Statistical Data Analysis: Lecture 2

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

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)

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

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

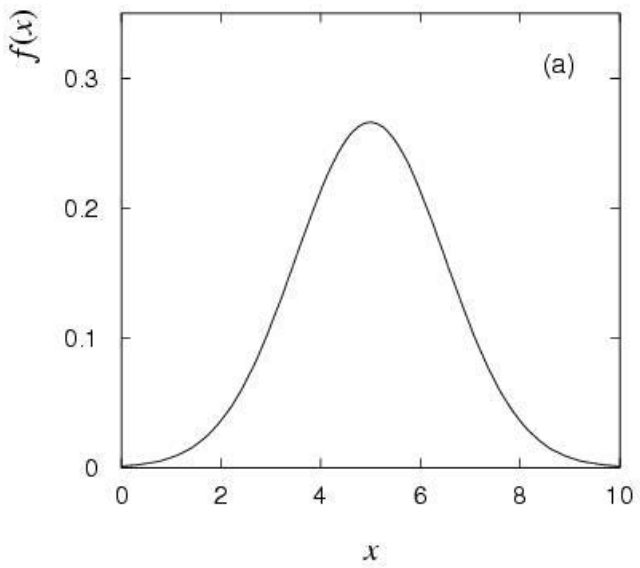
$$P(x_i) = p_i$$ probability mass function

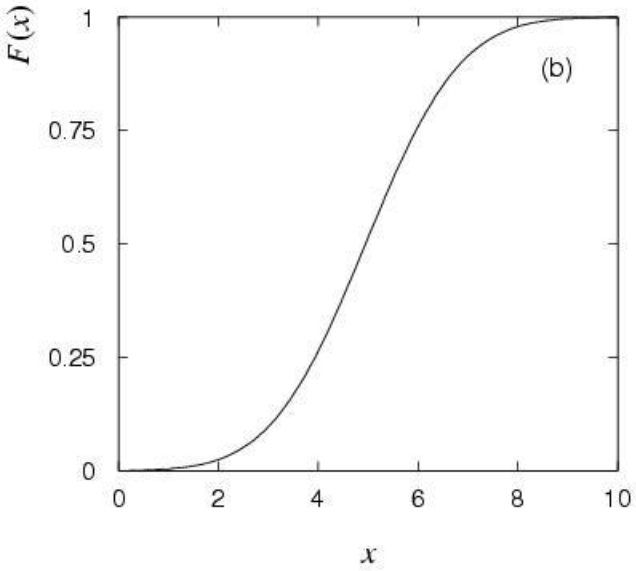
$$\sum_i P(x_i) = 1$$ x must take on one of its possible values

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 $f(x) = \dfrac{\partial F(x)}{\partial x}$

Histograms

pdf = histogram with

infinite data sample,

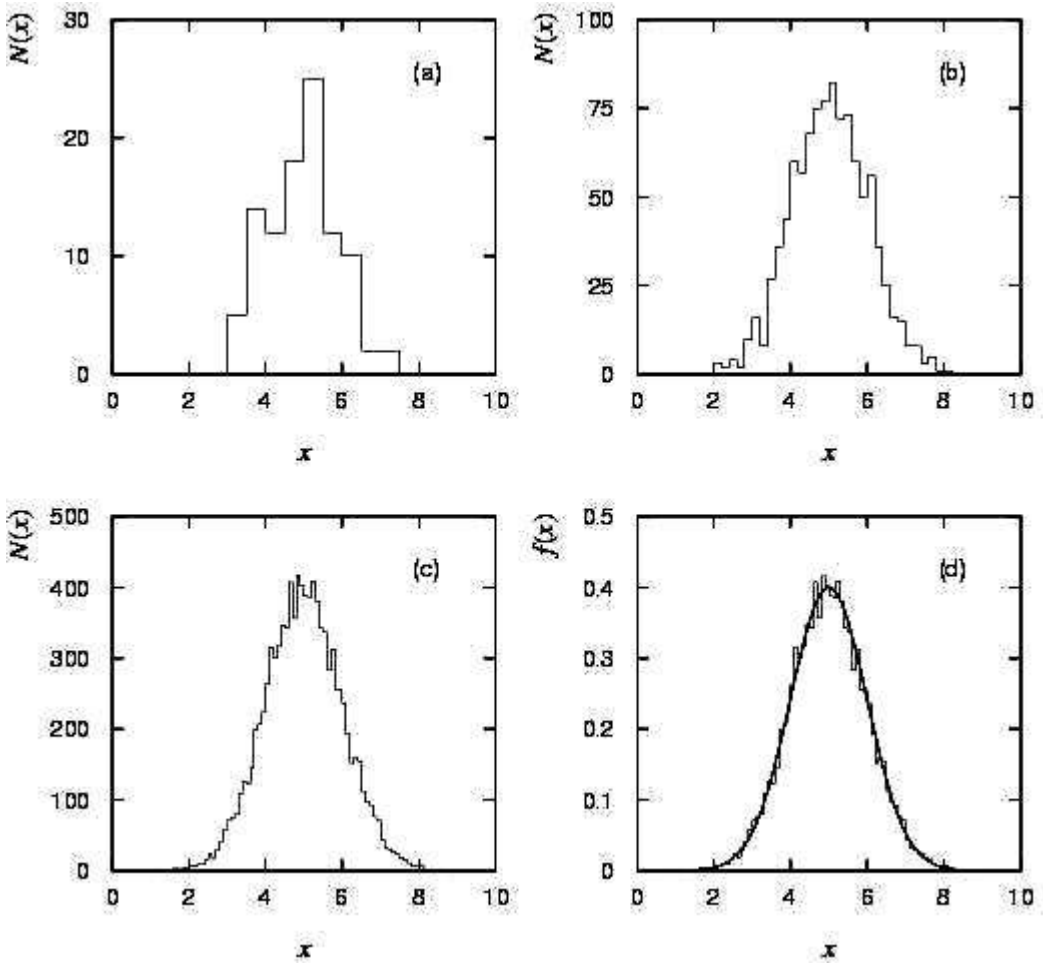
zero bin width,

normalized to unit area.

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

n = number of entries

Δx = bin width

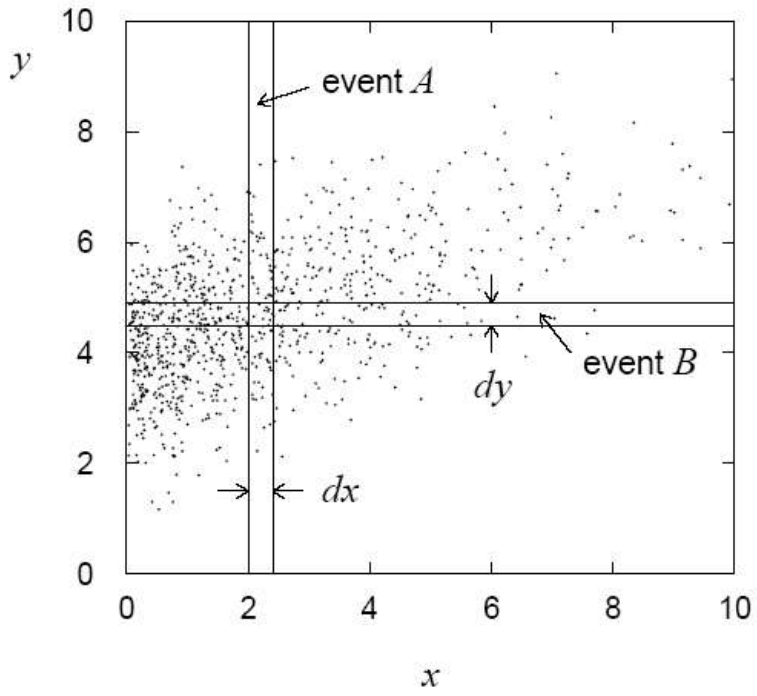


Multivariate distributions

Outcome of experiment characterized 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



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

Marginal pdf

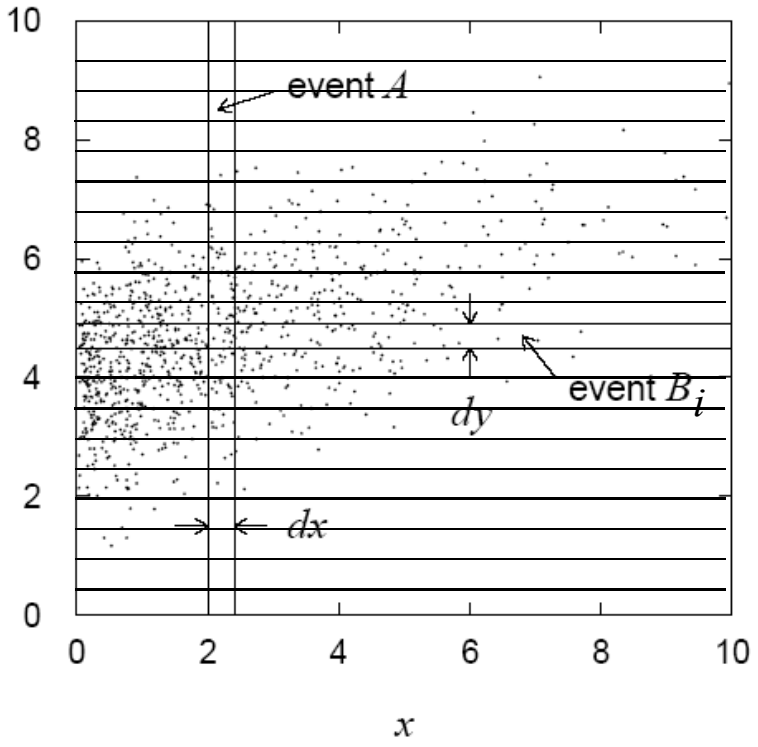
Sometimes we want only pdf of some (or one) of the components:

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

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

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

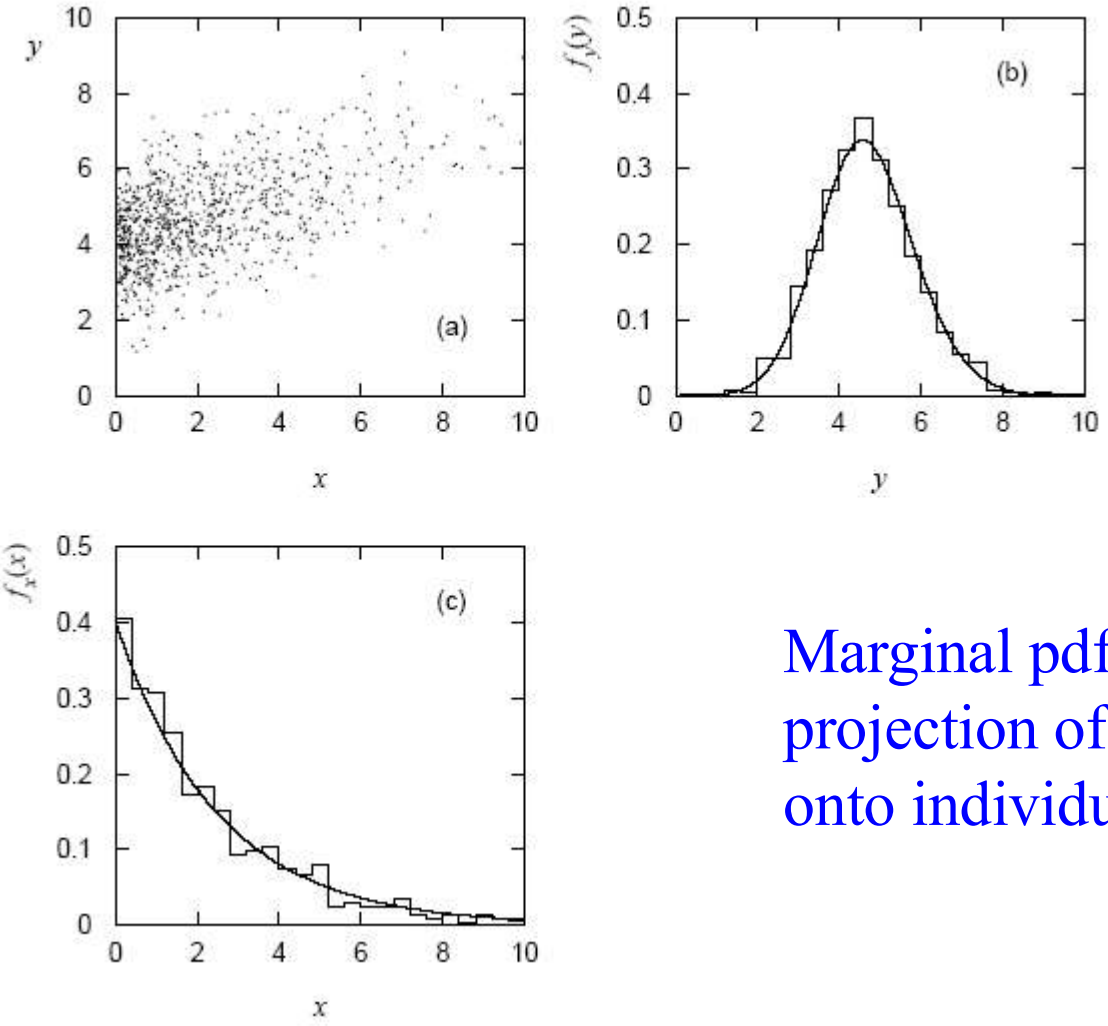
$$f_x(x) = \int f(x, y)\, dy$$



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

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

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

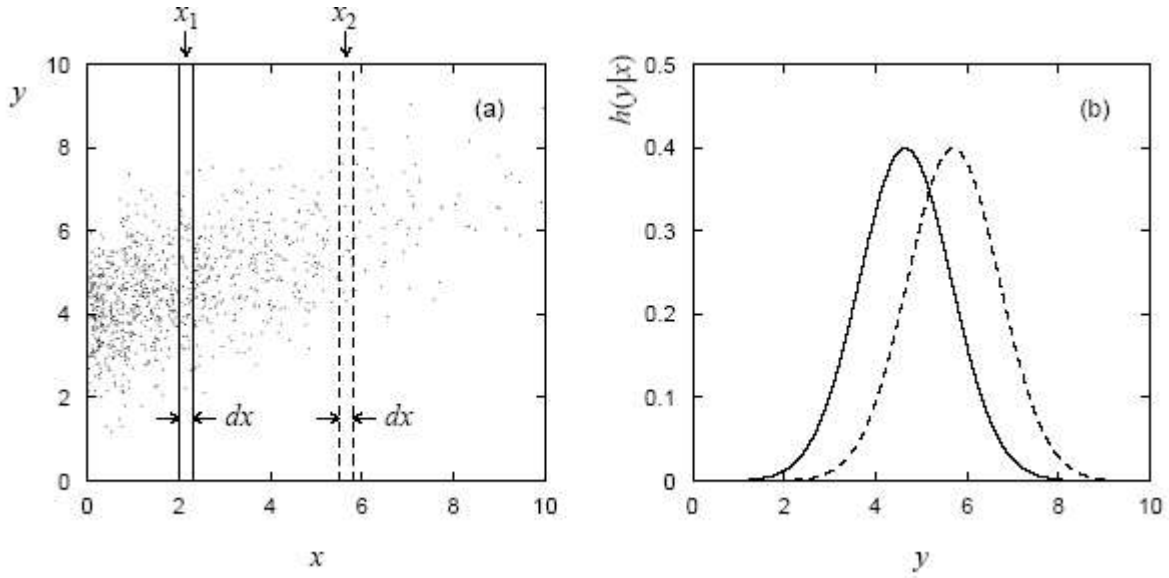
Bayes' theorem becomes: $g(x|y) = \frac{h(y|x) f_x(x)}{f_y(y)}$.

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



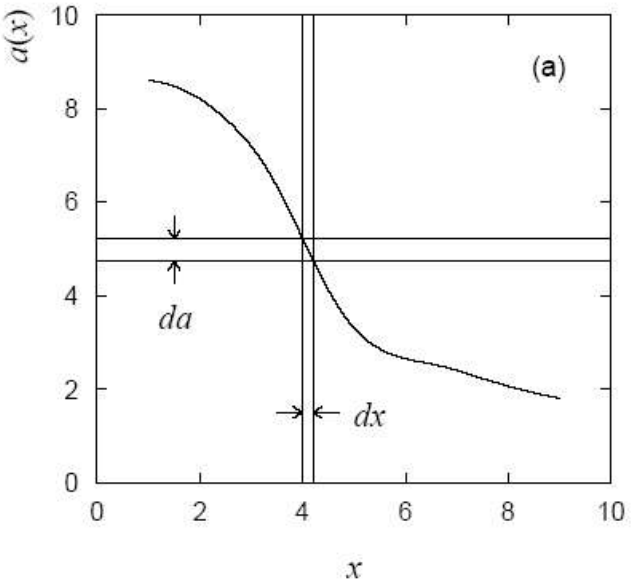
Basically treat some of the r.v.s as constant, then divide the joint pdf by the marginal pdf of those variables being held constant so that what is left has correct normalization, e.g., $\int h(y|x)\,dy = 1$.

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 $[a, a+da]$.

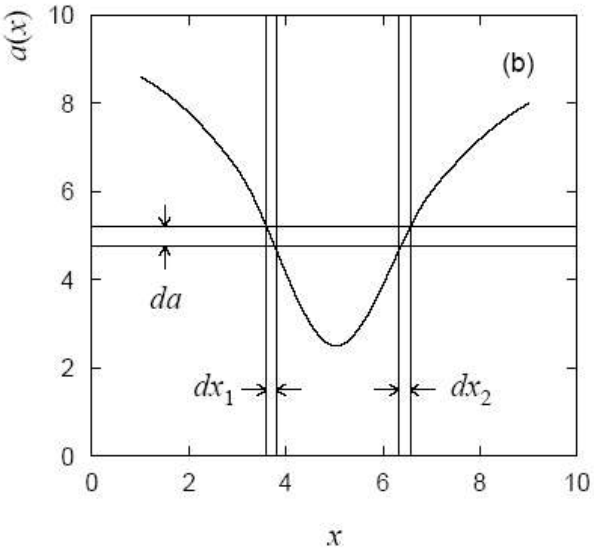
For one-variable case with unique inverse this is simply

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

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

Functions without unique inverse

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



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

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

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

Functions of more than one r.v.

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

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

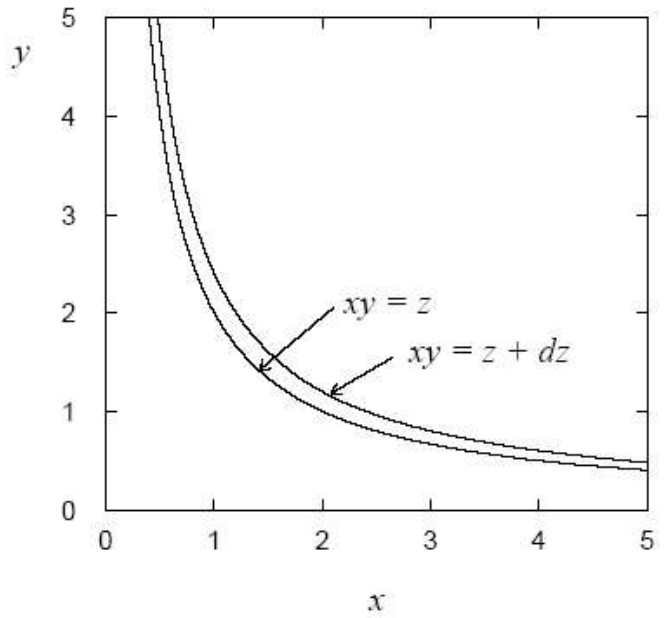
dS = region of x-space between (hyper)surfaces defined by

$$a(\vec{x}) = a', \; a(\vec{x}) = a' + da'$$

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 $g(z)$?



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

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

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

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

(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:

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

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.