Chapter 7

Hypothesis testing

7.1 Formulating a hypothesis

Up until now we have discussed how to define a measurement in terms of a central value, uncertainties, and units, as well as how to extend these concepts to encompass confidence levels (both one and two sided). A related aspect of performing a measurement is to test a theory or as it is usually phrased, a ***hypothesis***. For example, consider the case where a theorist writes a paper proposing the existence of some effect that can be tested via some physical process. It is then down to an experimentalist to develop a method that can be used to test the validity of that theory. In this example the default hypothesis (usually referred to as the ***null hypothesis*** and often denoted by H_0) would be that the theory is valid, and the experimenter would then embark on a measurement that could be used to test the null hypothesis.

Having defined a null hypothesis, by default the complement of that hypothesis exists as an ***alternative hypothesis*** often denoted by H_1. Given data we can test the null hypothesis to see if the data are in agreement or in disagreement with it. In order to qualify what is meant by data being in agreement with the null hypothesis, we need to quantify what we consider to be agreement. There are four possible outcomes to any given test with data: The null hypothesis can be compatible with or incompatible with the data, and the data can in reality either be based on the null hypothesis or not. If an effect is real, then we can make a measurement, and compare the results of that measurement with the null hypothesis. Some fraction of the time we do this, we will observe that the data agree with the experiment at some pre-defined confidence level. Thus some fraction of experiments where we do everything correctly we will obtain the wrong conclusion. A misclassification of this kind is called a ***type I error*** and the probability of obtaining a type I error is often denoted by α. The opposite can also be true - we can make a measurement to test that the null hypothesis is wrong, which is in reality the truth. On performing a correct measurement we can indeed find that the data support that there is no effect. Some fraction of the time however we will naturally reach the wrong conclusion, concluding that a non-existent hypothesis is actually true. This type of mistake is called a ***type II error***, which is often denoted by β. The rest of this chapter discusses such tests of the null hypothesis.

7.2 Testing if the hypothesis agrees with data

In order to determine if a hypothesis is compatible with data we must first specify to what degree of belief we aim to make a statement, or alternatively what is our acceptable level of drawing a wrong conclusion. If we want to be certain at the 90% CL that we are able to establish an effect proposed, then by definition 10% of the time we would expect to make a measurement of an effect that was real, and draw the conclusion that the data and theory were incompatible. This corresponds to a type I error rate $\alpha = 10\%$. If we incorrectly reject the hypothesis of a real effect we call this a ***false negative***. As the consequence of drawing such a conclusion can be quite profound, we must be precise and state what criteria are used to draw our conclusions.

Consider the following example: that of an ensemble of Monte Carlo based simulations of an expected outcome x from an experiment based on a theoretical prejudice (i.e. a null hypothesis), and the result of a measurement from data. The resulting distribution of the ensemble of Monte Carlo simulations is shown in Fig. 7.1, along with an arrow that indicates the measurement result. On further inspection of the simulation, we find that 4.89% of the simulated experiments have a value of x that is greater than that indicated by the arrow. So the estimate of the probability of obtaining a result greater than the one observed in data, assuming that the Monte Carlo simulation is a valid representation of the measurement is: $\alpha = 4.89\%$. Note that there is a computational uncertainty in the probability estimate that is related to the number of simulated experiments N, and this uncertainty may be reduced by increasing N. With regard to the data in Fig. 7.1, do we conclude that the data disagree with the theory, or do we conclude that the data are compatible? This brings us back to consider what is an acceptable level of false negatives we are willing to tolerate. If we wanted to ensure that the level of false negatives was 5%, then we would like to ensure that 95% of the time we would correctly identify an effect. As $\alpha < 5\%$, we would conclude in this case that the measurement was not compatible with an effect, and report a result inconsistent with H_0. Thus if H_0 was actually correct we would have been *unlucky* and reported a false negative. The chance of this happening would be about one in twenty. If on the other hand we are prepared to accept only a smaller rate of false negatives, say 1%, then we would conclude that this measurement is compatible with expectations of the simulation, and thus compatible with H_0. The conclusion drawn from the data will depend on the CL used for the test. It is important to set the CL, and hence maximum value of α prior to making the measurement to avoid biasing your conclusions according to any underlying prejudice you might have. To retrospectively decide upon on CL opens one up to experimenter bias, which is discussed in section 5.6.

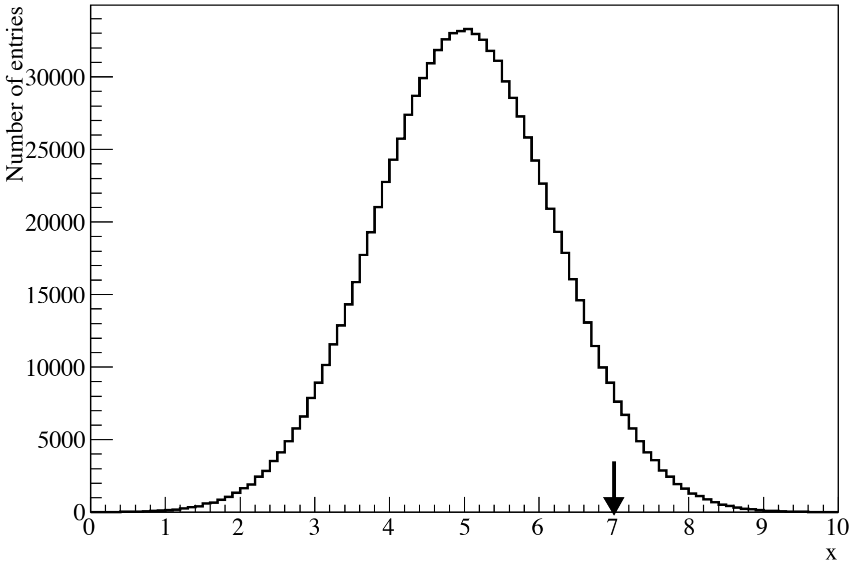

Figure 7.1: A histogram of simulated data compared with (arrow) the result of a measurement for the example discussed in the text.

One may consider how serious a false negative is, and this depends on the situation. In the example of an experiment, a false negative will usually be re-tested at a later point in time, and a result will be refuted (unless the experimenter is biased by the prior existence of the false negative result). While it would be desirable to never obtain a false negative when testing for a particular medical condition, unfortunately that is an unrealistic goal. Thus in a medical situation, as a false negative may result in a patient with a serious condition being incorrectly given a clean bill of health, and suffering greatly as a result, one would like to minimise the rate of such an outcome. The balancing situation is discussed in the following section.

At this point it is worth introducing the concept of a ***p-value***. The p-value is the probability of obtaining a fluctuation in data under the null hypothesis that is as, or more, extreme than that observed by a given experiment. If the p-value is small, then one might consider rejecting the null hypothesis. This situation will arise if the result obtained in data is found to be in the extreme tail of possible outcomes under the null hypothesis. For example if one were to find a p-value of 0.0027 for an observable with a Gaussian distribution, then this corresponds to a two tailed fluctuation at the level of 3σ or more. This is not indicative of supporting the null hypothesis and an appropriate conclusion of such a result would be to reject H_0. The p-value in the previous example was 0.0489.

7.3 Testing if the hypothesis disagrees with data

In the previous section we discussed how to approach the question of testing if a hypothesis agrees with data, and discovered that one necessary condition was to determine an acceptable level or probability of assigning a false negative, i.e. of rejecting a measurement that was compatible with the existence of an effect. Now we turn to the reverse situation, that of disproving a theory. In this case the null hypothesis H_0 is that some effect that we are searching for exists, where in reality it does not. If we correctly identify that the hypothesis is invalid, then we will have obtained a true negative test. However if we incorrectly establish the validity of H_0, then we will have obtained a ***false positive*** which is a type II error. As with the case of false negatives, the severity of a false positive depends on the scenario encountered.

7.4 Hypothesis comparison

Suppose we have two theories H_0 and H_1 proposed as descriptions of some data. We can compute $P(data|H_0)$ and $P(data|H_1)$ as defined by the theories, and we are able to specify a prior for both theories, $P(H_0)$ and $P(H_1)$. In general if H_0 and H_1 do not form the complete set of possible theories, then $P(data)$ is something that is not calculable. However we can note that (see Eq. 2.2.2)

$$P(H_0|data) \propto P(data|H_0)P(H_0), \tag{7.4.1}$$

$$P(H_1|data) \propto P(data|H_1)P(H_1). \tag{7.4.2}$$

So while we can't compute $P(H_0|data)$ or $P(H_1|data)$, we can compute the ratio R between these posterior probabilities

$$R = \frac{P(H_0|data)}{P(H_1|data)} = \frac{P(data|H_0)P(H_0)}{P(data|H_1)P(H_1)}. \tag{7.4.3}$$

The relative probability R of theory H_0 to H_1 can be used to compare the two theories, and one may be able to determine which theory is a better description of the data. For example

- If $R > 1$ then theory H_0 is preferred.
- If $R < 1$ then theory H_1 is preferred.
- If $R \simeq 1$ then there is insufficient data to discriminate between the two theories.

Example: Given a set of data resulting from the measurement of some observable

$$\Omega = \{-1.0, -0.9, -0.7, -0.1, 0.0, 0.1, 0.2, 0.5, 0.6, 1.0\}, \tag{7.4.4}$$

where the total number of data $N = 10$, determine which of the following models is a better description of the data:

- H_0: The data are distributed according to a Gaussian PDF with a mean of zero and a width of one.
- H_1: The data are uniformly distributed.

Given this information, for each element of Ω we are able to compute R where for the sake of illustration we assume uniform priors $P(H_0)$ and $P(H_1)$, while

$$\begin{aligned} P(data|H_0) &= G(\omega_i; 0, 1), && (7.4.5) \\ P(data|H_1) &= 1/N. && (7.4.6) \end{aligned}$$

Equation (7.4.3) can be used to compute R_i for a given element of data ω_i, and we are interested in the comparison of the total probability for the data, which is given by the product of the R_i,

$$\begin{aligned} R &= \prod_{i=1}^{N} R_i, && (7.4.7) \\ &= \prod_{i=1}^{N} \frac{P(data|H_0)P(H_0)}{P(data|H_1)P(H_1)}, && (7.4.8) \\ &= \prod_{i=1}^{N} NG(\omega_i; 0, 1). && (7.4.9) \end{aligned}$$

Table 7.1 shows the results of each step of this calculation, resulting in the final value of $R = 140290$. As $R > 1$, theory H_0 is the preferred description of the data. In this example the prior dependence in Eq. (7.4.3) cancels. Generally the priors would not cancel if the form of $P(H_0)$ was chosen to be different from that of $P(H_1)$.

Table 7.1: The values of $P(data|theory)$ used to compute R_i for the sample of data Ω.

ω_i	$P(data\|H_0)$	$P(data\|H_1)$	R_i
-1.0	0.242	0.1	2.42
-0.9	0.266	0.1	2.66
-0.7	0.312	0.1	3.12
-0.1	0.397	0.1	3.97
0.0	0.399	0.1	3.99
0.1	0.397	0.1	3.97
0.2	0.391	0.1	3.91
0.5	0.352	0.1	3.52
0.6	0.333	0.1	3.33
1.0	0.242	0.1	2.42

Consider now a second data sample

$$\Omega' = \{-4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}. \qquad (7.4.10)$$

Just as before we are able to choose a prior for each of the models (for example flat or Gaussian), and compute contributions R_i to each of the events in Ω'. The value of the ratio R in this case is 3.6×10^{-13}. Thus for this data sample we find that $R << 1$, and hence the preferred theoretical description of the data is H_1, i.e. the data are flat.

7.5 Testing the compatibility of results

Often, instead of being faced with a theory and a measurement, we may encounter a situation where we have two measurements of some observable and we want to understand if they are compatible or not. If two measurements of the same observable are compatible, then any differences would be the result of statistical fluctuations in the data. If one or both of the measurements is affected by systematic effects, then the corresponding uncertainty σ_m would be the sum (usually in quadrature[1]) of statistical and systematic uncertainties. If we assume uncorrelated uncertainties between two measurements $m_1 = x_1 \pm \sigma_1$, and $m_2 = x_2 \pm \sigma_2$, then the difference between the measurements is given by

$$\Delta m = x_1 - x_2 \pm \sqrt{\sigma_1^2 + \sigma_2^2}, \tag{7.5.1}$$

$$= \Delta x \pm \sigma_m. \tag{7.5.2}$$

In order to test the compatibility of these measurements we must compare the magnitude of Δx to that of σ_m.

There is an intrinsic possibility that two correct measurements will be in disagreement. The larger the disagreement the smaller the probability that the two results are compatible with each other. If we think in terms of Gaussian uncertainties, then it follows that by requiring two measurements to be within $\pm 1\sigma$ (68.3% CL) in order to consider them to be in agreement, we would classify 31.7% of legitimate results as being as incorrect or inconsistent. A more sensible way of looking at this problem is to define an acceptable level of mis-classification, given by $1 - CL$, and use this to test if the results are compatible. Typically we consider two results to be incompatible if they differ by more than 3σ from each other (where σ is the total uncertainty of the two measurements). Given a large number of compatible measurement comparisons we would expect this criteria to give an incorrect classification of agreement about three in every thousand times. In general if we find results that are between 2 and 3σ apart, we tend to scrutinise them a little more closely to understand if there are systematic effects that may have been overlooked in one or both of the measurements. If both results hold up to scrutiny, then we conclude that they must be compatible.

It is completely natural to expect that about three in every thousand measurements made of some observable will be more than 3σ from the true value of that observable. The consequence of this fact is that for about every three hundred correct measurements you make, one of them can be expected to be statistically inconsistent with previous or subsequent measurements at a level of 3σ. This does not mean that that measurement is wrong, this is simply a reflection of the syntax used to describe a measurement.

Example: If some observable O has a true value of μ, and it is possible to perform a measurement of that observable which would result in an expected standard deviation of σ, where systematic effects on the measurement are negligible, then we can perform an ensemble of measurements and observe how these results compare to the expectation of $\mathcal{O} = \mu$. If we perform one thousand measurements, then according to Table 5.1, we would expect on average to obtain 2.7 measurements that are more than 3σ from μ. Figure 7.2 shows the results of one thousand such measurements where $\mu = 0$ and $\sigma = 1$. We can see in this instance that there are two measurements that are between 3 and 4σ from μ, a result consistent with expectations.

7.6 Establishing evidence for, or observing a new effect

In general there are no globally accepted confidence levels that are used to establish the existence of a new effect. If we analyse data, with a null hypothesis that no effect exists, and find that the data are incompatible with that hypothesis at some confidence level, then we can interpret that result in a reasonable way. The reader should keep in mind that different fields use different confidence levels to test the significance of a result, and so "standard practice" for one specialisation is probably not directly applicable to another.

[1] By doing this one is making the tacit assumption that the systematic uncertainty is Gaussian in nature.

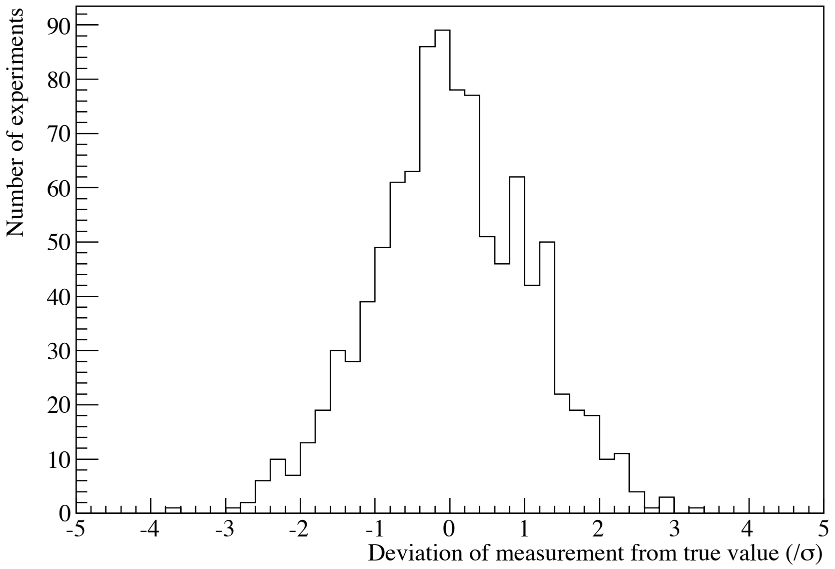

Figure 7.2: The resulting distribution on performing one thousand measurements of an observable $\mathcal{O}$.

For example, in particle physics, one often chooses to consider a deviation from the absence of an effect at the level of 3σ as "*evidence*" for the existence of that effect. Assuming that the measurement uncertainties are Gaussian in nature, this translates into a claim of evidence for some effect to be a false positive statement about three in every thousand correct measurements. One may require a more stringent classification for the existence of an effect than at the level of 3σ. If one finds a deviation from the absence of an effect at the level of 5σ, then we might regard this as "*observation*" of an effect. Only once every million measurements will we report a false positive result at this level. In contrast another field may choose to use a 2σ threshold to establish an effect. If you insist on describing a result with a phrase such as evidence for or observation of then you should remember that this label has little meaning, and the interpretation of the phrase is ambiguous. The proper way to address this is to specify the result in terms of the CL or p-value used in order to draw your conclusions.

7.6.1 Peak finding

Broadly speaking there are two types of peak finding problems (i) searching for a peak at some known position in parameter space, and (ii) searching for a peak at some unknown location in parameter space (see section 7.6.2). An example of the former scenario is the decay of an unstable particle into some identified final state, where that state is reconstructed from measurements of energy deposits in one or more detectors. For example a π^0 meson can decay into two electrons and a photon[2]. If one has accurate enough measurements of the energy and three-momenta of each of the decay products then one can compute the invariant mass of the π^0 from data. The mass of the π^0 is known to be approximately 135 MeV/c^2 (Beringer *et al.*, 2012), and the width of any peak in the reconstructed invariant mass distribution is given by detector resolution. Given this information it is possible to search through data and compute the invariant mass m_0 for all combinations of an $e^+e^-\gamma$ final state. This can be obtained using the relativistic energy-mass relation $E^2 = m_0^2c^4 + p^2c^2$, where c is the speed of light in vacuum, E is the sum of particle energies and p is the sum of particle three-momenta. If there are $\pi^0 \to e^+e^-\gamma$ events in the data, then this should be evident by looking at this histogram where one should see a peak at the appropriate location. The other combinations of events

[2] This particular decay, $\pi^0 \to e^+e^-\gamma$, is sometimes referred to by name as a π^0 Dalitz decay.

that enter the histogram would be from random combinations of candidate electrons, positrons and photons. These background events would have a flat distribution. In order to ascertain if one had a signal or not, one would have to determine if a peak at the appropriate mass was large enough to be unlikely to correspond to a fluctuation of the background. Hence the null hypothesis in this scenario would be that the histogram only contains background events, and there is no signal. An example this problem can be found in Figure 7.3, where there is a peak visible at 135 MeV/c^2 as expected. The width of this distribution is the result of the finite resolution of the detector. One can see the signal is significant without having to resort to a significance or p-value calculation.

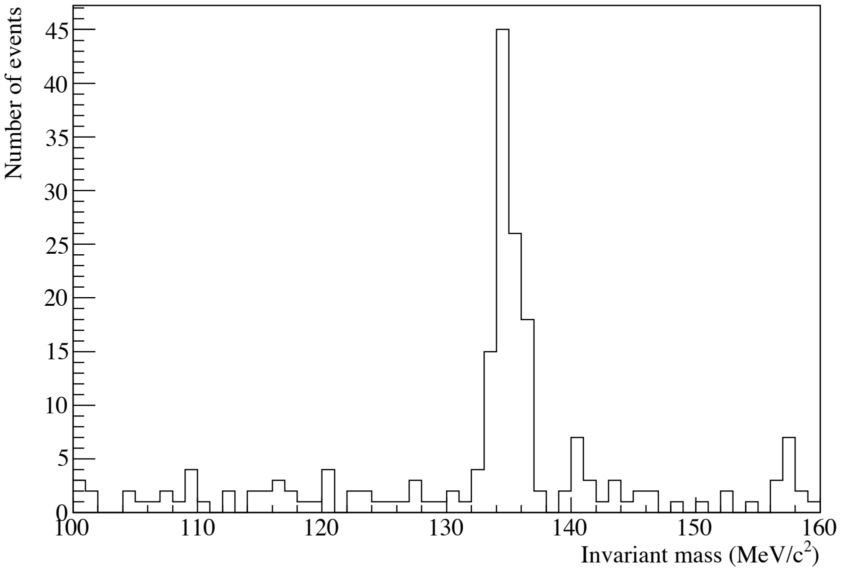

Figure 7.3: The distribution of events obtained from a simulated experiment illustrating the peak finding problem for $\pi^0 \to e^+e^-\gamma$ decays.

7.6.2 Trial factors and nuisance parameters

An example of the second type of search can be seen by once again considering Figure 7.3. The π^0 peak at 135 MeV/c^2 was expected and we can ignore that sample of events for the following discussion. Having done that one can see that there are two clusters of events which appear near 140 and 160 MeV/c^2. Are these also particles? In this simulation there are 100 background and 100 $\pi^0 \to e^+e^-\gamma$ signal events generated, and the histogram has 60 bins. Hence on average each bin has an expectation of 5/3 backgrounds events in it, and any excess beyond that could be considered a potential signal. Five background events are expected for both of the excesses, where each excess spans three adjacent bins. There are twelve events appearing in each cluster bins. One can ask what the probability is for the background level to fluctuate up to the observed level of events. The probability of observing 12 or more events given $\lambda = 5$ for a Poisson distribution is 0.55%. This p-value is quite small given the CL values often used to place bounds on unseen effects, and one might be tempted to report that another particle (or two) exists in the data. The problem is that while only one sample of data is being analysed, we are looking for peaks in many different sets of bins and we need to take into account that we are performing a number of simultaneous searches when we compute the p-value for any apparent excess. In general one can compute a trial factor, the ratio of the probabilities of observing an excess at a given point and observing an excess anywhere in a range, to help interpret such a situation. In order to determine the trial factor and hence compute the correct p-value for a given result one can perform an ensemble of background only simulations to determine the largest upward fluctuations of background, and compare that with data. This can be both time and resource intensive. It is also possible to approach the problem from a more formal approach as described by Gross and Vitells (2010) and references therein. This class of problem is referred to as hypothesis testing when a nuisance parameter is present only under the alternative, or colloquially as the look elsewhere effect. The term ***nuisance parameter*** refers to a parameter that is not the main result, but must be considered in the extraction of the main result from data, for example see Edwards (1992). In the example discussed above, the main parameter of interest when searching for a peak is the mass, and the width of the mass peak would be classed as a nuisance parameter. When computing a Bayesian constraint any nuisance parameters can be removed from the problem by integrating them out. This process is sometimes referred to as marginalisation, and the resulting probability distribution is called the marginal distribution.