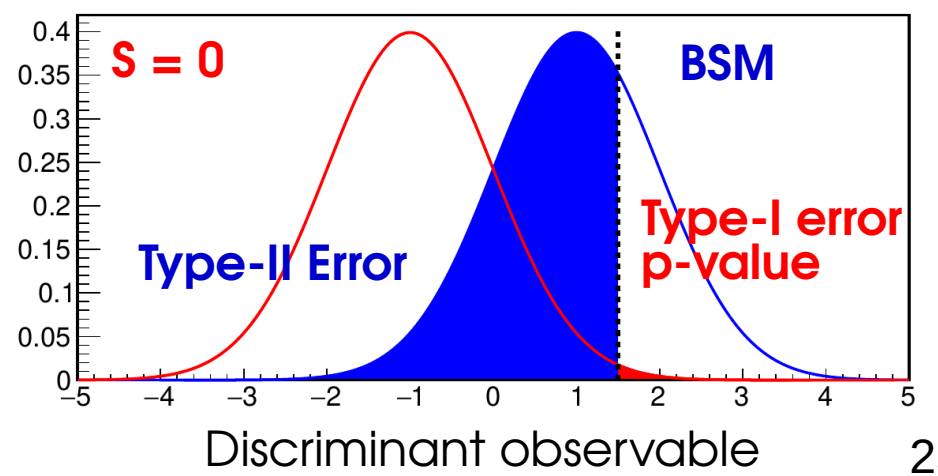
ROC Curves

"Receiver operating characteristic" (ROC) Curve:
→ Plot Type-I vs Type-II rates for different cut values
→ All curves monotonically decrease from (0,1) to (1,0)
→ Better discriminators more bent towards (1,1)



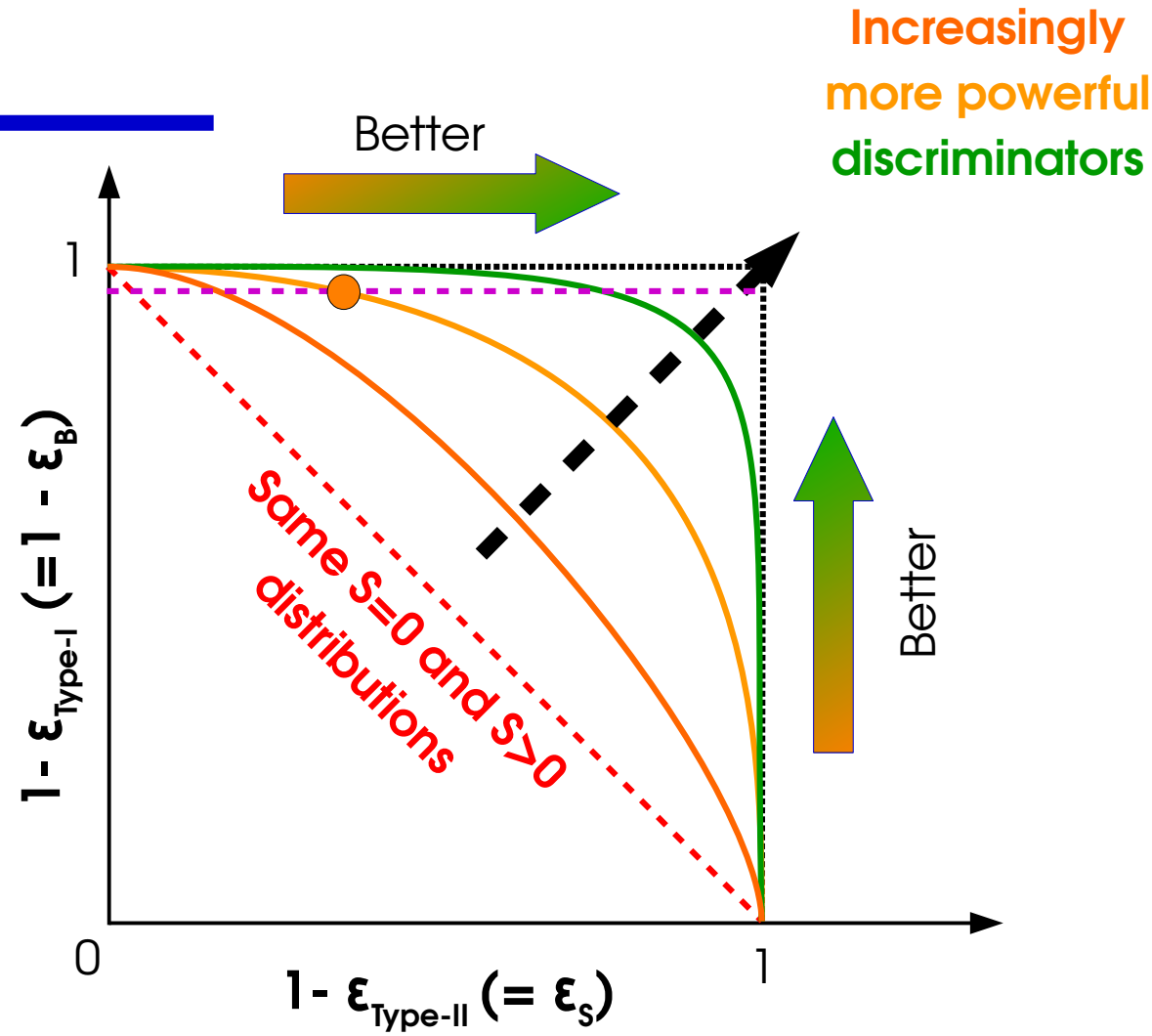
→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.

→ Usually set predefined level of **acceptable Type-I error** (e.g. "5σ")



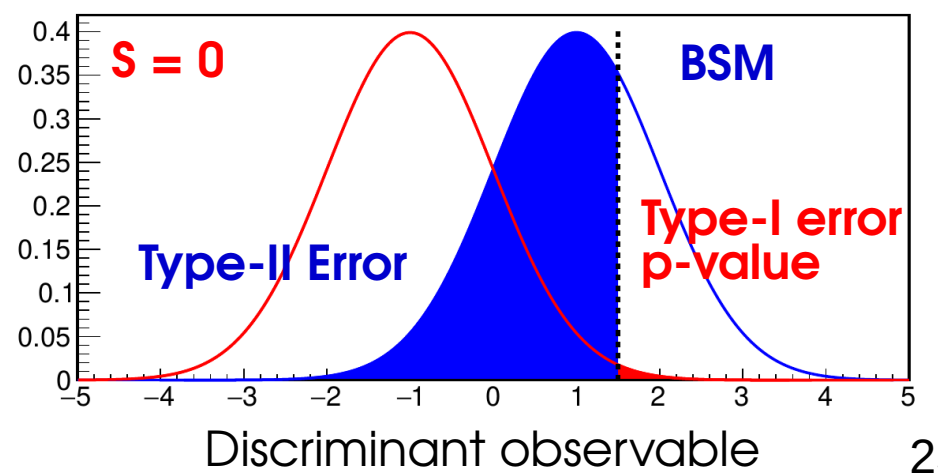
ROC Curves

"Receiver operating characteristic" (ROC) Curve:
→ Plot Type-I vs Type-II rates for different cut values
→ All curves monotonically decrease from (0,1) to (1,0)
→ Better discriminators more bent towards (1,1)



→ **Goal**: test that minimizes Type-II errors **for given level of Type-I error**.

→ Usually set predefined level of **acceptable Type-I error** (e.g. "5σ")



Hypothesis Testing with Likelihoods

Neyman-Pearson Lemma

When comparing two hypotheses H_0 *and* H_1*, the optimal discriminator is the* ***Likelihood ratio*** (*LR*)

$$\frac{L(H_1; data)}{L(H_0; data)}$$

e.g. $$\frac{L(S = 5; data)}{L(S = 0; data)}$$

Caveat: Strictly true only for *simple hypotheses* (no free parameters)

As for MLE, choose the hypothesis that is more likely **given the data we have**.

→ **Minimizes Type-II uncertainties** for given level of Type-I uncertainties
→ Always need an **alternate hypothesis** to test against.

→ **In the following**: all tests based on LR, will focus on p-values (Type-I errors), trusting that Type-II errors are anyway as small as they can be...

Outline

Computing statistical results

Estimating the value of a parameter

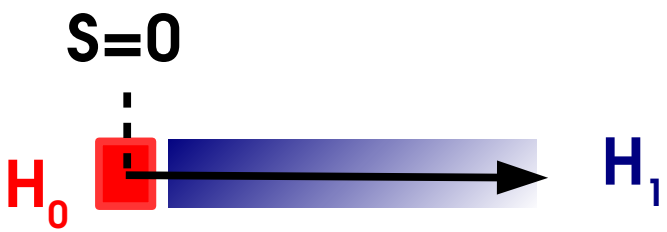
Testing hypotheses

Discovery significance

Discovery: Test Statistic

Cowan, Cranmer, Gross & Vitells,
Eur.Phys.J.C71:1554,2011

Discovery :



- H_0 : **background only (S = 0)** against
- H_1: presence of a signal (**S > 0**)

→ For H_1, any S>0 is possible, which to use ? **The one preferred by the data, $\hat{S}$.**

⇒ Use Likelihood ratio: $$\frac{L(S=0)}{L(\hat{S})}$$

→ In fact use the **test statistic** $$q_0 = -2\log\frac{L(S=0)}{L(\hat{S})}$$

Note: for $\hat{S} < 0$, set $q_0=0$ to reject negative signals ("*one-sided* test statistic")

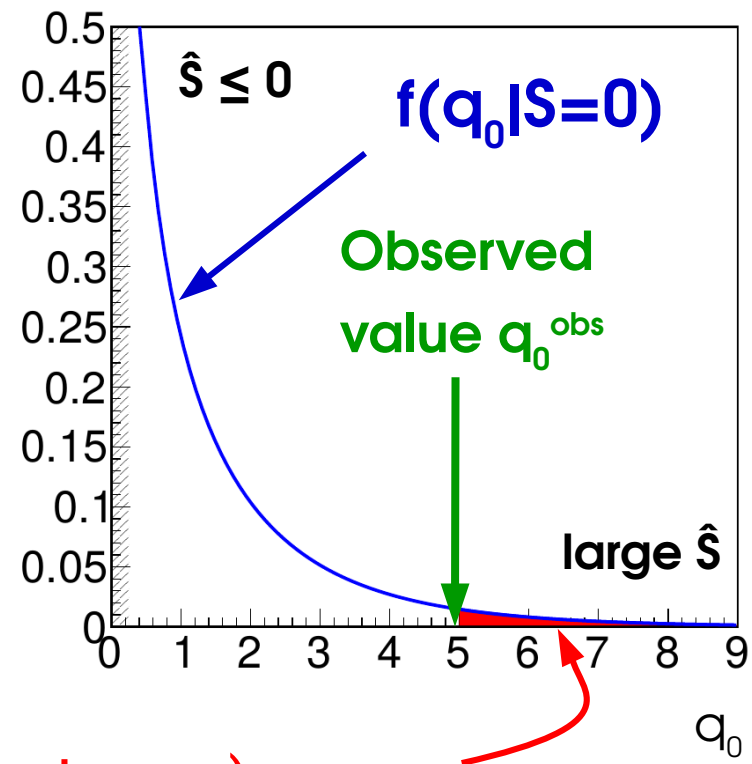
Discovery p-value

Large values of $-2\log\frac{L(S=0)}{L(\hat{S})}$ if:

⇒ observed $\hat{S}$ is far from 0

⇒ **$H_0(S=0)$** ***disfavored*** compared to **$H_1(S\neq 0)$**.

data prefer S = 0 ⟷ data prefer S > 0



How large q_0 before we can exclude H_0 ?

(and **claim a discovery!**)

→ Need small Type-I rate (falsely rejecting H_0)

→ Type-I rate, a.k.a. the ***p-value :*** $p_0 = \int_{q_0^{obs}}^{\infty} f(q_0|S=0)\, dq_0$

= *Fraction of outcomes that are*

at least as extreme (*signal-like*) ***as data***, *when* ***H_0 is true*** (*no signal*).

Asymptotic distribution of q_0

Cowan, Cranmer, Gross & Vitells
Eur.Phys.J.C71:1554,2011

Gaussian regime for $\hat{S}$ (e.g. large n_{evts}, Central-limit theorem) :

Wilk's Theorem (*) : for S = 0

q_0 is distributed as χ^2 (n_{par})

⇒ n_{par} = 1 : $\sqrt{q_0}$ is distributed as a Gaussian

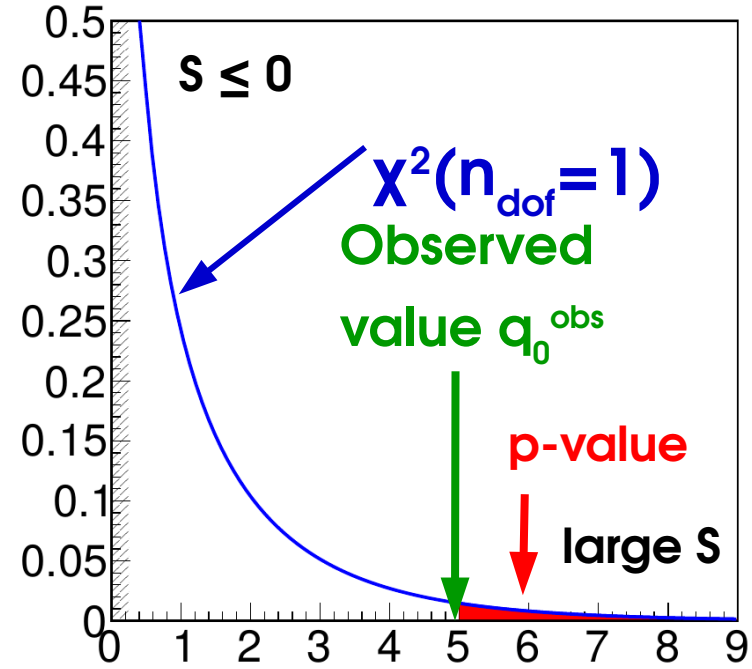
⇒ Can compute p-values from Gaussian quantiles

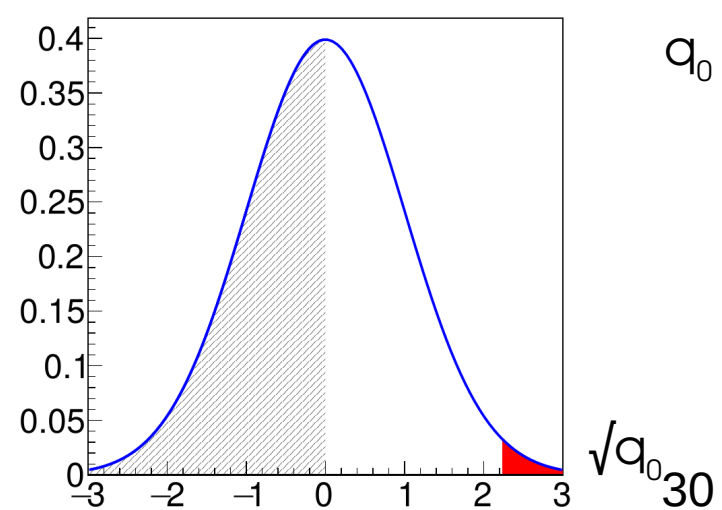
$$p_0 = 1 - \Phi\left(\sqrt{q_0}\right)$$

⇒ Even more simply, the significance is:

$$Z = \sqrt{q_0}$$

Typically works well already for for event counts of O(5) and above ⇒ Widely applicable





(*) 1-line "proof" : asymptotically L and S are Gaussian, so

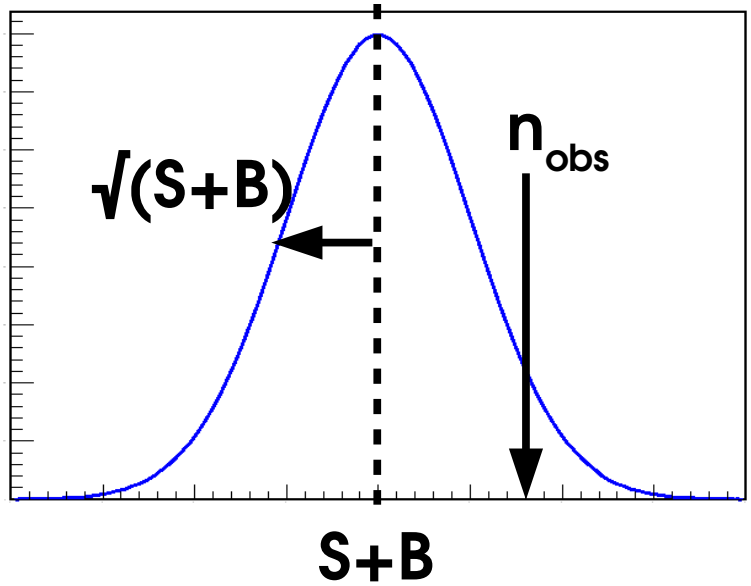
$$L(S) = \exp\left[-\frac{1}{2}\left(\frac{S-\hat{S}}{\sigma}\right)^2\right] \Rightarrow q_0 = \left(\frac{\hat{S}}{\sigma}\right)^2 \Rightarrow \sqrt{q_0} = \frac{\hat{S}}{\sigma} \sim G(0,1) \Rightarrow q_0 \sim \chi^2(n_{dof}=1)$$

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

→ Find the best-fit value (MLE) $\hat{S}$ for the signal

(can use $\lambda = -2 \log L$ instead of L for simplicity)

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

→ Find the expression for the significance

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

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

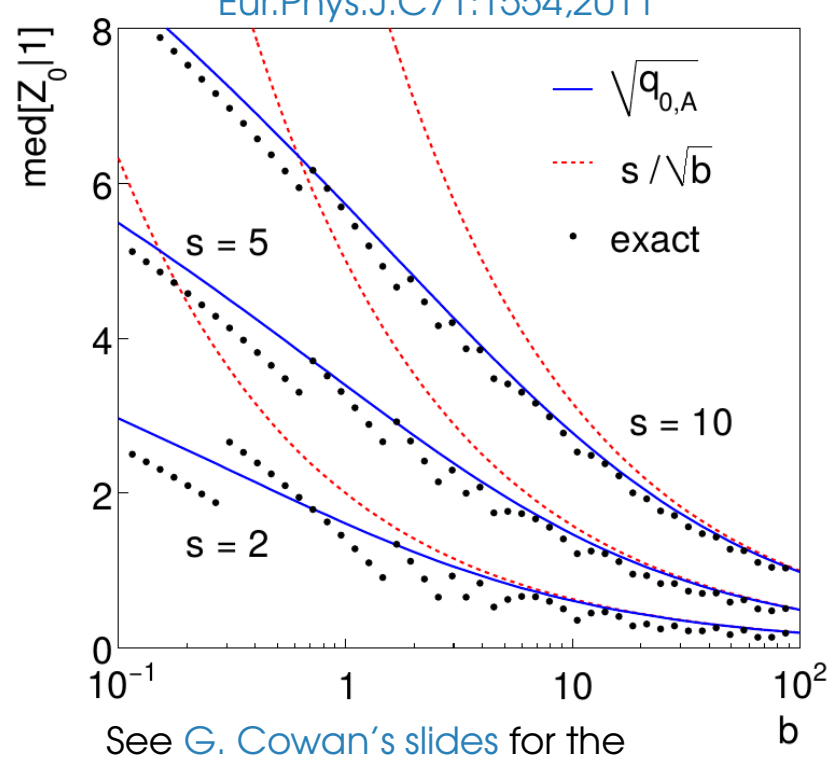
Solution:

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

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!)

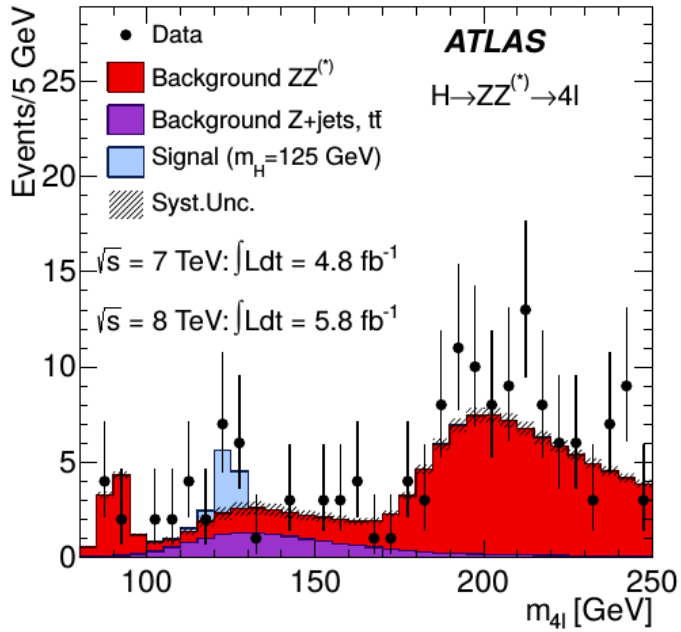
Eur.Phys.J.C71:1554,2011

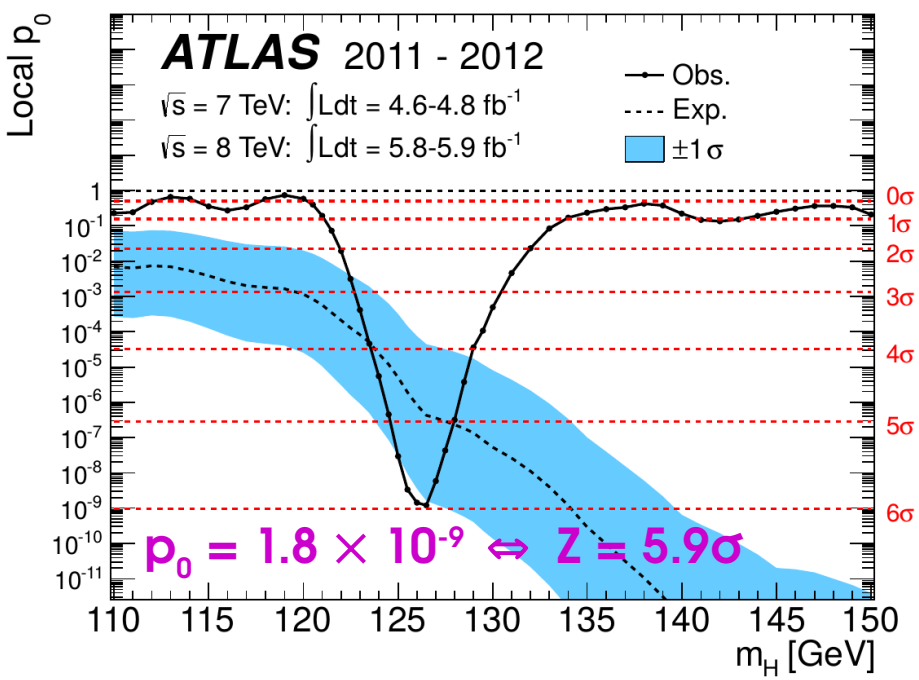


See G. Cowan's slides for the case with B uncertainty

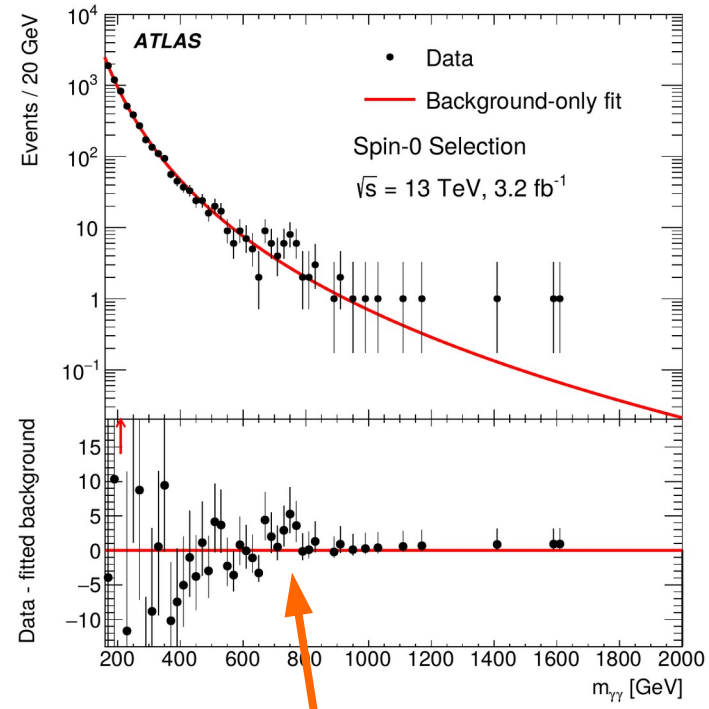
Some Examples

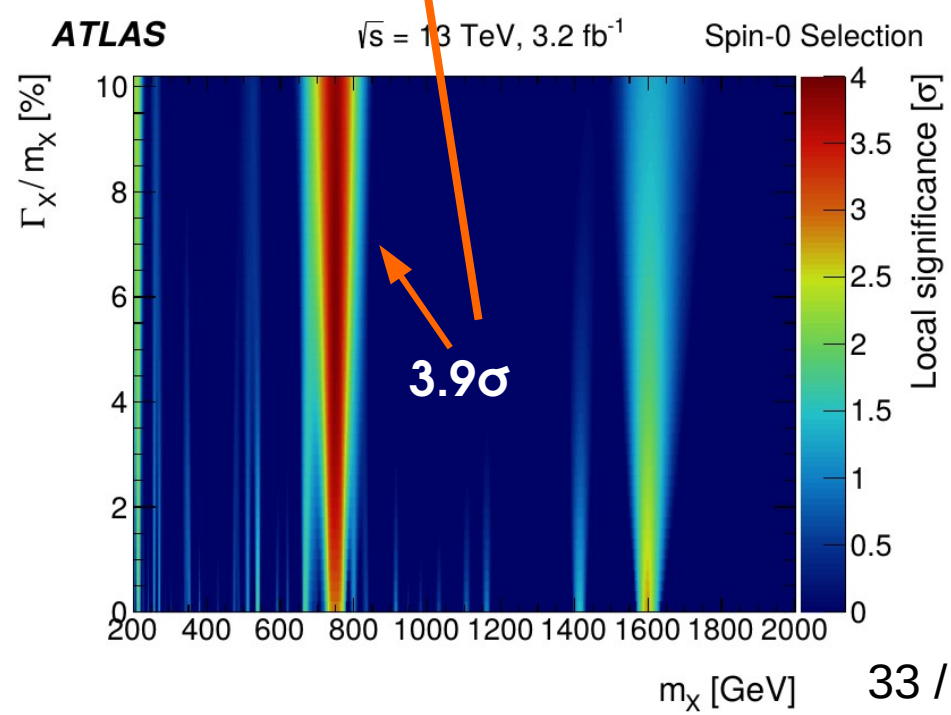
Higgs Discovery: Phys. Lett. B 716 (2012) 1-29





High-mass X→γγ Search: JHEP 09 (2016) 1





Discovery Thresholds

Evidence : **3σ ⇔ p_0 = 0.3% ⇔ 1 chance in 300**

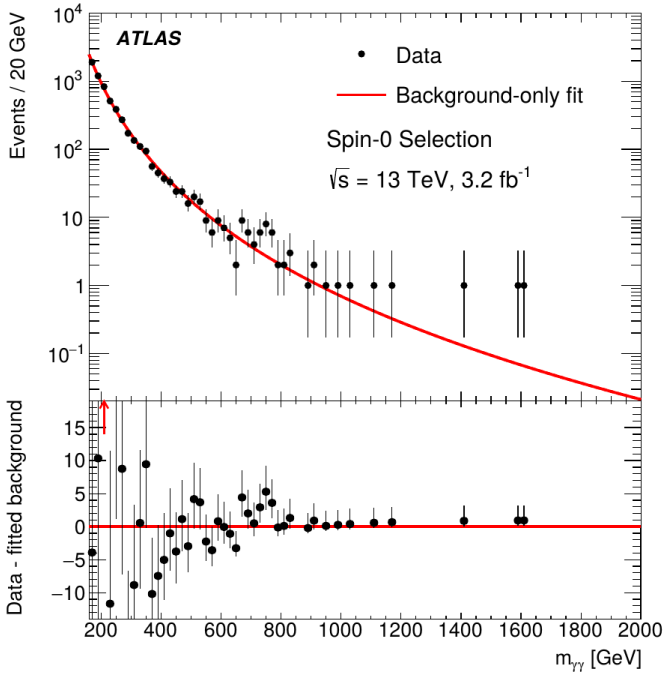
Discovery: **5σ ⇔ p_0 = 3 10-7 ⇔ 1 chance in 3.5M**

Why so high thresholds ? (from Louis Lyons):

- **Look-elsewhere effect** : searches typically cover multiple independent regions ⇒ Higher chance to have a fluctuation "somewhere"

  N_{trials} ~ 1000 : local 5σ ⇔ O(10^{-4}) more reasonable

- **Mismodeled systematics**: factor 2 error in syst-dominated analysis ⇒ factor 2 error on Z...



- **History**: 3σ and 4σ excesses do occur regularly, for the reasons above

Extraordinary claims require extraordinary evidence!

Takeaways

Given a statistical model P(data; μ), define likelihood **L(μ) = P(data; μ)**

To estimate a parameter, use the value **$\hat{\mu}$** that maximizes L(μ) → best-fit value

To decide between hypotheses H_0 and H_1, use the **likelihood ratio** $\frac{L(H_0)}{L(H_1)}$

To test for **discovery**, use $q_0 = -2\log\frac{L(S=0)}{L(\hat{S})} \quad \hat{S} \geq 0$

For large enough datasets (n >~ 5), $Z = \sqrt{q_0}$

For a **Gaussian** measurement, $Z = \frac{\hat{S}}{\sqrt{B}}$

For a **Poisson** measurement, $Z = \sqrt{2\left[(\hat{S}+B)\log\left(1+\frac{\hat{S}}{B}\right) - \hat{S}\right]}$