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=0.0218GeV-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 5[U+F073] that is, < 2.9 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 that is, < 2.9 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.

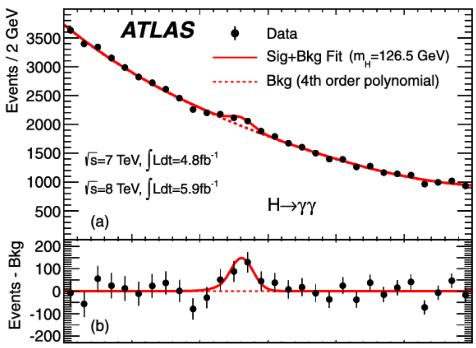


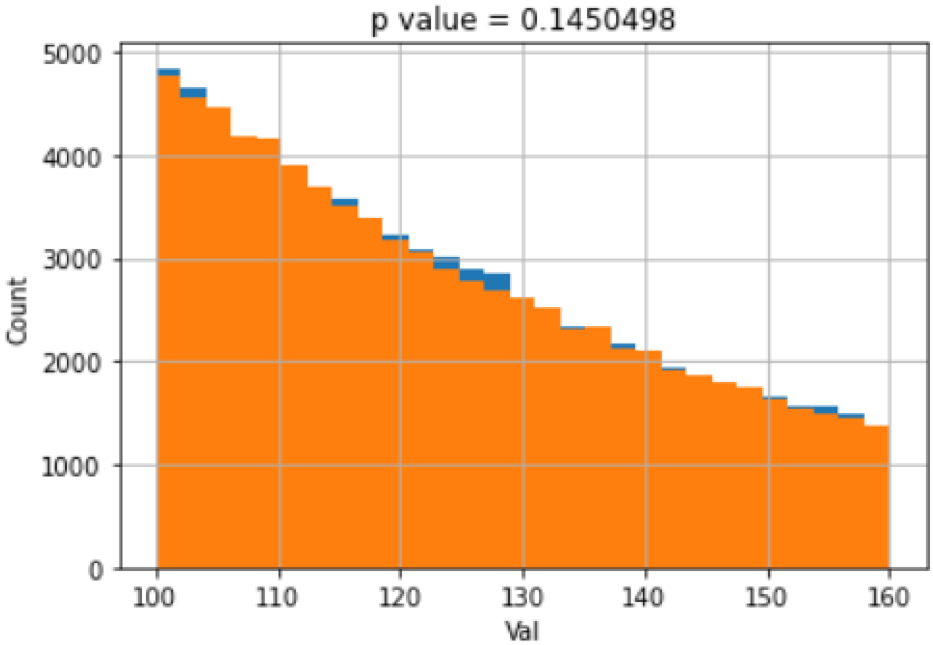
Figure 0.1: Plot showing the hidden gaussian under the decreasing exponential, for a mass around 126.5 GeV [1].

1 Building a X^2 estimator

For building this estimator, first I have adapted the given code to this, which plots the Higgs data (blue) collected in a some years. in front of only the decreasing exponential (orange). And then, we compare them with the estimator, which gives us the p-value 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)
Nsamples= 10000 #number of 1 year diferent experiments to compute each p
    value

# Max and min GeV of the sample
min= 100
max= 160
Nbins=int((max-min)/2)
bins= np.linspace(min,max,Nbins)

# Rate
tau= 0.0218

#GENERAR DATOS
# 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)


#LA DE COMPARAR
p_bins= [ Ntot*(bg_dist.cdf(bins[i+1])-bg_dist.cdf(bins[i])) for i in range
    (len(bins)-1)]
x= bg_dist.rvs(Ntot)
p= plt.hist(x,bins)

plt.grid(True)
plt.xlabel('Val')
plt.ylabel('Count')

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

t = plt.title("p value = {:.7f}".format(p_val))
```

```
plt.show()
```


p value = 0.1450498
5000
4000
3000
2000
1000
0
Count
100
110
120
130
140
150
160
Val

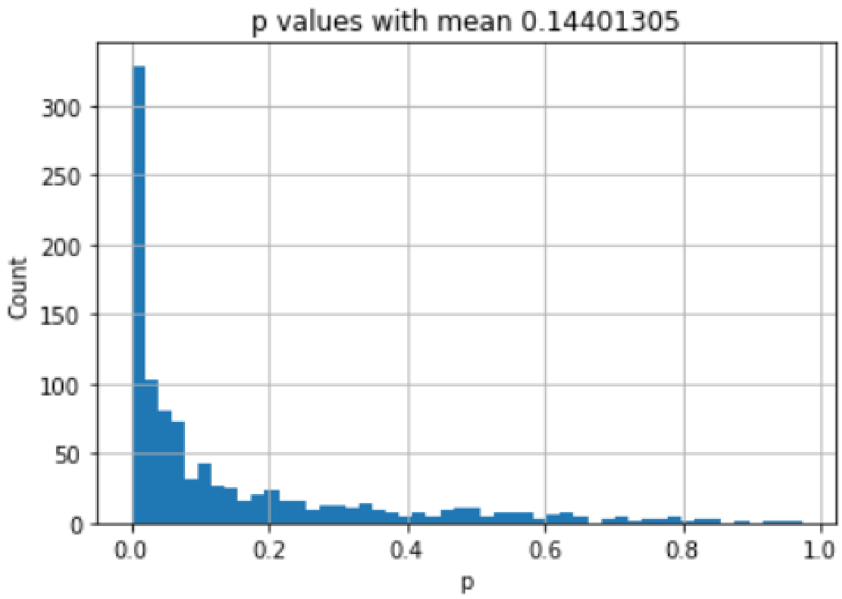
2 Expected significance

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:

```
%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=6
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
Nbins=int((max-min)/2)
bins= np.linspace(min,max,Nbins)

# Rate
tau= 0.0218

# Distribuciones que usa[U+FFFD]
bg_dist= scp.truncexpon(b=(max-min)*tau, loc=min, scale=1./tau) #
    Distribucion
sig_dist= scp.norm(loc=mH,scale=2)

#La de comparar
p_bins= [ Ntot*(bg_dist.cdf(bins[i+1])-bg_dist.cdf(bins[i])) for i in range
    (len(bins)-1)]

#Generar datos y recopilar pvalues
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)


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

    p_values.append(p_val)

```

```
#Miro totes les p que he obtinugt
h = plt.hist(p_values,50)
counter=0
p_mean= np.mean(p_values)
for i in range(Nsamples):
    if p_values[i] < 2.9e-7:
        counter+=1

print('')
print('In' , years, 'years of collected data:')

plt.grid(True)
plt.xlabel('p')
plt.ylabel('Count')

t = plt.title("p values with mean {:.8f}".format(p_mean))
plt.show()

print('Chance that the p-value is under 5-sigmas is:',counter/Nsamples*100,
    '%')
print('')
```

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

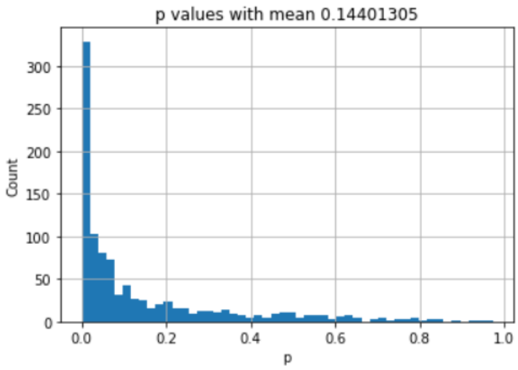



```
Chance that the p-value is under 5-sigmas is: 0.1 %
```

3 Expected significance at 5σ

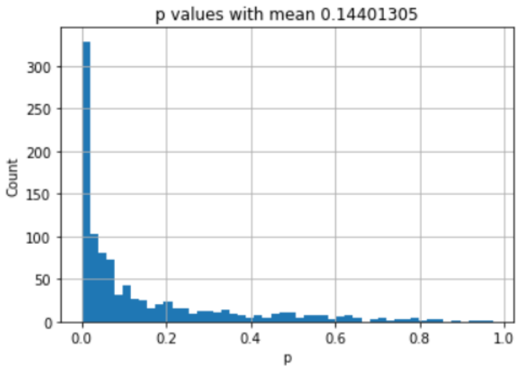
In order to obtain expected significance of the p-values of those orders, we need to go to 7-8 years of sampling. In the next pictures I'll show diverse results for 7 and 8 years where we can see it:

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



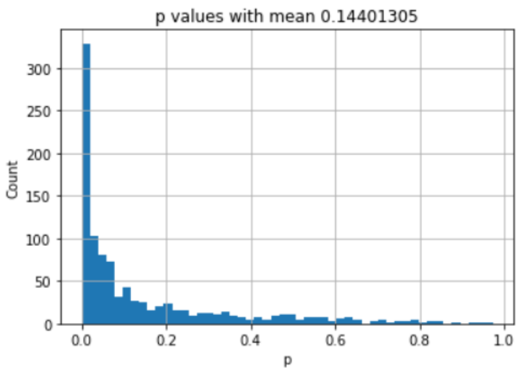

```
Chance that the p-value is under 5-sigmas is: 0.1 %
```

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



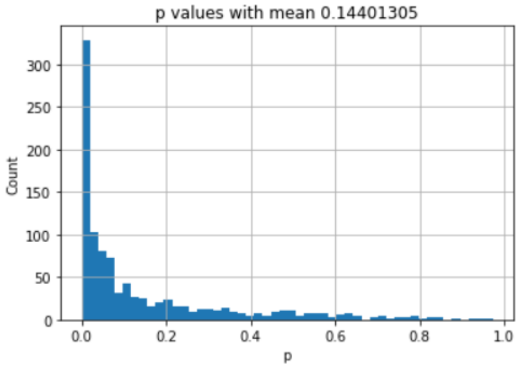

```
Chance that the p-value is under 5-sigmas is: 0.1 %
```

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




```
Chance that the p-value is under 5-sigmas is: 0.1 %
```

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

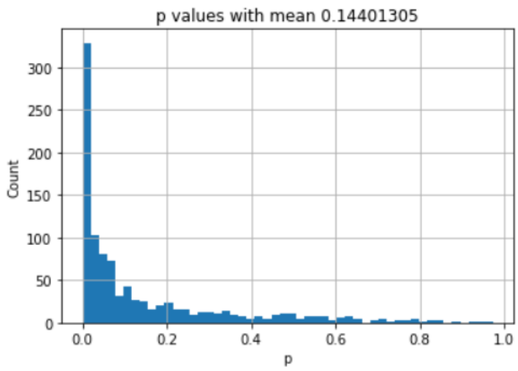



```
Chance that the p-value is under 5-sigmas is: 0.1 %
```

4 95% chance of p-value at 5σ

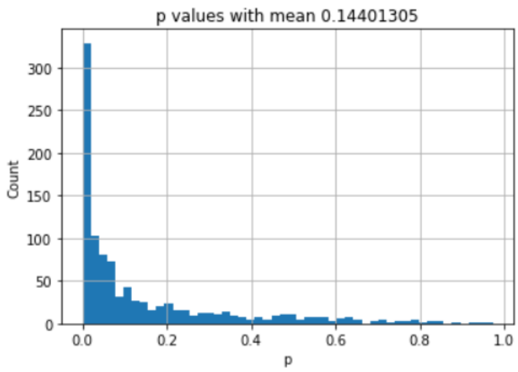
Now, for this we will only need 6 years, we can check that from the next Figures aswell:

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



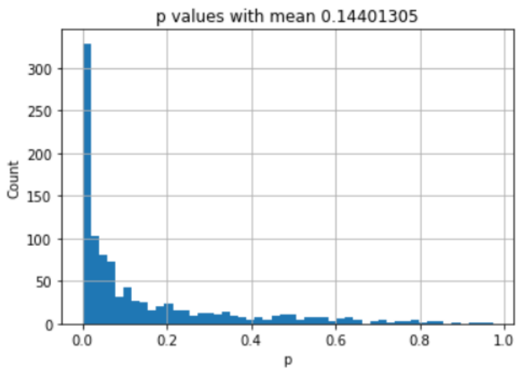

```
Chance that the p-value is under 5-sigmas is: 0.1 %
```

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




```
Chance that the p-value is under 5-sigmas is: 0.1 %
```

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




```
Chance that the p-value is under 5-sigmas is: 0.1 %
```

5 Kolmogorov-Smirnov test

Now we have to repeat everything but with this new test, which is done with this CODE

And the results are this ones.

6 Discussion

Now here we have to discuss in detail the interpretation of the p-values and the toy model of paper etc..

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.