Statistical Data Analysis
2017/18

London Postgraduate Lectures on Particle Physics;
University of London MSci course PH4515

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Course web page (moodle links to here):
www.pp.rhul.ac.uk/~cowan/stat_course.html
Course structure

The main lectures are from 3:00 to 5:00 and will cover statistical data analysis.

There is no assessed element in computing per se, although the coursework will use C++.

Through week 6 the hour from 5:00 to 6:00 will be a crash course in C++ (non-assessed, attend as needed).

From week 7, the hour from 5:00 to 6:00 will be used to discuss the coursework and go over additional examples.
Coursework, exams, etc.

9 problem sheets, provisionally due weeks 3 through 11.

Problems will only cover statistical data analysis, but for some problem sheets you will write simple C++ programs.

Please turn in your problem sheets on paper, Mondays at our lectures. Please staple the pages and indicate on the sheet your name, College and degree programme (PhD, MSc, MSci).

In general email or late submissions are not allowed unless due to exceptional circumstances and agreed with me.

For MSc/MSci students: problem sheets count 20% of the mark; written exam at end of year (80%).

For PhD students: assessment entirely through coursework; no material from this course in exam (~early next year).
Computing

The coursework includes C++ computing in a linux environment.

For PhD students, use your own accounts – usual HEP setup should be OK.

The computing problems require specific software (ROOT and its class library) – cannot just use e.g. visual C++.

Therefore for MSc/MSci students, you will get an account on the RHUL linux cluster. You then only need to be able to create an X-Window on your local machine, and from there you remotely login to RHUL.

For mac, install XQuartz from xquartz.macosforge.org and open a terminal window.

For windows, various options, e.g., mobaXterm or cygwin/X (see course page near bottom “information on computing”).
Statistical Data Analysis Outline

1. Probability, Bayes’ theorem
2. Random variables and probability densities
3. Expectation values, error propagation
4. Catalogue of pdfs
5. The Monte Carlo method
6. Statistical tests: general concepts
7. Test statistics, multivariate methods
8. Goodness-of-fit tests
9. Parameter estimation, maximum likelihood
10. More maximum likelihood
11. Method of least squares
12. Interval estimation, setting limits
13. Nuisance parameters, systematic uncertainties
14. Examples of Bayesian approach
Some statistics books, papers, etc.


C. Patrignani et al. (Particle Data Group), *Review of Particle Physics*, Chin. Phys. C, 40, 100001 (2016); see also [pdg.lbl.gov](http://pdg.lbl.gov) sections on probability, statistics, Monte Carlo
Data analysis in particle physics

Observe events of a certain type

Measure characteristics of each event (particle momenta, number of muons, energy of jets,...)

Theories (e.g. SM) predict distributions of these properties up to free parameters, e.g., $\alpha$, $G_F$, $M_Z$, $\alpha_s$, $m_H$, ...

Some tasks of data analysis:

- Estimate (measure) the parameters;
- Quantify the uncertainty of the parameter estimates;
- Test the extent to which the predictions of a theory are in agreement with the data.
Dealing with uncertainty

In particle physics there are various elements of uncertainty:

theory is not deterministic
quantum mechanics
random measurement errors
present even without quantum effects
things we could know in principle but don’t
e.g. from limitations of cost, time, ...

We can quantify the uncertainty using PROBABILITY
A definition of probability

Consider a set $S$ with subsets $A, B, ...$

For all $A \subset S$, $P(A) \geq 0$

$P(S) = 1$

If $A \cap B = \emptyset$, $P(A \cup B) = P(A) + P(B)$

From these axioms we can derive further properties, e.g.

$P(\overline{A}) = 1 - P(A)$

$P(A \cup \overline{A}) = 1$

$P(\emptyset) = 0$

if $A \subset B$, then $P(A) \leq P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Conditional probability, independence

Also define conditional probability of $A$ given $B$ (with $P(B) \neq 0$):

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}
\]

E.g. rolling dice:  
\[
P(n < 3 \mid n \text{ even}) = \frac{P((n<3) \cap n \text{ even})}{P(\text{even})} = \frac{1/6}{3/6} = \frac{1}{3}
\]

Subsets $A$, $B$ independent if:  
\[
P(A \cap B) = P(A)P(B)
\]

If $A$, $B$ independent,  
\[
P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)
\]

N.B. do not confuse with disjoint subsets, i.e.,  
\[
A \cap B = \emptyset
\]
Interpretation of probability

I. Relative frequency

\[ A, B, \ldots \text{ are outcomes of a repeatable experiment} \]

\[
P(A) = \lim_{n \to \infty} \frac{\text{times outcome is } A}{n}
\]

cf. quantum mechanics, particle scattering, radioactive decay...

II. Subjective probability

\[ A, B, \ldots \text{ are hypotheses (statements that are true or false)} \]

\[ P(A) = \text{degree of belief that } A \text{ is true} \]

- Both interpretations consistent with Kolmogorov axioms.
- In particle physics frequency interpretation often most useful, but subjective probability can provide more natural treatment of non-repeatable phenomena:
  
  systematic uncertainties, probability that Higgs boson exists,...
Bayes’ theorem

From the definition of conditional probability we have,

\[ P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)} \]

but \( P(A \cap B) = P(B \cap A) \), so

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

First published (posthumously) by the Reverend Thomas Bayes (1702–1761)

*An essay towards solving a problem in the doctrine of chances*, Philos. Trans. R. Soc. **53** (1763) 370; reprinted in *Biometrika*, **45** (1958) 293.
The law of total probability

Consider a subset $B$ of the sample space $S$, divided into disjoint subsets $A_i$ such that $\cup_i A_i = S$,

\[ B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i), \]

\[ P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i) \]

\[ P(B) = \sum_i P(B|A_i)P(A_i) \quad \text{law of total probability} \]

Bayes’ theorem becomes

\[ P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)} \]
An example using Bayes’ theorem

Suppose the probability (for anyone) to have a disease D is:

\[
P(D) = 0.001 \quad \text{← prior probabilities, i.e., before any test carried out}
\]

\[
P(\text{no D}) = 0.999
\]

Consider a test for the disease: result is + or −

\[
P(+|D) = 0.98 \quad \text{← probabilities to (in)correctly identify a person with the disease}
\]

\[
P(−|D) = 0.02
\]

\[
P(+|\text{no D}) = 0.03 \quad \text{← probabilities to (in)correctly identify a healthy person}
\]

\[
P(−|\text{no D}) = 0.97
\]

Suppose your result is +. How worried should you be?

G. Cowan
Bayes’ theorem example (cont.)

The probability to have the disease given a + result is

\[ p(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|\text{no } D)P(\text{no } D)} \]

\[ = \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} \]

\[ = 0.032 \quad \leftarrow \text{posterior probability} \]

i.e. you’re probably OK!

Your viewpoint: my degree of belief that I have the disease is 3.2%.

Your doctor’s viewpoint: 3.2% of people like this have the disease.
Frequentist Statistics – general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations (shorthand: $\bar{x}$). Probability = limiting frequency

Probabilities such as

$P (\text{Higgs boson exists}),$

$P (0.117 < \alpha_s < 0.121),$

etc. are either 0 or 1, but we don’t know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

The preferred theories (models, hypotheses, ...) are those for which our observations would be considered ‘usual’.
Bayesian Statistics – general philosophy

In Bayesian statistics, use subjective probability for hypotheses:

- probability of the data assuming hypothesis $H$ (the likelihood)
- prior probability, i.e., before seeing the data
- posterior probability, i.e., after seeing the data

Bayes’ theorem has an “if-then” character: If your prior probabilities were $\pi(H)$, then it says how these probabilities should change in the light of the data.

No general prescription for priors (subjective!)
Random variables and probability density functions

A random variable is a numerical characteristic assigned to an element of the sample space; can be discrete or continuous.

Suppose outcome of experiment is continuous value $x$

$$P(x \text{ found in } [x, x + dx]) = f(x) \, dx$$

$\rightarrow f(x) = \text{probability density function (pdf)}$

$$\int_{-\infty}^{\infty} f(x) \, dx = 1 \quad x \text{ must be somewhere}$$

Or for discrete outcome $x_i$ with e.g. $i = 1, 2, \ldots$ we have

$$P(x_i) = p_i \quad \text{probability mass function}$$

$$\sum_i P(x_i) = 1 \quad x \text{ must take on one of its possible values}$$
Cumulative distribution function

Probability to have outcome less than or equal to \( x \) is

\[
\int_{-\infty}^{x} f(x') \, dx' \equiv F(x)
\]

Alternatively define pdf with

\[
f(x) = \frac{\partial F(x)}{\partial x}
\]
**Histograms**

pdf = histogram with infinite data sample, zero bin width, normalized to unit area.

\[ f(x) = \frac{N(x)}{n \Delta x} \]

\( n \) = number of entries \( \Delta x \) = bin width
Multivariate distributions

Outcome of experiment characterized by several values, e.g. an $n$-component vector, $(x_1, \ldots, x_n)$

$$P(A \cap B) = \int f(x, y) \, dx \, dy$$

joint pdf

Normalization: $$\int \cdots \int f(x_1, \ldots, x_n) \, dx_1 \cdots dx_n = 1$$
Marginal pdf

Sometimes we want only pdf of some (or one) of the components:

\[ P(A) = \sum_i P(A \cap B_i) \]

\[ = \sum_i f(x, y_i) \, dy \, dx \]

\[ \rightarrow \int f(x, y) \, dy \, dx \]

\[ f_x(x) = \int f(x, y) \, dy \]

\[ \rightarrow \text{marginal pdf} \quad f_1(x_1) = \int \cdots \int f(x_1, \ldots, x_n) \, dx_2 \cdots dx_n \]

\[ x_1, x_2 \text{ independent if } f(x_1, x_2) = f_1(x_1) f_2(x_2) \]
Marginal pdf (2)

Marginal pdf ~ projection of joint pdf onto individual axes.
Conditional pdf

Sometimes we want to consider some components of joint pdf as constant. Recall conditional probability:

\[
P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{f(x, y) \, dx \, dy}{f_x(x) \, dx}
\]

→ conditional pdfs:

\[
h(y|x) = \frac{f(x, y)}{f_x(x)}, \quad g(x|y) = \frac{f(x, y)}{f_y(y)}
\]

Bayes’ theorem becomes:

\[
g(x|y) = \frac{h(y|x) \, f_x(x)}{f_y(y)}.
\]

Recall \(A, B\) independent if \(P(A \cap B) = P(A)P(B)\).

→ \(x, y\) independent if \(f(x, y) = f_x(x) \, f_y(y)\).
Conditional pdfs (2)

E.g. joint pdf $f(x,y)$ used to find conditional pdfs $h(y|x_1)$, $h(y|x_2)$:

Basically treat some of the r.v.s as constant, then divide the joint pdf by the marginal pdf of those variables being held constant so that what is left has correct normalization, e.g., $\int h(y|x) \, dy = 1$. 
Functions of a random variable

A function of a random variable is itself a random variable. Suppose \( x \) follows a pdf \( f(x) \), consider a function \( a(x) \).

What is the pdf \( g(a) \)?

\[
g(a) \, da = \int_{dS} f(x) \, dx
\]

\( dS = \) region of \( x \) space for which \( a \) is in \([a, a+da]\).

For one-variable case with unique inverse this is simply

\[
g(a) \, da = f(x) \, dx
\]

\[
\rightarrow g(a) = f(x(a)) \left| \frac{dx}{da} \right|
\]
Functions without unique inverse

If inverse of \( a(x) \) not unique, include all \( dx \) intervals in \( dS \) which correspond to \( da \):

\[
\text{Example: } \quad a = x^2, \quad x = \pm \sqrt{a}, \quad dx = \pm \frac{da}{2\sqrt{a}}.
\]

\[
dS = \left[ \sqrt{a}, \sqrt{a} + \frac{da}{2\sqrt{a}} \right] \cup \left[ -\sqrt{a} - \frac{da}{2\sqrt{a}}, -\sqrt{a} \right]
\]

\[
g(a) = \frac{f(\sqrt{a})}{2\sqrt{a}} + \frac{f(-\sqrt{a})}{2\sqrt{a}}
\]
Functions of more than one r.v.

Consider r.v.s \( \vec{x} = (x_1, \ldots, x_n) \) and a function \( a(\vec{x}) \).

\[
g(a')\,da' = \int \cdots \int_{dS} f(x_1, \ldots, x_n)\,dx_1 \cdots dx_n
\]

\( dS = \) region of \( x \)-space between (hyper)surfaces defined by

\[
a(\vec{x}) = a', \quad a(\vec{x}) = a' + da'
\]
Functions of more than one r.v. (2)

Example: r.v.s $x, y > 0$ follow joint pdf $f(x,y)$, consider the function $z = xy$. What is $g(z)$?

$$g(z) \, dz = \int \ldots \int_{dS} f(x, y) \, dx \, dy$$

$$= \int_0^\infty dx \int_{(z+dz)/x}^{z/x} f(x, y) \, dy$$

$$\rightarrow g(z) = \int_0^\infty f(x, \frac{z}{x}) \frac{dx}{x}$$

$$= \int_0^\infty f\left( \frac{z}{y}, y \right) \frac{dy}{y}$$

(Mellin convolution)
More on transformation of variables

Consider a random vector \( \vec{x} = (x_1, \ldots, x_n) \) with joint pdf \( f(\vec{x}) \).

Form \( n \) linearly independent functions \( \vec{y}(\vec{x}) = (y_1(\vec{x}), \ldots, y_n(\vec{x})) \)
for which the inverse functions \( x_1(\vec{y}), \ldots, x_n(\vec{y}) \) exist.

Then the joint pdf of the vector of functions is \( g(\vec{y}) = |J|f(\vec{x}) \)

where \( J \) is the

Jacobian determinant: \( J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \)

For e.g. \( g_1(y_1) \) integrate \( g(\vec{y}) \) over the unwanted components.
Expectation values

Consider continuous r.v. $x$ with pdf $f(x)$.

Define expectation (mean) value as

$$E[x] = \int x \, f(x) \, dx$$

Notation (often): $E[x] = \mu \sim \text{“centre of gravity” of pdf.}$

For a function $y(x)$ with pdf $g(y)$,

$$E[y] = \int y \, g(y) \, dy = \int y(x) \, f(x) \, dx \quad \text{(equivalent)}$$

Variance:

$$V[x] = E[x^2] - \mu^2 = E[(x - \mu)^2]$$

Notation:

$$V[x] = \sigma^2$$

Standard deviation:

$$\sigma = \sqrt{\sigma^2}$$

$\sigma \sim \text{width of pdf, same units as } x.$
**Covariance and correlation**

Define covariance $\text{cov}[x,y]$ (also use matrix notation $V_{xy}$) as

$$\text{cov}[x,y] = E[xy] - \mu_x \mu_y = E[(x - \mu_x)(y - \mu_y)]$$

Correlation coefficient (dimensionless) defined as

$$\rho_{xy} = \frac{\text{cov}[x,y]}{\sigma_x \sigma_y}$$

If $x, y$, independent, i.e., $f(x,y) = f_x(x)f_y(y)$, then

$$E[xy] = \int \int xy f(x,y) \, dx \, dy = \mu_x \mu_y$$

$$\rightarrow \text{cov}[x,y] = 0 \quad x \text{ and } y, \text{ ‘uncorrelated’}$$

N.B. converse not always true.
Correlation (cont.)

$\rho = 0.75$

$\rho = -0.75$

$\rho = 0.95$

$\rho = 0.25$
Error propagation

Suppose we measure a set of values $\vec{x} = (x_1, \ldots, x_n)$

and we have the covariances $V_{ij} = \text{COV}[x_i, x_j]$

which quantify the measurement errors in the $x_i$.

Now consider a function $y(\vec{x})$.

What is the variance of $y(\vec{x})$?

The hard way: use joint pdf $f(\vec{x})$ to find the pdf $g(y)$,
then from $g(y)$ find $V[y] = E[y^2] - (E[y])^2$.

Often not practical, $f(\vec{x})$ may not even be fully known.
Error propagation (2)

Suppose we had \( \bar{\mu} = E[\bar{x}] \)

in practice only estimates given by the measured \( \bar{x} \)

Expand \( y(\bar{x}) \) to 1st order in a Taylor series about \( \bar{\mu} \)

\[
y(\bar{x}) \approx y(\bar{\mu}) + \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{\bar{x}=\bar{\mu}} (x_i - \mu_i)
\]

To find \( V[y] \) we need \( E[y^2] \) and \( E[y] \).

\[
E[y(\bar{x})] \approx y(\bar{\mu}) \quad \text{since} \quad E[x_i - \mu_i] = 0
\]
Error propagation (3)

\[ E[y^2(\bar{x})] \approx y^2(\mu) + 2y(\mu) \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{\bar{x}=\mu} E[x_i - \mu] \]

\[ + E \left[ \left( \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{\bar{x}=\mu} (x_i - \mu_i) \right) \left( \sum_{j=1}^{n} \left[ \frac{\partial y}{\partial x_j} \right]_{\bar{x}=\mu} (x_j - \mu_j) \right) \right] \]

\[ = y^2(\mu) + \sum_{i,j=1}^{n} \left[ \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\bar{x}=\mu} V_{ij} \]

Putting the ingredients together gives the variance of \( y(\bar{x}) \)

\[ \sigma_y^2 \approx \sum_{i,j=1}^{n} \left[ \frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\bar{x}=\mu} V_{ij} \]
Error propagation (4)

If the $x_i$ are uncorrelated, i.e., $V_{ij} = \sigma_i^2 \delta_{ij}$, then this becomes

$$\sigma_y^2 \approx \sum_{i=1}^{n} \left[ \frac{\partial y}{\partial x_i} \right]_{{\bar{x} = \mu}}^2 \sigma_i^2$$

Similar for a set of $m$ functions $y(\bar{x}) = (y_1(\bar{x}), \ldots, y_m(\bar{x}))$

$$U_{kl} = \text{cov}[y_k, y_l] \approx \sum_{i,j=1}^{n} \left[ \frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \right]_{{\bar{x} = \mu}} V_{ij}$$

or in matrix notation $U = AVA^T$, where

$$A_{i,j} = \left[ \frac{\partial y_i}{\partial x_j} \right]_{{\bar{x} = \mu}}$$
Error propagation (5)

The ‘error propagation’ formulae tell us the covariances of a set of functions
\( \bar{y}(\bar{x}) = (y_1(\bar{x}), \ldots, y_m(\bar{x})) \) in terms of the covariances of the original variables.

Limitations: exact only if \( \bar{y}(\bar{x}) \) linear. Approximation breaks down if function nonlinear over a region comparable in size to the \( \sigma_i \).

N.B. We have said nothing about the exact pdf of the \( x_i \), e.g., it doesn’t have to be Gaussian.
Error propagation — special cases

\[ y = x_1 + x_2 \rightarrow \sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\text{cov}[x_1, x_2] \]

\[ y = x_1x_2 \rightarrow \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{\text{cov}[x_1, x_2]}{x_1x_2} \]

That is, if the \( x_i \) are uncorrelated:

- add errors quadratically for the sum (or difference),
- add relative errors quadratically for product (or ratio).

But correlations can change this completely...
Error propagation – special cases (2)

Consider $y = x_1 - x_2$ with

$$\mu_1 = \mu_2 = 10, \quad \sigma_1 = \sigma_2 = 1, \quad \rho = \frac{\text{COV}[x_1, x_2]}{\sigma_1 \sigma_2} = 0.$$  

$$V[y] = 1^2 + 1^2 = 2, \quad \rightarrow \quad \sigma_y = 1.4$$

Now suppose $\rho = 1$. Then

$$V[y] = 1^2 + 1^2 - 2 = 0, \quad \rightarrow \quad \sigma_y = 0$$

i.e. for 100% correlation, error in difference $\rightarrow 0$. 
Short catalogue of distributions

We will now run through a short catalog of probability functions and pdfs.

For each (usually) show expectation value, variance, a plot and discuss some properties and applications.

See also chapter on probability from pdg.lbl.gov

For a more complete catalogue see e.g. the handbook on statistical distributions by Christian Walck from

www.fysik.su.se/~walck/suf9601.pdf
## Some distributions

<table>
<thead>
<tr>
<th>Distribution/pdf</th>
<th>Example use in HEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial</td>
<td>Branching ratio</td>
</tr>
<tr>
<td>Multinomial</td>
<td>Histogram with fixed $N$</td>
</tr>
<tr>
<td>Poisson</td>
<td>Number of events found</td>
</tr>
<tr>
<td>Uniform</td>
<td>Monte Carlo method</td>
</tr>
<tr>
<td>Exponential</td>
<td>Decay time</td>
</tr>
<tr>
<td>Gaussian</td>
<td>Measurement error</td>
</tr>
<tr>
<td>Chi-square</td>
<td>Goodness-of-fit</td>
</tr>
<tr>
<td>Cauchy</td>
<td>Mass of resonance</td>
</tr>
<tr>
<td>Landau</td>
<td>Ionization energy loss</td>
</tr>
<tr>
<td>Beta</td>
<td>Prior pdf for efficiency</td>
</tr>
<tr>
<td>Gamma</td>
<td>Sum of exponential variables</td>
</tr>
<tr>
<td>Student’s $t$</td>
<td>Resolution function with adjustable tails</td>
</tr>
</tbody>
</table>
Binomial distribution

Consider $N$ independent experiments (Bernoulli trials):

- outcome of each is ‘success’ or ‘failure’,
- probability of success on any given trial is $p$.

Define discrete r.v. $n =$ number of successes ($0 \leq n \leq N$).

Probability of a specific outcome (in order), e.g. ‘ssfsf’ is

$$pp(1-p)p(1-p) = p^n (1-p)^{N-n}$$

But order not important; there are

$$\frac{N!}{n!(N-n)!}$$

ways (permutations) to get $n$ successes in $N$ trials, total probability for $n$ is sum of probabilities for each permutation.
Binomial distribution (2)

The binomial distribution is therefore

\[ f(n; N, p) = \frac{N!}{n!(N - n)!} p^n (1 - p)^{N-n} \]

random variable  parameters

For the expectation value and variance we find:

\[ E[n] = \sum_{n=0}^{N} n f(n; N, p) = Np \]

\[ V[n] = E[n^2] - (E[n])^2 = Np(1 - p) \]
Binomial distribution (3)

Binomial distribution for several values of the parameters:

Example: observe $N$ decays of $W^\pm$, the number $n$ of which are $W \rightarrow \mu \nu$ is a binomial r.v., $p =$ branching ratio.
Multinomial distribution

Like binomial but now $m$ outcomes instead of two, probabilities are

$\vec{p} = (p_1, \ldots, p_m)$, \hspace{1cm} \text{with} \hspace{1cm} \sum_{i=1}^{m} p_i = 1.$

For $N$ trials we want the probability to obtain:

- $n_1$ of outcome 1,
- $n_2$ of outcome 2,
- $\vdots$
- $n_m$ of outcome $m$.

This is the multinomial distribution for $\vec{n} = (n_1, \ldots, n_m)$

$$f(\vec{n}; N, \vec{p}) = \frac{N!}{n_1!n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$$
Multinomial distribution (2)

Now consider outcome $i$ as ‘success’, all others as ‘failure’.

→ all $n_i$ individually binomial with parameters $N, p_i$

$$E[n_i] = Np_i, \quad V[n_i] = Np_i(1 - p_i) \quad \text{for all } i$$

One can also find the covariance to be

$$V_{ij} = Np_i(\delta_{ij} - p_j)$$

Example: $\vec{n} = (n_1, \ldots, n_m)$ represents a histogram with $m$ bins, $N$ total entries, all entries independent.
Poisson distribution

Consider binomial $n$ in the limit

$$N \to \infty, \quad p \to 0, \quad E[n] = Np \to \nu.$$ 

$n$ follows the Poisson distribution:

$$f(n; \nu) = \frac{\nu^n}{n!} e^{-\nu} \quad (n \geq 0)$$

$$E[n] = \nu, \quad V[n] = \nu.$$

Example: number of scattering events $n$ with cross section $\sigma$ found for a fixed integrated luminosity, with $\nu = \sigma \int L \, dt$. 

![Graphs of Poisson distribution for different values of $\nu$]
Uniform distribution

Consider a continuous r.v. \( x \) with \( -\infty < x < \infty \). Uniform pdf is:

\[
f(x; \alpha, \beta) = \begin{cases} 
\frac{1}{\beta - \alpha} & \alpha \leq x \leq \beta \\
0 & \text{otherwise}
\end{cases}
\]

\[
E[x] = \frac{1}{2}(\alpha + \beta)
\]

\[
V[x] = \frac{1}{12}(\beta - \alpha)^2
\]

N.B. For any r.v. \( x \) with cumulative distribution \( F(x) \), \( y = F(x) \) is uniform in \([0, 1]\).

Example: for \( \pi^0 \to \gamma\gamma \), \( E_\gamma \) is uniform in \([E_{\text{min}}, E_{\text{max}}]\), with

\[
E_{\text{min}} = \frac{1}{2}E_\pi(1 - \beta) , \quad E_{\text{max}} = \frac{1}{2}E_\pi(1 + \beta)
\]
Exponential distribution

The exponential pdf for the continuous r.v. \( x \) is defined by:

\[
f(x; \xi) = \begin{cases} \frac{1}{\xi} e^{-x/\xi} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}
\]

\[
E[x] = \xi
\]

\[
V[x] = \xi^2
\]

Example: proper decay time \( t \) of an unstable particle

\[
f(t; \tau) = \frac{1}{\tau} e^{-t/\tau} \quad (\tau = \text{mean lifetime})
\]

Lack of memory (unique to exponential): \( f(t - t_0 | t \geq t_0) = f(t) \)
Gaussian distribution

The Gaussian (normal) pdf for a continuous r.v. $x$ is defined by:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$

$E[x] = \mu$  (N.B. often $\mu$, $\sigma^2$ denote mean, variance of any r.v., not only Gaussian.)

$V[x] = \sigma^2$

Special case: $\mu = 0$, $\sigma^2 = 1$  (‘standard Gaussian’):

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) = \int_{-\infty}^{x} \varphi(x') \, dx'$$

If $y \sim$ Gaussian with $\mu$, $\sigma^2$, then $x = (y - \mu) / \sigma$ follows $\varphi(x)$. 
Gaussian pdf and the Central Limit Theorem

The Gaussian pdf is so useful because almost any random variable that is a sum of a large number of small contributions follows it. This follows from the Central Limit Theorem:

For \( n \) independent r.v.s \( x_i \) with finite variances \( \sigma_i^2 \), otherwise arbitrary pdfs, consider the sum

\[
y = \sum_{i=1}^{n} x_i
\]

In the limit \( n \to \infty \), \( y \) is a Gaussian r.v. with

\[
E[y] = \sum_{i=1}^{n} \mu_i \quad V[y] = \sum_{i=1}^{n} \sigma_i^2
\]

Measurement errors are often the sum of many contributions, so frequently measured values can be treated as Gaussian r.v.s.
Central Limit Theorem (2)

The CLT can be proved using characteristic functions (Fourier transforms), see, e.g., SDA Chapter 10.

For finite \( n \), the theorem is approximately valid to the extent that the fluctuation of the sum is not dominated by one (or few) terms.

⚠️ Beware of measurement errors with non-Gaussian tails.

Good example: velocity component \( v_x \) of air molecules.

OK example: total deflection due to multiple Coulomb scattering. (Rare large angle deflections give non-Gaussian tail.)

Bad example: energy loss of charged particle traversing thin gas layer. (Rare collisions make up large fraction of energy loss, cf. Landau pdf.)
Multivariate Gaussian distribution

Multivariate Gaussian pdf for the vector \( \bar{x} = (x_1, \ldots, x_n) \):

\[
f(\bar{x}; \mu, V) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp \left[ -\frac{1}{2}(\bar{x} - \mu)^T V^{-1} (\bar{x} - \mu) \right]
\]

\( \bar{x}, \mu \) are column vectors, \( \bar{x}^T, \mu^T \) are transpose (row) vectors,

\[
E[x_i] = \mu_i, \quad \text{COV}[x_i, x_j] = V_{ij}.
\]

For \( n = 2 \) this is

\[
f(x_1, x_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \times \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[ \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left( \frac{x_1 - \mu_1}{\sigma_1} \right) \left( \frac{x_2 - \mu_2}{\sigma_2} \right) \right] \right\}
\]

where \( \rho = \text{COV}[x_1, x_2]/(\sigma_1\sigma_2) \) is the correlation coefficient.
Chi-square ($\chi^2$) distribution

The chi-square pdf for the continuous r.v. $z$ ($z \geq 0$) is defined by

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

$n = 1, 2, \ldots = \text{number of ‘degrees of freedom’ (dof)}$

$$E[z] = n, \quad V[z] = 2n.$$ 

For independent Gaussian $x_i, i = 1, \ldots, n$, means $\mu_i$, variances $\sigma_i^2$,

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

follows $\chi^2$ pdf with $n$ dof.

Example: goodness-of-fit test variable especially in conjunction with method of least squares.
Cauchy (Breit-Wigner) distribution

The Breit-Wigner pdf for the continuous r.v. \( x \) is defined by

\[
f(x; \Gamma, x_0) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}
\]

\((\Gamma = 2, x_0 = 0 \text{ is the Cauchy pdf.})\)

\[E[x]\] not well defined, \( V[x] \rightarrow \infty\).

\(x_0 = \text{mode (most probable value)}\)

\(\Gamma = \text{full width at half maximum}\)

Example: mass of resonance particle, e.g. \( \rho, K^*, \phi^0, \ldots \)

\(\Gamma = \text{decay rate (inverse of mean lifetime)}\)
Landau distribution

For a charged particle with $\beta = \nu / c$ traversing a layer of matter of thickness $d$, the energy loss $\Delta$ follows the Landau pdf:

$$f(\Delta; \beta) = \frac{1}{\xi} \phi(\lambda),$$

$$\phi(\lambda) = \frac{1}{\pi} \int_0^\infty \exp(-u \ln u - \lambda u) \sin \pi u \, du,$$

$$\lambda = \frac{1}{\xi} \left[ \Delta - \xi \left( \ln \frac{\xi}{\epsilon'} + 1 - \gamma_E \right) \right],$$

$$\xi = \frac{2\pi N_\Lambda e^4 z^2 \rho \sum Z \frac{d}{\beta^2}}{m_e c^2 \sum A}, \quad \epsilon' = \frac{I^2 \exp \beta^2}{2m_e c^2 \beta^2 \gamma^2}.$$

L. Landau, J. Phys. USSR 8 (1944) 201; see also
Landau distribution (2)

Long ‘Landau tail’
→ all moments $\infty$

Mode (most probable value) sensitive to $\beta$,
→ particle i.d.
Beta distribution

\[ f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \]

\[ E[x] = \frac{\alpha}{\alpha + \beta} \]

\[ V[x] = \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \]

Often used to represent pdf of continuous r.v. nonzero only between finite limits.
Gamma distribution

\[ f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha) \beta^\alpha} x^{\alpha-1} e^{-x/\beta} \]

\[ E[x] = \alpha \beta \]

\[ V[x] = \alpha \beta^2 \]

Often used to represent pdf of continuous r.v. nonzero only in \([0, \infty]\).

Also e.g. sum of \(n\) exponential r.v.s or time until \(n\)th event in Poisson process \(\sim\) Gamma
Student's $t$ distribution

\[
f(x; \nu) = \frac{\Gamma \left( \frac{\nu+1}{2} \right)}{\sqrt{\nu \pi} \Gamma(\nu/2)} \left( 1 + \frac{x^2}{\nu} \right)^{-\left( \frac{\nu+1}{2} \right)}
\]

\[
E[x] = 0 \quad (\nu > 1)
\]

\[
V[x] = \frac{\nu}{\nu - 2} \quad (\nu > 2)
\]

$\nu =$ number of degrees of freedom (not necessarily integer)

$\nu = 1$ gives Cauchy,

$\nu \to \infty$ gives Gaussian.
Student's \( t \) distribution (2)

If \( x \sim \text{Gaussian} \) with \( \mu = 0, \sigma^2 = 1 \), and
\[ z \sim \chi^2 \text{ with } n \text{ degrees of freedom}, \]then
\[ t = \frac{x}{(z/n)^{1/2}} \text{ follows Student's } t \text{ with } \nu = n. \]

This arises in problems where one forms the ratio of a sample mean to the sample standard deviation of Gaussian r.v.s.

The Student's \( t \) provides a bell-shaped pdf with adjustable tails, ranging from those of a Gaussian, which fall off very quickly, \((\nu \rightarrow \infty, \text{ but in fact already very Gauss-like for } \nu = \text{ two dozen})\), to the very long-tailed Cauchy \((\nu = 1)\).

Developed in 1908 by William Gosset, who worked under the pseudonym "Student" for the Guinness Brewery.
Extra slides
Theory ↔ Statistics ↔ Experiment

Theory (model, hypothesis):

\[ L = - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \bar{\psi} \gamma \psi + \ldots \]

\[ \sigma = \frac{G_F \alpha_s^2 m_H^2}{288 \sqrt{2} \pi} \times \]

+ simulation of detector and cuts

Experiment:

+ data selection
Data analysis in particle physics

Observe events (e.g., pp collisions) and for each, measure a set of characteristics:

- particle momenta, number of muons, energy of jets,...

Compare observed distributions of these characteristics to predictions of theory. From this, we want to:

- Estimate the free parameters of the theory: \( m_H = 125.4 \text{ GeV} \pm 0.4 \text{ GeV} \)
- Quantify the uncertainty in the estimates:
- Assess how well a given theory stands in agreement with the observed data:
  - \( 0^+ \) good, \( 2^+ \) bad

To do this we need a clear definition of PROBABILITY
Data analysis in particle physics: testing hypotheses

Test the extent to which a given model agrees with the data:

In general need tests with well-defined properties and quantitative results.