Introduction to Statistical Methods for High Energy Physics

2009 CERN Summer Student Lectures

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

- CERN course web page:
  www.pp.rhul.ac.uk/~cowan/stat_cern.html
- See also University of London course web page:
  www.pp.rhul.ac.uk/~cowan/stat_course.html

Outline

Lecture 1
- Probability
- Random variables, probability densities, etc.

Lecture 2
- Brief catalogue of probability densities
- The Monte Carlo method.

Lecture 3
- Statistical tests
- Fisher discriminants, neural networks, etc
- Significance and goodness-of-fit tests

Lecture 4
- Parameter estimation
- Maximum likelihood and least squares
- Interval estimation (setting limits)

Some statistics books, papers, etc.

G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998
see also **www.pp.rhul.ac.uk/~cowan/sda**

R.J. Barlow, Statistics, A Guide to the Use of Statistical in the Physical Sciences, Wiley, 1989
see also **hepwww.ph.man.ac.uk/~roger/book.html**

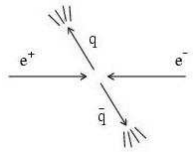
L. Lyons, *Statistics for Nuclear and Particle Physics*, CUP, 1986

F. James, *Statistical Methods in Experimental Physics*, 2nd ed., World Scientific, 2006; (W. Eadie et al., 1971).

S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998 (with program library on CD)

W.-M. Yao et al. (Particle Data Group), *Review of Particle Physics*, J. Physics G 33 (2006) 1; see also **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., α, G_F, M_Z, α_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 (→ presence of New Physics?)

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, ...

Kolmogorov axioms (1933)

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: $P(n < 3 \,|\, n \text{ even}) = \frac{P((n<3) \cap n \text{ even})}{P(\text{even})} = \frac{1/6}{3/6} = \frac{1}{3}$

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, ... 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, ... 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

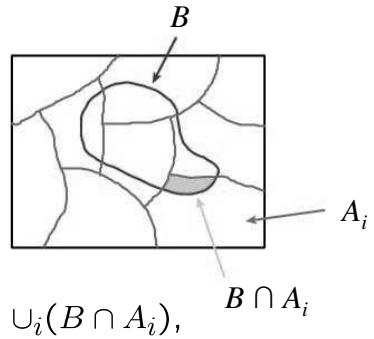
$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

Bayes' theorem

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 $\cup_i A_i = S$,

$$\rightarrow \; B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i),$$

$$\rightarrow \; P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$$

$$\rightarrow \; P(B) = \sum_i P(B|A_i)P(A_i)$$ law of total probability

Bayes' theorem becomes $\boxed{P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}}$

An example using Bayes' theorem

Suppose the probability (for anyone) to have AIDS is:

$P(\text{AIDS}) = 0.001$ ← prior probabilities, i.e., before any test carried out

$P(\text{no AIDS}) = 0.999$

Consider an AIDS test: result is + or −

$P(+|\text{AIDS}) = 0.98$ ← probabilities to (in)correctly identify an infected person

$P(-|\text{AIDS}) = 0.02$

$P(+|\text{no AIDS}) = 0.03$ ← probabilities to (in)correctly identify an uninfected person

$P(-|\text{no AIDS}) = 0.97$

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

Bayes' theorem example (cont.)

The probability to have AIDS given a + result is

$$P(\text{AIDS}|+) = \frac{P(+|\text{AIDS})P(\text{AIDS})}{P(+|\text{AIDS})P(\text{AIDS}) + P(+|\text{no AIDS})P(\text{no AIDS})}$$

$$= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999}$$

$$= 0.032$$ ← posterior probability

i.e. you're probably OK!

Your viewpoint: my degree of belief that I have AIDS is 3.2%

Your doctor's viewpoint: 3.2% of people like this will have AIDS

Frequentist Statistics – general philosophy

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

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),
P (0.117 < α_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

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

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 unique 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

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

→ $f(x)$ = probability density function (pdf)

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

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

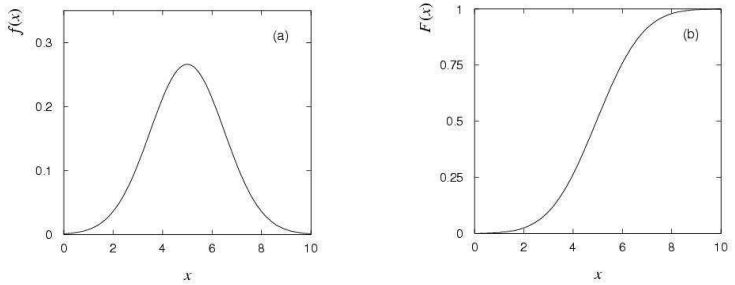
$$P(x_i) = p_i \qquad \text{probability mass function}$$

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

Cumulative distribution function

Probability to have outcome less than or equal to x is

$$\int_{-\infty}^{x} f(x')\,dx' \equiv F(x) \qquad \text{cumulative distribution function}$$

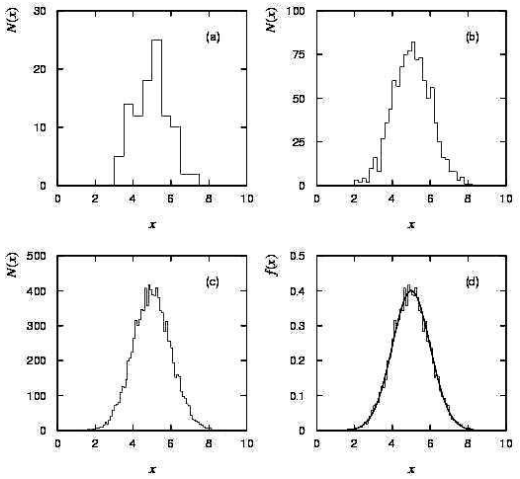


Alternatively define pdf with $f(x) = \dfrac{\partial F(x)}{\partial x}$

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

Δx = bin width

Other types of probability densities

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

→ joint pdf $f(x_1, \ldots, x_n)$

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

→ marginal pdf $f_1(x_1) = \int \cdots \int f(x_1, \ldots, x_n)\, dx_2 \ldots dx_n$

x_1, x_2 independent if $f(x_1, x_2) = f_1(x_1) f_2(x_2)$

Sometimes we want to consider some components as constant

→ conditional pdf $g(x_1|x_2) = \dfrac{f(x_1, x_2)}{f_2(x_2)}$

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): $E[x] = \mu$ ~ "centre of gravity" of pdf.

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

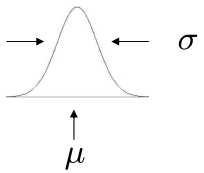
$$E[y] = \int y\, g(y)\, dy = \int y(x) f(x)\, dx \qquad \text{(equivalent)}$$

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

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

Standard deviation: $\sigma = \sqrt{\sigma^2}$

σ ~ width of pdf, same units as x.



Covariance and correlation

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

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

Correlation coefficient (dimensionless) defined as

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

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

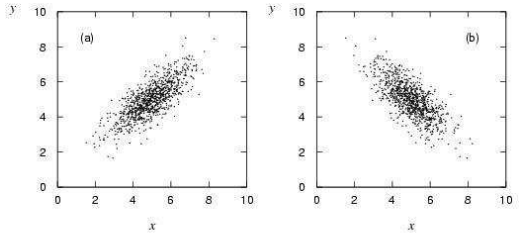
$$E[xy] = \int\int xy\, f(x, y)\, dx dy = \mu_x \mu_y$$

→ $\mathrm{cov}[x, y] = 0$ x and y, 'uncorrelated'

N.B. converse not always true.

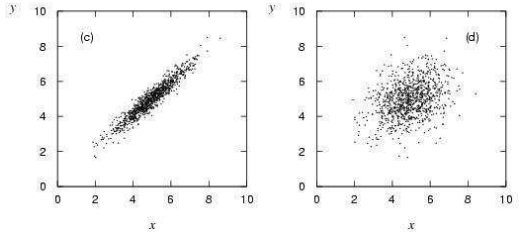
Correlation (cont.)

$\rho = 0.75$



$\rho = -0.75$

$\rho = 0.95$



$\rho = 0.25$

Error propagation

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

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

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 $\vec{\mu} = E[\vec{x}]$

in practice only estimates given by the measured $\vec{x}$

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

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

To find $V[y]$ we need $E[y^2]$ and $E[y]$.

$$E[y(\vec{x})] \approx y(\vec{\mu}) \quad \text{since} \quad E[x_i - \mu_i] = 0$$

Error propagation (3)

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

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

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

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

$$\sigma_y^2 \approx \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Error propagation (4)

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

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

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

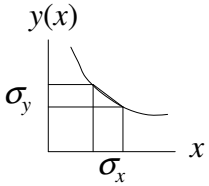
$$U_{kl} = \mathrm{cov}[y_k, y_l] \approx \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

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

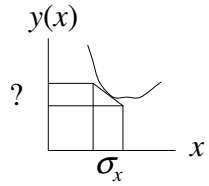
$$A_{ij} = \left[\frac{\partial y_i}{\partial x_j} \right]_{\vec{x}=\vec{\mu}}$$

Error propagation (5)

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



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

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

$$y = x_1 x_2 \quad \rightarrow \quad \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{\mathrm{cov}[x_1, x_2]}{x_1 x_2}$$

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

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

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

Now suppose $\rho = 1$. Then

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

i.e. for 100% correlation, error in difference → 0.

Wrapping up lecture 1

Up to now we've talked some abstract properties of probability:

- definition and interpretation,
- Bayes' theorem,
- random variables,
- probability density functions,
- expectation values,...

Next time we'll look at some probability distributions that come up in Particle Physics, and also discuss the Monte Carlo method, a valuable technique for computing probabilities.