Notes on statistics

D. Melconian

1 Background

1.1 Probability distribution functions

Continuous distributions Consider a repeatable experiment where the outcome is described by one continuous variable, x. The sample space spans all possible values that x can take. If one asks what the probability of observing a value within the interval $[x, x+dx]$, then the answer is given by the "probability distribution function" (p.d.f.), $f(x)$. *I.e.*:

$$\text{Probability that } x \text{ observed in interval } [x, x+dx] = f(x)dx. \tag{1}$$

Another way to interpret this (the frequentist approach) is to say that $f(x)dx$ is the fraction of times that an observation is found between x and $x+dx$, in the limit that the number of observations is very large.

The p.d.f. is normalized such that the total probability of *any* outcome is unity, so $\int f(x)dx = 1$.

Discrete distributions In some cases, the variable x can only take on discrete values x_i, where $i = 1, 2, \ldots, N$ (N can be infinite). In this case, the p.d.f. is defined as:

$$\text{Probability to observe the value } x_i = f(x_i). \tag{2}$$

with the normalization condition $\sum_{i=1}^{N} f(x_i) = 1$.

1.2 Cumulative distributions

The probability for a random variable to take on a value less than or equal to x is given by the "cumulative distribution", $F(x)$. It is related to the p.d.f. by:

$$F(x) = \begin{cases} \displaystyle\int_{-\infty}^{x} f(x')dx' & \text{continuous} \\ \displaystyle\sum_{x_i<x} f(x_i') & \text{discrete} \end{cases} \tag{3}$$

A very useful concept related to this is the "quantile of order α". The quantile x_α is defined as the value of the random variable x such that $F(x_\alpha) = \alpha$, with $0 \leq \alpha \leq 1$; that is, the quantile is simply the inverse function of the cumulative distribution:

$$x_\alpha = F^{-1}(\alpha) \tag{4}$$

and corresponds to the value of x such that the total probability for seeing a value up to and including x_α is α.

1.3 Expectation values

The expectation value (or "popultation mean", or simply "mean") of a random variable, x, distributed according to a p.d.f. $f(x)$, is

$$E[x] = \int_{-\infty}^{\infty} x\, f(x) dx \equiv \mu, \tag{5}$$

Note it is simply the integral of x weighted by the p.d.f. $f(x)$. If one has a function of the random variable x denoted by $a(x)$, then its expectation value is

$$E[a(x)] = \int_{-\infty}^{\infty} a(x)\, f(x) dx, \tag{6}$$

1.3.1 Central moments

The n^{th} central moment of x is defined as

$$E[(x - E[x])^n] = \int_{-\infty}^{\infty} (x-\mu)^n\, f(x) dx \equiv \mu_n, \tag{7}$$

and in particular, the second central moment,

$$E[(x - E[x])^2] = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x) dx \equiv \sigma^2 = V[x] \tag{8}$$

is called the "population variance" (or just variance) of x. Note that $E[(x - E[x])^2] = E[x^2] - \mu^2$, so the variance is a measure of how widely x is spread about its mean value. The square root of the variance, σ, is called the standard deviation of x.

1.3.2 Multi-variant expectation values

For the case of a function of more than one random variable, $a(\vec{x}) = a(x_1, x_2, \ldots, x_n)$, with a multivariant p.d.f. $f(\vec{x}) = f(x_1, x_2, \ldots, x_n)$, the expectation value and variance are

$$E[a(\vec{x})] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} a(\vec{x})\, f(\vec{x}) dx_1 dx_2 \ldots dx_n = \mu_a \tag{9}$$

$$V[a(\vec{x})] = E[(a-\mu_a)^2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} (a(\vec{x})-\mu_a)^2\, f(\vec{x}) dx_1 dx_2 \ldots dx_n = \sigma_a^2 \tag{10}$$

The "covariance" of the two random variables, say x and y, is

$$V_{xy} = E[xy] - \mu_x \mu_y = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f(x,y) dx\, dy \;-\; \mu_x \mu_y, \tag{11}$$

where $\mu_x = E[x]$ and $\mu_y = E[y]$. The "covariance matrix" (or "error matrix"), V_{ij} where i and j equal x and y, is in this case a symmetric 2×2 matrix which has the variances V_{ii} of x and y on it's diagonals, and the covariance between them on the off-diagonal. Often instead of using $\text{cov}[x,y] = V_{xy}$, one defines a dimensionless correlation coefficient:

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}, \tag{12}$$

where $-1 \leq \rho_{xy} \leq 1$ is a measure of how strongly correlated (or anti-correlated if negative) two parameters are. Figure 1 shows the situation for a few cases ranging from completely uncorrelated ($\rho_{xy} = 0$) to very highly correlated ($\rho_{xy} = 0.99$).

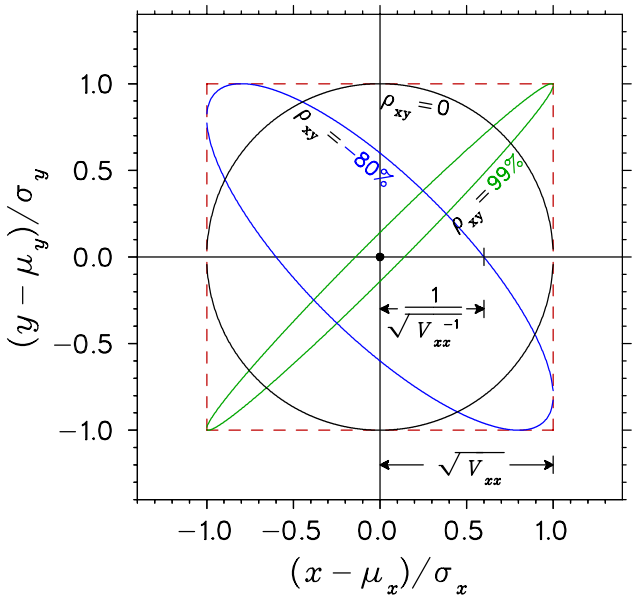

Figure 1: Plots of the error ellipses for two random variables x and y with different amounts of correlations between them.

2 Error propagation

Suppose one has a set of n random variables $\vec{x}$ distributed according to some joint p.d.f. $f(\vec{x})$. Suppose that the p.d.f. is not completely known, but the mean values, $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$, and covariance matrix, V_{ij}, are known (or at least have been estimated).

Now consider a function of the n random variables $a(\vec{x})$. Without knowing the p.d.f.'s of the x_i, we cannot determine the p.d.f. of a; however, one *can* approximate the expectation value of a and its variance by first expanding the function $a(\vec{x})$ to first order about the mean values of the x_i, which we do know:

$$a(\vec{x}) \approx a(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial a}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i). \tag{13}$$

The expectation value is then (as one would expect)

$$E[a(\vec{x})] \approx a(\vec{\mu}) \tag{14}$$

because $E[x_i - \mu_i] = 0$. Similarly, the expectation value of a^2 is

$$E[a^2(\vec{x})] \approx a^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial a}{\partial x_i}\frac{\partial a}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij} \tag{15}$$

so that the variance $V[a] = E[a^2] - \mu_a^2$ is

$$V[a(\vec{x})] \approx \sum_{i,j=1}^{n} \left[\frac{\partial a}{\partial x_i}\frac{\partial a}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij} \tag{16}$$

and similarly the covariance of two functions $a(\vec{x})$ and $b(\vec{x})$ is

$$V_{ab} \approx \sum_{i,j=1}^{n} \left[\frac{\partial a}{\partial x_i} \frac{\partial b}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}. \tag{17}$$

Eqs. (16) and (17) form the basis of *error propagation* (i.e. the variances, which are used as measures of statistical uncertainties, are propagated from the x_i to the functions a, b, etc.). For the case where the x_i are not correlated, *i.e.* $V_{ii} = \sigma_i^2$ and $V_{ij} = 0$ for $i \neq j$, the above reduce to

$$V[a(\vec{x})] = \sigma_a^2 \approx \sum_{i=1}^{n} \left[\frac{\partial a}{\partial x_i} \right]^2_{\vec{x}=\vec{\mu}} \sigma_i^2 \tag{18}$$

and

$$V_{ab} \approx \sum_{i=1}^{n} \left[\frac{\partial a}{\partial x_i} \frac{\partial b}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} \sigma_i^2. \tag{19}$$

The simplest case is if $a = x + y$ and the two are uncorrelated. Then $\frac{\partial a}{\partial x} = 1$ and $\frac{\partial a}{\partial y} = 1$ and Eq. (18) leads to $\sigma_a^2 = (1)^2\sigma_x^2 + (1)^2\sigma_y^2$; their uncertainties add in quadrature, as you may have already learned, or you can say their final variance is just the sum of the variances of the random variables x and y.

2.1 An example

Say you've measured two things which are 50% correlated (so $V_{xy} = 0.5$) and you've found $\mu_x = 8.0$ and $\mu_y = 10.1$ with uncertainties (standard deviations) $\sigma_x = 0.7$ and $\sigma_y = 0.8$. You're interested in the value $a = 2x - y$. Clearly the expectation value of a is 5.9; but how well do you *know* that it is 5.9? Based on Eq. (16), the variance of a is

$$\begin{aligned} \sigma_a^2 &= (2\sigma_x)^2 + \sigma_y^2 + 2(-2V_{xy}) \\ \Rightarrow \quad \sigma_a &= \sqrt{4\sigma_x^2 + \sigma_y^2 - 4\rho_{xy}\sigma_x\sigma_y} \\ &= 1.2 \end{aligned}$$

So based on the two measurements of $x = 8.0 \pm 0.7$ and $y = 10.1 \pm 0.8$ and given the accuracy of those measurements (the $\pm$ values are the standard deviations, σ_x and σ_y) as well as their degree of correlation, you know $a = 5.9 \pm 1.2$. If they were uncorrelated (as is often the [valid] assumption), you'll find $a = 5.9 \pm 1.6$. If 100% anti-correlated (so $\rho_{xy} = -1$), $a = 5.9 \pm 2.2$.

2.2 Another example

You've measured the half-life of a particle to be $t_{1/2} = (298 \pm 1)$ ms and an initial number of particles to be $(1.234 \pm 0.005) \times 10^6$. How many are there after 1 second?

Let $\Delta N_\circ (= \sigma_{N_\circ}) = 0.005 \times 10^6$ and let's convert the lifetime information to the decay constant via $\lambda = \ln 2/t_{1/2} = 2.326\ \mathrm{s}^{-1}$. Since $d\lambda/dt_{1/2} = -\ln 2/t_{1/2}^2$, $\Delta\lambda = \Delta t_{1/2}/t_{1/2}^2 = 0.0078\ \mathrm{s}^{-1}$. Note (or at least assume) our measurement of the number nuclei is independent of the lifetime, so these are uncorrelated random variables ($V_{N_\circ,\lambda} = \rho_{N_\circ\lambda} = 0$). From $N = N_\circ e^{-\lambda t}$, we know the mean

value of the number of atoms is $N = 1.205 \times 10^5$. To estimate the uncertainty from our imperfect knowledge of λ and $N_\circ$, we use Eq. (16) and find

$$\frac{\partial N}{\partial N_\circ} = e^{-\lambda t} \qquad \text{and} \qquad \frac{\partial N}{\partial \lambda} = -\lambda N_\circ e^{-\lambda t}$$

so

$$\begin{aligned}(\Delta N)^2 &= \left(\frac{\partial N}{\partial N_\circ}\Delta N_\circ\right)^2 + \left(\frac{\partial N}{\partial \lambda}\Delta\lambda\right)^2 + 2\left(\frac{\partial N}{\partial N_\circ}\frac{\partial N}{\partial \lambda}V_{N_\circ,\lambda}\right) \\ \Rightarrow \Delta N &= e^{-\lambda t}\sqrt{\Delta N_\circ^2 + (\lambda N_\circ \Delta\lambda)^2} \\ &= 2242.\end{aligned}$$

So the number of nuclei after 1 s is $(1.205 \pm 0.022) \times 10^5$.

Can you show that the activity after ten half-lives is $\mathcal{A} = 2800 \pm 43$ Bq?

3 Specific distribution functions

3.1 Binomial distribution

You've all heard of this one... for N independent observations for which there are two possible outcomes (e.g. "success" or "failure") where the probability for one ("success") is some constant value p, and the other ("failure") is $q = 1 - p$. One can define "sucess" if a measured quantity lands in a particular bin of a histogram (failure if not) with N total entries in the histogram. The set of trials can be regarded as a single measurement and is characterized by a discrete random variable, k, defined to be the total number of successes. Note that here the entire set of observations is treated as a single random measurement, not each individual trial. That is, the sample space is defined to be the set of possible values of k successes given N observations. If one were to repeat the entire experiment many times with N trials each time, the resulting values of k would occur with relative frequencies given by the so-called binomial distribution.

The binomial distribution gives the total probability to have k successes in N events according to

$$f(k; N, p) = \frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}, \tag{20}$$

for $k = 0, 1, \ldots, N$. One can show the expectation value of k is $E[k] = Np$ and the variance is $V[k] = Np(1-p)$. Let's not bother with the multinomial distribution, which is a generalization to where there are more than just "success" and "failure" results; there are m different possible outcomes.

3.2 Poisson distribution

The factorials in the binomial distribution are cumbersome and quickly become incalculable for large N; for example, $150! = 5.7 \times 10^{262}$ (). Consider the limit that N is very large and p is very small, but the expectation value of the number of successes (*i.e.* their product Np) is some finite value λ. It can be shown that in this limit, the binomial distribution approaches

$$f(k; \lambda) = \frac{\lambda^k}{k!}e^{-\lambda}, \tag{21}$$

which is known as the Poisson distribution. Here $k = 0, 1, \ldots, \infty$ and the p.d.f. has only one parameter, λ. The expectation value for the Poisson distribution is $E[k] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda$, and its variance is $V[k] = \sum_{k=0}^{\infty} (k-\lambda)^2 \frac{\lambda^k}{k!} e^{-\lambda}$ also equals λ.

3.3 The Gaussian distribution

The normal, or Gaussian, distribution is the p.d.f. of a continuous random variable, x, defined by

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{22}$$

where x can take on any value between $\pm\infty$. As expected, the two parameters represent the mean and variance: $E[x] = \mu$ and $V[x] = \sigma^2$.

The importance of the Gaussian distribution comes from the *Central Limit Theorem* which states that the sum of n independent continuous random variables x_i with means μ_i and variances σ_i^2 becomes a Gaussian random variable with mean $\mu = \sum_{i=1}^{n} \mu_i$ and variance $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$ in the limit that n approaches infinity. This holds regardless of the individual p.d.f.'s of the x_i, and this is the justification for treating measurement uncertainties as Gaussian random variables; this holds to the extent that the total uncertainty is the sum of a large number of small contributions (although "large" is a somewhat subjective term).

Figure 2 shows a comparison of the binomial, Poission and Gaussian distributions for $N = 150$ and a few values of p such that the means are $2, 10, 25$ and 75. Note the limiting cases of applicability: when Np is small (2), the Gaussian is quite different from the others (is not a good approximation and shouldn't be used!). Because in this case p is small and N is large, the Poisson distribution reproduces well the binomial. By $Np = 10$, the three distributions are pretty close to each other. For $Np = 25$, the Poisson and Gaussian are aligning even better (as a consequence of the Central Limit Theorem), however the binomial is different because now $p = 1/6$ isn't very small as it should be for the Poisson to be a good approximation to the binomial. Finally, with $p = 1/2$, the Poisson and Gaussian are almost the same, and clearly with p so large the binomial is dramatically different.

The N-dimensional generalization of the Gaussian distribution is the multivariant Gaussian:

$$f(\vec{x}; \vec{\mu}, V) = \frac{1}{(2\pi)^{N/2}|V|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1} (\vec{x}-\vec{\mu})\right), \tag{23}$$

where $\vec{x}$ and $\vec{\mu}$ are column vectors, $\vec{x}^T$ and $\vec{\mu}^T$ are the corresponding row vectors, and V is a symmetric $N \times N$ matrix. The expectation values and (co)variances are found to be $E[x_i] = \mu_i, V[x_i] = V_{ii}$, and $\mathrm{cov}[x_i, x_j] = V_{ij}$.

In the 2D case, the p.d.f. becomes, with $\rho = \mathrm{cov}[x,y]/\sigma_x\sigma_y$ the correlation coefficient,

$$\begin{aligned} f(x, y; \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right\}. \end{aligned} \tag{24}$$

It is this expression which defines the error ellipses of Fig. 1. The contours plotted are the ones that correspond to containing 68.27% of the area under this 2D multivariant Gaussian surface.

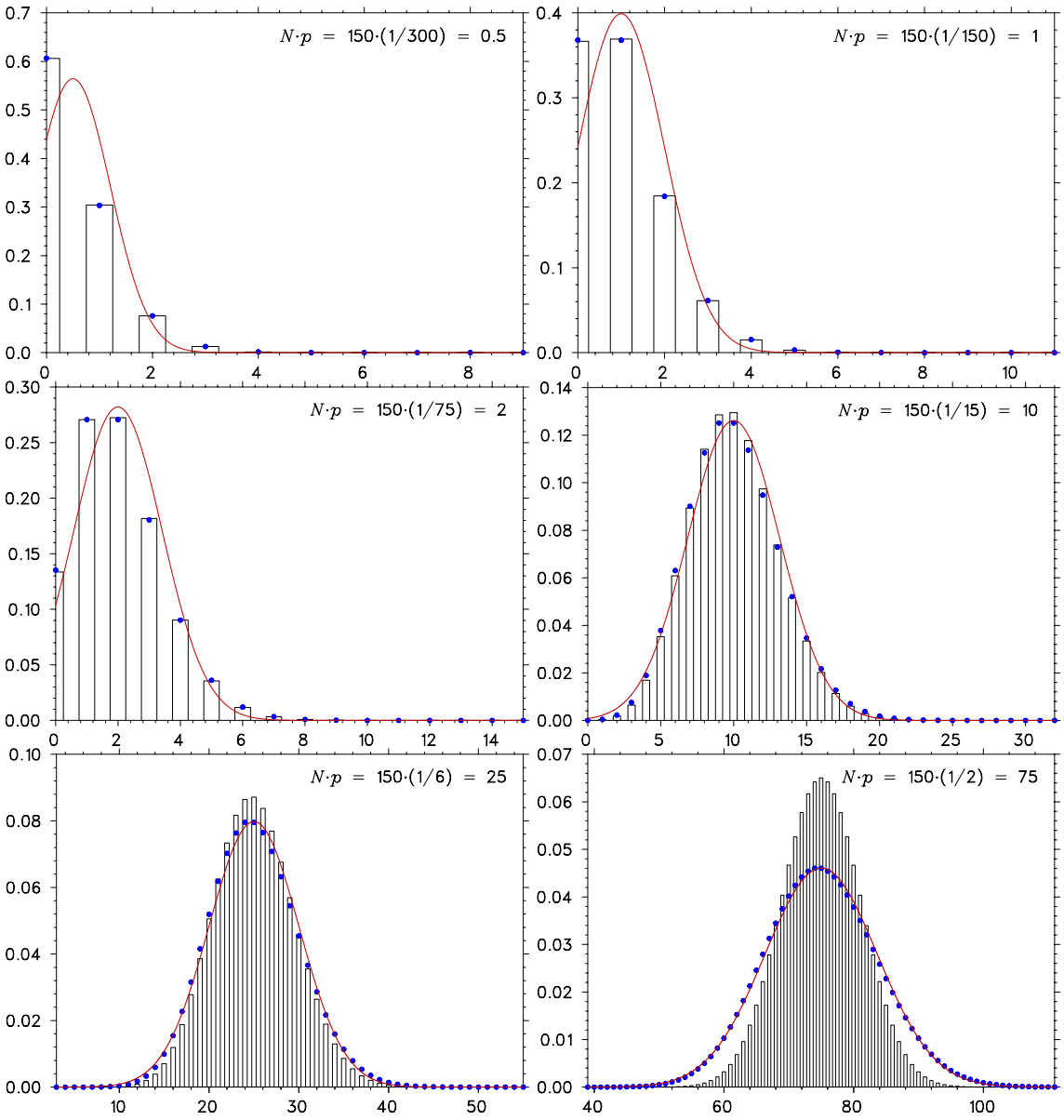

Figure 2: Comparison of binomial (histogram), Poisson (filled circles) and Gaussian (solid line) distributions with different means.

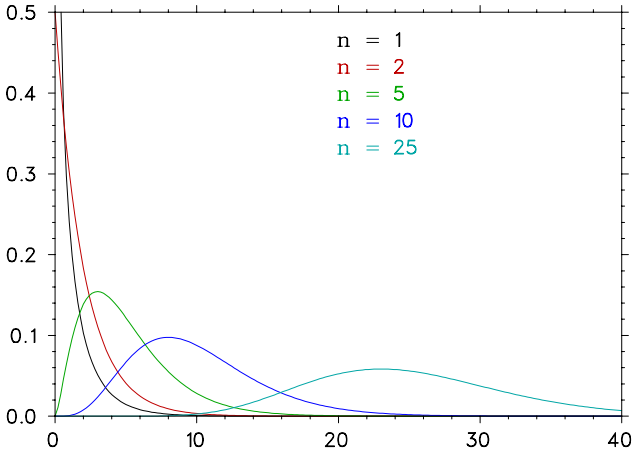

Figure 3: The χ^2 probability density for various values of the parameter n (the degrees of freedom).

3.4 The χ^2 distribution

The χ^2 (chi-square) distribution of the continuous variable z ($0 \leq z < \infty$) is defined by

$$f(z;n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} \tag{25}$$

where $n = 1, 2, \ldots$ is called the number of degrees of freedom. The gamma function $\Gamma(x)$ is in many math libraries. If $x = n$ is an integer, $\Gamma(n) = n!$; in general, $\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. The mean and variance of the χ^2 distribution can be found to be $E[x] = n$ and $V[z] = 2n$.

This distribution derives its importance from the following: given N independent Gaussian random variables x_i with known mean μ_i and variance σ_i^2, it can be shown that the random variable

$$z = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{26}$$

is distributed according to the χ^2 distribution for N degrees of freedom. More generally, if the x_i are not independent but are described by an N-dimensional Gaussian p.d.f., the variable

$$z = (\vec{x} - \vec{\mu})^T \, V^{-1} \, (\vec{x} - \vec{\mu}) \tag{27}$$

is a χ^2 random variable for N degrees of freedom. This is an important part of hypothesis-testing and determining the quality of fits. Figure 3 shows this distribution for a few different degrees of freedom.

4 Parameter Estimation

Suppose one has a sample of size n of a random variable x: $x_1, x_2, \ldots, x_n$. It is assumed that x is distributed according to some p.d.f. $f(x)$ which is not known. We would like to construct a function of the x_i to be an estimator for the expectation value of x, $E[x] = \mu$. One possibility is the arithmetic mean of the x_i, defined by

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{28}$$

The arithmetic mean of the elements of a sample is called the "sample mean"; it should not be confused with the expectation value ("population mean") of x. The latter is denoted by μ or $E[x]$, for which $\overline{x}$ is an *estimator*. The expectation value of our estimator $\overline{x}$ is

$$E[\overline{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \mu \tag{29}$$

since one can show that $E[x_i] = \mu$ for all i. Thus we can say that the sample mean $\overline{x}$ is an *unbiased* estimator for the population mean μ.

The "sample variance", s^2, of this sample of size n is defined by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{30}$$

By computing the expectation value of s^2, one can show that the sample variance is also an unbiased estimator of the population variance σ^2. If the mean is known, one would of course use that information and instead define

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{31}$$

for an unbiased estimator of the population variance.

One can estimate the covariance of two random variables, x and y, of unknown means via

$$\hat{V}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{32}$$

which can also be shown to be an unbiased estimator of the true covariance V_{xy}.

The variance of $\overline{x}$ is

$$\begin{aligned} V[\overline{x}] = E[\overline{x}^2] - (E[\overline{x}])^2 &= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_j\right)\right] \; - \; \mu^2 \\ &= \frac{1}{n^2}\sum_{i,j=1}^{n} E[x_i x_j] - \mu^2 \\ &= \frac{1}{n^2}[(n^2-\mu)\mu^2 + n(\mu^2+\sigma^2)] - \mu^2 \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{33}$$

where we have used the fact that $E[x_i x_j] = \mu^2$ for $i \neq j$ and, for $i = j$, $E[x_i^2] = \mu^2 + \sigma^2$. This expresses the fact that the standard deviation of the mean of n measurements of x is equal to the standard deviation of $f(x)$ itself (σ) divided by $\sqrt{n}$. ***The more counts you have, the better you measure something***, and the improvement goes like $1/\sqrt{n}$.

The variance of s^2 can be shown to be

$$V[s^2] = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4), \tag{34}$$

where μ_4 is the fourth central moment of x. For Gaussianly distributed x_i, this becomes

$$V[s^2]_{\text{Gauss}} = \frac{2\sigma^4}{n-1} \tag{35}$$

for any $n > 1$. For large n, the standard deviation of s^2 (the "uncertainty on the uncertainty") is $\sigma/\sqrt{2n}$.

Finally, if the x_i have different, known variances σ_i^2, then the weighted average

$$\overline{x} = \frac{1}{2}\sum_{i=1}^{n} w_i x_i \tag{36}$$

is an unbiased estimator for μ with a smaller variance than an unweighted average; here the weighting factors are $w_i = 1/\sigma_i^2$ and $w = \sum_{i=1}^{n} w_i$. In this case, the variance of $\overline{x}$ is $1/w$ so the standard deviation is $1/\sqrt{w}$.

4.1 Method of maximum likelihood

4.2 Method of least squares

4.3 Hypothesis testing

4.4 Confidence levels

4.5 Bayesian intervals