Some basic statistics and curve fitting techniques

Some important concepts:

- Data
- Statistical description of data (data reduction, independence)
- The use of statistics to make a point:
  1. Statistics never proves a point.
  2. If you need fancy statistic to support a point, your point is, at best, weak…

Statistical description of data

Statistical moments (1st and 2nd):

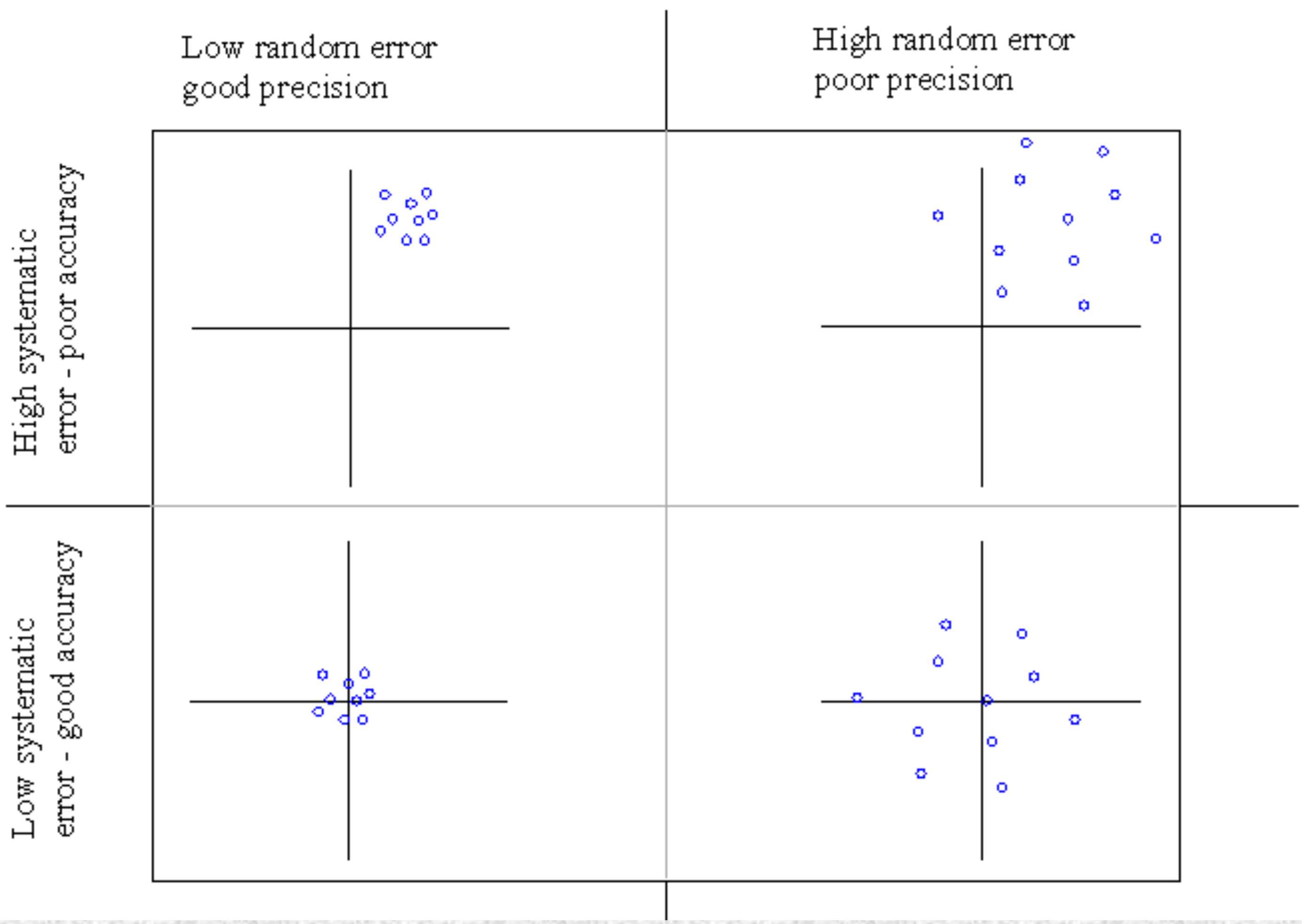
- Mean: $$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$$
- variance: $$Var = \frac{1}{N-1}\sum_{j=1}^{N}\left(x_j - \bar{x}\right)^2$$
- Standard deviation: $$\sigma = \sqrt{Var}$$
- Average deviation:

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

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

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

When is the uncertainty not reduced by sampling more?



Statistical description of data

Probability distribution:

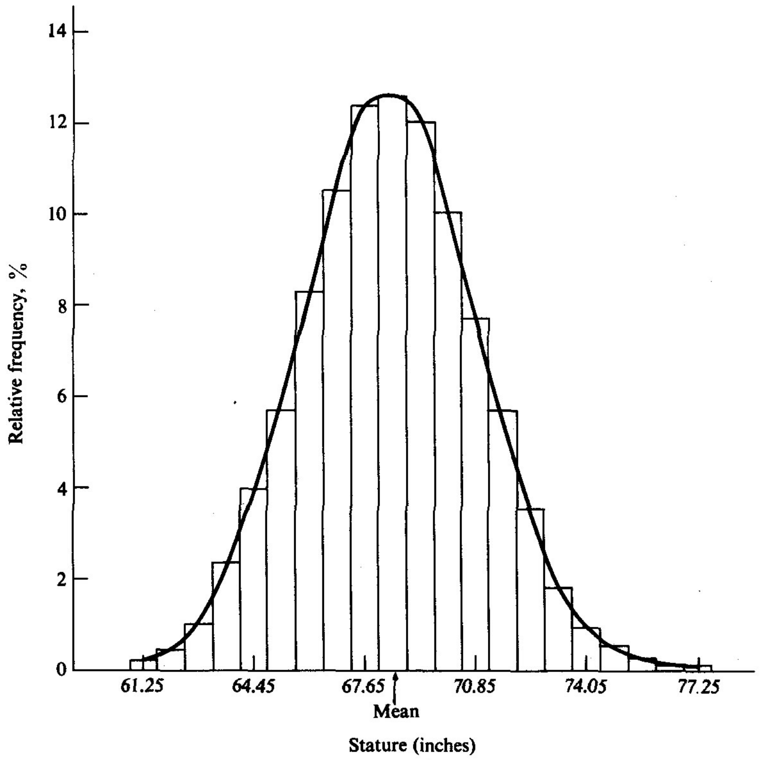


Fig. 1-2 Histogram of frequency distribution of stature of 24,404 U.S. Army males. Adapted from data of Newman and White.

Non-normal probability distribution:

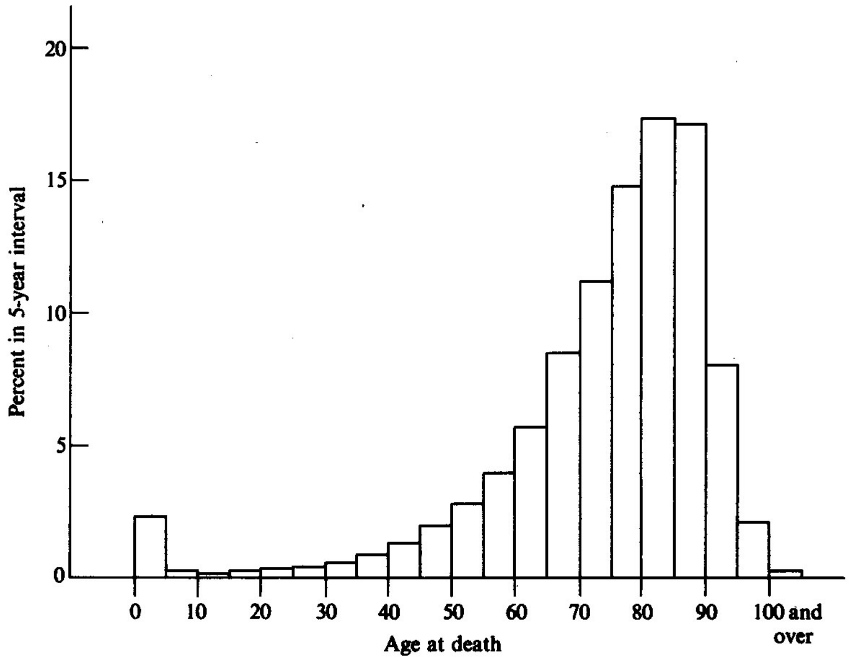


Fig. 1–3 U.S., female, 1965: percent dying in each 5-year age interval (the 100–105 interval includes all deaths after 100 rather than only those occurring in the interval). Data from N. Keyfitz and W. Flieger, *World Population: An Analysis of Vital Data*. Chicago: University of Chicago Press, 1968,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

Statistical description of data

Nonparametric statistics (when the distribution is unknown):

- rank statistics

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

- Median

- percentile

- Deviation estimate

- The mode

Issue: *robustness*

Statistical description of data

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

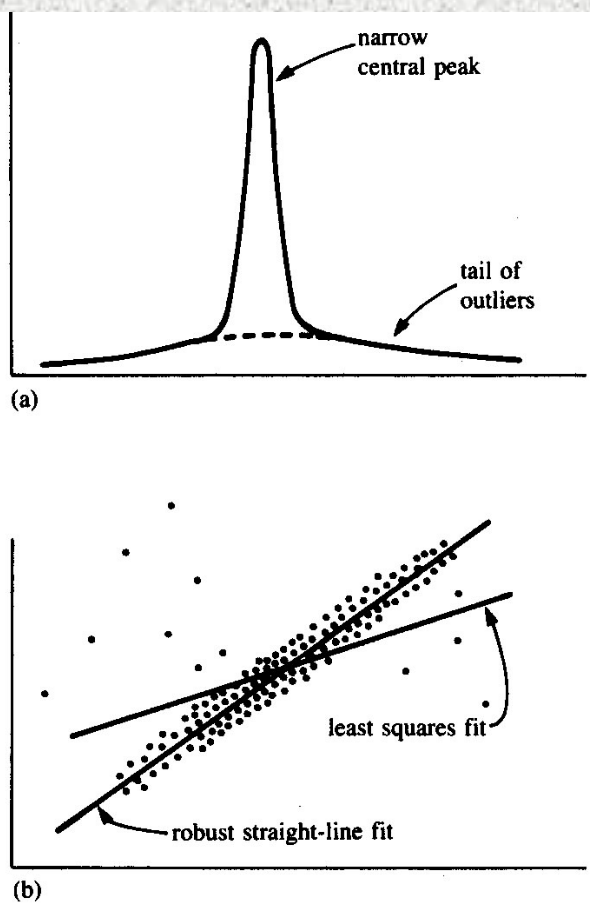
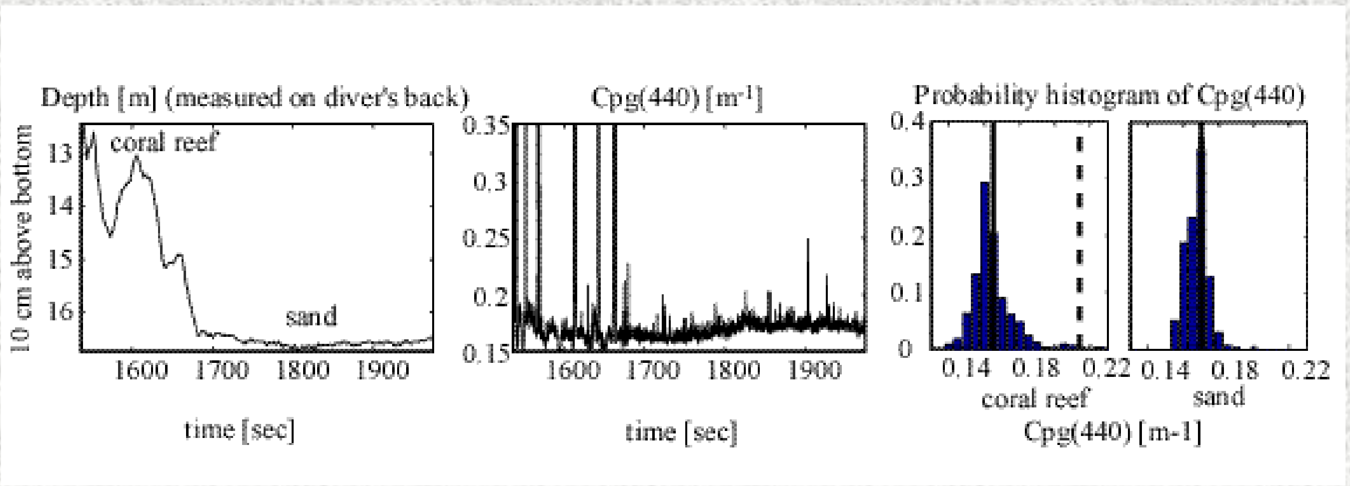
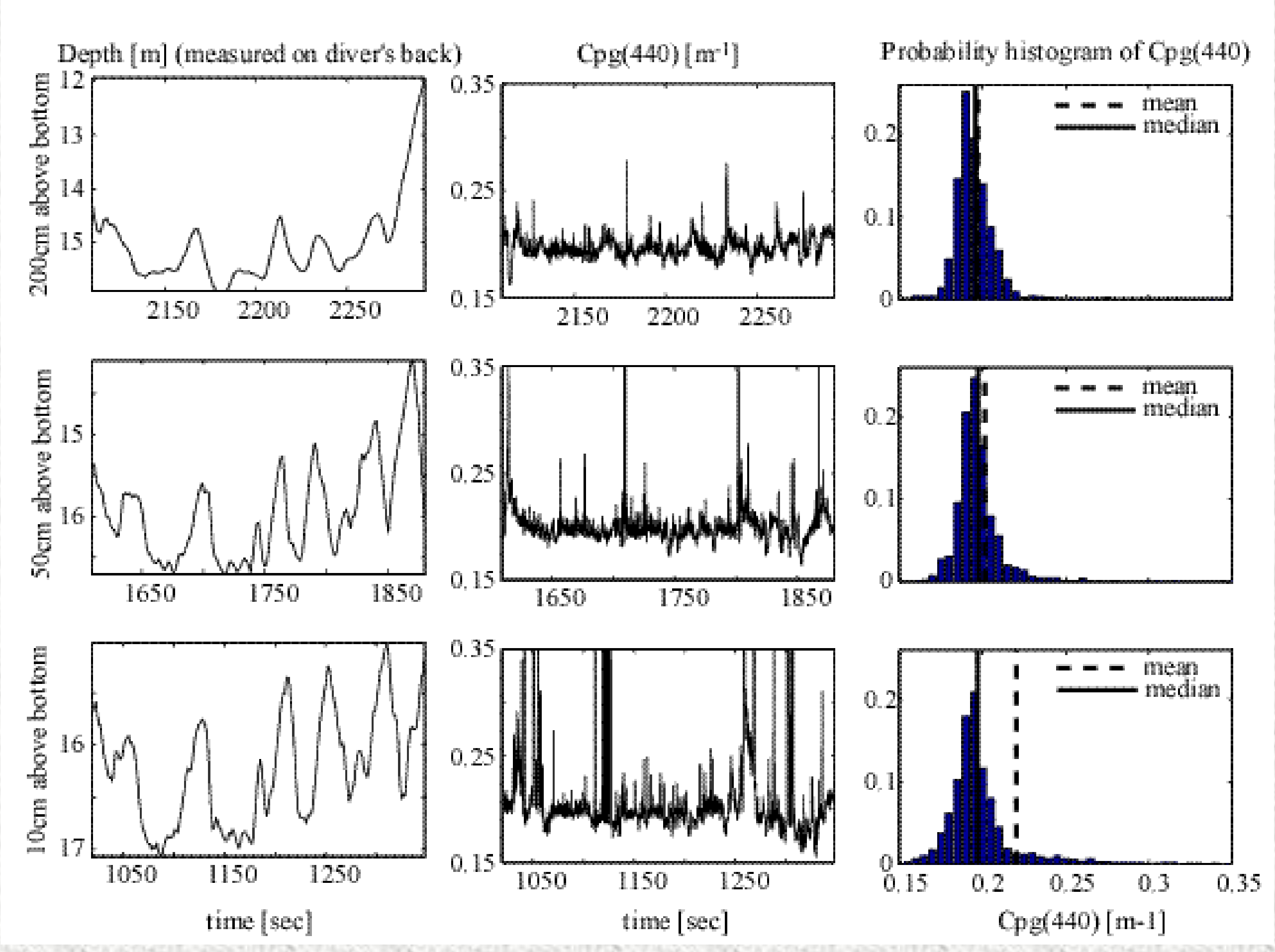


Figure 14.6.1. Examples where robust statistical methods are desirable: (a) A one-dimensional distribution with a tail of outliers; statistical fluctuations in these outliers can prevent accurate determination of the position of the central peak. (b) A distribution in two dimensions fitted to a straight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

Statistical description of data

Examples from COBOP:





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:

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

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?

Press et al.'s approach:

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

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

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

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

The coefficient of determination

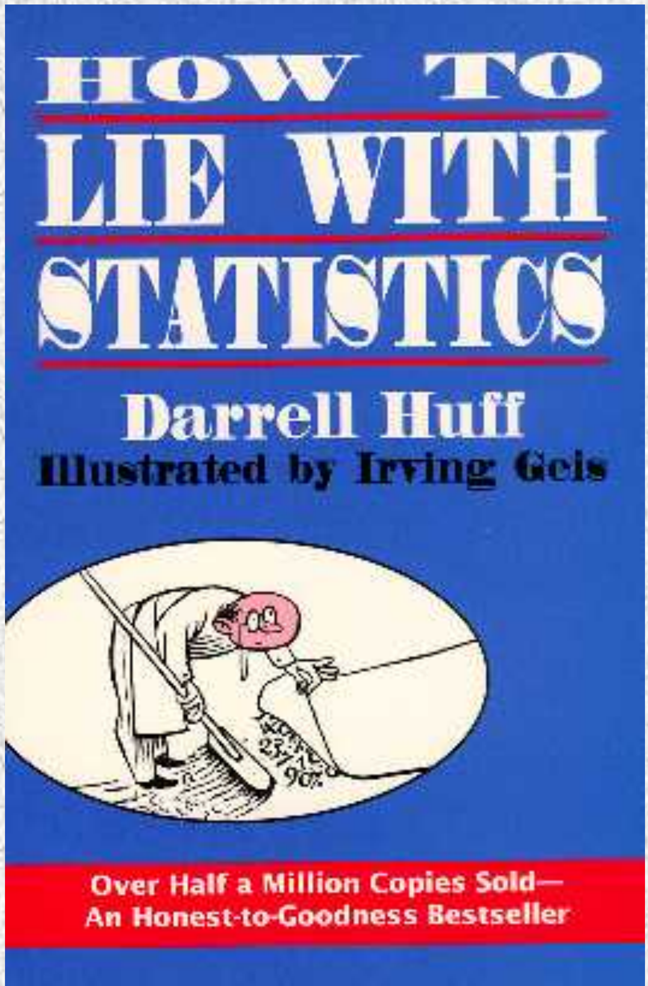
$R^2 = 1 - MSE/Var(y)$.

What variance does it explain?

Can it reveal cause and effect?

How is it affected by dynamic range?

When is R related to 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

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.

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

$$\Rightarrow \boldsymbol{a} = \lfloor \tilde{a}_g , s \rfloor$$

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

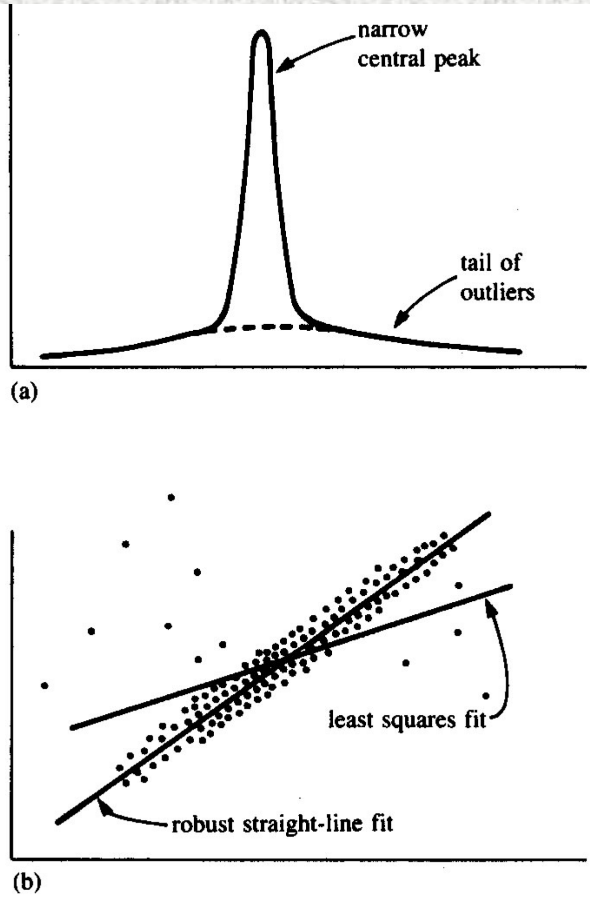


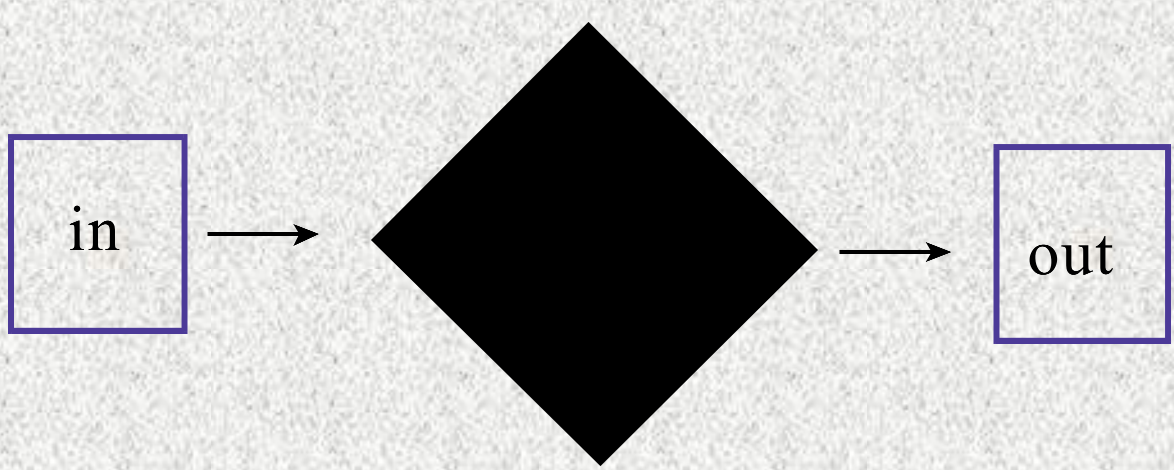
Figure 14.6.1. Examples where robust statistical methods are desirable: (a) A one-dimensional distribution with a tail of outliers; statistical fluctuations in these outliers can prevent accurate determination of the position of the central peak. (b) A distribution in two dimensions fitted to a straight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

Monte-Carlo/Bootstrap methods

Need to establish confidence intervals in:

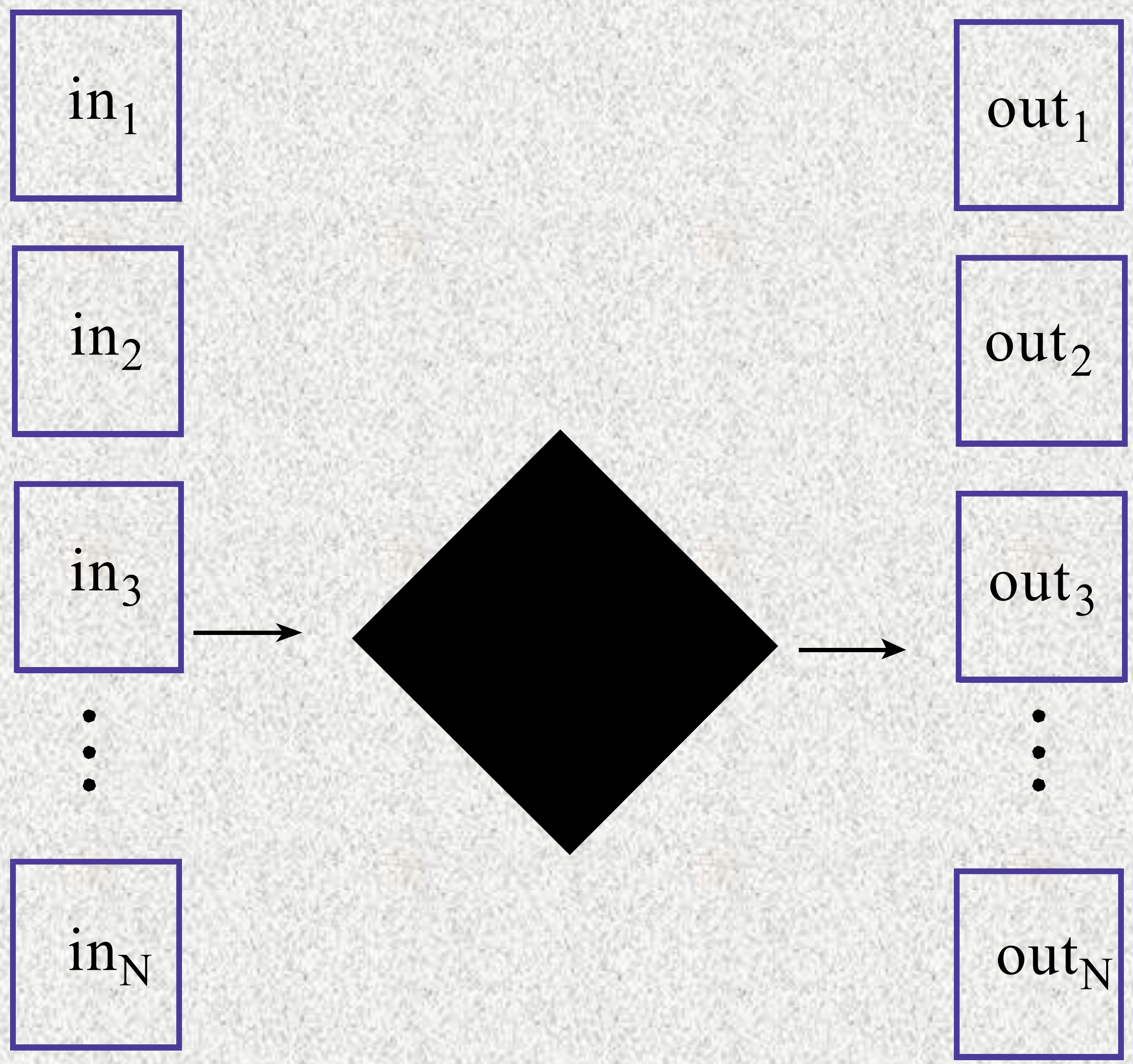
1. Fitting-model parameters (e.g. CDM fit).

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

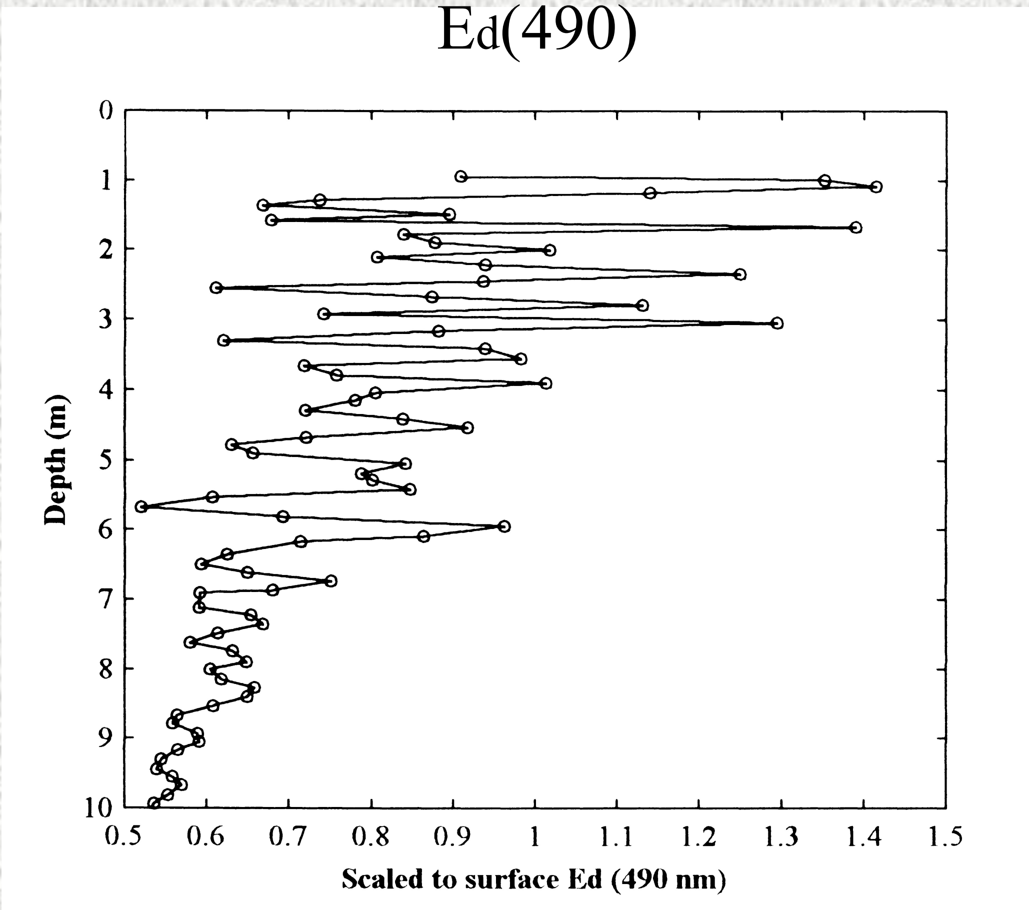


When there is an uncertainty (or possible error) associated with the input:

Vary inputs with random errors and observe effect on output:

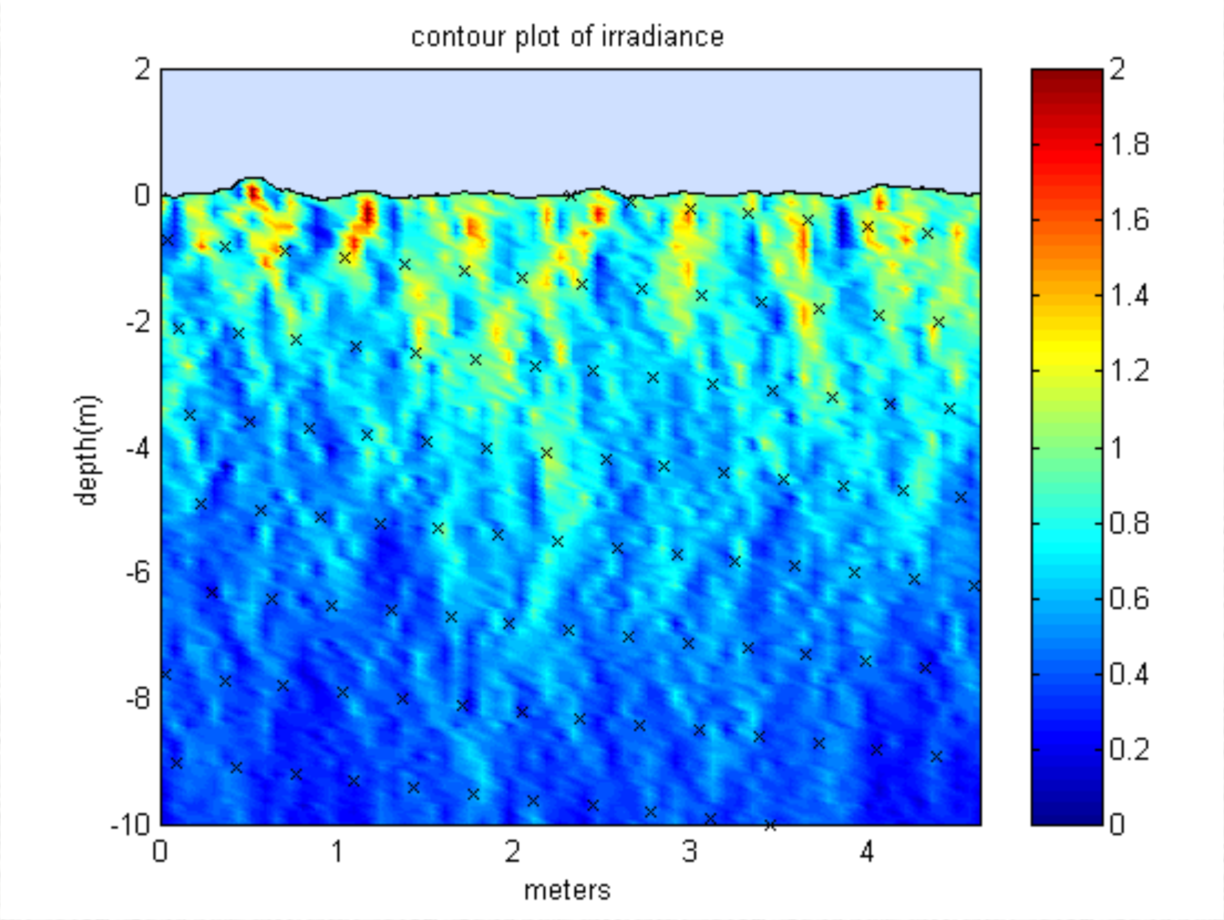


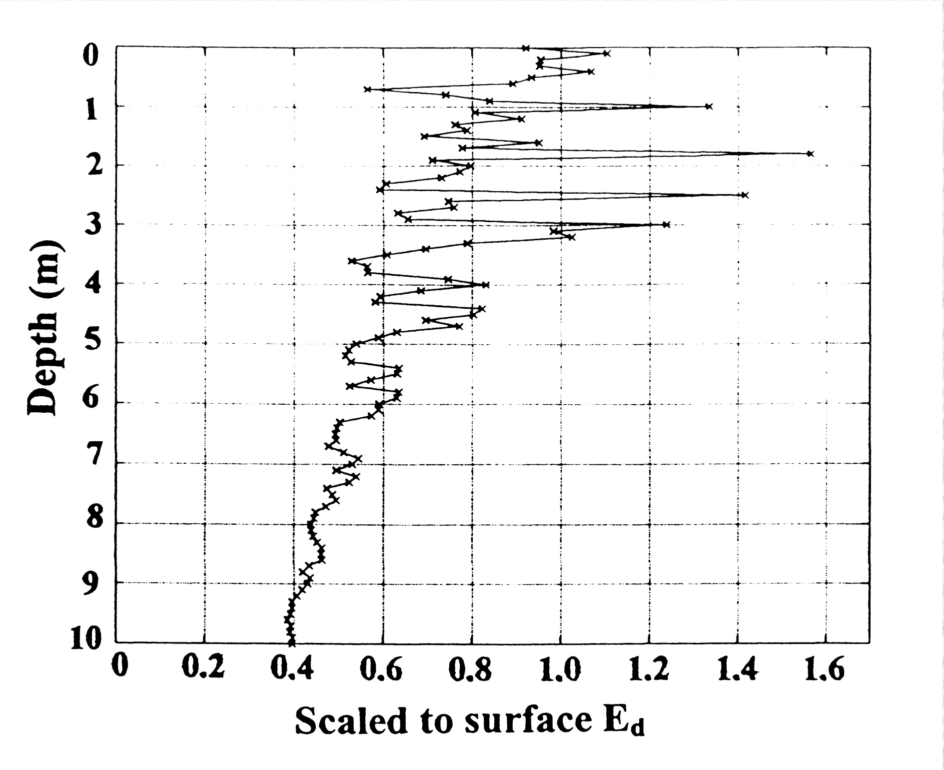
Fitting K_d to ‘noisy’ downwelling irradiance data:



Oregon coast, Wind<2m/sec

Model results:





Fitting K_d to 'noisy' downwelling irradiance data:

Assume:

$$\langle E_d(z)\rangle = \langle E_d(0)\rangle \exp(-K_d z)$$

K_d is what we are after.

Define:

$$I(z) \equiv \int_{z_0}^{z} E_d(z)dz =$$

$$\frac{\langle E_d(0)\rangle}{K_d}\left[\exp(-K_d z_0) - \exp(-K_d z)\right]$$

Where z_0 is in the zone of no/little fluctuations.

Fitting K_d to 'noisy' downwelling irradiance data:

$$I(z) \equiv \int_{z_0}^{z} E_d(z')dz' =$$

$$\frac{\langle E_d(0) \rangle}{K_d}\left[\exp(-K_d z_0) - \exp(-K_d z)\right]$$

Nonlinearly fit to the integrated irradiance data:

$$I(z;[A, B, K_d]) = A + B\exp(-K_d z)$$