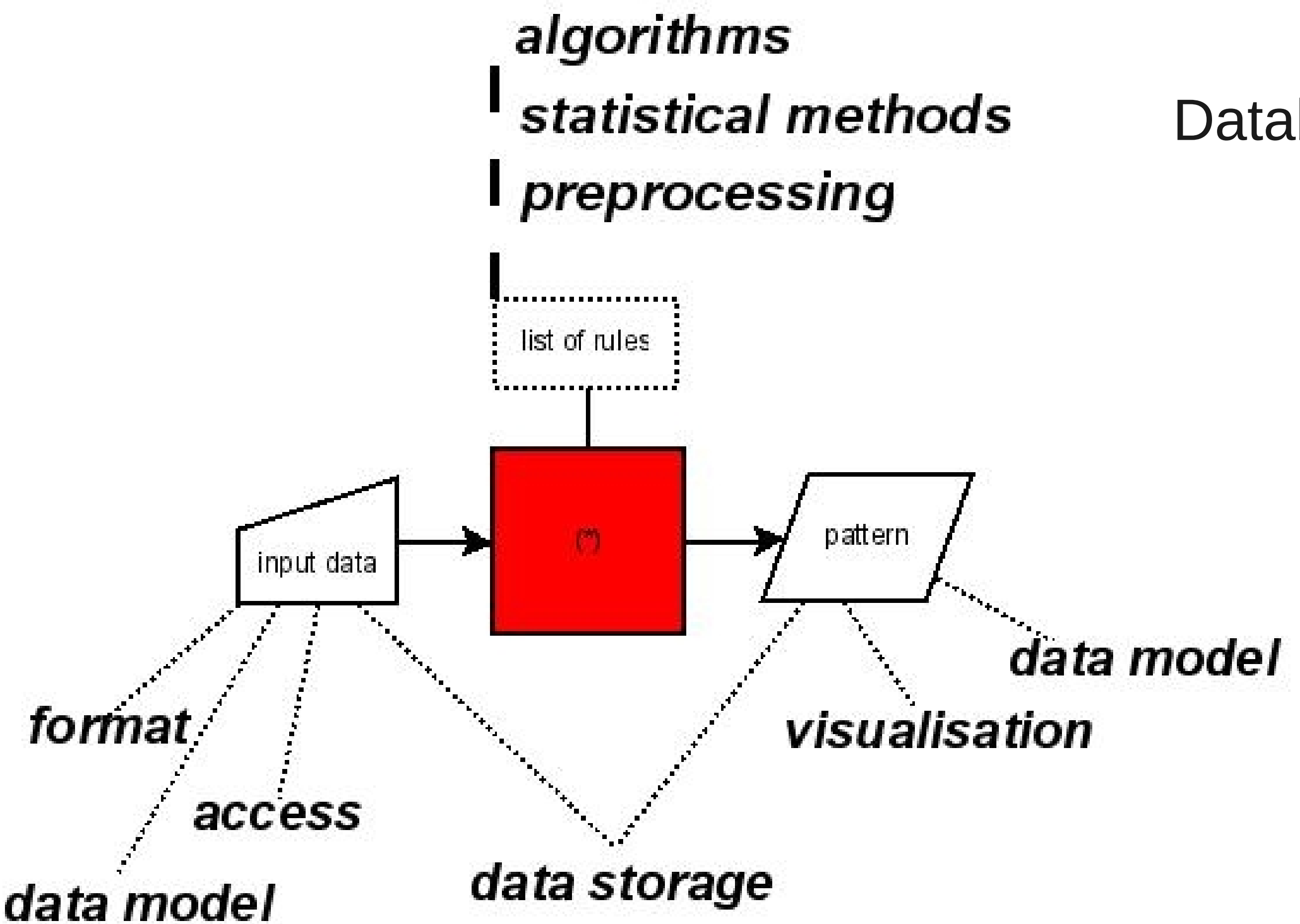
Virtual Observations 2012

Statistics in Data Mining

Data Mining



Database: Data format
Access
Data model
Preprocessing
Processing

Introduction to statistics

- List of rules: verification of model or parametric search
- Verification of model: parametric and non-parametric
- PDF
- Bayesian statistic
- Tests

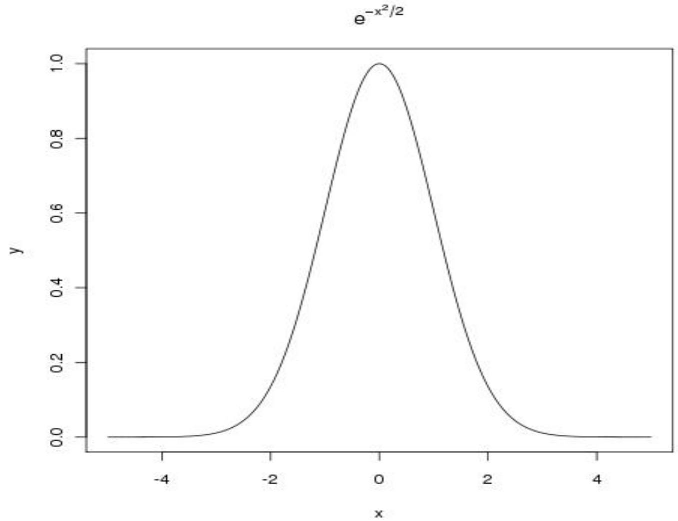
PDF, CDF

$$dP = f(x)dx$$

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)\,dx$$

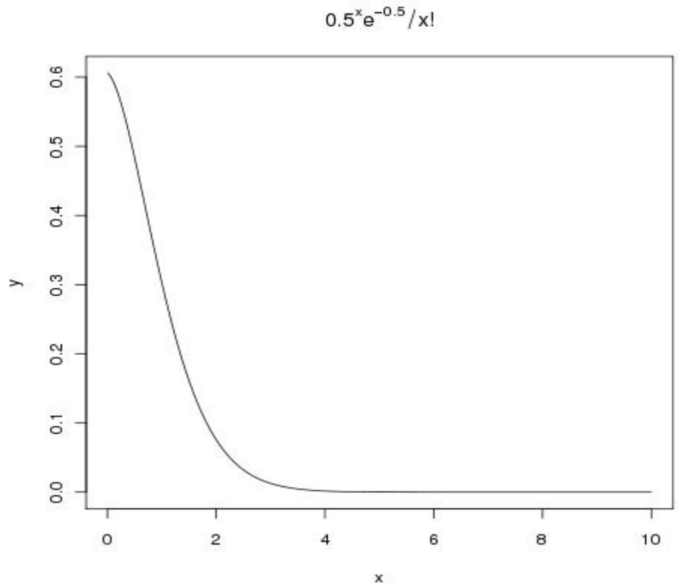
$$P(x < x_1) = \int_{-\infty}^{x_1} f(x)\,dx$$

PDF



$$f(x) = \frac{1}{\sqrt{2\,\pi\sigma^2}} e^{-\frac{(x-x_0)^2}{2\,\sigma^2}}$$

$$f(x) = \lambda^x e^{-x} / x!$$



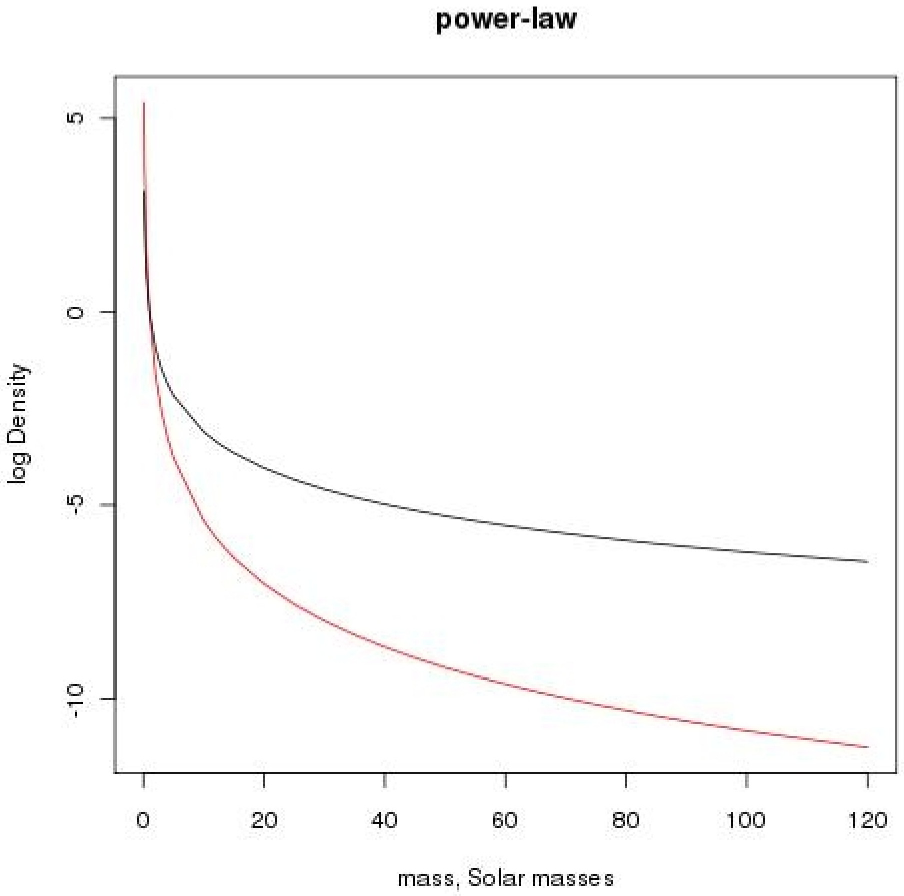
- Gauss
- Poisson

Central Limit Theorem

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = N(\bar{x}, \sigma^2/n)$$

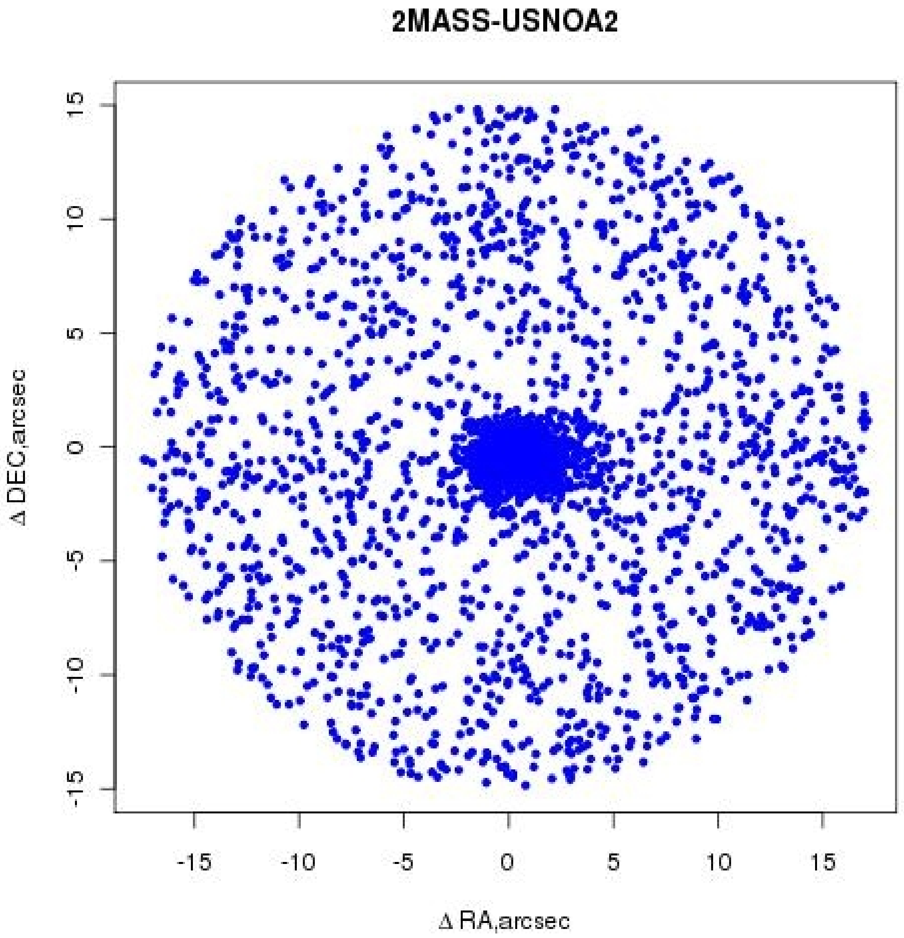
$$\lim_{n\to\infty} \frac{1/n \sum_{i=1}^{n} X_i - \bar{x}}{\sigma/\sqrt{n}} = N(0,1)$$

Power-law distribution



$$dN = m^{-\alpha}\, dm$$

2MASS-USNO



- Cross-identify 2MASS and USNO-A2
- Select coordinate differences
- Find mean, median, variance, MAD

Mean, median

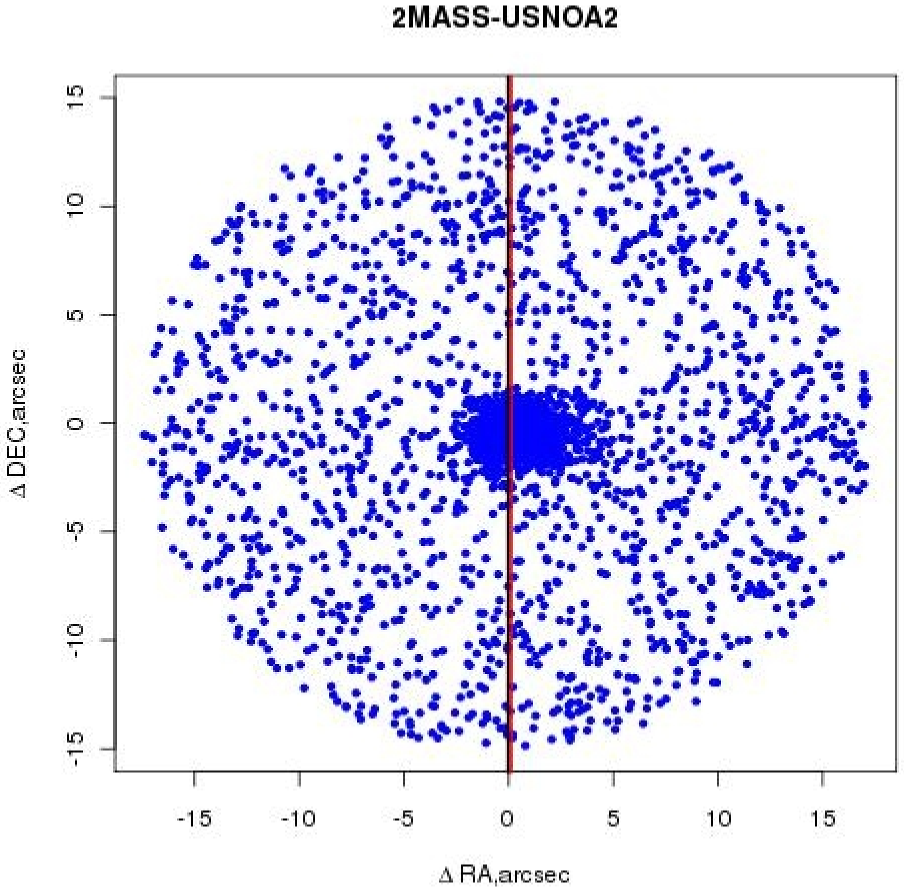
$$\bar{x} = E(x) = \int_{-\infty}^{\infty} f(x)\, x\, dx$$

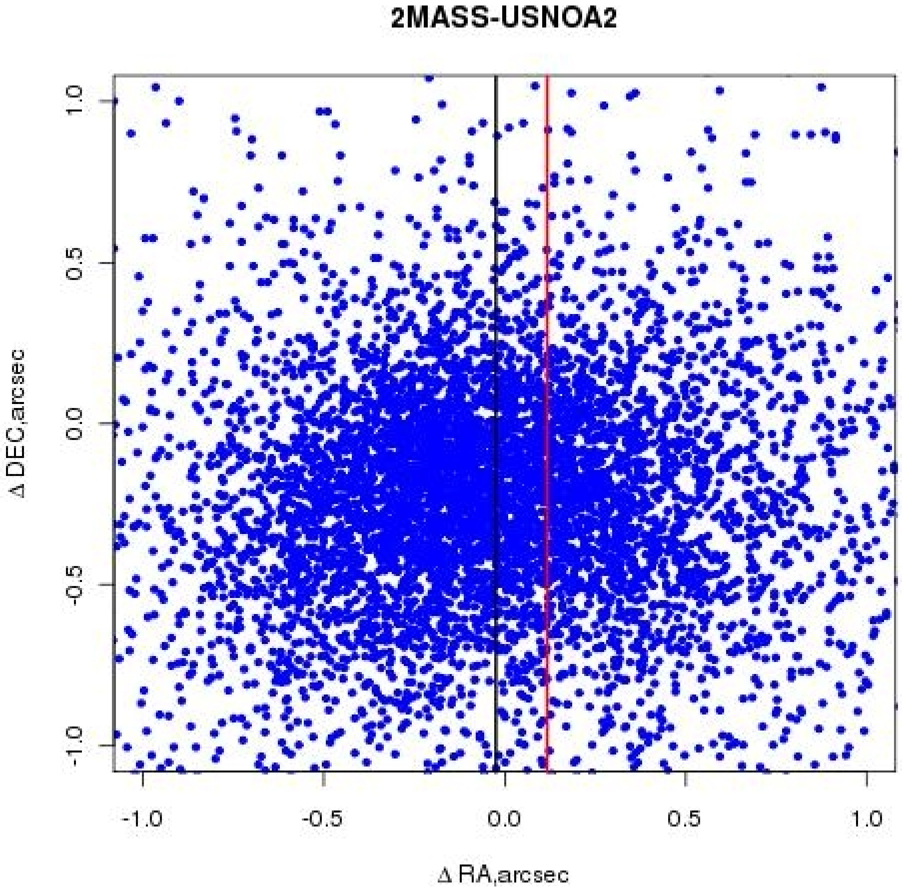
$$\bar{x} = \sum_i f(x_i)\, x_i$$

$$median(x) = \tilde{x} : P(x < \tilde{x})) = \int_{-\infty}^{\tilde{x}} f(x)\, dx = 1/2$$

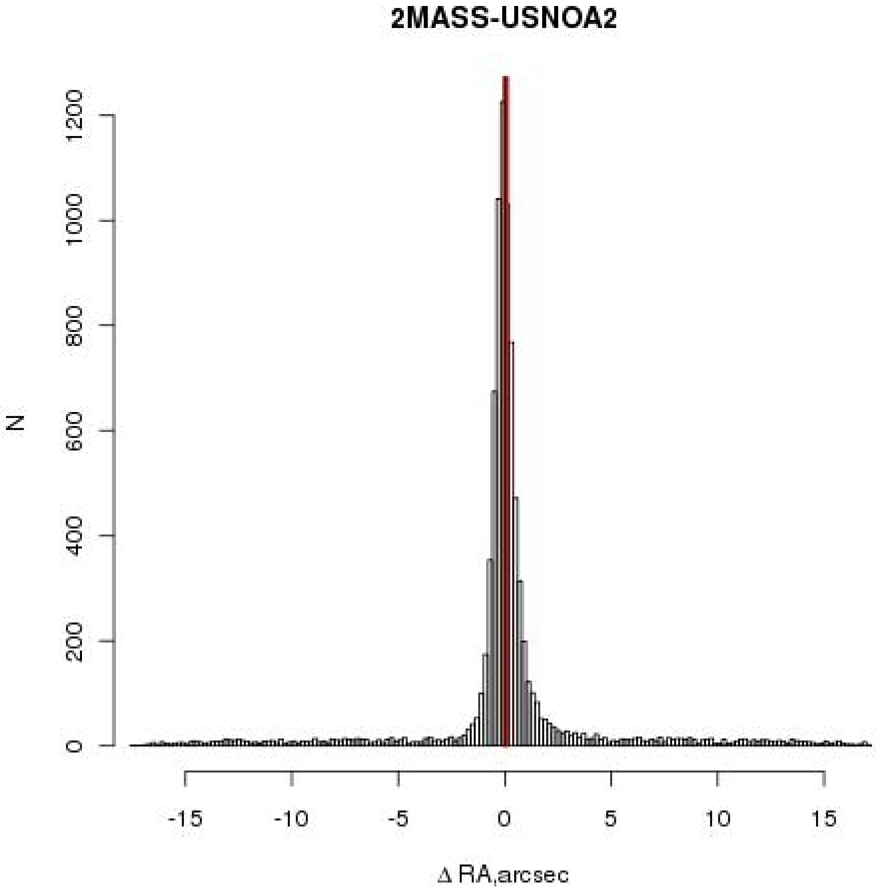
$$P(x > \tilde{x}) = \int_{\tilde{x}}^{\infty} f(x)\, dx = 1/2$$

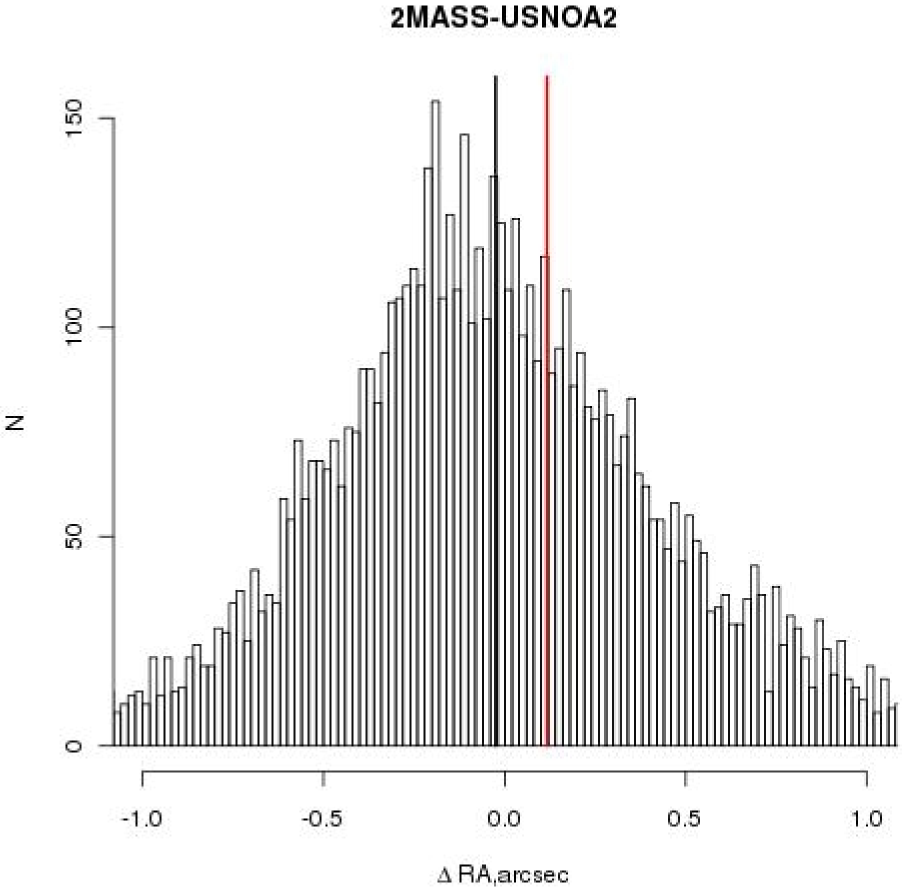
Mean, median


2MASS-USNOA2
Δ DEC,arcsec
Δ RA,arcsec
15
10
5
0
-5
-10
-15


2MASS-USNOA2
Δ DEC,arcsec
Δ RA,arcsec
1.0
0.5
0.0
-0.5
-1.0

Mean, median


2MASS-USNOA2
N
0
200
400
600
800
1000
1200
-15
-10
-5
0
5
10
15
Δ RA,arcsec


2MASS-USNOA2
N
0
50
100
150
-1.0
-0.5
0.0
0.5
1.0
Δ RA,arcsec

Deviation, Variance

$$\sigma^2 = \int (x - \bar{x})^2 \, f(x) \, dx$$

$$\sigma^2 = \overline{x^2} - \bar{x}^2$$

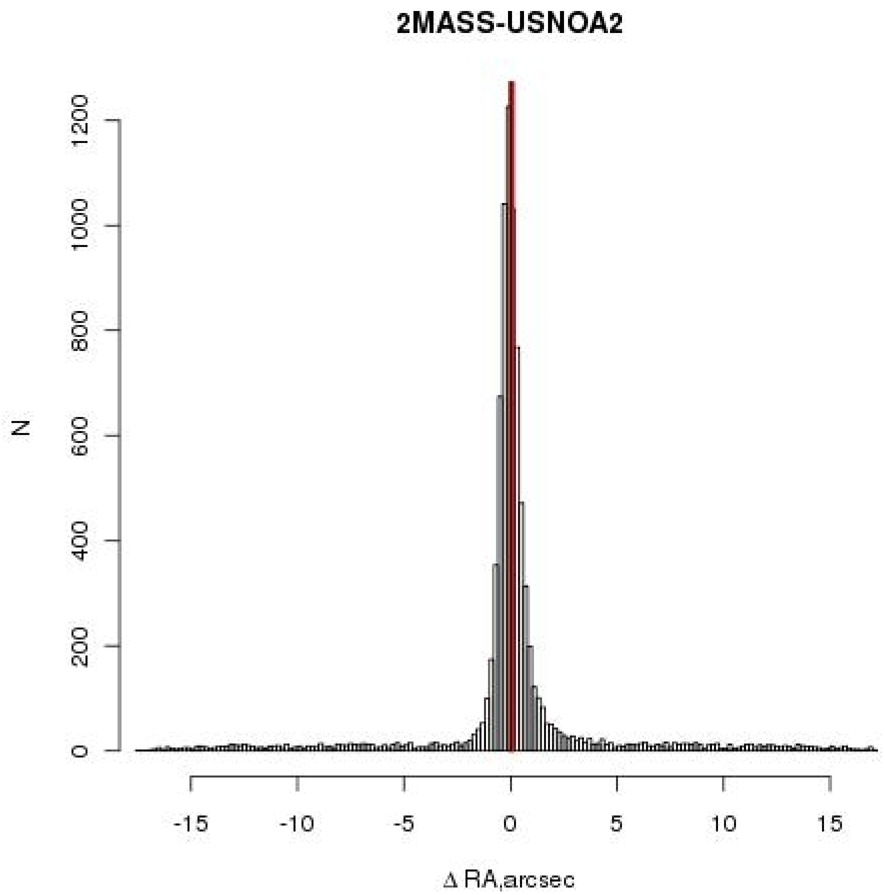
$$\mathrm{Var(x)} = E(x - \bar{x})^2 = \int (x - \bar{x})^2 \, f(x) \, dx = \sigma^2$$

RMS

$$x^2_{RMS} = \int x^2\, f(x)dx$$

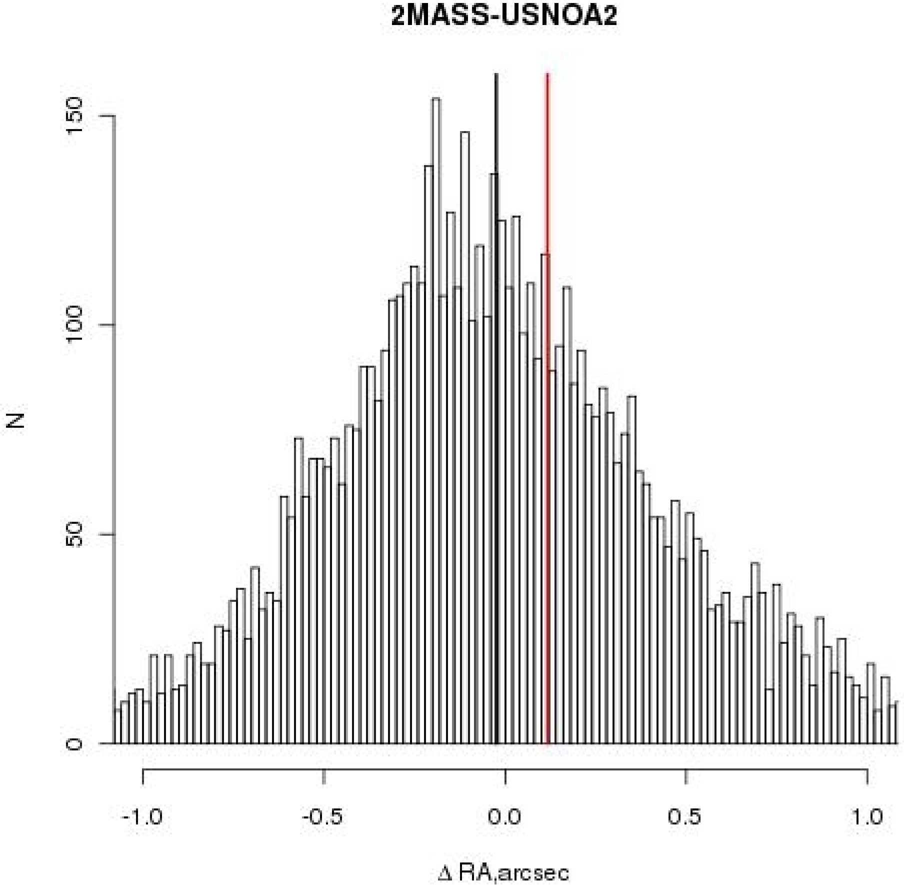
$$x^2_{RMS} = \overline{x^2} = \sigma^2 + \bar{x}^2$$

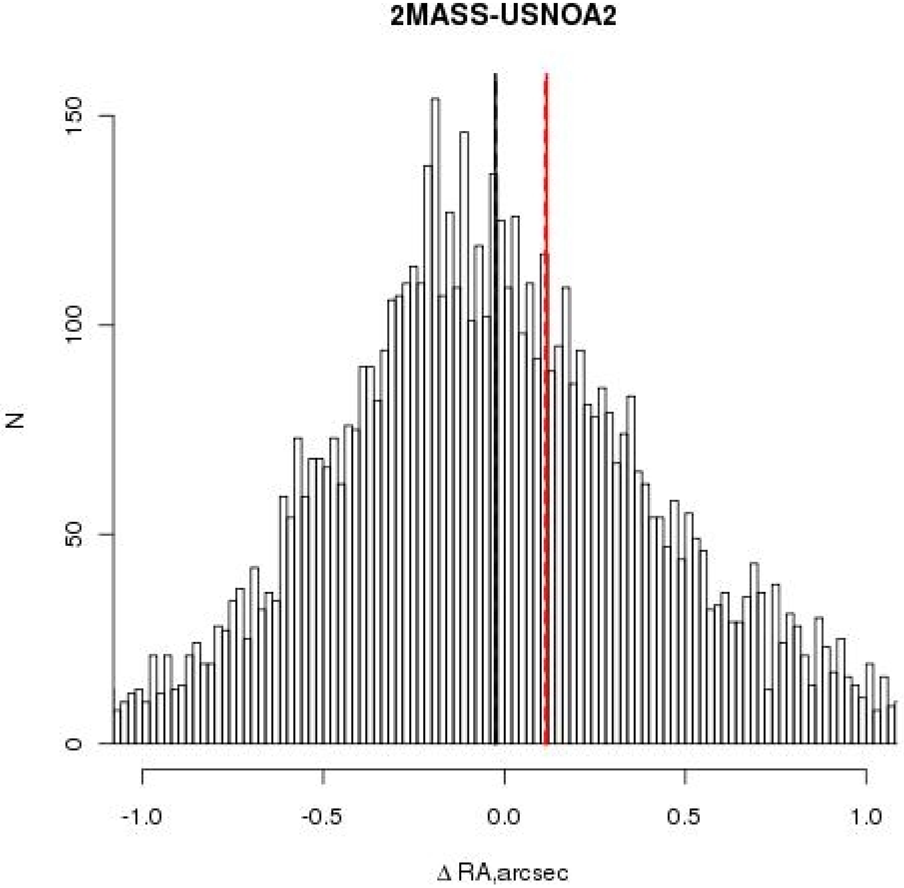
K-sigma clipping



$$|x_i - \bar{x}| > k\,\sigma$$

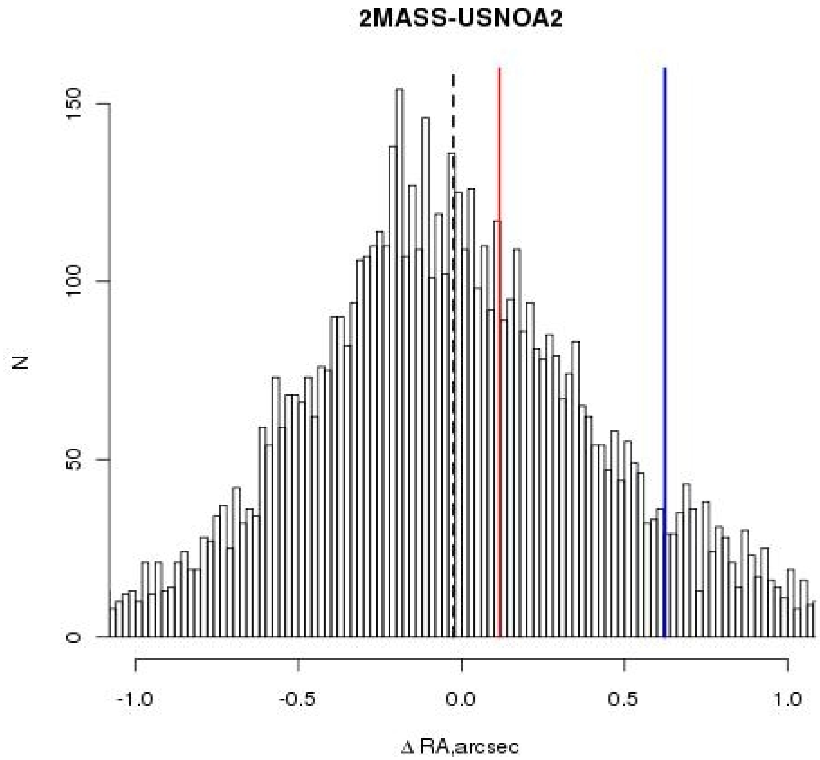
K-sigma clipping





MAD

$$\mathrm{MAD(x)} = \int_{-\infty}^{\infty} (x - \tilde{x})\, f(x)\, dx$$



SQL

- MySQL:
  - AVG()   - mean
  - COUNT() - N
  - STD(), STDDEV(), STDDEV_POP()
  - STDDEV_SAMP()
  - VARIANCE(), VAR_POP()
  - VAR_SAMP()

Distributions to test

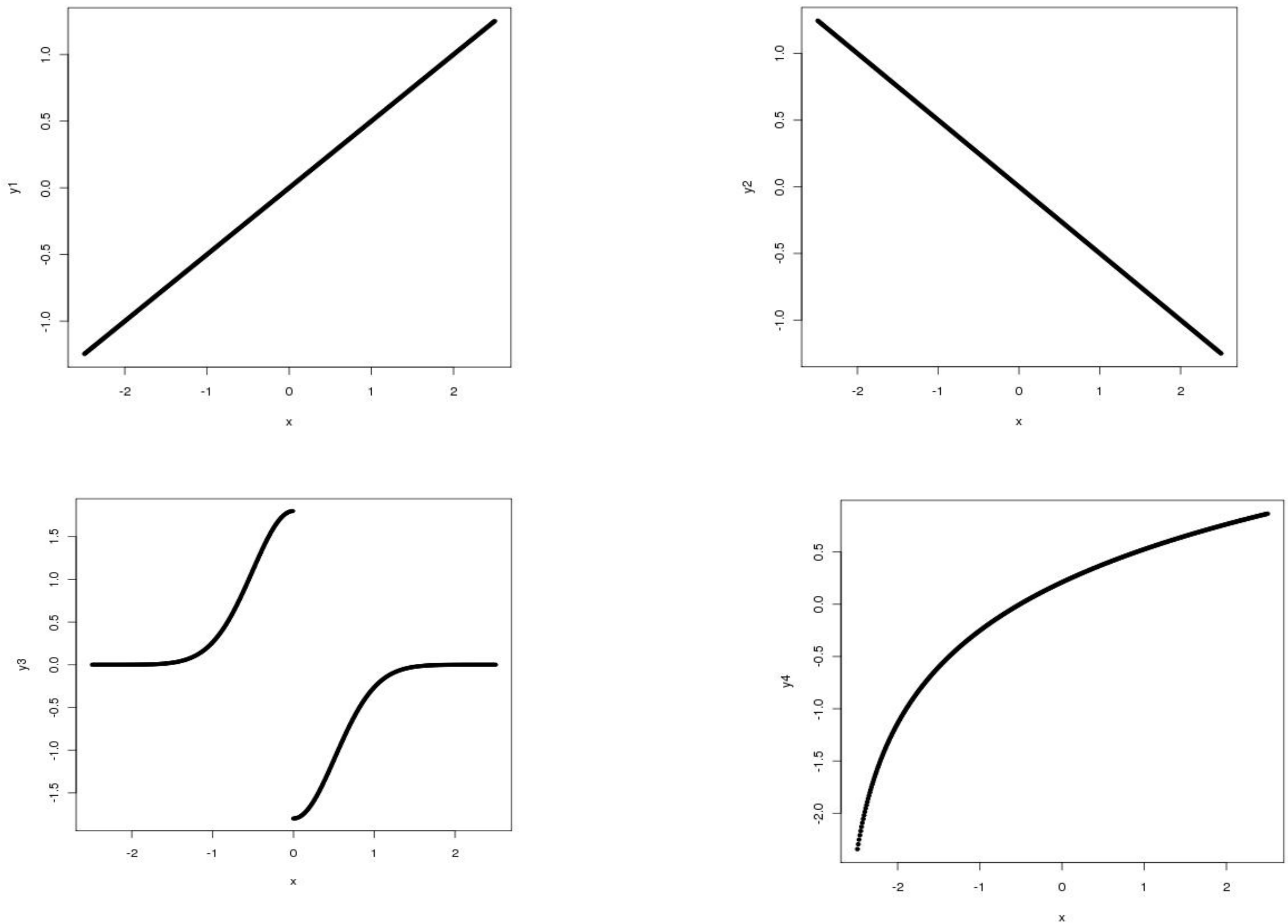
0.025
0.52

-0.025
0.52

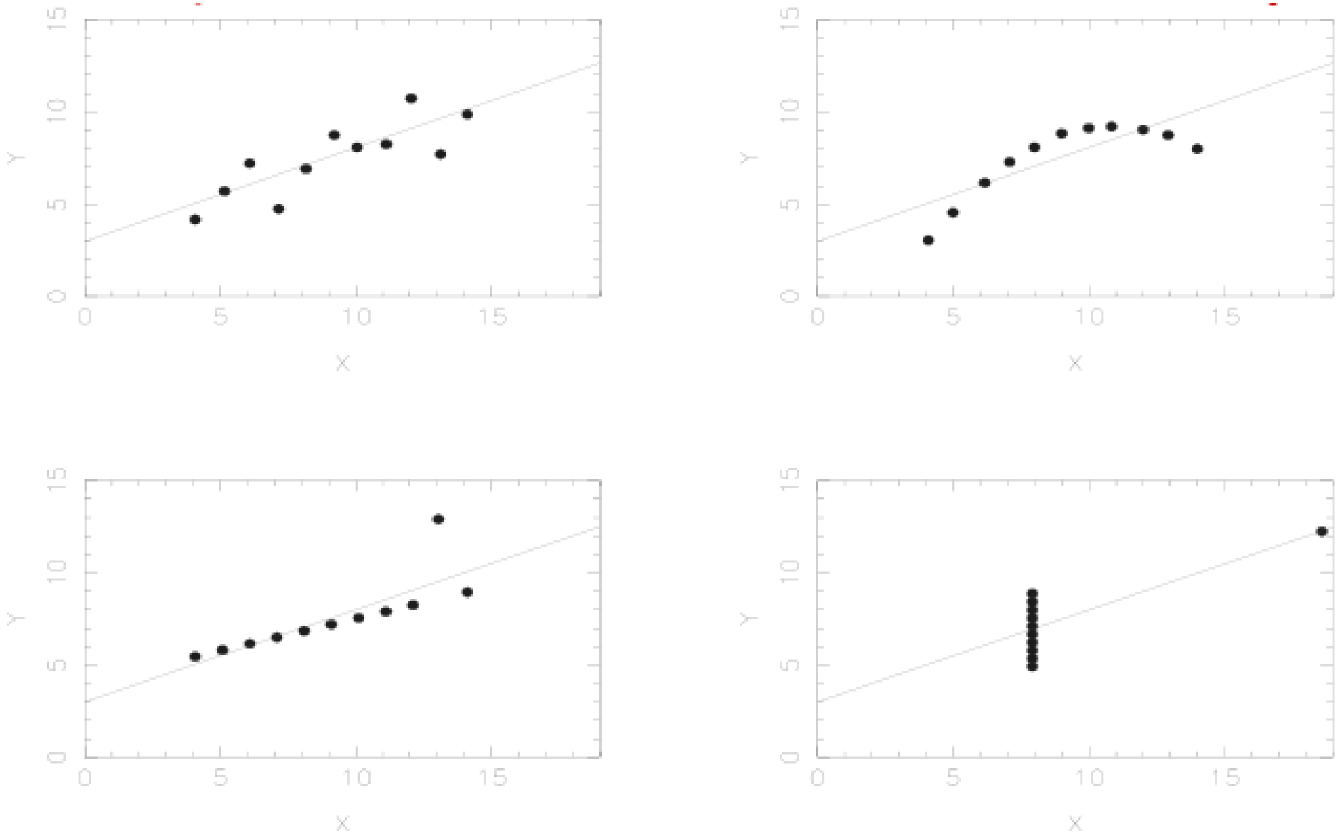
-0.003
0.52

0.0003
0.52

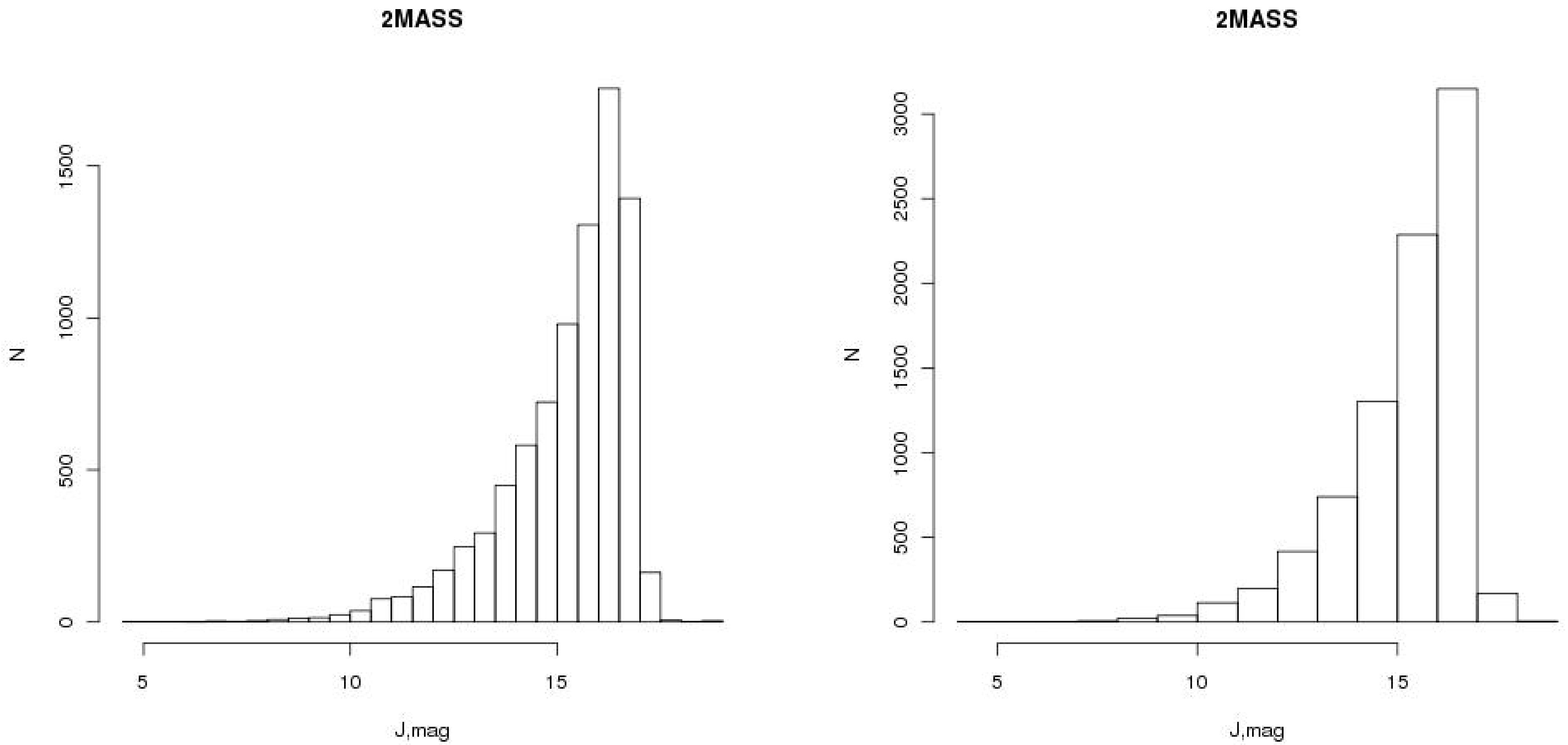
Real situation



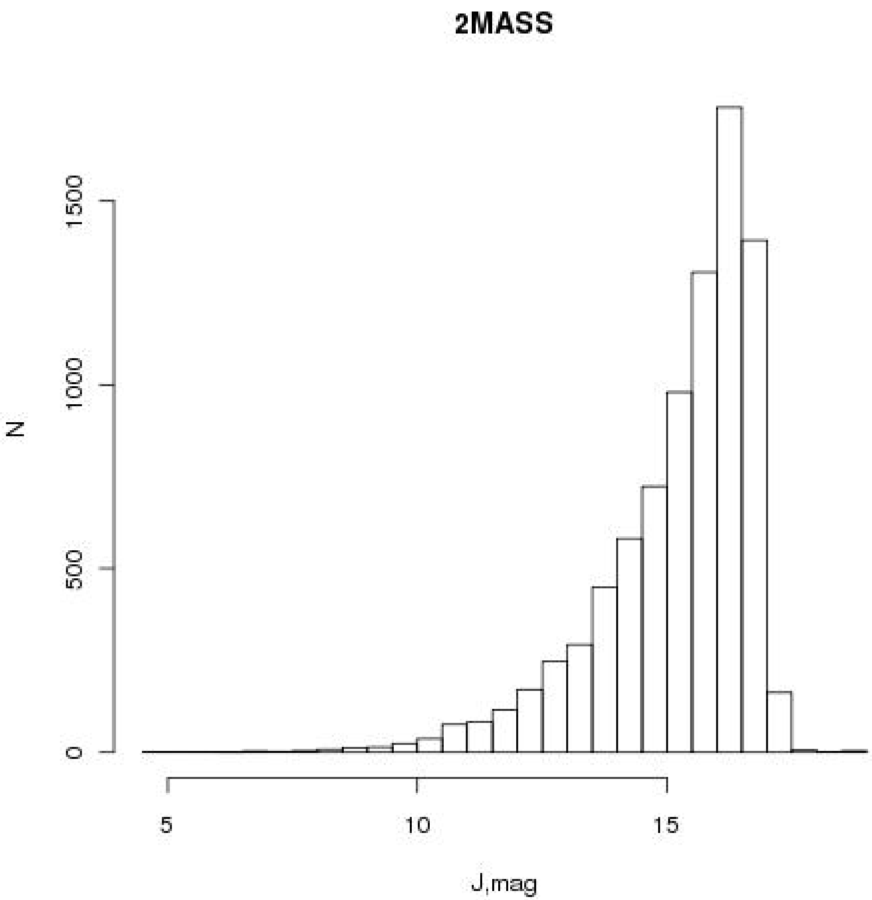
Anscombe quartet

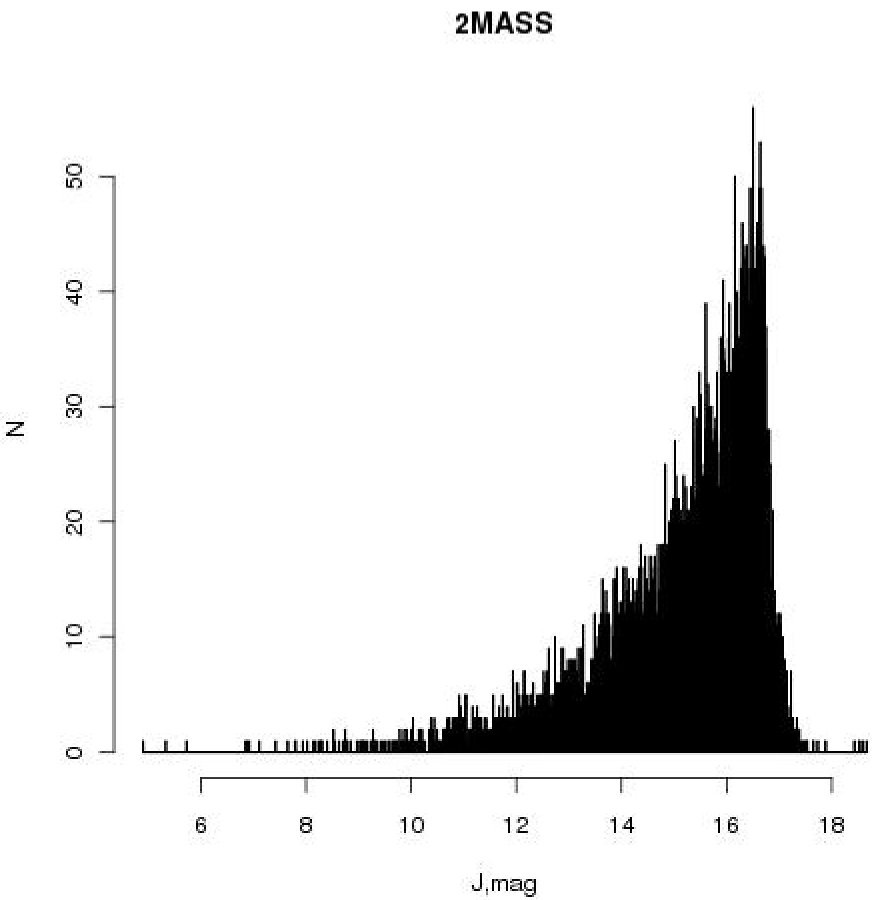


Histogram


2MASS
N
J,mag
0
500
1000
1500
5
10
15
2MASS
N
J,mag
0
500
1000
1500
2000
2500
3000
5
10
15

Histogram


2MASS
N
0
500
1000
1500
5
10
15
J,mag


2MASS
N
0
10
20
30
40
50
6
8
10
12
14
16
18
J,mag

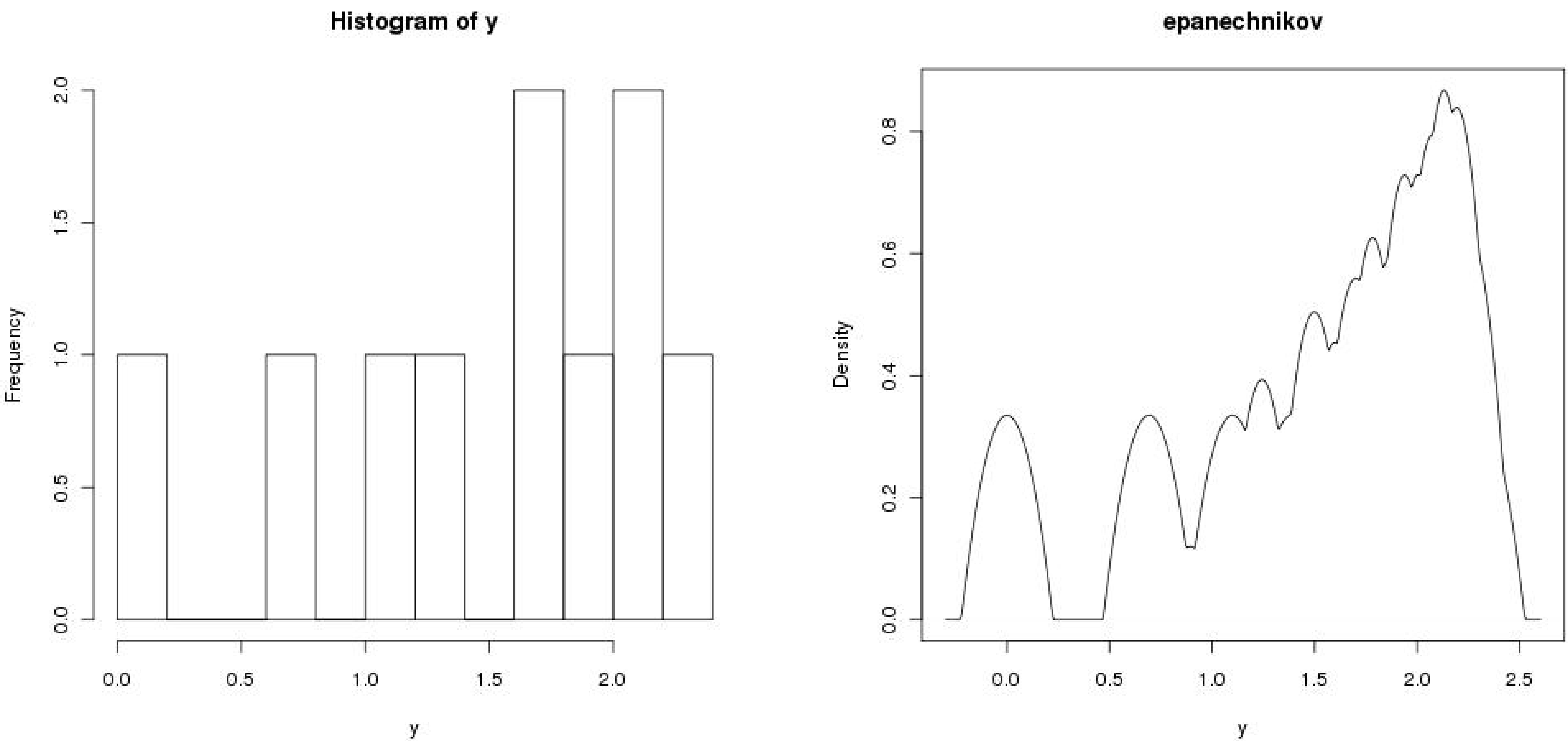
Histogram

- Biased
- Oversmoothed/Undersmoothed

$$p_i = \frac{n_i}{\Delta_i \sum_i n_i}$$

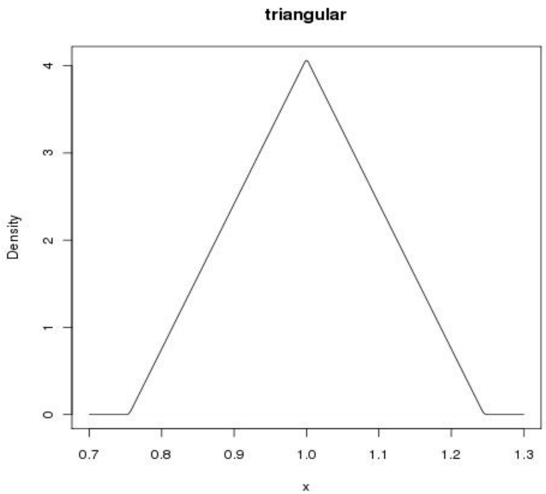
$$h_{opt} = \frac{3.5\,\sigma}{n^{1/3}}$$

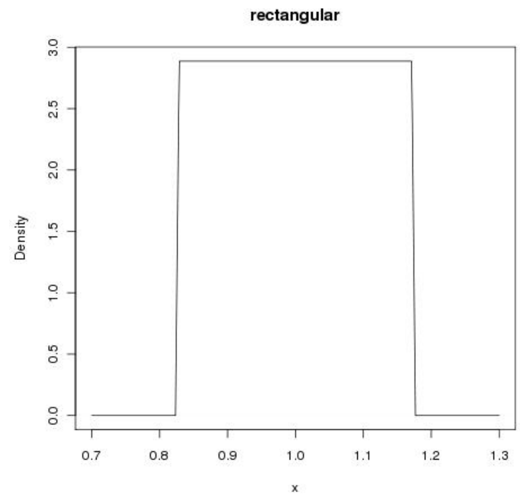
Kernel smoothing

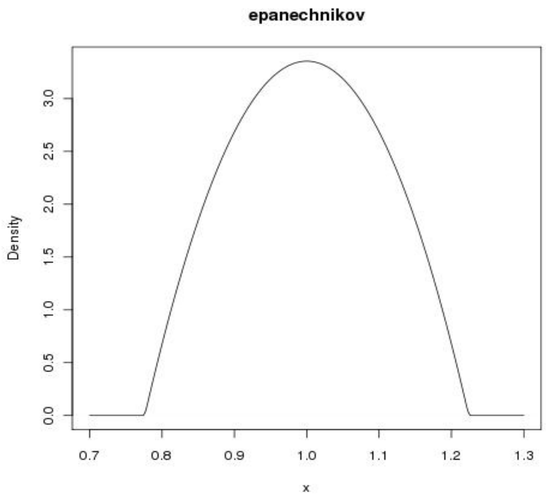


$$\hat{f}(x;h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

Kernel smoothing



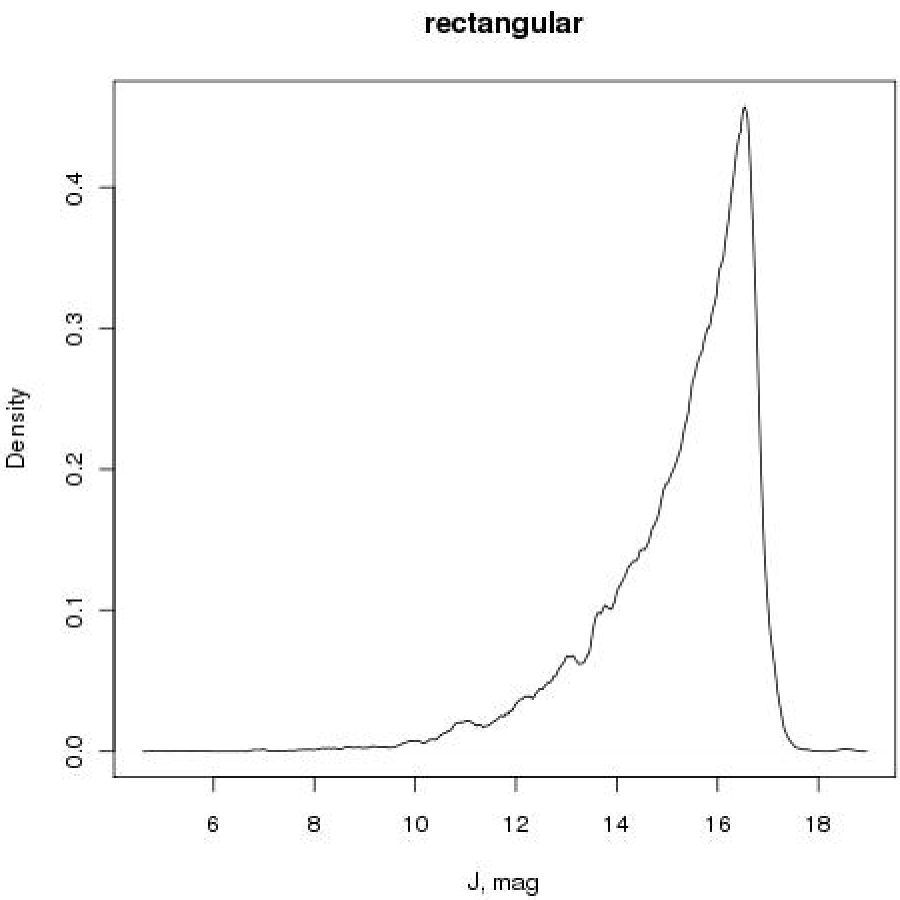


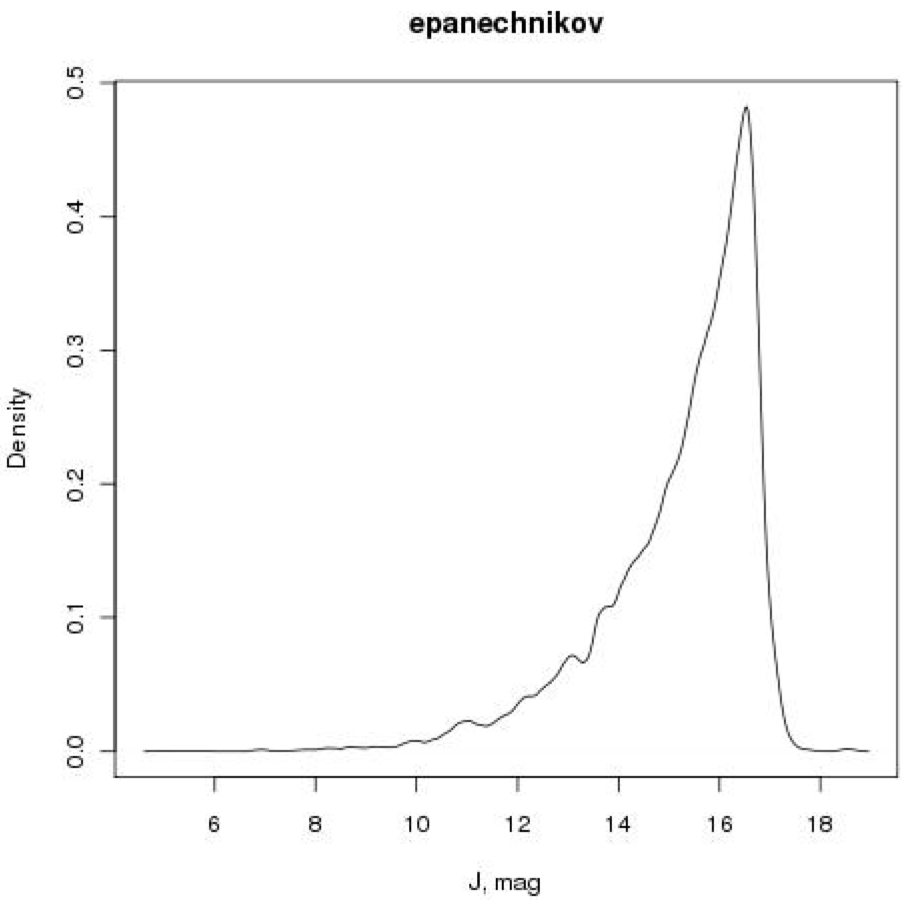


Coordinates are shifted on 0.5 on figures!

$$K(x) = 1/2, |x| < 1 \qquad K(x) = (1 - |x|), |x| < 1 \qquad K(x) = \frac{3}{4}(1 - x^2), |x| < 1$$

Kernel smoothing





Optimal bandwidth

$$h_{opt} = [\frac{R(K)}{\mu_2(K)^2\, R(\ddot{f})n}]^{1/5}$$

$$\mu_2(K) = \int x^2\, K(x)\, dx$$

$$R(f) = \int f^2(x)\, dx$$

Bayesian statistic

- $P(B|A)=P(A|B)P(B)/P(A)$
  - $P(A)$ – normalization
  - $P(B)$ – prior probability
  - $P(A|B)$ – likelihood
  - $P(B|A)$ – posterior probability

Maximum likelihood

$$f(p_k|p_1,\ldots,p_{k-1},p_{k+1},\ldots,p_n) = \int_{p_1,\ldots,p_{k-1},p_{k+1},\ldots,p_n} f(\vec{P})dp_1,\ldots,dp_{k-1},dp_{k+1},\ldots,dp_n$$

$$L = \prod_{i=1}^{n} f(p_i)$$

$$\chi_k^2 = -2\,\ln\frac{L_{-k}}{L}$$

Tests

$$\chi^2 = \sum_i \frac{(O_i - T_i)^2}{T_i}$$

$$\chi^2 < \chi^2(\alpha, n - k - 1)$$

Step-by-step

- Data sample
- Mean, median, variance, rms
- Visualisation
- Model
- Test

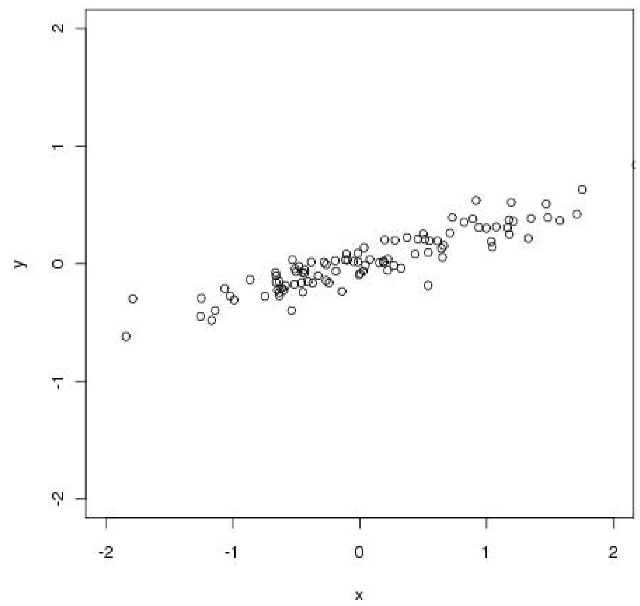
Covariance

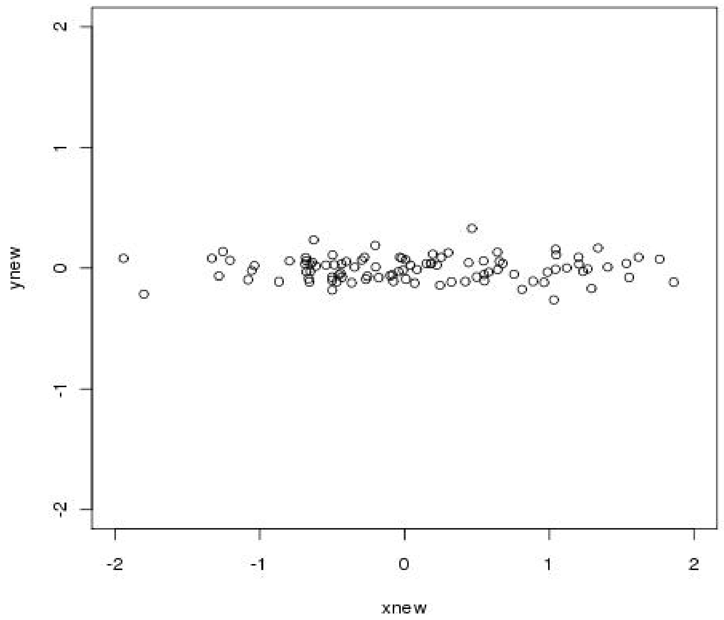
$$X = \{X_1, X_2, X_3, ...X_n\}$$

$$\bar{X} = \{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n\}$$

$$cov(X_i, X_j) = E((X_i - \bar{X}_i)(X_j - \bar{X}_j))$$

PCA





$$cov(X, Y) = C$$

$$TCT^{-1} = \Lambda$$

Lutz-Kelker effect

$$g(\pi_0 \mid \pi) = \frac{1}{\sqrt{(2\,\mathrm{pi})}\sigma} \exp\left(- \frac{(\pi_0 - \pi)^2}{2\sigma^2} \right)$$

$$g(\pi \mid \pi_0) \propto \frac{1}{\pi^4} \exp\left(- \frac{(\pi - \pi_0)^2}{2\sigma^2} \right)$$

$$Z \equiv \frac{\pi}{\pi_0}$$

$$g(\pi \mid \pi_0) \propto G(Z) \equiv \frac{1}{Z^4} \exp\left(- \frac{(Z-1)^2}{2(\sigma/\pi_0)^2} \right)$$

$$\Delta M = M_{\mathrm{true}} - M_{\mathrm{observed}} = 5 \log \frac{\pi}{\pi_0} = 5 \log Z$$

$$\langle \Delta M \rangle = \frac{5 \int_0^\infty \log Z\, G(Z) dZ}{\int_0^\infty G(Z)\, dZ}$$