Data analysis methods in weather and climate research

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

4. Probability distributions

- Random variables and mathematical notation
- Distributions of discrete random variables
- Distributions of continuous random variables
- Expectation, Variance, Covariance and Correlation

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. Probability distributions/models

Discrete distributions

- Bernoulli distribution X~Be(pi)
- Binomial distribution X~Bin(n,pi)
- Poisson distribution X~Poisson(mu)

Continuous distributions

- Uniform distribution X~U(a,b)
- Normal (Gaussian) X~N(mu,sigma)
- Gamma distribution X~Gamma(alpha,beta)

where "~" means "distributed as"

4. Distributions of discrete variables

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

Bernoulli distribution
X~Be(p=0.4)

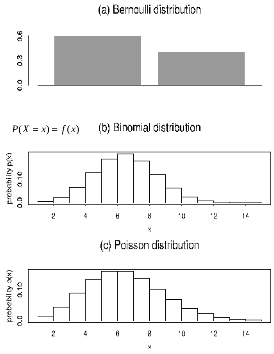
$$f(x) = p^x (1-p)^{1-x}$$
$$x = 0,1$$

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

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

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)

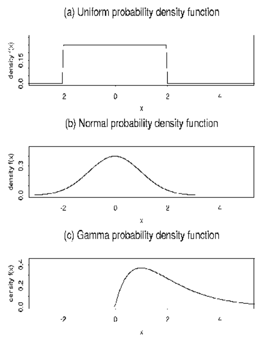
$$f(x) = (b-a)^{-1}$$
$$a \le x \le b$$

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)

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

Also referred to as the **population mean** μ_X

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

Useful properties:

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$
$$E(XY) = E(X)E(Y) \text{ if } X,Y \text{ independent}$$

4. Covariance and Variance

$$Cov(X,Y) = E((X - E(X))(Y - E(Y)))$$
$$= E(XY) - E(X)E(Y)$$

$Cov(X,X) = Var(X) = \sigma_X^2$ population variance

Properties:

$$Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X,Y)$$

4. Example: variance of sample mean

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = E\left(\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} x_i x_j\right) - \left[E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\right]^2$$
$$= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} E(x_i x_j) - \left[\frac{1}{n}\sum_{i=1}^{n} E(x_i)\right]^2$$
$$= \left(\mu^2 + \frac{\sigma^2}{n}\right) - \mu^2 = \frac{\sigma^2}{n}$$

using

$E(x_i) = \mu$

$E(x_i x_j) = \mu^2 + \sigma^2 \delta_{ij}$ (assumes independent x)

4. (Product moment) correlation

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

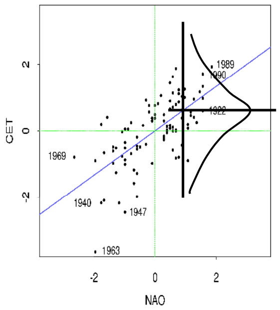
Properties:
- Measure of linear association varying between –1 and +1
- Independent of mean and variance of X and Y
- Symmetric in X and Y

4. Probability models

Dependency can be modelled by allowing the parameters of a distribution to depend on other explanatory variables:

For example:

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



linear regression of Y on X

(mean Y depends linearly on X)

4. Summary

- Random events can be labelled with numbers known as *random variables*
- The probabilities for a set of values of a random variable defines a *probability distribution*
- Probability distributions can be modelled by a wide range of theoretical curves (*probability models*)
- The population (true) mean of a random variable is known as the expectation – it is analogous to the sample mean.

Sample	$\{x_1, x_2, \ldots, x_n\}$	mean $\bar{x}$	variance s_x^2
Population	X	$\mu = E(X)$	$\sigma^2 = Var(X)$