Some basic statistics and curve fitting techniques

'**Statistics** is the discipline concerned with the study of variability, with the study of uncertainty, and with the study of decision-making in the face of uncertainty' (Lindsay et al., 2004).

Statistics is the science of **collecting**, **organizing**, **analyzing** and **interpreting data**.

Nominal data – categories that are not ordered (e.g. taxa).

Ordinal data – fits in categories that are ordered but level between orders has no objective measure (e.g. pain level).

Scale data – fits in categories that are ordered with units measures between levels (e.g. units such as m/s)

Why do we need statistics?

Statistics helps to provide answers to questions such as:

1. What is the concentration of plankton at the dock right now (given past measurements)?

2. Will species x be in the water tomorrow?

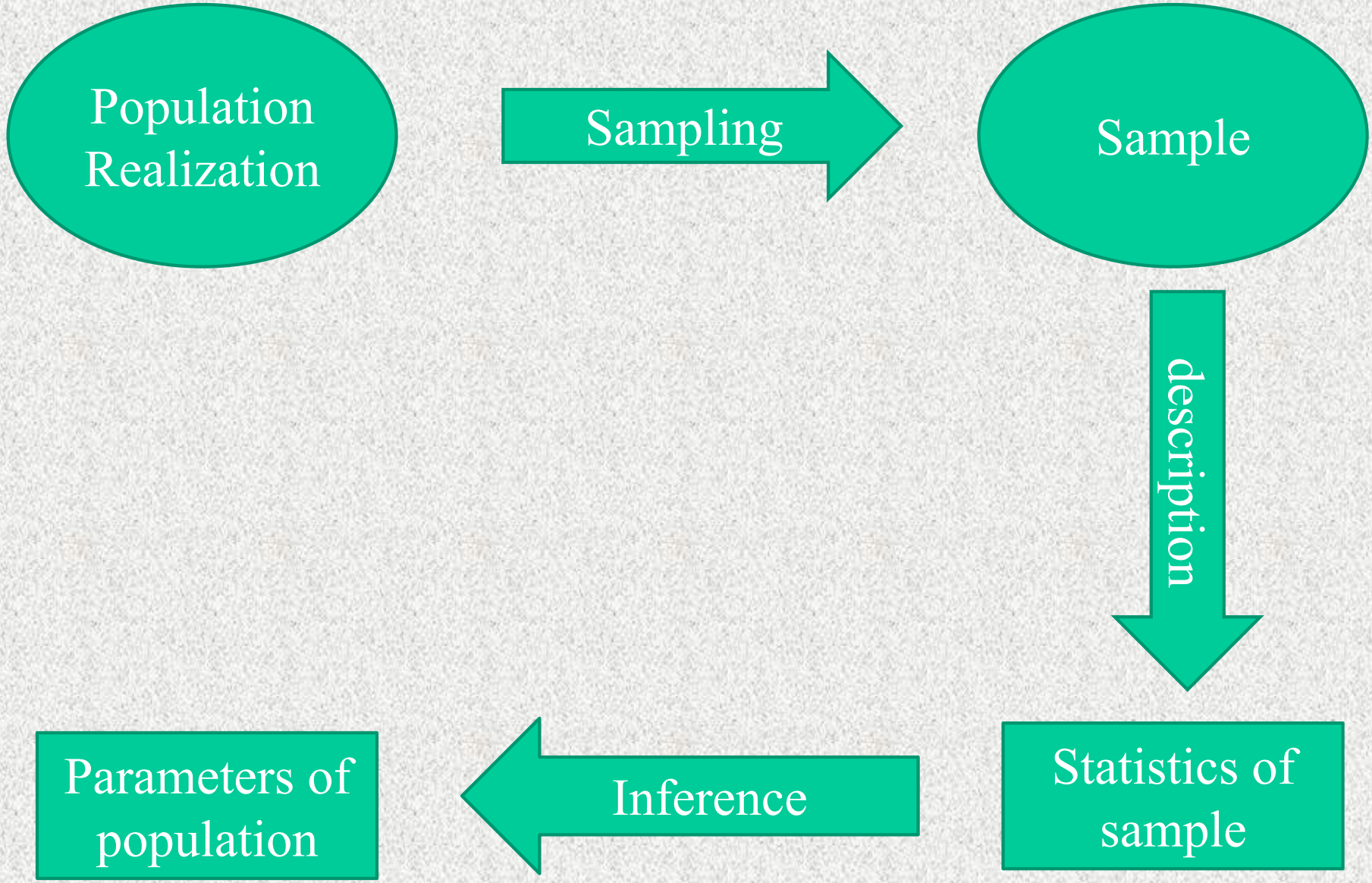
We are interested in the likelihood of the answer and help reduce large datasets into their salient characteristics.

The use of statistics to make a point:

1. Statistics never proves a point (it says something about likelihood).

2. If you need fancy statistic to support a point, your point is, at best, weak… (Lazar, 1991, personal communication)

Why do we need statistics?



Statistical description of data

Statistical moments (1st and 2nd):

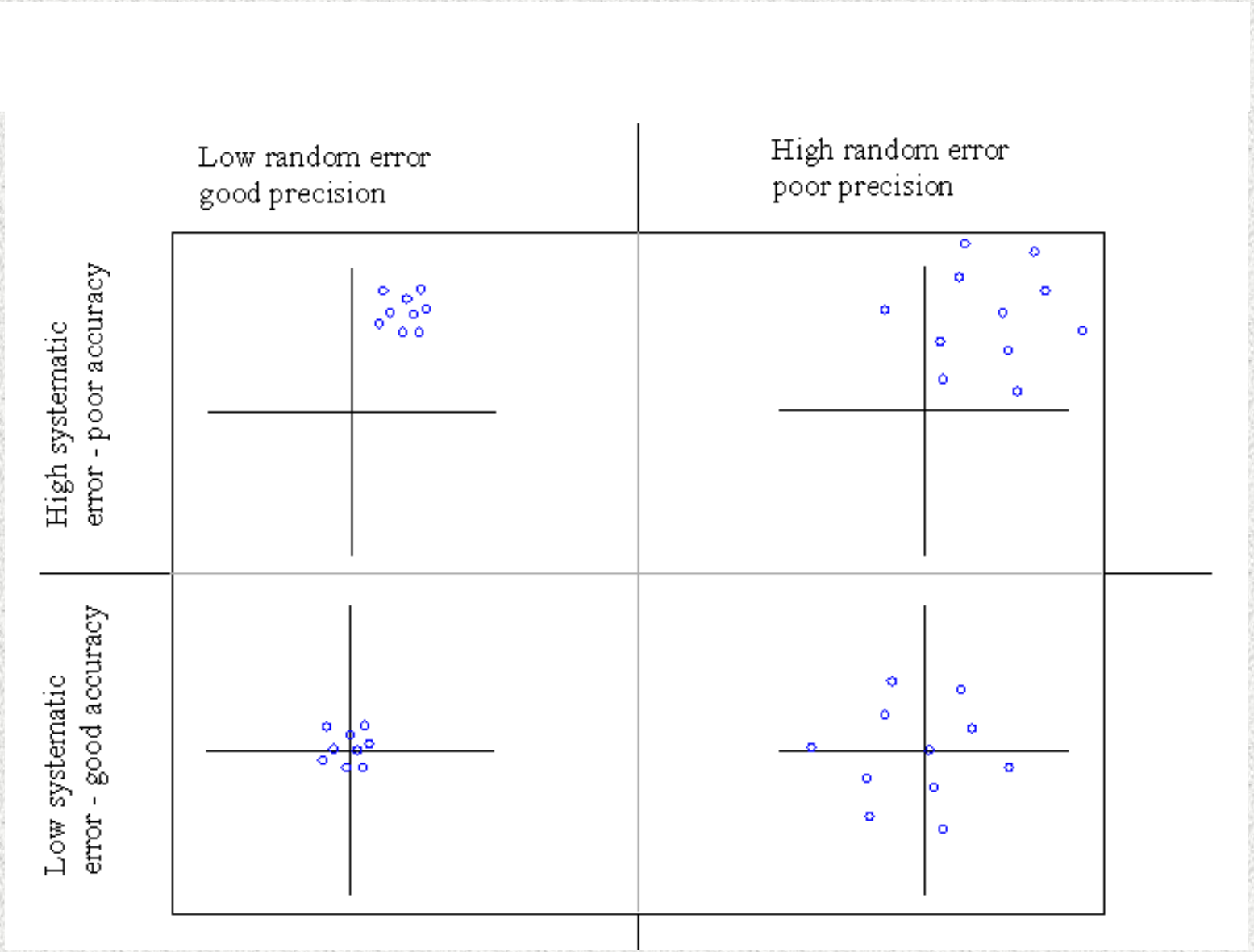
- Mean: $\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$
- variance: $Var = \frac{1}{N-1}\sum_{j=1}^{N}\left(x_j - \bar{x}\right)^2$
- Standard deviation: $\sigma = \sqrt{Var}$
- Average deviation:

$$Adev = \frac{1}{N}\sum_{j=1}^{N}\left|x_j - \bar{x}\right|$$

- Standard error: $s_{error} = \sigma/\sqrt{N}$

- Standard error: $s_{error} = \sigma / \sqrt{N}$

When is the uncertainty not reduced by additional sampling?



Statistical description of data

Probability distribution:

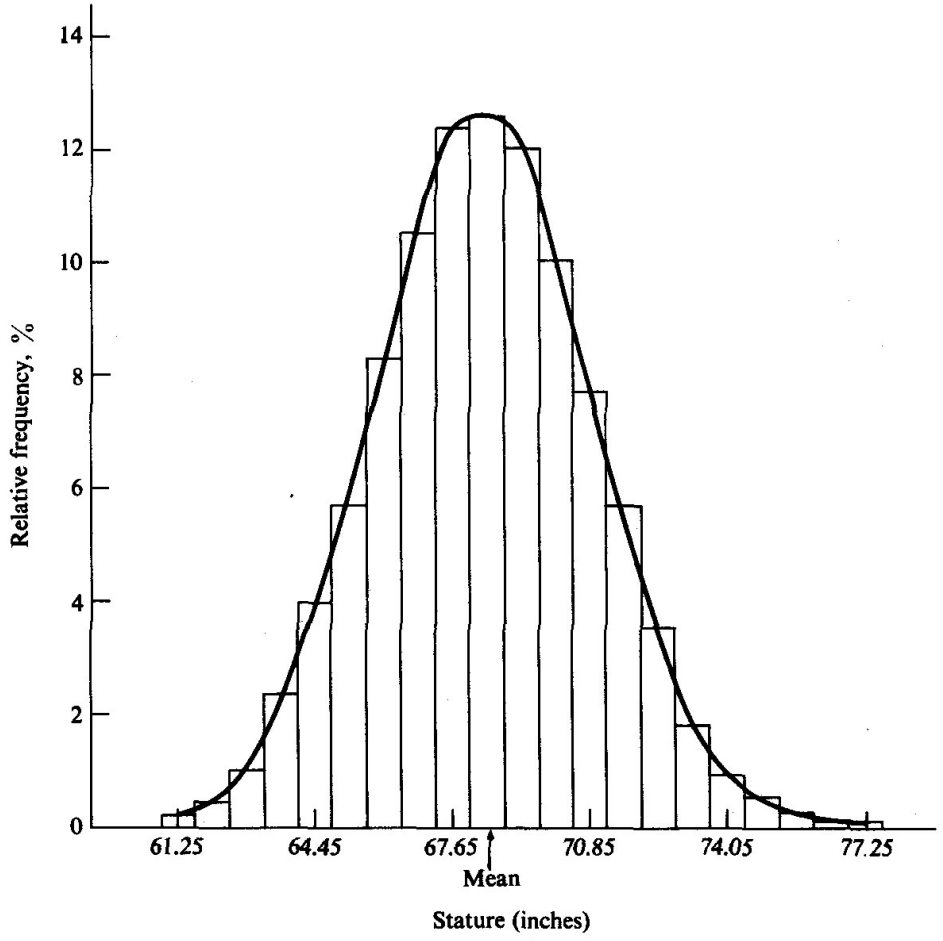


Fig. 1–2 Histogram of frequency distribution of stature of 24,404 U.S. Army males. Adapted from data of Newman and White.

Non-normal probability distribution:

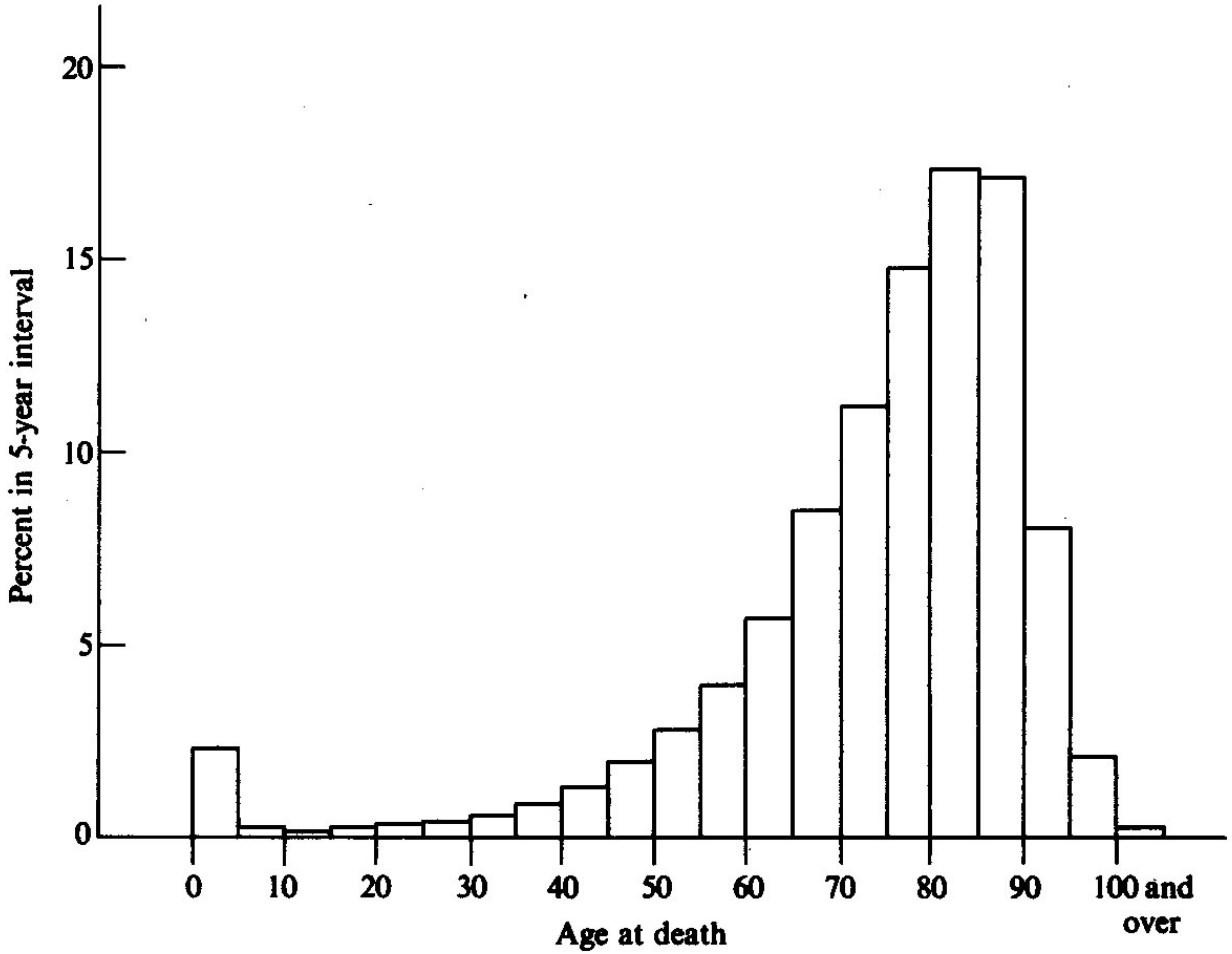


Fig. 1–3 U.S., female, 1965: percent dying in each 5-year age interval (the 100–105 interval includes all deaths after 100 rather than only those occurring in the interval). Data from N. Keyfitz and W. Flieger, *World Population: An Analysis of Vital Data.* Chicago: University of Chicago Press, 1968, p. 45.

Statistical description of data

Nonparametric statistics (when the distribution is unknown):

- rank statistics

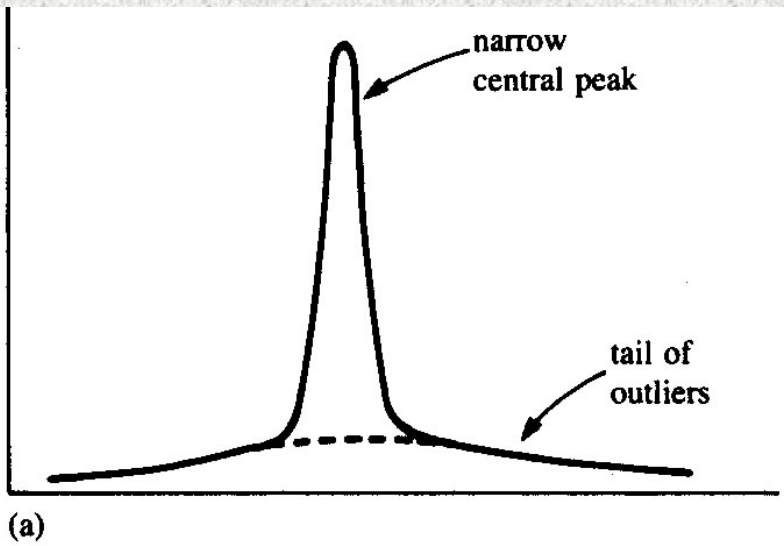
$$x_1, x_2, \ldots, x_N \rightarrow 1,2,\ldots,N$$

- Median
- percentile
- Deviation estimate
- The mode

Issue: *robustness,* sensitivity to outliers

Statistical description of data

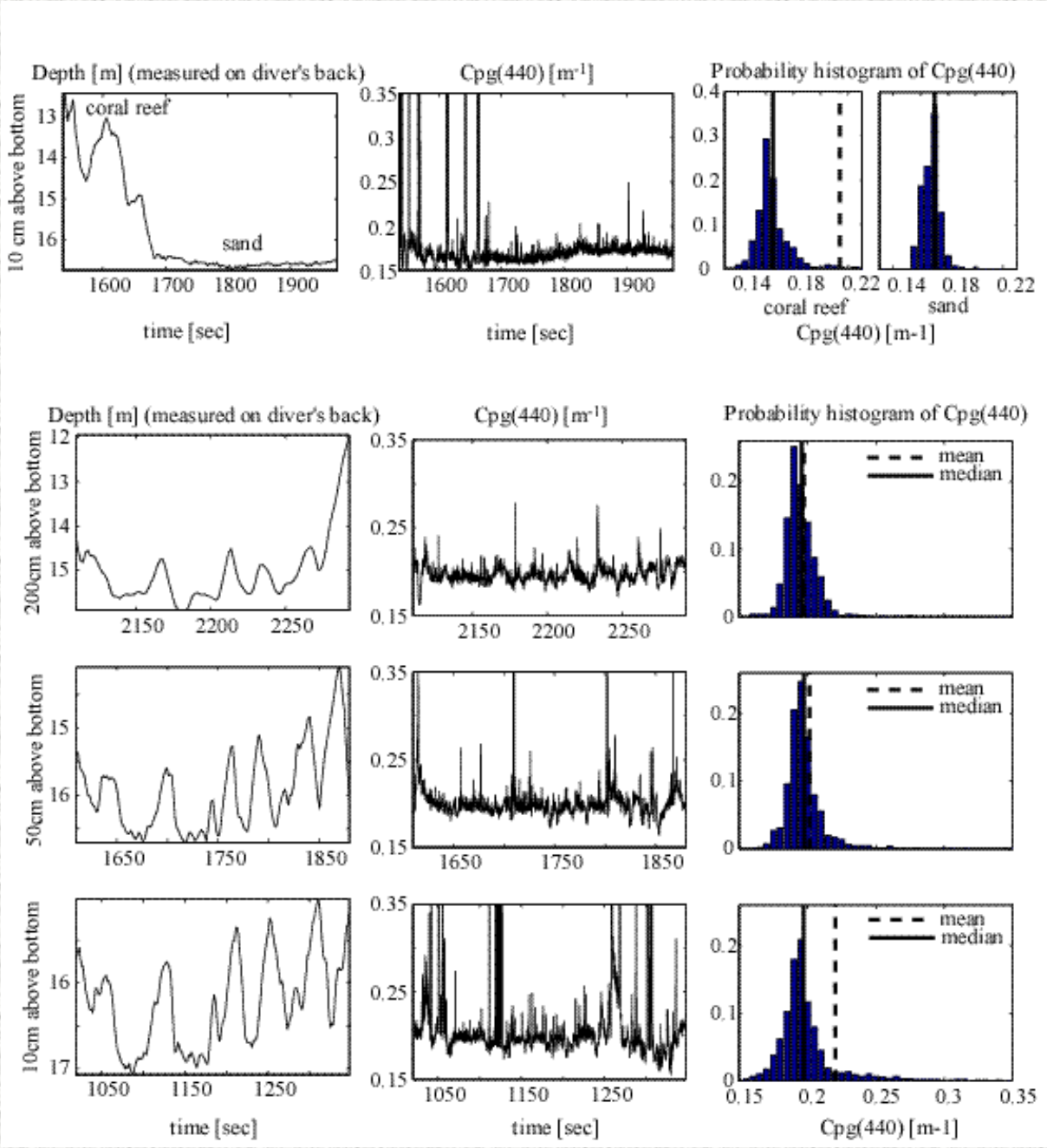
Robust: "insensitive to small departures form the idealized assumptions for which the estimator is optimized."



Press et al., 1992,
Numerical recipe

Statistical description of data

Examples from COBOP, Linking variability in IOPs to substrate:



Boss and Zaneveld, 2003 (L&O)

What do we care about in research to which statistics can contribute?

- Relationships between variables (e.g. do we get blooms when nutrients are plentiful?)

- Contrast between conditions (e.g. is diatom vs. dinoflagellate domination associated with fresh water input?).

Relationship between 2 variables

Linear correlation:

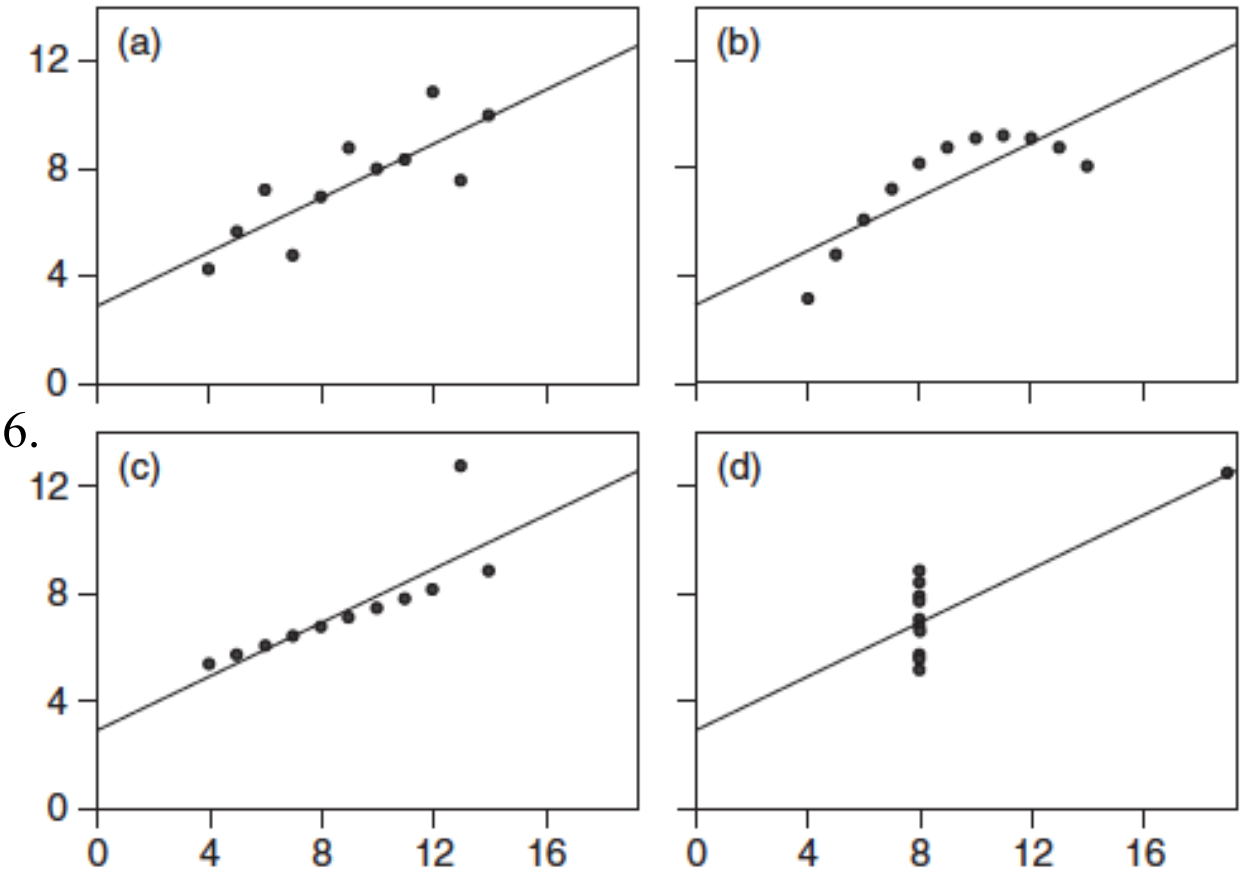
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

Rank-order correlation:

$$r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2} \sqrt{\sum_i (S_i - \bar{S})^2}}$$

Relationship between 2 variables

Same mean,
Stdev, and r=0.816.



Wilks, 2011

FIGURE 3.16 "Anscombe's quartet," illustrating the ability of graphical EDA to discern data features more powerfully than can a few numerical summaries. Each horizontal (x) variable has the same mean (9.0) and standard deviation (11.0), as does each of the vertical (y) variables (mean 7.5, standard deviation 4.12). Both the ordinary (Pearson) correlation coefficient ($r_{xy} = 0.816$) and the regression relationship ($y = 3 + x/2$) are the same for all four of the panels.

Regressions (models)

$y = f(x)$

Dependent and independent variables:

- Absorption spectra.
- Time series of scattering.

What about chlorophyll vs. size?

Regressions of type I and type II

Uncertainties in y only:

$$y(x) = ax + b$$

$$\chi^2 = \sum_{i=1:N} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

Minimize χ^2 by taking the derivative of χ^2 wrt *a* and *b* and equal it to zero.

What if we have errors in both x and y?

$$y(x) = ax + b$$

$$\chi^2 = \sum_{i=1:N} \frac{(y_i - ax_i - b)^2}{\sigma^2{}_{yi} + a^2\sigma^2{}_{xi}}$$

$$Var(y_i - ax_i - b) = \sigma^2{}_{yi} + a^2\sigma^2{}_{xi}$$

Minimize χ^2 by taking the derivative of χ^2 wrt *a* and *b* and equal it to zero.

The coefficient of determination

$R^2 = 1 - MSE/Var(y)$.

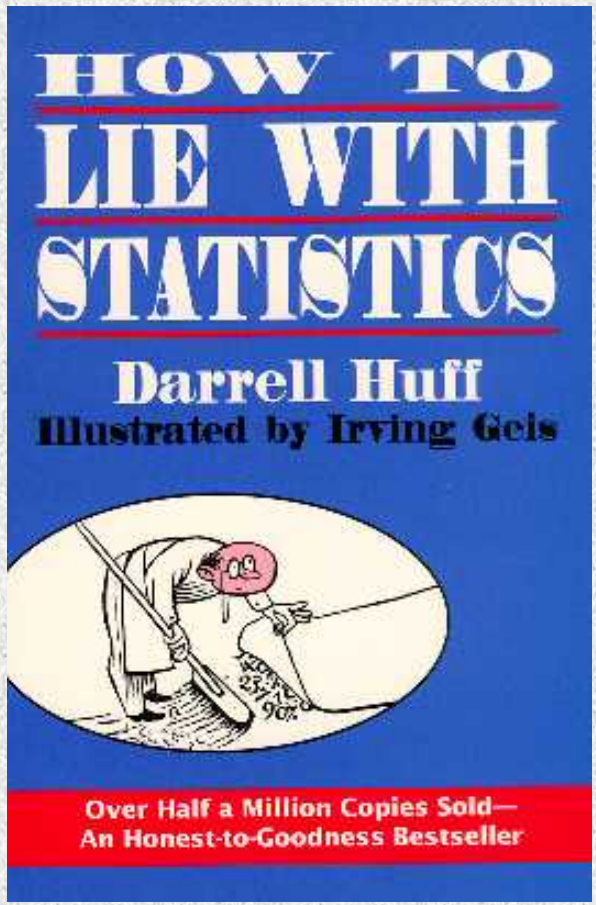
MSE=mean square error=average error of model^2/variance.

What variance does it explain?

Can it reveal cause and effect?

How is it affected by dynamic range?

R is the ‘correlation coefficient’.



Regressions of type I and type II

Classic type II approach (Ricker, 1973):

The slope of the type II regression is the geometeric mean of the slope of y vs. x and the inverse of the slope of x vs. y.

$$y(x) = ax + b$$
$$x(y) = cy + d$$
$$a_{II} = \sqrt{a/c} = \pm \sigma_y / \sigma_x$$
$$\pm = sign\left\{\sum_i x_i y_i\right\}$$

Smoothing of data

Filtering noisy signals.

What is noise?

- instrumental (electronic) noise.
- Environmental ‘noise’.

“one person’s *noise* may be another person’s *signal*”

Matlab: filtfilt

Method of fluctuation

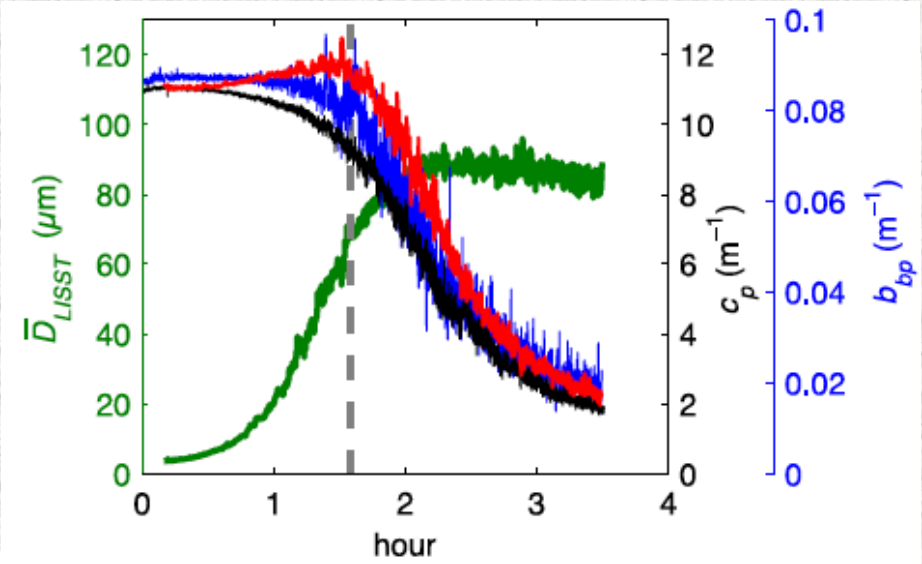
Lab aggregation exp.:

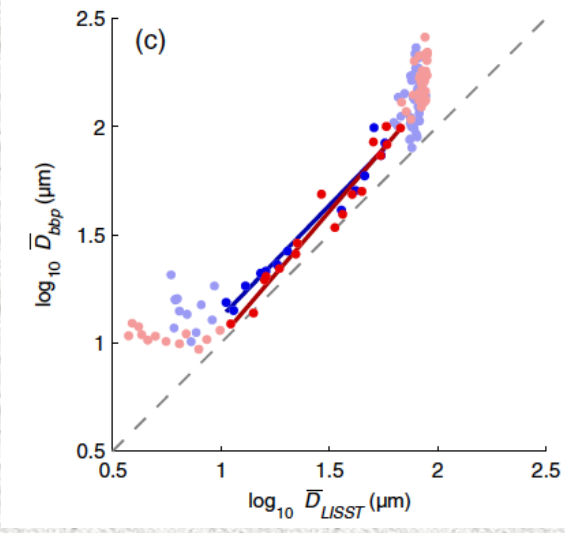
Sample volume

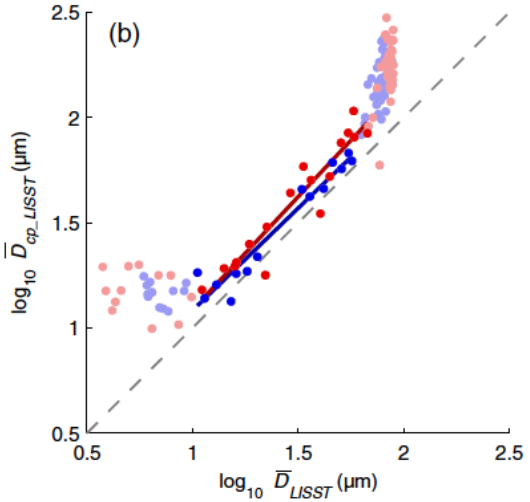
$$\bar{A}_{cp} = \frac{\mathrm{Var}[c_p(t)]}{E[c_p(t)]} \frac{V}{Q_c} \frac{1}{\alpha(\tau)}$$

$$\bar{D} = 2\sqrt{\bar{A}\pi^{-1}}$$

Measurement time







Briggs et al., 2013

Modeling of data

Condense/summarize data by fitting it to a model that depends on adjustable parameters.

Example, CDM spectra:

$$a_g(\lambda) = \tilde{a}_g \exp(-s(\lambda - \lambda_0))$$

particulate attenuation spectra:

$$c_p(\lambda) = \tilde{c}_p \left(\frac{\lambda}{\lambda_0} \right)^{-\gamma}$$

Modeling of data

Example: CDM spectra.

$$a_g(\lambda) = \tilde{a}_g \exp(-s(\lambda - \lambda_0))$$

$$\Rightarrow \boldsymbol{a} = \left[\tilde{a}_g, s\right]$$

Merit function:

$$\chi^2 = \sum_{i=1}^{9} \left[\frac{a_g(\lambda_i) - \tilde{a}_g \exp(-s(\lambda - \lambda_0))}{\sigma_i} \right]^2$$

- For non-linear models, there is no guarantee to have a single minimum.
- Need to provide an initial guess.

Matlab: fminsearch

Modeling of data

Lets assume that we have a model

$$y = y(\lambda; \boldsymbol{a})$$

A more robust merit function:

$$\tilde{\chi} = \sum_{i=1}^{N} \left| \frac{y(\lambda_i) - y(\lambda_i; \boldsymbol{a})}{\sigma_i} \right|$$

Problem: derivative is not
continuous. Can be used to fit lines.

Statistical description of data

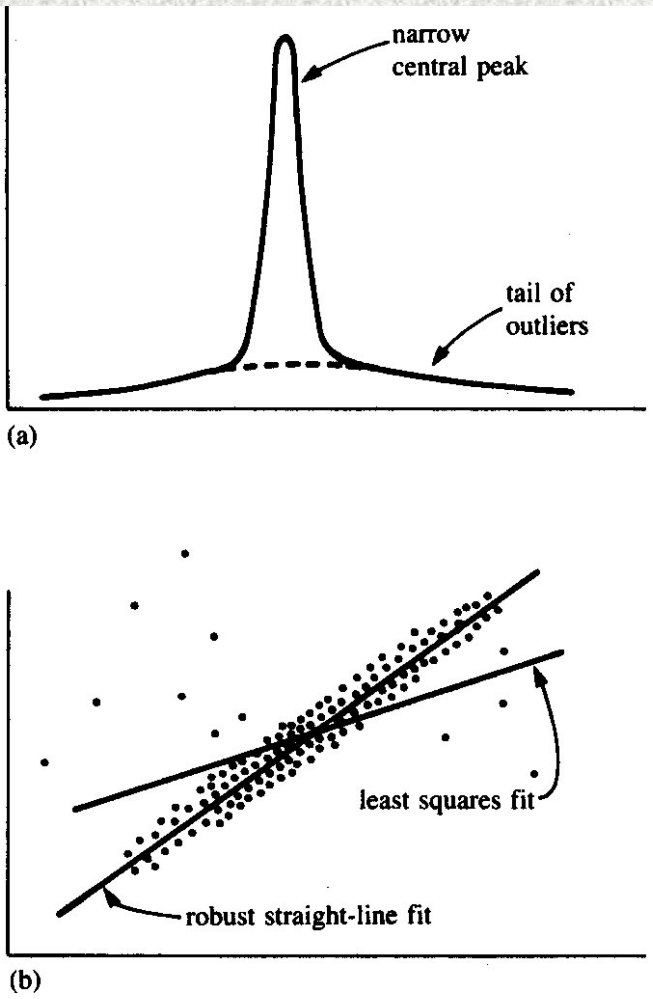


Figure 14.6.1. Examples where robust statistical methods are desirable: (a) A one-dimensional distribution with a tail of outliers; statistical fluctuations in these outliers can prevent accurate determination of the position of the central peak. (b) A distribution in two dimensions fitted to a straight line; non-robust techniques such as least-squares fitting can have undesired sensitivity to outlying points.

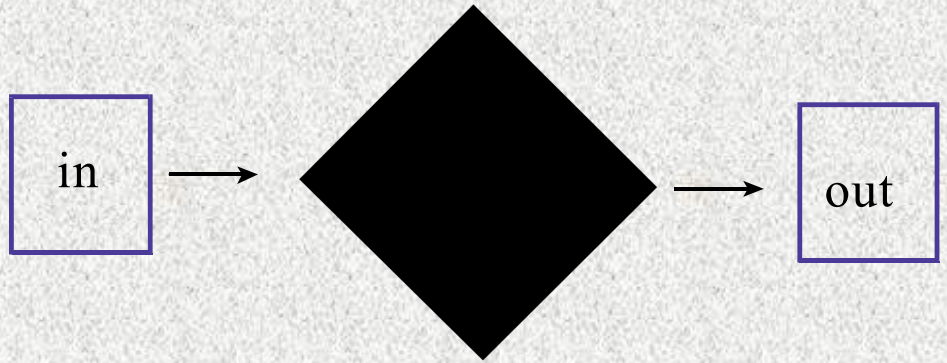
Press et al., 1992

Monte-Carlo/Bootstrap methods

Need to establish confidence intervals in:

1. Fitting-model parameters (e.g. CDM fit).

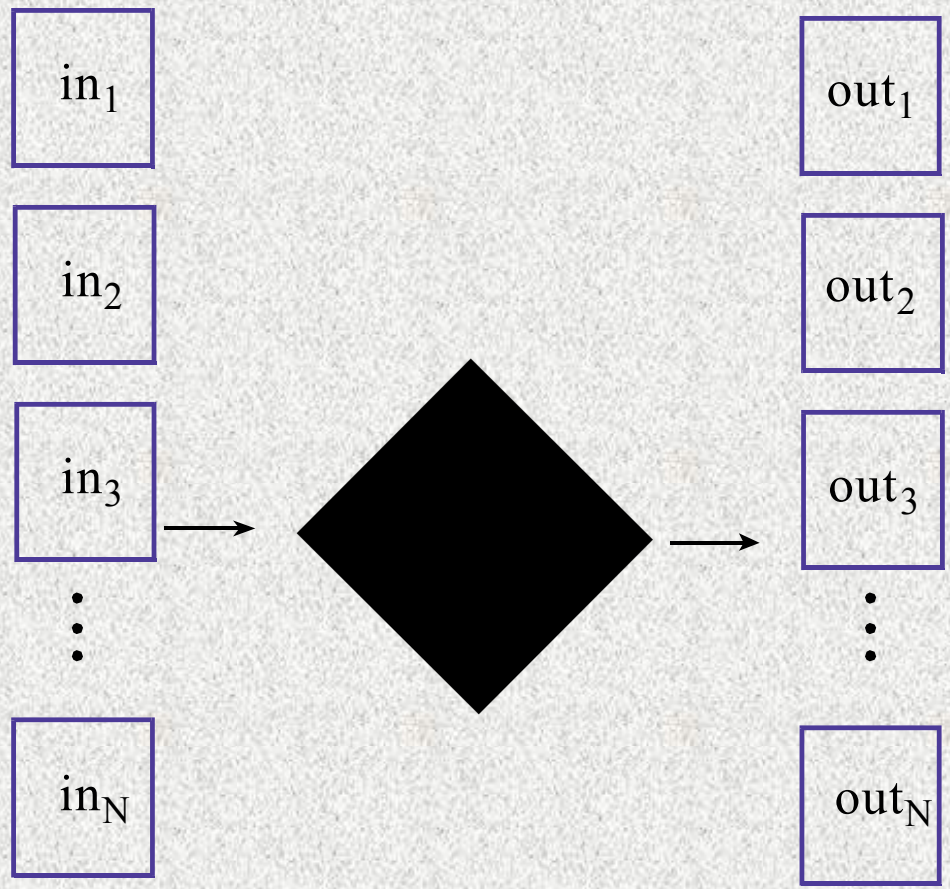
2. Model output (e.g. Hydrolight).



Bootstrap

When there is an uncertainty (or possible error) associated with the input:

Vary inputs with random errors and observe effect on output:



Bootstrap

Example: how to assign uncertainties in derived spectral slope of CDOM.

Merit function:

$$\chi^2 = \sum_{i=1}^{9} \left(a_g(\lambda_i) \pm \Delta_i - \tilde{a}_g \exp\left(-s\left(\lambda - \lambda_0\right)\right)\right)^2$$

Randomly add uncertainties (Δ_i) to each measurement, each time performing the fit (e.g. using randn.m in Matlab, RAND in Excel).

Then do the stats for the different *s*.

Summary

Use statistics logically. If you don't know the underlying distribution use non-parametric stats.

Statistics does not prove anything but can give you a sense of the likelihood of a hypothesis (about relationships).

I strongly encourage you to study hypothesis tests and Baysian methods. Beware that they are often misused…

frontiers in Physics

MINI REVIEW
published: 04 March 2016
doi: 10.3389/fphy.2016.00006

REPRODUCIBILITY

Statisticians issue warning on *P* values

Statement aims to halt missteps in the quest for certainty.