

Lectures on Statistical Data Analysis



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:

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

Statistical Data Analysis: Outline by Lecture

- 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

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

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*, Journal of Physics G 33 (2006) 1; see also pdg.lbl.gov sections on probability statistics, Monte Carlo

Dealing with uncertainty

We encounter uncertainty all the time.



We can quantify the uncertainty using **PROBABILITY**

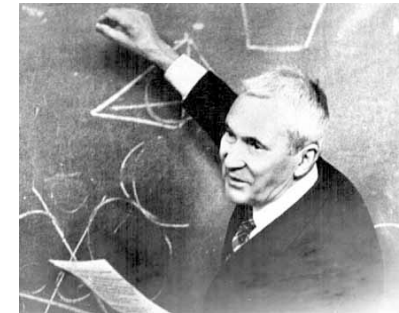
A definition of probability

Consider a set S with subsets A, B, \dots

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 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, \dots are outcomes of a repeatable experiment

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

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

II. Subjective probability

A, B, \dots 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

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

Bayes' theorem



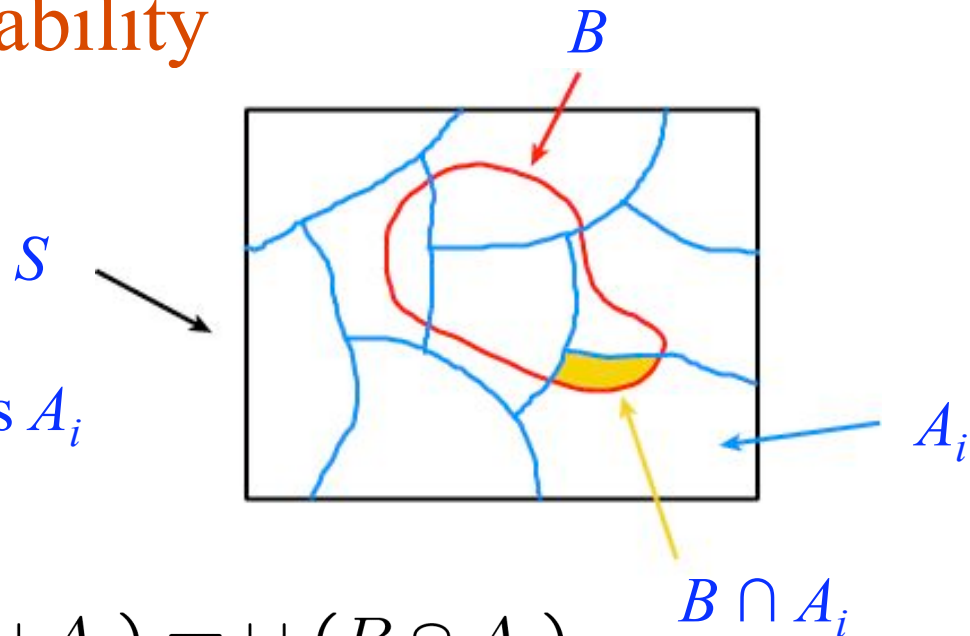
First published (posthumously) by the
Reverend Thomas Bayes (1702–1761)

*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 $\cup_i A_i = S$,



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

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

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

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 AIDS is:

$$P(\text{AIDS}) = 0.001$$

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

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

Consider an AIDS test: result is + or –

$$P(+|\text{AIDS}) = 0.98$$

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

← probabilities to (in)correctly
identify an infected person

$$P(+|\text{no AIDS}) = 0.03$$

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

← probabilities to (in)correctly
identify an uninfected person

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

Bayes' theorem example (cont.)

The probability to have AIDS given a + result is

$$\begin{aligned}P(\text{AIDS}|+) &= \frac{P(+|\text{AIDS})P(\text{AIDS})}{P(+|\text{AIDS})P(\text{AIDS}) + P(+|\text{no AIDS})P(\text{no AIDS})} \\&= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} \\&= 0.032 \quad \leftarrow \text{posterior probability}\end{aligned}$$

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),

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

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

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

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!)

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)

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

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

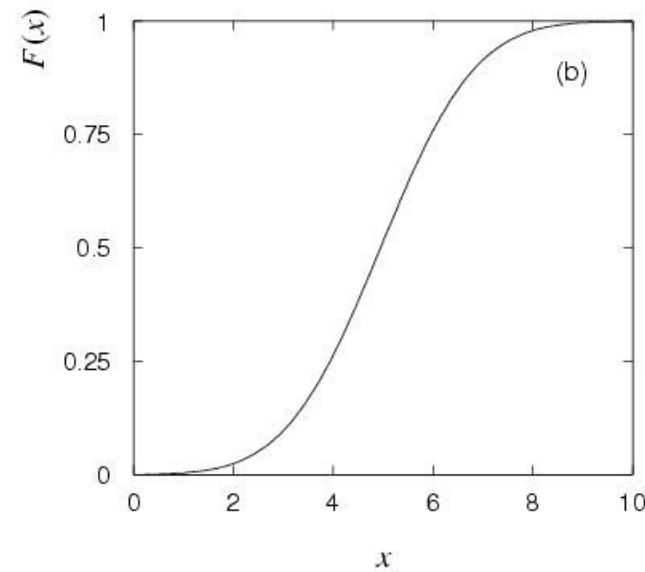
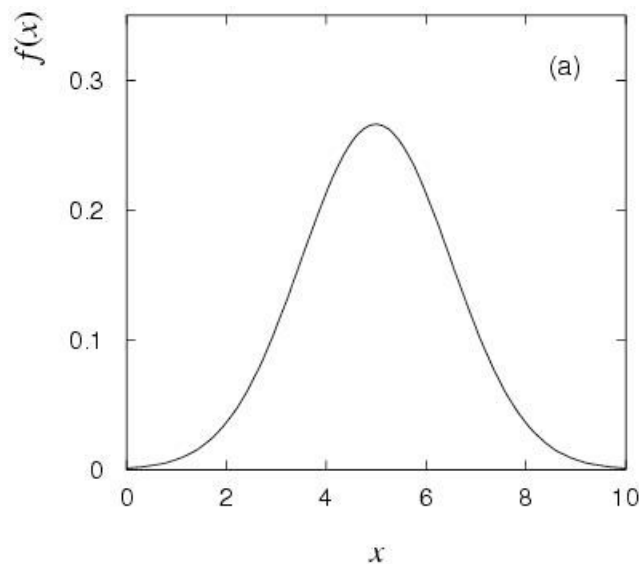
$$P(x_i) = p_i \quad \text{probability mass function}$$

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

Cumulative distribution function

Probability to have outcome less than or equal to x is

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



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

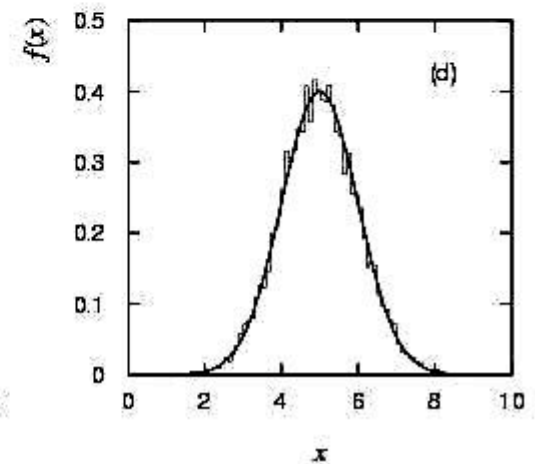
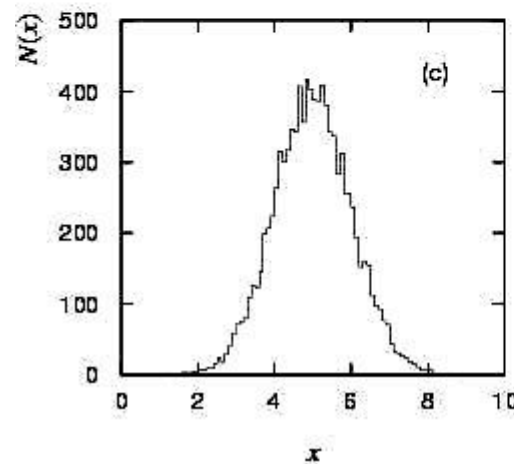
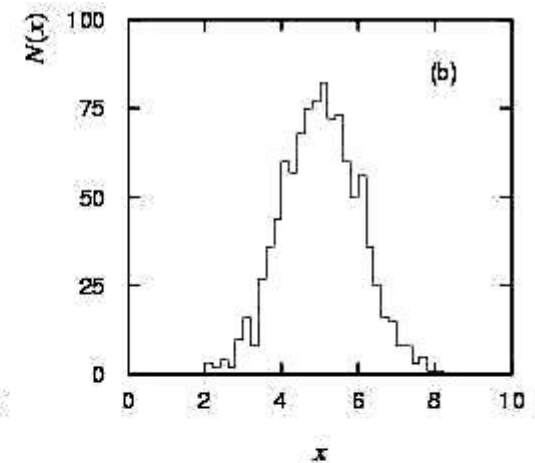
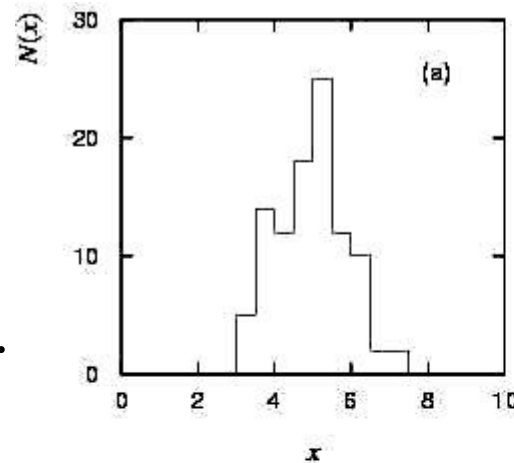
Histograms

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

Δx = bin width

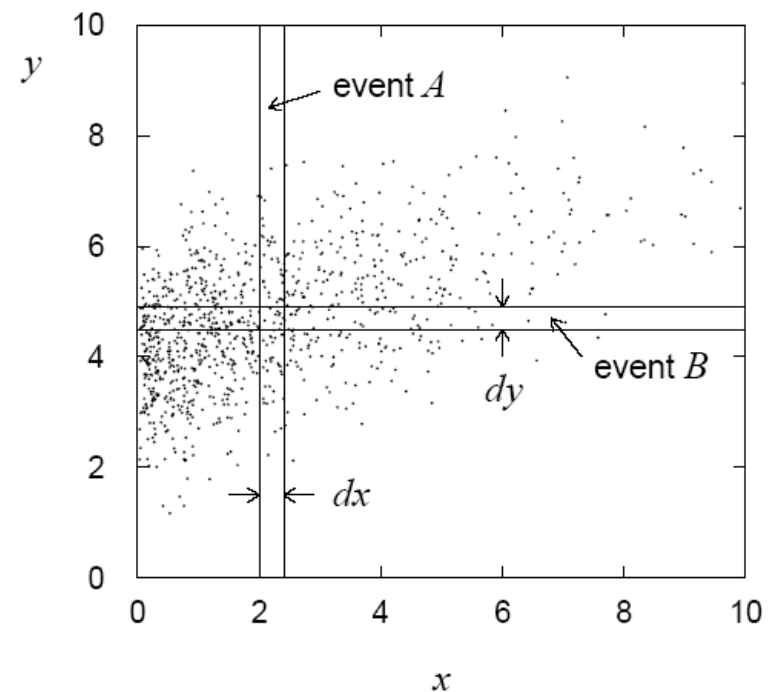


Multivariate distributions

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

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

joint pdf



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

Marginal pdf

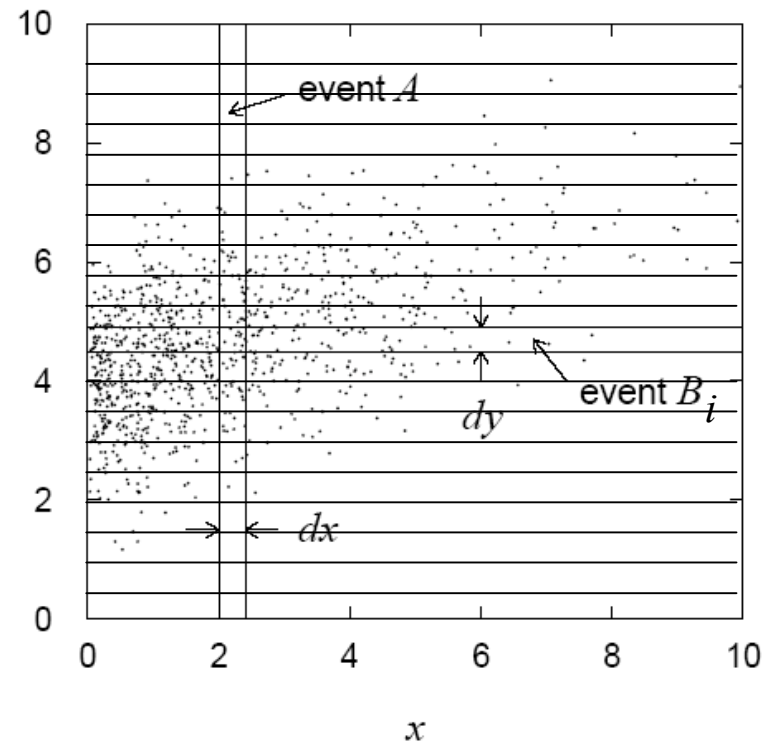
Sometimes we want only pdf of y some (or one) of the components:

$$\begin{aligned} P(A) &= \sum_i P(A \cap B_i) \\ &= \sum_i \int f(x, y_i) dy dx \\ &\rightarrow \int f(x, y) dy dx \end{aligned}$$

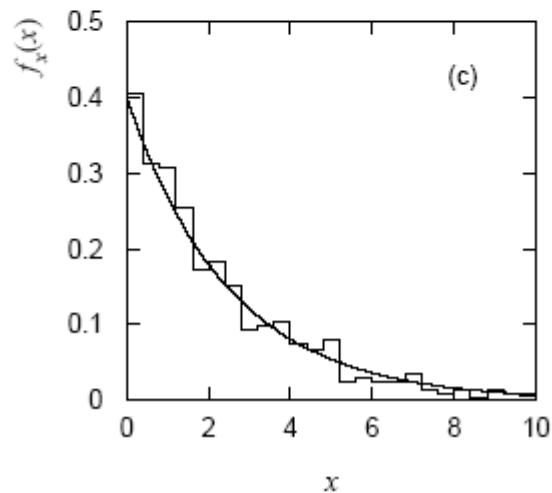
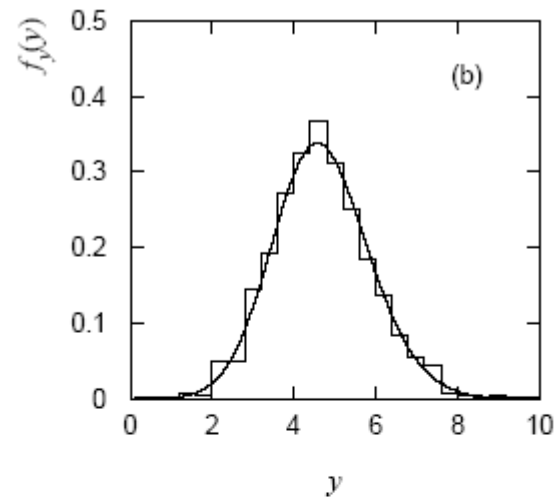
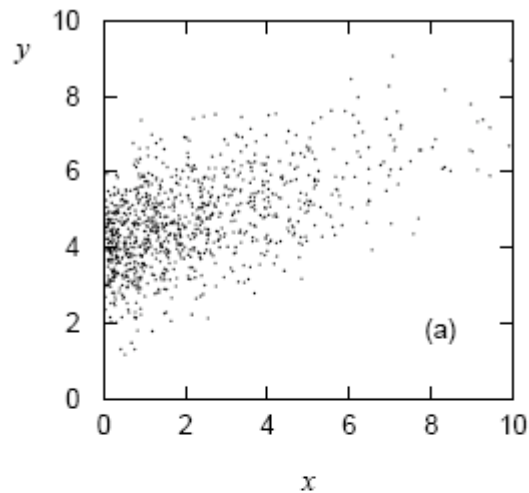
$$f_x(x) = \int f(x, y) dy$$

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

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



Marginal pdf (2)



Marginal pdf \sim
projection of joint pdf
onto individual axes.

Conditional pdf

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

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

→ conditional pdfs: $h(y|x) = \frac{f(x, y)}{f_x(x)}$, $g(x|y) = \frac{f(x, y)}{f_y(y)}$

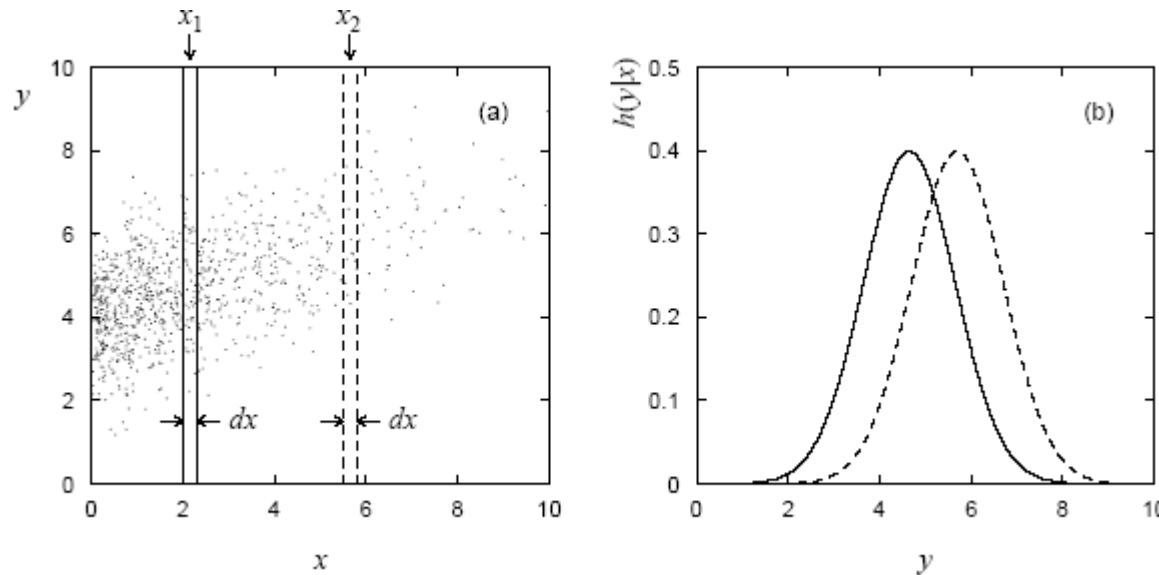
Bayes' theorem becomes: $g(x|y) = \frac{h(y|x)f_x(x)}{f_y(y)}$.

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 $h(y|x_1)$, $h(y|x_2)$:



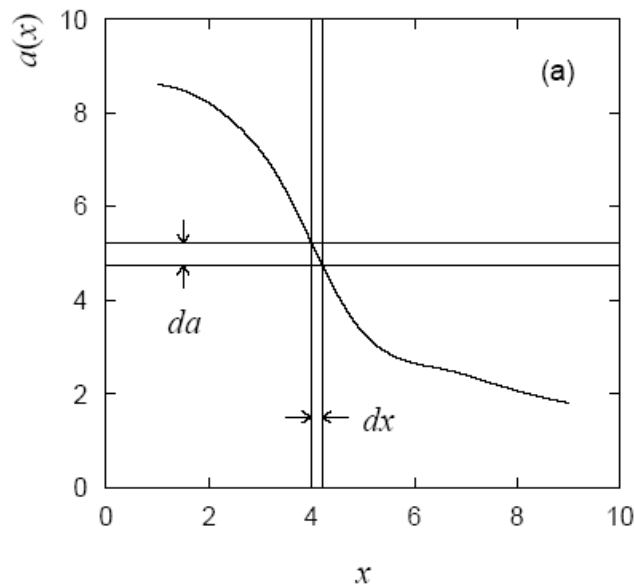
Basically treat some of the r.v.s as constant, then divide the joint pdf by the marginal pdf of those variables being held constant so that what is left has correct normalization, e.g., $\int h(y|x) dy = 1$.

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 $[a, a+da]$.

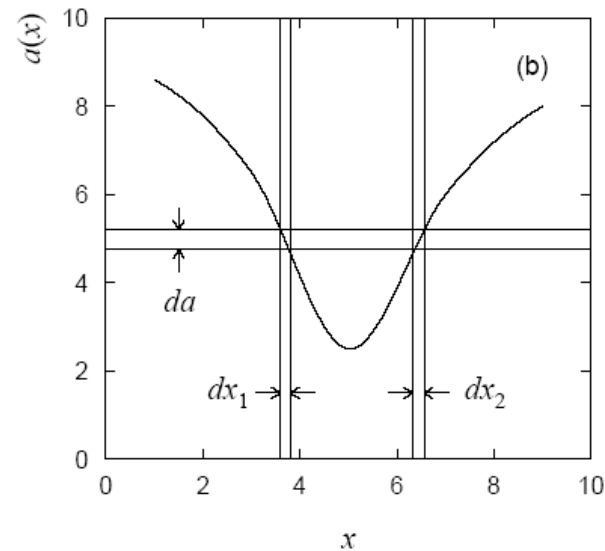
For one-variable case with unique inverse this is simply

$$g(a) da = f(x) dx$$

$$\rightarrow g(a) = f(x(a)) \left| \frac{dx}{da} \right|$$

Functions without unique inverse

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



Example: $a = x^2$, $x = \pm\sqrt{a}$, $dx = \pm\frac{da}{2\sqrt{a}}$.

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

$$g(a) = \frac{f(\sqrt{a})}{2\sqrt{a}} + \frac{f(-\sqrt{a})}{2\sqrt{a}}$$

Functions of more than one r.v.

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

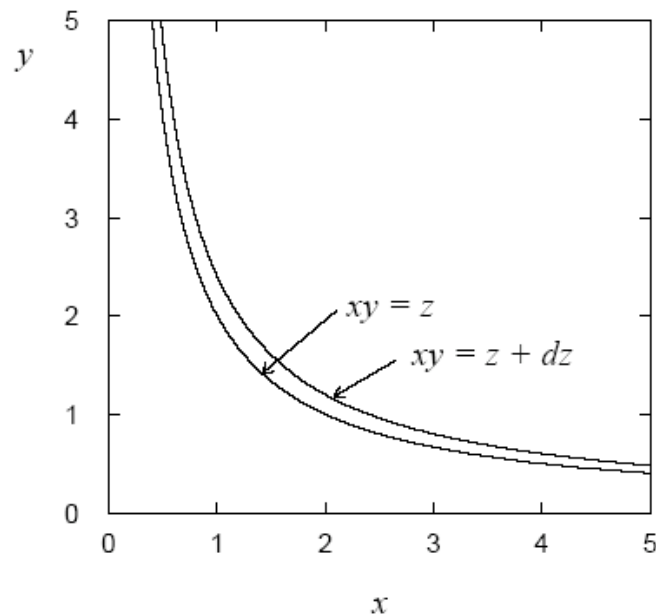
$$g(a')da' = \int \dots \int_{dS} f(x_1, \dots, x_n) dx_1 \dots dx_n$$

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 $g(z)$?



$$\begin{aligned} g(z) dz &= \int \dots \int_{dS} f(x, y) dx dy \\ &= \int_0^\infty dx \int_{z/x}^{(z+dz)/x} f(x, y) dy \\ \rightarrow g(z) &= \int_0^\infty f\left(x, \frac{z}{x}\right) \frac{dx}{x} \\ &= \int_0^\infty f\left(\frac{z}{y}, y\right) \frac{dy}{y} \end{aligned}$$

(Mellin convolution)

More on transformation of variables

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

Form n linearly independent functions $\vec{y}(\vec{x}) = (y_1(\vec{x}), \dots, y_n(\vec{x}))$

for which the inverse functions $x_1(\vec{y}), \dots, x_n(\vec{y})$ exist.

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

where J is the
Jacobian determinant:

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & & & \vdots \\ & & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

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

Wrapping up lecture 2

We are now familiar with:

random variables

probability density function (pdf)

cumulative distribution function (cdf)


joint pdf, marginal pdf, conditional pdf,...

And we know how to determine the pdf of a function of an r.v.

single variable, unique inverse: $g(a) = f(x(a)) \left| \frac{dx}{da} \right|$

also saw non-unique inverse and multivariate case.

Statistical Data Analysis: Lecture 3

- 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

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): $E[x] = \mu \sim$ “centre of gravity” of pdf.

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

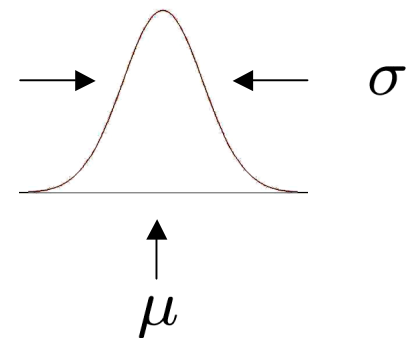
$$E[y] = \int y g(y) dy = \int y(x) f(x) dx \quad (\text{equivalent})$$

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

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

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

$\sigma \sim$ 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

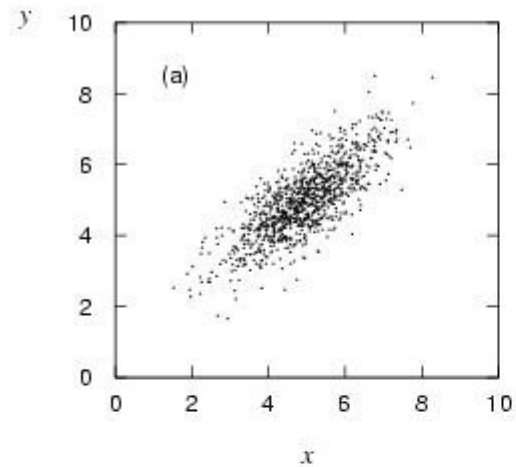
$$E[xy] = \int \int xy f(x, y) dx dy = \mu_x \mu_y$$

→ $\text{COV}[x, y] = 0$ x and y , ‘uncorrelated’

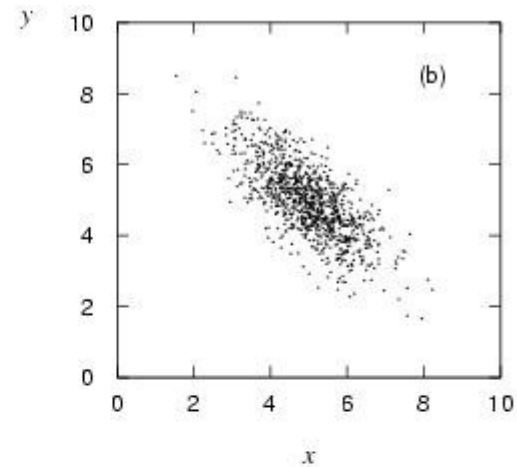
N.B. converse not always true.

Correlation (cont.)

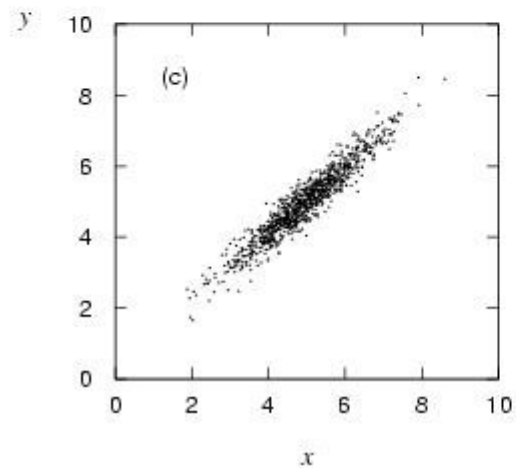
$$\rho = 0.75$$



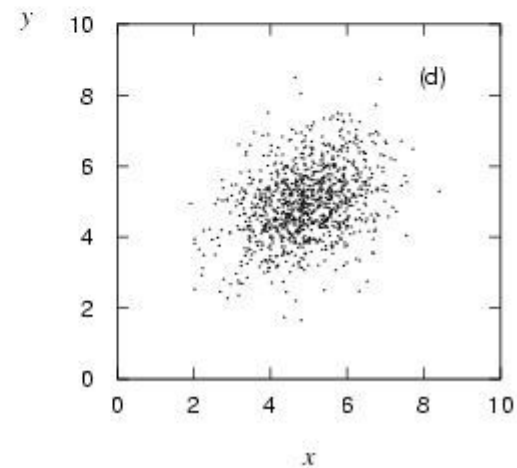
$$\rho = -0.75$$



$$\rho = 0.95$$



$$\rho = 0.25$$



Error propagation

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

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

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)

$$\begin{aligned} E[y^2(\vec{x})] &\approx y^2(\vec{\mu}) + 2y(\vec{\mu}) \sum_{i=1}^n \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} E[x_i - \mu_i] \\ &\quad + E \left[\left(\sum_{i=1}^n \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i) \right) \left(\sum_{j=1}^n \left[\frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j) \right) \right] \\ &= y^2(\vec{\mu}) + \sum_{i,j=1}^n \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij} \end{aligned}$$

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

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

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

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

or in matrix notation $U = A V A^T$, where

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

Error propagation (5)

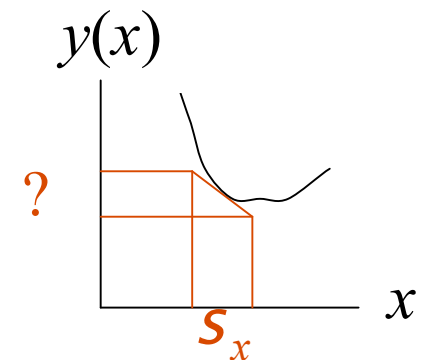
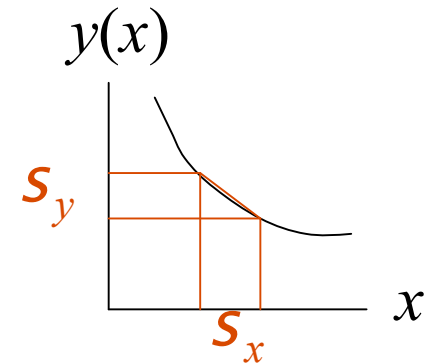
The ‘error propagation’ formulae tell us the covariances of a set of functions

$\vec{y}(\vec{x}) = (y_1(\vec{x}), \dots, y_m(\vec{x}))$ 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 s_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

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

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

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

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

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

Now suppose $r = 1$. Then

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

i.e. for 100% correlation, error in difference $\rightarrow 0$.

Wrapping up lecture 3


We know how to describe a pdf using

expectation values (mean, variance),
covariance, correlation, ...

Given a function of a random variable, we know how
to find the variance of the function using error propagation.

also for covariance matrix in multivariate case;
based on linear approximation.

Statistical Data Analysis: Lecture 4

- 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 distributions

<u>Distribution/pdf</u>	<u>Example use in HEP</u>
Binomial	Branching ratio
Multinomial	Histogram with fixed N
Poisson	Number of events found
Uniform	Monte Carlo method
Exponential	Decay time
Gaussian	Measurement error
Chi-square	Goodness-of-fit
Cauchy	Mass of resonance
Landau	Ionization energy loss

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

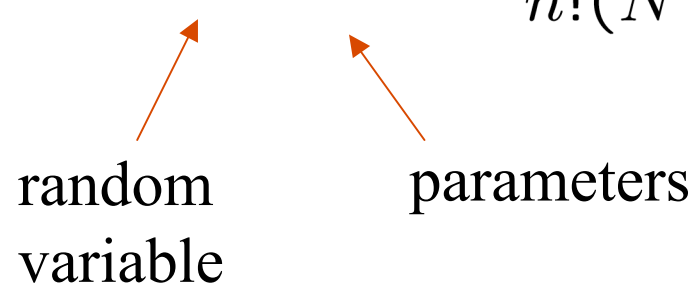
$$pp(1-p)p(1-p) = p^n(1-p)^{N-n}$$

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

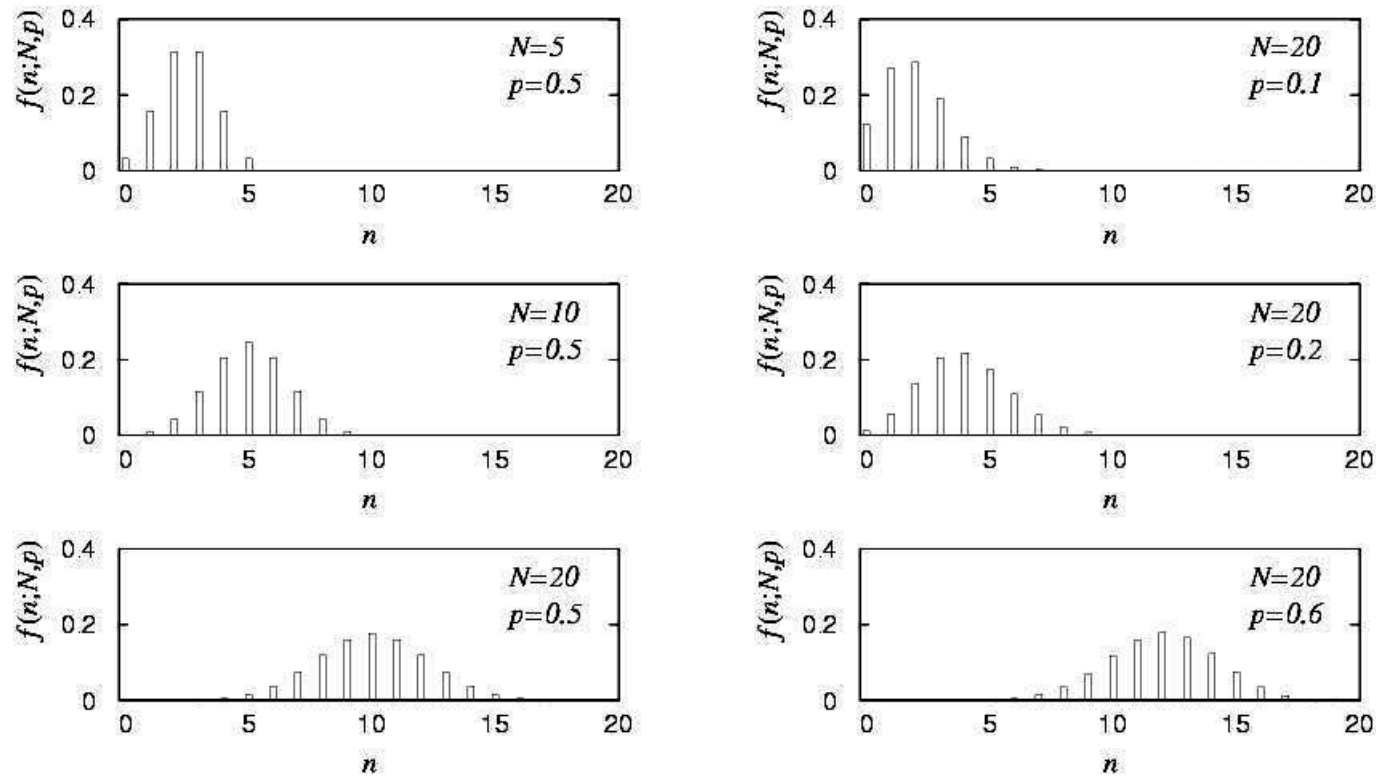
For the expectation value and variance we find:

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

$$V[n] = E[n^2] - (E[n])^2 = Np(1 - p)$$

Binomial distribution (3)

Binomial distribution for several values of the parameters:



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

Multinomial distribution

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

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

For N trials we want the probability to obtain:

n_1 of outcome 1,
 n_2 of outcome 2,
...
 n_m of outcome m .

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

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

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

Poisson distribution

Consider binomial n in the limit

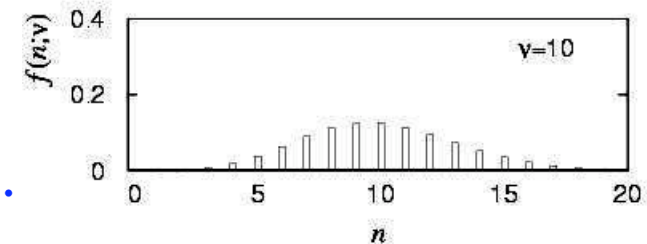
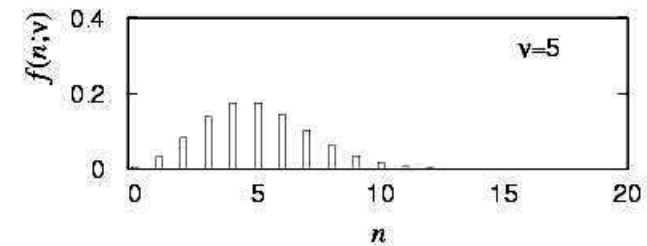
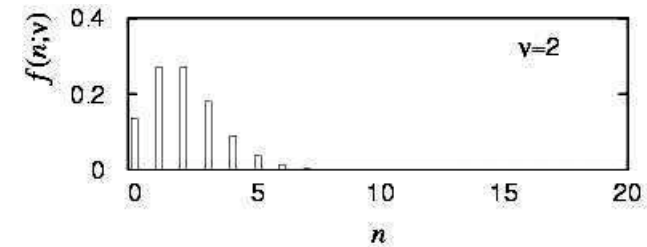
$$N \rightarrow \infty, \quad p \rightarrow 0, \quad E[n] = Np \rightarrow \nu .$$

→ 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 σ found for a fixed integrated luminosity, with $\nu = \sigma \int L dt$.



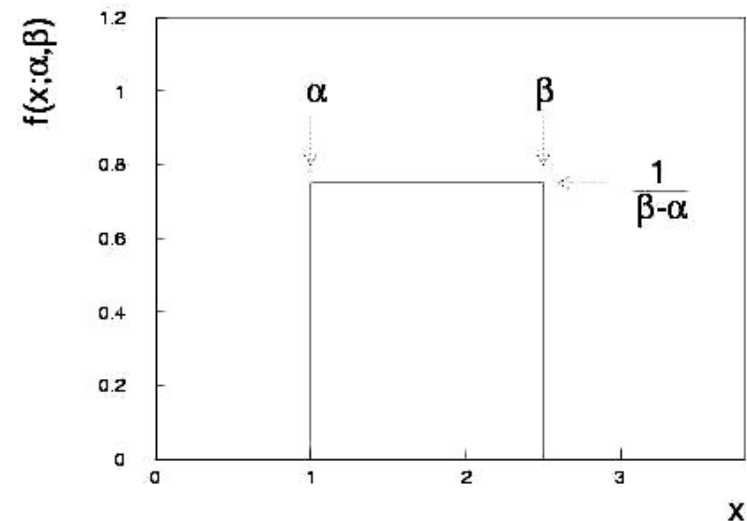
Uniform distribution

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 $p^0 \rightarrow gg$, E_g is uniform in $[E_{\min}, E_{\max}]$, with

$$E_{\min} = \frac{1}{2}E_{\pi}(1 - \beta), \quad E_{\max} = \frac{1}{2}E_{\pi}(1 + \beta)$$

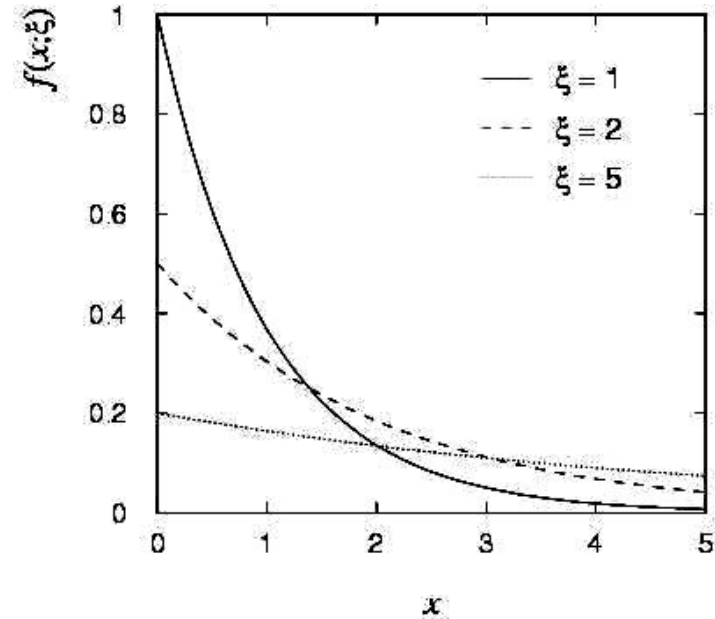
Exponential distribution

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

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

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

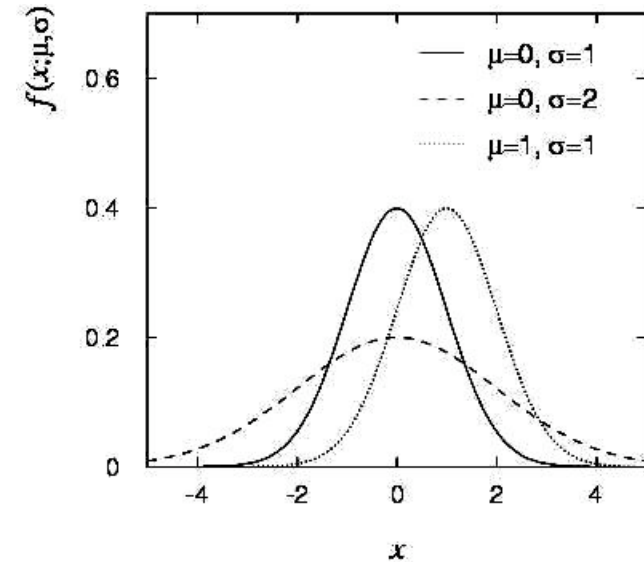
Gaussian distribution

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

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

$$E[x] = \mu \quad (\text{N.B. often } \boldsymbol{m}, s^2 \text{ denote mean, variance of any}$$

$$V[x] = \sigma^2 \quad \text{r.v., not only Gaussian.})$$



Special case: $\boldsymbol{m} = 0, s^2 = 1$ ('standard Gaussian'):

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

If $y \sim \text{Gaussian with } \boldsymbol{m}, s^2$, then $x = (y - \boldsymbol{m})/s$ 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 σ_i^2 , otherwise arbitrary pdfs, consider the sum

$$y = \sum_{i=1}^n x_i$$

In the limit $n \rightarrow \infty$, 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 $\vec{x} = (x_1, \dots, x_n)$:

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

\vec{x} , $\vec{\mu}$ are column vectors, \vec{x}^T , $\vec{\mu}^T$ 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 $r = \text{cov}[x_1, x_2]/(\sigma_1\sigma_2)$ is the correlation coefficient.

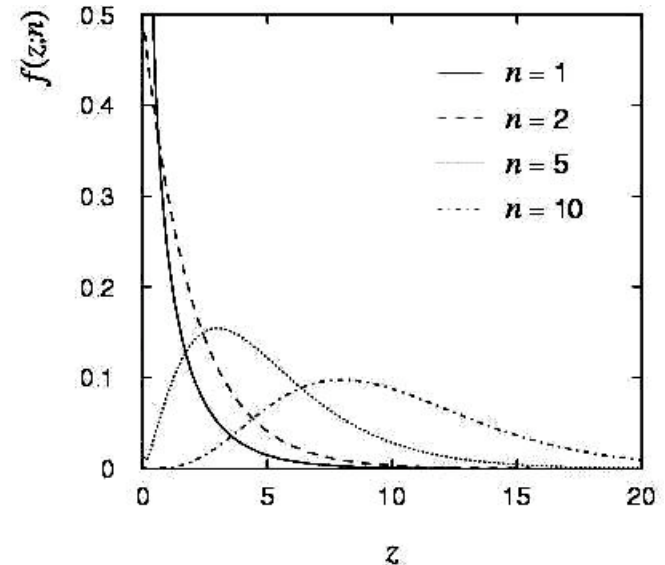
Chi-square (\mathcal{C}^2) distribution

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

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

$n = 1, 2, \dots$ = number of ‘degrees of freedom’ (dof)

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



For independent Gaussian x_i , $i = 1, \dots, n$, means μ_i , variances σ_i^2 ,

$$z = \sum_{i=1}^n \frac{(x_i - \mu_i)^2}{\sigma_i^2} \quad \text{follows } \mathcal{C}^2 \text{ pdf with } n \text{ dof.}$$

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

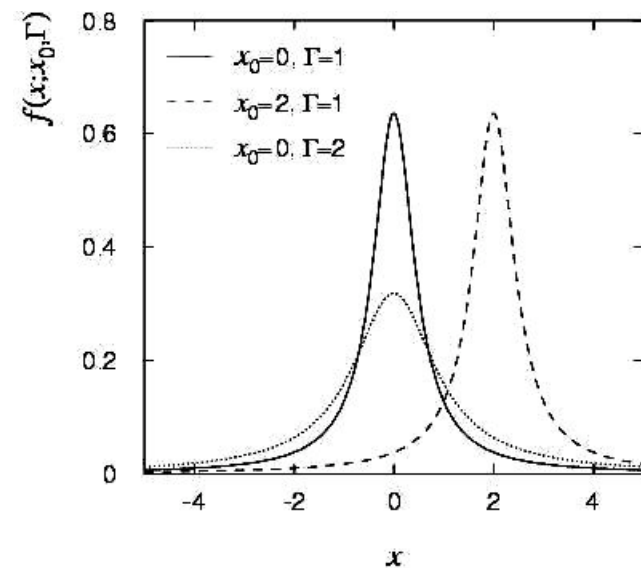
$$f(x; \Gamma, x_0) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}$$

($\Gamma = 2$, $x_0 = 0$ is the Cauchy pdf.)

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

x_0 = mode (most probable value)

Γ = full width at half maximum



Example: mass of resonance particle, e.g. ρ , K^* , f^0 , ...

Γ = decay rate (inverse of mean lifetime)

Landau distribution

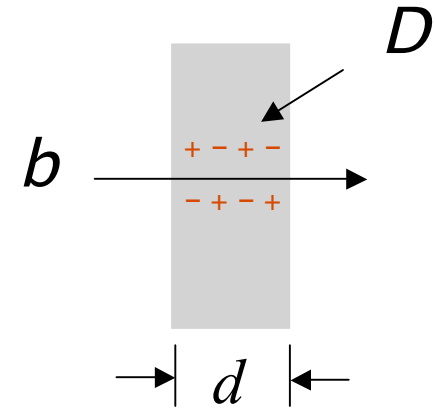
For a charged particle with $b = v/c$ traversing a layer of matter of thickness d , the energy loss D follows the Landau pdf:

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

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

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

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



L. Landau, J. Phys. USSR **8** (1944) 201; see also
W. Allison and J. Cobb, Ann. Rev. Nucl. Part. Sci. **30** (1980) 253.

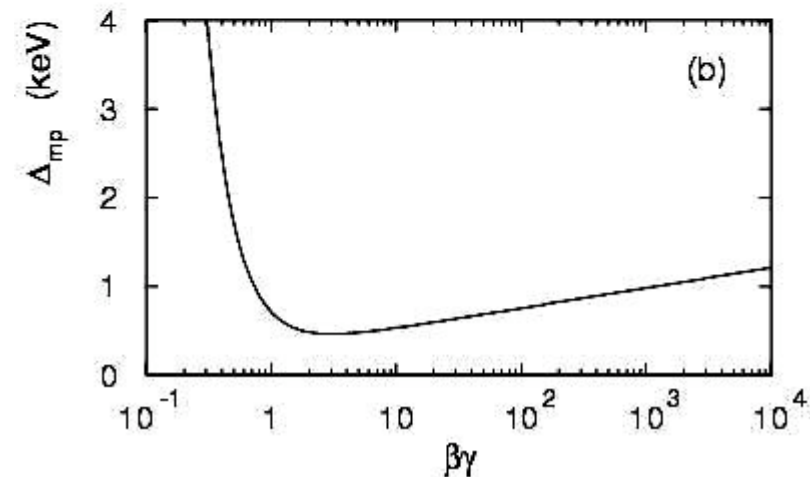
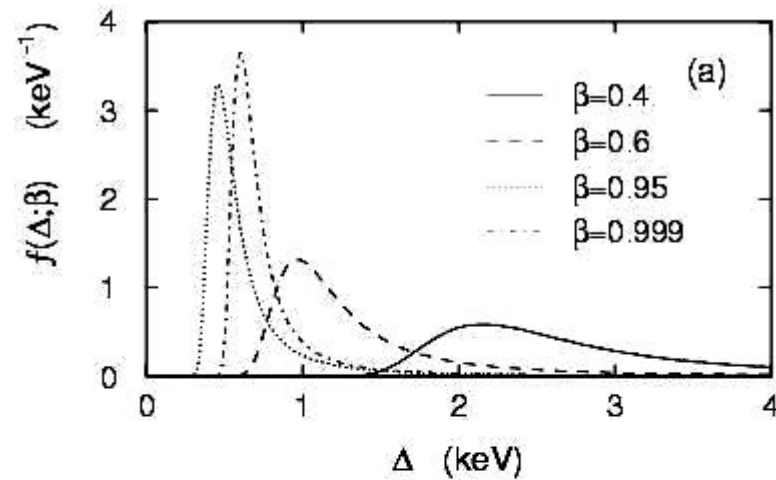
Landau distribution (2)

Long ‘Landau tail’

→ all moments ∞

Mode (most probable value) sensitive to b ,

→ particle i.d.



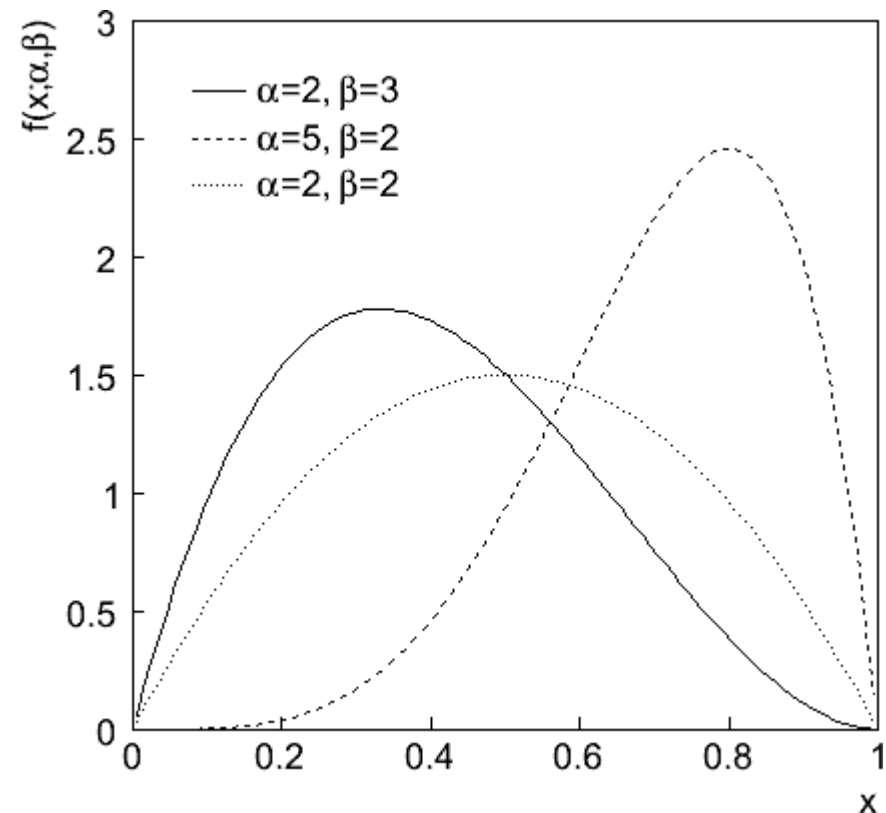
Beta distribution

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

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

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

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



Gamma distribution

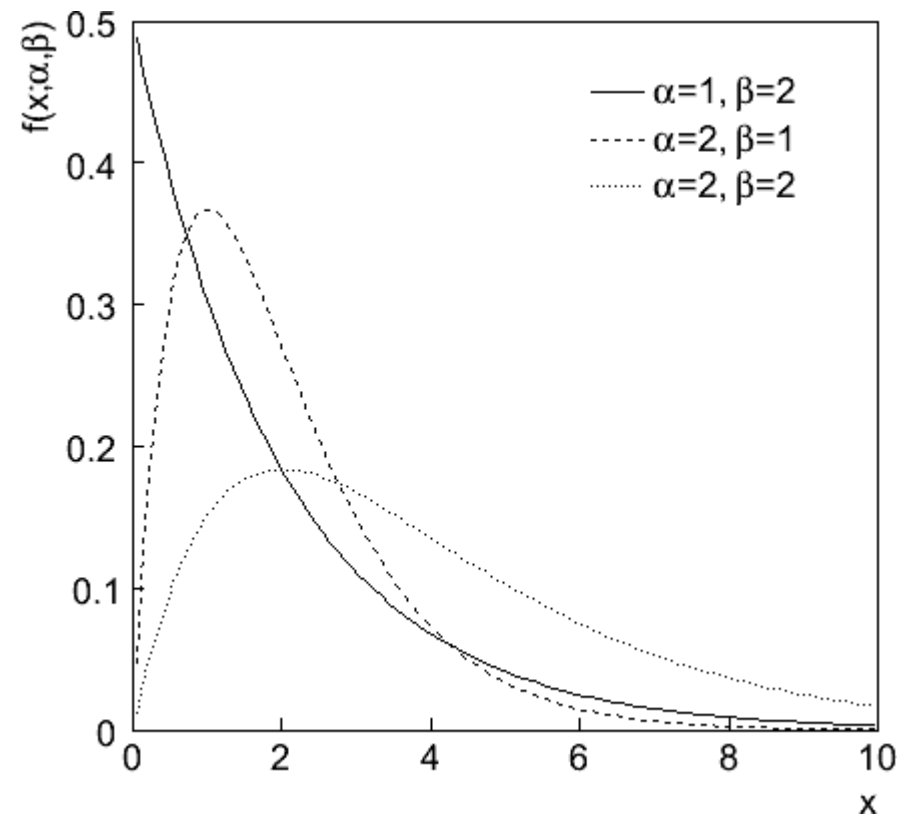
$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

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

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

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

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

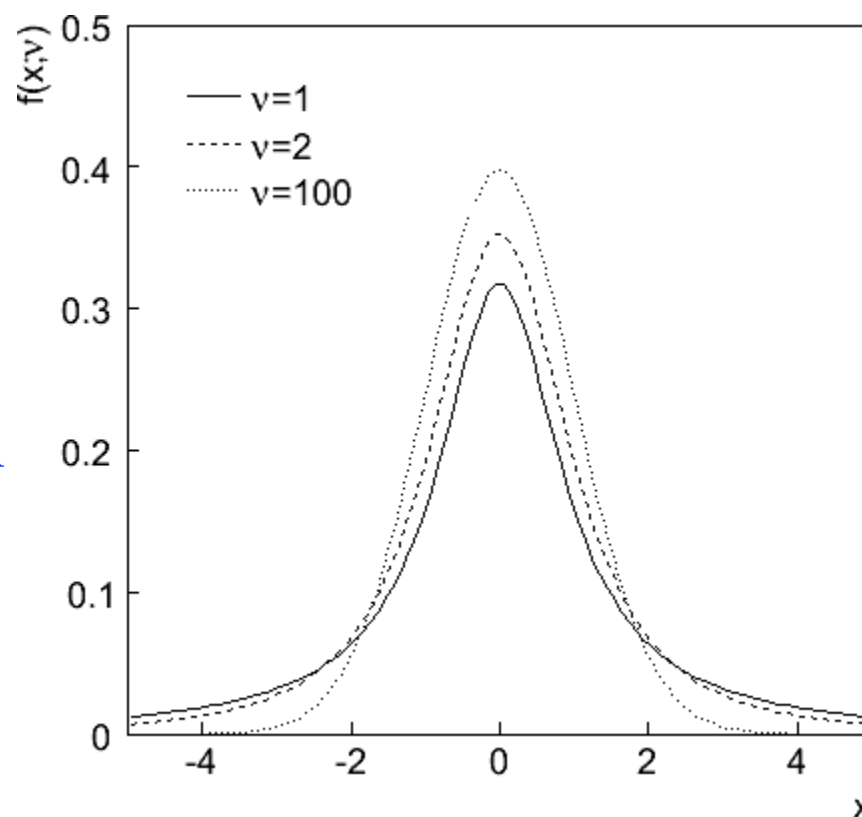
$$E[x] = 0 \quad (\nu > 1)$$

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

n = number of degrees of freedom
(not necessarily integer)

$n = 1$ gives Cauchy,

$n \rightarrow \infty$ gives Gaussian.



Student's t distribution (2)

If $x \sim \text{Gaussian}$ with $m = 0$, $s^2 = 1$, and

$z \sim \chi^2$ with n degrees of freedom, then

$t = x / (z/n)^{1/2}$ follows Student's t with $n = n$.

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, ($n \rightarrow \infty$, but in fact already very Gauss-like for $n = \text{two dozen}$), to the very long-tailed Cauchy ($n = 1$).

Developed in 1908 by William Gosset, who worked under the pseudonym "Student" for the Guinness Brewery.

Wrapping up lecture 4

We've looked at a number of important distributions:

Binomial, Multinomial, Poisson, Uniform, Exponential
Gaussian, Chi-square, Cauchy, Landau, Beta,
Gamma, Student's t

and we've seen the important Central Limit Theorem:

explains why Gaussian r.v.s come up so often

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

`http://www.physto.se/~walck/`