Lecture 4:

- Student's t
- Correlation test
- Non-parametric tests
- What is 'Normal?'
- Robust parameter estimation

Student's t

(Often misinterpreted as referring to being from or for "a student," rather than the fact that the name of the author happens to be "Student")

(Except this was actually a pseudonym used by William Sealy Gosset in his 1908 paper, who was couching himself as "a student"!)

Recall that the rms deviation in the estimated mean from a set of n samples is given by :

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

← rms of the full distribution.

But what if we don't know σ *a priori* and all we have are the sampled estimators?

$$t \equiv \frac{\bar{x} - \mu}{\left(s/\sqrt{n}\right)}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

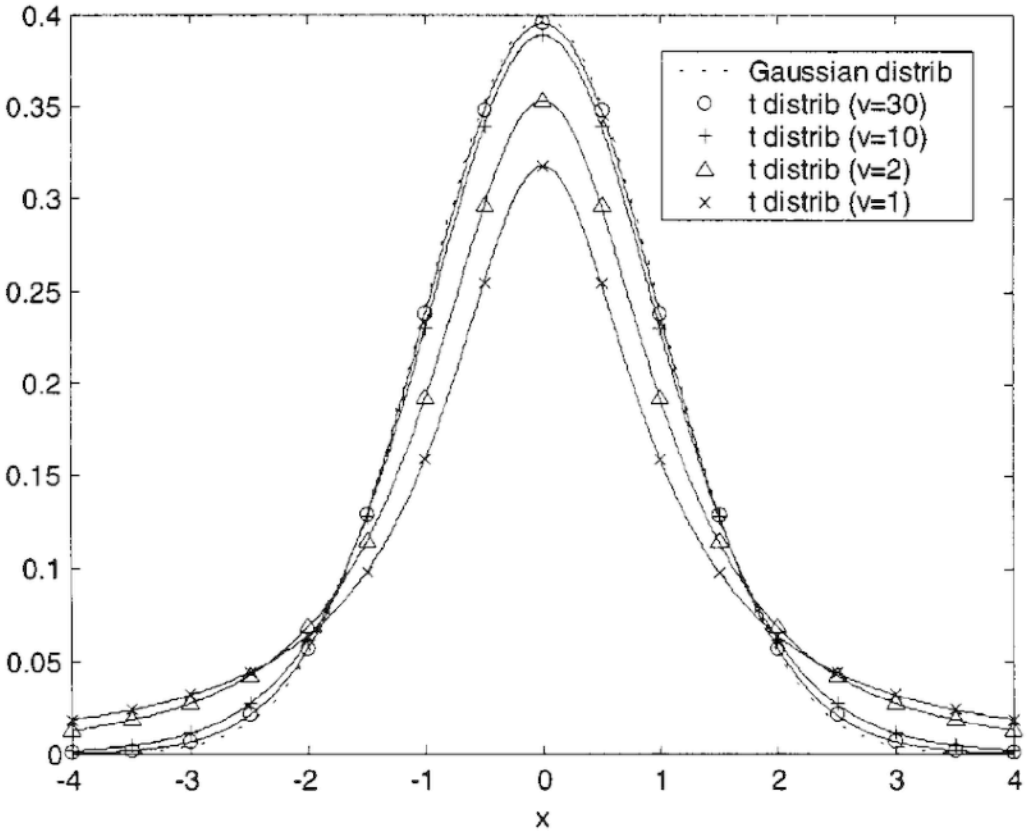
Want to find the distribution of t

$$t \equiv \frac{\bar{x} - \mu}{\left(s/\sqrt{n} \right)}$$

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

ν = # degrees of freedom

As you would expect, this approaches the shape of a Gaussian distribution as the sample size grows:



Pearson Correlation Coefficient

A test of linear correlation between two sets of data

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2} \sqrt{\sum_i y_i^2 - n\bar{y}^2}}$$

This is just the covariance normalised by the sample rms deviations.

The value of this quantity runs from 1 (completely correlated) to -1 (completely anti-correlated), with zero indicating no correlation.

The statistics provides a *relative* measure of linear correlation but, in general, the probability distribution for r will depend on the distributions of x and y.

IF x and y are uncorrelated and each drawn from a normal distribution (such that, jointly, they can be described by a 2-D Gaussian), then:

$$\sigma_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

DoF for 2 free parameters in linear fit

From which it is possible to define a t statistic for r:

$$t_r = r\sqrt{\frac{n - 2}{1 - r^2}}$$

Spearman Rank-Order Correlation Coefficient

A non-parametric test of correlation between two sets of data (*i.e.* linearity is not assumed)

Define R_i as the 'rank' of x_i (i.e. the numerical position in an ordered list of the n data points from lowest to highest x value).

Define S_i as the 'rank' of y_i (i.e. the numerical position in an ordered list of the n data points from lowest to highest y value).

Then define the rank coefficient as:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{R_i - \bar{R}}{s_R} \right) \left(\frac{S_i - \bar{S}}{s_S} \right)$$

Similarly, the probability distribution can be approximated by the t statistic:

$$t_r = r \sqrt{\frac{n-2}{1-r^2}}$$

Generally pretty good and no longer depends on the actual distributions of x & y

Kolmogorov-Smirnov (and the like)

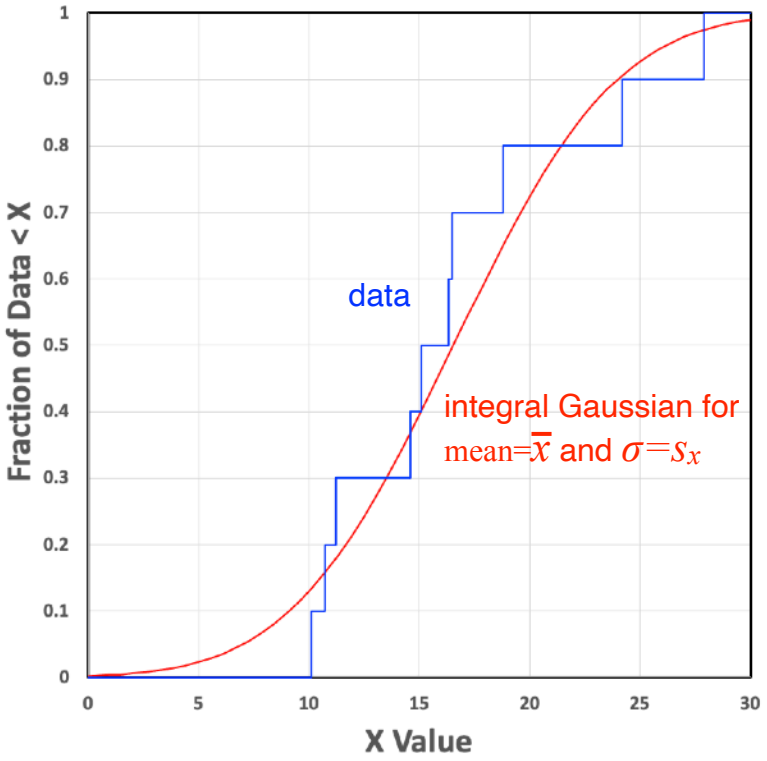
A non-parametric test of distributions

Plot the cumulative fraction of events less than or equal to a particular value of x as a function of x, along with the cumulative distribution for some model:

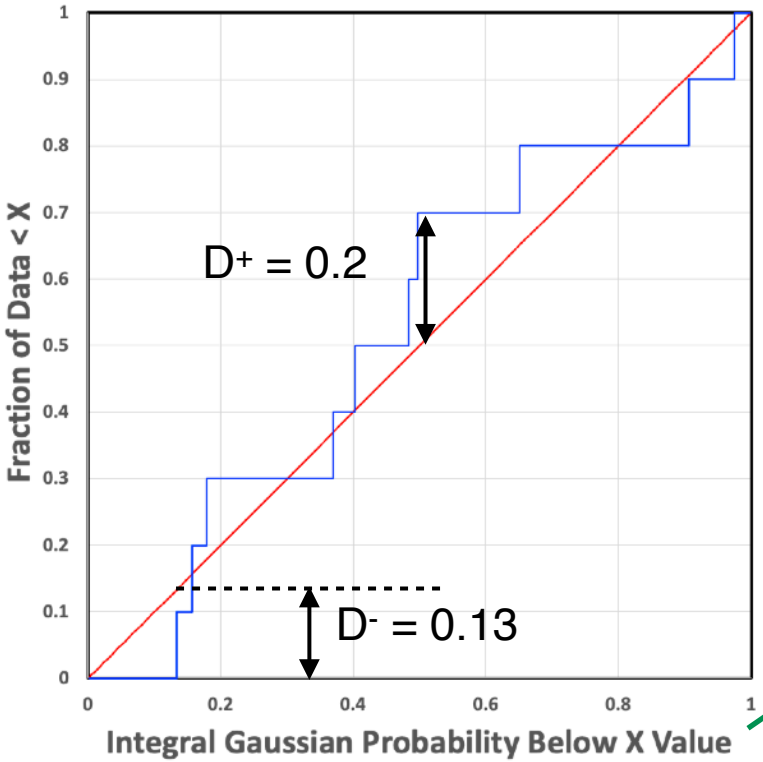
data point	x value	fraction ≤ x
1	10.1	0.1
2	10.7	0.2
3	11.2	0.3
4	14.6	0.4
5	15.1	0.5
6	16.3	0.6
7	16.5	0.7
8	18.8	0.8
9	24.2	0.9
10	27.9	1.0

$$\bar{x} = 16.54$$
$$s_x = 5.8$$

For example: Is this data Normally distributed?



Equivalently:



A more clearly defined span on the x-axis with a visually simple model expectation that is independent of the test distribution

y_i = integral of model probability distribution below the value of x_i

There are several statistics that can be used to assess the level of agreement:

K-S statistics (Clustering)

D^+ = maximum positive deviation from the model line

D^- = maximum negative deviation from the model line

$D = \max(D^+, D^-)$

$V = D^+ + D^-$

Cramer-von Mises (Variance)

$$W^2 = \sum_{i=1}^{n} \left(y_i - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}$$

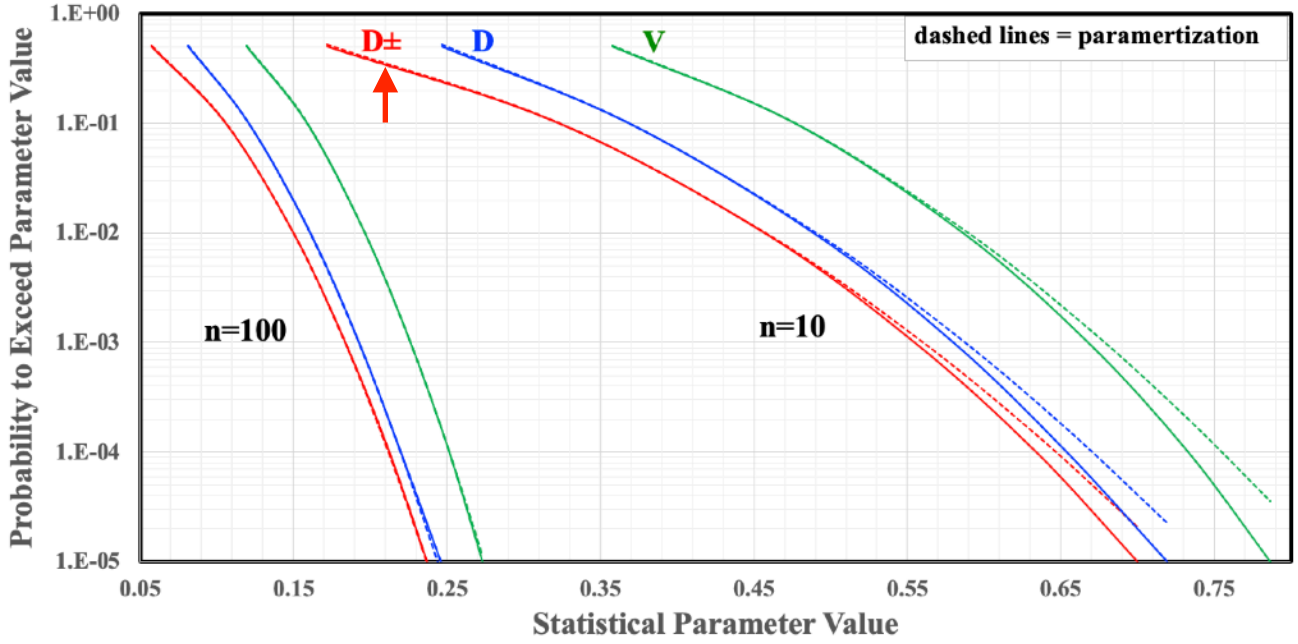
$$U^2 = W^2 - n\left(\bar{y} - \frac{1}{2} \right)^2$$

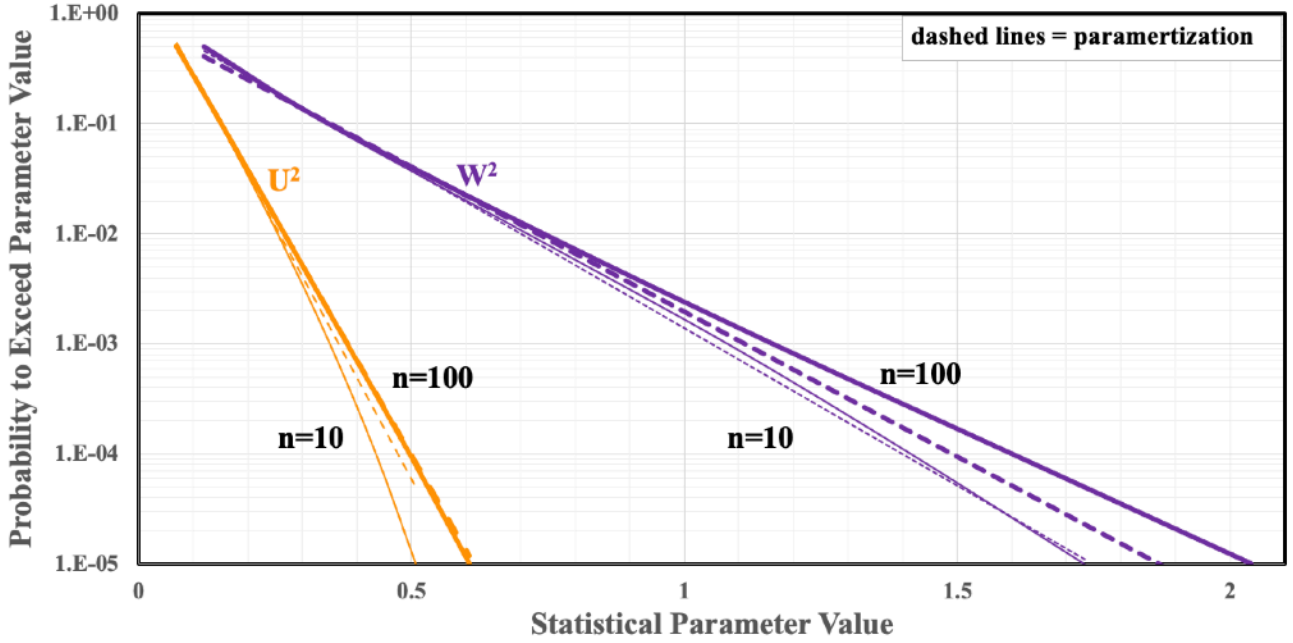
In general, the probability distributions for these statistics need to be determined by Monte Carlo calculations. However, for continuous variables tested against a well-defined model distribution under the null hypothesis, tables and approximate parameterisations exist to obtain p-values:

Test Statistic (T)	Modified Test Statistic (T*)	"High Tail" Approximate Parameterisation for P(T* > z)
D^+	$D^+(\sqrt{n}+0{\cdot}12+0{\cdot}11/\sqrt{n})$	$\exp(-2z^2)$
D^-	$D^-(\sqrt{n}+0{\cdot}12+0{\cdot}11/\sqrt{n})$	$\exp(-2z^2)$
D	$D(\sqrt{n}+0{\cdot}12+0{\cdot}11/\sqrt{n})$	$2\exp(-2z^2)$
V	$V(\sqrt{n}+0{\cdot}155+0{\cdot}24/\sqrt{n})$	$(8z^2-2)\exp(-2z^2)$
W^2	$(W^2-0{\cdot}4/n+0{\cdot}6/n^2)(1{\cdot}0+1{\cdot}0/n)$	$0{\cdot}05\exp(2{\cdot}79-6z)$
U^2	$(U^2-0{\cdot}1/n+0{\cdot}1/n^2)(1{\cdot}0+0{\cdot}8/n)$	$2\exp(-2z^2\pi^2)$

trial factor for choosing best!

M.A. Stephens, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 32, No. 1. (1970), pp. 115-122


Probability to Exceed Parameter Value
1.E+00
1.E-01
1.E-02
1.E-03
1.E-04
1.E-05
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
Statistical Parameter Value
D±
D
V
dashed lines = paramertization
n=100
n=10


Probability to Exceed Parameter Value
1.E+00
1.E-01
1.E-02
1.E-03
1.E-04
1.E-05
0
0.5
1
1.5
2
Statistical Parameter Value
dashed lines = paramertization
U²
W²
n=100
n=10
n=100
n=10

Is That Normal?

We frequently encounter vague statements about the assumption that distributions are “sufficiently Normal,” but exactly what does that mean and how do you check that things are Normal enough?

It depends on what you're trying to do:

- For example, if you're fitting a function to a set of data, so long as the probability distributions for the data points are reasonably symmetric and tails are not very large, the derived central values for the fit parameters will generally be pretty good.

- If you want to make a precise measurement and quote Gaussian error bars, the probability distribution for the parameters should be Normal to at least ~2σ or more, as this is a tacit assumption by the reader when you quote ±1σ error bars. If this is not the case, the details should be given.

- If you want to exclude models at high confidence based on Gaussian error bars, the relevant distribution should obviously be Normal to at least that confidence level.

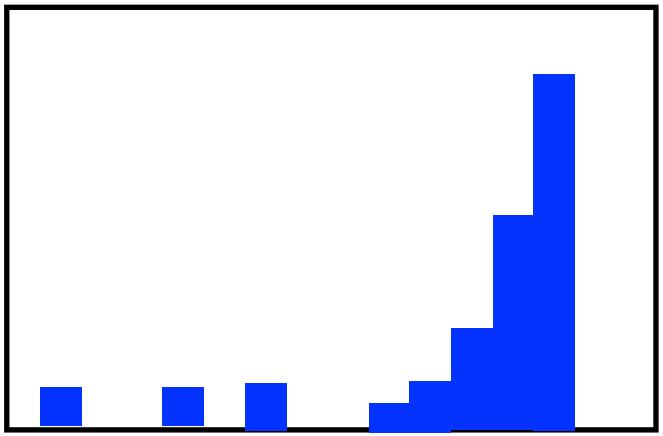
Note: the requirement on the precise Gaussian nature of individual data points may be less restrictive, since the variance of fit parameters generally arises from the accumulation of smaller deviations from the data points.

So the nature of Gaussian requirements is necessarily pragmatic, but is generally logically straight-forward.

The real issue is about:

1) Notably asymmetric distributions that can lead to systematic biases

2) Long distribution tails, whereby large deviations from the expected mean ("outliers") occur much more frequently than assumed, which can skew fits and lead to misinterpretation.



Goodness of fit parameters, such as chi-squared, can be useful indicators of issues, but these don't catch everything and won't diagnose the issue

It is always advisable to look at the distributions!

But how do you deal with very large and complex data sets, where visually inspecting every distribution is not very practical?

- If distributions are symmetric, then mean = median = peak (mode)

- Different ways to compute the standard deviation:
  1) Perform an explicit Gaussian fit
  2) Compute the sampled RMS deviation
  3) Find the peak and then the FWHM = 2.35σ for Gaussian
  4) The central ±1σ should contain 68% of the events

Building some of these checks into your analysis is an extremely useful way to flag potential issues that warrant further investigation

Robust Parameter Estimation

The idea is to minimise the effect of distribution tails and asymmetries on the determination of derived parameters

For example, the median (or 50th percentile) is much more robust in this regard than the mean:

n odd → $x_{med} = x_{(n+1)/2}$

n even → $x_{med} = \frac{x_{n/2} + x_{n/2+1}}{2}$

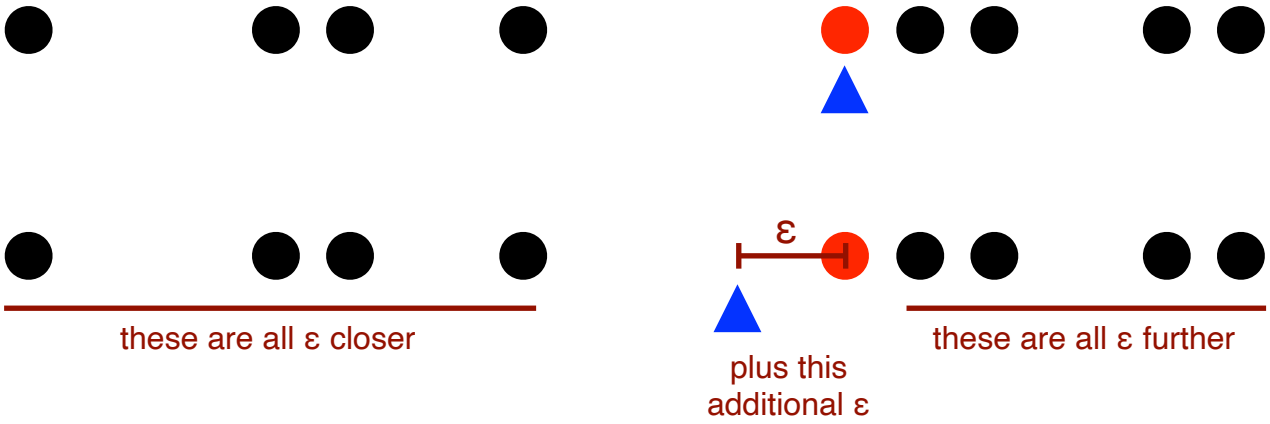
In general, distribution percentiles are robust. So, for example, one could define an equivalent distribution "width" by the 84th percentile (*i.e.* the value below which contains 84% of the distribution) minus the 16th percentile to give a region containing 68% of the distribution (roughly ±1σ for a Gaussian distribution) centred on the median.

A fit to parameters based on minimising the sum of RMS deviations provides an unbiased estimator for the mean:

$$\frac{d}{d\alpha}\sum_{i=1}^{n}(x_i-\alpha)^2=0 \quad =-2\sum_{i=1}^{n}(x_i-\alpha)=-2\left[\sum_{i=1}^{n}x_i-n\alpha\right]$$

$$\alpha=\frac{1}{n}\sum_{i=1}^{n}x_i$$

To instead provide an unbiased estimator for the median, minimise with respect to the sum of the absolute deviations:



So the minimum sum of absolute deviations finds the median!

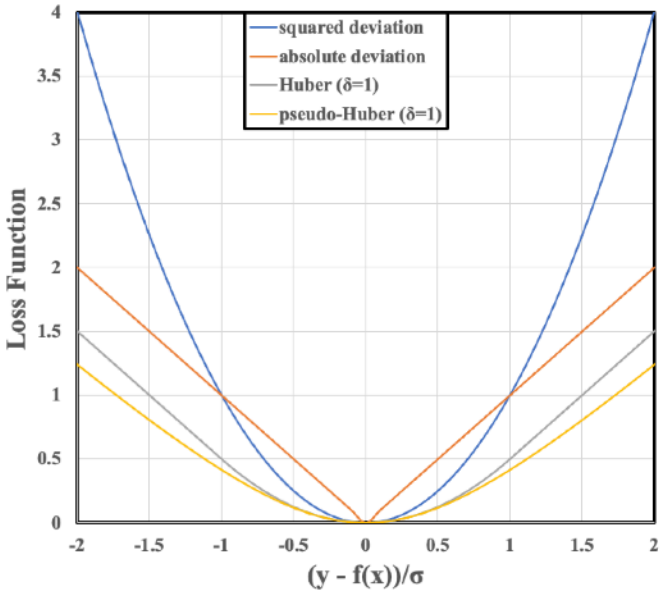
In general, the function to be minimised in order to find the best set of parameters is called the "Loss Function"

An alternative loss function suggested by Huber* provides smooth convergence in the vicinity of the minimum, while maintaining robustness from the distribution tails:

$$L_\delta = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(y_i - f(x_i)\right)^2 \qquad |y_i - f(x_i)| \le \delta$$

$$L_\delta = \frac{1}{n}\sum_{i=1}^{n}\delta\left(|y_i - f(x_i)| - \frac{\delta}{2}\right) \qquad |y_i - f(x_i)| > \delta$$

where δ is a tuneable parameter that would equal σ for a Gaussian distribution



A "Pseudo Huber Loss Function" provides a more convenient form that has continuous derivatives at all degrees:

$$L_\delta = \frac{1}{n}\sum_{i=1}^{n}\delta^2\left(\sqrt{1+\left(\frac{y_i - f(x_i)}{\delta}\right)^2} - 1\right)$$

*P. Huber, "Robust Estimation of a Local Parameter," Ann. Math. Statist. 35(1): 73-101 (1964)