Statistical Data Analysis 2020/21
Lecture Week 1

London Postgraduate Lectures on Particle Physics

University of London MSc/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 via RHUL moodle (PH4515) and also

www.pp.rhul.ac.uk/~cowan/stat_course.html

Statistical Data Analysis
Lecture 1-1

- Course structure and policies
- Outline of topics
- Some resources

Course Info (1)

This year completely online:

2 hours per week of recorded lectures (the core material), will be available by start of every Monday.

1 hour per week discussion session (schedule tbc): live video via MS Teams with additional examples, discussion of problem sheets, computing tools, etc.

You should view the lectures before the discussion session.

There will be 9 problem sheets.

Some paper & pencil, some computing problems.

Due weekly Mondays 5 pm starting lecture week 3 through 11.

Late submissions according to College policy (10% off for 24h, then no credit unless agreed).

Course Info (2)

For problem sheets, scan/merge into a single pdf file.

Filename: yourName_stat_probsheet_*n*.pdf (n = 1,2,...)

Please no hi-res photos from phone (use iScanner or similar).

MSc/MSci students: upload into Turnitin on moodle.

PhD students: email single pdf attachment to **g.cowan@rhul.ac.uk**

Subject line: statistics problem sheet *n*

Put your name in the pdf file (not just in the email).

Also write on the problem sheet your College and degree programme (MSc, MSci, PhD).

For MSc/MSci students, written exam at end of year (May 2021); format (e.g., open/closed-book) will depend on covid situation.

Msc/Msci: Exam worth 80%, problem sheets 20%.
For PhD students, no statistics exam.

Computing

The coursework includes short computer programs.

Some choice of language – best to use python or C++ in a linux environment (some flexibility here).

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

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

For MSc/MSci students, you will get an account on the RHUL linux cluster. You create an X-Window on your local machine (e.g. laptop), and from there you remotely login to RHUL.

> For mac, install XQuartz from **`www.xquartz.org`** and open a terminal window. For windows, various options, e.g., mobaXterm or cygwin/X (more info on web page).

Course 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.

G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998

R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989

Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014.

Luca Lista, *Statistical Methods for Data Analysis in Particle Physics*, Springer, 2017.

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

F. James., *Statistical and Computational Methods in Experimental Physics*, 2nd ed., World Scientific, 2006

S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998.

P.A. Zyla et al. (Particle Data Group), Prog. Theor. Exp. Phys. 2020, 083C01 (2020); **`pdg.lbl.gov`** sections on probability, statistics, MC.

Statistical Data Analysis
Lecture 1-2

- Tasks of statistical data analysis in science
- The role of uncertainty
- Definition of probability

Theory ↔ Statistics ↔ Experiment

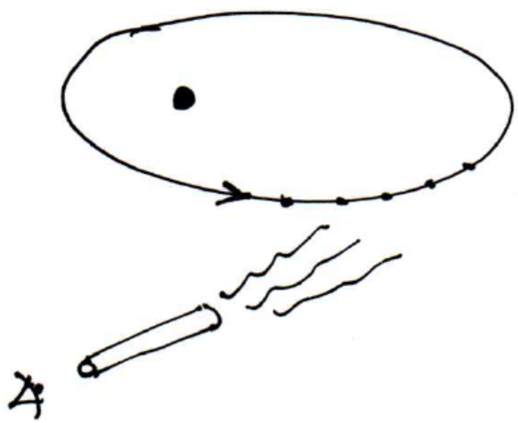
Theory (model, hypothesis):

$$F = -G\frac{m_1 m_2}{r^2}, \ldots$$

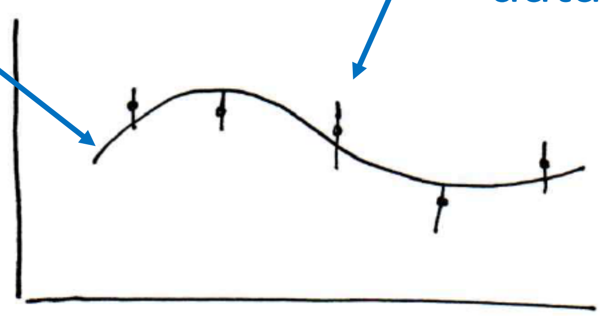
+ response of measurement apparatus

= model prediction

Experiment (observation):



data



Some tasks of statistical data analysis

Compare data to predictions of competing models

Most models contain adjustable parameters
(e.g., particle physics: G_F, M_Z, α_s, m_H,...)

Estimate (measure) the unknown parameters

Quantify uncertainty in parameter estimates

Test and quantify the extent to which the model is in agreement with the data.

Uncertainty

Uncertainty enters on several levels

Measurements not in general exactly reproducible

- Quantum effects
- Random effects (even without QM)

Model prediction uncertain

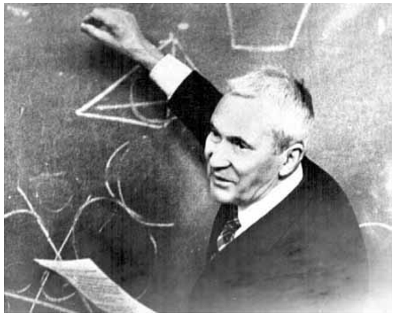
- Approximations used to extract theoretical prediction
- Modeling of apparatus

Quantify the uncertainty using PROBABILITY

A definition of probability

Consider a set *S* (the sample space)

Interpretation of elements left open, *S* could be e.g. set of outcomes of a repeatable observation.



Kolmogorov (1933)

Label subsets of *S* as *A*, *B*, ...

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

$$P(S) = 1$$

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

Properties of Probability

From the axioms of probability, further properties can be derived, e.g.,

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

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

$$P(\emptyset) = 0$$

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

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Conditional probability

Start with sample space S (e.g., set of outcomes), then restrict to a subset B (with $P(B) \neq 0$).

Define conditional probability of A given B (~4th axiom):

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

E.g. rolling die, outcome $n = 1,2,\ldots,6$:

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

Independence

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

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

I.e. if A, B, independent, imposing one has no effect on the probability of the other.

N.B. do not confuse with disjoint subsets, i.e., $A \cap B = \emptyset$

E.g. dice: $A = n$ even, $B = n$ odd, $A \cap B = \emptyset$

$P(A) = ½$

$P(A|B) = 0$

Requiring *B* affects probability of *A*, so *A*, *B* not independent.

Statistical Data Analysis
Lecture 1-3

- Interpretation of probability
- Bayes' theorem
- Law of total probability

Interpretation of Probability

I. Relative frequency (→ "frequentist statistics")

A, *B*, ... are outcomes of a repeatable experiment

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

II. Subjective probability (→ "Bayesian statistics")

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 magnetic monopoles exist,...

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

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

Bayes' theorem

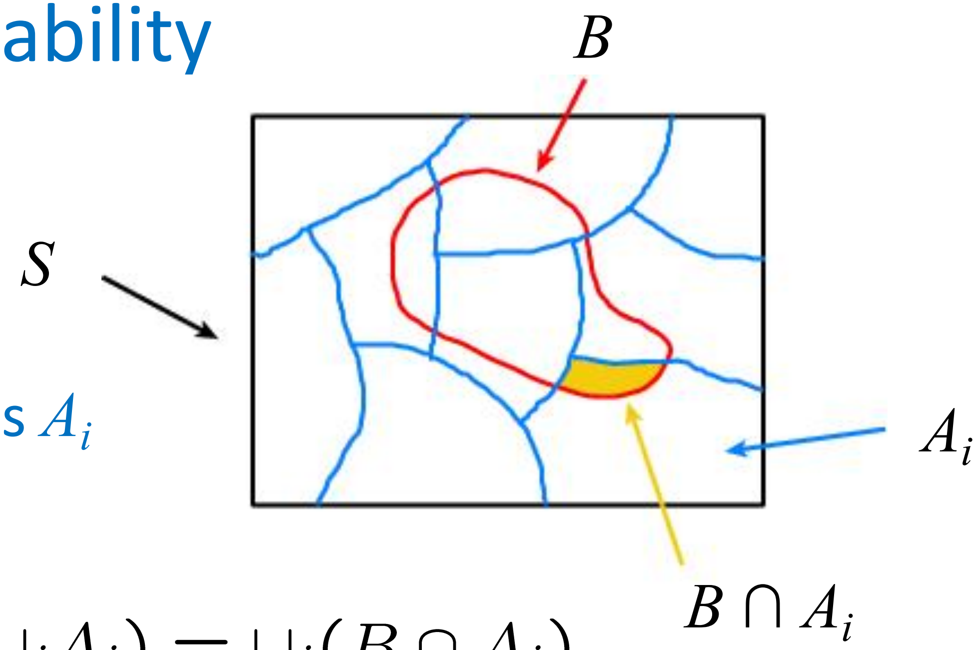
First published (posthumously) by the Reverend Thomas Bayes (1702−1761)

An essay towards solving a problem in the doctrine of chances, Philos. Trans. R. Soc. **53** (1763) 370

The law of total probability

Consider a subset B of the sample space S,

divided into disjoint subsets A_i such that $\cup_i A_i = S$,



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

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

$$\rightarrow \quad P(B) = \textstyle\sum_i P(B|A_i) P(A_i)$$

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:

$$P(\mathrm{D}) = 0.001$$
$$P(\text{no D}) = 0.999$$

← prior probabilities, i.e., before any test carried out

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

$$P(+|\mathrm{D}) = 0.98$$
$$P(-|\mathrm{D}) = 0.02$$

← probabilities to (in)correctly identify a person with the disease

$$P(+|\text{no D}) = 0.03$$
$$P(-|\text{no D}) = 0.97$$

← probabilities to (in)correctly identify a healthy person

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

Bayes' theorem example (cont.)

The probability to have the disease given a + result is

$$p(\mathrm{D}|+) = \frac{P(+|\mathrm{D})P(\mathrm{D})}{P(+|\mathrm{D})P(\mathrm{D}) + P(+|\mathrm{no\ D})P(\mathrm{no\ D})}$$

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

$$= 0.032$$ ← posterior probability

i.e. you're probably OK!

Your viewpoint: my degree of belief that I have 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: $\boldsymbol{x}$).

Probability = limiting frequency

Probabilities such as

P (string theory is true),
P ($0.117 < \alpha_s < 0.119$),
P (Biden wins in 2020),

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.

Preferred theories (models, hypotheses, ...) are those that predict a high probability for data "like" the data observed.

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 general prescription for priors (subjective!)

Statistical Data Analysis
Lecture 1-4

- Random variables
- Probability (density) functions:
  - joint pdf
  - marginal pdf
  - conditional pdf

Random variables and probability density functions

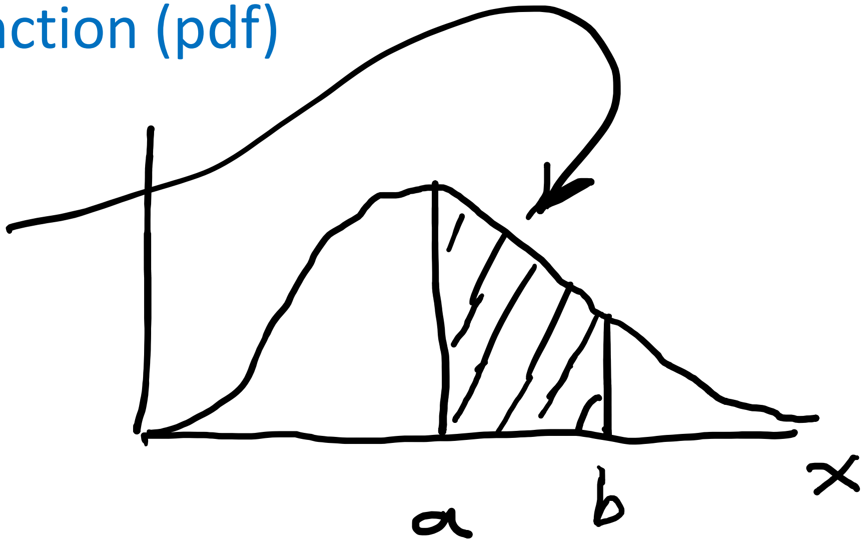
A random variable is a numerical characteristic assigned to an element of the sample space; can be discrete or continuous.

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)

$$\int_a^b f(x)\,dx = P(a \le x \le b)$$



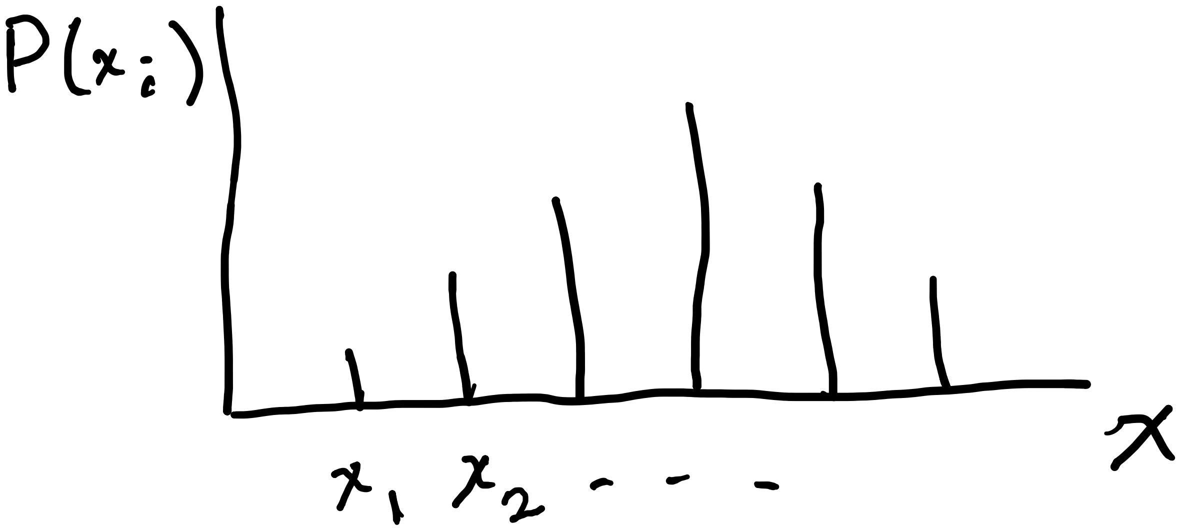
$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$ x must be somewhere

Probability mass function

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

$$P(x_i) = p_i$$ probability (mass) function

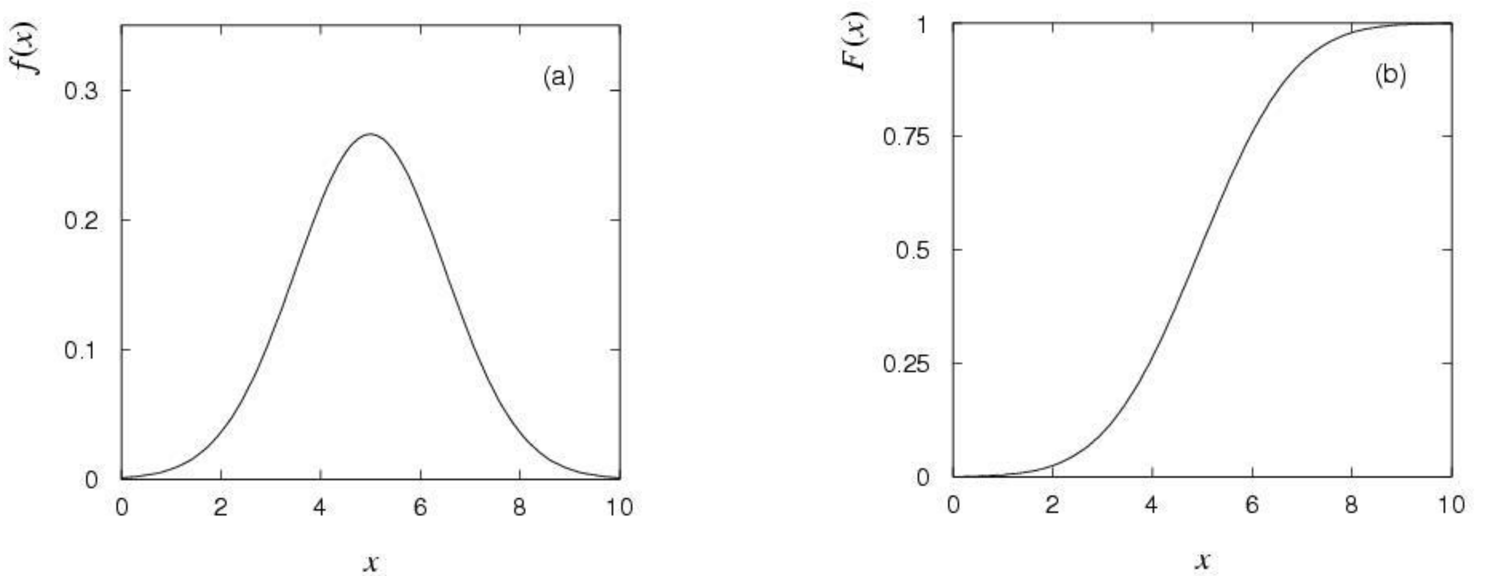
$$\sum_i P(x_i) = 1$$ x must take on one of its possible values



Cumulative distribution function

Probability to have outcome less than or equal to x is

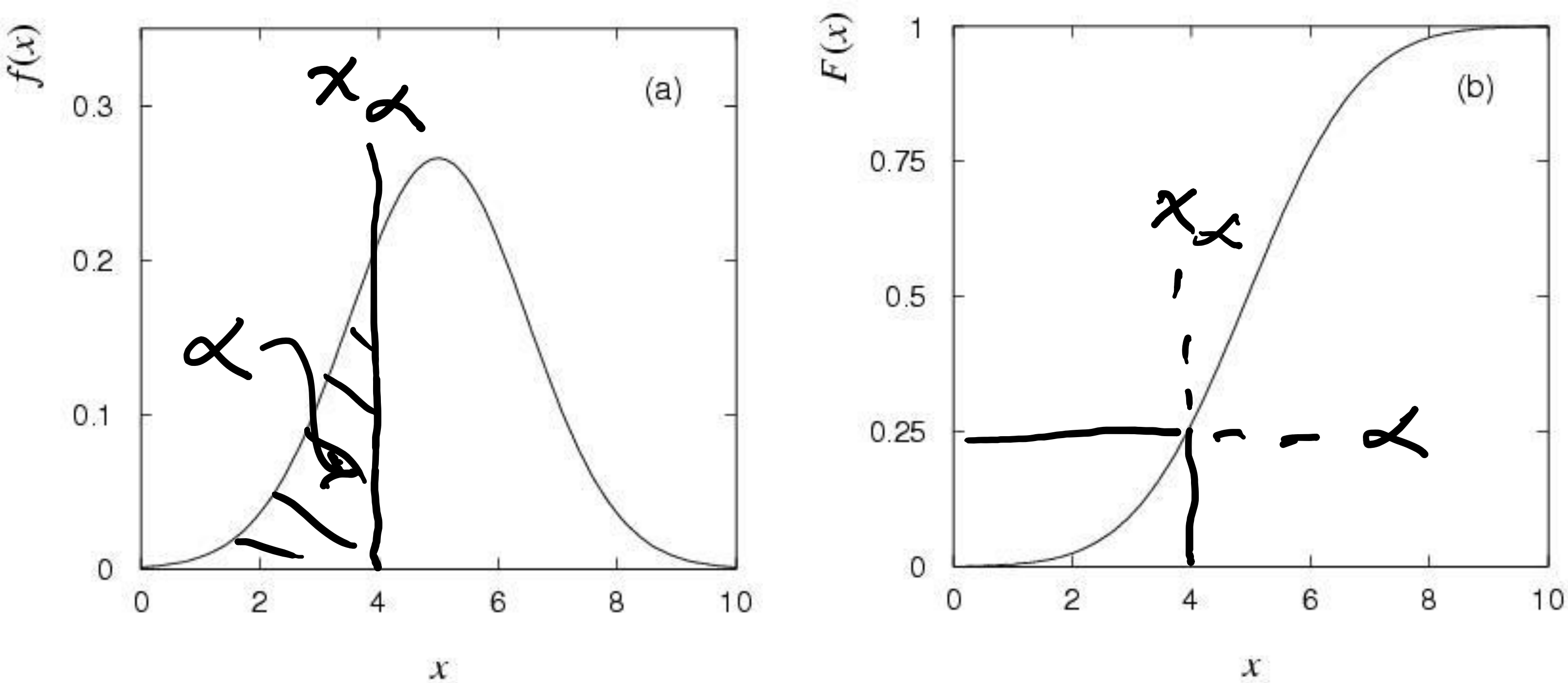
$$\int_{-\infty}^{x} f(x')\, dx' \equiv F(x)$$ cumulative distribution function



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

Quantiles

Define quantile (α-point) x_α by $F(x_\alpha) = \alpha \ (0 \leq \alpha \leq 1)$



i.e., quantile x_α is inverse of cumulative distribution: $x_\alpha = F^{-1}(\alpha)$,

Special case of quantile: $x_{1/2}$ = median

(compare to peak of pdf = mode)

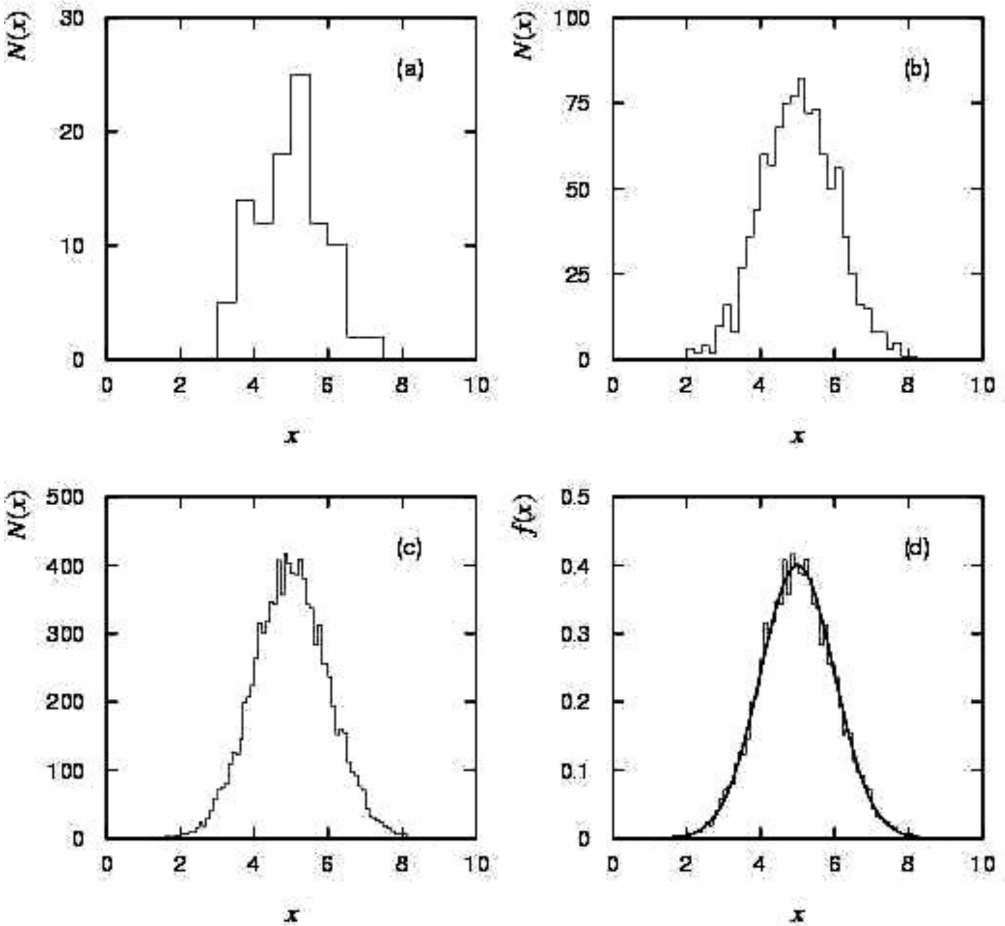
Histograms

Data sample $\boldsymbol{x} = (x_1,\ldots, x_n)$,
\# events n could be very large.

$\rightarrow$ Histogram $\boldsymbol{N} = (N_1,\ldots, N_M)$
M bins, bin size Δx.

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

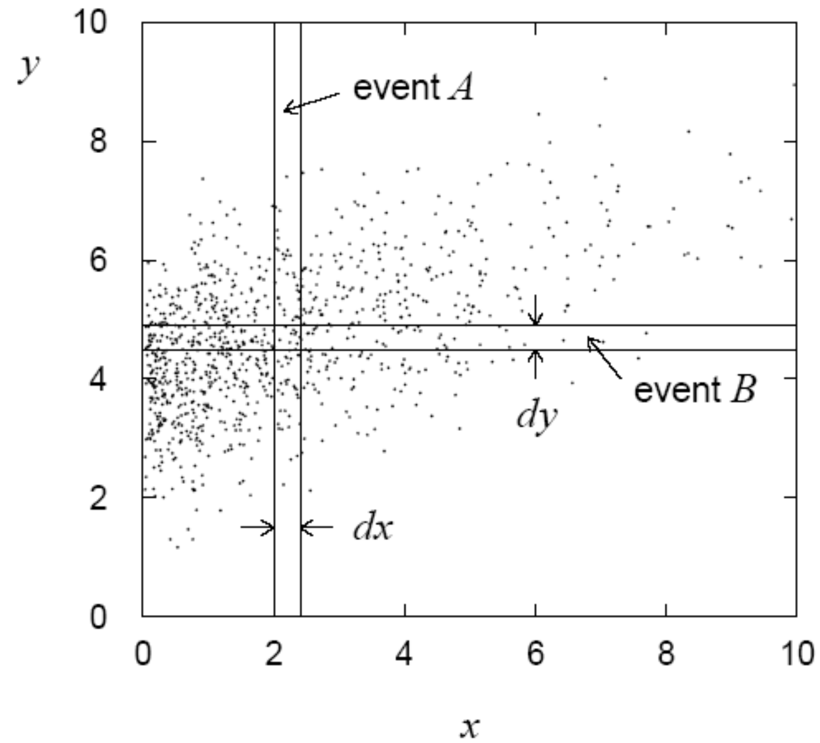
$$f(x) = \frac{N(x)}{n\Delta x}$$



Often normalize histogram to unit area, compare directly to pdf.

Multivariate distributions, joint pdf

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



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

joint pdf

$$\int \cdots \int_R f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n = P(\mathbf{x} \in R)$$

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

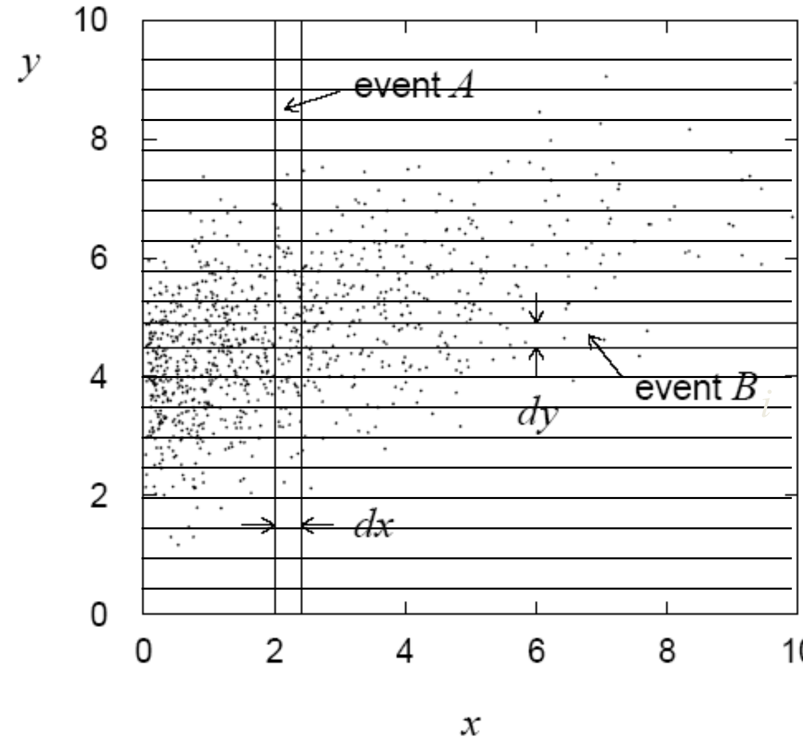
Marginal pdf

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

$$P(A) = \sum_i P(A \cap B_i)$$

$$= \sum_i f(x, y_i)\, dy\, dx$$

$$\rightarrow \int f(x, y)\, dy\, dx$$

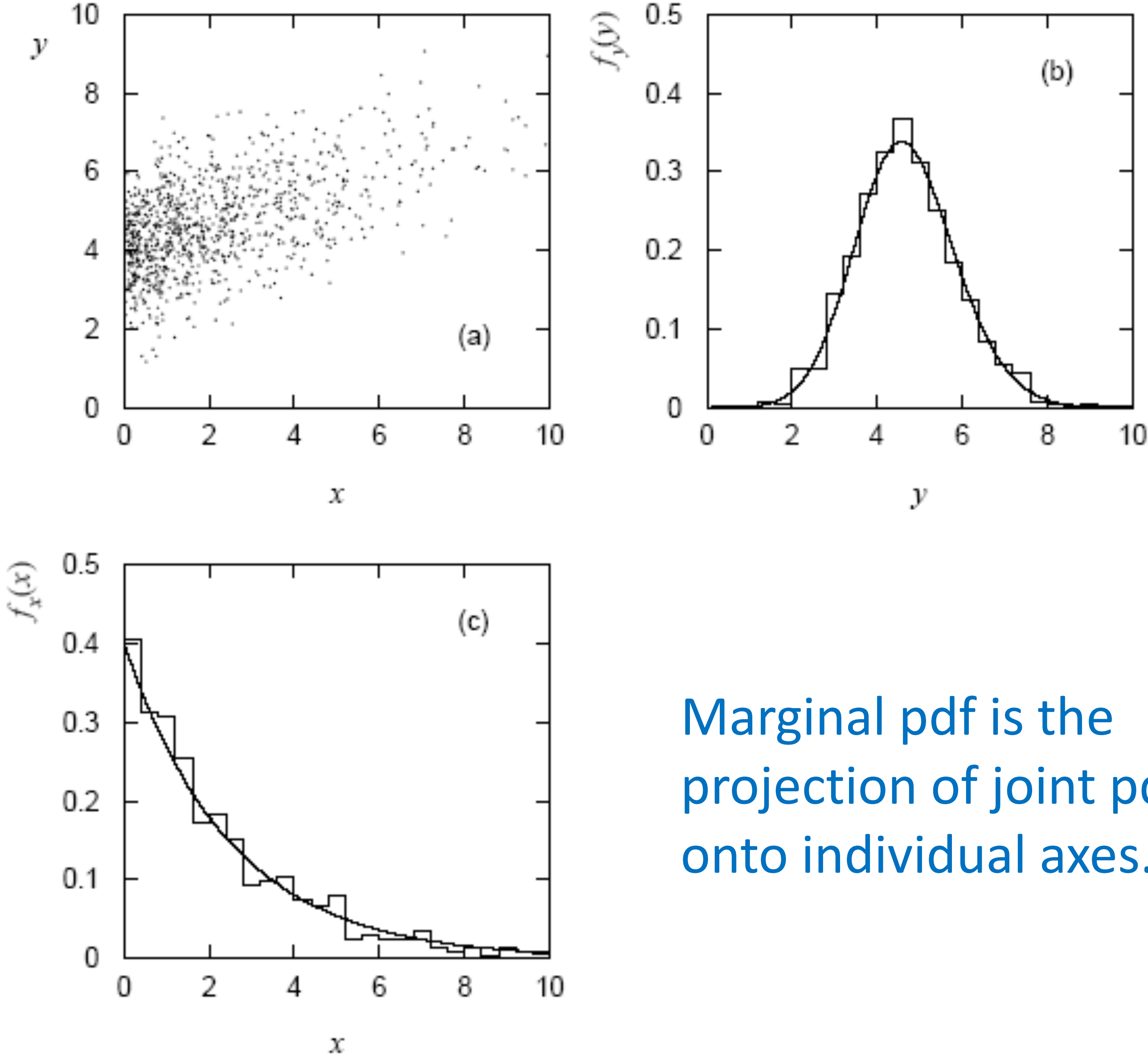


$$f_x(x) = \int f(x, y)\, dy$$ marginal pdf

E.g. to find marginal pdf of x_1 from n-dim. joint pdf, integrate over all variables except x_1 :

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

Marginal pdf (2)



Marginal pdf is the 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: $$h(y|x) = \frac{f(x,y)}{f_x(x)}$$

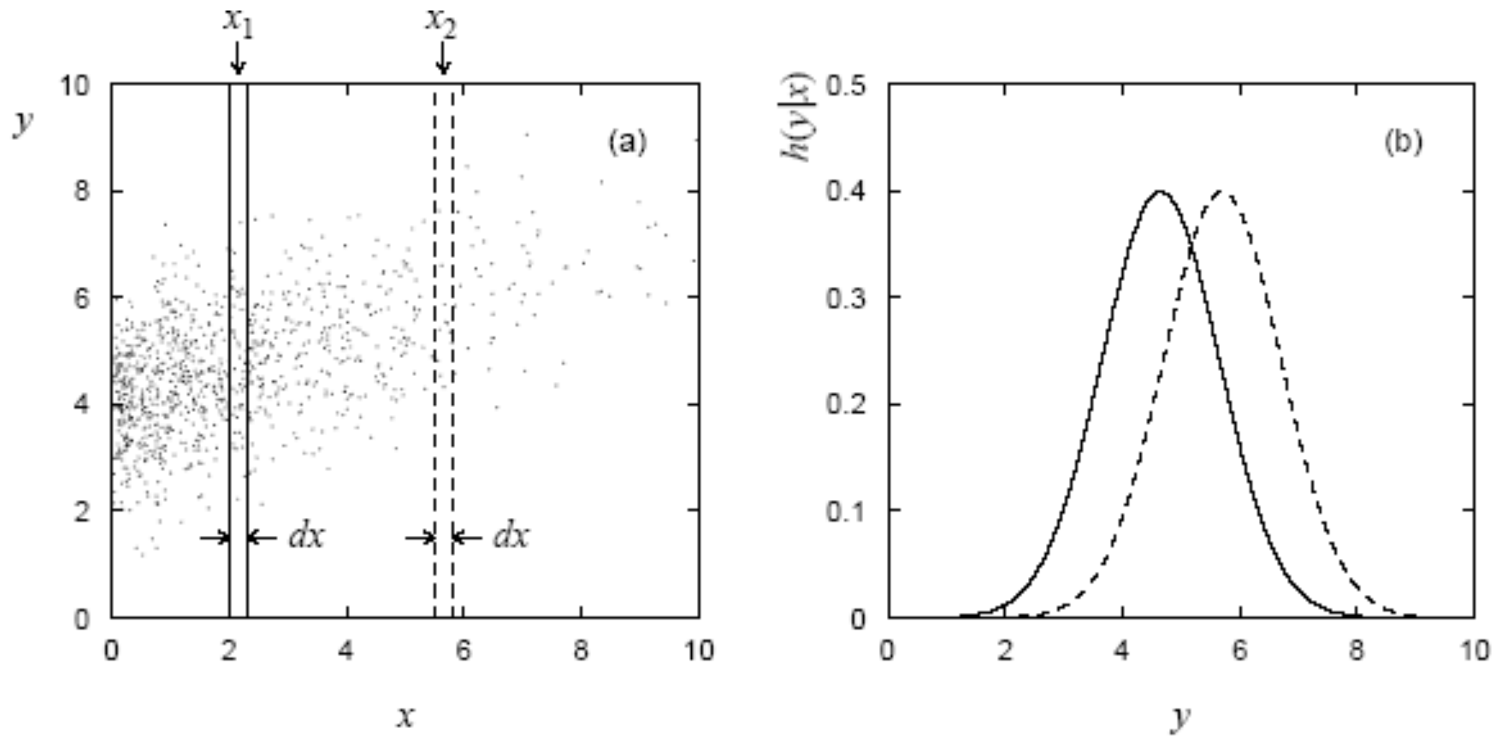
$$g(x|y) = \frac{f(x,y)}{f_y(y)}$$

E.g. $h(y|x)$ is a pdf for y, here x is fixed.

The denominator fixes normalization so that $\int h(y|x)\,dy = 1$

Conditional pdf (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., $\int h(y|x)\,dy = 1$.

Bayes' theorem, independence for conditional pdf

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 $f(x,y) = f_x(x) f_y(y)$.

Then e.g. fixing y has no effect on pdf of x:

$$g(x|y) = \frac{f_x(x) f_y(y)}{f_y(y)} = f_x(x)$$