Análisis estadístico de datos

1 Measurements in sciences: errors

It is well-established fact of scientific investigation that the first time an experiment is performed the results often bear all too little resemblance to the "truth"being sought. As the experiment is repeated, with successive refinements of technique and method, the results gradually and asymptotically approach what we may accept with some confidence to be a reliable description of the events. It is certainly true that for all physical experiments, errors and uncertainties exist that must be reduced by improved experimental techniques and repeated measurements, and those errors remaining must always be estimated to establish the validity of our results.

Error is defined in common language as "the difference between an observed or calculated value and the true value". But usually we **do not** know the "truth" value; otherwise there would be no reason for performing the experiment. We may know approximately what it should be from earlier experiments or from theoretical predictions. The systematic determination of the data through repeated measurements will determine how much confidence we can have in our experimental results.

Our interest is in *uncertainties* introduced by random fluctuations in our measurements, and *systematic errors* that limit the precision and accuracy of our results in more or less well-defined ways. Generally, we refer to the uncertainties as the errors in our results, and the procedure for estimating them as *error analysis*.

1.1 Accuracy vs. precision

It is important to distinguish between the term *accuracy* and *precision*. The accuracy of an experiment is a measure of how close the result of the experiment is to the true value; the precision is a measure of how well the result has been determined, without reference to its agreement with the true value. The precision is also a measure of the reproducibility of the result in a given experiment. Figure 1 shows both concepts: In left panel, the data has been measured to a high degree of precision as illustrated by the small error bars, and are in excellent agreement with the expected variation of y with x, but are clearly inaccurate, deviating from the line by a constant offset. On the other hand, the data points on the right are imprecise as illustrated by the large error bars, but are scattered about the predicted distribution.

It is obvious that we must consider the accuracy and precision simultaneously for any experiment. In general, when we quote the *uncertainty* or *error* in a experiment result, we are referring to the precision with which that result has been determined. *Absolute* precision indicates the magnitude of the uncertainty in the result in the same units as the result; *relative* precision indicates the uncertainty in terms of a fraction of the result.

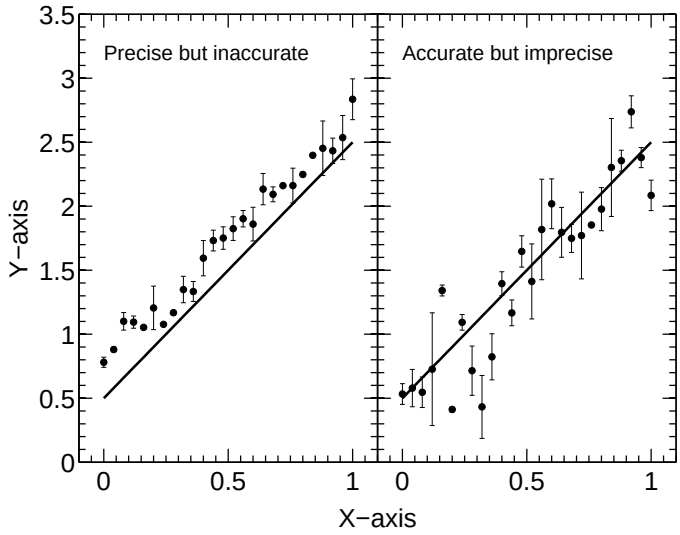

Figure 1: Left: Precise but inaccurate data. Right: Accurate but imprecise data. True values are represented by the straight lines

1.2 Systematic errors

The accuracy of an experiment is generally dependent on how well we can control or compensate for systematic errors, errors that will make our results different from the "truth" values with reproducible discrepancies. Errors of this type are not easy to detect and not easily studied by statistical analysis. They may result from fault calibration of equipment or from bias on the part of the observer. The must be estimated from an analysis of the experimental conditions and techniques. A major part of the planning of an experiment should be devoted to understanding and reducing sources of systematic errors.

1.3 Random errors

The precision of an experiment depends upon how well can overcome random errors, fluctuations in observations that yield different results each time the experiment is repeated, and thus require repeated experimentation to yield precise results. A given accuracy implies an equivalent precision and, therefore, also depends to some extent on random errors.

The problem of reducing random errors is essentially one of improving the experimental method and refining the techniques, as well as simply repeating the experiment. If the random errors result from instrumental uncertainties, the may be reduced by using more reliable and more precise measuring instruments. If the random errors result from statistical fluctuations in a limited number of measurements, they may be reduced by making more measurements.

1.4 Uncertainties

Uncertainties in experimental results can be separated into two categories: those that result from fluctuations in measurements, and those associated with the theoretical description of our result.

Usually we cannot know what the "true" is, and can only estimate the errors inherent in the experiment. If we repeat an experiment, the results may well differ from those of the first attempt. We express this difference as *discrepancy* between the two results. Discrepancies arise because we can determine a result only with a given *uncertainty*. For example, when we compare different measurements of a standard physical constant, or compare our result with the accepted value, we should refer to the differences as discrepancies, not errors or uncertainties.

In general, we shall be interested in obtaining the maximum amount of useful information from the data on hand without being able either to repeat the experiment with better equipment or to reduce the statistical uncertainties by making more measurements. It is reasonable to expect that the most reliable results we can calculate from a given set

of data will be those for which the estimated errors are the smallest. It must be noted, however, that even our best efforts on reducing the errors will yield only estimates of the quantities investigated.

1.5 Parent and sample distributions

If we make a measurement of x_i of a quantity x, we expect our observation to approximate the quantity, but we don not expect the experimental data point to be exactly equal to the quantity. If we make another measurement, we expect to observe a discrepancy between the two measurements because the random errors, and we do not expect either determination to be exactly correct, that is, equal to x. As we make more and more measurements, a pattern will emerge from the data. Some of the measurements will be too large, some will be too small. On the average, however, we expect them to be distributed around the correct value, assuming we can neglect or correct for systematic errors.

If we could make an infinite number of measurements, then we could describe exactly the distribution of the data points. This is not possible in practice, but we can hypothesize the existence of such a distribution that determines the probability of getting any particular observation in a single measurement. This distribution is called the *parent distribution*. Similarly, we can hypothesize that the measurements we have made are samples from the parent distribution and they form the *sample distribution*. In the limit of an infinite number of measurements, the sample distribution becomes the parent distribution.

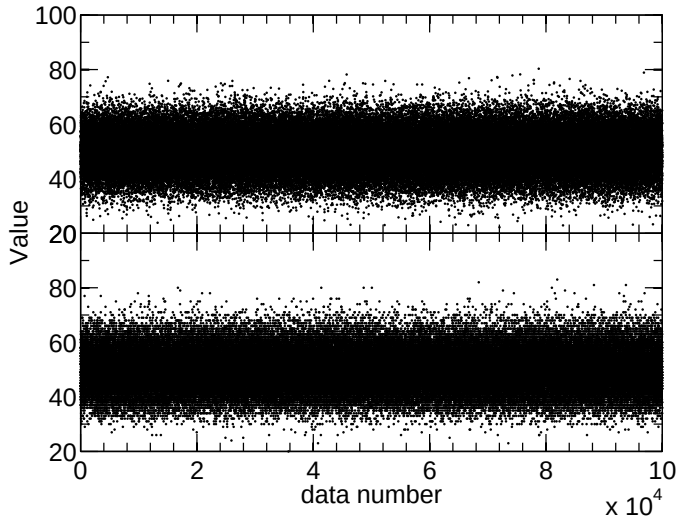

Figure 2: Top: Data sample from a Normal distribution. Bottom: Data sample from a Poisson distribution.

It is convenient to think in terms of a *probability density function p(x)* normalized to unit area (i.e. so that the integral of the entire curve is equal to 1) and defined such that in the limit of a very large number N of observations, the number ΔN of observations of the variable x between x and $x + \delta x$ is given by $\Delta N = Np(x)\delta x$.

In order to determine the parameters of the parent distribution, we assume that the results of experiments asymptotically approach the parent quantities as the number of measurements approach infinite; that is, the parameters of the experimental distribution equal the parameters of the parent distribution in the limit of an *infinite number of measurements*. If we specify that there are N observations in a given experiment, then we can denote this by

$$(parent\ parameter) = \lim_{N\to\infty} (experimental\ parameter)$$

If we make N measurements and label them x_1, x_2, x_3, and so forth, up to a final measurement x_N, then we can identify the sum of all these measurements as

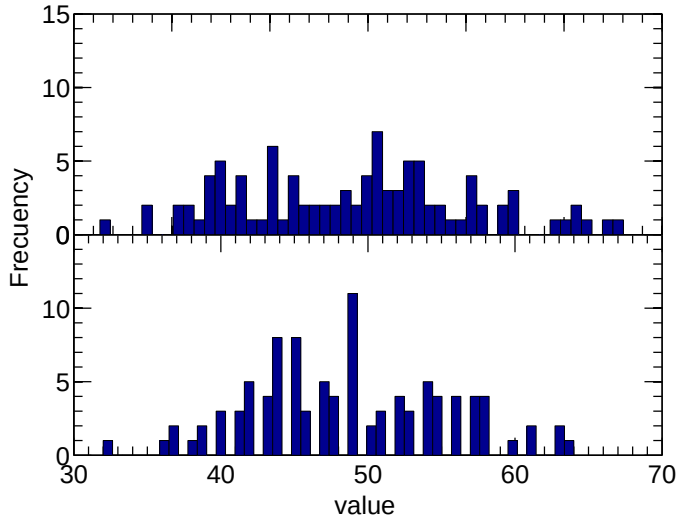

Figure 3: Histogram for a low number of data for a Normal distribution (top) and for a Poisson distribution (bottom).

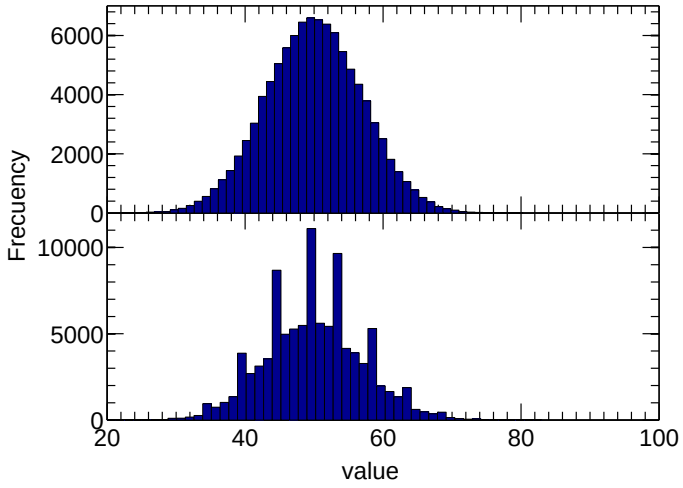

Figure 4: In Top: Histogram from a Normal distribution. Bottom: Histogram from a Poisson distribution.

$$\sum_{i=1}^{N} x_i \equiv x_1 + x_2 + x_3 + \ldots x_N$$

where the left-hand side is interpreted as the sum of the observations x_i over the index i from $i = 1$ to $i = N$ inclusive.

We can simply the notation by writing:

$$\sum x_i \equiv \sum_{i=1}^{N} x_i$$

1.5.1 Mean, median, and mode

The mean $\bar{x}$ of the experimental distribution is given as the sum of N determinations x_i of the quantity x divided by the number of determinations:

$$\bar{x} \equiv \lim \frac{1}{N} \sum x_i \tag{1}$$

and the mean μ of the parent population is defined as the limit:

$$\mu \equiv \lim_{N\to\infty}\left(\frac{1}{N}\sum x_i\right) \tag{2}$$

The mean is therefore equivalent to the centroid or *average* value of the quantity x.

The *median* of the parent population $\mu_{1/2}$ is defined as that value for which, in the limit of an infinite number of determinations x_i, half the observations will be less than the median and half will be greater. In terms of the parent distribution, this means that the probability is 50% that any measurement x_i will be larger or smaller than the median:

$$P(x_i < \mu_{1/2}) = P(x_i > \mu_{1/2}) = 1/2 \tag{3}$$

so that the median line cuts the area of the probability density distribution in half.

The *mode* or *most probable value* μ_{max}, of the parent population is that value for which the parent distribution has the greatest value. In the limit of a large number of observations, this value will probably occur most often:

$$P(\mu_{max}) \geq P(x \neq \mu_{max}) \tag{4}$$

For a symmetrical distribution these parameters would all be equal by the symmetry of their definitions. For an asymmetric distribution the median generally falls between the most probable value and the mean. The most probable value correspond to the peak of the distribution, and the areas on either side of the median are equal.

1.5.2 Deviations

The *deviation* d_i of any measurement x_i from the mean μ of the parent distribution is defined as the difference between x_i and μ:

$$d_i \equiv x_i - \mu \tag{5}$$

If μ is the true value of the quantity, d_i is also the true error in x_i.

The average of the deviations d must vanish by virtue of the definition of the mean in Eq. 2

$$\lim_{N\to\infty}\bar{d} = \lim_{N\to\infty}\left(\frac{1}{N}\sum(x_i-\mu)\right) = \lim_{N\to\infty}\left(\frac{1}{N}\sum x_i\right) - \mu = 0 \tag{6}$$

The average deviation α, therefore, is defined as the average of the absolute values of the deviations:

$$\alpha \equiv \lim_{N\to\infty}\left(\frac{1}{N}\sum \mid x_i - \mu \mid\right) \tag{7}$$

The average deviation is a measure of the *dispersion* of the expected observations about the mean. The presence of the abslute value sing makes its use inconvenient for statistical analysis.

A parameter that is easier to use analytically and that can be justified fairly well on theoretical grounds to be a more approximate measure of the dispersion of the observations is the *standard deviation* σ. The variance σ^2 is defined as de limit of the average of the squares of the deviations from the mean μ:

$$\sigma^2 \equiv \lim_{N\to\infty}\left(\frac{1}{N}\sum(x_i-\mu)^2\right) = \lim_{N\to\infty}\left(\frac{1}{N}\sum x_i^2\right) - \mu^2 \tag{8}$$

and the standard deviation σ is the square root of the variance. The standard deviation is then the root mean square of the deviations, and is associated with the *second moment* of the data about the mean. The corresponding expression for the variance s^2 of the sample population is given by

$$s^2 \equiv \frac{1}{N-1}\sum(x_i-\bar{x})^2 \tag{9}$$

where the factor $N-1$, rather than N is required in the denominator to account for the fact that the parameter $\bar{x}$ has been determined from the data and not independently. We note that the symbol σ (instead of s) is often used to represent the best estimate of the standard deviation of the parent distribution determined from a sample distribution.

1.5.3 Significance

The mean μ and the standard deviation, as well as the median, the most probable value, and the average deviation, are all parameters that characterize the information we are seeking when we preform an experiment.

In general, the best we can say about the mean is that it is one of the parameters that specifies the probability distribution: It has the same units as the "true" value and, in accordance with convention, we shall consider it to be the best estimate of the "true"value under the prevailing experiment conditions.

The variance s^2 and the standard deviation s characterize the uncertainties associated with our experimental attempts to determine the "true" values. For a given number of observations, the uncertainty in determining the mean of the parent distribution is proportional to the standard deviation of that distribution. The standard deviation s is, therefore, an appropriate measure of the uncertainty due to fluctuations in the observations in our attempt to determine the "true" value.

1.6 Discrete and continuous distributions

We can define the mean μ and the standard deviation σ in terms of the distribution $p(x)$ of the parent population. The probability density $p(x)$ is defined such that in the limit of a very large number of observations, the fraction dN of observations of the variable x that yield values between x and $x+dx$ is given by $dN = Np(x)dx$.

The mean μ is the expectation value $\langle x\rangle$ of x, and the variance σ^2 is the expectation value $\langle (x-\mu)^2\rangle$ of the square of deviations of x from μ. The expectation value $\langle f(x)\rangle$ of any function of x is defined as the weighted average of $f(x)$, over all possible values of the variable x, with each value of $f(x)$ weighted by the probability density distribution $p(x)$

Discrete distributions: If the probability function is a discrete function $P(x)$ of the observed value x, we replace the sum over the individual observations $\sum x_i$ in Eq. 2 by a sum over the values of the possible observations multiplied by the number of times these observations are expected to occur. If there are n possible different observable values of the quantity x, which we denote by x_j (where the index j runs from 1 to n with no two values of x_j equal), we should expect from a total of N observations to obtain each observable $NP(x_j)$ times. The mean can then be expressed as

$$\mu = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} x_i = \lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}(x_j NP(x_j)) = \lim_{N\to\infty}\sum(x_j P(x_j)) \tag{10}$$

Similarly, the variance σ in Eq. 8 can be expressed in terms of the probability function $P(x)$:

$$\sigma^2 = \lim_{N\to\infty}\sum_{j=1}^{N}((x_j-\mu)^2 P(x_j)) = \lim_{N\to\infty}\sum_{j=1}^{N}(x_j^2 P(x_j)) - \mu^2 \tag{11}$$

In general, the expectation value of any function of f(x) is given by:

$$\langle f(x)\rangle = \sum_{j=1}^{N}(f(x_j)P(x_j)) \tag{12}$$

Continuous distributions: If the probability density function is a continuous smoothly varying function $p(x)$ of the observable value x, we replace the sum over the individual observations by an integral over all values of x multiplied

by the probability $p(x)$. The mean μ becomes the first moment of the parent distribution:

$$\mu = \int_{-\infty}^{\infty} xp(x)dx \tag{13}$$

and the variance σ^2 becomes the second central product moment:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \int_{-\infty}^{\infty} x^2 p(x)dx - \mu^2 \tag{14}$$

The expectation value of any function of x is:

$$\langle f(x) \rangle = \int_{-\infty}^{\infty} f(x)p(x)dx \tag{15}$$

By considering the data to be a sample form the parent population with the values of the observations distributed according to the parent population, we can estimate the shape and dispersion of the parent distribution to obtain useful information on the precision and reliability of our results. Thus, we consider the sample mean $\bar{x}$ to be our best estimate from the data of the mean μ, and we consider the sample variance s^2 to be our best estimate from the data of the variance σ^2, from which we can estimate the uncertainty in our estimate of μ.

1.7 Exercises

1. a) Generate 100 random numbers using a Normal Distribution with parameters mean $\mu = 0.5$ and standard deviation $\sigma = 0.2$, b) Verify the mean, the standard deviation and the variance of the distribution, c) Display the normalized histogram of the samples using 5, 10, and 20 bins.

2. a) Generate 100 random numbers using a Poisson Distribution with parameter $l = 5$, b) Verify the mean, median, the standard deviation and the variance of the distribution, c) Display the normalized histogram of the samples using 20 bins, c) Display the normalized histogram of the samples using 5, 10, and 20 bins.

3. Verify Eq. 11 (i.e., for $N \to \infty$ the variance can be written as $\sigma^2 = \bar{x^2} - \mu^2$) by increasing the sample population in a Normal Distribution.

4. Calculate the probability that a value x be within μ and $\mu \pm n \times \sigma$, (being n=1,2,3...) in a Normal distribution. Use different values for the sample population.

Solution to Ex. 1

```
import numpy as np

mu, sigma = 0.5, 0.2    # mean and standard deviation of the distribution.
gs = np.random.normal(mu, sigma, 100) # Generate the distribution.

abs(np.mean(gs) - mu) < 0.01 # Verify the mean.
(np.std(gs) - sigma) < 0.01 # Verify the standard deviation.
(np.var(gs) - sigma**2) < 0.01 # Verify the variance.

import matplotlib
from pylab import *
ion()
hist(gs, 20, normed=True)
```

Solution to Ex. 2

```
import numpy as np
l = 5    # mean and standard deviation of the distribution.
ps = np.random.poisson(l, 100) # Generate the distribution.

abs(np.mean(ps) - l) < 0.01 # Verify the mean.
abs(np.median(ps) - l) < 0.01 # Verify the median.
(np.std(ps) - sqrt(l)) < 0.01 # Verify the standard deviation.
(np.var(ps) - l) < 0.01 # Verify the variance.

import matplotlib
from pylab import *
ion()
hist(ps, 20, normed=True)
```

Solution to Ex. 3

```
import numpy as np

mu, sigma = 1.5, 0.3
sam=np.logspace(1, 7, num=7, endpoint=True, base=10.0)

for i in sam:
    gs = np.random.normal(mu, sigma, i) # Generate the distribution.
    vgs=np.mean(gs**2) - mu**2
    print((np.var(gs)-vgs)<0.01)
```

Solution to Ex. 4

```
i=1e6
n=1
import numpy as np
mu, sigma = 1.5, 0.3
gs = np.random.normal(mu, sigma, i) # Generate the distribution.
size(where (abs(gs-mu) < n*sigma))/i
```

2 Probability distributions

Of the many probability distributions that are involved in the analysis of experimental data, three play a fundamental role: the *binomial distribution*, The *Poisson distribution*, and the *Gaussian distribution*. Of these, the Gaussian, or normal error, distribution is undoubtedly the most important in statistical analysis of data. Practically, it is useful because it seems to describe the distribution of random observations for many experiments, as well as describing the distributions obtained when we try to estimate the parameters of most other probability distributions.

The Poisson distribution is generally appropriate for counting experiments where the data represent the number of items or events observed per unit interval. It is important in the study of random processes such as those associated with the radioactive decay of elementary particles or nuclear states, and is also applied to data that have been sorted into ranges to form a frequency table or a histogram.

The binomial distribution is generally applied to experiments in which the result is one of a small number of possible final states, such as the number of "heads"or "tails"in a series of coin tosses, or the number of particles scattered forward or backward relative to the direction of the incident particle in a particle physics experiment. The Poisson and the Gaussian distribution can be considered as limiting cases of the binomial distribution.

2.1 Binomial distribution

Suppose we toss n coins into the air, where n is some integer. Alternatively, suppose that we toss one coin n times. What is the probability that exactly x of these coins will land heads up, without distinguishing which of the coins actually belongs to which group? We can consider the probability $P(x;n)$ to be a function of the number n of coins tossed and of the number x of coins that land heads up. For a given experiment in which n coins are tossed, this probability $P(x;n)$ will vary as a function of x. Of course, x must be an integer for any physical experiment, but we can consider the probability to be smoothly varying with x as a continuous variable for mathematical purposes.

Probability.- The probability $P(x;n)$ that we should observe x coins with heads up and $n-x$ with tails up is the product of the number of different combinations $C(n,x)$ that contribute to that set of observations multiplied by the probability for each of the combinations to occur, which is $(1/2)^n$. $C(n,x)$ is the number of permutations $Pm(n,x)$ divide by the degeneracy factor $x!$ of the permutations:

$$C(n,x) = \frac{Pm(n,x)}{x!} = \frac{n!}{x!(n-x)!} = \binom{n}{x} \tag{16}$$

We should separate the probability for each combination into two parts: one part is the probability $p^x = (1/2)^x$ for x coins to be heads up; the other part is the probability $q^{n-x} = (1-1/2)^{n-x} = (1/2)^{n-x}$ for the other $n-x$ coins to be tails up. For symmetrical coins, the product of these two parts $p^x q^{n-x} = (1/2)^2$ is the probability of the combinations with x coins heads up and $n-x$ coins tail up. With these definitions of p and q, the probability $P_B(x;n,p)$ for observing x of the n items to be in the state with probability p is given by the *binomial distribution*

$$P_B(x;n,p) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \tag{17}$$

where $q = 1-p$

Mean and Standard Deviation.- The mean of the binomial distribution is evaluated by combining the definitions of μ in Eq. 10 with the formula for the probability function of Eq. 17:

$$\mu = \sum_{x=0}^{n} \left(x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \right) = np \tag{18}$$

We interpret this to mean that if we perform and experiment with n items and observe the number of x successes, after a large number of repeated experiments the average $\bar{x}$ of the number of successes will approach a mean value μ given by the probability for for success of each item p times the number of items n.

The variance σ^2 of a binomial distribution is similarly evaluated by combining Eq. 11 and 17:

$$\sigma^2 = \sum_{x=0}^{n} \left((x-\mu)^2 \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \right) = np(1-p) \tag{19}$$

If the probability for a single success p is equal to the probability for failure $p = q = 1/2$, then the distribution is symmetric about the mean μ, and the median $\mu_{1/2}$ and the most probable value are both equal to the mean. In this case, the variance σ^2 is equal to half the mean: $\sigma^2 = \mu/2$. If p and q are not equal, the distribution is asymmetric with smaller variance.

Binomial distribution in python

```
# Draw samples from the Binomial distribution:

import numpy as np

n, p = 1, .5 # number of trials, probability of each trial

b=np.random.binomial(n, p, size=1000)  # result of flipping a coin 1 time, tested 1000 times.

# What is the probabilty that flipping a coin 5 times all the result be the same?
# Let's do 20,000 trials of the model, and count the number that generate 5 positive results.

sum(np.random.binomial(5,0.5,20000)==0)/20000.
# answer = 0.032 or 3.2%
```

2.2 Poisson distribution

The Poisson distribution represents an approximation to the binomial distribution for the special case where the average number of successes is much smaller that the possible number; that is, when $\mu \ll n$ because $p \ll 1$. For such experiments the binomial distribution correctly describes the probability $P_B(x; n, p)$ of observing x events per time interval out of n possible events, each of which has a probability p of occurring, but the large number n of possible events makes exact evaluation from the binomial distribution impractical. Furthermore, neither the number of n of possible events nor the probability p for each is usually known. What may be known instead is the average number of events μ expected in each time interval or its estimate $\bar{x}$. The Poisson distribution provides an analytical form appropriate to such investigations that describes the probability distribution in terms of just the variable x and the parameter μ.

The probability of observing x events in the time interval t is given by:

$$P_p(x; \mu) = \frac{\mu^x}{x!} e^{-\mu} \tag{20}$$

where τ is a constant proportionally factor that is associated with the mean time between events and $\mu = t/\tau$ is the average number of events observed in the time interval t.

The Poisson distribution, like the binomial distribution, is a *discrete* distribution. That is, it is defined only at integral values of the variable x, although the parameter μ is a positive, real number. The mean of the Poisson distribution is actually the parameter μ. The expectation value $\langle x \rangle$ of x is μ, while the variance is $\sigma^2 = \mu$, that is, the standard deviation σ is equal to the square root of the mean μ and the Poisson distribution has only a single parameter, μ.

Poisson distribution in python

```
#Draw samples from the distribution:

import numpy as np
import math as mt
mu=5
s = np.random.poisson(mu, 10000)

# Display histogram of the sample:

import matplotlib.pylab as plt
plt.ion()
count, bins, ignored = plt.hist(s, 10, normed=True)


# Draw the density probability function:

bin=np.arange(max(s))
n=[]
for i in bin:
    n.append(mt.factorial(i))        # This is necessary to calculate the factorial of an array
    .

plt.plot(bin, mu**bin/n * np.exp(-mu), linewidth=2, color='r')
```

2.3 Gaussian or normal error distribution

The Gaussian distribution is an approximation to the binomial distribution for the special limiting case where thee number of possible different observations n becomes infinitely large and the probability of success for each is finitely large so $np \ll 1$. It is also the limiting case for the Poisson distribution as μ becomes large. The Gaussian Probability density is defined as

$$p_G = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{21}$$

This is a continuous function describing the probability of obtaining the value x in a random observation from a parent distribution with parameters μ and σ, corresponding to the mean and standard deviation, respectively. Because the distribution is continuous, we must define an interval in which the value of the observation x will fall. The probability density function is properly defined such that the probability $dP_G(x;\mu,\sigma)$ that the value of a random observation will fall within an interval dx around x is given by: $dP_G(x;\mu,\sigma) = p_G(x;\mu,\sigma)dx$

We can characterize a distribution by its *full-width at half maximum* Γ, often referred to as the *half-width*, defined as the range of x between the values at which the probability $p_G(x;\mu,\sigma)$ is half its maximum value. With this definition, we can determine that $\Gamma = 2.354\sigma$

Gaussian distribution in python

```
import numpy as np

mu, sigma = 0, 0.1 # mean and standard deviation
s = np.random.normal(mu, sigma, 1000)

# Display the histogram of the samples, along with the probability density function:

import matplotlib.pylab as plt
count, bins, ignored = plt.hist(s, 30, normed=True)  # Normalized histogram with counts and
    bins saved.

plt.plot(bins, 1/(sigma * np.sqrt(2 * np.pi)) *  np.exp( - (bins - mu)**2 / (2 * sigma**2) ),
    linewidth=2, color='r')
```

2.4 Lorentz distribution

There are many other distributions that appear in scientific research. Some are phenomenological distributions, created to parameterize certain data distributions. Others are well grounded in theory. One such distribution in the latter category is the Lorentzian distribution, similar but unrelated to the binomial distribution. The Lorentzian distribution is an appropriate distribution for describing data corresponding to resonant behavior, such as the variation with energy of the cross section of a nuclear or particle reaction or absorption of radiation in the the Mössbauer effect. The *Lorentzian probability density* function $P_L(x;\mu,\Gamma)$, also called the *Cauchy distribution*, is defined as:

$$p_L(x;\mu,\Gamma) = \frac{1}{\pi}\frac{\Gamma/2}{(x-\mu)^2+(\Gamma/2)^2} \tag{22}$$

This distribution is symmetric about its mean μ with a width characterize by its half-width Γ. The most striking difference between it and the Gaussian distribution is that it does not diminish to to 0 as rapidly; the behavior for large deviations is proportional to the inverse square of the deviation, rather than exponentially related to the square of the deviation.

As with the Gaussian distribution, the Lorentzian distribution function is a continuous function, and the probability of observing a value x must be related to the interval within which the observation may fall. The probability $dP_L(x;\mu,\Gamma)$ for an observation to fall within an infinitesimal differential interval dx around x is given by the product of the probability density function $p_L(x;\mu,\Gamma)$ and the size of the interval dx: $dP_L(x;\mu,\Gamma) = p_L(x;\mu,\Gamma)dx$

The mean μ of the Lorentzian distribution is given as one of the parameters in 21. It is obvious from the symmetry of the distribution that μ must be equal to the mean as well as to the median and to the most probable value. The standard deviation is not defined for the Lorentzian distribution as a consequence of its slowly decreasing behavior for large deviations. If we attempt to evaluate the expectation value for the square of the deviations:

$$\sigma^2 = \langle(x-\mu)^2\rangle = \frac{1}{\pi}\frac{\Gamma^2}{4}\int_{-\infty}^{+\infty}\frac{z^2}{1+z^2}dz \tag{23}$$

we find that the integral is unbounded: the integral does not converge for large deviations. Although it is possible to calculate a *sample standard deviation* by evaluating the average value of the square of the deviations from the sample mean, this calculation has no meaning and will not converge to a fixed value as the number of samples increases.

The width of the Lorentzian distribution is instead characterized by the *full-width at half maximum* Γ, generally called the *half-width*. This parameter is defined such that when $x = \mu \pm \Gamma/2$, the probability density function is equal to one-half its maximum value. Thus, the half-width Γ is the full width of the curve measured between the levels of half maximum probability.

Lorentz distribution in python

```
import numpy as np
import matplotlib.pyplot as plt
plt.ion()

l = np.random.standard_cauchy(10000)  # Lorentz distribution for mu=0 and gamma=1 (gamma=)

# Display the histogram of the samples, along with the probability density function:
l = l[(l>-25) & (l<25)]  # truncate distribution so it plots well

count, bins, ignored = plt.hist(l, 50, normed=True)

plt.plot(bins, 1/np.pi*1/(bins**2+1), linewidth=2, color='r')
```

2.5 Exercises

1. A company drills 9 wild-cat oil exploration wells, each with an estimated probability of success of 0.1. All nine wells fail. What is the probability of that happening?

2. Evaluate and plot the two following Poisson distribution: a) with $\mu = 1.69$ and b) with $\mu = 11.48$. Plot on each graph the corresponding Gaussian distribution with the same mean and standard deviation.

3. Verify that, for the Poisson distribution, if μ is an integer, the probability for $x = \mu$ is equal to the probability for $x = \mu - 1$.

4. Show by numerical calculation that, for the Gaussian probability distribution, the full-width at half maximum Γ is related to the standard deviation by $\Gamma = 2.354\sigma$.

5. Download files in "Data for exercise #5" (dist1, dist2, dist3 and dist4) and determine the name of the distribution sampled in each of them. Give the main parameters for each distribution.

2.6 Solutions

Solution to Ex. 1

```
# Let's do 20,000 trials of the model, and count the number that generate zero positive results
    .
import numpy as np

sum(np.random.binomial(9,0.1,20000)==0)/20000.
answer = 0.38885, or 38%.
```

Solution to Ex. 2

```
import numpy as np
p1=np.random.poisson(1.69,10000)
g1=np.random.normal(1.69,np.sqrt(1.69),10000)
p2=np.random.poisson(11.48,10000)
g2=np.random.normal(11.48,np.sqrt(11.48),10000)

import matplotlib.pyplot as plt
plt.ion()

cp1, bp1, ignored = plt.hist(p1, 50, normed=True)
cg1, bg1, ignored = plt.hist(g1, 50, normed=True)
cp2, bp2, ignored = plt.hist(p2, 50, normed=True)
cg2, bg2, ignored = plt.hist(g2, 50, normed=True)
```

Solution to Ex. 3

```
import numpy as np
mu=3
l=100000
p=np.random.poisson(3,l)  # Poission distribution with mu=100
sum(p==mu)/l                # Probability for x=100
sum(p==mu-1)/l            # Probability for x=100-1
```

Solution to Ex. 4

```
import numpy as np
import matplotlib.pylab as plt
mu,sigma=0,0.2
s=np.random.normal(mu,sigma,100000)
count, bins, ignored = plt.hist(s, 101, normed=True)  # Normalized histogram with counts and
    bins saved.

pdf=1/(sigma * np.sqrt(2 * np.pi)) *  np.exp( - (bins - mu)**2 / (2 * sigma**2))
m=max(pdf)
l=2*abs(bins[(np.abs(pdf - m/2) <=0.001)])   # full-width at half-maximum
print(l-2.354*sigma<=0.0001)
True
```

Solution to Ex. 5

```
import numpy as np
import math as mt
import matplotlib.pylab as plt
plt.ion()

d=np.loadtxt('dist1.dat')
mu=np.mean(d)
sigma=np.std(d)

count, bins, ignored = plt.hist(d, 30, normed=True)


# Normal?
plt.plot(bins, 1/(sigma * np.sqrt(2 * np.pi)) *  np.exp( - (bins - mu)**2 / (2 * sigma**2) ),
    linewidth=2, color='r')

# Poission?
bin=np.arange(max(d))
n=[]
for i in bin:
    n.append(mt.factorial(i))        # This is necessary to calculate the factorial of an array
    .
plt.plot(bin, mu**bin/n * np.exp(-mu), linewidth=2, color='c')

# Lorentz
plt.plot(bins, 1/np.pi*1/(bins**2+1), linewidth=2, color='g')

# Binomial?
# It is not a discrete distribution.

dist1.dat: Best fit for a Normal distribution with mu=50.0 and sigma=0.10
dist2.dat: Poisson distribution with mu=50 (sigma=sqrt(mu))
dist3.dat: Lorentz distribution with mu=0 and Gamma=2
dist4.dat: Normal distribution with mu=0 and sigma=0.1
```