Computing and Statistical Data Analysis

2005/06 University of London Lectures

PH4515 and HEP PhD students

Glen Cowan
Physics Department
Royal Holloway, University of London
(01784) 44 3452
`g.cowan@rhul.ac.uk`
`http://www.pp.rhul.ac.uk/~cowan`

- Course web page:

  `http://www.pp.rhul.ac.uk/~cowan/stat_course`

- Tentative schedule for 2005:

  Mostly Mondays 12:00 to 13:00 and 14:00 to 15:00
  (with a few exceptions to be announced).

Course aims

→ Understand role of uncertainty and probability in relating experiment and theory.

→ Understand statistical tools needed for analysis of experimental data.

→ Practice using statistics on the computer.

→ Learn computing tools for High Energy Physics.

Books

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

see also `alephwww.cern.ch/~cowan/stat`

R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods 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

W. Eadie et al., *Statistical Methods in Experimental Physics*, North-Holland, 1971

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

comes with FORTRAN and C program library on CD

S. Eidelman et al., Physics Letters B592, 1 (2004); see also `pdg.lbl.gov`.

sections on probability, statistics, Monte Carlo

Exercises (almost every week)

Tools (flexible):

- C++
- ROOT, MINUIT, etc.
- gnuplot?
- other (???)
- non-computer exercises

Half-day tutorial/workshop for HEP PhD students

At a central venue, date to be decided

Assessment

for PhD students: exercises (100%)

for MSc/MSci students: exercises and written exam

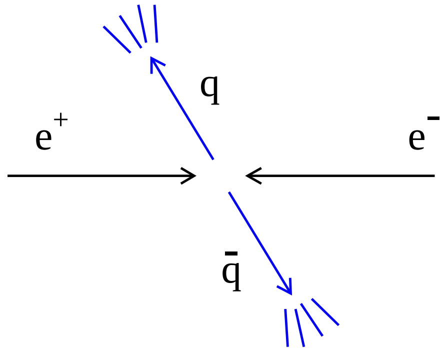
Statistical Data Analysis Course Outline

- Probability. Definition and interpretation, Bayes' theorem, random variables, probability density functions, expectation values, transformation of variables, error propagation.
- Examples of probability functions. Binomial, multinomial, Poisson, uniform, exponential, Gaussian, chi-square, Cauchy distributions.
- The Monte Carlo method. Random number generators, the transformation method, the acceptance-rejection method.
- Statistical tests. Significance and power of a test, choice of the critical region. Constructing test statistics: the Fisher discriminant, neural networks. Testing goodness-of-fit, χ^2-test, P-values.
- Parameter estimation: general concepts. Samples, estimators, bias. Estimators for mean, variance, covariance.
- The method of maximum likelihood. The likelihood function, ML estimators for parameters of Gaussian and exponential distributions. Variance of ML estimators, the information inequality, extended ML, ML with binned data.
- The method of least squares. Relation to maximum likelihood, linear least squares fit, LS with binned data, testing goodness-of-fit, combining measurements with least squares.
- Interval estimation. Classical confidence intervals: with Gaussian distributed estimator, for mean of Poisson variable. Setting limits, limits near a physical boundary.
- Unfolding. Formulation of the problem: response function and matrix. Inversion of the response matrix, bin-by-bin correction factors. Regularized unfolding: regularization functions, bias and variance of estimators, choice of regularization parameter.

Lecture 1 outline

1. **Probability**
   (a) definition
   (b) interpretation
   (c) Bayes' theorem
2. **Random variables**
   (a) probability densities and derived quantities

Data analysis in particle physics



Observe n events of a certain type

Measure characteristics of each event (angles, event shapes particle multiplicity, number found for a given $\int L dt, \ldots$)

Theories (e.g. SM) predict distributions of these properties up to free parameters, e.g. α, $G_{\rm F}$, $M_{\rm Z}$, $\alpha_{\rm s}$, $m_{\rm H}, \ldots$

Some tasks of statistical data analysis:

Estimate the parameters.

Quantify the uncertainty of the parameter estimates.

Test to what extent the predictions of a theory are in agreement with the data.

There are various elements of uncertainty :

theory is not deterministic,

random measurement errors,

things we could know in principle but don't,...

$\rightarrow$ quantify using PROBABILITY

Definition of probability

Consider a set S with subsets $A, B, \ldots$

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

$P(S) = 1$

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

Kolmogorov axioms (1933)

From these axioms one 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 $P(A) \leq P(B)$

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

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

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

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

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

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

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

(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

→ Data analysis in HEP: frequency interperation often most natural, but subjective probability has some attractive features, e.g. more natural treatment of phenomena that are not repeatable:

Systematic errors (same upon repetition of experiment)

The particle in this event was a positron

Nature is supersymmetric

Billionth digit of π is 7

It will rain tomorrow (uncertain future event)

It rained in Cairo on March 8, 1587 (uncertain past event)

Frequentist vs. subjective probability

What is $P(0.118 \leq \alpha_s \leq 0.122)$?

Frequentist: 0 or 1 (but I don't know which)

Subjectivist (Bayesian): 68% (statement of knowledge)

i.e. $P(0.118 \leq \alpha_s \leq 0.122) = 0.68$ (subjective) means:

my uncertainty that $0.118 \leq \alpha_s \leq 0.122$ is same as uncertainty to draw white ball out of container of 100 balls, 68 of which are white. (cf. G. D'Agostini, CERN Yellow Report 99-03, July 1999)

→ Calibration by relation to frequency (or symmetry, betting, etc.)

If a large group of Bayesians say things like:

P(Brazil will win 2002 World Cup) = 68%

$P(0.118 \leq \alpha_s \leq 0.122) = 68\%$

P(Al Gore president in 2001) = 68%

then 68% of these statements should wind up being true.

N.B. Calibration not always feasible, e.g.

P(Ivanov will win chess tournament in Tomsk in 2017) = ???

Attempt to rescue freqency: can $P(0.118 \leq \alpha_s \leq 0.122) = 68\%$ mean,

Consider an ensemble of universes in which Nature assigns different values of α_s; 68% of these will have α_s in $[0.118, 0.122]$ (???)

Fine . . . but this is just a way of phrasing degree of belief.

Bayes' theorem

From the definition of conditional probability,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)} \;,$$

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

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

Bayes' theorem

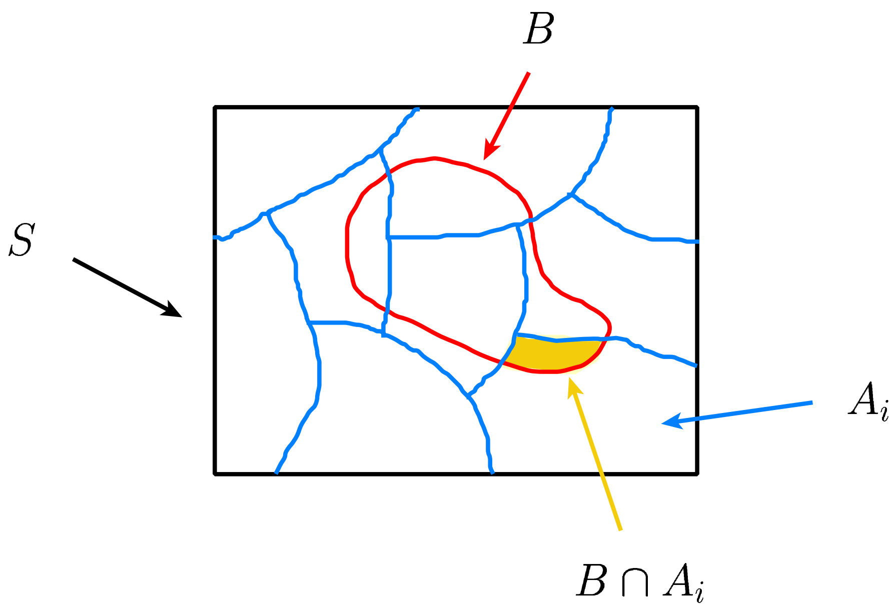
First published (posthumously) by the Reverend Thomas Bayes (1702–1761)

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

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 $\cup_i A_i = S$,

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

$\rightarrow P(B) = P(\cup_i (B \cap A_i)) = \Sigma_i P(B \cap A_i)$ (since $B \cap A_i$ disjoint)

$\rightarrow P(B) = \Sigma_i \, P(B|A_i) \, P(A_i)$ (law of total probability)

Bayes' theorem becomes

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

An example using Bayes' theorem

Suppose the probabilities (for anyone) to have AIDS are:

$P(\text{AIDS}) = 0.001$ ← prior probabilities, i.e.

$P(\text{no AIDS}) = 0.999$ before any test carried out

Consider an AIDS test: result is $+$ or $-$

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

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

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

$P(-|\text{no AIDS}) = 0.97$ identify person without AIDS

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

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

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

$$= 0.032$$ ← 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 guy will have AIDS

Random variables

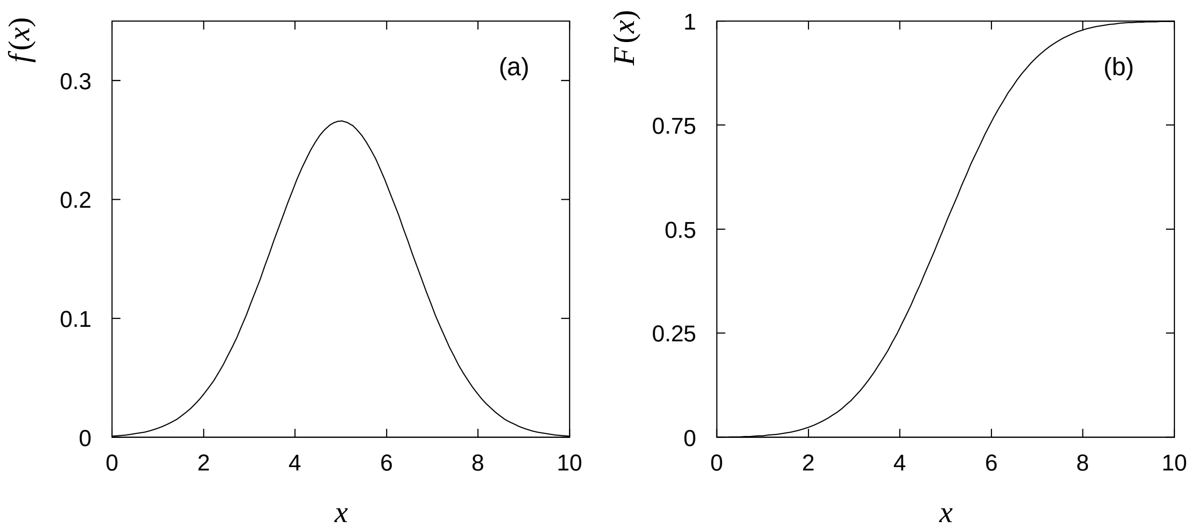
Suppose outcome of experiment is x (label for element of sample space)

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

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

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

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



For discrete case:

$$f_i = P(x_i)$$

$$\sum_i f_i = 1$$

$$F(x) = \sum_{x_i \le x} P(x_i)$$

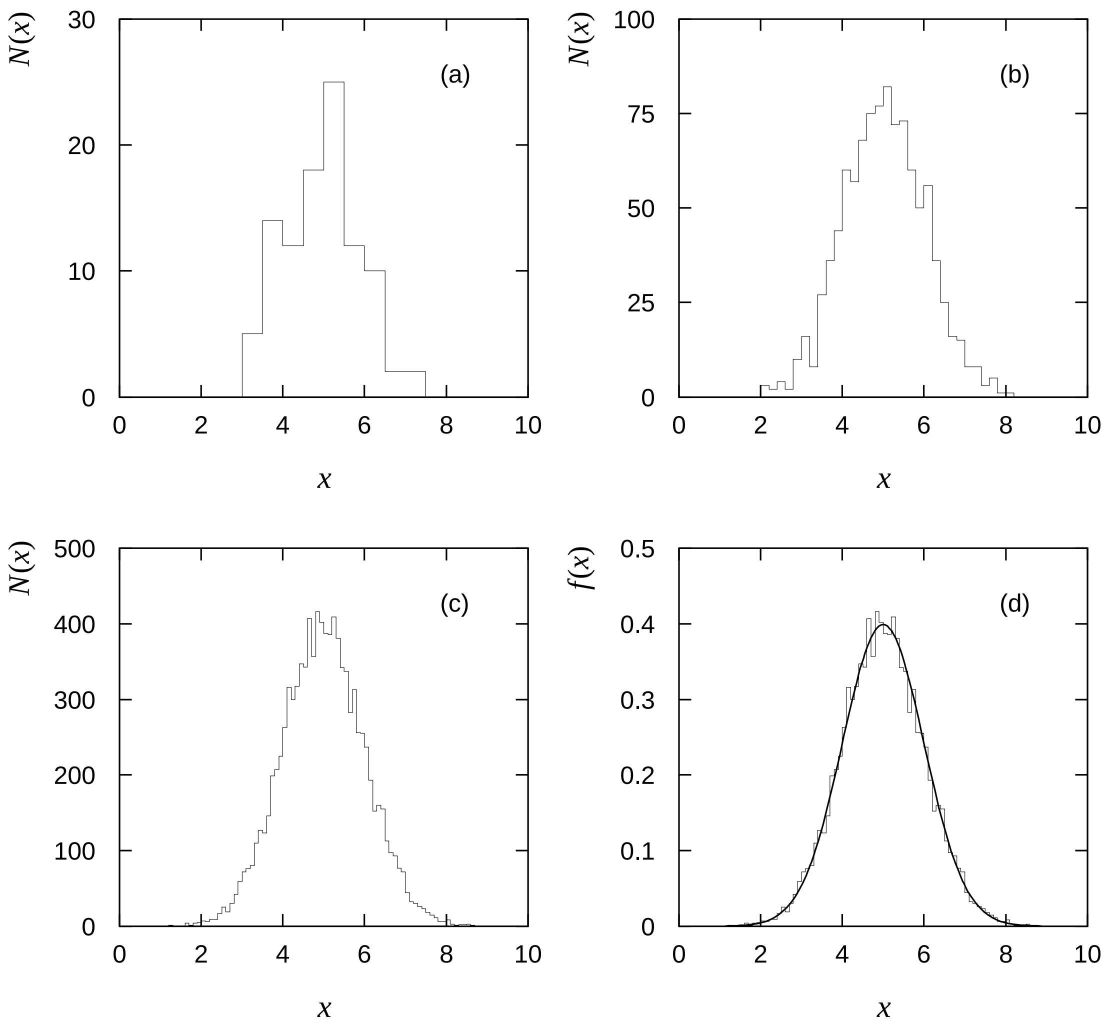
Histograms

pdf = histogram with:

infinite data sample

zero bin width

normalized to unit area



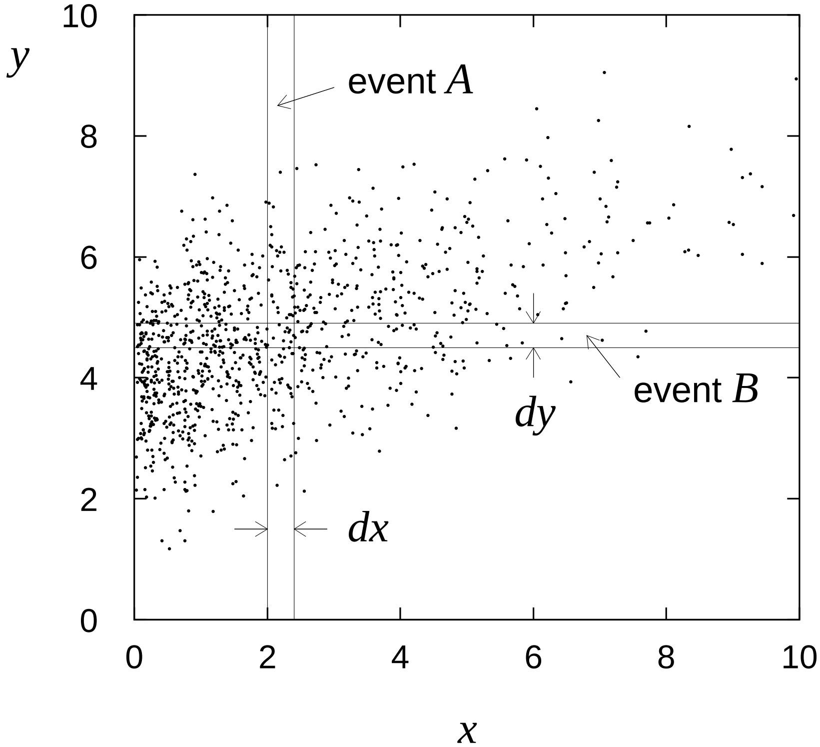
$$f(x) = \frac{N(x)}{n\Delta x}$$

n = number of entries

Δx = bin width

Multivariate case

Outcome characterized by > 1 quantity, e.g. x and y



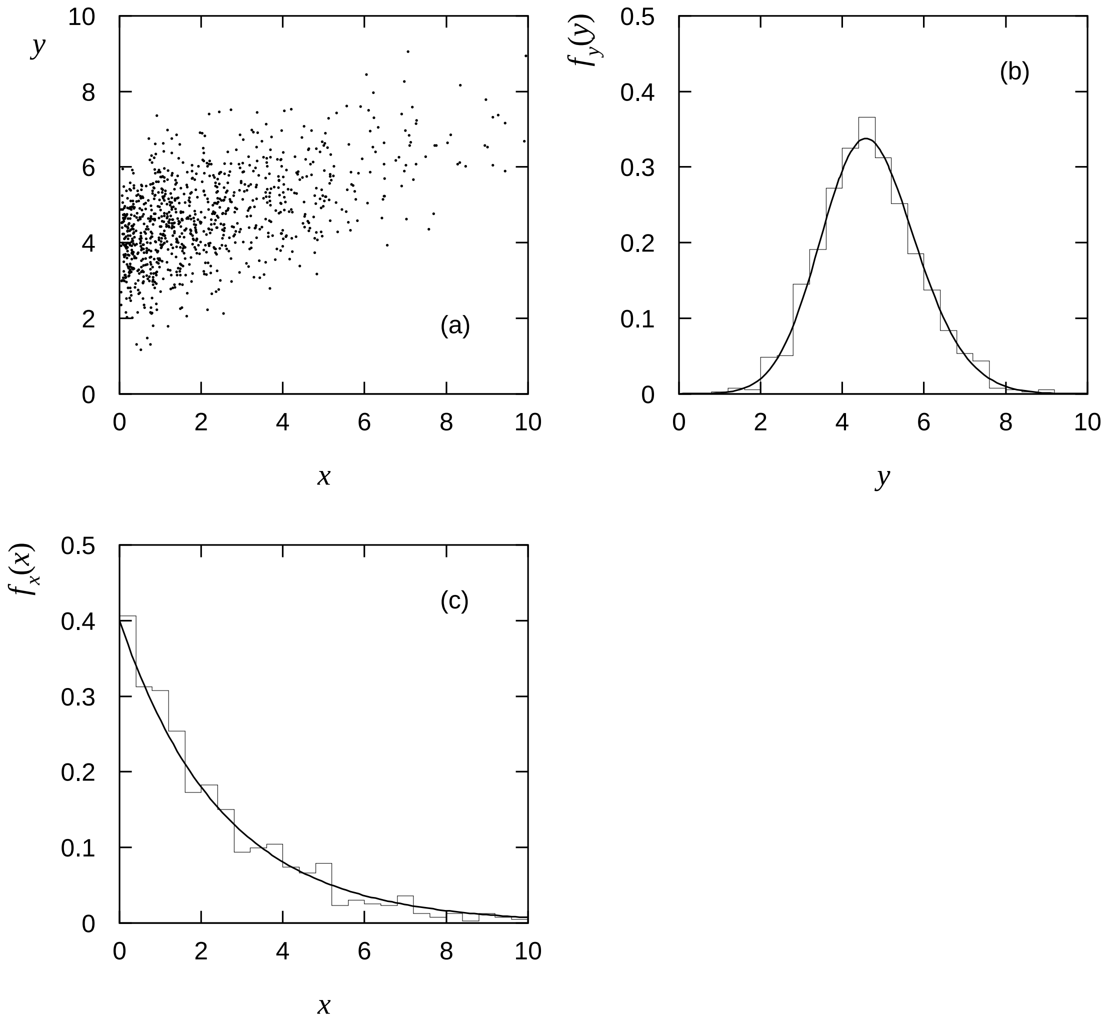
$$P(A \cap B) = f(x, y)\, dx\, dy$$

$$\rightarrow f(x, y) = \text{joint pdf}$$

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

Marginal distributions

Projections of joint pdf (scatter plot) onto x, y axes:



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

$$f_y(y) = \int f(x,y)\, dx$$

$\rightarrow$ $f_x(x),\ f_y(y)$ = marginal pdfs

Conditional pdf

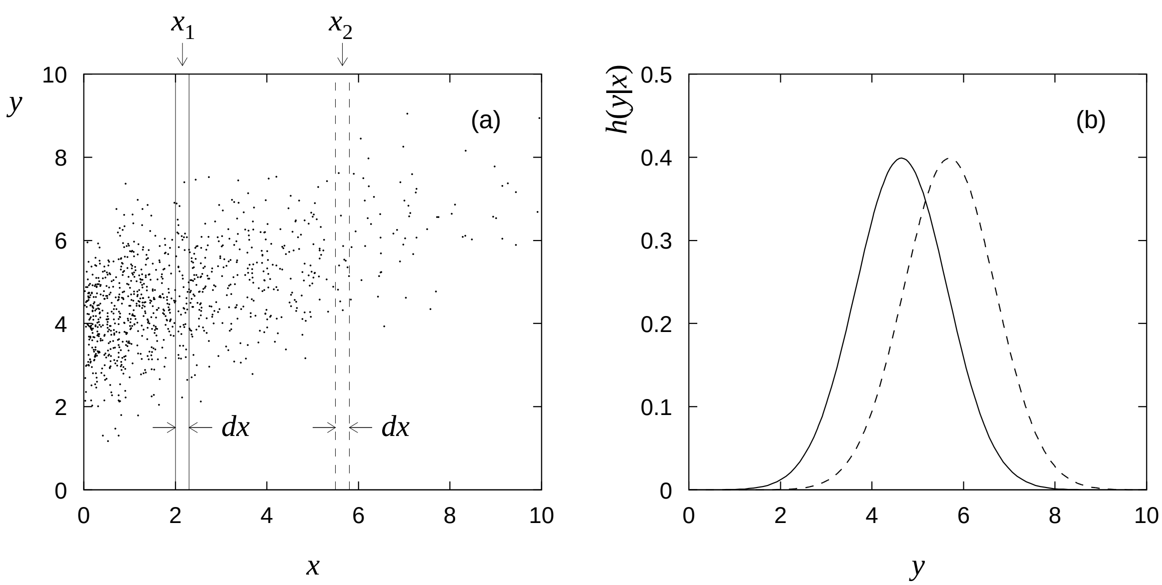
Recall conditional probability:

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

Define $h(y|x) = \dfrac{f(x,y)}{f_x(x)}$

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

← conditional pdfs



Bayes' theorem becomes

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

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

$\Rightarrow$ x, y independent if $f(x,y) = f_x(x) f_y(y)$

Lecture 1 summary

1. **Probability**
   (a) definition: Kolmogorov axioms + conditional probability
   (b) interpretation: frequency or degree of belief
   (c) Bayes' theorem

2. **Random variables**
   (a) probability density functions (pdf), e.g. $f(x)$
   (b) cumulative distribution functions, $F(x) = \int_{-\infty}^{x} f(x')\,dx'$
   (c) joint pdf, e.g. $f(x,y)$
   (d) marginal pdf, e.g. $f_x(x) = \int f(x,y)dy$
   (e) conditional pdf, e.g. $g(x|y) = f(x,y)/f_y(y)$