Distribution Analyses and its Applications

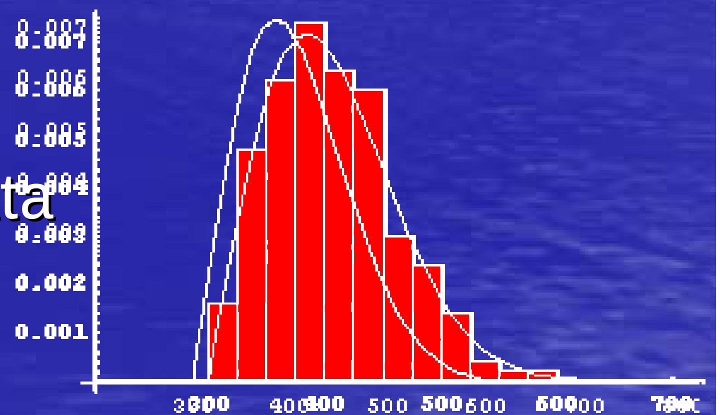
Denis Cousineau
Université de Montréal

This talk and the demos will be available at:
www.mapageweb.umontreal.ca/cousined/home/talks.html

1- Why look at distributions and how

Introduction

- Why look at distributions? (mostly RT distributions)
  - Screening the outliers
  - Getting better descriptive statistics
  - Testing models
- What is a distribution?
  - the empirical distribution of the data
  - the theoretical distribution
  - both



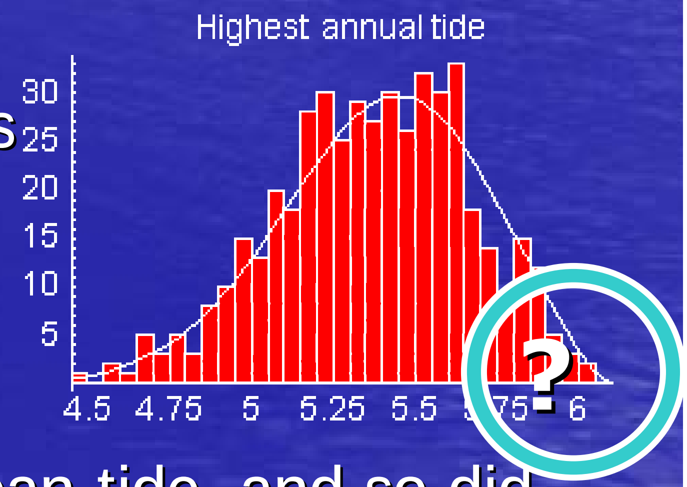
What is fitting a model of RT distribution?

- Techniques to estimate the population parameters.
- With the Normal (Gaussian) distribution, there are "straightforward" recipes (i.e. direct computations) to obtain the population parameters
  - μ is given by the sample mean
  - σ is the sample standard deviation (corrected for the bias)
- Thus, when you estimate the mean μ of a population, you are in fact fitting a model: the Normal model!

What is fitting a model of RT distribution?

- With other distributions, there may not exist direct computations to get the population parameters.
  - They must be estimated
  - The estimates must be evaluated through fitting

- Example of the Netherlands
  - They are really concerned with tides
  - They have accurate records dating back to 1534 ➔

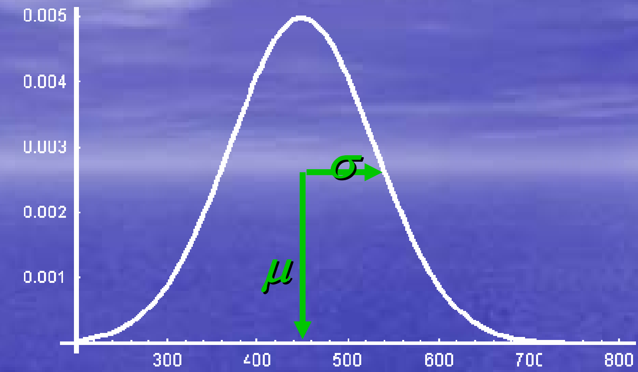


  - They were not interested by the mean tide, and so did not use the Normal model

If you need a model with | Use | Which looks like

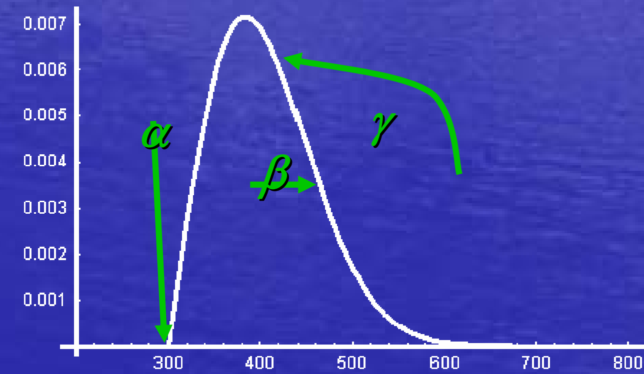
a central tendency parameter
a spread parameter

Normal distribution
- The position μ
- The spread σ

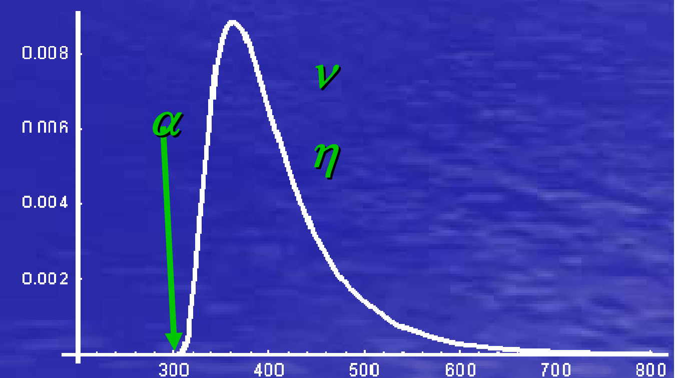


a lower limit to how fast a person can be
a spread parameter

Weibull distribution
- The position α
- The spread β
- The asymmetry γ



LogNormal distribution
- The position α
- The spread and asymmetry ν & η



The ExGaussian is undistinguishable from the LogNormal

How to fit a model of RT distribution?

- In order to fit a distribution, two things are required:
  - An objective function
    - A function that gives the fit of the parameters to the data
    - The best choice is the likelihood of the data given the parameters
  - A search procedure
    - e.g. the simplex (Nelder-Mead method) which plays with the parameters until the objective function is as large as possible.
    - Exists in many computer programs, e.g. *Matlab* (fmin), *Mathematica* (NMinimize), *Excel* (Solver), etc.

How to fit a model of RT distribution?

- The likelihood function requires:
  - f, the equation of the distribution, its shape
  - θ, the parameter set of the distribution

$$L(X \mid \theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

between 0 (bad fit)
and 1 (perfect fit)

$$LL(X \mid \theta) = -\mathrm{Log}(L(X \mid \theta))$$

$$= -\mathrm{Log}\left(\prod_{i=1}^{n} f(X_i \mid \theta)\right)$$

$$= -\sum_{i=1}^{n} \mathrm{Log}(f(X_i \mid \theta))$$

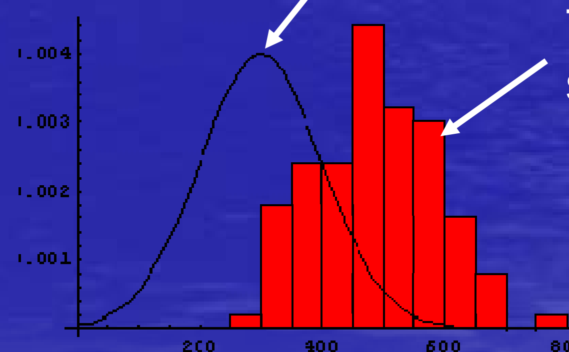
between ∞ (bad fit)
and 0 (perfect fit)

The model

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

$\mu = 300$

$\sigma = 100$



The data set **X**

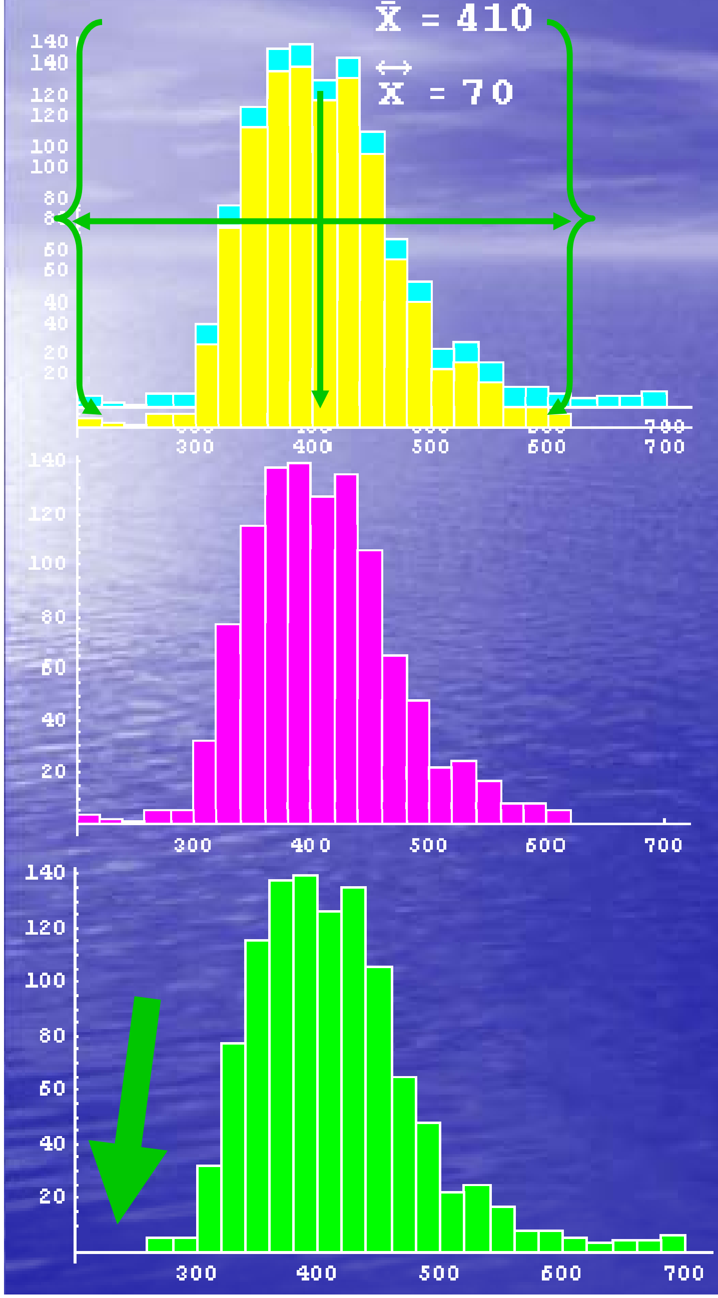
2- Examples of situations where distributions can be used

a. For screening outliers

Screening outliers

- Outliers are RTs that are either too small or too large
  - They can be correct RTs
  - or –
  - Caused by unrelated activities
- There exists three techniques to remove outliers:
  - Visual inspection of the distribution
  - Single cut at ± 3 standard deviations from the mean
  - Iterative cut at ± 3 standard deviations from the mean

Screening outliers



- <u>Single</u> truncation at ± 3 std
  - the left tail is untouched
  - the right tail is truncated
- <u>Iterative</u> truncation at ± 3 std
  - the results are undistinguishable
  - not worth the trouble
- Visual inspection
  - the left tail is problematic
  - Because of the asymmetry, no symmetrical process will detect them

Screening outliers

- The best technique at this moment is visual inspection
- RT data are always asymmetrical and techniques that weigh both sides identically around the mean are doomed to failed
- There might exist an alternative based on the most probable smallest/highest observation... Next year?

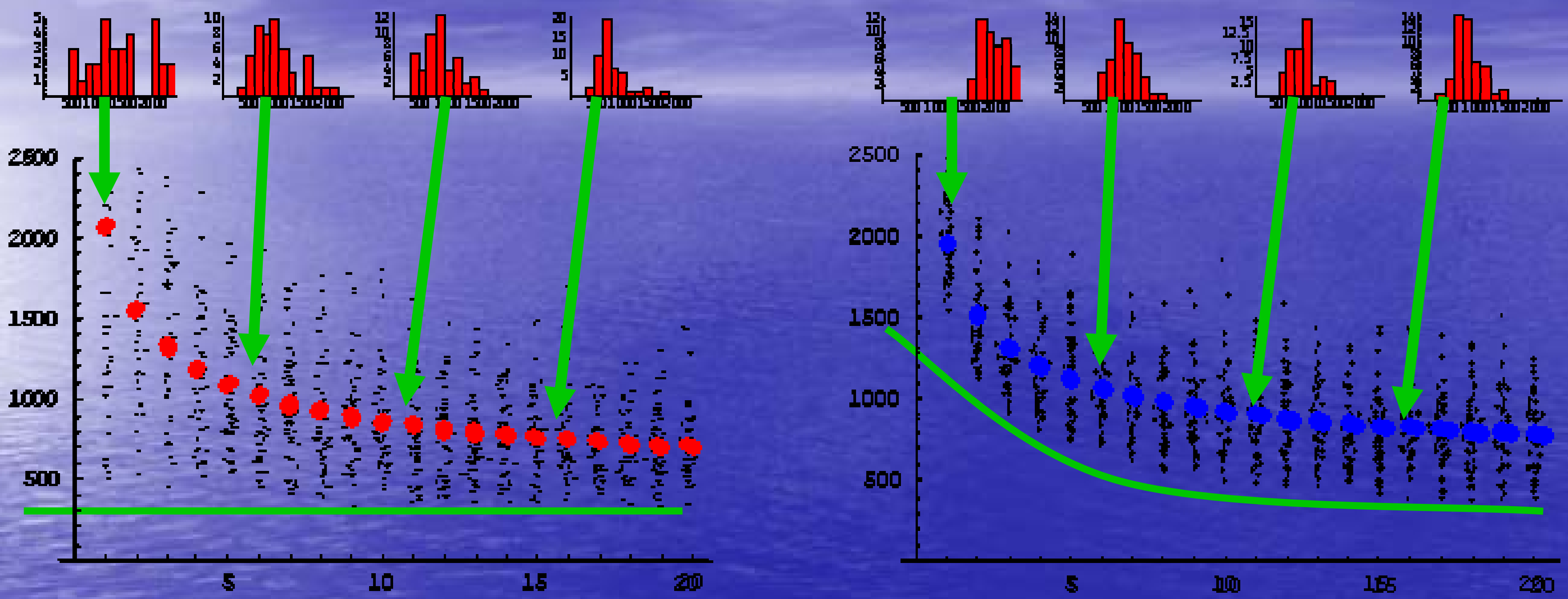
b. For getting descriptive statistics

Getting descriptive statistics

- The most usual descriptive statistics are
  - the mean
  - the mean
  - the mean
  - the median or the geometric/harmonic mean

  - the standard deviation – or equivalently –
  - the standard error of the mean
- Does the mean hold the key to all the questions? or should we look at some results through different lenses?

Getting descriptive statistics



- The learning curve showing mean RT as a function of training session.
- What is the meaning of the mean in this context?

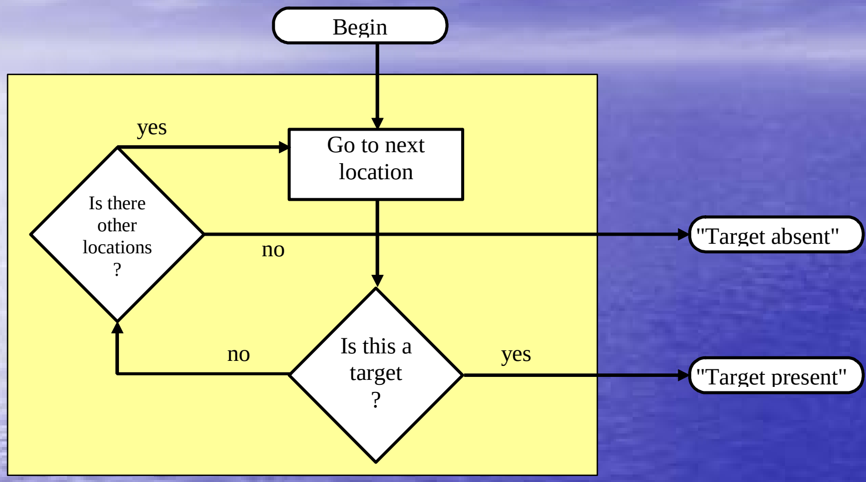
Getting descriptive statistics

- Despite the appearance, the mean may not always be a relevant statistic
- Ask the distribution of your data what are the best way to describe them

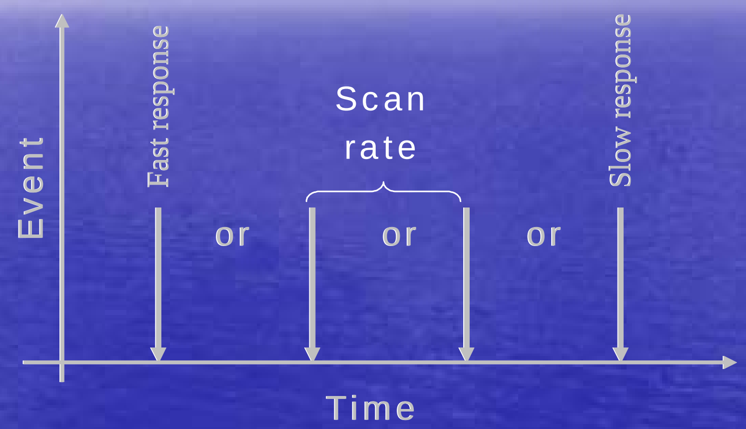
c. For testing models

Testing a model of visual search

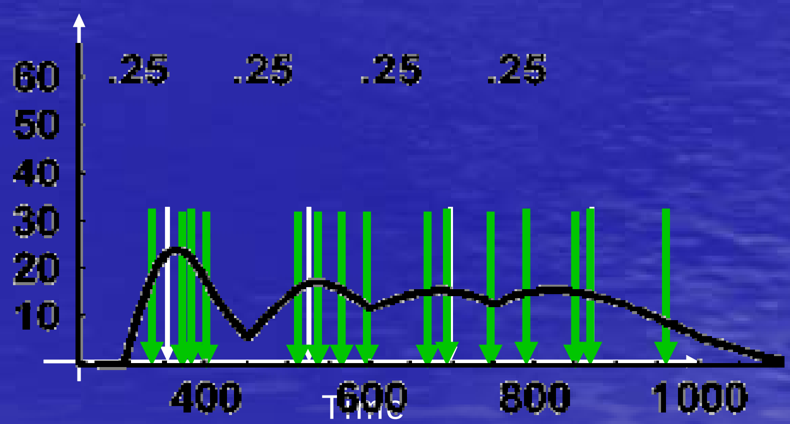
The serial (random-order) self-terminating search



Four items, target is present



Four items, target present, variability



- The responses are more spread out for the slow responses because the variability of the previous responses is additive…

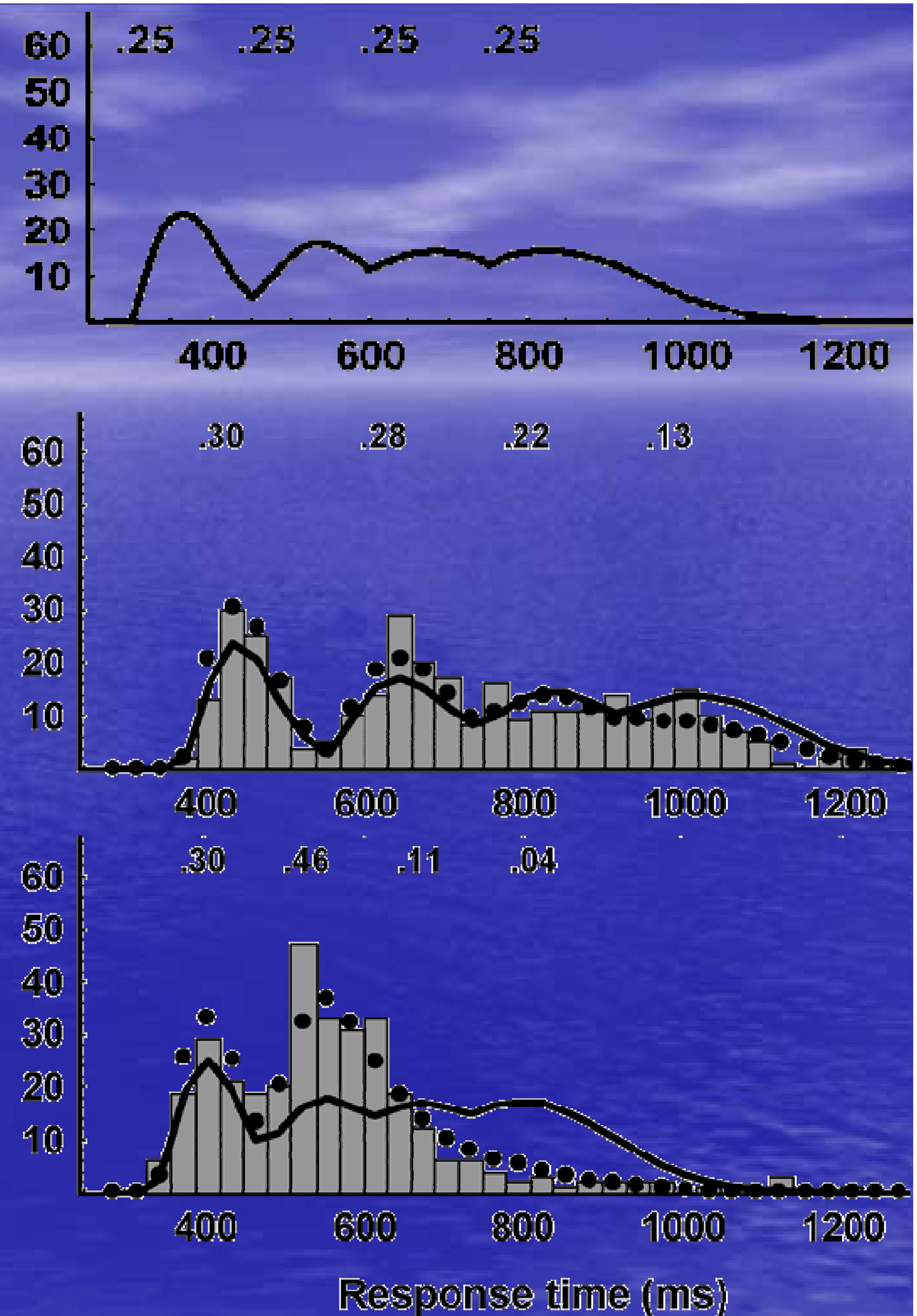
Testing models

Results (Cousineau & Shiffrin 2004)

- The participants were well trained (45 hours).

- Targets are found more often on the first or second scan than expected.

- ➔ The order of the search is not random.



Testing a model of visual search
The serial (random-order) self-terminating search

- The above results are definitive
  - No random-order model could mimic such pattern of results
- Looking at means only:
  - the slopes and the 2:1 slope ratios could favor a serial search model or a parallel search model
  - This is called mimicking (different models predicting the same means)
  - Whole distributions cannot be mimicked easily
- Whereas means are relevant in the context of search models, they have no power to discriminate between models.

3- Doing it with Mathematica or Excel

Conclusion

Conclusions

- Samples should be reasonably large:
  - greater than 100 per subject per condition with L (Cousineau & Larochelle, 1997)
  - greater than 40 per subject per condition with QL (Cousineau, Brown & Heathcote, 2004)
  - greater than 25 per subject with distribution averaging (Cousineau & Lacouture, submitted)

Conclusions

- Beware of the means
  - Is it really what you want?
  - Is it what the data deserve?
- Never miss a chance to look at the BIG picture
  - The empirical distribution shows everything from the mean to the asymmetry

Thank you

This talk and the demos will be available at:
www.mapageweb.umontreal.ca/cousined/home/talks.html