# DAML (Week-9, CP5): Hypothesis testing: Type-I and Type-II errors; Likelihood ratios

Christos Leonidopoulos*

*University of Edinburgh*

November 10, 2023

## 1 Introduction

In the last lecture we saw how to create (pseudo)data distributions that draw from background and/or signal models, and how to apply Wilk's theorem in order to do hypothesis-testing ($H_0$ *vs.* $H_1$). In this lecture, we will discuss in more detail the mechanics behind hypothesis testing; in particular, we will review methods for quantifying the corresponding error probabilities when ruling out either the $H_0$ or the $H_1$ hypotheses, and ways to optimise the discriminating power of the employed *test statistic*.

The workshop uses small snippets of `python` code, and the `Jupyter Notebook` web application environment [1] to display the expected code output.

## 2 Hypothesis testing and test statistic

*Hypothesis testing* is the methodology employed to discriminate between two or more competing hypotheses (e.g. *null* or background-only: $H_0$, and *alternative* or signal-plus-background: $H_1$) on the basis of the observed experimental data.

Examples of questions that address alternate hypotheses to interpret experimental data:

- Is the defendant innocent or guilty?
- Is the detected particle a pion or a muon?

*Christos.Leonidopoulos@ed.ac.uk

- Does the anomaly in the data suggest a signal that can be attributed to a Higgs boson or is it caused by a background fluctuation?

Note that a failure to reject *e.g.* the $H_0$ hypothesis does not mean that the null hypothesis is true. There is no formal methodology that leads to the conclusion "Hypothesis $H_i$ is true"[1]. It only means that we do not have sufficient evidence to support $H_1$ (at the expense of $H_0$). In the first example above, the hypothesis of innocence ($H_0$) is rejected if the hypothesis of guilt ($H_1$) is supported by evidence beyond "reasonable doubt". Failure to prove that the defendant is guilty (or: reject $H_0$) does not imply their innocence, only that the evidence is not sufficient to reject it.

The observed data sample typically consists of measurements of many variables ($\vec{x} = (x_1, \cdots, x_n)$), randomly distributed according to a probability density function $f(\vec{x})$, which is different under the $H_0$ and $H_1$ hypotheses: $f(\vec{x}\,|H_0)$ *vs.* $f(\vec{x}\,|H_1)$. The goal is to determine if the observed data sample agrees better with the $H_0$ or the $H_1$ hypotheses, *i.e.* effectively whether $f(\vec{x}\,|H_0)$ or $f(\vec{x}\,|H_1)$ can better describe the observed experimental data.

Determining the $f(\vec{x}\,|H_i)$ is very often impractical, especially for a large number of variables $\vec{x}$. Instead, it is customary to define and employ a *test statistic* variable, constructed from the variables (or: measurements) $\vec{x}$, which summarises the information contained in the event sample: $t = t(\vec{x})$. The test statistic $t$ can be used to distinguish between the two hypotheses in a more efficient (*i.e.* simpler) way by considering the PDFs for the two scenarios: $g(t\,|H_0)$ *vs.* $g(t\,|H_1)$.

By using a test-statistic we reduce a multi-dimensional ($\vec{x}$) problem for which the PDF $f(\vec{x})$ may be impossible or very difficult to calculate, to one of lower dimension ($t$) for which the PDF $g(t)$ still provides us with the discriminating power to distinguish between the different hypotheses under consideration ($H_0$ *vs.* $H_1$).

Examples of experimental measurements and test-statistic variables:

- $\vec{x}$: all measured quantities (*e.g.* kinematics of all reconstructed decay products) in experimental data
- $t$: event counts, invariant mass of decay products, or fit $\chi^2$ for a given model

# 3 Type-I and Type-II errors

Once a test-statistic $t$ has been chosen for the hypothesis testing of a particular experimental outcome, we apply a selection requirement on the measured value of $t$, $t_{\text{meas}}$ (typically: $t_{\text{meas}} < t_{\text{cut}}$ or $t_{\text{meas}} > t_{\text{cut}}$, for a 1-dimensional problem) which translates into a judgement on the favoured-by-the-data ($H_0$ *vs.* $H_1$) hypothesis.

Fig. 1 shows an example of the test-statistic distributions for the signal and background hypotheses of a physics problem. In this particular example, we have chosen to classify

[1] Karl Popper: "You can only prove a model wrong, never right."

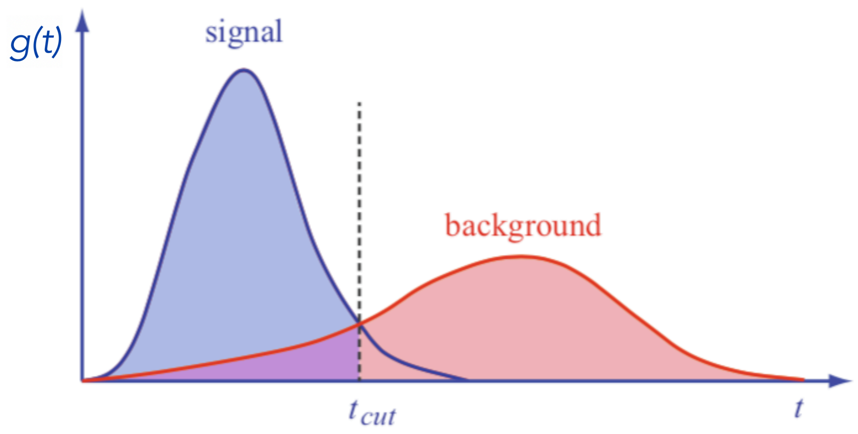


Figure 1: Example of probability distribution functions for a test-statistic of signal (blue) and background (red) origin, and the value $t_{\text{cut}}$ chosen to separate the two hypotheses. Figure taken from Ref. [2].

the outcome of the measurement as background (signal) if $t_{\text{meas}} > t_{\text{cut}}$ ($t_{\text{meas}} < t_{\text{cut}}$). Even though it is true that most of the time a $t_{\text{meas}} > t_{\text{cut}}$ ($t_{\text{meas}} < t_{\text{cut}}$) measurement will correctly identify the outcome of the experiment as being consistent with the background (signal) hypothesis, it is also true that because of the partial overlap of the $g_{\text{sig}}$ and $g_{\text{bgd}}$ PDFs the test-statistic measurement will occasionally give the wrong answer.

For normalised PDFs $g_{\text{sig}}$ and $g_{\text{bgd}}$, we see that

- the probability to reject $H_0$ if $H_0$ is true (**"Type-I error"**) is: $\alpha \equiv \int_{-\infty}^{t_{\text{cut}}} g_{\text{bgd}}(t)\,dt$ (also called significance level)
- the probability to accept $H_0$ if $H_0$ is false (**"Type-II error"**) is the signal inefficiency: $\beta \equiv 1 - \epsilon_{\text{sig}} = \int_{t_{\text{cut}}}^{\infty} g_{\text{sig}}(t)\,dt$, where $\epsilon_{\text{sig}}$ is the signal efficiency.

All the possible outcomes are summarised in Table 1.

| Analysis outcome | $\boldsymbol{H_0}$ **is true** | $\boldsymbol{H_0}$ **is false** |
|---|---|---|
| Rejected $H_0$ | Type-I error ($P = \alpha$) | Correct decision ($P = 1 - \beta$) |
| Did not reject $H_0$ | Correct decision ($P = 1 - \alpha$) | Type-II error ($P = \beta$) |

Table 1: Type-I and Type-II errors, and their corresponding occurrence probabilities, $P$.

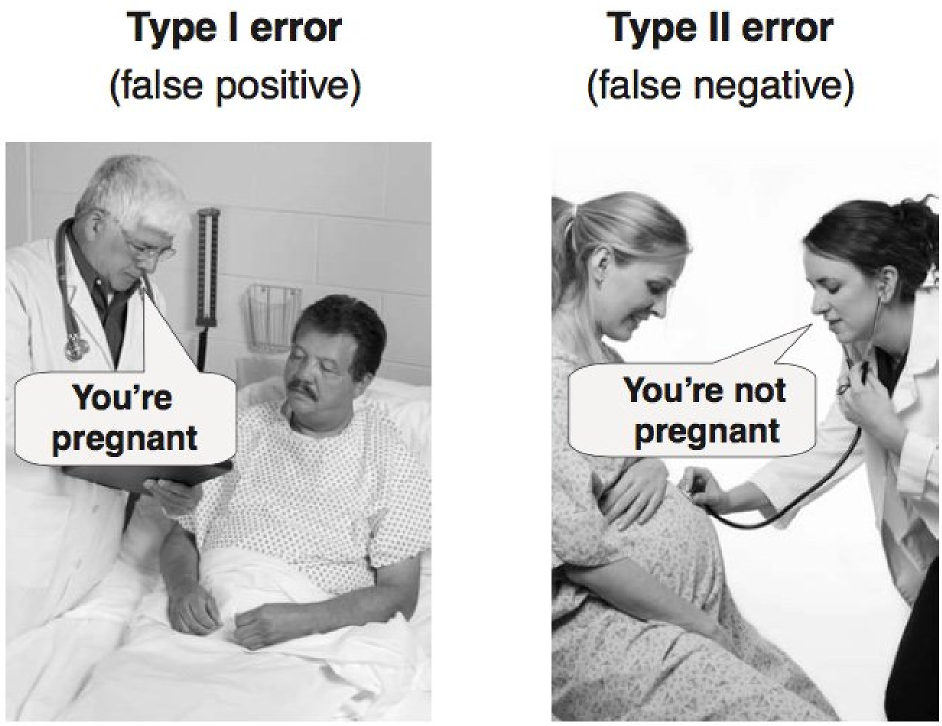


Figure 2: Examples of Type-I and Type-II errors. Credit: "Essential Guide to Effect Sizes", by Paul D. Ellis (2010).

Both $\alpha$ and $\beta$ will ideally be small. Typical values used in High Energy Physics are:

- $\alpha \simeq 3 \times 10^{-7}$ (*i.e.* the threshold for claiming discovery, corresponding to $5\sigma$). To achieve the desired accuracy, we typically use Monte Carlo simulation.
- $\beta = 5$ or 10% (*i.e.* for excluding new physics at 95% or 90% C.L.)

By modifying the threshold $t_{\text{cut}}$, one can change the values $\alpha$ and $\beta$ in an effort to optimise the discriminating power of the chosen test statistic. One can thus obtain a curve of possible $\alpha$ *vs.* $\beta$ values for a fixed test statistic (*Receiver Operating Characteristic*, or ROC curve). Examples of different ROC curves can be seen in Fig. 4.

It is obvious that $\alpha$ and $\beta$ are related for a given test statistic: decreasing one of them generally increases the other. In order to avoid introducing a bias in the statistical inference procedure, it is important to determine both the test statistic and the threshold ($t_{\text{cut}}$) before performing the measurement, *i.e.* before the data are looked at.

More advanced problems of test-statistic choices that involve multiple variables require multi-dimensional thresholds. In such cases, it is not always easy (or potentially even possible) to determine the correct model for multi-dimensional PDFs, in which case we opt for approximate solutions. Examples of test statistics in two dimensions can be seen in Fig. 5. Higher-dimension examples are generally not trivial to visualise. We typically use a neural network or other machine-learning variant for carrying out the optimisation of the discriminant and the threshold selection.

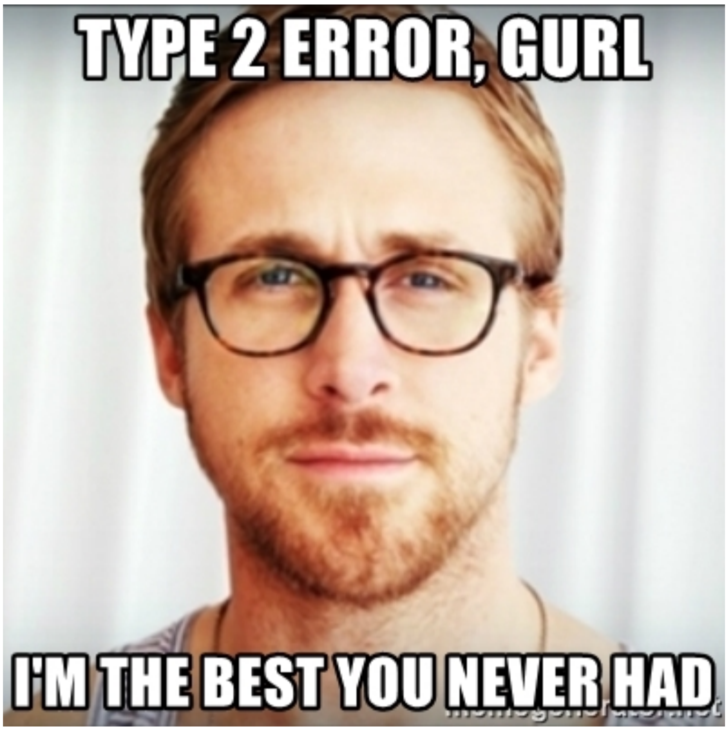


Figure 3: Example of a Type-II error. Credit: internet meme, original author unknown.

# 4 The Neyman-Pearson lemma

According to the Neyman-Pearson lemma [4], the optimal test statistic is given by the ratio of the likelihood functions $L(\vec{x}\,|H_1)$ and $L(\vec{x}\,|H_0)$, evaluated for the observed data sample $\vec{x}$ under the two hypotheses $H_1$ and $H_0$:

$$\lambda(\vec{x}) = \frac{L(\vec{x}\,|H_1)}{L(\vec{x}\,|H_0)} \tag{1}$$

The test is optimal in the sense that, for a fixed background misidentification probability $\alpha$, the selection obtained corresponds to the largest possible signal selection efficiency $1-\beta$.

We have seen this ratio in the last problem of the (Week-8) CP4: We used the (log of the) ratio of the $H_1$ and $H_0$ likelihoods (equivalently: the difference of the corresponding $\chi^2$'s) in order to carry out hypothesis testing. Remember that in the specific case of "nested" hypotheses, we can apply Wilk's theorem [5] and employ $\Delta\chi^2 \equiv \chi^2_{H_0} - \chi^2_{H_1}$ in order to quantify the magnitude of the deviation observed in the data.

More generally, it should be noted that the likelihoods $L$ for the different hypotheses are not always available analytically. In these cases, we typically use (large-statistics) histograms, some multi-dimensional discriminant or a neural network as approximations.

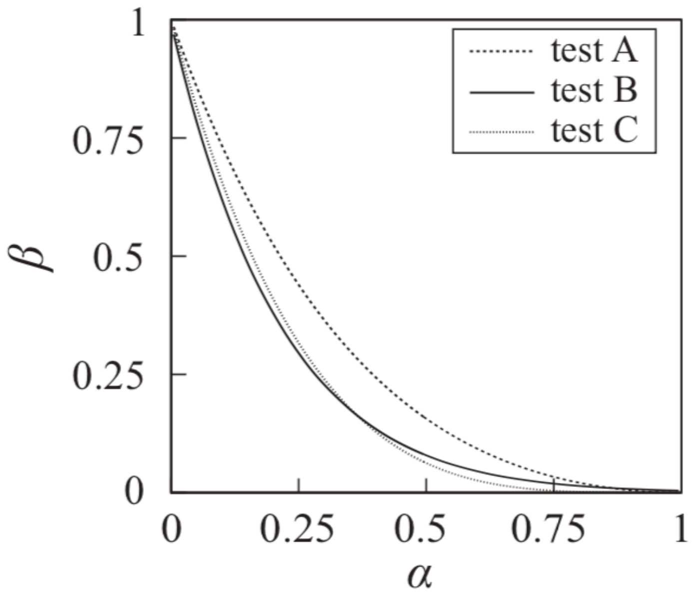


Figure 4: Example curves of $\beta$ *vs.* $\alpha$ for three tests. Test A (dashed line) is clearly the worst one since for a given size $\alpha$ it has the lowest power $1-\beta$, while test B (solid line) and test C (dotted line) lie close to each other. Neither of the two tests can be considered better than the other: while test B is more powerful for small $\alpha$, test C is more powerful for large $\alpha$. As $\alpha$ is usually chosen to be small one would prefer test B in most cases. Figure and caption taken from Ref. [3].

# 5 Summary

We have seen in this lecture how to carry out hypothesis testing:

- Define the null hypothesis ($H_0$) and an alternative hypothesis ($H_1$).
- Choose a test statistic $t$. There may be more than once choices here, and the optimal choice will depend on the specifics of the analysis.
- Determine the expected distributions for the test statistics for the null and alternative hypotheses, $g(t|H_0)$ and $g(t|H_1)$.
- Consider the type-I and type-II errors and determine the threshold $t_{\mathrm{cut}}$ in order to achieve the desired $\alpha$ and $1-\beta$ values.
- Determine $t_{\mathrm{meas}}$ from the experimental data.
- Use (previously defined) selection criterion ($t_{\mathrm{cut}}$) to determine if the null hypothesis can be rejected (*i.e.* case of "discovery").

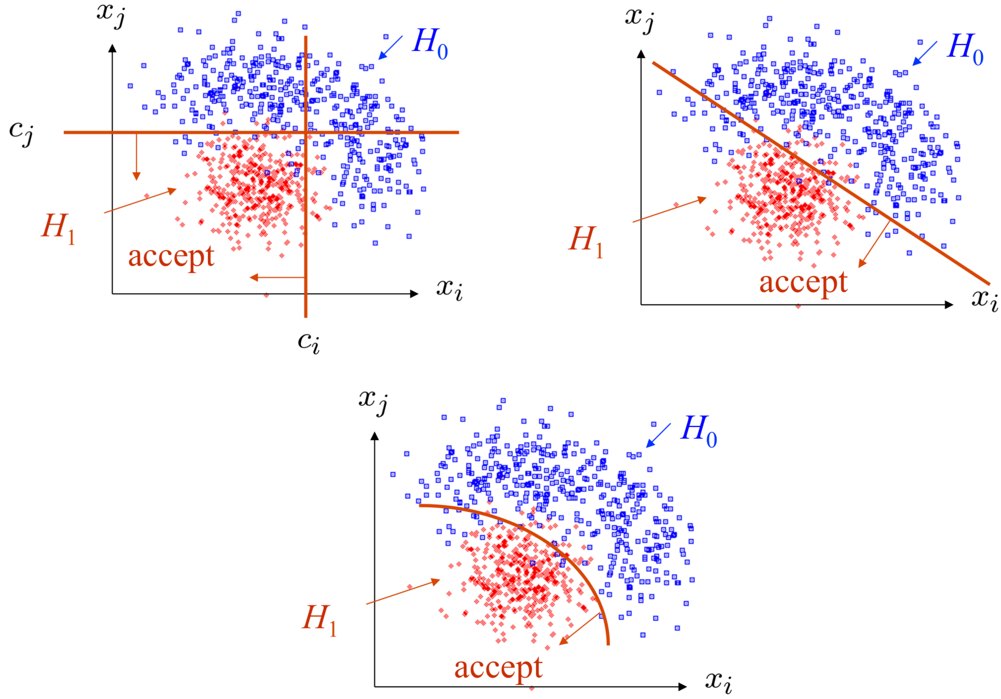


Figure 5: Examples of two-dimensional selections for distinguishing between $H_0$ and $H_1$ hypotheses in a region containing both signal (red points) and background (blue points): square (top left), linear (top right), and non-linear (bottom) selections. Credit: Glen Cowan.

# References


[1] The `Jupyter Notebook` open-source web application, `http://jupyter.org/`

[2] "Statistical Methods for Data Analysis in Particle Physics", Luca Lista, DOI: 10.1007/978-3-319-62840-0

[3] "Data Analysis in High Energy Physics", Behnke, Kröninge, Schott, and Schörner-Sadenius, DOI: 10.1002/9783527653416

[4] "On the problem of the most efficient tests of statistical hypotheses", J. Neyman and E. Pearson, Philos. Trans. R. Soc. Lond. Ser. A 231, 289–337 (1933)

[5] S.S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypothesis", Annals of Math. Stat. **9** (1938), 60

[6] S. Baker, and R.D. Cousins, NIM 221 (1984) 437-442

# A Problem set:

- **Problem #1 (2 points):**

  A company manufacturing computer monitors claims that the faulty rate of the screen population is 5%. We want to test if the claim is true. We have ordered a sample of 100 monitors to test. We choose $t_{\text{cut}} = 9$ as the maximum number of faulty monitors that we are willing to have and still accept that the manufacturer's claim is true.

  (a) What is the significance level (Type-I error), $\alpha$, of the chosen threshold?

  (b) What is the probability $\beta$ of a Type-II error if the true faulty rate is 15%?

  NB: we can only compute the Type-II error for a concrete $H_1$ scenario (*i.e.* fixed faulty rate), but not if the faulty rate is unknown!

  Hint: Use `scipy.stats.binom` (`scipy`'s binomial distribution) and its method `pmf`.

- **Problem #2 (3 points):**

  A Time-of-Flight (ToF) system designed to separate kaons ($m_K = 493.7$ MeV/$c^2$) from pions ($m_\pi = 139.6$ MeV/$c^2$) consists of two scintillation counters that are a distance $L = 20$ m apart. For a particle with mass $m$ and momentum $p$, the time needed to travel between the two scintillators is

$$t = \frac{L}{c} \times \sqrt{1 + \left(\frac{mc}{p}\right)^2}$$

  where $c = 3 \times 10^8$ m/s is the speed of light in vacuum.

  The time resolution of the ToF system is $\sigma = 400$ ps (*i.e.* for an average time $t$, the time reported by the system follows a Gaussian distribution with mean $t$ and width $\sigma$).

  (a) Write a `Gaussian` class that calculates the integral between an arbitrary point `xval` and $\pm\infty$. Name these methods `integralAbove` and `integralBelow`, to be used for calculating $\alpha$ and $\beta$ values, as discussed below.

  Hint: You should try to recycle some of the code developed for the Week-8, CP #4 in order to save time. Be careful to choose practical values for $\pm\infty$!

  (b) Create another class `ROC` that calculates ($\alpha_i$, $\beta_i$) pairs of ToF performance for distinguishing between pions and kaons for a given momentum $p$ and an arbitrary threshold $t^i_{\text{cut}}$. Use the class to produce 100 performance points evenly spaced between the average travel times for kaons and pions.

  (c) Create a single plot that overlays the ROC kaon-pion separation curves for $p = 3$ GeV/$c$, $p = 4$ GeV/$c$ and $p = 6$ GeV/$c$. Which momentum value gives better performance and why?

  Hint: it is more practical to use natural units than SI in the code implementation.

- **Problem #3 (2 points for 3.1, and 3 points for 3.2):**

  In 1992, the ARGUS $e^+e^-$ experiment reported the observation of the charmed and doubly strange baryon $\Omega_c$ through its decay channel $\Xi^- K^- \pi^+ \pi^+$. The obtained mass spectrum is shown in the figure below.

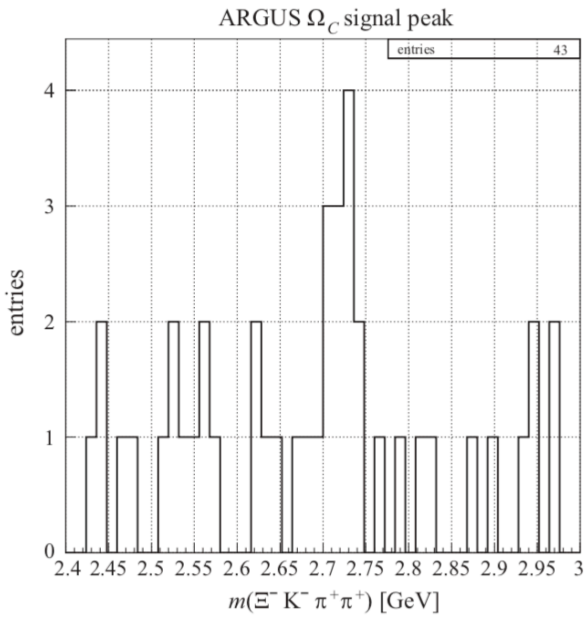


Figure 6: The invariant-mass spectrum reported by the ARGUS experiment.

You can reproduce the plot by using the following snippet:

```
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline

nbins = 50
XMIN = 2.4 # in GeV
XMAX = 3.0 # in GeV
bins = np.linspace(XMIN, XMAX, nbins)
counts = np.array([0, 0, 1, 2, 0, 1, 1, 0, 0, 1, 2, 1, 1, 2, 1, 0, 0, 0, 2, 1, 1,
                   0, 1, 1, 1, 3, 3, 4, 2, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0,
                   0, 1, 2, 0, 2, 0, 0]).astype(float)
tot = 0
for i in counts:
    tot += i
print("total # of events =",tot, "# of bins=",len(counts))
plt.hist(bins, bins=len(counts),weights=counts, range=(min(bins), max(bins)))
plt.show()
```

You should get a total of 43 events and 50 bins in the printout.

3.1 (a) Assuming that all the events are caused by background, calculate the average number of backgrounds events per bin.

(b) Use method `numpy.argmax` to find the location of the peak in the mass spectrum (in GeV).

(c) Define a $\pm 2.5\sigma$ window around the peak ($\sigma = 12$ MeV, the width of the histogram bin), and count the total number of events $N_{\rm obs}$ in this window (use 5 bins in total, with the middle bin containing the peak).

(d) Estimate the number of expected background events within the window $N_{\rm bgd}$, and calculate the probability for a Poisson distribution with mean $N_{\rm bgd}$ to produce $N_{\rm obs}$ or more events, and the number of standard deviations it corresponds to.
Hint #1: Use `scipy.stats.poisson` (`scipy`'s Poisson distribution), and its method `pmf` or `sf`.
Hint #2: Use `scipy.special.erfinv(1 - pvalue) * np.sqrt(2)` to convert a $p$-value into the corresponding number of standard deviations.

3.2 We will repeat the significance evaluation, this time by doing a signal-plus-background ($H_1$) and a background-only ($H_0$) fits. Most of the code we will need here has been developed in (and can be recycled from) the Week-8, CP#4.

(a) Write two classes, `Flat` (to describe the flat background), and `Gaussian` (to describe the hypothetical signal). Class `Flat` should be a simplified version of class `Linear` developed for the Week-8, CP#5.

(b) We will need a minimiser that returns the $\chi^2$ (as minimised by the fit). As discussed in previous weeks, you are welcome to use your favourite minimiser (and you should really have one available by now). Examples: `iminuit`, your own custom implementation of the log-likelihood, or the ($\chi^2$-equivalent of the) log-likelihood for a binned fit, as described in the Week-8 lecture notes (and in Ref. [6]).

(c) Unlike what we had done in Week-8, here we will assume that we do not know the location (*i.e.* Gaussian mean) of the hypothetical signal, but we do know its width (Gaussian sigma, equal to the width of the histogram bin). We will perform $1 + N$ fits: the first one for the $H_0$ hypothesis, and the remaining $N$ fits will **scan** the mass spectrum by assuming each time that the location of the signal is fixed at the centre of the $i$-th bin. For each of the $N$ fits, calculate the $\chi^2(H_0) - \chi^2(H_1)$ difference. Put all these values into a histogram with the mass value indicating the Gaussian mean as the abscissa (*i.e.* the $x$-coordinate), and plot it.

(d) Find the maximum value of the $\chi^2(H_0)-\chi^2(H_1)$ array (using `numpy.amax`). Use (Wilk's theorem, and) previously seen `scipy` methods `stats.chi2.cdf` and `special.erfinv` to calculate the significance of the deviation.