Practical Statistics for LHC Physicists

Descriptive Statistics, Probability and Likelihood

Harrison B. Prosper

Florida State University

CERN Academic Training Lectures

7 April, 2015

Outline

- Lecture 1
  - Descriptive Statistics
  - Probability & Likelihood
- Lecture 2
  - Frequentist Inference
- Lecture 3
  - Bayesian Inference

Descriptive Statistics

Descriptive Statistics: Samples

Definition: A **statistic** is any function of the data,

x = $x_1, x_2, \ldots x_n$. Here are some simple examples:

the **sample average** $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

the **sample moments** $$m_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

and the **sample variance** $$S^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Descriptive Statistics: Samples

It is often useful to order the data so that

$x_{(1)} < x_{(2)} < \ldots < x_{(n)}$

$x_{(k)}$ is called the kth order statistic

$x_{(k)}$ is also the α-quantile, where $\alpha = k / n$

If $\alpha = 0.5$, then $x_{(k)}$ is called the median.

All of these quantities, and many more, can be computed because the sample is *known*.

Descriptive Statistics: Populations

Now consider an *infinitely* large sample, called a population.

This is clearly an *abstraction*, which exists only in the sense that the set of real numbers exist.

Like many abstractions, however, we can study this one mathematically.

But, since all we have is a sample, we need a way to connect it to its associated population. One goal of a theory of statistical inference is to use a sample to say something about its associated population.

Descriptive Statistics: Populations

Expected Value	$E[x]$
Mean	μ
Error	$\varepsilon = x - \mu$
Mean Square Error	$\mathrm{MSE} = E[\varepsilon^2]$
Bias	$b = E[x] - \mu$
Variance	$V[x] = E[(x - E[x])^2]$

Descriptive Statistics – 3

$$\text{MSE} = E[\varepsilon^2]$$
$$= V + b^2$$

Exercise 1: Show this

The **MSE** is the most widely used measure of how close an ensemble of statistics {**x**} is to the mean (or true value) $\boldsymbol{\mu}$.

The **root mean square** (RMS) is

$$\text{RMS} = \sqrt{\text{MSE}}$$

Descriptive Statistics – 4

Consider the expected value of the *sample* **variance**

$$E[S^2] = E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2]$$

$$= E[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2}{n}\sum_{i=1}^{n} x_i\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[x_i^2] - E[\bar{x}^2]$$

$$= E[x^2] - E[\bar{x}^2]$$

Descriptive Statistics – 5

The expected value of the sample variance (as we have defined it) is biased

$$E[S^2] = E[x^2] - E[\bar{x}^2]$$

$$= E[x^2] - \frac{1}{n}E[x^2] - \left(\frac{n-1}{n}\right)E[x]^2$$

$$= V[x]\left(\frac{n-1}{n}\right)$$

The bias is $-V/n$

Exercise 2:
Show this

Probability

Probability – 1

Objects

1. Sample space: the set S of outcomes of an experiment
2. Event: a subset E of S*
3. Function: P associates a real number to E

Rules (Kolmogorov Axioms)

1. $P(E) \geq 0$
2. $P(S) = 1$
3. $P(E_1 + E_2 + \ldots) = P(E_1) + P(E_2) + \ldots$ where $E_i E_j = \emptyset$

and the rules of Boolean algebra.

* With a technical restriction on the collection of subsets of S

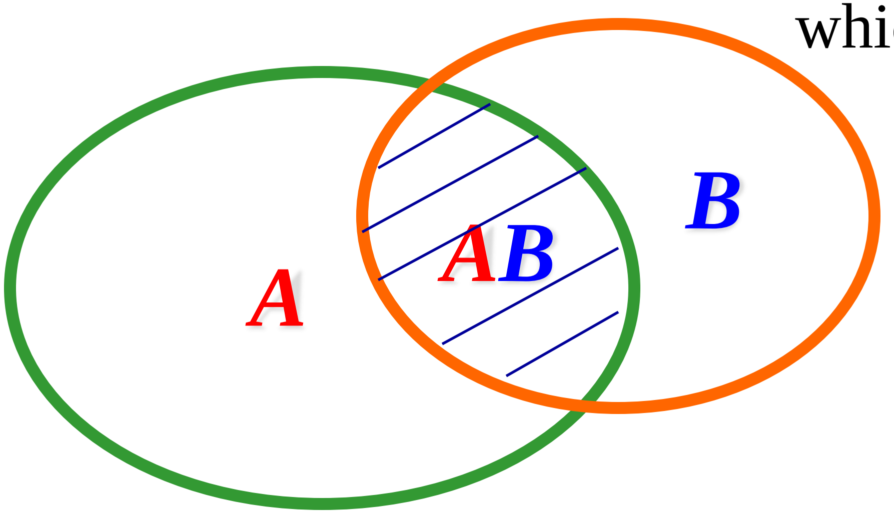
Probability – 2

By definition, the **conditional probability** of A *given* B is

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

$P(B)$ is the probability of B *without the restriction* imposed by A.

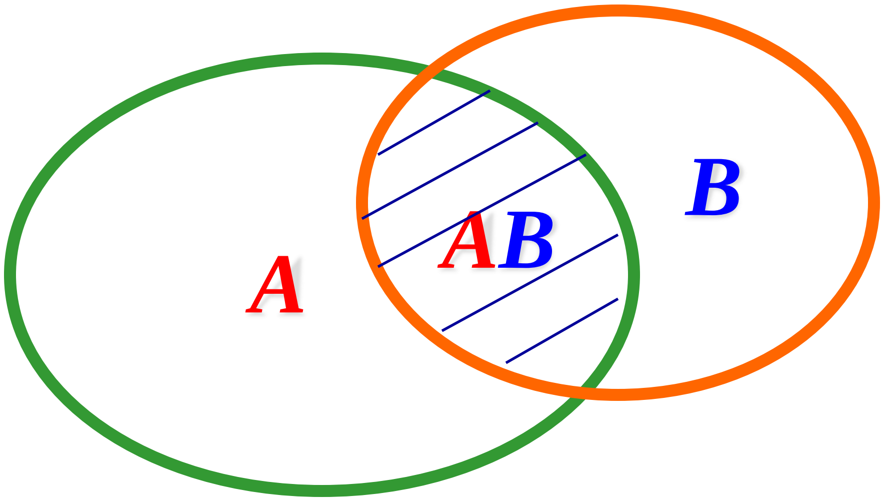
$P(A|B)$ is the probability of A when we *restrict* to the conditions under which B is true.



Probability – 3

A and B are mutually exclusive if

$$P(AB) = 0$$



A and B are exhaustive if

$$P(A) + P(B) = 1$$

Theorem

$$P(A+B) = P(A) + P(B) - P(AB)$$

Exercise 3: Prove theorem

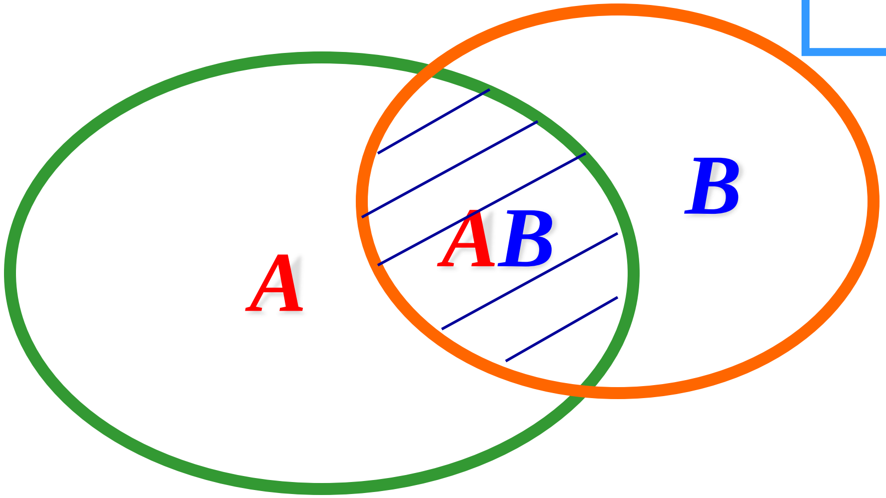
Probability – 4

By definition: $P(AB) = P(A \mid B)P(B)$

$$P(BA) = P(B \mid A)P(A)$$

But, since AND commutes, i.e., $AB = BA$, we immediately deduce Bayes Theorem:

$$\boxed{P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}}$$



Bayes Theorem: Are You Doomed?

Diagnostic Example (Michael Goldstein)

You are Diseased (event D)

You are Healthy (event H)

A test result is either positive (event +) or negative (event –)

Let $P(+ \mid D) = 0.99$ and $P(+ \mid H) = 0.01$.

Your test result is positive. Are you doomed? It all depends…

Suppose the incidence of disease is 1 in a 1000, i.e., $P(D) = 0.001$, then Bayes theorem yields

$$
\begin{aligned}
P(D \mid +) &= P(+ \mid D)\, P(D) \,/\, P(+) \\
&= P(+ \mid D)\, P(D) \,/\, [P(+ \mid D)\, P(D) + P(- \mid H)\, P(H)] \\
&= 0.09
\end{aligned}
$$

Probability: Some Definitions

Suppose we have some function $f(x)$, for example,

$$f(x) = x,$$

$$f(x) = (x - \mu)^2$$

then its expected value is the functional

$$E[f] = \sum_i f(x_i) P(x_i)$$

If x is continuous, this becomes

$$E[f] = \int f(x)\, dP(x) = \int f(x)\, p(x)\, dx$$

$p(\mathrm{x}) = dP / dx$ is called a probability density function (pdf).

Probability: Some Definitions

Suppose we have potential observations (random variables) x and y, then their covariance is the functional

$$Cov[f,g] = \iint f(x)g(y)\,p(x,y)\,dx\,dy$$

where

$f(x) = x - E[x]$ and $g(y) = y - E[y]$ and

$p(x, y)$ is the joint probability density of x and y.

If we can write $p(x, y) = p(x)\,p(y)$ then x and y are said to be statistically independent, in which case $Cov[f, g] = 0$. But note that, in general, $Cov[f, g] = 0$ does *not* imply statistical independence.

Probability: What Exactly Is It?

There are at least two *interpretations* of probability:

1. **Degree of belief** in, or assigned to, a proposition, e.g.,

   "A tsunami will flood Geneva tomorrow"

2. **Relative frequency** of outcomes in an *infinite* sequence of trials, e.g.,

   proton-proton collisions at the LHC with

   outcome the creation of Higgs bosons.

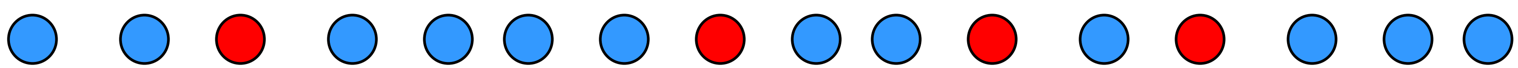
Binomial & Poisson Distributions

Binomial & Poisson Distributions – 1

A **Bernoulli** trial has two outcomes:

S = success or ***F*** = failure.

Example: Each collision between protons at the LHC is a Bernoulli trial in which either something interesting happens (***S***) or does not (***F***).

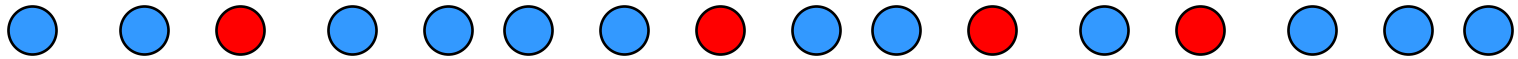


Binomial & Poisson Distributions – 2

Let ***p*** be the probability of a success, which is assumed to be the *same at each trial*. Since ***S*** and ***F*** are exhaustive, the probability of a failure is **1 –** ***p***.

For a given order ***O*** of ***n*** trails, the probability $P(k, O, n)$ of *exactly* ***k*** successes and ***n*** **–** ***k*** failures is

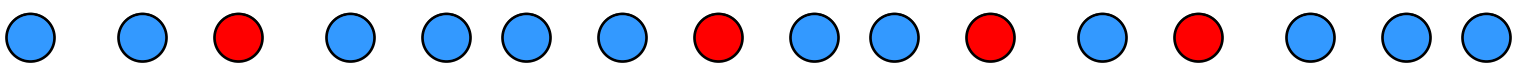
$$P(k,O,n) = p^k(1-p)^{n-k}$$



Binomial & Poisson Distributions – 3

If the order ***O*** of successes and failures is assumed to be irrelevant, we can eliminate the order from the problem by *summing* over all possible orders

$$P(k,n) = \sum_{O} P(k,O,n) = \sum_{O} p^{k}(1-p)^{n-k}$$



This yields the **binomial distribution**

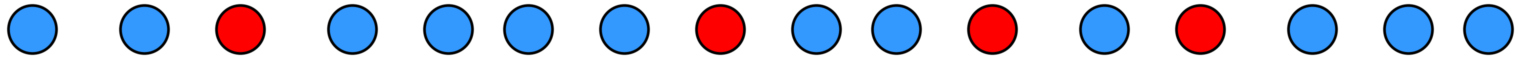
$$P(k,n) = \text{Binomial}(k,n,p) \equiv \binom{n}{k} p^{k}(1-p)^{n-k}$$

Binomial & Poisson Distributions – 4

We can prove that the mean number of successes ***a*** is

$\boldsymbol{a} = \boldsymbol{p}\,\boldsymbol{n}$. **Exercise 4**: Prove it

Suppose that the probability, ***p***, of a success is very small,



then, in the limit $\boldsymbol{p} \to 0$ and $\boldsymbol{n} \to \infty$, such that ***a*** is *constant*,

Binomial$(k, n, p) \to$ **Poisson**(k, a).

The Poisson distribution is generally regarded as a good model for a **counting experiment**

Exercise 5: Show that Binomial$(k, n, p) \to$ Poisson(k, a)

Common Densities and Distributions

$\mathrm{Uniform}(x,a)$	$1/a$
$\mathrm{Gaussian}(x,\mu,\sigma)$	$\exp[-(x-\mu)^2/(2\sigma^2)]/(\sigma\sqrt{2\pi})$
$\mathrm{LogNormal}(x,\mu,\sigma)$	$\exp[-(\ln x-\mu)^2/(2\sigma^2)]/(x\sigma\sqrt{2\pi})$
$\mathrm{Chisq}(x,n)$	$x^{n/2-1}\exp(-x/2)/[2^{n/2}\Gamma(n/2)]$
$\mathrm{Gamma}(x,a,b)$	$x^{b-1}a^b\exp(-ax)/\Gamma(b)$
$\mathrm{Exp}(x,a)$	$a\exp(-ax)$
$\mathrm{Binomial}(k,n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$
$\mathrm{Poisson}(k,a)$	$a^k\exp(-a)/k!$
$\mathrm{Multinomial}(k,n,p)$	$\frac{n!}{k_1!\ldots k_K!}\prod_{i=1}^{K}p_i^{k_i},\ \sum_{i=1}^{K}p_i=1,\ \sum_{i=1}^{K}k_i=n$

Likelihood

Likelihood – 1

The likelihood function is simply the probability, or probability density function (pdf), evaluated at the *observed* data.

Example 1: Evidence for electroweak production of $W^{\pm}W^{\pm}jj$ (ATLAS, PRL 113, 141803 (2014))

$p(D|\, d) = \text{Poisson}(D\,|d)$ *probability* to observe a count D

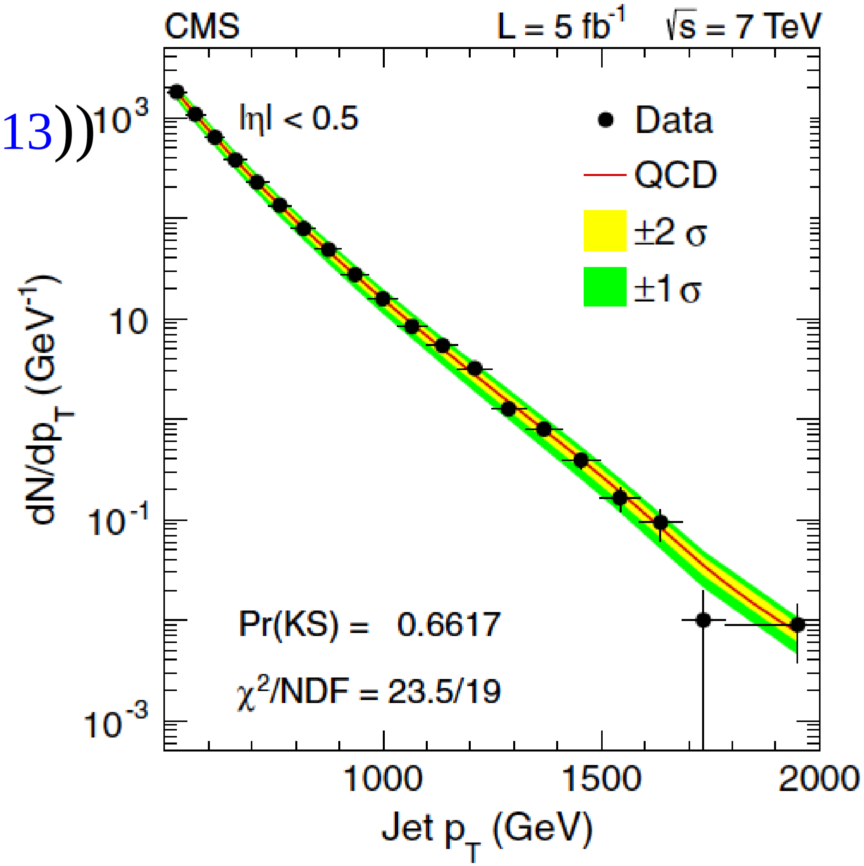
$p(12|d) = \text{Poisson}(12|d)$ *likelihood* of observation $D = 12$

where $d = E[D]$ is the expected count.

Likelihood – 2

Example 2:

(CMS, Phys. Rev. D 87, 052017 (2013))



Observed counts D_i

$$p(D \mid p) = \text{Multinomial}(D, N, p)$$

$$D = D_1, \ldots, D_K, \quad p = p_1, \ldots, p_K$$

$$\sum_{i=1}^{K} D_i = N, \quad \sum_{i=1}^{K} p_i = 1$$

This is an example of a *binned* likelihood

PHYSICAL REVIEW D **87**, 052017 (2013)

Search for contact interactions using the inclusive jet p_T spectrum in pp collisions at $\sqrt{s} = 7$ TeV

S. Chatrchyan *et al.**
(CMS Collaboration)
(Received 21 January 2013; published 26 March 2013)

Likelihood – 3

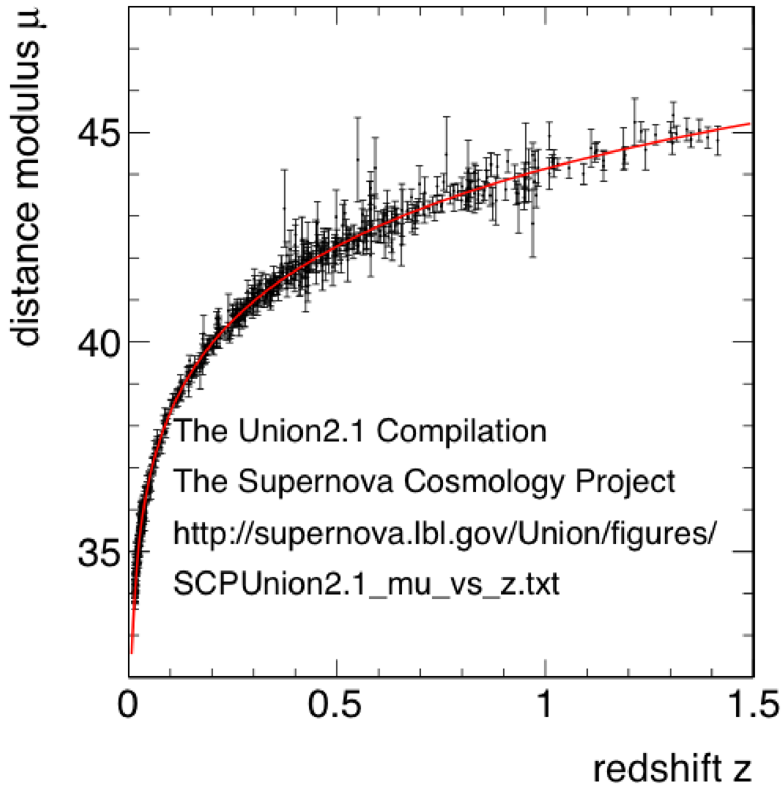
Example 3: (Union2.1 Compilation, SCP)

Red shift and distance modulus measurements of $N = 580$ Type Ia supernovae

$$p(D \mid \Omega_M, \Omega_\Delta, Q) =$$

$$\prod_{i=1}^{N} \text{Gaussian}(x_i, \mu(z_i, \Omega_M, \Omega_\Delta, Q), \sigma_i)$$

$$D = z_i, x_i \pm \sigma_i$$



This is an example of an *un-binned* likelihood for *heteroscedastic* data.

Likelihood – 4

Example 4: Higgs to γγ (CMS & ATLAS, 2012 – 15)

The analyses of the di-photon final states use an *un-binned* likelihood of the form,

$$p(x \mid s, m, w, b) = \exp[-(s+b)] \prod_{i=1}^{N} \left[s\, f_s(x_i, m, w) + b\, f_b(x_i) \right]$$

where x = measured di-photon masses

m = mass of particle

w = expected width

s = expected signal

b = expected background

f_s = signal model

f_b = background model

Exercise 6: Show that a binned multi-Poisson likelihood yields an un-binned likelihood of this form as the bin widths go to zero

Likelihood – 5

Given the likelihood function, we can answer several questions including:

1. How do I estimate a parameter?
2. How do I quantify its accuracy?
3. How do I test an hypothesis?
4. How do I quantify the significance of a result?

Writing down the likelihood function requires:

1. Identifying all that is *known*, e.g., the observations
2. Identifying all that is *unknown*, e.g., the parameters
3. Constructing a probability model *for both*

Example 1: $W^\pm W^\pm jj$ Production (ATLAS)

Evidence for electroweak production of $W^\pm W^\pm jj$ (2014)

PRL 113, 141803 (2014)

knowns:

$D = 12$ observed events ($\mu^\pm\mu^\pm$ mode)

$B = 3.0 \pm 0.6$ background events

unknowns:

b expected background count

s expected signal count

$d = b + s$ expected event count

Note: we are uncertain about *unknowns*, so **12 ± 3.5** is a statement about d, *not about the observed count* **12**!

Example 1: $W^{\pm}W^{\pm}jj$ Production (ATLAS)

Probability:

$$p(D \mid s, b) = \text{Poisson}(D, s+b)\ \text{Poisson}(Q, bk)$$

$$= \frac{(s+b)^D e^{-(s+b)}}{D!} \frac{(bk)^Q e^{-bk}}{\Gamma(Q+1)}$$

Likelihood:

$$p(12 \mid s, b)$$

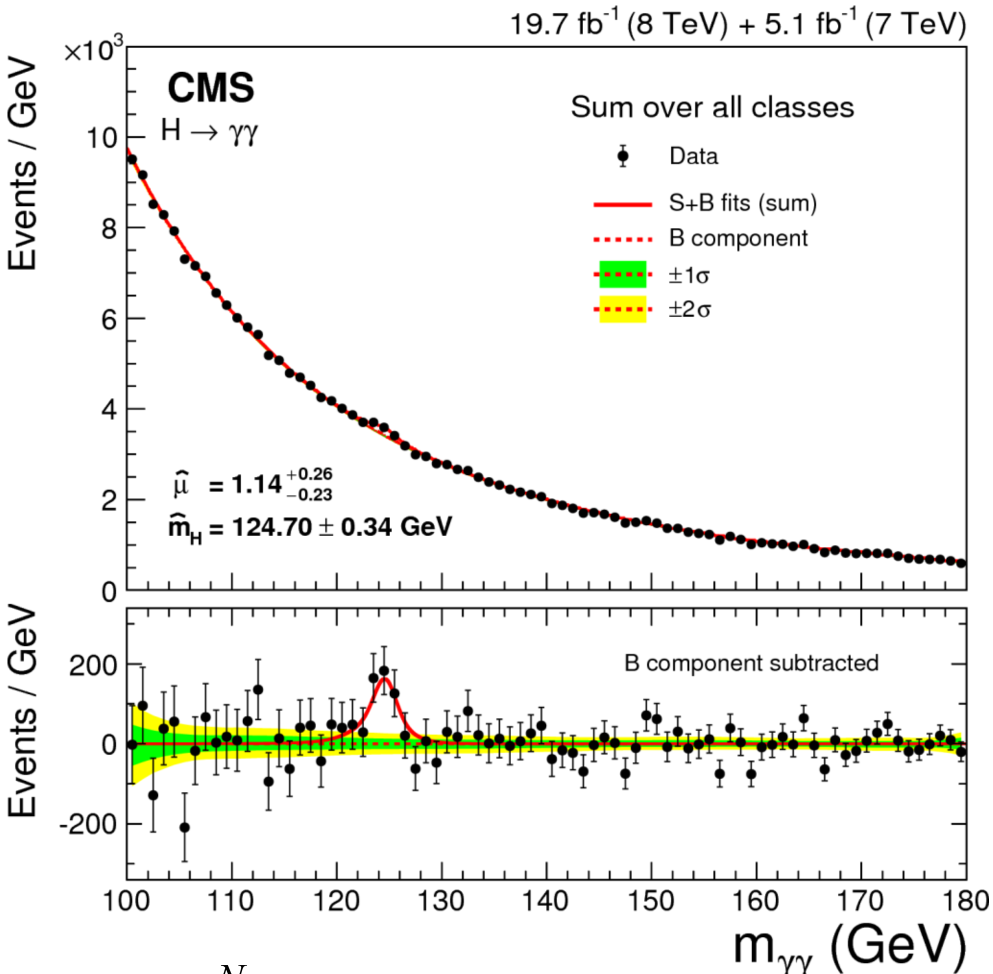
where

$$B \quad = Q / k \qquad\qquad Q = (B / \delta B)^2 = (3.0 / 0.6)^2 = 25.0$$

$$\delta B \quad = \sqrt{Q} / k \qquad\qquad k = B / \delta B^2 = 3.0 / 0.6^2 = 8.33$$

Example 4: Higgs to γγ (CMS)

Eur.Phys.J. C74 (2014) 3076



background model

$$f_b(x,a), \quad x = m_{\gamma\gamma}$$

signal model

$$f_s(x \,|\, m, w)$$

$$p(x \,|\, s, m, w, b, a) = \exp[-(s+b)] \prod_{i=1}^{N} \left[s\, f_s(x_i, m, w) + b\, f_b(x_i, a) \right]$$

Example 4: Higgs to γγ (CMS)

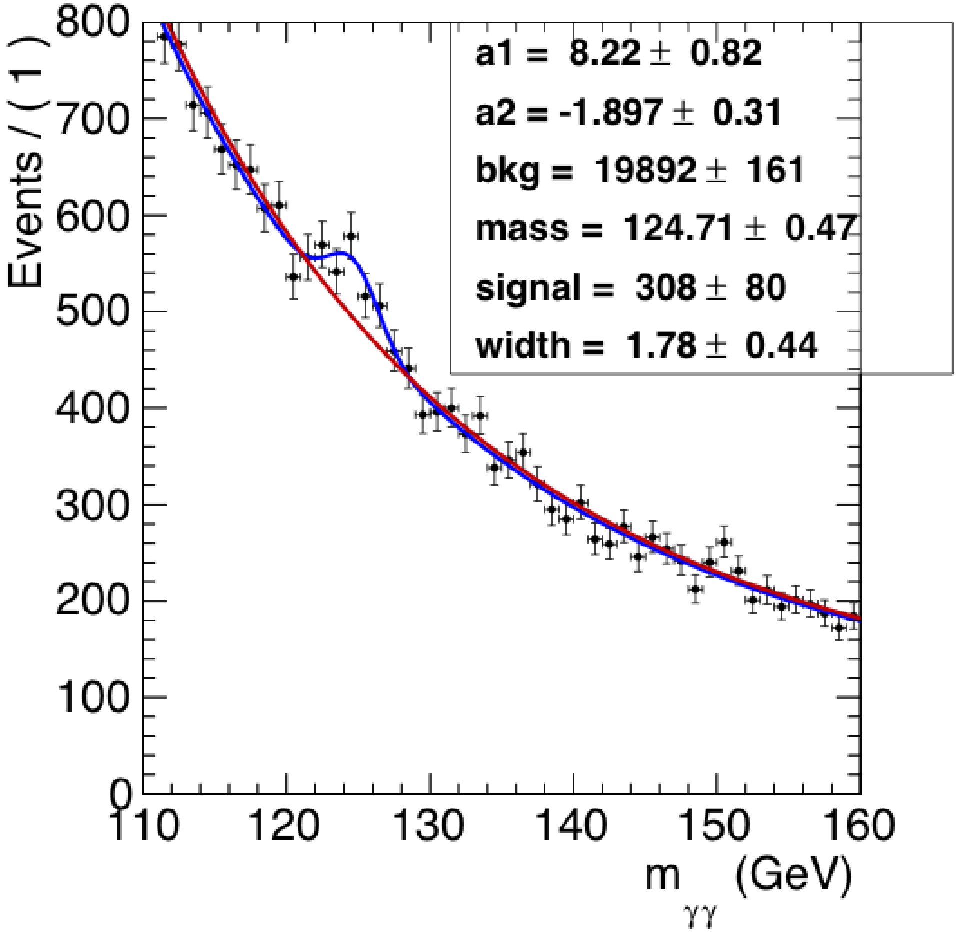
Tomorrow, we shall study a toy version of the likelihood:

background model

$$f_b(x,a) = A\exp[-(a_1 x + a_2 x^2)]$$

signal model

$$f_s(x,m,w) = \text{Gaussian}(x, m, w)$$



$$p(x \mid s,m,w,b,a) = \exp[-(s+b)]\prod_{i=1}^{N}\left[s f_s(x_i,m,w) + b f_b(x_i,a)\right]$$

Summary

Statistic

A statistic is *any* calculable function of potential observations

Probability

Probability is an *abstraction* that must be interpreted

Likelihood

The likelihood is the probability (or probability density) of potential observations *evaluated at the observed data*