Robust statistics

and some non-parametric statistics

Stéphane Paltani

ISDC Data Center for Astrophysics
Astronomical Observatory of the University of Geneva

Statistics Course for Astrophysicists, 2010–2011

Outline

Caveat

The presentation aims at being user-focused and at presenting usable recipes

Do not expect a fully mathematically rigorous description!

This has been prepared in the hope to be useful even in the stand-alone version

Please provide me feedback with any misconception or error you may find, and with any topic not addressed here but which should be included

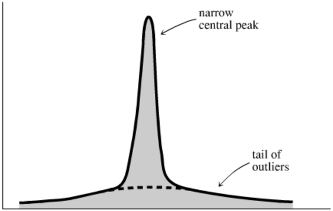
Outline

Example: The average weight of opossums

- Find a group of N opossums
- Weight them: $w_i,\ i = 1, \ldots, N$
- Calculate the average weight as $1/N\ \sum_i w_i$
- You should get about 3.5 kg
- But you better not use this group!

Outline

General concepts



Robust statistics:

- are not (less) affected by the presence of outliers or deviations from model assumptions
- are related, but not identical to non-parametric statistics, where we drop the hypothesis of underlying Gaussian distribution

Main approaches

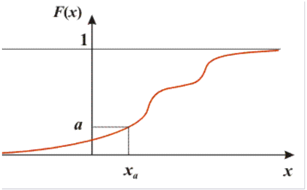
Four main approaches:

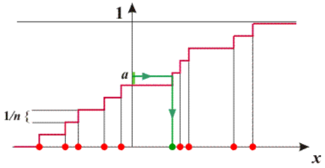
- Discard data
- L-estimators: Use linear combinations of order statistics
- R-estimators: Use rank instead of values
- M-estimators: Estimators based on Maximum-likelihood argument

The breakdown point is the fraction of outliers above which the robust statistics fails. It can reach 50 %

Outline

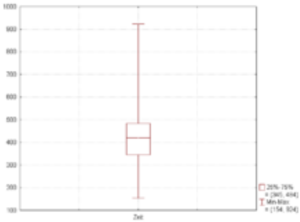
Quantiles





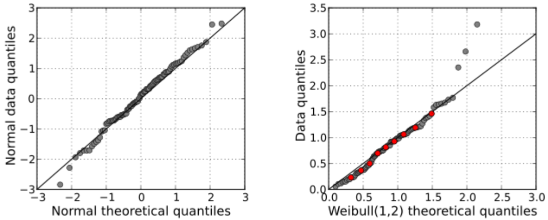
- q-quantiles ($q \in \mathbb{N}$) are the set of value $x_{[a]}, \ a = k/q, k = 1, \ldots, q-1$ from a probability distribution for which $P(x < x_{[a]}) = a$
- 4-quantiles, or quartiles are then $x_{[0.25]}, \ x_{[0.5]}, \ x_{[0.75]}$
- q-quantiles from a sorted sample $\{x_i\}, i = 1, \ldots, N$: $x_{[a]} = \min_i x_i \mid i/N \geq a$
- Example: if $N = 12$, quartiles are: x_3, x_6, x_9; if $N = 14$, quartiles are: x_4, x_7, x_{11}

Box plot



- A box plot is a synthetic way to look at the sample properties
- It represents in a single plot, the minimum, the quartiles (hence also the median) and the maximum
- that is, the quantiles $x_{[0]}$, $x_{[0.25]}$, $x_{[0.5]}$, $x_{[0.75]}$, $x_{[1.0]}$

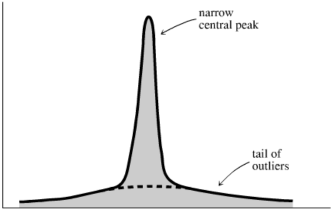
Q-Q plot



- A Q-Q plot (quantile-quantile plot) is powerful way to compare sample properties with an assumed underlying distribution
- The QQ-plot is the plot of x_i vs f_i, where f_i are the values for which $P(x < f_i) = i/(N+1)$, when x_i is sorted in increasing order
- One usually also draw f_i vs f_i for comparison with expected values
- Outliers appear quite clearly

Outline

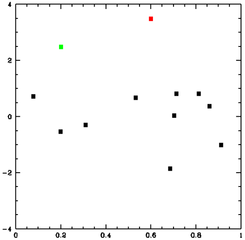
What are these outliers?



- They can be due to errors in the measurement
- They may be *bona fide* measurements, but the distribution is heavy-tailed, so these points may be the black swan and contain important information
- Outliers can be difficult to identify with confidence, and they can hide each other
- There is the risk of data manipulation

Outline

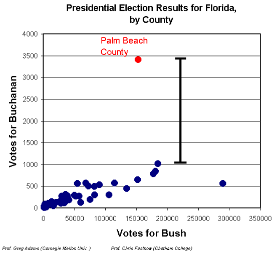
Chauvenet's criterion



- Assuming a data set $\{x_i\} \sim \mathcal{N}(\mu, \sigma),\ i = 1, \ldots, N$
- Calculate for each i $P(x_i) \equiv P(|x| > |x_i|)$
- Discard the point if $N \cdot P(x_i) < 0.5$, i.e. if the probability to have such an extreme value is less than 50%, taking the number of trials into account
- Moderate outliers may mask more extreme outliers
- Grubbs's test uses absolute maximum deviation $G = \max_i \frac{|x_i - \mu|}{\sigma}$

Outline

Dixon's Q test



- Dixon's Q test: Find the largest gap in a sorted sample, and divide it by the total range
- In this case: $Q = \frac{2500}{3500} \simeq 0.71$
- Critical values:

Number of values:	3	4	5	6	7	8	9	10
Q90%:	0.941	0.765	0.642	0.560	0.507	0.468	0.437	0.412
Q95%:	0.970	0.829	0.710	0.625	0.568	0.526	0.493	0.466
Q99%:	0.994	0.926	0.821	0.740	0.680	0.634	0.598	0.568

Outline

Trimmed estimators

- Trimming is a generic method to to make an estimator robust
- The *n*% trimmed estimator is obtained by calculating the estimator on the sample limited to the range $[x_{[n\%]}, x_{[1-n\%]}]$
- This is not equivalent to removing outliers, as the trimmed estimators have the same expectations if there are no outliers
- The trimmed mean (or truncated mean) is a robust alternative to the mean, which is more dependent on the distribution than the median

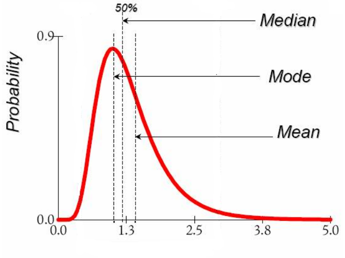
Outline

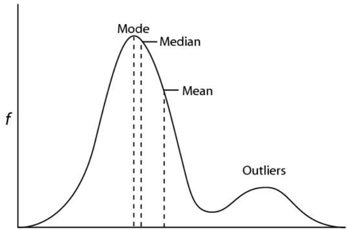
Winsorized estimators

- Winsorizing is another generic method to to make an estimator robust which is very similar to trimming
- The *n% winsorizing estimator* is obtained by replacing in the sample all values below $x_{[n\%]}$ by $x_{[n\%]}$ and all values above $x_{[1-n\%]}$] by $x_{[1-n\%]}$]
- I have no idea why you would want to do that...

Outline

Median





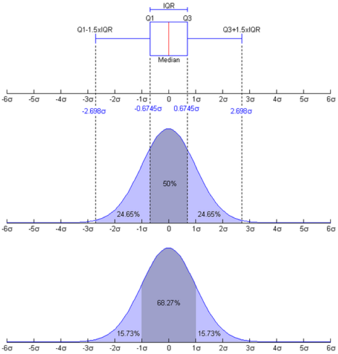
- The median $\tilde{x}$ is the 2-quantile, i.e. $x_{[0.5]}$, i.e. $x_{N/2}$ if N is even or $x_{N+1/2}$ if N is odd
- If we have 10 "opossums", the mean will be about one ton! It has a breakdown point of 0 %
- The median has a breakdown point of 50 %, i.e. you would get a roughly correct weight estimations even if there are 5 mammoths in your populations of 10 "opossums"

Outline

Median average deviation

- Assuming a sample $\{x_i\}$, $i = 1, \ldots, N$, and its median $\tilde{x}$, the median average deviation is $\mathrm{MAD}(\{x_i\}) = \mathrm{median}(|x_i - \tilde{x}|)$
- Example: $\{x_i\} \equiv \{1, 1, 2, 2, 4, 6, 9\}$ $\tilde{x} = 2$, $\{|x_i - \tilde{x}|\} \equiv \{0, 0, 1, 1, 2, 4, 7\}$, so $\mathrm{MAD}(\{x_i\}) = 1$
- Note that 6 and 9 are completely ignored
- Relation to standard deviation is distribution-dependent:
  - For a Gaussian distribution, $\sigma_x \simeq 1.4826\,\mathrm{MAD}(\{x_i\})$
  - For a uniform distribution, $\sigma_x = \sqrt{4/3}\,\mathrm{MAD}(\{x_i\})$

Inter-quartile range



- Assuming a sample $\{x_i\}, i = 1, \ldots, N$
- The inter-quartile range is the difference between the third and first quartiles: $\mathrm{IQR}(\{x_i\}) = x_{[0.75]} - x_{[0.25]}$
- For a Gaussian distribution, $\sigma_x \simeq \mathrm{IQR}(\{x_i\})/1.349$
- This can be used as a (non-robust) normality test

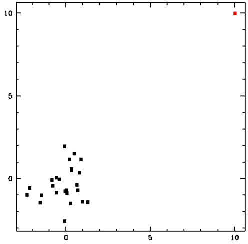
Outline

Ranks

- R-estimators are based on rank
- Given a sample $\{x_i\}$, $i = 1, \ldots, N$
- The rank of x_i is i if x_i is sorted in increasing order
- Example: $\{x_i\} \equiv \{4, 9, 6, 21, 3, 11, 1\}$
- The ranks are: $\{3, 5, 4, 7, 2, 6, 1\}$, because the sorted $\{x_i\}$ is $\{1, 3, 4, 6, 9, 11, 21\}$

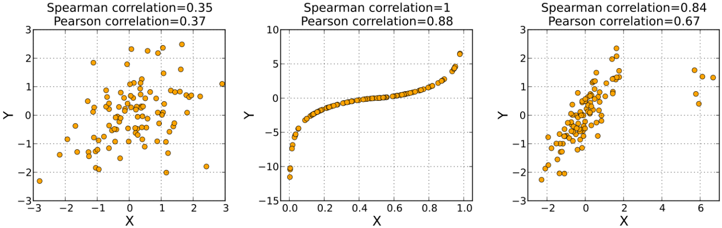
Outline

Pearson's r: a non-robust estimator



- Pearson's r correlation coefficient between two random variables $\{x_i, y_i\}$, $i = 1, \ldots, N$ is: $r = \frac{\overline{xy}}{\sigma_x \, \sigma_y}$
- But it is not robust. In this case, $r = 0.85$
- Of course, on can use L-estimators to compute $\overline{xy}$, σ_x, σ_y
- But we would have to figure out the critical values

Spearman's correlation coefficient



- Spearman's coefficient: Replace $\{x_i\}$ and $\{y_i\}$ with their ranks, and calculate s as the the Pearson's correlation coefficient of the ranks
- Consistent with Pearson's r in "good" conditions
- It is insensitive to the shape of y vs x and it is robust
- Significance: $t = s\sqrt{\frac{N-2}{1-s^2}}$ is approximately Student-t distributed with $N-2$ degrees of freedom
- Ties should have the same (average) rank, i.e. if $x_6 = x_7$, give them a rank of 6.5

Kendall's τ rank correlation

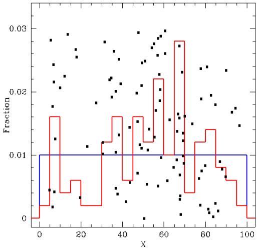
- Rank still contain quantitative values
- Kendall's τ test removes all quantities
- Form $N(N-1)/2$ pairs $\{\{x_i, y_i\}; \{x_j, y_j\}\}$, $j > i$

$$\tau = \frac{\sum_{i,j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j) - \sum_{i,j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_j - y_i)}{N(N-1)/2}$$

- For relatively large sample, $\tau \sim \mathcal{N}\left(0, \frac{2(2N+5)}{9N(N-1)}\right)$
- No single recipe in case of ties

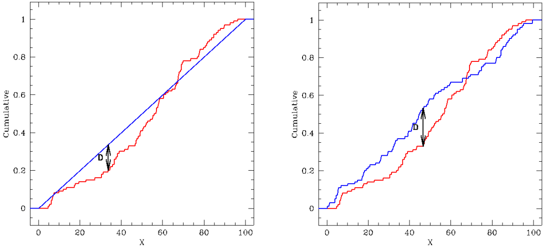
Outline

Do two distributions differ?



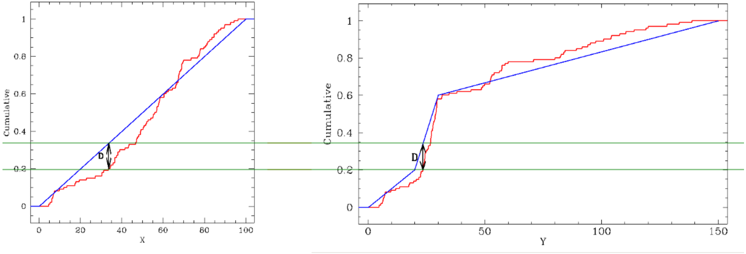
- Assuming a sample $\{x_i\}$, $i = 1, \ldots, N$, is it a probable outcome from a draw of N random variables from a given distribution (say $\mathcal{U}(a, b)$)?
- Similarly, assuming two samples $\{x_i\}$, $i = 1, \ldots, N$ and $\{y_j\}$, $j = 1, \ldots, M$, how probable is it that they have the same parent population?

Cumulative comparison



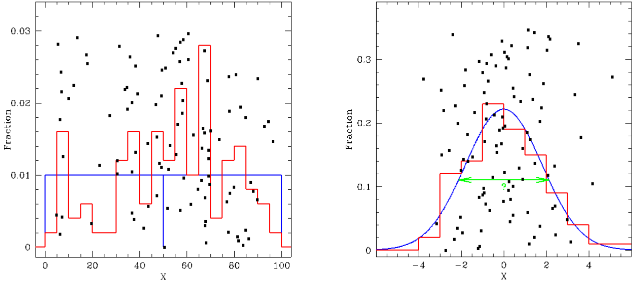
- A (good) idea is to compare cumulative distributions
- $F(x) = \int_a^b f(x)\,\mathrm{d}x$ for a continuous distribution $f(x)$
- $C_{\{x_i\}}(x) = \frac{1}{N}\sum_i H(x - x_i)$ for a sample ($H(x)$ is the Heaviside function, i.e., 1 if $x \geq 0$, 0 if $x < 0$)
- The KS test is the simplest quantitative comparison:
  $D = \max_x |C_{\{x_i\}}(x) - F(x)|$ or
  $D = \max_x |C_{\{x_i\}}(x) - C_{\{y_i\}}(x)|$

Independence on underlying distribution



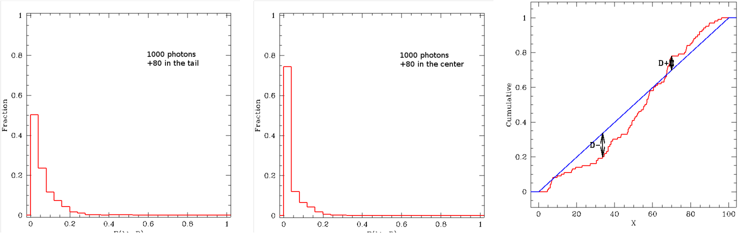
- D is preserved under any transformation $x \to y = \psi(x)$, where $\psi(x)$ is an arbitrary strictly monotonic function
- Thus KS test works with any underlying distribution
- The null-hypothesis distribution of D is: $P(\lambda > D) = 2\sum_{k=1}^{\infty} (-1)^{k-1}\, \mathrm{e}^{-2k^2\mu^2}$, with $\mu = \left(\sqrt{N} + 0.12 + 0.11/\sqrt{N}\right) \cdot D$
- When comparing two samples, use: $N_\mathrm{e} = \frac{N \cdot M}{N+M}$

Caveats



- In weird cases, KS-test might be extremely inefficient. KS test makes hidden assumptions
- KS-test will not work if you derive parameters from the data (see “Monte-Carlo methods” course)

Kuiper test



- KS test is more sensitive in the tails than in the center
- Among several solutions, the simplest: Kuiper test!
  $V = D^+ + D^-$
- $P(\lambda > V) = 2\sum_{k=1}^{\infty}\left(4k^2\mu^2 - 1\right)\mathrm{e}^{-2k^2\mu^2}$, with
  $\mu = \left(\sqrt{N} + 0.155 + 0.24/\sqrt{N}\right)\cdot V$
- Kuiper test can be used for distributions on a circle (see Paltani 2004)

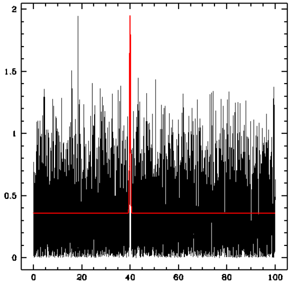
Outline

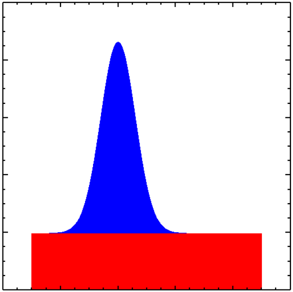
Concepts

- M-estimators are a generalization of maximum-likelihood estimators
- So, maximum likelihood is an M-estimator
- It allows to give arbitrary weight to data points
- Weights can be chosen so that they decrease for too far-off points, opposite to Gaussian weights in least square-estimation

Outline

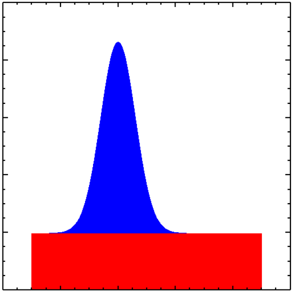
A case of known outlier distribution





- We search for a Gaussian peak in a very noisy environment
- In a fraction f of the cases the right peak is found. It has a Gaussian uncertainty
- In a fraction $1 - f$ of the cases a spurious peak is found. It can be anywhere in the searched range
- We know the real distribution of our measurements

Maximum-likelihood estimation



- We obtain a sample $\{x_i\}$, $i = 1, \ldots, N$
- The distribution of x_i is $f\,\mathcal{N}(\mu, \sigma) + (1 - f)\,\mathcal{U}(0, 100)$
- One can use a maximum likelihood to find parameters μ, σ and f!

$$L = \prod_i \frac{f}{\sqrt{2\pi}\sigma} \exp\left(-(x_i - \mu)^2/2\sigma^2\right) + \frac{1-f}{100}$$

Outline

Model fitting with M-estimators

- Let's take sample $\{x_i; y_i\}$, $i = 1, \ldots, N$, whose errors on y_i we know are not normally distributed, and a model $y(x; \mathbf{p})$, where $\mathbf{p}$ is the parameters of the model. As above, we have:

$$L = \prod_i \exp\left(-\rho(y_i, y(x_i; \mathbf{p}))\right)$$

  ρ is the negative logarithm of the probability
- We want then to minimize:

$$\sum_i \rho\left(y_i, y(x_i; \mathbf{p})\right)$$

- Let's assume that ρ is local, i.e. $\rho(y_i, y(x_i; \mathbf{p})) \equiv \rho(z)$, with $z = (y_i - y(x_i; \mathbf{p}))/\lambda_i$, i.e. ρ depends only on the difference with the model scaled by a factor λ_i

Parameter estimation (cont.)

- Let's write $\psi(z) \equiv \frac{\mathrm{d}\rho(z)}{\mathrm{d}z}$
- The minimum of L is obtained when:

$$0 = \sum \frac{1}{\lambda_i} \psi \left(\frac{y_i - y(x_i; \mathbf{p})}{\lambda_i} \right) \left(\frac{\partial y(x_i; \mathbf{p})}{\partial \mathbf{p}_k} \right)$$

- We can solve this equation, or we can minimize $\sum_i \rho \left(\frac{y_i - y(x_i; \mathbf{p})}{\lambda_i} \right)$
- $\psi(z)$ acts as a weight in the above equation

Some weights one can think of

- In the Gaussian case, put $\lambda_i = \sigma_i$, and we get the least-squares estimates
- But we have then: $\rho(z) = z^2/2$ and $\psi(z) = z$
- Two-sided exponential, $P(z) \sim \exp(-z)$:
  $\rho(z) = |z|$ and $\psi(z) = \mathrm{sgn}(z)$
- Lorentzian, $P(z) \sim \frac{1}{1+z^2/2}$:
  $\rho(z) = \log(1 + z^2/2)$ and $\psi(z) = \frac{z}{1+z^2/2}$