2ND LA SERENA DATA SCHOOL

Introduction to probability and statistics - I

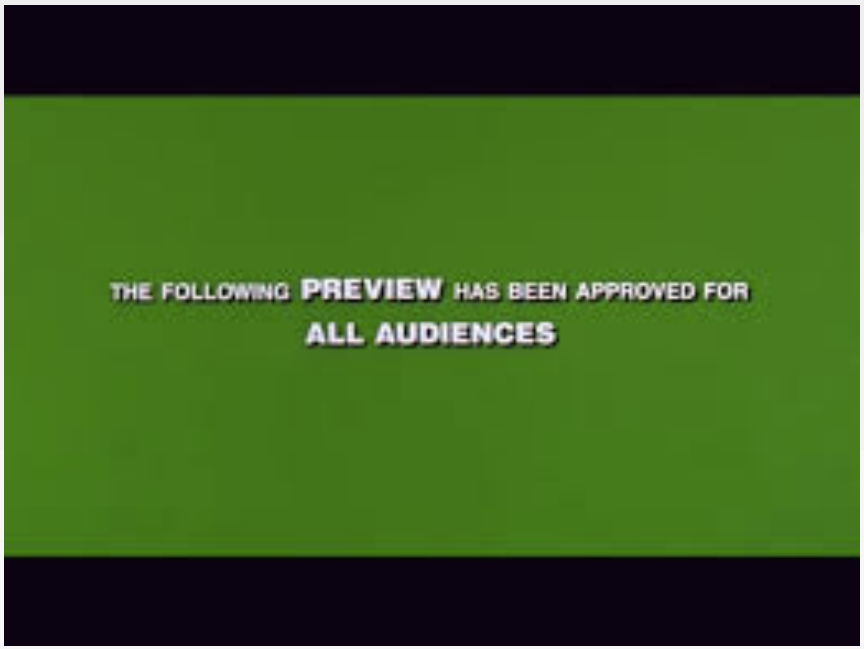
Matthew J. Graham, Caltech

Aug 16, 2014

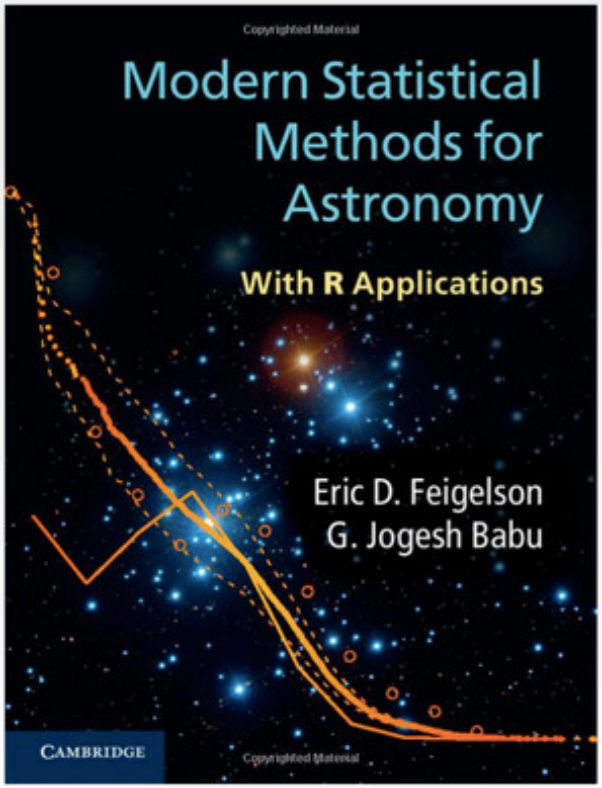
Overview

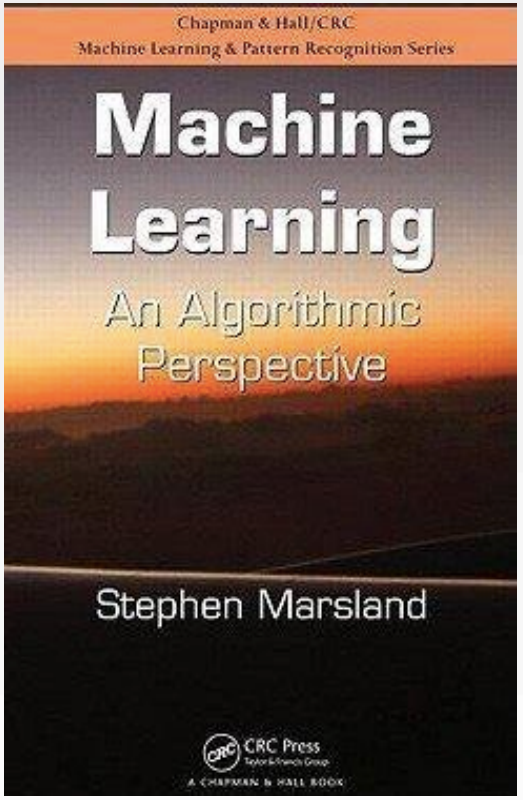
- Basic terminology
- Describing data
- Distributions
- Linear regression

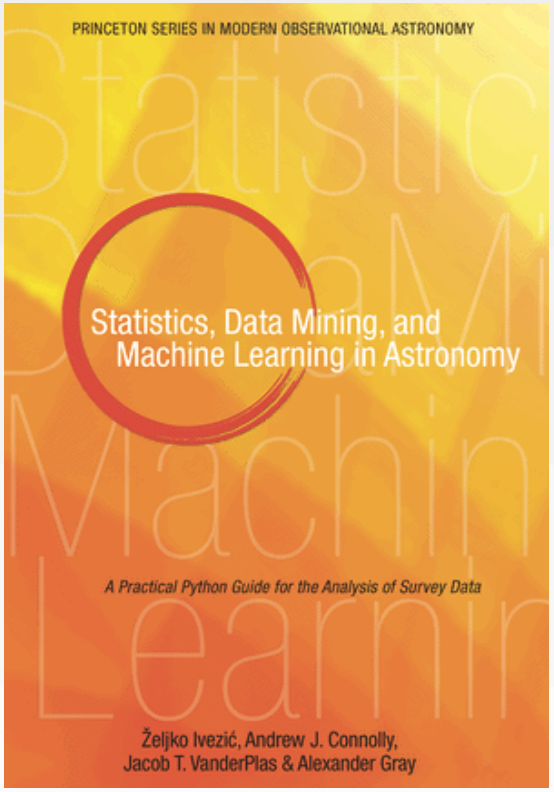
- Confidence intervals
- Histograms
- Modeling/fitting data



Recommended reading



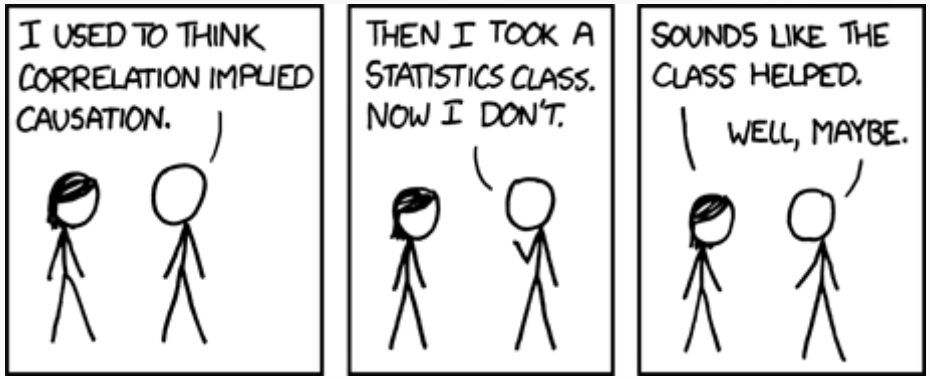




Typical data questions

- What is the best estimate for a model parameter *x* given the available data?
- How confident are we about the value of model parameter *x*?
- Is this particular set of data consistent with a given hypothesis or model?

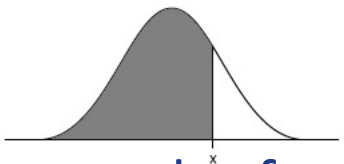
- To answer these we use probability and statistics



Basic terminology - I

- Discrete variable: maps the sample space to a countable set with probability ≥ 0
- Continuous variable: maps the sample space to a noncountable set with probability of any specific value = 0
- Cumulative distribution function, F(x): gives the probability that a variable X has a value less than or equal to *x*
- Probability density function, f(x): for a continuous variable:

$$F(x) = \int_{-\infty}^{x} f(y)\,dy$$



- Significance: a measure of how unlikely the result of a hypothesis test is to have occurred by change. The p-value is the probability of observing a result at least as extreme as the test statistic.

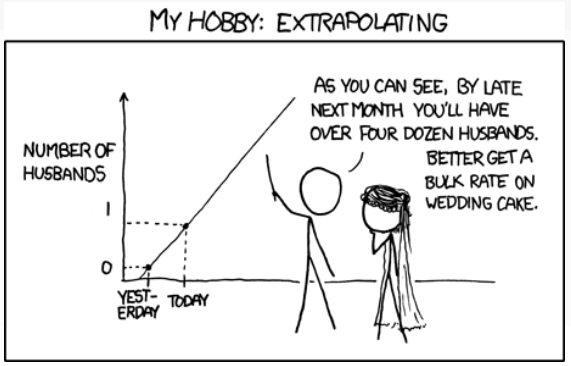
Basic terminology - II

- Outlier: a data point that is *distant* from other data points
- Nonparametric: methods that do not involve parametric assumptions or require the data to belong to a particular parametric family of distributions
- Robust estimator: insensitive to slight deviations
- Prior: probability before any evidence is taken into account
- Posterior: probability after relevant evidence taken into account
- Marginalizing: the distribution of a subset of variables in a multidimensional distribution

$$p(x) = \int_y p(x,y)\,dy = \int_y p(x|y)p(y)\,dy$$

Basic terminology - III

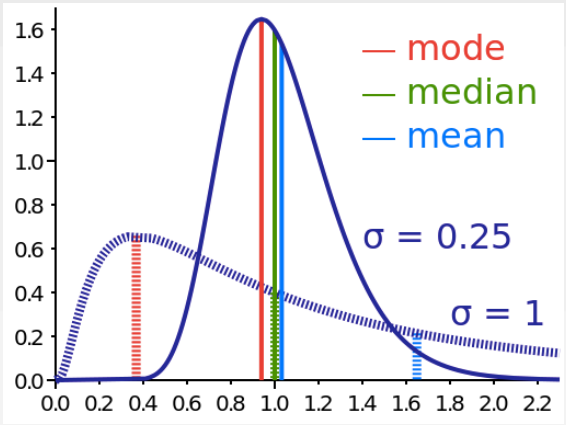
- Covariance: A measure of how much two random variables change together
- Estimator bias: the difference between the expected value of an estimator and the true value of the parameter
- Interpolation: Using a model fit to observed data to estimate values within the observed range
- Extrapolation: Using a model fit to observed data to estimate values outside the observed range
- Heteroscedastic: A sample of variables where the dispersion varies between subsamples



Descriptive statistics - I

Location – some notion of a central value

Mean	Median	Mode
$\int_{-\infty}^{\infty} xh(x)\,dx$	$\int_{-\infty}^{x_m} xh(x)\,dx = 0.5$	$\left(\frac{dh(x)}{dx}\right)_{x_m} = 0$



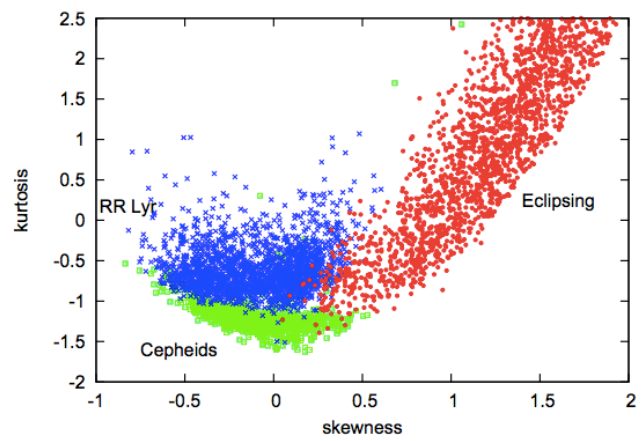
Scale – some notion of spread or dispersion

Variance	IQR	MAD	Biweight midvariance
$\int_{-\infty}^{\infty} x^2 h(x)\,dx - \mu^2$	$Q_3 - Q_1$ $\int_{-\infty}^{Q_n} xh(x)\,dx = 0.25n$	$\text{med}(\lvert x - \text{med}(x)\rvert)$ $\sigma = 1.4826\,\text{MAD}$	$\frac{n\sum_{i=1}^{n}(x_i - Q)^2(1-u_i^2)^4}{\left(\sum_i (1-u_i^2)(1-5u_i^2)\right)^2}$ for $\lvert u_i\rvert < 1$ $u_i = \frac{x_i - Q}{9\,\text{MAD}}$, $Q = \text{med}(x)$

Descriptive statistics - II

Shape – some notion of asymmetry or peakedness

Skew	Kurtosis
$\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^3 h(x)\,dx$	$\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^4 h(x)\,dx - 3$



Graczyk & Eyer (2010)

L-moments

$$\lambda_r = r^{-1}\binom{n}{r}^{-1}\sum_{x_1<\cdots<x_j<\cdots x_r}(-1)^{r-j}\binom{r-1}{j}x_j$$

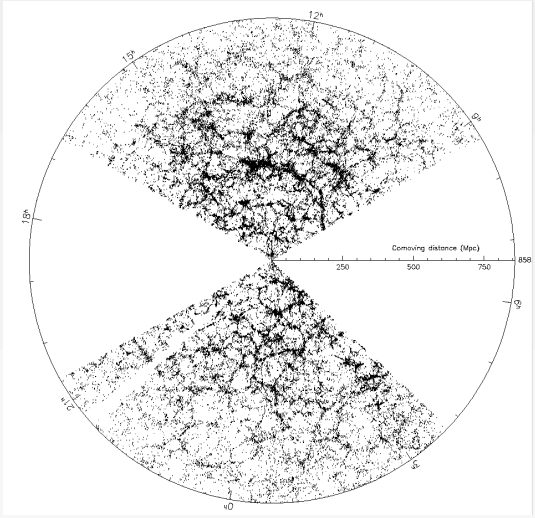
$$\tau_r = \frac{\lambda_r}{\lambda_2}$$

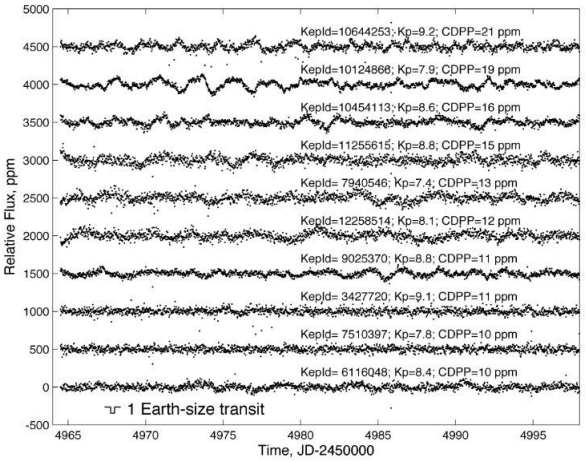
- L-mean, L-scale, L-skew, L-kurtosis

Descriptive statistics - III

Statistical measures characterize particular patterns of behaviour

- Spatial statistics to quantify clustering and structure:
  - Density estimation
  - Correlation function
  - Power spectrum
- Temporal statistics to describe time:
  - Variability
  - Periodicity
  - Autocorrelation
  - Stochasticity

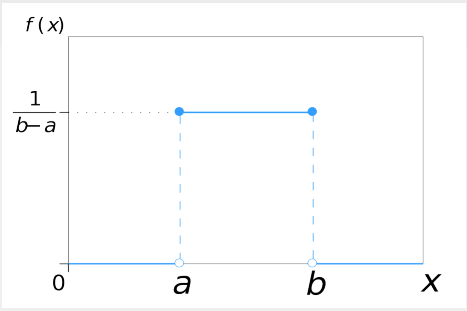




Distributions - I

Uniform (top-hat, box)

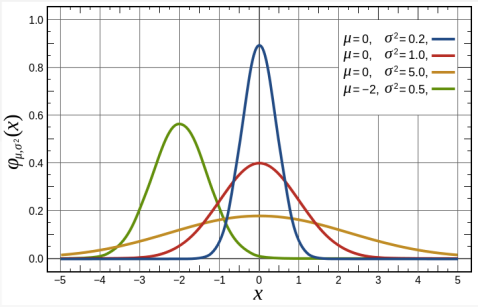
$$f(x) = \frac{1}{b-a} \text{ for } a < x < b$$



- Important for simulations and sampling

Gaussian

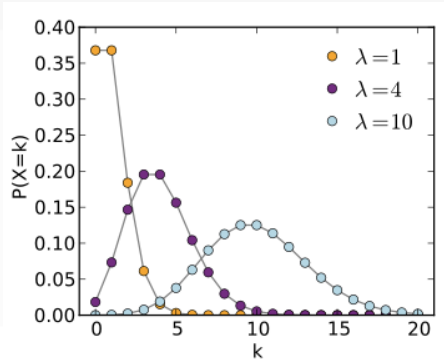
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$



- Most significant distribution in science

Poisson

$$p(x = k) = \frac{\lambda^k}{k!} \exp(-\lambda)$$

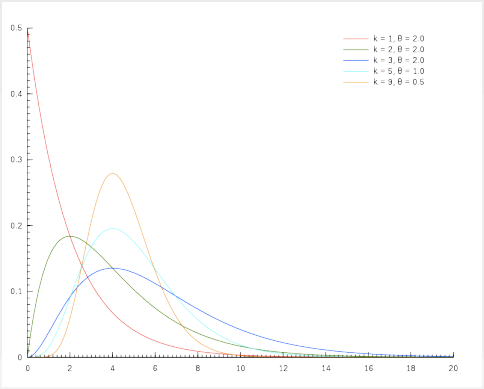


- Distribution of number of photons in a given interval

Distributions - II

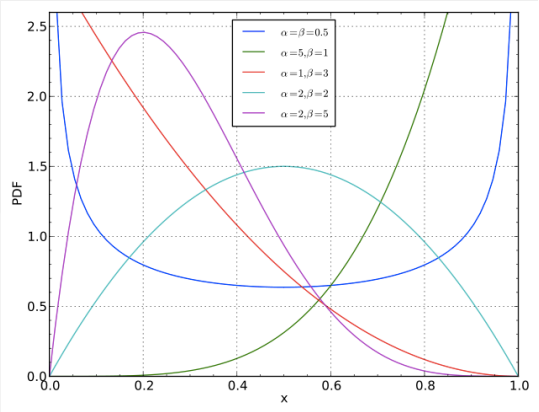
Gamma

$$f(x) = \frac{x^{k-1}\exp(-x/\theta)}{\theta^k \Gamma(k)}$$



Beta

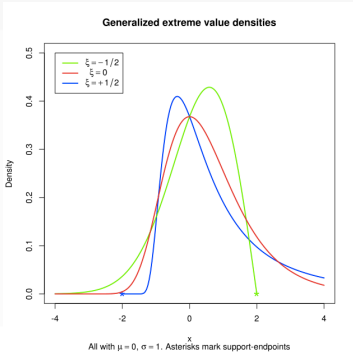
$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$



- Random variable limited to intervals of finite length

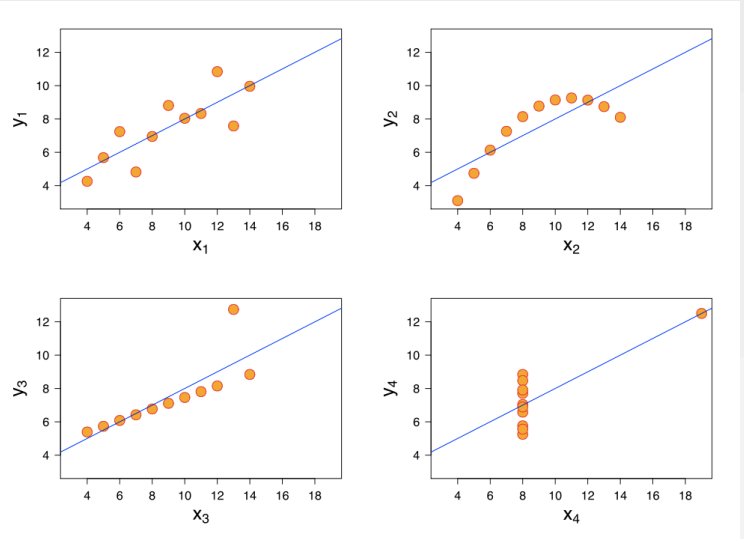
Generalized extreme value (Gumbell, Frechet, Weibull)

$$f(x) = \frac{1}{\sigma} t(x)^{\xi+1} \exp(-t(x)) \text{ where } t(x) = \begin{cases} \left(1 + \left(\frac{x-\mu}{\sigma}\right)\xi\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ \exp\left(\frac{-(x-\mu)}{\sigma}\right) & \text{if } \xi = 0 \end{cases}$$



Linear regression (trends)

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

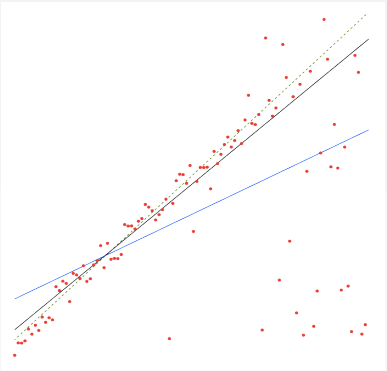


Ordinary least squares

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \left(\sum \mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\left(\sum \mathbf{x}_i\mathbf{y}_i\right)$$

Thiel-Sen

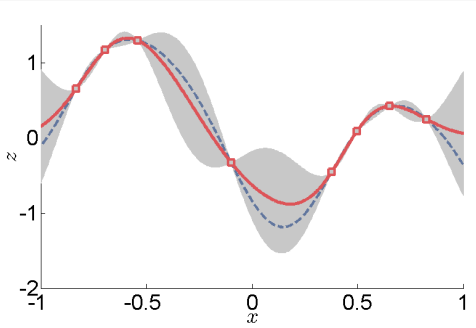
$$\hat{\beta} = \text{med}\left(\frac{y_j - y_i}{x_j - x_i}\right)$$



Gaussian process (kriging)

- Best linear unbiased predictor with squared-exponential covariance function

$$\mathrm{N}\left(\mu,\Sigma\right)\ ;\ \mu = K_{12}K_{11}^{-1}\mathbf{y}\ ;\ \Sigma = \mathrm{K}_{22} - K_{12}^T K_{11}^{-1}K_{12}$$



2ND LA SERENA DATA SCHOOL

Introduction to probability and statistics - II

Matthew J. Graham, Caltech

Aug 16, 2014

Confidence intervals

$$P(X_L < \theta < X_U) = 1 - \alpha$$

Jackknife

$$\theta_J = \theta_N + \Delta\theta$$

$$\Delta\theta = (N-1)\left(\theta_N - \frac{1}{N}\sum_{i=1}^{N}\theta^*\right)$$

Bootstrapping

$$f(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - x_i)$$

- Select *j* new values from data set $\{x_n\}$ *B* times with replacement
- Distribution of values gives confidence estimates

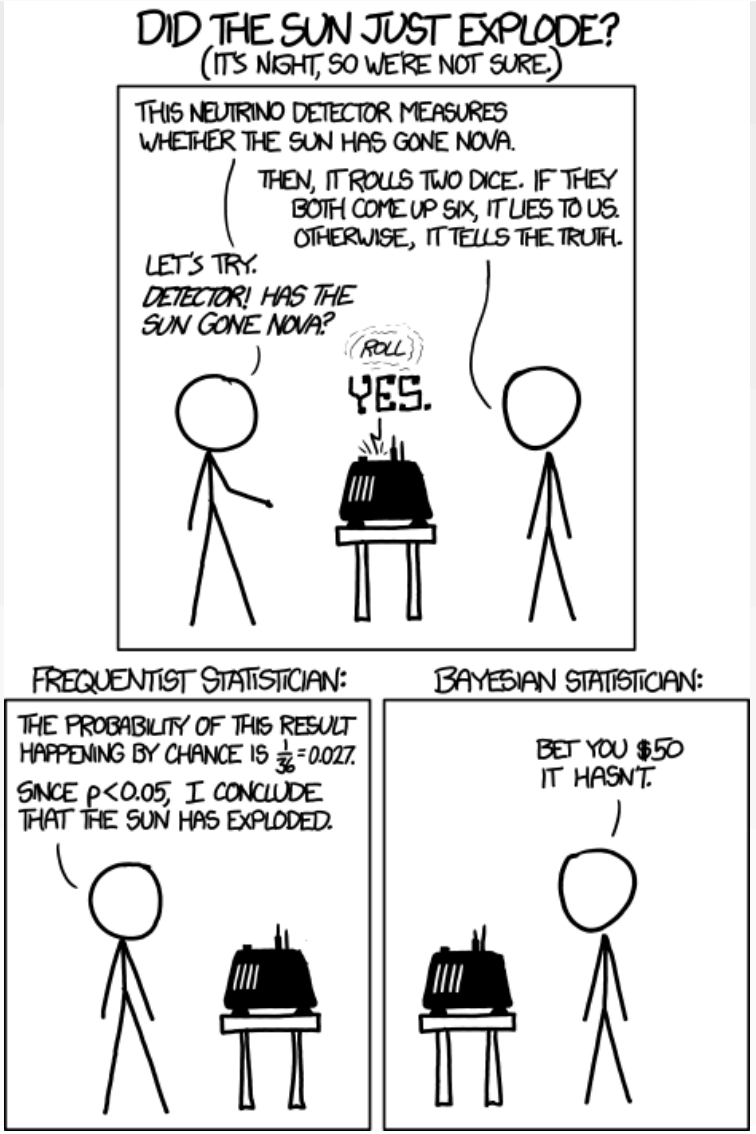
Frequentist vs. Bayesian

Frequentist:

- Relative frequencies of events
- A 95% confidence interval is the range in which the mean will occur 95% of the time with repeated sampling
- Only based on data

Bayesian:

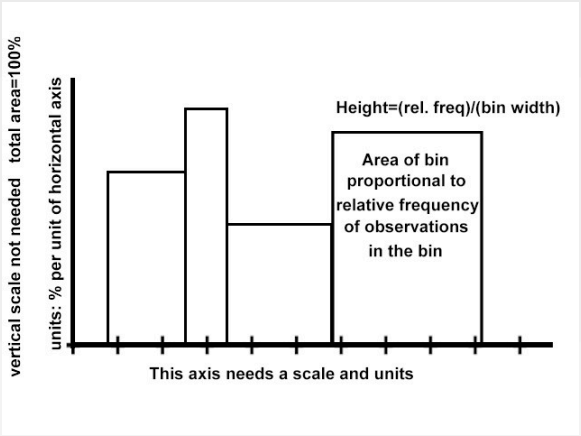
- Degree of subjective belief
- A 95% credible interval says that 95% of the population lies in that interval
- Incorporates information from prior



Histograms

Optimal bin width

$$\Delta_x = \frac{3.5\sigma}{n^{1/3}} \text{ or } \frac{2IQR}{n^{1/3}}$$



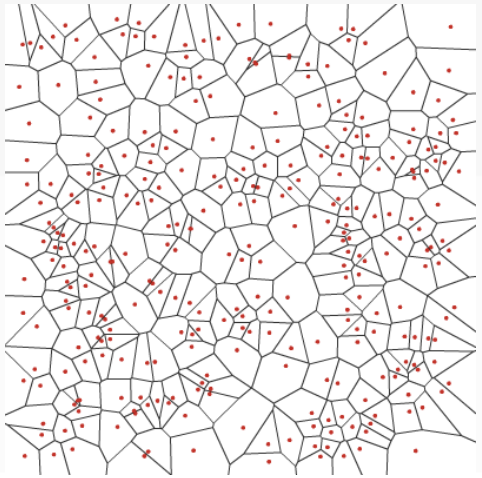
Uneven widths

- Minimum number of data points per bin: 5 or 11

Bayesian blocks

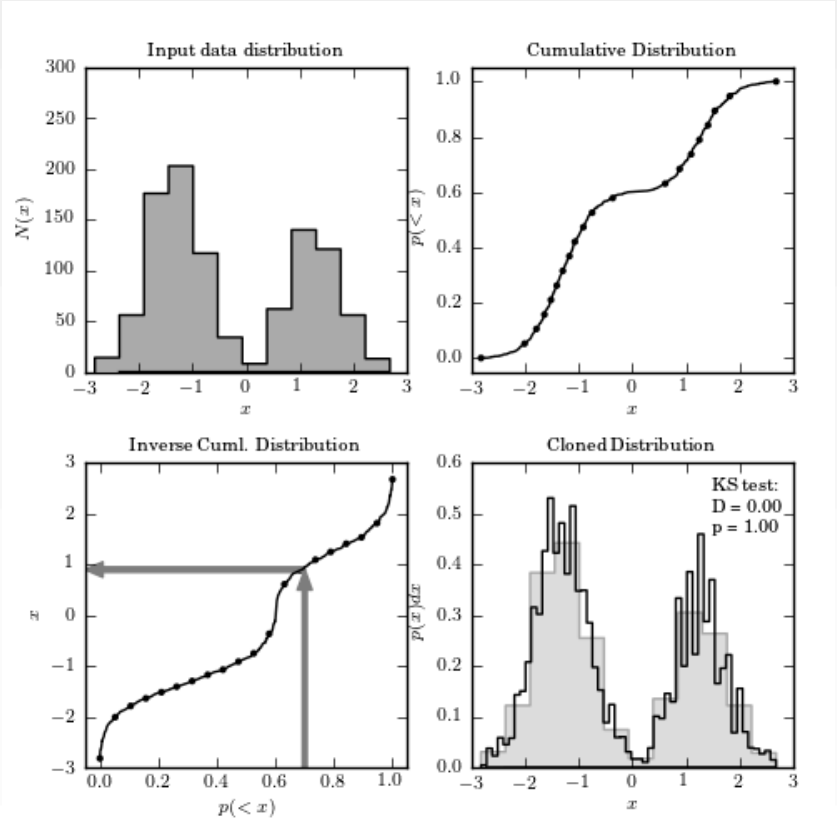
- For each block:

$$F(N_i, T_i) = N_i(\log N_i - \log T_i)$$



Sample an arbitrary 1D distribution

- Construct cumulative distribution function
- Invert and fit with spline
- Sample randomly over [0, 1]



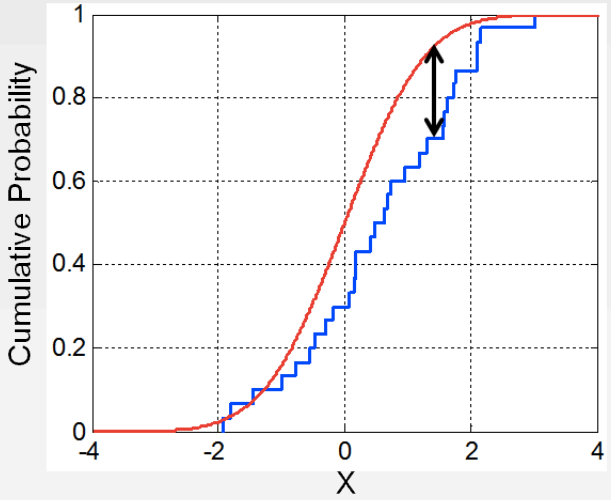
Goodness-of-fit

Chi-squared

$$\chi^2_{red} = \frac{1}{\nu}\sum\frac{(O-E)^2}{\sigma^2}$$

Kolmogorov-Smirnov

$$D_n = \sup_x \left|F_n(x) - F(x)\right|$$



Quadratic tests

$$n\int_{-\infty}^{\infty}\left(F_n(x)-F(x)\right)^2 w(x)\,dF(x)$$

$$w(x) = \frac{1}{F(x)(1-F(x))}$$ Anderson-Darling

Information criterion

$$-2\ln(L(M)) + 2k + \frac{2k(k+1)}{N-k-1}\ \text{(AIC)};\ -2\ln(L(M)) + k\ln N\ \text{(BIC)}$$

Method of moments

$$\mu_k(x) = \int x^k f(x) dx$$

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

Example

- Exponential distribution: $f(x) = \lambda \exp(-\lambda x)$

$$\mu_1(x) = \frac{1}{\lambda}$$

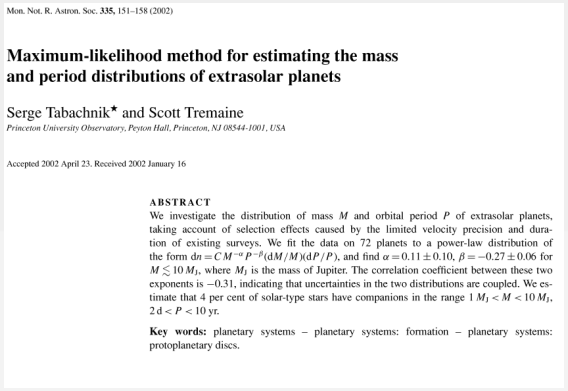
$$\hat{\mu}_1 = \bar{x}$$

Maximum Likelihood Estimation (MLE)

- The *likelihood* of the data set $\{x_i\}$ is:

$$L \equiv p(\{x_i\}|M(\theta)) = \prod_{i=1}^{n} p(x_i|M(\theta))$$

$$l \equiv \ln L = \prod_{i=1}^{n} \ln(p(x_i|M(\theta)))$$



Mon. Not. R. Astron. Soc. **335,** 151–158 (2002)

Maximum-likelihood method for estimating the mass and period distributions of extrasolar planets

Serge Tabachnik★ and Scott Tremaine
Princeton University Observatory, Peyton Hall, Princeton, NJ 08544-1001, USA

Accepted 2002 April 23. Received 2002 January 16

ABSTRACT

We investigate the distribution of mass M and orbital period P of extrasolar planets, taking account of selection effects caused by the limited velocity precision and duration of existing surveys. We fit the data on 72 planets to a power-law distribution of the form $dn = C\,M^{-\alpha}\,P^{-\beta}(dM/M)(dP/P)$, and find $\alpha = 0.11 \pm 0.10$, $\beta = -0.27 \pm 0.06$ for $M \lesssim 10\,M_J$, where M_J is the mass of Jupiter. The correlation coefficient between these two exponents is -0.31, indicating that uncertainties in the two distributions are coupled. We estimate that 4 per cent of solar-type stars have companions in the range $1\,M_J < M < 10\,M_J$, $2\,\mathrm{d} < P < 10\,\mathrm{yr}$.

Key words: planetary systems – planetary systems: formation – planetary systems: protoplanetary discs.

- The *maximum likelihood estimator* of a parameter θ is given by:

$$\frac{\partial l}{\partial \theta} = 0$$

- MLE is the most probable Bayesian estimator given a flat prior for θ

MLE example

- Consider the power law (Pareto) distribution:

$$p = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}} \text{ for } x \geq x_m$$

$$L = \prod_{i=1}^{n} \frac{\alpha x_m^{\alpha}}{x_i^{\alpha+1}} \; ; \; l = n \ln \alpha + n\alpha \ln x_m - (\alpha+1) \sum_{i=1}^{n} \ln x_i$$

- MLE:

$$\frac{\partial l}{\partial \alpha} = \frac{n}{\alpha} + n \ln x_m - \sum_{i=1}^{n} \ln x_i = 0$$

The Monty Hall problem

	A	B	C	Action	Stick	Switch
Case 1	Gold	Goat	Goat	B or C	WIN	LOSE
Case 2	Goat	Gold	Goat	C	LOSE	WIN
Case 3	Goat	Goat	Gold	B	LOSE	WIN

Bayes' Theorem

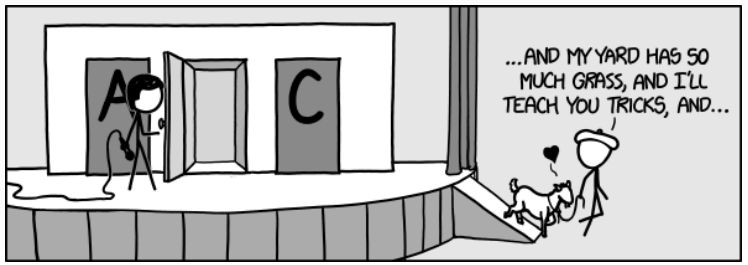
$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

$$p(\text{gold behind A}|\text{goat shown behind B}) =$$

$$\frac{p(\text{goat shown behind B}|\text{gold behind A})p(\text{gold behind A})}{p(\text{goat shown behind B})}$$

$$= \frac{1/2 \times 1/3}{1/2} = 1/3$$

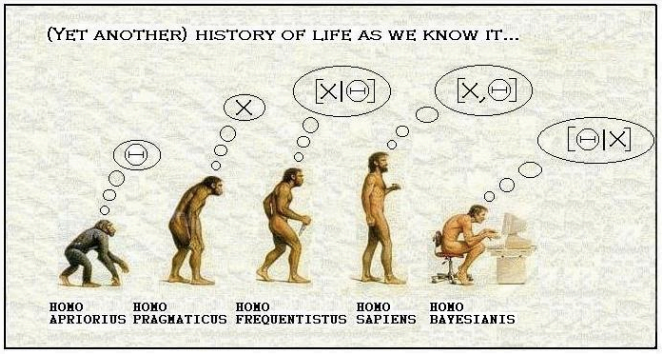
- Better to switch, unless…



Bayesian inference

$$p(\theta|x) = \frac{p(\theta)L(x|\theta)}{\sum_j p(\theta_j)L(x|\theta_j)} \text{ if } \theta \text{ is discrete}$$

$$= \frac{p(\theta)L(x|\theta)}{\int p(u)L(x|u)\,du} \text{ if } \theta \text{ is continuous}$$



- p(θ|x) is the posterior probability
- p(θ) is the prior
- p(u)L(x|u) is the marginal likelihood or evidence

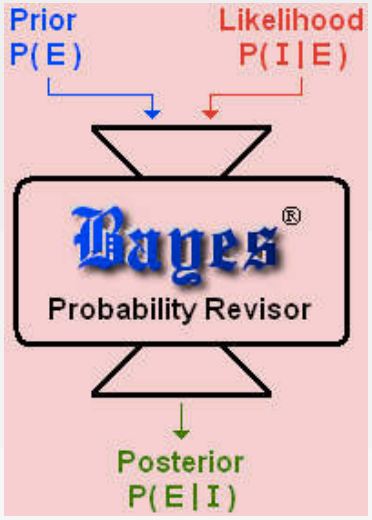
Choice of priors

Flat or non-informative prior

$$p(\theta) = const., a < \theta < b$$

Scale-invariant prior

$$p(\theta) \propto \theta^{-1} \therefore p(\ln\theta) \propto const.$$



Maximum entropy priors

- Known mean and variance: Gaussian
- Known mean: exponential

Data-based priors

- Gamma distribution gives significant probability to a broad range of values

Markov Chain Monte Carlo (MCMC)

$$I(\theta) = \int g(\theta)p(\theta)d\theta \approx \frac{V_\theta}{M}\sum_{j=1}^{M} g(\theta_j)p(\theta_j)$$

- A Markov chain is a sequence of possible states where the prob. of a given state at time t only depends on the state at time t-1
- With such a chain of length *M*:

$$I = \frac{1}{M}\sum_{j=1}^{M} g(\theta_j)$$

```
int getRandomNumber()
{
    return 4;  // chosen by fair dice roll.
               // guaranteed to be random.
}
```

- The transition probabilities must not change the distribution:

$$p(\theta) = \sum_{y} T(y,\theta)p(\theta)$$

- The chain must also be reversible (detailed balance condition):

$$p(\theta)T(\theta,\theta') = p(\theta')T(\theta',\theta)$$

MCMC - II

- Since $\sum_y T(\theta, y) = 1$: $\sum_y p(y)T(y,\theta) = p(x)$

Metropolis-Hastings algorithm

- A proposal distribution q(x) of an arbitrary form
- Sample x^* from $q(x_i | x_{i-1})$
- Sample *u* from the uniform distribution
- If $u < \min\left(1, \frac{p(x^*)q(x_i|x^*)}{p(x_i)q(x^*|x_i)}\right)$, $x_{i+1} = x^*$

otherwise $x_{i+1} = x_i$

- Repeat until …

"I'd like to meet the algorithm that thought we'd be a good match."

Practical – I (see Chris Miller's session)

- Andreon & Hurn (2010), MNRAS, 404, 1922
- Estimate (log) mass of a galaxy cluster from measurements via caustic analysis involving distances and velocities

$$p(lgM|obslgM200) = \frac{p(obslgM200|lgM)p(lgM)}{p(obslgM200)}$$

$$obslgM200 \sim N(lgM200, \sigma_i^2)$$

$$lgM200 \sim N(\alpha + 14.5 + \beta(\log(n200) - 1.5, \sigma_{scat}^2)$$

$$\alpha \sim N(0.0, 10^4), \beta \sim t_1$$

$$n200 \sim U(0,\infty), nbkg \sim U(0,\infty)$$

$$1/\sigma_i^2 \sim \Gamma(\varepsilon,\varepsilon), \quad 1/\sigma_{scat}^2 \sim \Gamma(\varepsilon,\varepsilon)$$

Practical - II

Image acknowledgements

- Wikipedia
- xkcd
- SMBC
- New Yorker