Bootstrap Method

- From the bootstrap dataset Z_i, calculate the parameters of interest θ_i (it can be mean, median, parameters for linear regression, etc.). For example, for the linear regression case, $\theta_i = \{a_i, b_i\}$ which can be calculated using maximum likelihood method or least-square.

- The two steps before are repeated a large number of times, say $i = 1, \ldots, N_{boot}$.

- At the end of the process, the parameter values θ_i are used to approximate the sampling distribution of the parameters, and therefore obtain an estimate of the best-fit values and confidence intervals. We will be able to visualize the histograms of $\hat{\theta} = \{\hat{a}, \hat{b}\}$ and make inferences.

$$\bar{\theta} = \frac{1}{N_{boot}} \sum_{i=1}^{N_{boot}} \theta_i$$

$$s_\theta^2 = \frac{1}{N_{boot}} \sum_{i=1}^{N_{boot}} (\theta_i - \bar{\theta})^2$$

Bootstrapping is not 100% free of bias (particularly with small sample sizes). Bootstrapping has many variations and extensions to mitigate bias have been proposed.

Such re-sampling methods can be used even when the errors on the data points are not available!

Maximum Likelihood

Normally distributed measurements are very common in data analysis. In this case, data are usually reported in the form

$$(x_i, y_i \pm \sigma_i) \text{ with } i = 1, \dots, N$$

with the meaning that the Gaussian-distributed Y variable was estimated to have a mean of y_i and a variance of σ_i^2.

Suppose that the true value is given as a function x as $\mu = f\left(x, \{\vec{\theta}\}\right)$ which depends on unknown parameters $\{\theta_0, \theta_1, \dots, \theta_m\}$. The main of least squares is to estimate the parameters $\{\vec{\theta}\}$ and the basis of the method is founded in the maximum likelihood principle:

$$\mathcal{L}(y_1, \dots y_N, \mu_1, \dots, \mu_N, \sigma_1, \dots \sigma_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{\left(y_i - f\left(x_i, \{\vec{\theta}\}\right)\right)^2}{2\sigma_i^2}\right\}$$

Joint Probability Density Function (PDF) of N Gaussians

Maximum Likelihood

Taking the logarithm of the joint PDF and neglecting the additive terms that do not depend on the parameters $\{\vec{\theta}\}$, we get the **log-likelihood**:

$$\log \mathcal{L}(\{\vec{\theta}\}) = -\frac{1}{2}\sum_{i=1}^{N} \frac{\left(y_i - f(x_i, \{\vec{\theta}\})\right)^2}{\sigma_i^2}$$

This is maximized by finding the parameters $\{\vec{\theta}\}$ that minimize the quantity

$$\chi^2(\{\vec{\theta}\}) = \sum_{i=1}^{N} \frac{\left(y_i - f(x_i, \{\vec{\theta}\})\right)^2}{\sigma_i^2}$$

What is the relation with the so-called Chi-square distribution?

Chi-Square Distribution

Given a dataset consisting of N points (x_i, y_i), $i = 1, \ldots, N$ and a model $f\left(x|\vec{\theta}\right)$, dependent on n unknown parameters $\vec{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$, one can evaluate the goodness of a fit based on the χ^2 function defined as

$$\chi^2 = \sum_{i=1}^{N} \frac{\left[y_i - f\left(x|\vec{\theta}\right)\right]^2}{\sigma_i^2}$$

where σ_i^2 are the errors on the n measurements y_i. **For measurements governed by Gaussian statistics, that is, when repeated measurements at a given value of x yield values of y distributed according to a normal distribution, the probability density function of the χ^2 values is given by the χ^2-distribution** defined as

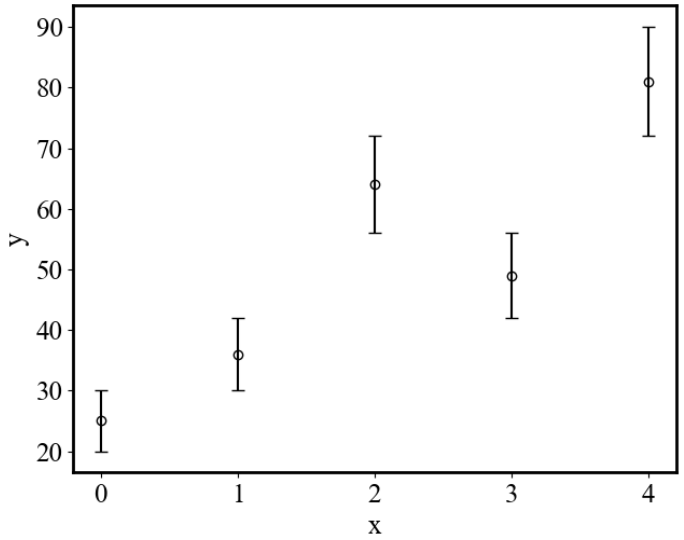
$$p_{\chi^2}(z|k) = \frac{1}{2^{k/2}\Gamma(k/2)} z^{k/2-1} e^{-z/2} \quad \text{for } k = 1,2,3, \ldots$$

in which z is a continuous random variable and k is an integer called **number of degrees of freedom**. This number is a measure of the number of independent variables in a system or model, i.e., N residuals that depend on $\vec{\theta}$ that can be rearranged into a system of equations (derivatives to find the χ^2 minimum) that can be solved to determine n predictors. In this way, the number of degrees of freedom of a dataset consisting of N random values fitted to a model involving n parameters is $k = N - n$.

Pruneau, Data Analysis Techniques for Physical Sciences, Cambridge

Chi-Square Distribution

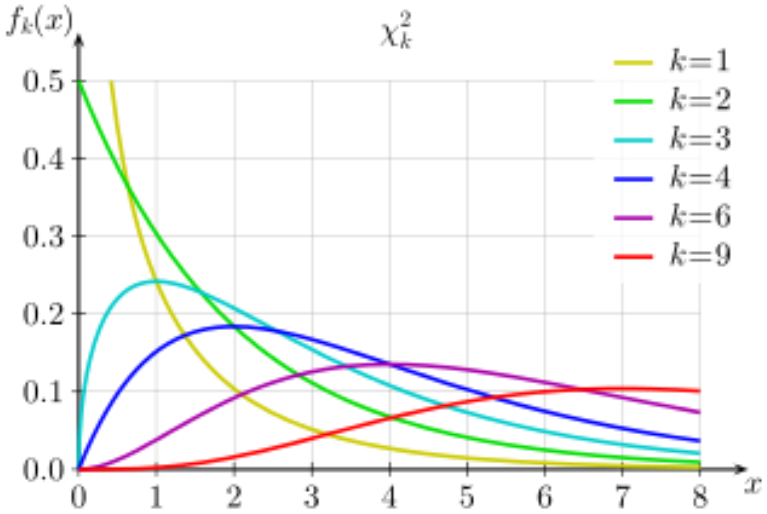
Example



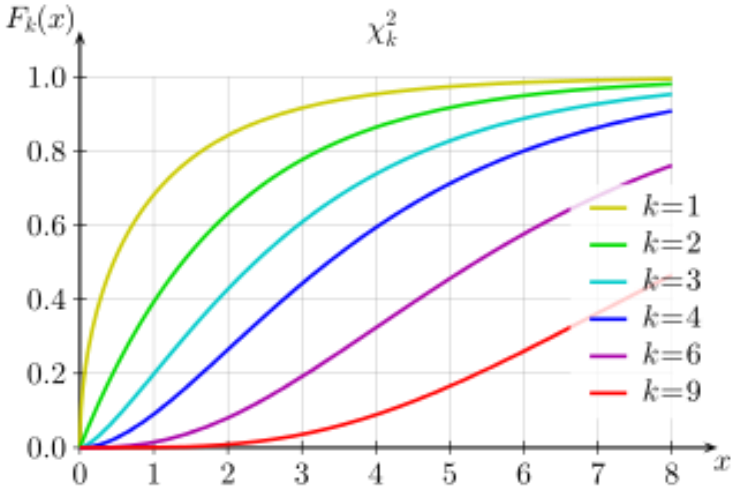
All distributions exhibit a peak near k!

This translates in having a value of χ^2 per degree of freedom near the unity!

Probability density function



Cumulative density function



The x versus y plot above can be seen as the points being the mean of several measurements of y and the error bars the standard deviation. But we can plot the individual measurements in separated plots and obtain multiple χ^2_{min}.

Given the uncertainties in y_i, we can imagine that we have many distinct x versus y plots with the points falling within the uncertainty from which we can obtain many χ^2_{min}. The latter is chi-distributed!

Chi-Square Distribution
Other properties

Characteristic function: $(1-2it)^{k/2}$

Mode: $\max(k-2,0)$

Median: $\approx k(1-2/9k)^3$

Mean: $\mu = k$

Variance: $\sigma^2 = 2k$

Skewness excess: $\gamma_1 = \sqrt{8/k}$

Kurtosis excess: $\gamma_2 = 12/k$

$$\chi^2(\{\vec{\theta}\}) = \sum_{i=1}^{N} \frac{\left(y_i - f(x_i, \{\vec{\theta}\})\right)^2}{\sigma_i^2}$$

Let's now expand χ^2 around the minimum up to second order…

Maximum Likelihood

$$\chi^2(\vec{\theta}) = \chi^2_{min}(\vec{\theta}_{best}) + \frac{1}{2}(\vec{\theta} - \vec{\theta}_{best})^T \underbrace{\{D^2\chi^2_{min}(\vec{\theta})\}}_{\text{Hessian matrix}}(\vec{\theta} - \vec{\theta}_{best})$$

2 parameters case example:

$$\chi^2(\{a,b\}) = \chi^2_{min}(\{a_{best}, b_{best}\}) + \frac{1}{2}(a - a_{best} \quad b - b_{best}) \begin{pmatrix} \frac{\partial^2\chi^2_{min}}{\partial a^2} & \frac{\partial^2\chi^2_{min}}{\partial a \partial b} \\ \frac{\partial^2\chi^2_{min}}{\partial b \partial a} & \frac{\partial^2\chi^2_{min}}{\partial b^2} \end{pmatrix} \begin{pmatrix} a - a_{best} \\ b - b_{best} \end{pmatrix}$$

Given that we have found the minimum, we can study how χ^2 changes from the minimum and the shape of the surface. **Wilks' Theorem:** the uncertainty of a parameter can be obtained by taking $\Delta\chi^2 = \chi^2(\theta) - \chi^2_{min} \sim \chi^2(p)$ (from the minimum) with p being the number of free parameters in the model. From the Wilks' Theorem, it is possible to obtain

$$\sigma^2_{ij} = \left(-\frac{\partial^2 \log \mathcal{L}}{\partial\theta_i \partial\theta_j}\right)^{-1} = 2\left(\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\right)^{-1}$$

Wilks Theorem

For one parameter, the 1-sigma uncertainty (68.3% confidence) corresponds to:

$$\Delta\chi^2 = 1$$

For two parameters, the 1-sigma uncertainty (68.3% confidence) corresponds to:

$$\Delta\chi^2 = 2.30$$

For three parameters, the 1-sigma uncertainty (68.3% confidence) corresponds to:

$$\Delta\chi^2 = 3.53$$

Generalizing…

$$\Delta\chi^2(p, 68.3\%) \sim \chi^2_{p,1\sigma}$$

Maximum Likelihood

If the parameters are uncorrelated, we will have:

$$\chi^2(\{a,b\}) = \chi^2_{min}(\{a_{best}, b_{best}\}) + \frac{1}{2}(a - a_{best} \quad b - b_{best}) \begin{pmatrix} \frac{\partial^2 \chi^2_{min}}{\partial a^2} & 0 \\ 0 & \frac{\partial^2 \chi^2_{min}}{\partial b^2} \end{pmatrix} \begin{pmatrix} a - a_{best} \\ b - b_{best} \end{pmatrix}$$

$$\chi^2(\{a,b\}) = \chi^2_{min} + \frac{1}{2}\left((\Delta a)^2 \frac{\partial^2 \chi^2_{min}}{\partial a^2} + (\Delta b)^2 \frac{\partial^2 \chi^2_{min}}{\partial b^2}\right) = \chi^2_{min} + \left(\frac{(\Delta a)^2}{\sigma_a^2} + \frac{(\Delta b)^2}{\sigma_b^2}\right)$$ **Ellipse**

Maximum Likelihood

If $\frac{\partial^2 \chi^2_{min}}{\partial a \partial b} \neq 0$, that indicates correlation between the parameters. With uncorrelated parameters, we had a total uncertainty from the χ^2 given by

$$\sigma_{tot}^2 = \sigma_a^2 + \sigma_b^2$$

For correlated parameters, we have the full ellipse form

$$\sigma_{tot}^2 = \sigma_a^2 + \sigma_b^2 + 2\sigma_{ab}$$

with $\sigma_{ab} = COV(a,b)$. The error matrix in 2D can then be written as

$$\begin{pmatrix} \sigma_a^2 & COV(a,b) \\ COV(a,b) & \sigma_b^2 \end{pmatrix} = \sum_{i=1}^{N} \begin{pmatrix} (a_i - \bar{a})^2 & (a_i - \bar{a})(b_i - \bar{b}) \\ (a_i - \bar{a})(b_i - \bar{b}) & (b_i - \bar{b})^2 \end{pmatrix}$$

We can also write the matrix above as the *correlation matrix* where we normalize its terms by the uncertainties:

Maximum Likelihood

Normalizing…

$$\begin{pmatrix} \frac{1}{\sigma_a} & \frac{1}{\sigma_b} \end{pmatrix}^T \begin{pmatrix} \sigma_a^2 & COV(a,b) \\ COV(a,b) & \sigma_b^2 \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_a} \\ \frac{1}{\sigma_b} \end{pmatrix}$$

$$\rho = corr(a,b) = \begin{pmatrix} 1 & \frac{COV(a,b)}{\sigma_a \sigma_b} \\ \frac{COV(a,b)}{\sigma_a \sigma_b} & 1 \end{pmatrix}$$

Remember that if any set of variables are correlated, we can use the correlation to propagate the uncertainties as:

$$\sigma_f^2 \approx \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)\sigma_{xy}$$

Maximum Likelihood

If the N measurements are not independent but described by an N-dimensional Gaussian PDF with known covariance matrix $\hat{V}$, the likelihood can be generalized to

$$\mathcal{L}(\vec{y},\vec{\mu},\hat{V}) = \frac{1}{(2\pi)^{N/2}\left|\hat{V}\right|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{y}-\vec{\mu})^T\hat{V}^{-1}(\vec{y}-\vec{\mu})\right\}$$

and the **log-likelihood** can be written as (dropping the terms that do not depend on the parameters)

$$\log\mathcal{L}(\{\vec{a}\}) = -\frac{1}{2}\sum_{i,j=1}^{N}\left(y_i - f(x_i,\{\vec{a}\})\right)\left(\hat{V}^{-1}\right)_{ij}\left(y_j - f\left(x_j,\{\vec{a}\}\right)\right) \qquad V_{ij} = \mathrm{cov}(y_i,y_j)$$

The **log-likelihood** is maximized by minimizing the quantity:

$$\chi^2(\{\vec{a}\}) = \sum_{i,j=1}^{N}\left(y_i - f(x_i,\{\vec{a}\})\right)\left(\hat{V}^{-1}\right)_{ij}\left(y_j - f\left(x_j,\{\vec{a}\}\right)\right)$$

Examples: time series, sensor measurements with spatial correlation, correlated experimental measurements (particle detectors), …

"Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC,"

ATLAS Collaboration

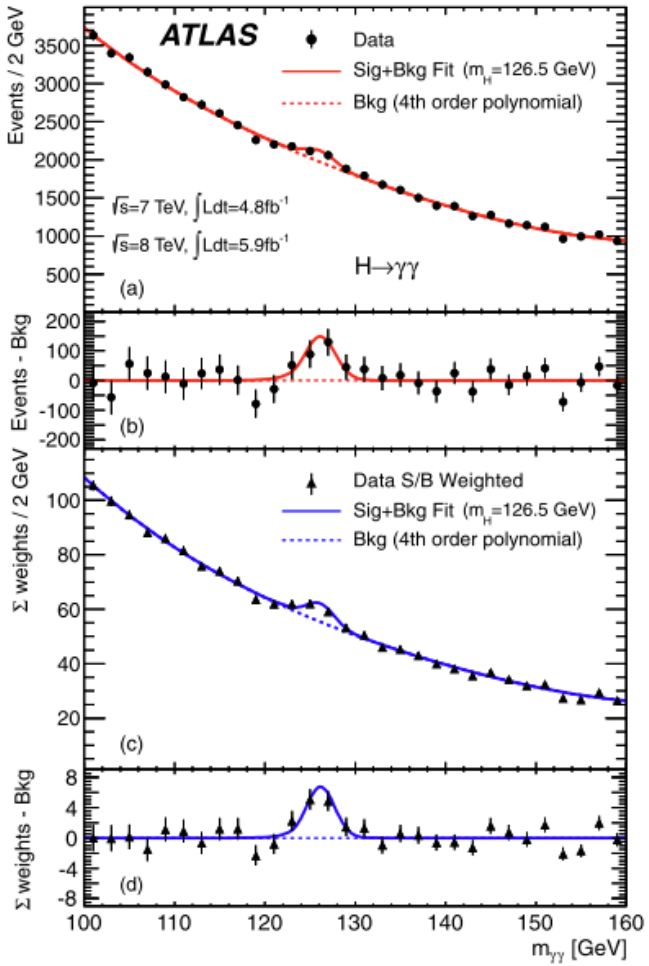


Fig. 4. The distributions of the invariant mass of diphoton candidates after all selections for the combined 7 TeV and 8 TeV data sample. The inclusive sample is shown in (a) and a weighted version of the same sample in (c); the weights are explained in the text. The result of a fit to the data of the sum of a signal component fixed to $m_H = 126.5$ GeV and a background component described by a fourth-order Bernstein polynomial is superimposed. The residuals of the data and weighted data with respect to the respective fitted background component are displayed in (b) and (d).

Physics Letters B
Volume 716, Issue 1, 17
September 2012, Pages 1-29

https://home.cern/resources/faqs/five-sigma