Data analysis methods in weather and climate research

Dr. David B. Stephenson
Department of Meteorology
University of Reading
www.met.rdg.ac.uk/cag

10. Spatio-temporal methods for gridded datasets

- ??
- ??
- ??

4. Probability distributions

1. Continuous and discrete distributions
2. Discrete distributions
   1. Bernoulli distribution X~Be(pi)
   2. Binomial distribution X~Bin(n,pi)
   3. Poisson distribution X~Poisson(mu)
3. Continuous distributions
   1. Uniform distribution X~U(a,b)
   2. Normal (Gaussian) X~N(mu,sigma)
   3. Gamma distribution X~Gamma(alpha,beta)

4. Random variables and notation

Random variable (r.v) = a number X associated with a random event A. Instead of having to talk of P(A) one can then talk about distributions P(X=x).

- Discrete random variables: e.g. integer counts X=0,1,2,3,…
- Continuous random variables: e.g. temperature, rainfall amount, etc.

Statistical notation:

Random variables – large Roman	e.g. X
Specific values – small Roman	e.g. x
Model parameters – Greek	e.g. α, β, θ

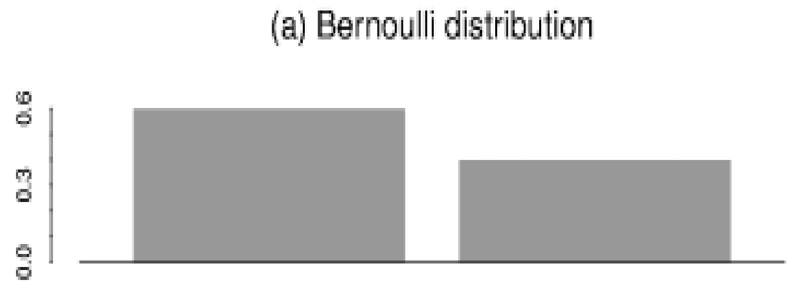
4. Distributions of discrete variables

$$P(X = x) = f(x)$$

Bernoulli distribution

X~Be(p=0.4)

$$f(x) = p^x (1-p)^{1-x}$$
$$x = 0,1$$

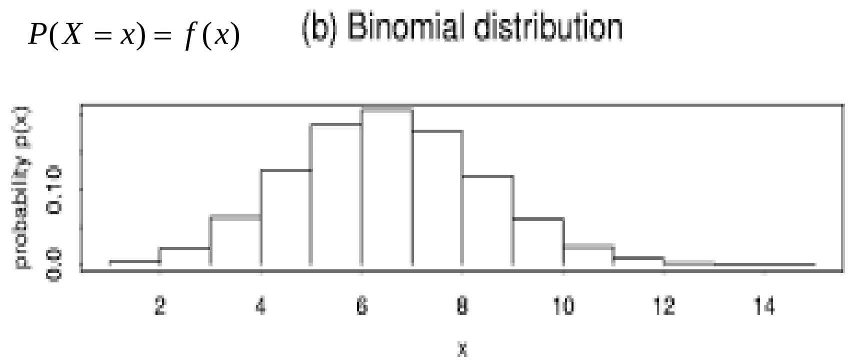


Binomial distribution

X~Bin(n=15,p=0.4)

$$f(x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{1-x}$$
$$x = 0,1,\ldots,n$$

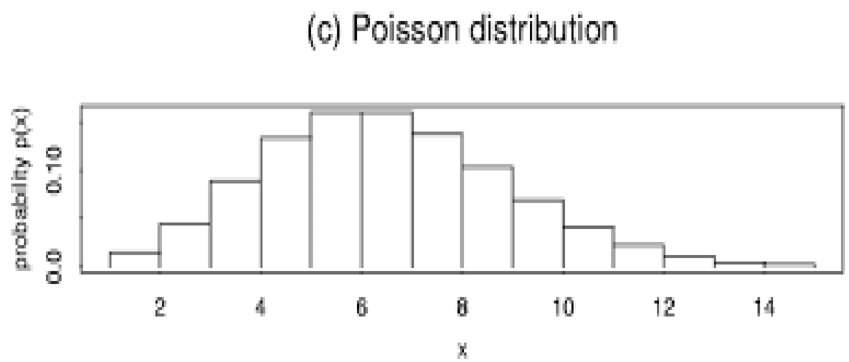
$P(X = x) = f(x)$



Poisson distribution

X~Poisson(mu=6)

$$f(x) = \frac{e^{-\mu}\mu^x}{x!}$$
$$x = 0,1,2,\ldots$$



4. Distributions of continous variables

$$P(X \le x) = F(x) = \int^{x} f(x')dx'$$

Uniform distribution
X~U(a=-2,b=2)

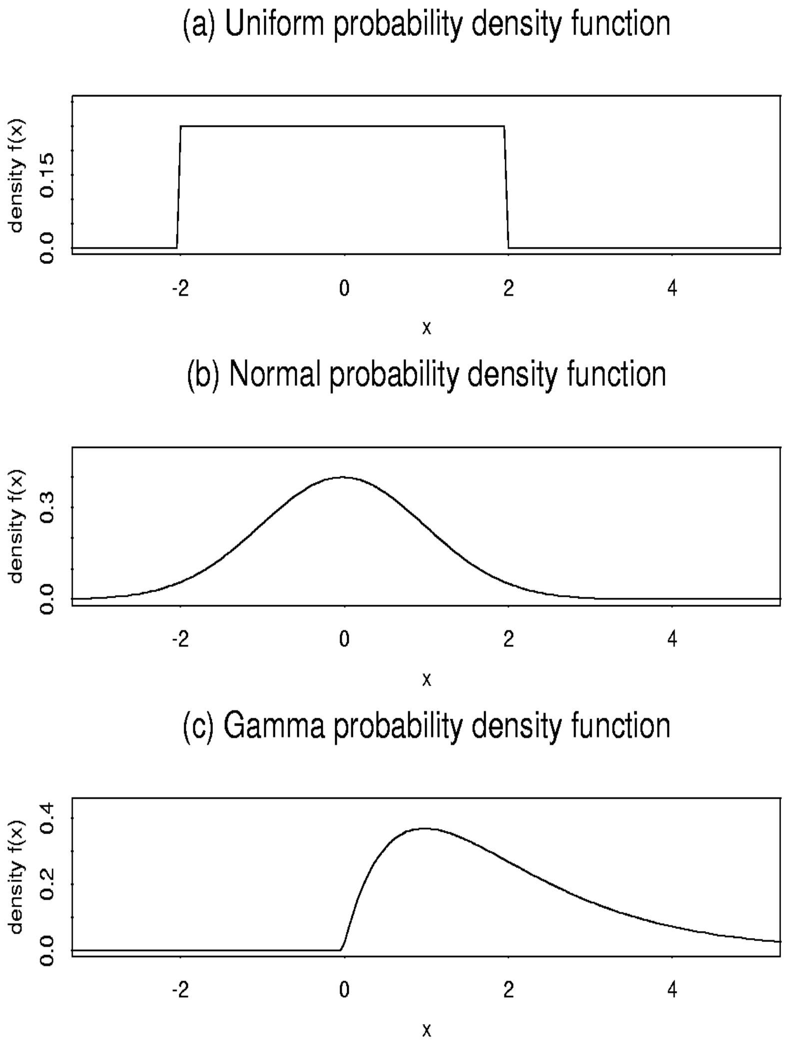
$$f(x) = (b-a)^{-1}$$
$$a \le x \le b$$

Normal distribution
X~N(0,1)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Gamma distribution
X~Gamma(2,1)

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$
$$x \ge 0$$



4. Expectation (population mean)

$$E(X) = \sum_{j=1}^{m} X_j \Pr\{X = X_j\}$$

Sometimes referred to as the population mean μ_X

Not to be confused with the sample mean $\overline{x}$

4. Covariance

$$\mathrm{var}(X,Y) = E((X - E(X))(Y - E(Y)))$$

Properties:

- When X=Y gives population variance
- Independent of means of X and Y

$$\mathrm{var}(X,X) = \sigma_X^2$$

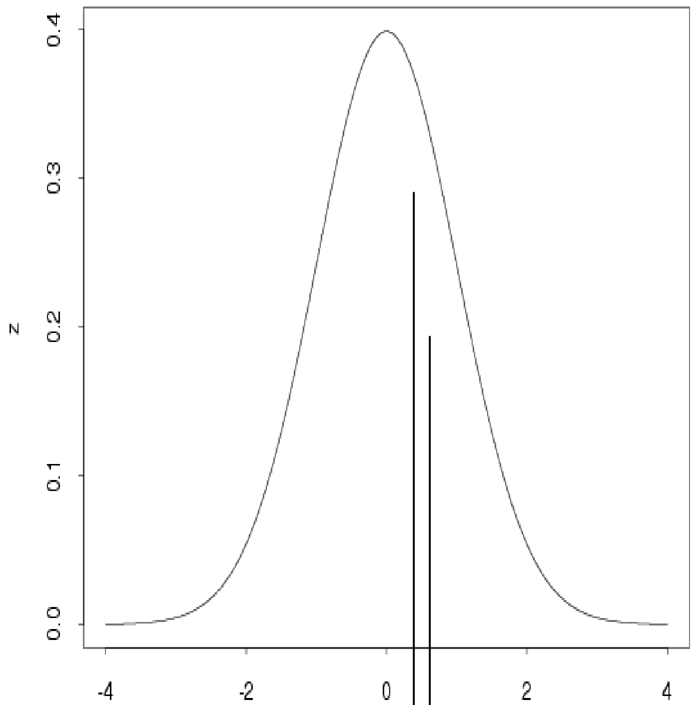
4. Correlation

$$cor(X,Y) = \frac{\text{var}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}$$

Properties:

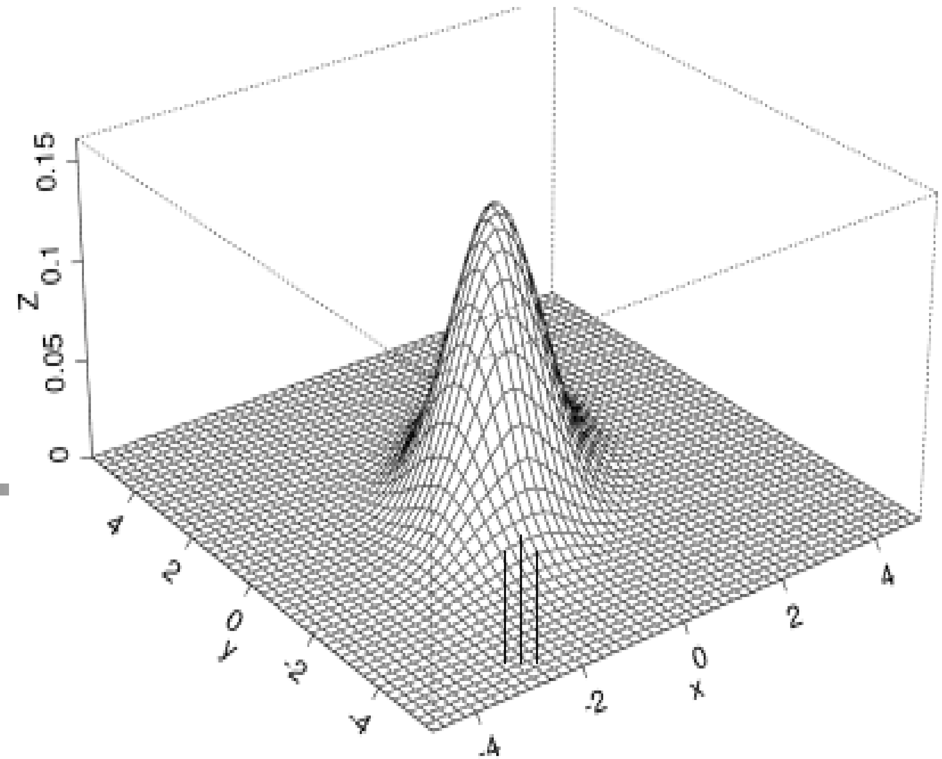
- Measure of linear association varying between –1 and +1
- Independent of mean and variance of X and Y
- Symmetric in X and Y

Probability density functions (distributions)



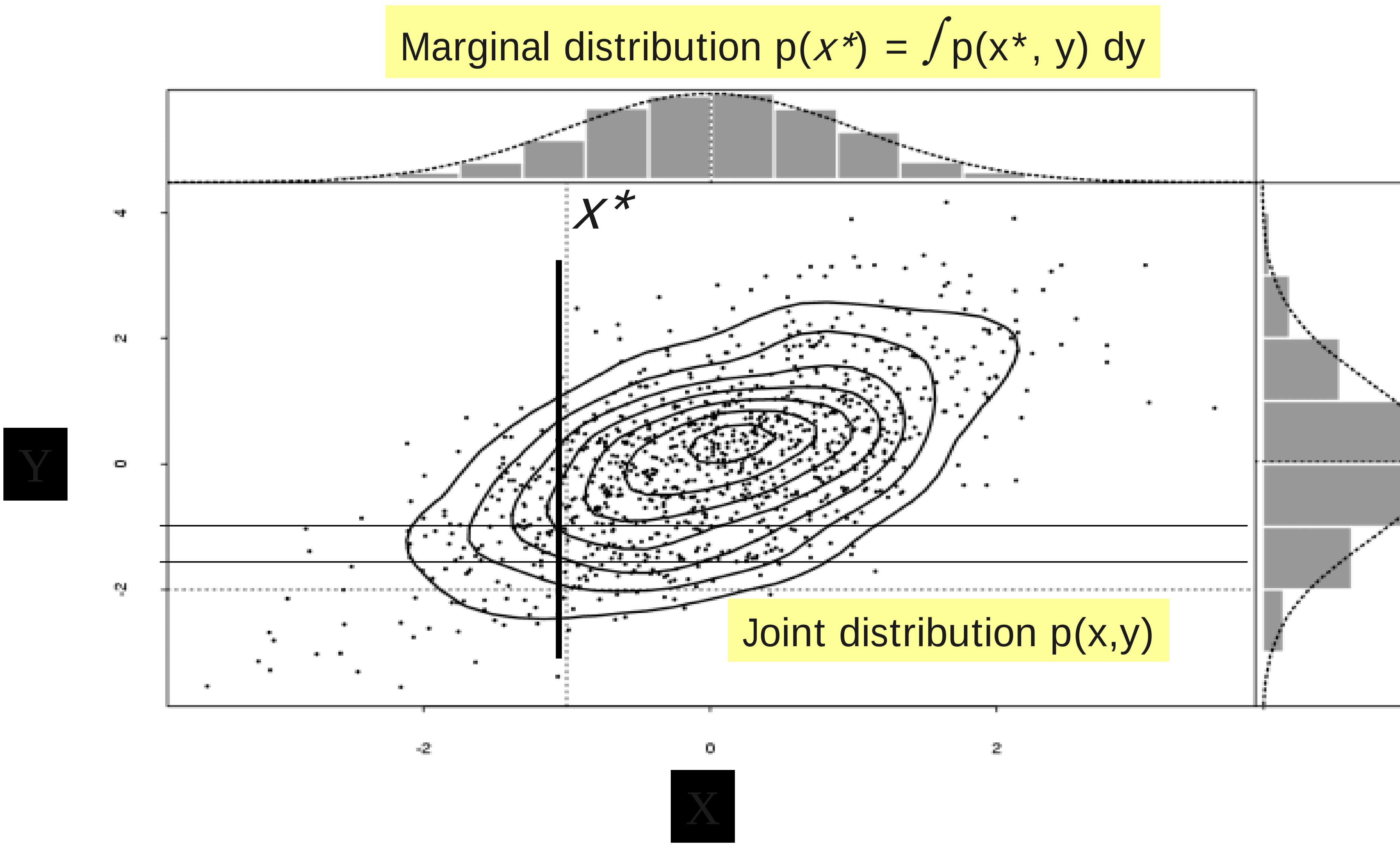
Univariate

e.g. Normal (Gaussian)



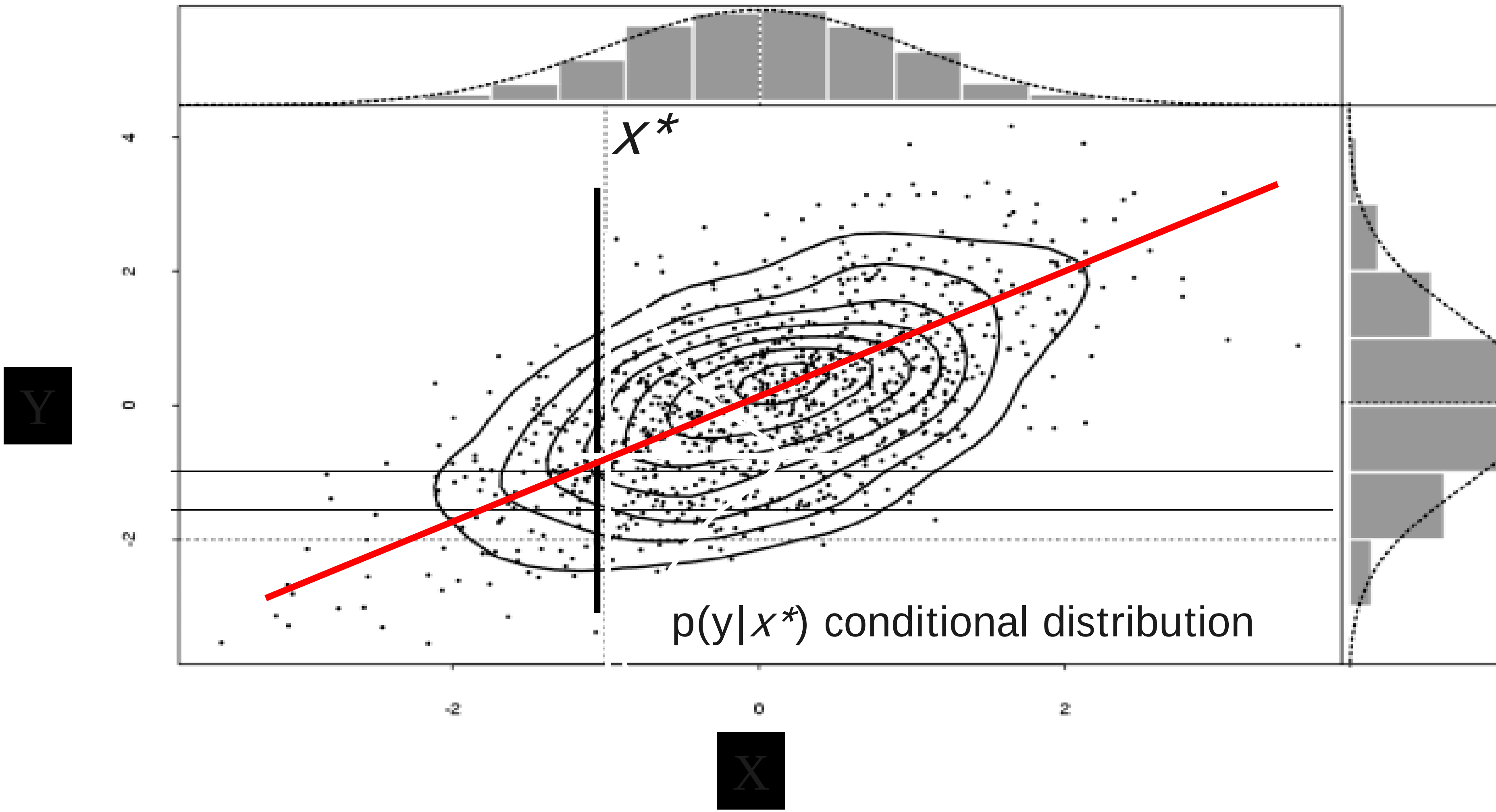
bivariate
joint distribution of X & Y

Distributions of more than one variable



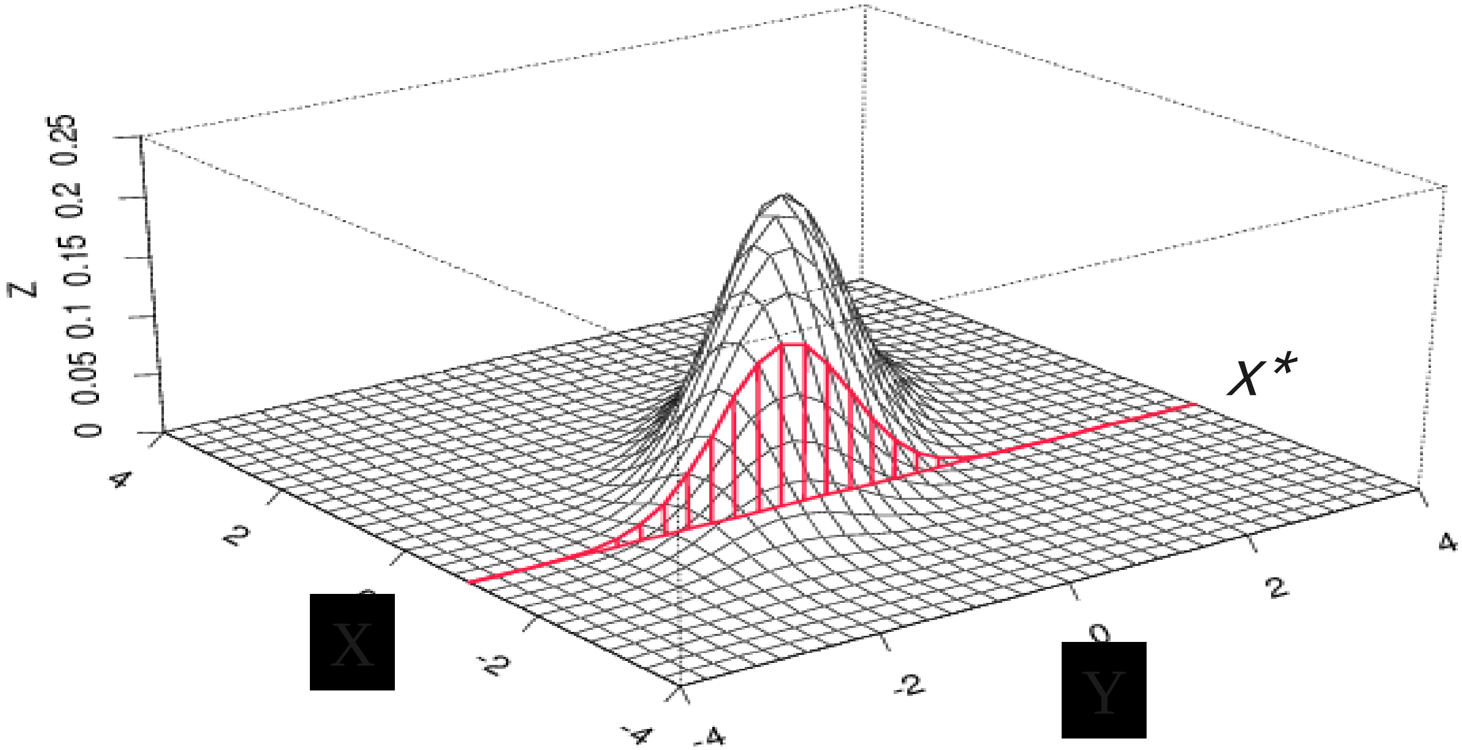
Distributions of more than one variable

Marginal distribution $p(x^*) = \int p(x^*, y)\,dy$



Conditional distribution

$$p(y \mid x^*) = p(x^*, y) / p(x^*)$$



4. Regression model interpretation

- Data-analytic – least-squares minimisation

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\text{Minimise} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- Probability model

$$Y \mid X \sim N(\alpha + \beta X, \sigma_\varepsilon^2)$$

$$\Rightarrow$$

$$E(Y \mid X) = \alpha + \beta X$$

$$Var(Y \mid X) = \sigma_\varepsilon^2$$

4. Summary

- The concept of probability is fundamental for summarising uncertainty

- Probability is a fascinating concept

- Regression models are conditional probability models and hence regression forecasts can be used to issue probability forecasts.

Probability problem

You meet a person who tells you that they have 2 childre
and that 1 (or more) of them is a girl.

What is the probability that the other child is also a girl?

The person now tells you that the girl has a very rare name with probability p close to zero.

Would you revise your probability estimate? If so, what

Summary