Probability basics

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

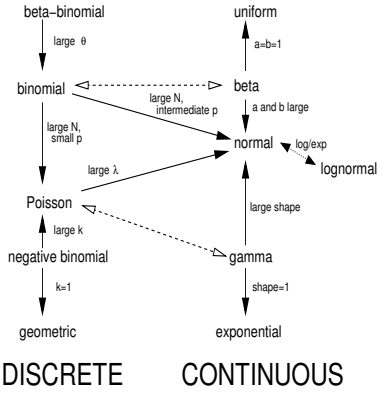
Discrete: probability distribution, cumulative probability distribution. Continuous: cumulative distribution function, probability *density* function ($p(x) =$ limit of $\text{Prob}(x < X < x + \Delta x)/\Delta x$ as $\Delta x \to 0$). Describe by **moments**. *Mean*: $\sum x_i/N = \sum \text{count}(x)x/N = \sum p(x)x$ (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 ...)

- Binomial (*Bernoulli*: $N = 1$)
- Poisson: (limit of Binomial with $N \to \infty$, Np constant)
- Negative binomial: coin-flipping, or *overdispersed* version of Poisson (Gamma-Poisson)
- Normal (Gaussian): continuous, symmetric; central limit theorem.
- Gamma: waiting time (*exponential*: shape=1).
- Lognormal: exponential-transformed Gaussian.



(Insanely thorough version: [3].)

Method of moments

Quick estimation of parameters. Solve $\{\bar{x} =$ (theor. mean), $s^2 =$ (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: $E_p[f(x)] \approx E_p[f(\bar{x})] + E_p[f'(x)|_{x=\bar{x}}(x-\bar{x})] + 1/2E_p[f''(x)|_{x=\bar{x}}(x-\bar{x})^2] = f(\bar{x}) + 1/2f''(\bar{x})\text{Var}(x)$.

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.