Stochastic modelling

Definitions

1. For two events A and B, the probability of A **or** B is $\text{Prob}(A \cup B) = \text{Prob}(A) + \text{Prob}(B) - \text{Prob}(A \cap B)$, where the last bit term is the *joint probability* of A and B.

2. For *mutually exclusive* events (joint prob. = 0, e.g., "individual is male/female"); probability of A *or* B $\equiv \text{Prob}(A \cup B) = \text{Prob}(A) + \text{Prob}(B)$.

3. The sum of probabilities of all possible outcomes of an observation or experiment = 1.0. (E.g.: *normalization constants*.)

4. *Conditional probability* of A *given* B, $\text{Prob}(A|B)$: $\text{Prob}(A|B) = \text{Prob}(A \cap B)/\text{Prob}(B)$. (Compare the *unconditional* probability of A: $\text{Prob}(A) = \text{Prob}(A|B) + \text{Prob}(A|\text{not } B)$.)

5. If $\text{Prob}(A|B) = \text{Prob}(A)$, A is *independent* of B. Independence $\iff$ $\text{Prob}(A \cap B) = \text{Prob}(A)\text{Prob}(B)$ (or $\log \prod_i \text{Prob}(A_i) = \sum_i \log \text{Prob}(A_i)$).

Probability distributions

Discrete: probability distribution, cumulative probability distribution. Continuous: cumulative distribution function, probability *density* function ($p(x)$ = limit of $\text{Prob}(x < X < x + \Delta x)/\Delta x$ as $\Delta x \to 0$). Describe by **moments**. *Mean*: $\sum x_i/N = \sum \text{count}(x)x/N = \sum p(x)x$ (discrete), $\int p(x)x\,dx$. *Variance*: $\sum p(x)(x - \bar{x})^2$, $\int p(x)(x - \bar{x})^2\,dx$. *Higher moments*: skew, kurtosis. Also: median, mode.

Bestiary

Pretty good summaries on Wikipedia (`http://en.wikipedia.org/wiki/List_of_probability_distributions`). R help pages. Books: [1, 2], Johnson, Kotz, Balakrishnan et al.

Characteristics (discrete vs continuous; range (positive, bounded, ...); symmetric or skewed ...)

$\Omega = \{0, 1, ..., N\}$

- Binomial
  - Number of successes in a given number of trials
  - parameters: p – probability of success; n – number of trials.
  - mean: Np
  - variance: $Np(1 - p)$
- Beta-binomial
  - Number of successes in a given number of trails with variation (overdispersion) in the probability of success
  - parameters: p – average probability per trial; θ – overdispersion parameter (how much does the probability of success vary?)
  - mean: Np
  - variance: $Np(1 - p)(1 + (N - 1)/(\theta + 1))$

$\Omega = \{0, 1, ...\}$

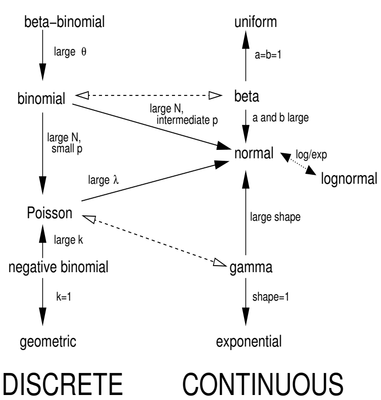
- Geometric
  - Number of trials until a single failure
  - parameters: p – probability of failure
  - mean: $1/p$
  - variance: $(1 - p)/p^2$
- Poisson
  - Counts of events if events are independent
  - parameters: λ – expected count
  - mean: λ
  - variance: λ
- Negative binomial
  - Poisson distribution with variation (overdispersion) in λ
  - parameters: μ – expected number of counts; k – overdispersion parameter
  - mean: μ
  - variance: $\mu + \mu^2/k$

$\Omega = \Re$

- Normal distribution
  - Good for many many things!
  - parameters: μ – mean; σ – standard deviation
  - mean: μ
  - variance: σ^2

$\Omega = \Re^+$

- Gamma
  - Waiting times until a certain number of events
  - parameters: s – scale (average time between events); a – shape (number of events)
  - mean: as
  - variance: as^2
- Exponential
  - Gamma with shape parameter one, $a = 1$
- Lognormal
  - Distribution of e^X where X is distributed normally
  - parameters: μ – mean of the logarithm; σ – standard deviation of the logarithm
  - mean: $\exp(\mu + \sigma^2/2)$
  - variance: $\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$



(Insanely thorough version: [3].)

Method of moments

Quick estimation of parameters. Solve $\{\bar{x} = \text{(theor. mean)}, s^2 = \text{(theor. variance)}$: sometimes biased.

Jensen's inequality

$E[f(x)] \neq f(E[x])$, unless f is linear (notation: expectation of function $f(x)$ over PDF $p(x)$: $E_p[f(x)] \equiv \int p(x) f(x)\, dx$. Analytic integration (if possible); numeric integration; or *delta method*.

Delta method

To approximate the expected value and variance of a function using only the expected value of its argument, we make a Taylor-series approximation: $E_p[f(x)] \approx E_p[f(\bar{x})] + E_p[f'(x)|_{x=\bar{x}}(x - \bar{x})] + 1/2 E_p[f''(x)|_{x=\bar{x}}(x - \bar{x})^2]$. Therefore:

- $E_p[f(x)] \approx f(E_p[x]) + 1/2 f''(E_p[x]) \text{Var}(x)$
- $\text{Var}(f(x)) \approx f'(E_p[x])^2 \text{Var}(x)$

Markov-chains

Markov models are multivariate linear models where the columns sum to 1. Dominant eigenvalue now equal to 1; the dominant eigenvector is the stable distribution. The reason these are stochastic models is that the elements of the state vector can be used to model the probability of being in a particular state. Therefore, the sum of the state vector is 1, because you must be in a particular state.

In particular, $\boldsymbol{x}(n+1) = \boldsymbol{P}^\top \boldsymbol{x}(n)$, where $\boldsymbol{x}$ is the state vector and $\boldsymbol{P}$ is called the transition matrix. The element in the ith row of the jth column gives the probability of moving from state i to state j at each time step. This model makes the following assumptions:

1. Markov property: each move is independent of previous moves.
2. Probabilities out of a state must sum to one (i.e. rows of $\boldsymbol{P}$ sum to one, or columns of $\boldsymbol{P}^\top$ must sum to one)

Steady-state solution

From linear theory, $\boldsymbol{x}(n) = (\boldsymbol{P}^\top)^n \boldsymbol{x}(0)$.

Fixed points

There are two kinds of fixed points we will deal with.

Definition The system is called regular if some power of $\boldsymbol{P}^\top$ has all positive entires.

For all regular systems, the dominant eigenvalue of $\boldsymbol{P}^\top$ is one. Therefore, if $d\boldsymbol{v} = \boldsymbol{P}^\top \boldsymbol{v}$, where d and $\boldsymbol{v}$ is an eigenpair of $\boldsymbol{P}^\top$, then if $\boldsymbol{v}$ is the dominant eigenvector, $d = 1$, and $\boldsymbol{v} = \boldsymbol{P}^\top \boldsymbol{v}$. Therefore, $\boldsymbol{x}\star = \boldsymbol{v}$ is a fixed point.

Here is the second kind of fixed point.

Definition The system is called absorbing if $\boldsymbol{P}^\top$ can be rearranged into the form,

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{B} & \boldsymbol{I} \end{pmatrix}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices, $\boldsymbol{0}$ is a matrix of all zeros, and $\boldsymbol{I}$ is the identity matrix. Why is this system called absorbing? Hint: recall the drunkard's walk!

In this case, as time-dependent solution is,

$$\boldsymbol{x}(n) = \begin{pmatrix} \boldsymbol{A}^n & \boldsymbol{0} \\ \boldsymbol{B}(\boldsymbol{I} - \boldsymbol{A})^{-1} & \boldsymbol{I} \end{pmatrix} \boldsymbol{x}(0)$$

In the limit, we can further reduce this expression because $\boldsymbol{A}^n$ goes to a matrix of zeros as n increases without bound (why?),

$$\boldsymbol{x}\star = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{B}(\boldsymbol{I} - \boldsymbol{A})^{-1} & \boldsymbol{I} \end{pmatrix} \boldsymbol{x}(0)$$

which is the fixed point for an absorbing Markov chain.

References

[1] B. M. Bolker. *Ecological Models and Data in R*. Princeton University Press, 2008.

[2] M. Evans, N. Hastings, C. Forbes, and J. B. Peacock. *Statistical Distributions*. John Wiley & Sons, 2010.

[3] L. M. Leemis and J. T. McQueston. Univariate distribution relationships. *The American Statistician*, 62(1):45–53, 2008.