TME – HIPOTHESIS TESTING ASSIGNMENT

The Higgs notebook simulates the distribution of the invariant mass of pairs of photons (m$\gamma\gamma$) found in events collected by the ATLAS experiment at the LHC [1].

The model used for the simulation contains:

- A decreasing exponential background with rate parameter, corresponding to pairs of photons produced independently: τ=0.0218GeV-1
- A Gaussian distribution corresponding to the decay H $\rightarrow \gamma\gamma$ with H being a Higgs boson of a mass of 126.5 GeV, and with a standard deviation given by the experimental resolution of the measurement of m$\gamma\gamma$ (2GeV).

Let us assume that the normalization set by default in the program (Ntot=80000) corresponds to the data collected by the ATLAS detector in one year, and that the performances of the collider and the detector are constant in time: the amount of data to analyze is just proportional to the time used to collect the data.

Let our null hypothesis be "the data follows the exponential distribution with a constant τ=0.0218 GeV^{-1}". Then the p-value of our null hypothesis will quantify the statistical significance of the discovery of a new particle (the Higgs boson).

- 1) Build a χ^2 estimator that tests whether data follows the null hypothesis. Obtain its sampling distribution for many one-year experiments.
- 2) Let us define the expected significance as the expectation value of the p-value of the null hypothesis. What is the expected significance after one year of data taking?
- 3) How many years of data do we need for the expected significance to be at the level of 5σ that is, $< 2.9 \cdot 10^{-7}$?
- 4) How many years of data do we need in order to have a 95% probability of the p-value being at the level of 5σ that is, $< 2.9 \cdot 10^{-7}$?

Repeat the above numerical experiments using the Kolmogorov-Smirnov test instead of the χ^2 test. Are more or less years of data required?

Discuss in detail the interpretation of the P-values and its distribution as obtained above. Discuss also in detail the implications of the results for the design of the experiment and the different efficiency of the two estimators.

The above test is a toy model, while the test used in the paper [1] is much more sophisticated. Provide a short review and comparison of the methodology used in the paper.

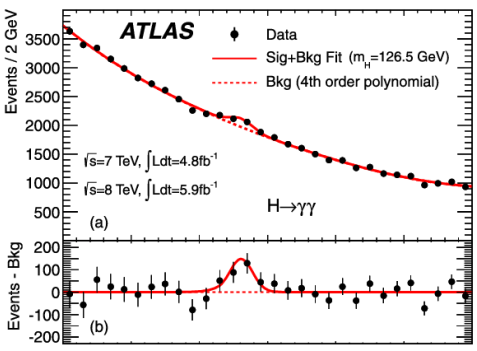


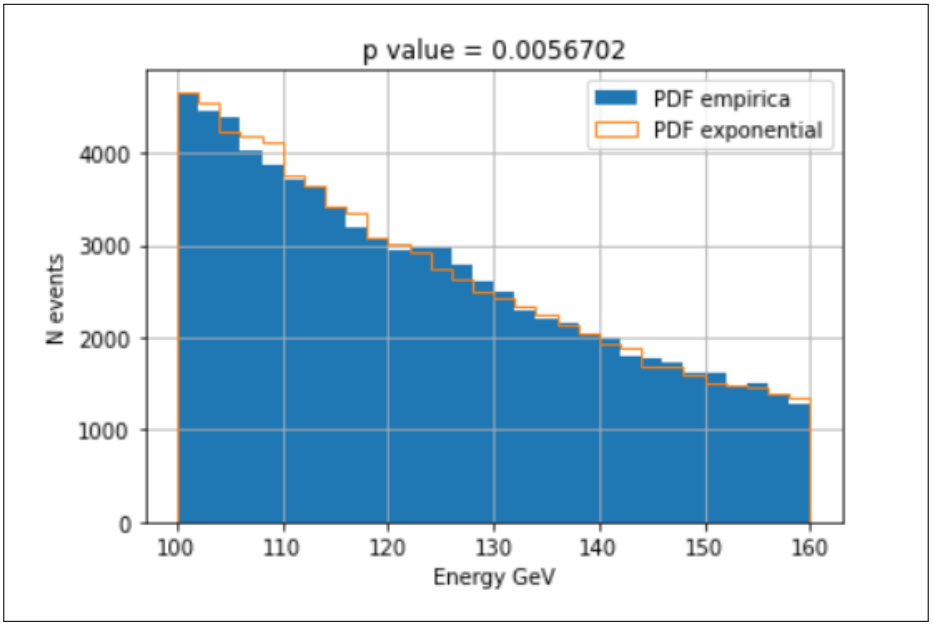
Figure 0.1: Plot showing the hidden gaussian under the decreasing exponential, for a mass around 126.5 GeV [1].

1 Building a X^2 estimator

For building this estimator, first I have adapted the given code to one which plots the collected Higgs data (blue), in front of only the decreasing exponential (orange). And then, we compare them with the estimator, which gives us the p-value of that case, in the graphic title:

```
%matplotlib inline
import scipy.stats as scp
import matplotlib.pyplot as plt
import numpy as np

# Higgs' boson mass in GeV
mH= 126.5

#Total number of events generated and ratios background/signal
years=1
Ntot = 80000*years
Nbg = int(Ntot*0.995)
Ns = int(Ntot*0.005)

# Max and min GeV of the sample
min= 100
max= 160

# Bins for plot
SizeBins=2
Nbins=int((max-min)/SizeBins)+1
bins= np.linspace(min,max,Nbins)

# Rate
tau= 0.0218

#GENERATING DATA
# Generating background (truncated exponential)
bg_dist= scp.truncexpon(b=(max-min)*tau, loc=min, scale=1./tau)
bg_points= bg_dist.rvs(Nbg)

# Generating signal (gaussian) around mH. We take a sigma of 2.
sig_dist= scp.norm(loc=mH,scale=2)
sig_points= sig_dist.rvs(Ns)

# Join background and signal in a single sample
all_points= np.concatenate( (bg_points,sig_points) )

# Histogram of global sample
# Note: the binning here is the one used in Figure 4
hh= plt.hist(all_points,bins,label='PDF empirica')


#GENERATING THE COMPARISON NULL HYPOTHESIS (exponential)
#p_bins= [ Ntot*(bg_dist.pdf(bins[i])+bg_dist.pdf(bins[i+1]))*SizeBins/2
    for i in range(len(bins)-1)]
p_bins= [ Ntot*(bg_dist.cdf(bins[i+1])-bg_dist.cdf(bins[i])) for i in range
    (len(bins)-1)]

# Histogram of the comparison (exponential)
x= bg_dist.rvs(Ntot)
p= plt.hist(x,bins, histtype=u'step', label='PDF exponential')

# Run test
c2_stat, p_val = scp.chisquare(hh[0],p_bins)

```

```
#PLOT
plt.grid(True)
plt.xlabel('Energy GeV')
plt.ylabel('N events')
plt.title("p value = {:.7f}".format(p_val))
plt.legend(loc=1)
plt.show()
```



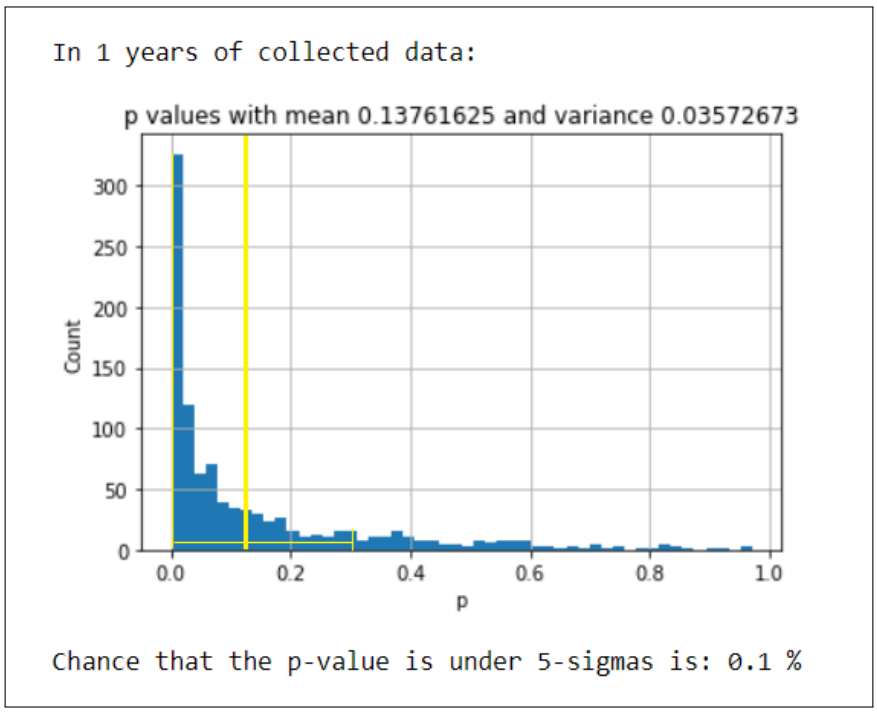
In the next section we will change the code to make many one-years experiments, obtaining the p value for each case, from which, we can do its sampling distribution and obtain the expected p value.

2 Expected significance

Now to obtain the expected value of the p value we need to do lots of χ^2 test, and take the mean of the diverse p values. For that I have included a loop on the code to do the test, "Nsamples" times. And then I have plotted the distribution of the p-values, with it's mean in the title:

```
%matplotlib inline
import scipy.stats as scp
import matplotlib.pyplot as plt
import numpy as np

# Higgs' boson mass in GeV
mH= 126.5

#Total number of events generated and ratios background/signal
years=1
Ntot = 80000*years
Nbg = int(Ntot*0.995)
Ns = int(Ntot*0.005)
Nsamples= 1000 #number of diferent experiments to compute each p value

# Max and min GeV of the sample
min= 100
max= 160

# Bins for plot
SizeBins=2
Nbins=int((max-min)/SizeBins)+1
bins= np.linspace(min,max,Nbins)

# Rate
tau= 0.0218

# Distributions i'll use
bg_dist= scp.truncexpon(b=(max-min)*tau, loc=min, scale=1./tau) #
    Distribucion
sig_dist= scp.norm(loc=mH,scale=2)


#GENERATE DATA AND COMPILE P VALUES
p_values= []
for i in range(Nsamples):

    # Generating background (truncated exponential)
    bg_points= bg_dist.rvs(Nbg)

    # Generating signal (gaussian) around mH. We take a sigma of 2.
    sig_points= sig_dist.rvs(Ns)

    # Join background and signal in a single sample
    all_points= np.concatenate( (bg_points,sig_points) )

    # Histogram of global sample
    # Note: the binning here is the one used in Figure 4
    hh= np.histogram(all_points,bins)


    #GENERATE THE COMPARISON NULL HYPOTHESIS (exponential)
    p_bins= [ Ntot*(bg_dist.pdf(bins[i])+bg_dist.pdf(bins[i+1]))*SizeBins/2
     for i in range(len(bins)-1)]
    #p_bins= [ Ntot*(bg_dist.cdf(bins[i+1])-bg_dist.cdf(bins[i])) for i in
    range(len(bins)-1)]
```

```
54
55      # Run test
56      c2_stat, p_val = scp.chisquare(hh[0],p_bins)
57
58      p_values.append(p_val)
59
60  #Histogram of p values
61  h = plt.hist(p_values,50)
62  #Computing expected p value
63  counter=0
64  p_mean= np.mean(p_values)
65  for i in range(Nsamples):
66      if p_values[i] < 2.9e-7:
67          counter+=1
68
69  #PLOT
70  print('')
71  print('In' , years, 'years of collected data:')
72
73  plt.grid(True)
74  plt.xlim(0,1)
75  plt.xlabel('p')
76  plt.ylabel('Count')
77  plt.title("p values with mean {:.8f}".format(p_mean))
78  plt.show()
79
80  print('Chance that the p-value is under 5-sigmas is:',counter/Nsamples*100,
        '%')
81  print('')
```



So the expected significance after one year of data taking is 0.1376, with a variance (1σ) of 0.0357. So 68,3% of the p values will be within:

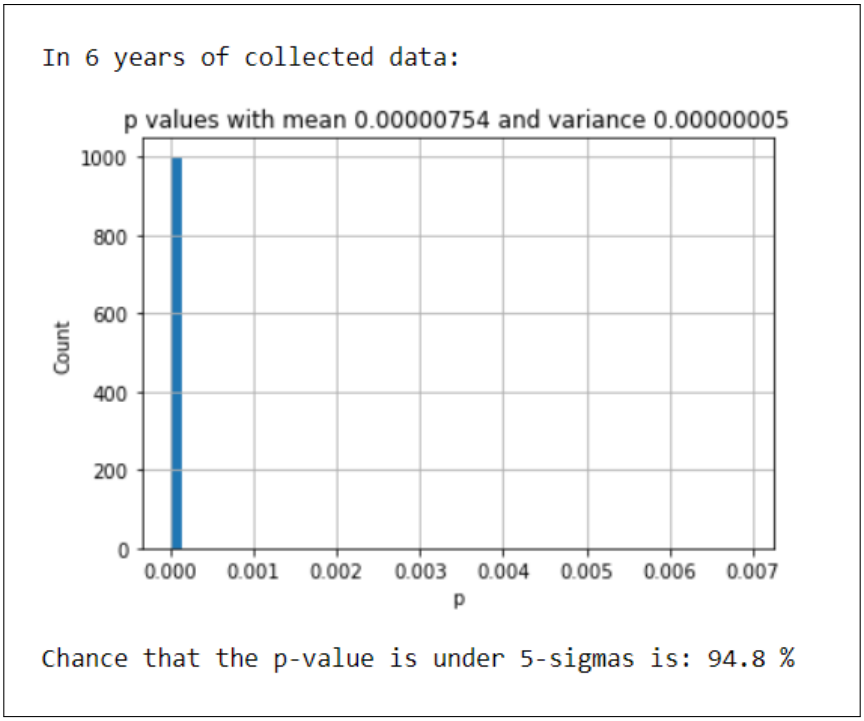
$$p = 0.1376 \pm 0.1890$$

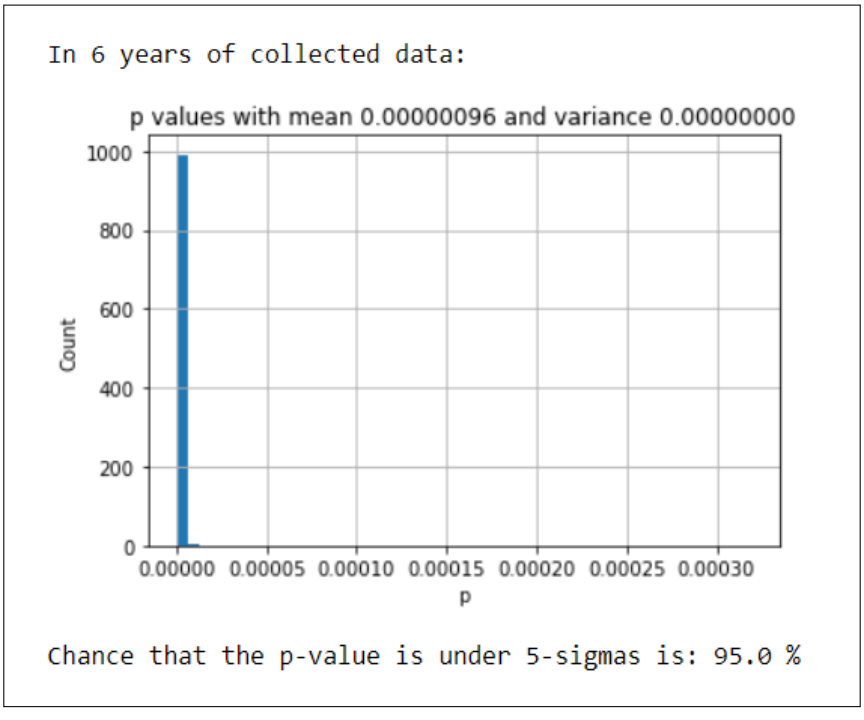
(it's clear that negative p values make no sense, so in reality it should that 68,3% of the p values are within 0 and 0.3266, as shown in the picture with the yellow lines.)

3 Expected significance at 5σ

Now we are going to change the variable "years" in the previous codes, in order to obtain experiment of diverse lengths instead of the previous fixed value of 1 year.

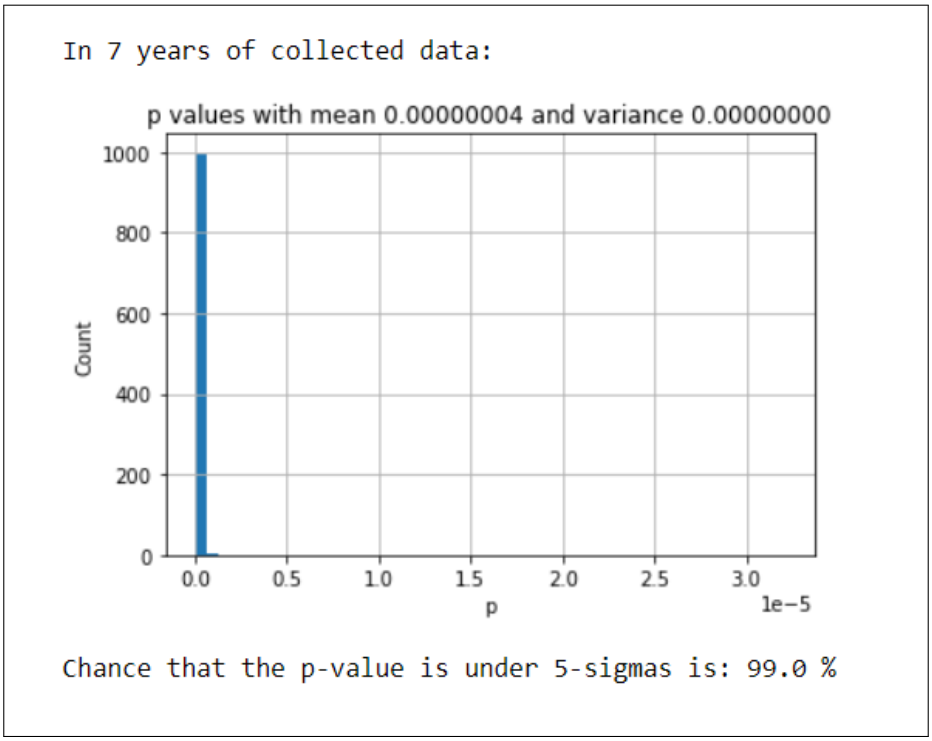
Doing this, we observe that in order to obtain expected significance of the p-values at the level of 5σ ($2.9 \cdot 10^{-7}$), we need to go to 7-8 years of sampling. In the next pictures I'll show diverse results for 6, 7 and 8 years where we can see this:

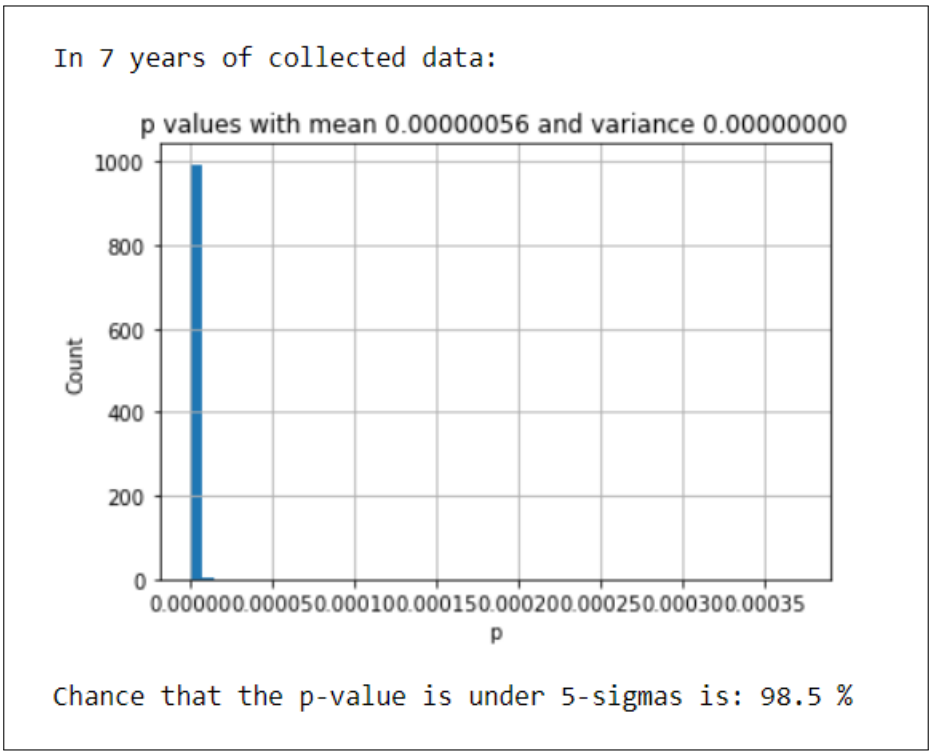




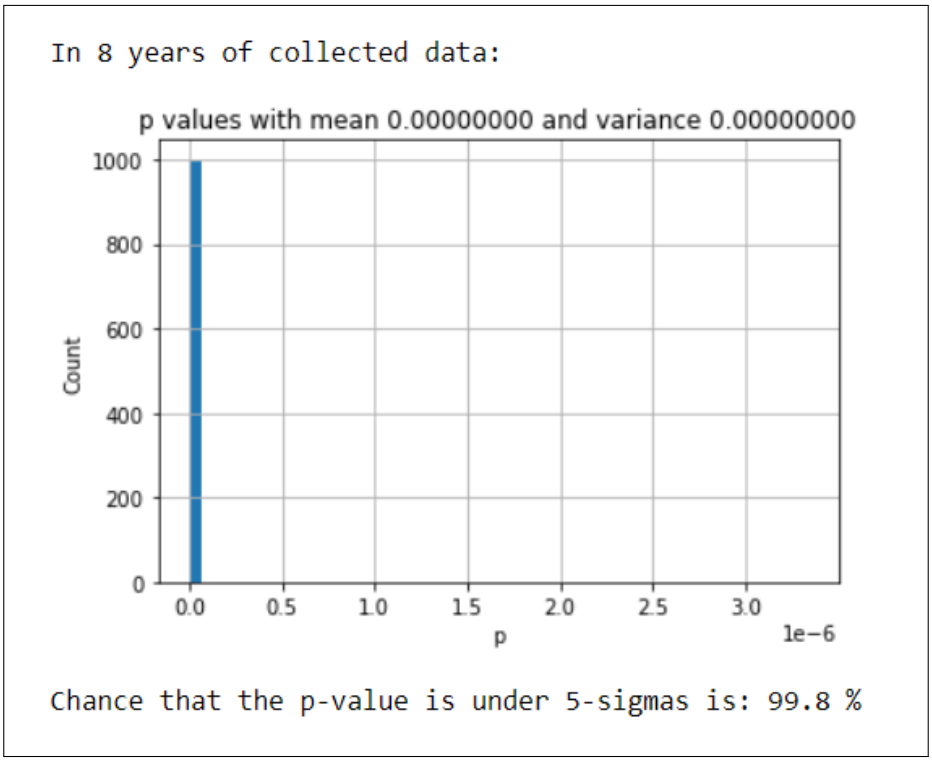
First, we see that in 6 years, we don't get the p value mean $< 2.9 \cdot 10^{-7}$ as we wanted,

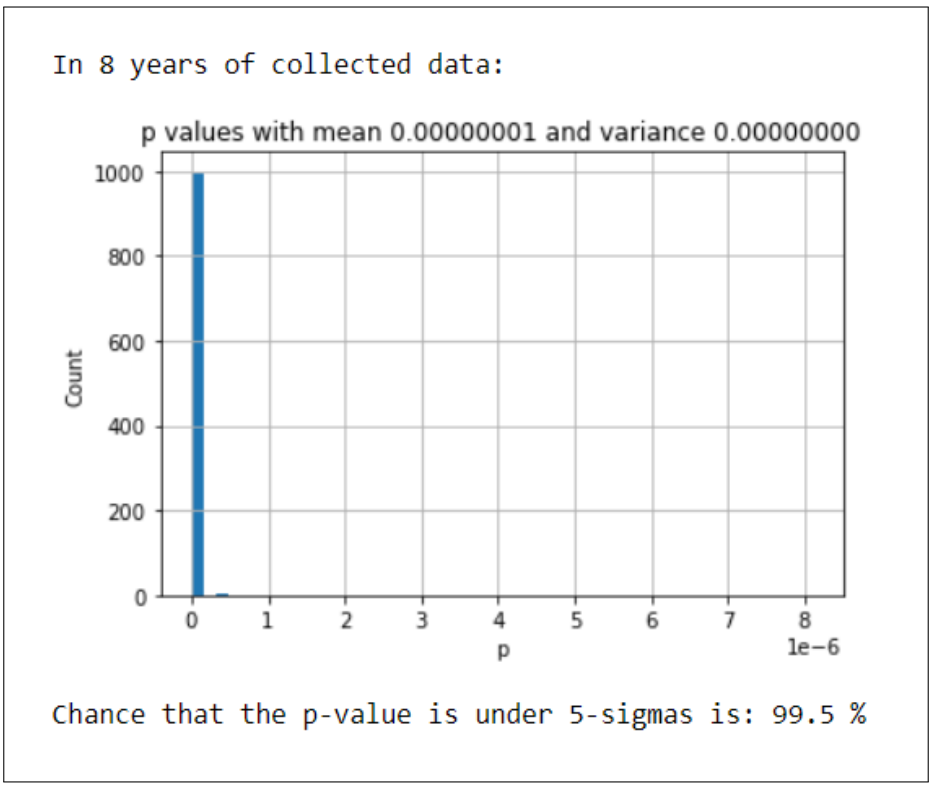
we obtain results of the order of $2 \cdot 10^{-6}$.





Now, we see that for 7 years we obtain results for the p value mean around the $2 \cdot 10^{-7}$. Which means that normally, we get the p mean $< 2.9 \cdot 10^{-7}$, but some other times it doesn't get under the 5σ level we want.





And finally, we obtain that in 8 years, every time we have checked (around 20 times), the p mean is $< 2.9 \cdot 10^{-7}$, getting results for the p value mean of the order of 10^{-8}.

So to conclude, we should wait from 7 to 8 years in order to the expected significance to be at the level of 5σ, that is $< 2.9 \cdot 10^{-7}$.

4 95% chance of p-value at 5σ

If now we focus in each of the p values, instead than in the mean, to obtain a 95% chance of those individual p values being at 5σ level, we will need to wait only around 6 years, instead of from 7 to 8 years.

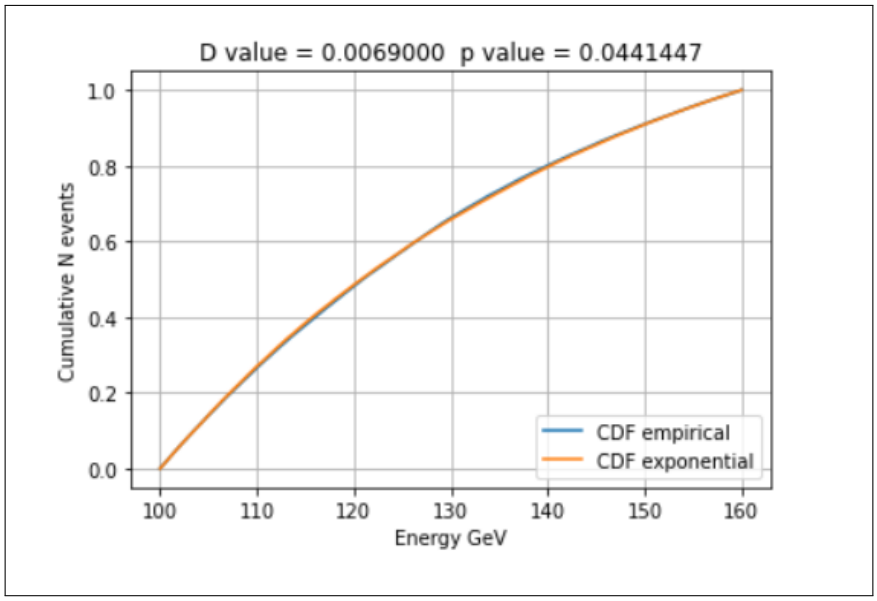
We were already computing this with the code of section 2, and also under each graphic of the previous section, we were already showing the results, the chances that each individual p value is under the 5σ level.

From all the simulations we have done, we have obtained that the chance that the individuals p values are under 5σ is 95%, is accomplished in most of the cases at 6 years. Sometimes we only get to 94% or so, which means that we might need a bit more, but the right time is clearly nearer to 6 years than to 7 years.

5 Kolmogorov-Smirnov test

Now we have to repeat the numerical experiments using the Kolmogorov-Smirnov test, to do this test, we are going to use this other code:

```
%matplotlib inline
import scipy.stats as scp
import matplotlib.pyplot as plt
import numpy as np
import statsmodels.distributions.empirical_distribution as stm

# Higgs' boson mass in GeV
mH= 126.5

#Total number of events generated and ratios background/signal
years=1
Ntot = 80000*years
Nbg = int(Ntot*0.995)
Ns = int(Ntot*0.005)

# Max and min GeV of the sample
min= 100
max= 160

# Bins for plot
SizeBins=2
Nbins=int((max-min)/SizeBins)+1
bins= np.linspace(min,max,Nbins)

# Rate
tau= 0.0218

#GENERATING DATA
# Generating background (truncated exponential)
bg_dist= scp.truncexpon(b=(max-min)*tau, loc=min, scale=1./tau)
bg_points= bg_dist.rvs(Nbg)

# Generating signal (gaussian) around mH. We take a sigma of 2.
sig_dist= scp.norm(loc=mH,scale=2)
sig_points= sig_dist.rvs(Ns)

# Join background and signal in a single sample
all_points= np.concatenate( (bg_points,sig_points) )

# Ploting experimental cdf of the global sample
ecdf=stm.ECDF(all_points)
plt.plot(ecdf.x,ecdf.y,label='CDF empirical')

#GENERATING THE COMPARISON NULL HYPOTHESIS (exponential) and ploting the
    CDF
bg_points2= bg_dist.rvs(Ntot)
ecdf=stm.ECDF(bg_points2)
plt.plot(ecdf.x,ecdf.y,label='CDF exponential')

# Run test
D_value, p_value = scp.ks_2samp(all_points,bg_points2)

#PLOT
plt.grid(True)
plt.xlabel('Energy GeV')
plt.ylabel('Cumulative N events')
plt.title("D value = {:.7f}  p value = {:.7f}".format(D_value,p_value))
plt.legend(loc=4)
plt.show()
```

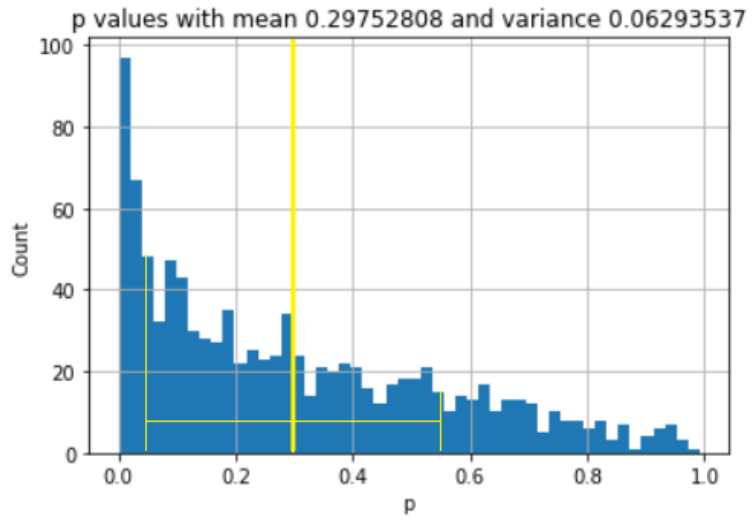


But as happened in the previous case, we wanna do this test multiple times to obtain a p value sampling distribution. So the final code we are going to use will be the previous adding a loop:

```
%matplotlib inline
import scipy.stats as scp
import matplotlib.pyplot as plt
import numpy as np

# Higgs' boson mass in GeV
mH= 126.5

#Total number of events generated and ratios background/signal
years=1
Ntot = 80000*years
Nbg = int(Ntot*0.995)
Ns = int(Ntot*0.005)
Nsamples= 1000 #number of 1 year diferent experiments to compute each p
    value

# Max and min GeV of the sample
min= 100
max= 160

# Bins for plot
SizeBins=2
Nbins=int((max-min)/SizeBins)+1
bins= np.linspace(min,max,Nbins)

# Rate
tau= 0.0218

# Distributions i'll use
bg_dist= scp.truncexpon(b=(max-min)*tau, loc=min, scale=1./tau)
sig_dist= scp.norm(loc=mH,scale=2)

#GENERATE DATA AND COMPILE P VALUES
p_values= []
D_values= []
for i in range(Nsamples):
    # Generating background (truncated exponential)
    bg_points= bg_dist.rvs(Nbg)
```

```
38
39      # Generating signal (gaussian) around mH. We take a sigma of 2.
40      sig_points= sig_dist.rvs(Ns)
41
42      # Join background and signal in a single sample
43      all_points= np.concatenate( (bg_points,sig_points))
44
45      #GENERATE THE COMPARISON NULL HYPOTHESIS (exponential)
46      bg_points2 = bg_dist.rvs(Ntot)
47
48      # Run test
49      D_value, p_value = scp.ks_2samp(all_points,bg_points2)
50      p_values.append(p_value)
51
52  #Histogram of p values
53  h = plt.hist(p_values,50)
54  #Computing expected p value
55  counter=0
56  p_mean= np.mean(p_values)
57  p_var = np.var(p_values)
58  for i in range(Nsamples):
59      if p_values[i] < 2.9e-7:
60          counter+=1
61
62  #PLOT
63  print('')
64  print('In' , years, 'years of collected data:')
65
66  plt.grid(True)
67  plt.xlabel('p')
68  plt.ylabel('Count')
69
70  plt.title("p values with mean {:.8f} and variance {:.8f}".format(p_mean,
        p_var))
71  plt.show()
72
73  print('Chance that the p-value is under 5-sigmas is:',counter/Nsamples*100,
        '%')
```

```
In 1 years of collected data:
```




```
Chance that the p-value is under 5-sigmas is: 0.0 %
```

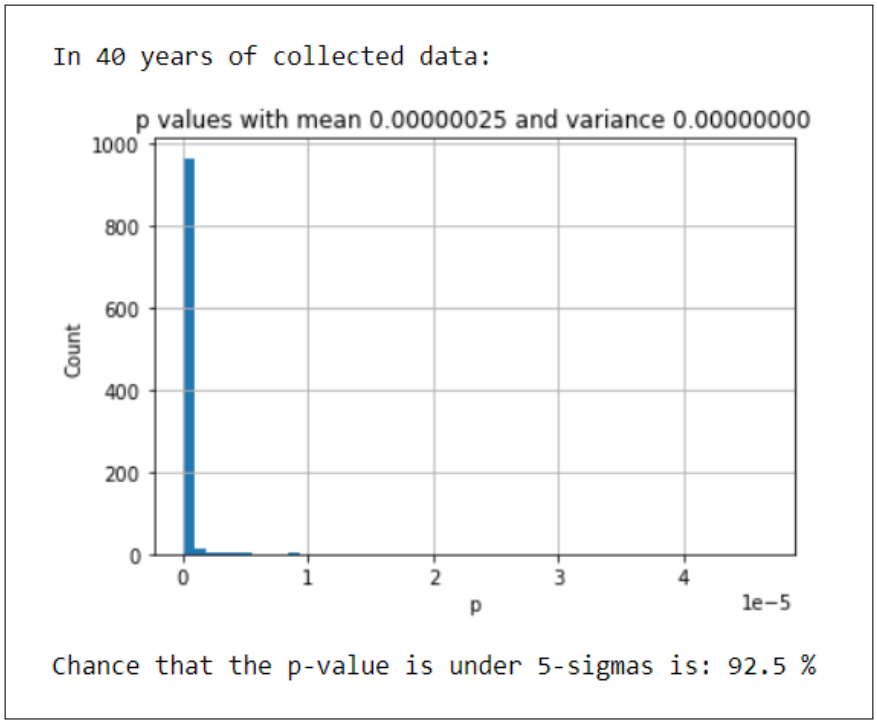
And the results of the previous sections for this test are:

- From the previous image, we see that the expected significance after one year of data taking is 0,2975, with a variance (1σ) of 0.0629. So 68,3% of the p values will be within:

  $$p = 0,2975 \pm 0,2508$$

  where obviously the left side contributes much more to that 68,3% than the right side, as shown in the previous picture with the yellow lines.

- For the expected significance to be at the level of 5σ we will need around 40 years of data, as we can see in the next picture:



- But to have the individuals p values at 5σ level, 95% of the times, we will need around 42 to 43 years, which is more than what we needed for the expected significance to be at the level of 5σ, opposite with what happened in the chi square test.

In conclusion, with the Kolmogorov-Smirnov test, we need more years of data to arrive to the same confidence of our p values.

So, the fact that we need lots more years of data to establish the same confidence level on the p values, with this test, tells us, that at first sight we would like to favor the usage of the chi square test instead, in the real experiments.

But we have also to consider, that in the chi square test, we had to bin our comparison pdf, losing some resolution in our final result, in contrast with the Kolmogorov-Smirnov test, where we can compare directly with the cdf (at least until we arrive to the discretization of our software), or the fact that some test work better for variations near the mean, and other with variations in the tails, etc... before we can make a real decision, on which test is better, or which one should we use.

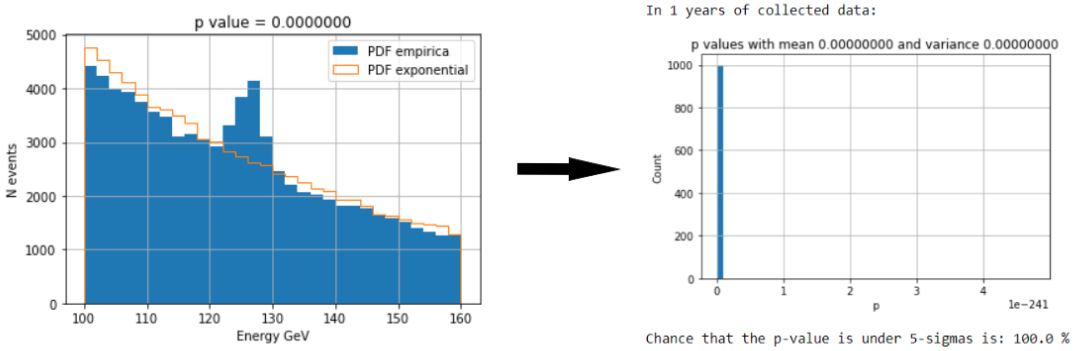
6 Discussion

To end, we are going to discuss the interpretation of the p-values distribution and the toy model we used vs the method used in [1].

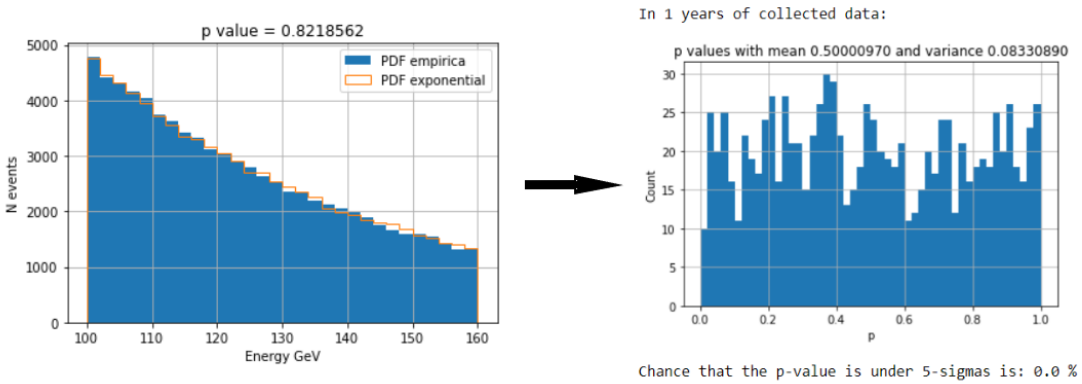
First, respect the distribution of the p values, we clearly see what we expected from our "allmighty" position, where we knew we were comparing two different distributions, we expected that the p values had to go to 0 at the limit of having more data, and that is exactly what we see.

Whats more, if we play a bit with the weight of the exponential and gaussian parts of the experimental data, when we make the gaussian contribution bigger, the difference on the experimental data and the comparison exponential becomes obvious, and thats exactly what happens in the simulations, their p values tend to 0 quicker. And the same happens in the opposite case, when we give more weight to the exponential part of the experimental data, the difference between the data and a exponential behaviour becomes harder to notice, and the p values get more homogeneously dispersed, as one would expect.

In the next pictures we will try to show this, from our original 99,5% and 0,5% proportion of exponential to the gaussian, we are going to make the proportions become 95% and 5% instead, and we directly get all the p values at 0, with only 1 year of data:



And now the opposite, if we make the proportions of gaussian to exponential become 99,95% and 0,05% instead, it's impossible to distinguish them with only 1 year of data:



So clearly, to make our next experiments more efficient, we would like to observe a process, where the Higgs contribution is bigger than that 0,5% or we would like to increase the number of events happening near the 126,5 GeV range. Maybe increasing the total number of events, with more luminosity on the beams, or trying to "focus" more the measurements of the experiment to that regime.

Finally, respect the methodology used in the paper, because they didn't know from previous hand the distributions of their experimental data, what they did instead of fixing the strength of the gaussian signal, they used it as a variable to fit the data in, meaning that if they got a 0 on that strengh, the data would correspond to only the exponential background, and if they got a 1, the data would include the Higgs gaussian.

Doing this limits yourself to some test, which are harder to use when you have to fit data, but you don't take things for granted beforehand, which makes the whole process, more reliable.

References

[1] G. Aad, T. Abajyan, B. Abbott, *et al.*, "Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc," *Physics Letters B*, vol. 716, no. 1, pp. 1–29, Sep. 2012, ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.020. [Online]. Available: http://dx.doi.org/10.1016/j.physletb.2012.08.020.