Statistical properties of data

• In this chapter we deal with the properties of data sets.
  - Based on L. E. Reichl: *A Modern Course in Statistical Physics*.
  - Another good source is[1]:
    P.R.Bevington, D.K.Robinson,
    Data reduction and error analysis for the physical sciences, 2nd edition,
    McGraw-Hill, 1992.

• More thoroughly this subject is handled in the lecture course
  Statistical Methods I - Tilastolliset Menetelmät I
  http://www.kumpula.helsinki.fi/~www_sefo/statistics/

• A data set is a set of number obtained from experiments or simulations.

• Ways to characterize data sets and the most important probability distributions will be presented.

• Using the probability theory we can more or less quantitatively describe the outcome of an 'experiment'.

• If the probability of the event A is $P(A)$ then the expected number of events A in N identical experiments is $NP(A)$.
  - In the limit $N \to \infty$ the number of events $A \to NP(A)$.
  - The result of the 'experiment' determines a quantity called **stochastic variable**.
  - Stochastic variable X of a **sample space** S is a function that maps elements of S to a set of real numbers.

1. The first edition of the book has Bevington as the only author.

Statistical properties of data

- In every experiment variable X can obtain one of the values $\{x_i\}$. E.g.
  1. Number of tails after three tosses of a coin.
  2. Maximum number obtained after tossing the dice four times.

- Let our stochastic variable X be defined in space S and let the allowed values be $X(S) = \{x_1, x_2, \ldots\}$.

  - We can make $X(S)$ a sample space by assigning every x_i a probability.
  - These probabilities $f(x_i)$ define the **probability distribution** of S and they fulfill the following conditions

$$f(x_i) \geq 0$$

$$\sum_i f(x_i) = 1 .$$

- In many cases we know only the **moments** of the distribution f:

$$\langle X^n \rangle = \sum_i x_i^n f(x_i) .$$

- The first moment $n = 1$ is the mean $\langle X \rangle$ and the standard deviation of the distribution is expressed in terms of the first two moments as

$$\sigma_X \equiv \sqrt{\langle X^2 \rangle - \langle X \rangle^2} .$$

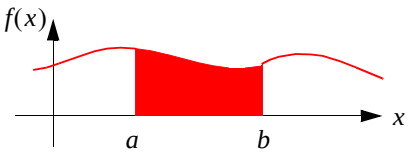
Statistical properties of data

- A stochastic variable can also be a continuous variable.

  - E.g. interval $[a, b]$ may correspond to one event.

  - In the continuous case the probability distribution is defined such that

$$P(a \le X \le b) = \int_a^b f_X(x)dx$$



  is the probability for event $a \le X \le b$.

  - The distribution also fulfills the following conditions

$$f_X(x) \ge 0$$

$$\int f_X(x)dx = 1.$$

  - Moments are defined as

$$\langle X^n \rangle = \int x^n f_X(x)dx$$

  - If we know all the moments of f_X then we know the distribution completely.

Statistical properties of data

  - This can be shown with the help of so called **characteristic function** $\phi_X(k)$:

$$\phi_X(k) = \langle e^{ikX} \rangle = \int e^{ikx} f_X(x)dx = \sum_{n=0}^{\infty} \frac{(ik)^n \langle X^n \rangle}{n!}.$$

  - Probability distribution is the Fourier transform of the characteristic function:

$$f_X(x) = \frac{1}{2\pi} \int e^{-ikx} \phi_X(k)dk.$$

  - Moments are obtained as the derivatives of the characteristic function:

$$\langle X^n \rangle = \frac{1}{i^n} \left[\frac{\mathrm{d}^n}{\mathrm{d}k^n} \phi_X(k) \right]_{k=0}.$$

  - Sometimes so called **cumulant expansion** is used

$$\phi_X(k) = \exp\left[\sum_{n=1}^{\infty} \frac{(ik)^n}{n!} C_N(X) \right]$$

$$C_1(X) = \langle X \rangle, \quad C_2(X) = \langle X^2 \rangle - \langle X \rangle^2, \quad C_3(X) = \langle X^3 \rangle - 3\langle X \rangle \langle X^2 \rangle + 2\langle X \rangle^3, \text{ etc.}$$

  - Cumulant expansion is often used because many distributions have only a few dominant cumulants.

Statistical properties of data

- We may have many stochastic variables.

- Assume we have variables $X(S) = \{x_1, x_2, \ldots\}$ and $Y(S) = \{y_1, y_2, \ldots\}$.

- Cartesian product of these $X(S) \times Y(S) = \{(x_1, y_1), (x_1, y_2), \ldots, (x_i, y_j), \ldots\}$ defines a sample space when we define the probability of pair $\{x_i, y_j\}$ to be

$$P(X = x_i, Y = y_j) = f(x_i, y_j) \ .$$

- In the case of continuous variables the probability distribution is $f(x, y)$; it fulfills the following conditions:

$$f(x, y) \geq 0$$
$$\int f(x, y)\, dx dy = 1 .$$

- **Covariance** of the variables is defined as

$$\begin{aligned} \mathrm{cov}(X, Y) &= \int (x - \langle X \rangle)(y - \langle Y \rangle) f(x, y)\, dx dy \quad . \\ &= \int xy f(x, y)\, dx dy - \langle X \rangle \langle Y \rangle \\ &= \langle XY \rangle - \langle X \rangle \langle Y \rangle \end{aligned}$$

- And **correlation**

$$\mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Statistical properties of data

- Correlation has the following properties:

1. $\mathrm{cor}(X, Y) = \mathrm{cor}(Y, X)$
2. $-1 \leq \mathrm{cor}(X, Y) \leq 1$
3. $\mathrm{cor}(X, X) = 1, \mathrm{cor}(X, -X) = -1$
4. $\mathrm{cor}(aX + b, cY + d) = \mathrm{cor}(X, Y)$, $a, c \neq 0$.

- If variables X and Y are independent then

1. $f(x, y) = f_X(x) f_Y(y)$
2. $\langle XY \rangle = \langle X \rangle \langle Y \rangle$
3. $\langle (X + Y)^2 \rangle - \langle X + Y \rangle^2 = \langle X^2 \rangle - \langle X \rangle^2 + \langle Y^2 \rangle - \langle Y \rangle^2$
4. $\mathrm{cov}(X, Y) = 0$.

- Other chracteristics of distributions are

median: $\int_a^{x_{\mathrm{med}}} f(x)\, dx = \int_{x_{\mathrm{med}}}^{b} f(x)\, dx = \frac{1}{2}$

mode: $x_{\mathrm{mode}} = \max_{x \in [a, b]} \{f(x)\}$

Statistical properties of data: important distributions

• Often we have a situation with a large number of measurements (N) each of which has two possible outcomes (+1 and -1).

- One example might be whether a particle scatters or not while traversing a certain path length.

- Let the probabilities of these outcomes be p and q; clearly

$$p + q = 1.$$

- Probability that after N measurements we get n_1 times +1 and n_2 times -1 is

$$P_N(n_1) = \frac{N!}{n_1! n_2!} p^{n_1} q^{n_2}.$$

- This is the **binomial distribution**.

- The mean and the standard deviation of the distribution are

$$\langle n_1 \rangle = pN$$

$$\sigma_N^2 = Npq.$$

- When $N \to \infty$ and $pN \to \infty$ we obtain the **Gaussian distribution**

$$P_N(n_1) = \frac{1}{\sigma_N \sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(n_1 - \langle n_1 \rangle)^2}{\sigma_N^2} \right], \qquad \sigma_N = \sqrt{Npq}.$$

- Gaussian distribution is determined by two moments $\langle n_1 \rangle$ and σ_N.

Statistical properties of data: important distributions

- Random numbers with Gaussian distribution can not be generated by the inversion algorithm.
- So called Box-Muller algorithm utilizes the method to generate RNGs obeying a desired multidimensional distribution.
  - If variables $x_1, x_2, \ldots, x_N$ have a joint probability distribution

$f_x(x_1, x_2, \ldots, x_N) dx_1 dx_2 \ldots dx_N$ then variables $y_1, y_2, \ldots, y_N$ have a distribution

$$f_y(y_1, y_2, \ldots, y_N) dy_1 dy_2 \ldots dy_N = f_x(x_1, x_2, \ldots, x_N) \left| \frac{\partial(x_1, x_2, \ldots, x_N)}{\partial(y_1, y_2, \ldots, y_N)} \right| dy_1 dy_2 \ldots dy_N,$$

where the determinant of the Jacobian is $\left| \frac{\partial(x_1, x_2, \ldots, x_N)}{\partial(y_1, y_2, \ldots, y_N)} \right| = \det \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_N}{\partial y_N} \\ \cdots & \cdots & \cdots \\ \frac{\partial x_N}{\partial y_1} & \cdots & \frac{\partial x_N}{\partial y_N} \end{bmatrix}$

Univariate case:

$$f_y(y) dy = \left| \frac{dx}{dy} \right| f_x(x) dx$$

  - By transforming two uniform ($[0, 1]$) RNs x_1 and x_2 in the following way

$$y_1 = \sqrt{-2 \ln x_1} \cos(2\pi x_2)$$

$$y_2 = \sqrt{-2 \ln x_1} \sin(2\pi x_2)$$

two normally distributed RNs are generated.

- The algorithm can be made a bit faster by minimizing the number of calls of transcendental functions:
  1. Obtain two evenly distributed random numbers v_1 and v_2 between -1 and 1, then calculate $w = v_1^2 + v_2^2$
  2. If $w \geq 1$ return to step 1°
  3. Calculate $r = \sqrt{-2 \log w}$
  4. Calculate $x = r v_1 / \sqrt{w}$ and $y = r v_2 / \sqrt{w}$
  5. Return x and on next step y

Statistical properties of data: important distributions

- In the limit $N \to \infty$ and $p \to 0$ such that $Np = a \ll N$ (where a is a finite constant) the binomial distribution can be approximated by the **Poisson distribution**

$$P_N(n_1) = \frac{a^{n_1} e^{-a}}{n_1!}.$$

- Poisson distribution is determined by the first moment

$$\langle n_1 \rangle = a.$$

- As an example of the binomial distribution we take the 1D random walker.
  - Probability that the walker takes a step to left is $p = 1/2$ and to the right is $q = 1/2$.
  - Number of steps is N, number of steps to left is n_1 and to right n_2.
  - Distance from the origin is $m = n_1 - n_2$.
  - Because $N = n_1 + n_2 \to m = 2n_1 - N$.
  - For large values of N the probability distribution of the distance m is obtained by substituting $n_1 = (m+N)/2$ to Gaussian distribution:

$$P_N(m) = \left(\frac{2}{\pi N}\right)^{1/2} e^{-m^2/2N}.$$

Statistical properties of data: important distributions

- If the step length is l then the distance from the origin is $x = ml$.
- Probability that the walker is in the interval $[x, x+\Delta x]$ after N steps is $P_N(x) = P_N(m)(\Delta x/2l)$ or

$$P_N(x) = \frac{1}{(2\pi N l^2)^{1/2}} e^{-x^2/2Nl^2}.$$

- If the walker takes n steps in a unit of time we obtain the probability density at time t:

$$P(x,t) = \frac{1}{2(\pi D t)^{1/2}} e^{-x^2/4Dt}; \qquad N = nt, \quad D = nl^2/2 \text{ (diffusion constant)}$$

- The importance of Gaussian distribution is based on the **central limit theorem**.
  - We want to know the distribution of a stochastic variable Y. Values of Y correspond to measurements of N variables X:

$$y_N = \frac{x_1 + x_2 + \ldots + x_N}{N}.$$

  - What is the distribution $f_Y(y_n - \langle X \rangle)$?

Statistical properties of data: important distributions

- Its characteristic function can be written in the form

$$\begin{aligned}\Phi(k) &= \int e^{ik(y_N - \langle X\rangle)} f_Y(y_N - \langle X\rangle) dy_N \\ &= \int e^{i(k/N)((x_1 - \langle X\rangle) + \ldots + (x_N - \langle X\rangle))} f_X(x_1)\ldots f_X(x_N) dx_1 \ldots dx_N \\ &= \left[\phi\left(\frac{k}{N}\right)\right]^N\end{aligned}$$

- If $\sigma^2 = \langle X^2\rangle - \langle X\rangle^2$ then

$$\phi\left(\frac{k}{N}\right) = \int e^{i(k/N)(x_1 - \langle X\rangle)} f_X(x_1) dx_1 = 1 - \frac{1}{2}\frac{k^2}{N^2}\sigma^2 + \ldots .$$

- This function decreases quickly when k grows because of the oscillatory term $e^{i(k/N)x_1}$.

- $[\phi(k/N)]^N$ goes to zero even faster.

- Assuming that the **moments of the distribution $f_X(x_1)$ are finite** we get

$$\Phi(k) = \left[1 - \frac{1}{2}\frac{k^2}{N^2}\sigma^2 + O\left(\frac{k^3}{N^3}\right)\right]^N \to e^{-k^2\sigma^2/2N},$$

when $N \to \infty$.

Statistical properties of data: important distributions

- Finally we get the probability distribution

$$\begin{aligned}f_Y(y_N - \langle X\rangle) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ik(y_N - \langle X\rangle)}\Phi(k) dk \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ik(y_N - \langle X\rangle)} e^{-k^2\sigma^2/2N} dk \\ &= \sqrt{\frac{N}{2\pi}}\frac{1}{\sigma}\exp\left[-\frac{N(y_N - \langle X\rangle)^2}{2\sigma^2}\right]\end{aligned}$$

- This is an important result because it says that the distribution is Gaussian regardless if the distribution f_X.

- The only condition is the existence of the moments of f_X.

Statistical properties of data: important distributions

- Another important theorem is the **law of large numbers**:

  - The fraction of the number of events A (probability p) in N independent experiments approaches p when $N \to \infty$.

  - Assume we have an average

$$y_N = \frac{x_1 + x_2 + \ldots + x_N}{N}.$$

  - According to the law of large numbers the probability that y_N deviates from $\langle X \rangle$ approaches zero when $N \to \infty$:

$$\lim_{N \to \infty} P(|y_N - \langle X \rangle| \geq \varepsilon) = 0,$$

    where $\varepsilon > 0$.

  - For proof see L. E. Reichl: *A Modern Course in Statistical Physics*.

Statistical properties of data: important distributions

• Finally we could mention a couple of important distributions.

- **Exponential distribution** describes e.g. radioactive decay or the free path of a photon in matter:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0.$$

- **Lorentz distribution** is of the form

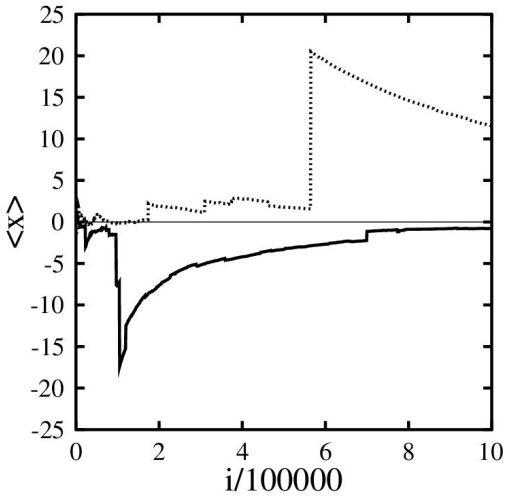
$$f(x) \propto \frac{a}{a^2 + x^2}, \quad -\infty < x < \infty.$$

  - It anomalous in the sense that its second moment diverges.
  - Moreover, calculation of its mean can not be done directly by integration but can be deduced from the symmetry.
  - Lorentz distribution is obtained e.g. from a quotient of two Gaussian stochastic variables.
  - Lorentz-distributed RNs are easily generated by the inversion method.

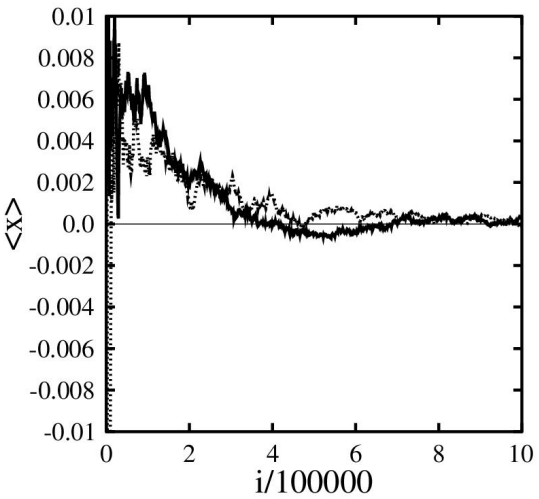
Statistical properties of data: important distributions

- The following figures demonstrate the anomalous character of the Lorentz distribution:

Lorentz



Gaussian



Statistical properties of data: comparing data sets

• Sometimes we have to compare two data sets and try to deduce whether the sets are from the same probability distribution.

• Assume we have the sets

$$\{x_{\mathrm{A}1}, x_{\mathrm{A}2}, \ldots, x_{\mathrm{A}N_\mathrm{A}}\}$$
$$\{x_{\mathrm{B}1}, x_{\mathrm{B}2}, \ldots, x_{\mathrm{B}N_\mathrm{B}}\}.$$

- Averages and variances of the sets are

$$\bar{x}_\mathrm{A} = \frac{1}{N_\mathrm{A}} \sum_{i=1}^{N_\mathrm{A}} x_{\mathrm{A}i} \qquad \mathrm{Var(A)} = \frac{1}{N_\mathrm{A}-1} \sum_{i=1}^{N_\mathrm{A}} (x_{\mathrm{A}i} - \bar{x}_\mathrm{A})^2$$
$$\bar{x}_\mathrm{B} = \frac{1}{N_\mathrm{B}} \sum_{i=1}^{N_\mathrm{B}} x_{\mathrm{B}i} \qquad \mathrm{Var(B)} = \frac{1}{N_\mathrm{B}-1} \sum_{i=1}^{N_\mathrm{B}} (x_{\mathrm{B}i} - \bar{x}_\mathrm{B})^2 .$$

- Standard deviation is the square root of the variance:

$$\sigma_\mathrm{A} = \sqrt{\mathrm{Var(A)}}$$
$$\sigma_\mathrm{B} = \sqrt{\mathrm{Var(B)}}.$$

Statistical properties of data: comparing data sets

- The first idea might be to compare the distance of the averages of the data $|\bar{x}_\mathrm{A} - \bar{x}_\mathrm{B}|$. and compare it to the standard deviations.

  - Another thing is whether this distance is statistically significant.
  - The distance can be very small compared to standard deviations but if the sets are large this distance may be significant.

  - Instead of using standard deviations it is better to use the **standard error of mean** ($\sigma_{\bar{x}_\mathrm{A}}$, $\sigma_{\bar{x}_\mathrm{B}}$).

  - It is the standard deviation of the mean of a data set.
  - It tells how are the means of different data sets distributed.
  - One can show that

$$\sigma_{\bar{x}_\mathrm{A}} = \frac{\sigma_\mathrm{A}}{\sqrt{N_\mathrm{A}}}.$$

  - The mean computed from a data set becomes more and more accurate when we increase the size of the set.

  - The above equation is easy to proof.
  - The mean of the data set can be written as

$$\bar{x} = \sum_{i=1}^{N} \frac{1}{N} x_i.$$

  - From the definition of variance we can derive the following

$$\begin{aligned}&\mathrm{Var}(a_1 x_1 + a_2 x_2 + \ldots + a_N x_N)\\ &= a_1^2 \mathrm{Var}(x_1) + a_2^2 \mathrm{Var}(x_2) + \ldots + a_N^2 \mathrm{Var}(x_N)\end{aligned}$$

Statistical properties of data: comparing data sets

  - Because the points of the data set are drawn from the same distribution with variance $\mathrm{Var}(x)$ we obtain

$$\mathrm{Var}(\bar{x}) = \sum_{i=1}^{N} \frac{1}{N^2} \mathrm{Var}(x) = \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{Var}(x) = \frac{\mathrm{Var}(x)}{N}.$$

  - Finally we get the connection between the two standard deviations

$$\sigma(\bar{x}) = \frac{\sigma(x)}{\sqrt{N}}.$$

- Using the **Student's t test** one can compare the similarity of the averages of two data sets.
- The standard error of the difference of the averages is estimated as

$$s_D = \sqrt{\frac{\sum_{i=1}^{N_\mathrm{A}} (x_{\mathrm{A}i} - \bar{x}_\mathrm{A})^2 + \sum_{i=1}^{N_\mathrm{B}} (x_{\mathrm{B}i} - \bar{x}_\mathrm{B})^2}{N_\mathrm{A} + N_\mathrm{B} - 2} \left(\frac{1}{N_\mathrm{A}} + \frac{1}{N_\mathrm{B}} \right)}.$$

- One can show that the quantity

$$t = \frac{\bar{x}_\mathrm{A} - \bar{x}_\mathrm{B}}{s_D}$$

has the Student distribution $A(t|\nu)$, where ν is the number of degrees of freedom.

- $A(t_0|\nu)$ gives the probability that the value of t calculated for two samples with the same average is no more than t_0.

Statistical properties of data: comparing data sets

• Maybe one of the most used tests (in physics) is the **chi-square (χ^2) test**.

- With this test one can study binned data i.e. histograms.
- I.e. we have numbers N_i that tell the number of 'events' in bin i.

- We compare the histogram $\{N_i\}$ with a known distribution $\{n_i\}$ by computing the value of χ^2:

$$\chi^2 = \sum_{i=1}^{N_B} \frac{(N_i - n_i)^2}{n_i}.$$

- A large χ^2 tells that it is not probable that our data is a sample of the known distribution $\{n_i\}$.

- χ^2 has the distribution $Q(\chi^2|\nu)$ where ν is the number of degrees of freedom.
- $Q(\chi^2|\nu)$ tells the probability that the sum in the above equation is larger than χ^2.

- It is assumed that the terms in the above sum are distributed as Gaussian and their mean is zero and variance one.
- Number of degrees of freedom is the number of bins in the histogram N_B.
- If the n_i are normalized to the same as the histogram then the number of degrees of freedom is $N_\mathrm{b} - 1$.

Statistical properties of data: comparing data sets

- Let's look at the χ^2 distribution more carefully.

- Let's write χ^2 in the following way

$$\chi^2 = \sum_{i=1}^{N} z_i^2,$$

where z_i have Gaussian distribution, their mean is zero and variance one.

- If it were no the case we could scale the data: $z_i^2 \to (z_i - \bar{z}_i)^2/\sigma_i^2$

- We want to know the probability distribution of χ^2 $f(\chi^2)$.

- Let $\mathbf{z} = (z_1, z_2, \ldots, z_N)$. Now the probability density function can be written as

$$f(\mathbf{z})d^N\mathbf{z} \propto e^{-\chi^2/2}d^N\mathbf{z}.$$

- Here we think χ^2 as the norm of a vector in a N dimensional space.

- Volume element $d^N\mathbf{z}$ is proportional to

$$\chi^{N-1}d\chi = \frac{1}{2}(\chi^2)^{(N-2)/2}d\chi^2.$$

Statistical properties of data: comparing data sets

- So the probability density as a function of χ^2 is

$$f(\chi^2)d\chi^2 = C \cdot (\chi^2)^{N/2-1} e^{-\chi^2/2} d\chi^2 ,$$

where C is a normalizing constant obtained fomr equation

$$\int_0^\infty f(\chi^2)d\chi^2 = 1 .$$

- The final result is thus

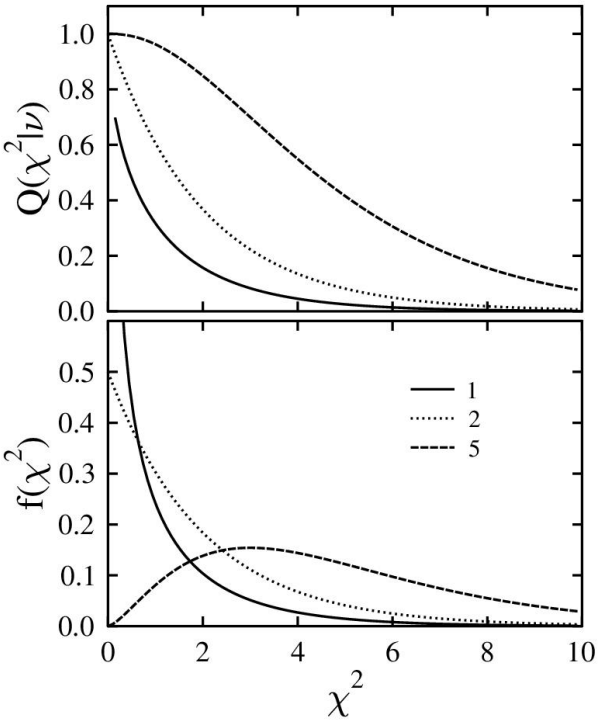
$$f(\chi^2)d\chi^2 = \frac{1}{\Gamma(N/2)} \left(\frac{\chi^2}{2}\right)^{N/2-1} e^{-\chi^2/2} d\left(\frac{\chi^2}{2}\right),$$

where Γ is the gamma function.

- The cumulative probability distribution function $Q(\chi^2|N)$ is obtained from $f(\chi^2)$ by integration.

Statistical properties of data: comparing data sets

- Below are given examples of the distributions:



Statistical properties of data: comparing data sets

• With the **Kolmogorov-Smirnov test** one can compare non-binned data with a know distribution or with another non-binned data set.

- Assume the data points are $\{x_1, x_2, \ldots, x_N\}$.
- Compute the estimate of the cumulative distribution function from the data:

  $S(x)$ gives the fraction of data points left of a given value x.
- When comparing the data set with a known probability distribution $P(x)$ we calculate

  $D = \max_{x \in [a,\, b]} |S(x) - P(x)|$.
- In case we compare two data sets:

  $D = \max_{x \in [a,\, b]} |S_1(x) - S_2(x)|$.

- The distribution of the K-S statistics (D) can be calculated enabling us to estimate the significance level of observed D_{obs}

$$P(D > D_{\text{obs}}) = Q_{\text{KS}}(\sqrt{N} D_{\text{obs}})$$ (for data set + known distribution)

$$P(D > D_{\text{obs}}) = Q_{\text{KS}}\left(\sqrt{\frac{N_1 N_2}{N_1 + N_2}} D_{\text{obs}}\right)$$ (for two data sets)

where $Q_{\text{KS}}(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2}$