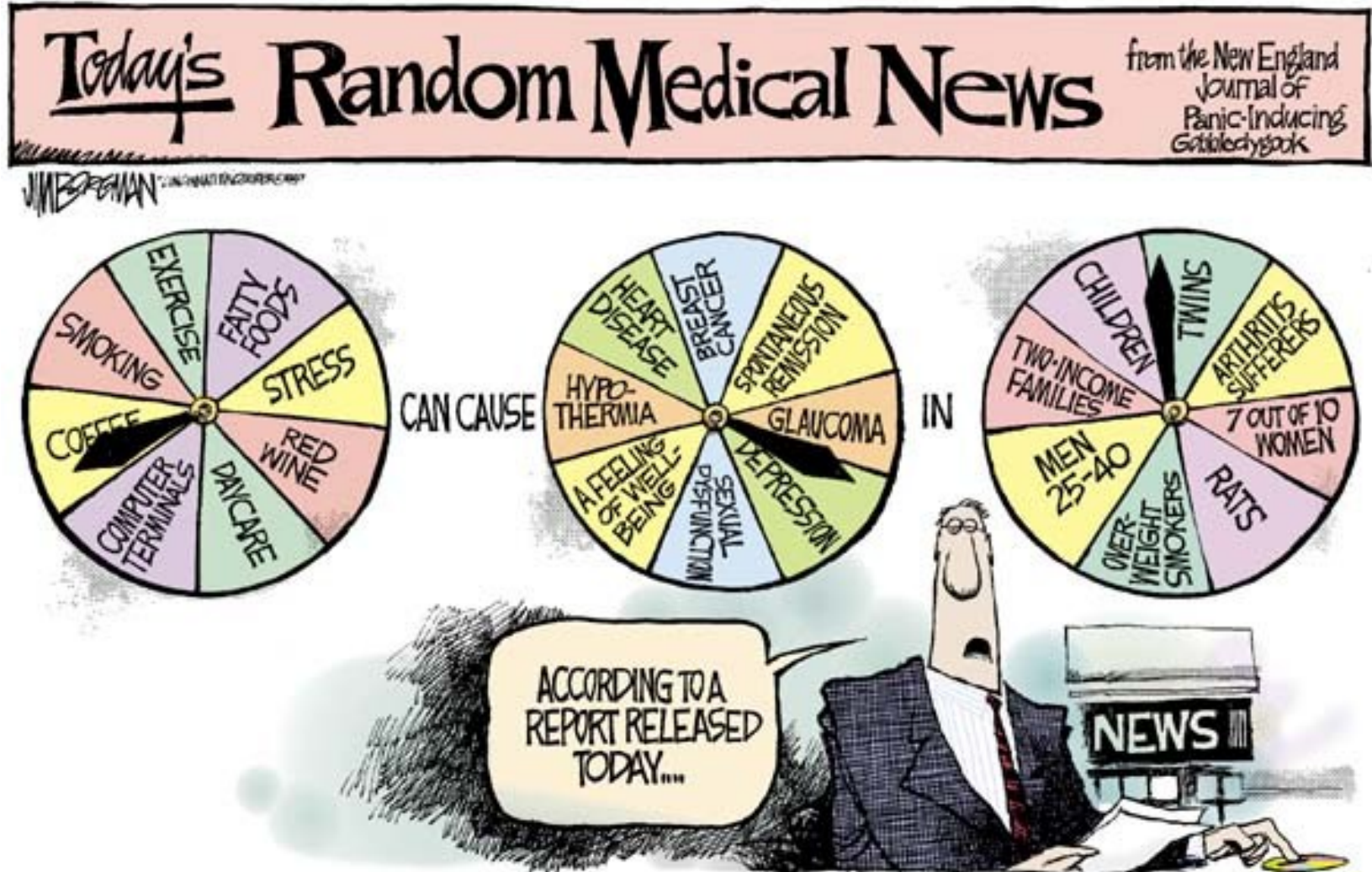


The importance of statistics And error analysis



An Introduction to Error Analysis

The Study of Uncertainties in Physical Measurements



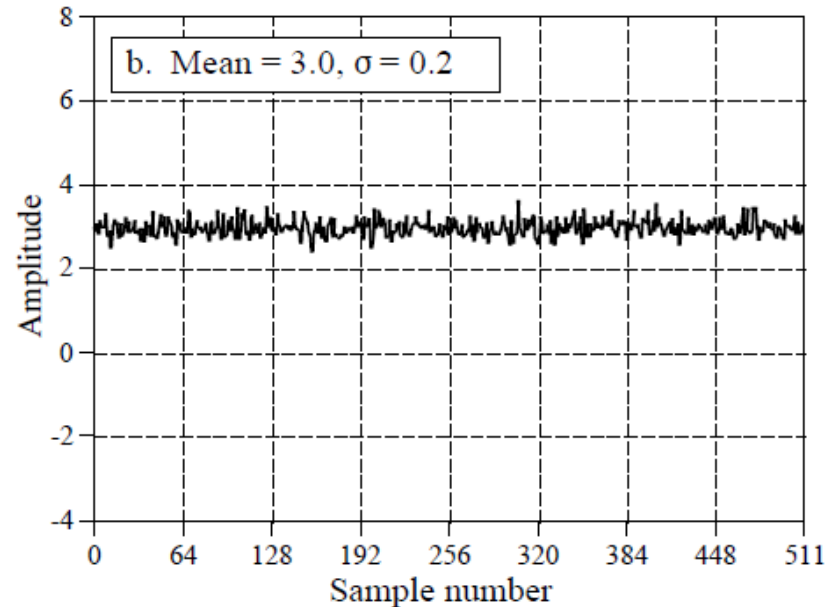
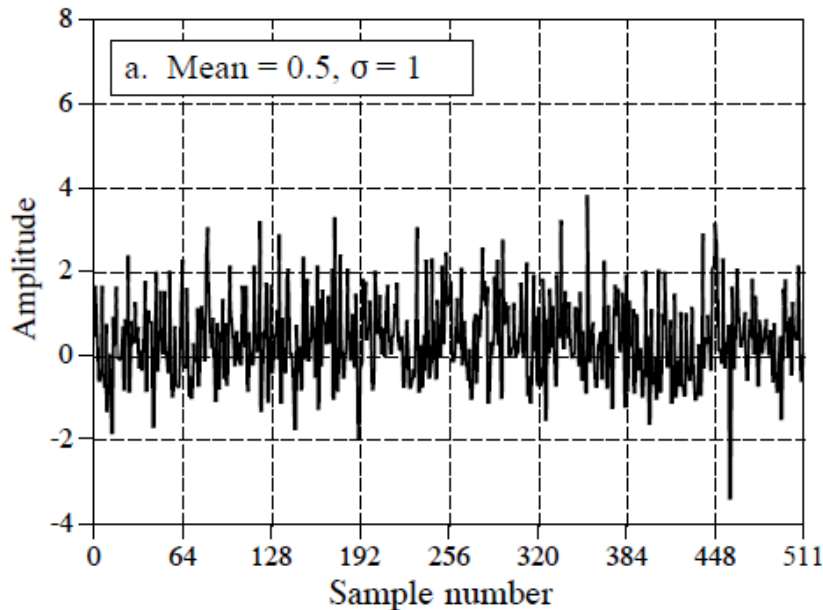
John R. Taylor

Errors and Data Analysis

Types of errors:

- 1) **Precision errors** – these are random errors. These could also be called repeatability errors. They are caused by fluctuations in some part (or parts) of the data acquisition. These errors can be treated by statistical analysis.
- 2) **Bias errors** – These are systematic errors. Zero offset, scale errors (nonlinear output vs input) , hysteresis, calibration errors, etc. If these are hidden, they are essentially impossible to correct. These are often negligible in instruments used for calibration for a long time. But new instruments and devices can easily have bias errors. For instance, when reducing scales from meters and millimeters to a scale of nanometers bias errors can creep in due to unforeseen new effects.
- 3) **Analysis errors** – wrong theory or wrong analysis applied to data, which are used to "fit" the data. This is usually not considered as a error in the data acquisition, but nevertheless can waste a lot of time.

Examples of a constant signal and random noise from time acquired data



Where does the “randomness” come from?

- Counting statistics – small numbers (radioactive decay and photon counting)
- Electronic noise from an electronic circuit
- Small number fluctuations in number of molecules or nano-sized objects

Some helpful “rules” when dealing with errors of an experimental set-up

- 1: As soon as an error from a particular source is seen to be significantly smaller than other errors present, it is given no further consideration.**
- 2: The major concern of most error analyses is the** quantitative estimate of bias errors, and correction of data accordingly when possible.
- 3: Whenever feasible, precision errors should be estimated** from repeated tests or from observed scatter in graphed results.
- 4: In planning an experiment where it appears that** significant bias errors will be present, an effort should be made to ensure that precision errors are much smaller.

How to handle data samples of multiple measurements taken of the same configuration.

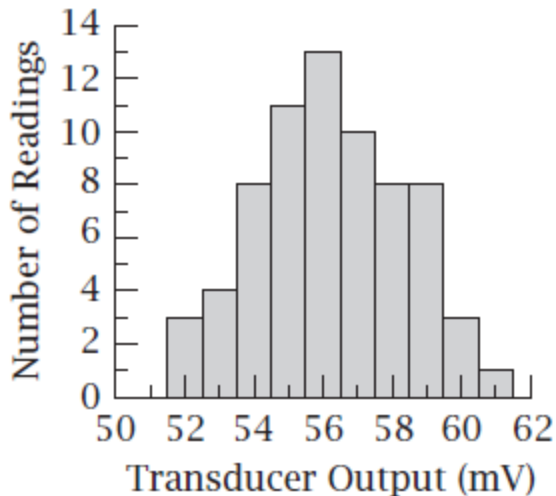
The mean value of the sample values is:

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

The usual measure of the scatter is **the standard deviation**, which is the **square root of the variance**:

$$S_x = \left[\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \right]^{1/2}$$

Example:



Histogram of a large data sample.

Notice that the shape of the histogram is similar to the familiar normal (Gaussian) probability distribution. Indeed, most precision errors have the characteristic that, as the sample size becomes large, the shape of the histogram tends to that of the normal distribution. This characteristic allows many powerful methods of statistical analysis to be applied to the analysis of precision errors.

Running Statistics

Calculation trick using the two definitions for μ and σ :

$$\mu = \frac{1}{N} \sum_{i=0}^{N-1} x_i$$

$$\sigma^2 = \frac{1}{N-1} \sum_{i=0}^{N-1} (x_i - \mu)^2$$

You can show the following, which is a faster way to keep a running calculation of the variance, and has less digital round-off

$$\sigma^2 = \frac{1}{N-1} \left[\sum_{i=0}^{N-1} x_i^2 - \frac{1}{N} \left(\sum_{i=0}^{N-1} x_i \right)^2 \right]$$

or using a simpler notation,

$$\sigma^2 = \frac{1}{N-1} \left[\text{sum of squares} - \frac{\text{sum}^2}{N} \right]$$

While moving through the signal, a running tally is kept of three parameters: (1) the number of samples already processed, (2) the sum of these samples, and (3) the sum of the squares of the samples (that is, square the value of each sample and add the result to the accumulated value). After any number of samples have been processed, the mean and standard deviation can be efficiently calculated using only the current value of the three parameters.

The standard deviation of the mean is:

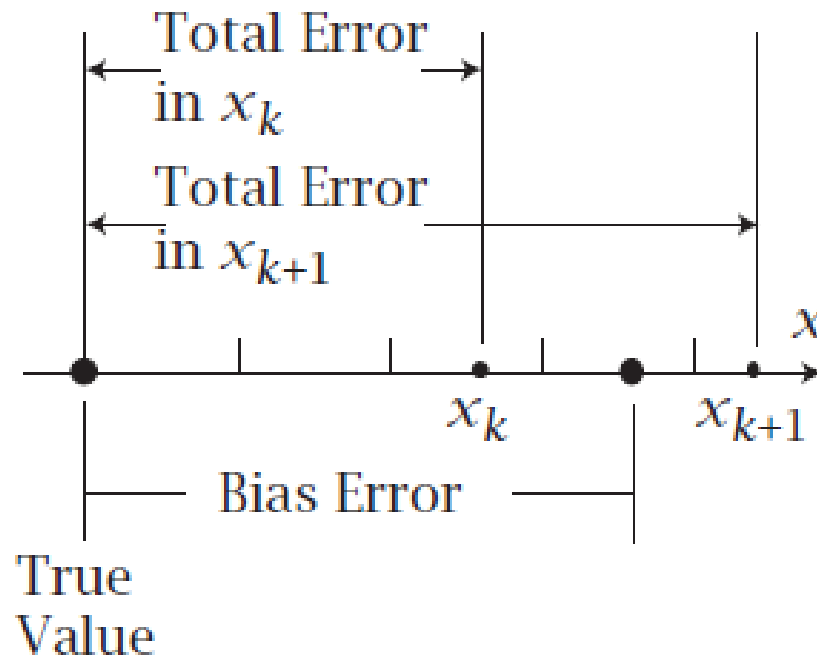
$$S_{\bar{X}} = \frac{S_X}{N^{1/2}}$$

This is NOT the standard deviation of one measurement from the mean of one set of experiments! If the experiment is carried out in many times data sets, and in each set of data many measurements are taken, the standard deviation of the mean values of the sets of data have a much lower standard deviation than the standard deviation of the values of the individual sets. That is, there is always less precision error in a sample mean than in the individual measurements, and if the sample size is large enough the error can be negligible.

Remember this is only for the statistical precision error – NOT the bias error.

A statistical analysis of a sample tells a lot about precision errors, having a sample tells us nothing about *bias errors*.

The total error in a measurement is the difference between the measured value and the true value. BUT we do not know what the true value is! If we take a large enough sample we could say that a good estimate of the bias error is $\bar{x} - x_{true}$. But the catch is that we do not know x_{true} *a priori*: x_{true} is the unknown we want to determine. Thus, determination of bias errors has nothing to do with samples of data and statistical analysis. To find the bias errors you have to compare with data from similar instruments, or with standard measurements, or patiently find the bias in your instrument.



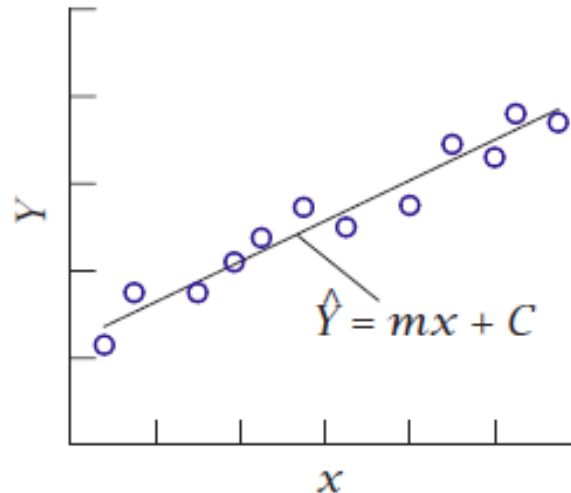
Total and bias errors in a measurement.

How about **least square curve fits** – that is, one parameter depends on another.

Take the example of a **straight line dependence**.

$$y = Mx + C \quad (x_i, y_i); i = 1, 2, \dots, N$$

assume that y has significant precision error,
but the x precision error is negligible



$$\sum D_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - mx_i - C)^2 \quad \text{Sum of squared of differences}$$

How to determine the slope and intercept

$$\sum D_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - mx_i - C)^2$$

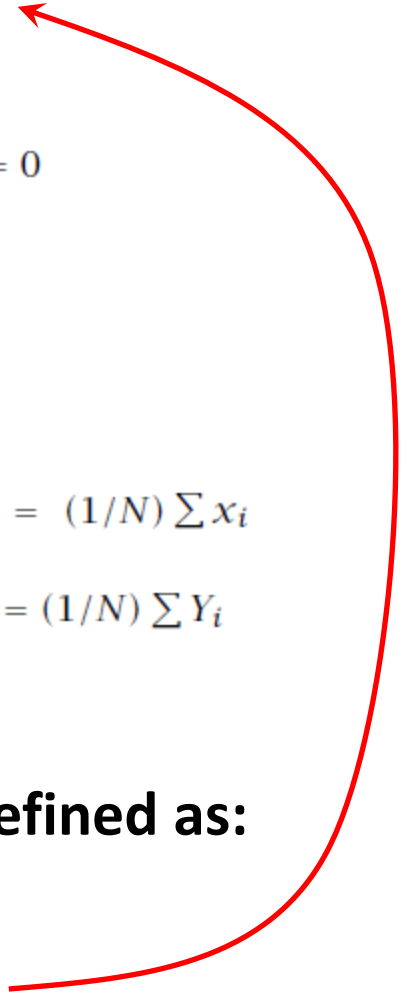
$$\frac{\partial \sum D_i^2}{\partial m} = \sum x_i Y_i - C \sum x_i - m \sum x_i^2 = 0$$

$$\frac{\partial \sum D_i^2}{\partial C} = \sum Y_i - NC - m \sum x_i = 0.$$

$$m = \frac{\sum x_i Y_i - N\bar{x}\bar{Y}}{\sum x_i^2 - (\sum x_i)^2 / N}; \quad \bar{x} = (1/N) \sum x_i$$

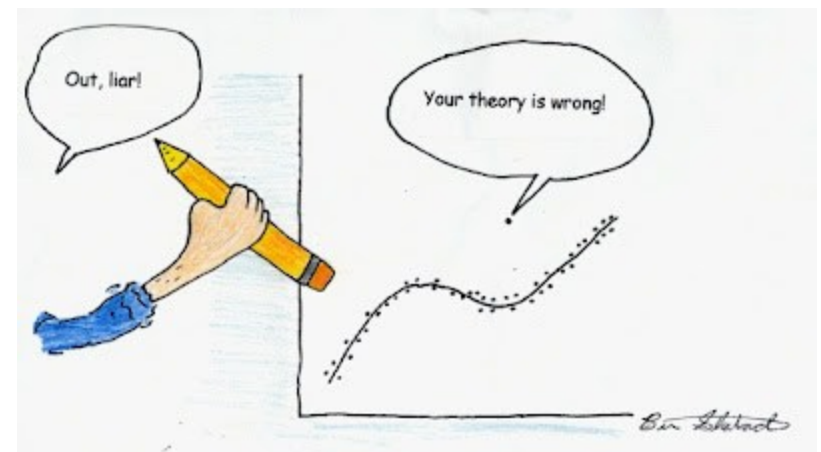
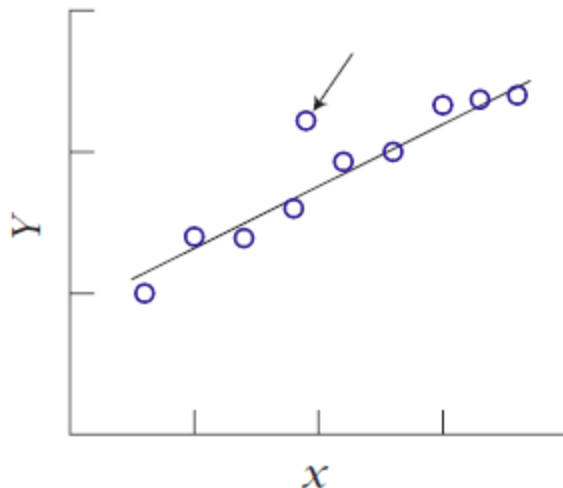
$$C = \frac{\bar{Y} \sum x_i^2 - \bar{x} \sum x_i Y_i}{\sum x_i^2 - (\sum x_i)^2 / N} = \bar{Y} - m\bar{x}. \quad \bar{Y} = (1/N) \sum Y_i$$

Standard error for the curve fit is defined as:

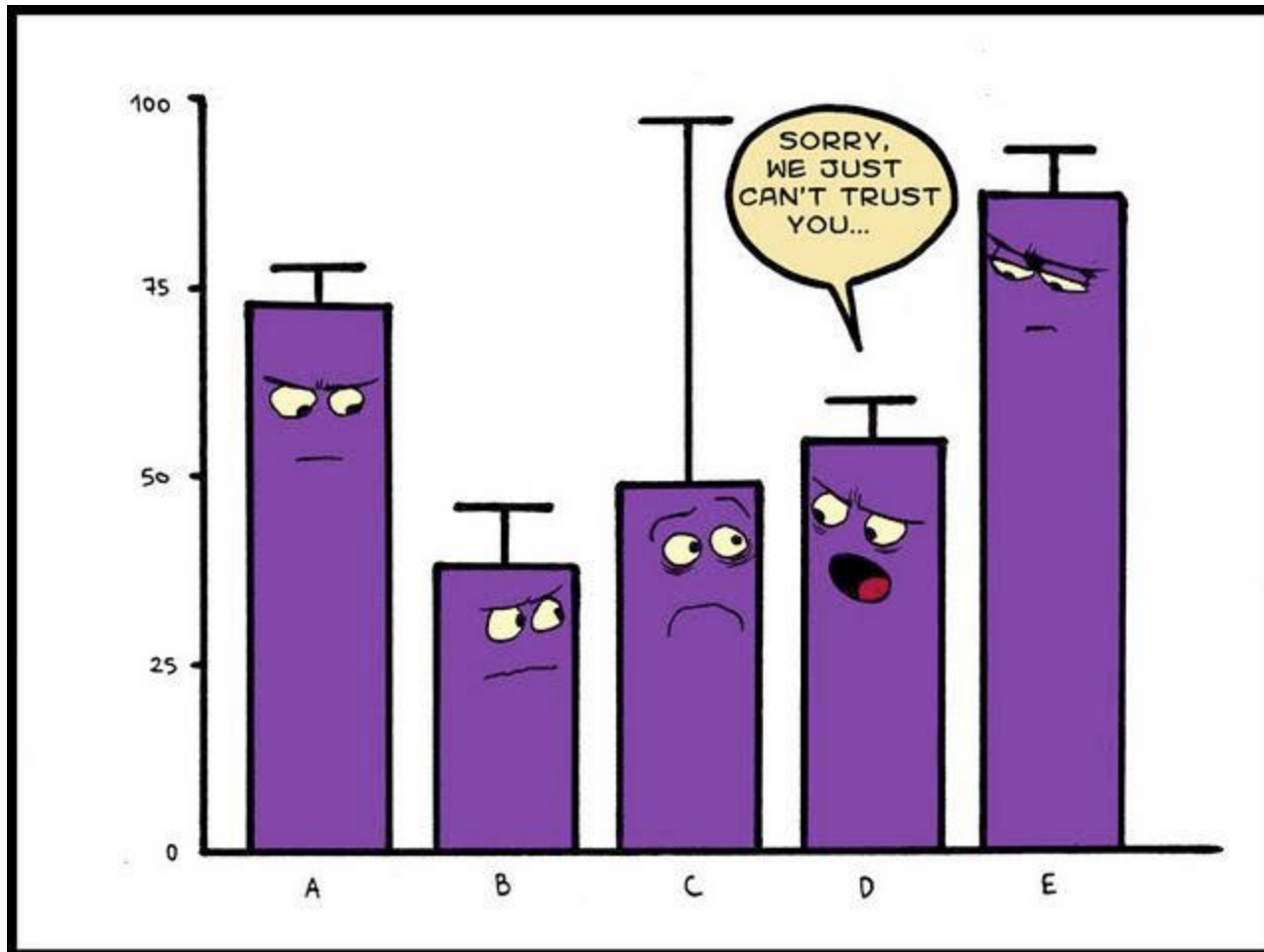
$$S_Y = \left[\frac{1}{N-2} \sum D_i^2 \right]^{1/2}$$


Comments:

- It was assumed that all the variance was in “y”. If “x” also has significant variance, the expressions are more complex.
 - If the plot is seen to be **nonlinear**, maybe we can **linearize** the data: for instance
If $y = ae^{-kx}$, then $\ln(y) = \ln a - kx$; plot $\ln(y)$ vs x ; slope = $-k$, and intercept = $\ln a$.
- If $y = ax^n$; then $\ln y = \ln a + n \ln x$; plot $\ln y$ vs $\ln x$
- Often the data points can be fit to **several models**. If you are testing a theory you know the model; or maybe you are searching for a hint for a theory.
 - How do you handle **outliers** (see figure below and later)?



Another type of “outlier”



Uncertainty

We do not know the actual value of the parameter(s) we are measuring – we only know an estimate of this value. So we have to deal with estimated – or probable - errors. If we say we are C% confident that the true value X_{true} of a measurement X_i lies within the interval $X_i \pm P_x$: then P_x is called the precision uncertainty at a confidence level of C%. This means that if we specify a 95% confidence level estimate of P_x , we would expect X_{true} to be in the interval $X_i \pm P_x$ about 95 times out of a 100.

We **usually assume a normal distribution** if $N > 10$; then P_x is approximately 2x the standard deviation for 95% confidence:

$$P_x \cong 2S_x (C = 95\%, N > 10)$$

This is the uncertainty at 95% confidence for individual samples drawn from a normal population and the total sample is large

For small samples this must be amended – so always try to keep $N > 10$.

Now what about **the precision in the uncertainty of the value of the mean of repeated sets of measurements**, each set consisting of a certain number of individual measurements?

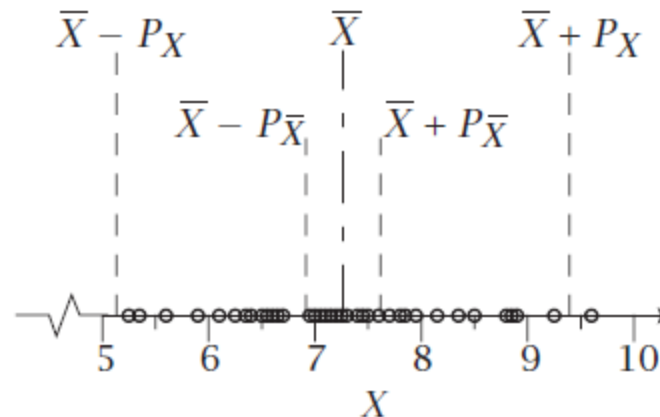
Remember:
$$S_{\bar{x}} \cong \frac{S_X}{N^{1/2}}$$

Then the corresponding **precision uncertainty in the sample mean** is:

$$P_{\bar{X}} \cong 2S_{\bar{X}} \quad (C = 95\%, N > 10)$$

So, The probable error in a sample mean is much less than in the individual measurements. Why is this important?

- We usually average individual measurements over a time interval before recording the averaged values.
- When precision error is important, we usually are interested in the sample mean, not in individual measurements in any particular set of measurements.



Can we know estimates of our error when we only take a single measurement?

Yes, if we have independent data for the variance of the measurement from previous measurements, or from an examination of the instrument from the factory or from control measurements. But in general it is best to take several measurements.

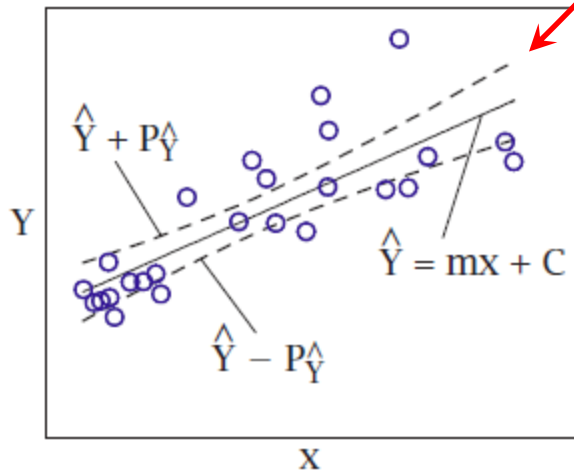
How about the precision error for a curve fit? Then one can show:

$$S_Y = \left[\frac{1}{N-2} \sum D_i^2 \right]^{1/2} \quad S_{xx} = \sum x_i^2 - \left(\frac{1}{N} \right) (\sum x_i)^2$$

\hat{Y} for a curve-fit is like a “mean” value analogous to \bar{X} for a sample of values of a single variable.

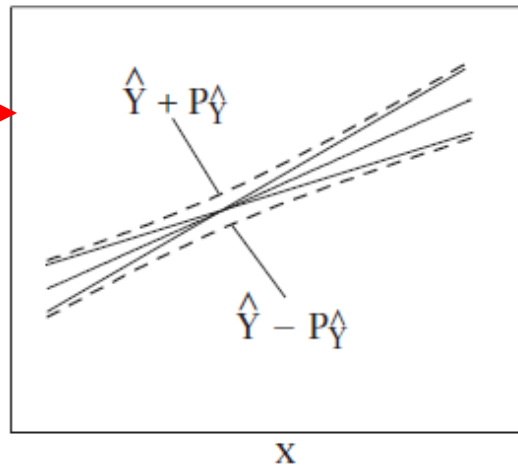
$$P_{\hat{Y}} = 2 \left\{ S_Y^2 \left[\frac{1}{N} + \frac{(x - \bar{x})^2}{S_{xx}} \right] \right\}^{1/2}$$

(C = 95%, N > 10)



$P_{\hat{Y}}$ depends on how far x is away from \bar{x} : it is a minimum at $x = \bar{x}$

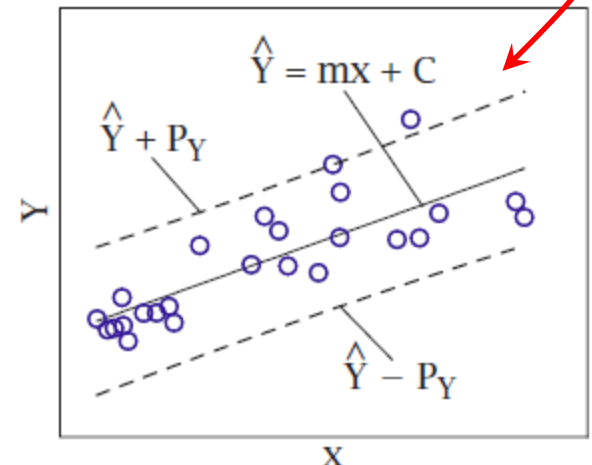
P_Y is always larger than $P_{\hat{Y}}$,
just like P_x is larger than $P_{\bar{x}}$



The range where the curve fits will fall 95% of the time for repeated sets of measurements

$$P_Y = 2 \left\{ S_Y^2 \left[1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{S_{xx}} \right] \right\}^{1/2}$$

(C = 95%, N > 10)



The range in which we are 95% confident a **single data point** will fall

Bias uncertainty differs from precision uncertainty:

- We are usually concerned with the precision uncertainty of a sample mean or a curve-fit.
- Precision uncertainties can be reduced by increasing the number of data points used.
- Bias uncertainty is independent of sample size: it is the same for one data point as for a sample of 100 data points.

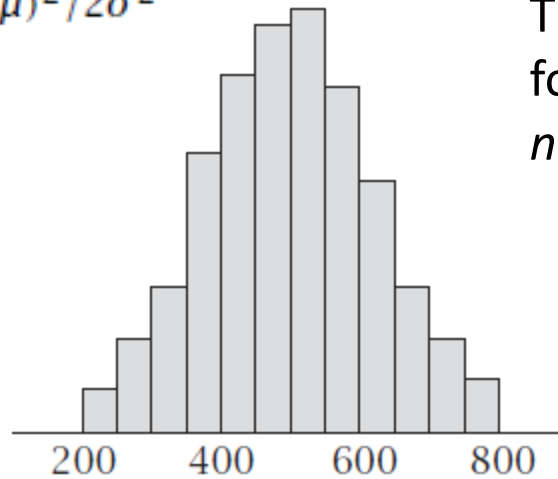
The Normal Probability Distribution

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

$$\int_{-\infty}^{\infty} f(X) dX = 1.0$$

$$\mu = \int_{-\infty}^{\infty} X f(X) dX$$

$$\sigma^2 = \int_{-\infty}^{\infty} (X - \mu)^2 f(X) dX$$



A histogram of a sample from a normal population.

The probability density function for a random variable X having a normal distribution

A single measurement that is assumed to be from a normal parent population.

$$P(\mu - \Delta X \leq X \leq \mu + \Delta X) \longrightarrow z = \frac{X - \mu}{\sigma} \longrightarrow P(-z_1 \leq z \leq z_1) = \frac{1}{\sqrt{2\pi}} \int_{-z_1}^{z_1} e^{-z^2/2} dz$$

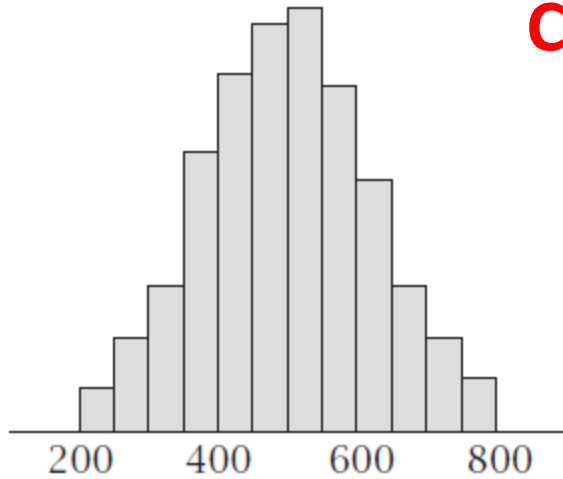
$$= \int_{\mu - \Delta X}^{\mu + \Delta X} \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2} dX$$

$$\frac{1}{\sqrt{2\pi}} \int_{-z_1}^{z_1} e^{-z^2/2} dz = \operatorname{erf} \left(\frac{z_1}{\sqrt{2}} \right)$$

$$\operatorname{erf} z_1 = \frac{2}{\sqrt{\pi}} \int_0^{z_1} e^{-z^2} dz$$

Confidence levels

E.g. probability that a measurement will fall within 1 standard deviation.



$$P(-1 \leq z \leq 1) = \frac{1}{\sqrt{2\pi}} \int_{-1}^1 e^{-z^2/2} dz$$

A histogram of a sample from a normal population.

The probability that a measurement will fall within a certain fraction of standard deviations (σ 's) of the mean:

Probability	Range about mean value
50	$\pm 0.675\sigma$
68.3	$\pm 1.0\sigma$
95	$\pm 1.96\sigma$
99.7	$\pm 3\sigma$
99.99	$\pm 4\sigma$

t-statistics – small number of samples

Remember

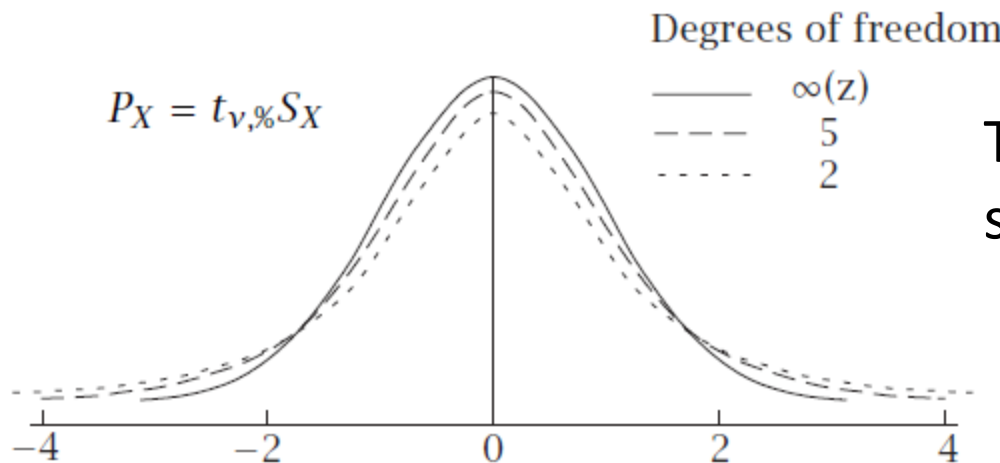
$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad S_x = \left[\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \right]^{1/2} \quad \sigma_x = \frac{\sigma}{N^{1/2}}$$

$N-1 = \nu$ = degrees of freedom

The precision uncertainty P_x of an individual measurement at a confidence level C% is defined such that we are C% sure that the population mean μ lies in the interval $X_i \pm P_x$. BUT we do not know the population standard deviation σ .

ν	$t_{95\%}$	ν	$t_{95\%}$
4	2.770	15	2.131
5	2.571	20	2.086
7	2.365	30	2.042
10	2.228	60	2.000

$$\bar{X} = \mu \pm t_{\nu, \%} S_{\bar{X}}(C\%)$$



The smaller the # of samples, the larger is “t”

How well does the sample mean \bar{X} estimate the population mean μ ?

Because the sample means are normally distributed, the *t-distribution can be used*:

$$\bar{X} = \mu \pm t_{v,\%} S_{\bar{X}} (C\%) \quad S_{\bar{X}} = \frac{S_X}{N^{1/2}}$$

NOTE: Sample means are normally distributed even when the parent population is not Gaussian.

That is, one can say with $C\%$ confidence that the population mean μ is within $\pm t_{v,\%} S_{\bar{X}}$ of \bar{X} .

What do you do with outliers? Find the problem, or if there is no reason found use:

Chauvenet's criterion is recommended: It states that points should be discarded if the probability (calculated from the normal distribution) of obtaining their deviation from the mean is less than $1/2N$.

Ratio of the maximum acceptable deviation to the standard deviation is given as a function of N .



N	$\frac{(X_{\max} - \bar{X})}{S_X}$
5	1.65
7	1.80
10	1.96
15	2.13
20	2.24
50	2.57
100	2.81

Standard error of a fit to a straight line

$$\hat{Y} = mx + C$$

$$S_Y = \left[\frac{1}{N-2} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \right]^{1/2}$$

Y_i is a random variable and can be taken to have a normal distribution for each value of x_i .

$$P_{\hat{Y}} = t_{v,\%} \left\{ S_Y^2 \left[\frac{1}{N} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] \right\}^{1/2}$$

$$S_{xx} = \sum x_i^2 - (1/N) (\sum x_i)^2$$

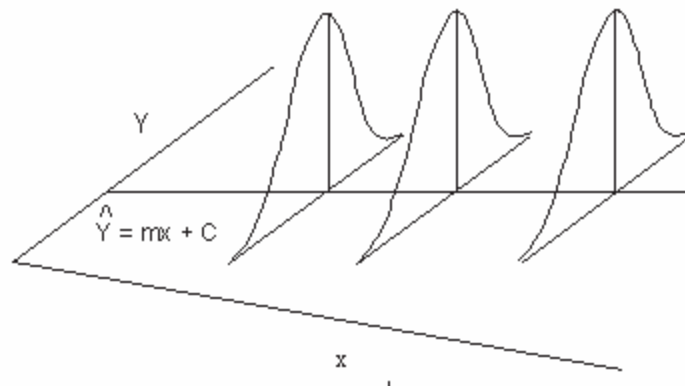
$$v = N - 2$$

For N large and a 95% confidence level, we set

$$t_{v,\%} = 2$$

$$P_{\hat{Y}} = 2 \left\{ S_Y^2 \left[\frac{1}{N} + \frac{(x - \bar{x})^2}{S_{xx}} \right] \right\}^{1/2}$$

($C = 95\%, N > 10$)



Standard error of a fit to a straight line

Standard deviation for the **slope**

$$S_m = \left(\frac{S_Y^2}{S_{xx}} \right)^{1/2}$$

Standard deviation for the **intercept**

$$S_C = \left[S_Y^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}} \right) \right]^{1/2}$$

Precision uncertainty for the **slope**

$$P_m = t_{v,\%} S_m$$

Precision uncertainty for the **intercept**

$$P_C = t_{v,\%} S_C$$

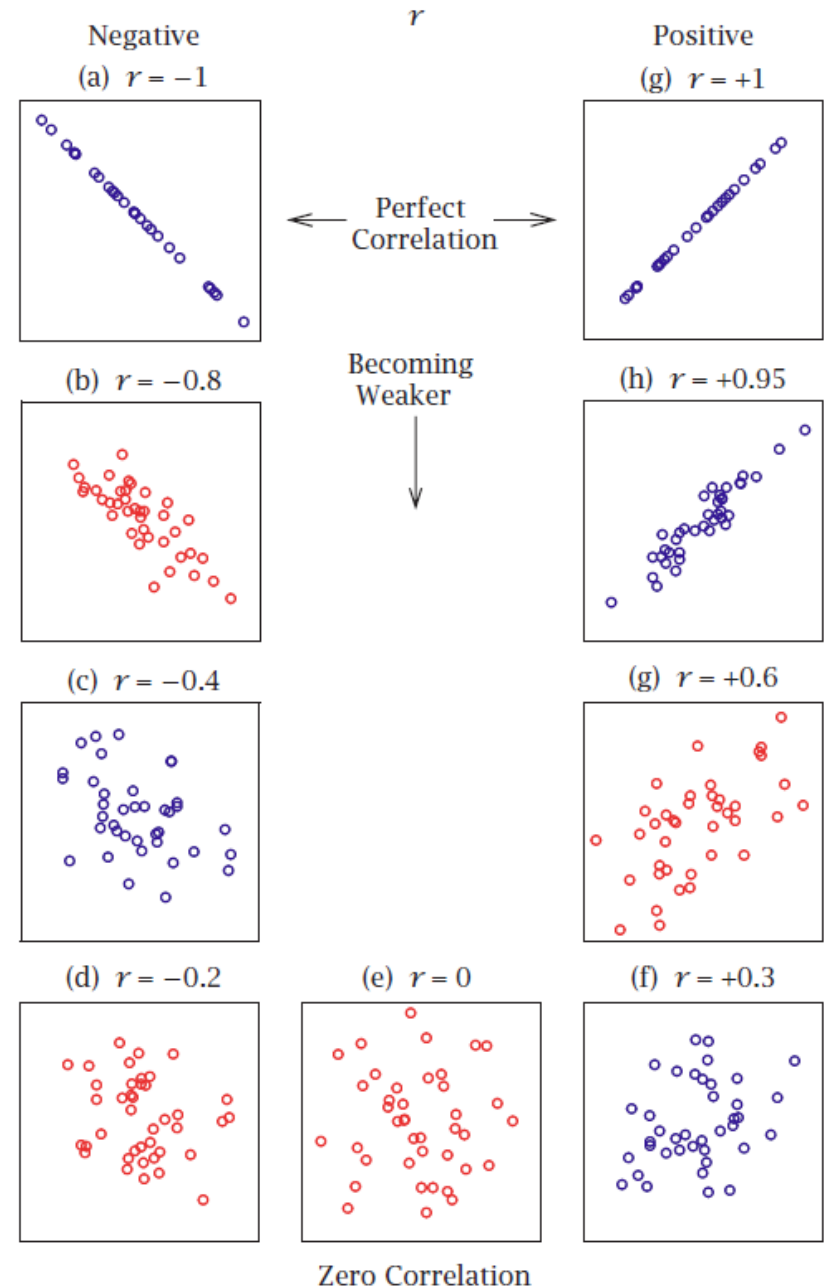
*And for N large and a 95% confidence level,
we set $t_{v,\%} = 2$*

The Correlation Coefficient

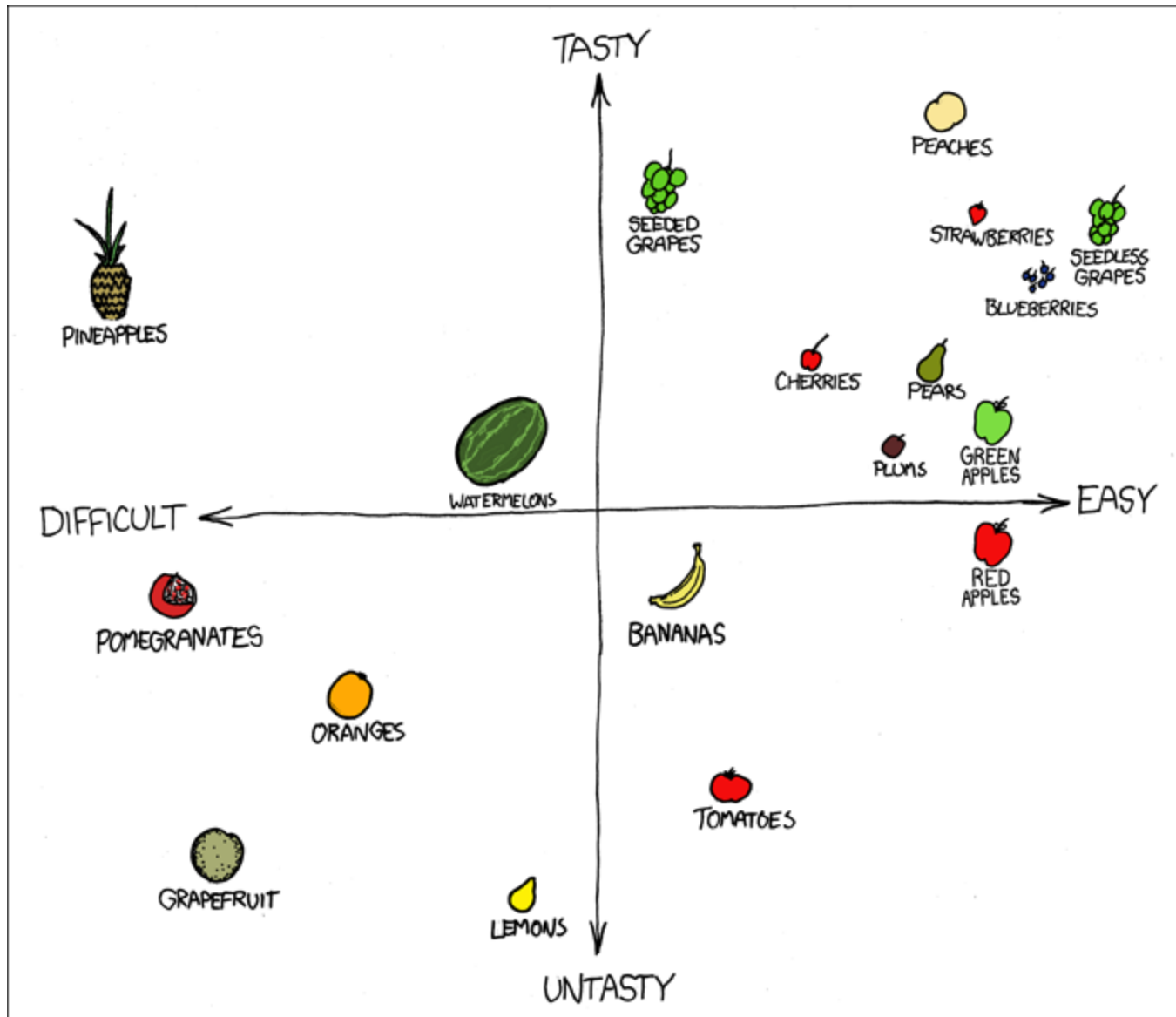
$$r = \frac{1}{N-1} \sum \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

In statistics practice a straight-line curve-fit is considered reliable for $\pm 0.9 \leq r \leq \pm 1$ (the sign indicates that Y increases or decreases with X).

The correlation coefficient is useful when precision errors are large, such as in experiments in the life sciences and medicine. Then the central question is whether there is any correlation whatsoever. In physics and engineering experiments the precision errors are usually much smaller and the precision uncertainties of \hat{Y} , m , and C are more useful.



But be careful! You can correlate anything, even if ill or subjectively defined.



Autocorrelation shows how similar data is over certain distances

correlation between observations separated by k time steps

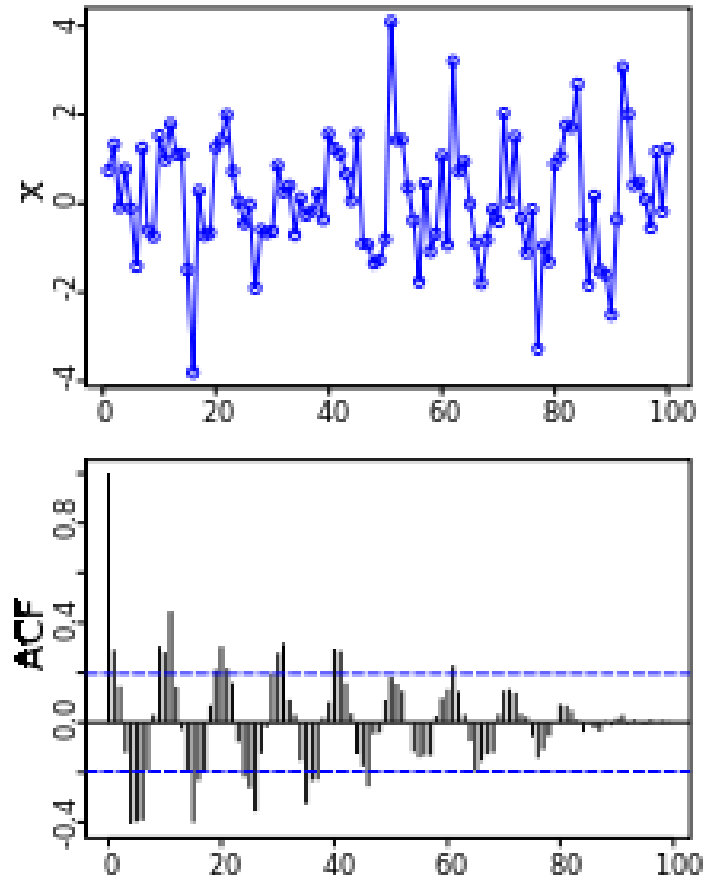
$$r_k = \frac{\sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

Autocovariance

$$c_k = \frac{1}{N} \sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x})$$

$$r_k = c_k / c_0$$

c_0 is the variance



A plot showing 100 random numbers with a "hidden" [sine](#) function, and an autocorrelation ([correlogram](#)) of the series on the bottom.

<http://en.wikipedia.org/wiki/Autocorrelation>

Propagation of Precision Uncertainties

Say Y is a function of N independent measurements X_i . If the uncertainties P_i are small enough we can use a first order Taylor expansion of Y to write

$$Y(X_1 + P_1, X_2 + P_2, \dots, X_N + P_N) \\ \cong Y(X_1, X_2, \dots, X_N) + \frac{\partial Y}{\partial X_1} P_1 + \frac{\partial Y}{\partial X_2} P_2 + \dots + \frac{\partial Y}{\partial X_N} P_N.$$

Since Y is a linear function of the independent variables, a theorem of mathematical statistics says:

$$P_Y = \left[\sum_{i=1}^n \left(\frac{\partial Y}{\partial X_i} P_i \right)^2 \right]^{1/2} \quad \text{or} \quad P_Y = \left[\sum_{i=1}^N \Delta Y_i^2 \right]^{1/2}$$

All the uncertainties in the X_i must be at the same confidence level.

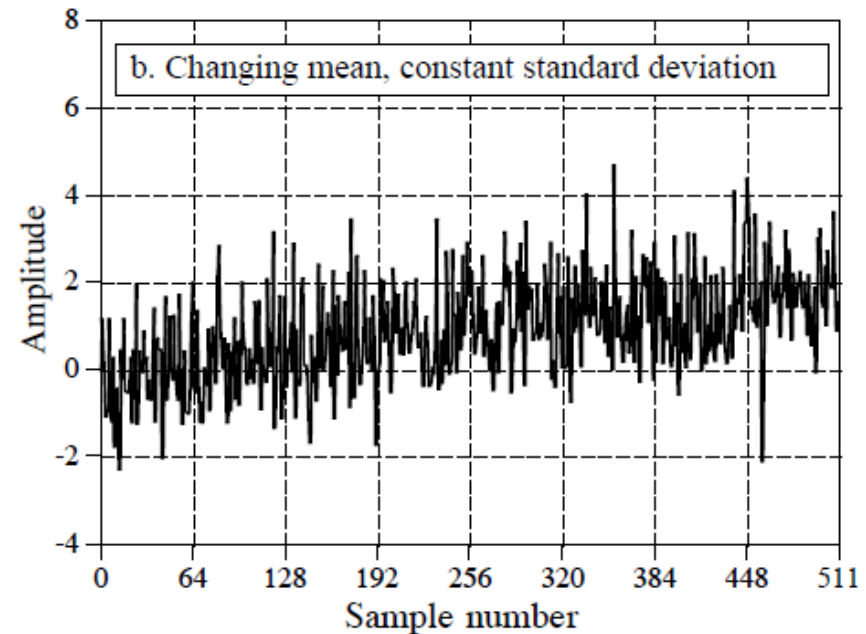
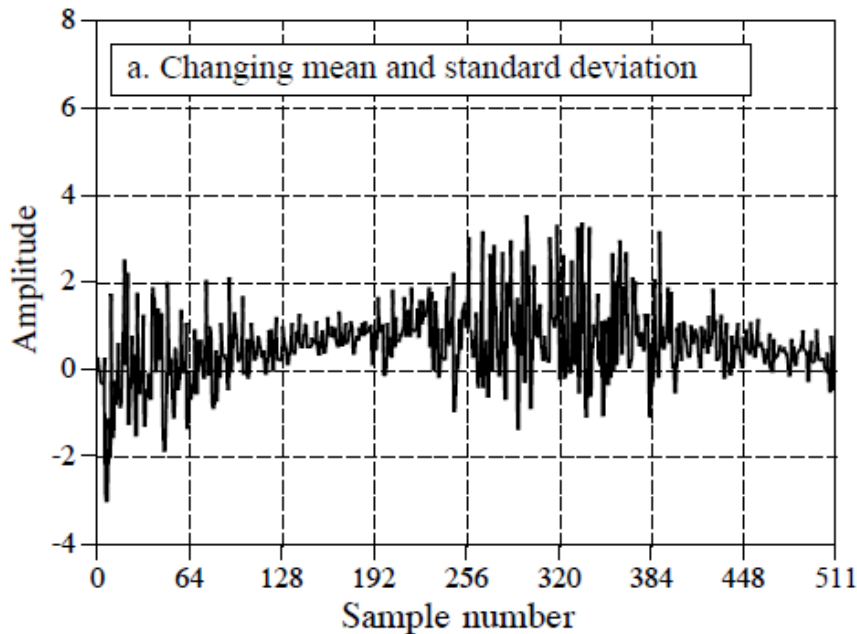
If Y depends only on a product of the independent measurements X_i ,

$$Y = C X_1^{m_1} X_2^{m_2} \dots \quad \text{then} \quad \frac{P_Y}{Y} = \left[\sum_i \left(m_i \frac{P_i}{X_i} \right)^2 \right]^{1/2}$$

What about:

- **weighting,**
- **Precision and accuracy**
- **Histograms**
- **Poisson statistics**
- **Non-linear fitting**
- **Chi-square analysis**

Weighting



Examples of signals generated from non-stationary processes. In (a), both the mean and standard deviation change. In (b), the standard deviation remains a constant value of one, while the mean changes from a value of zero to two. It is a common analysis technique to break these signals into short segments, and calculate the statistics of each segment individually.

Least Square fitting of a straight line $\xrightarrow{\text{minimize}}$
$$S = \sum_i w_i (Y_i - y_i)^2 = \sum_i \frac{1}{\sigma_i^2} (Y_i - y_i)^2$$

If the variance varies, you want to minimize chi-square

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - y_{fit})^2}{\sigma_i^2}$$

Goodness of fit parameter that should be unity for a “fit within error”

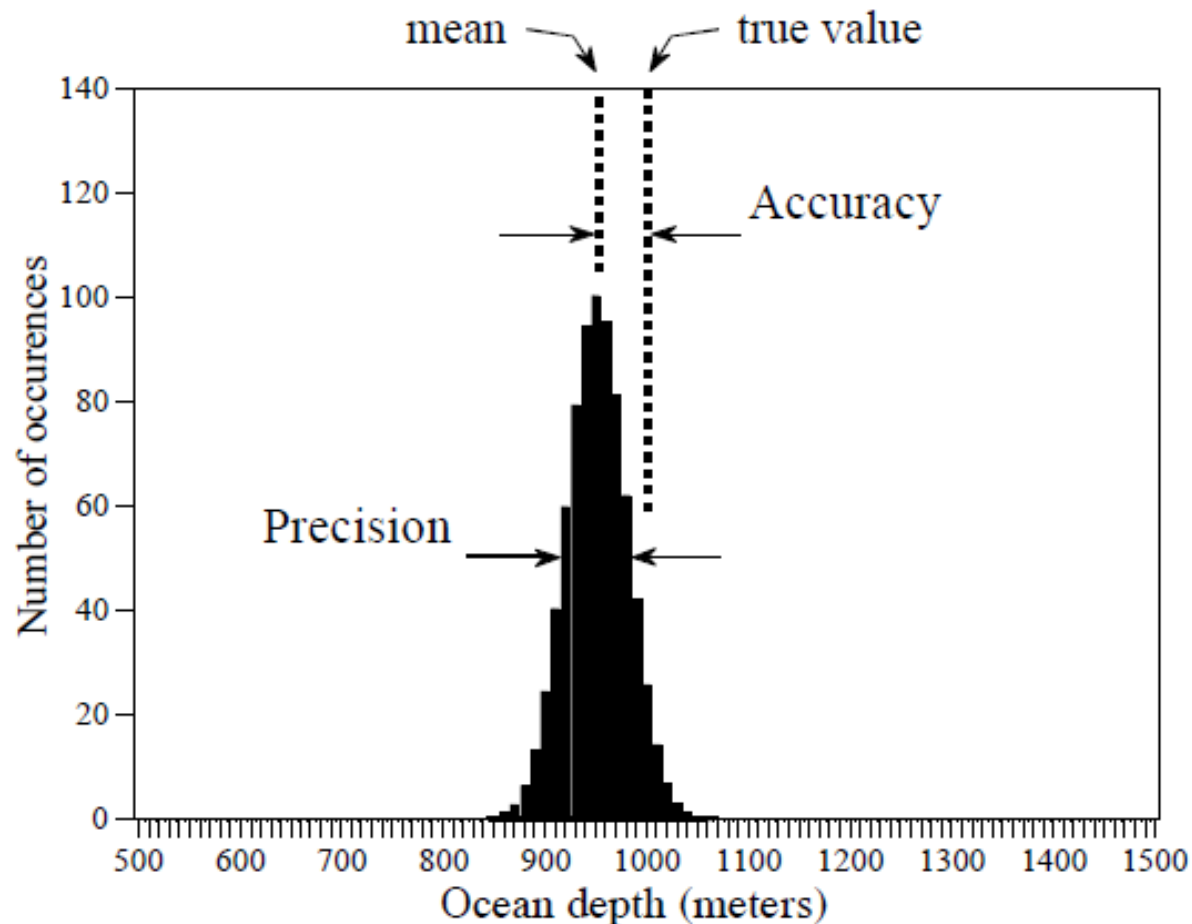
$$\chi_{reduced}^2 = \frac{1}{\nu} \sum_{i=1}^n \frac{(y_i - y_{fit})^2}{\sigma_i^2}$$

ν is the # of degrees of freedom
 $\nu \cong n - \#$ of parameters fitted

χ^2 caveats

- Chi-square lower than unity is meaningless...if you trust your s^2 estimates in the first place.
- Fitting too many parameters will lower c^2 but this may be just doing a better and better job of fitting the noise!
- A fit should go smoothly THROUGH the noise, not follow it!
- There is such a thing as enforcing a “parsimonious” fit by minimizing a quantity a bit more complicated than c^2 . This is done when you have *a-priori* information that the fitted line must be “smooth”.

Graphical description of precision and accuracy



Poor **accuracy** results from **systematic errors**.

Precision is a measure of random noise. Averaging several measurements will always improve the precision.

Poisson distribution:

λ = mean value,

k = # of times observed

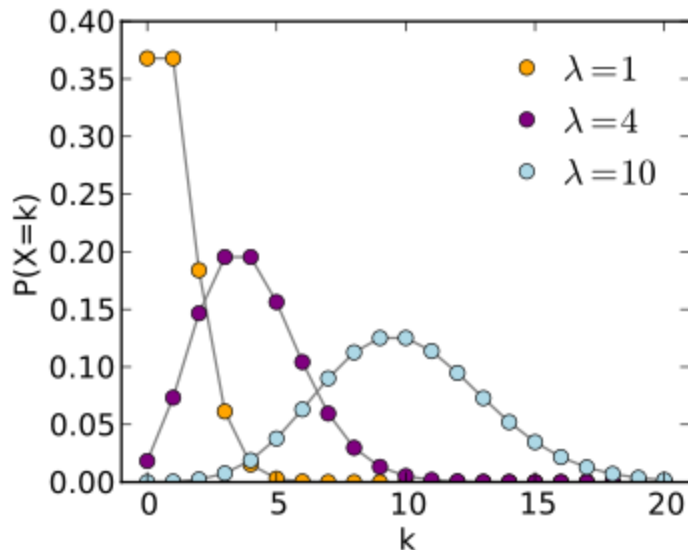
$$\longrightarrow f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$



Probability of observing k occurrences in time t , where λ is the average rate per time

$$\longrightarrow \Pr(N_t = k) = f(k; \lambda t) = \frac{e^{-\lambda t} (\lambda t)^k}{k!}$$

The variance is equal to the mean



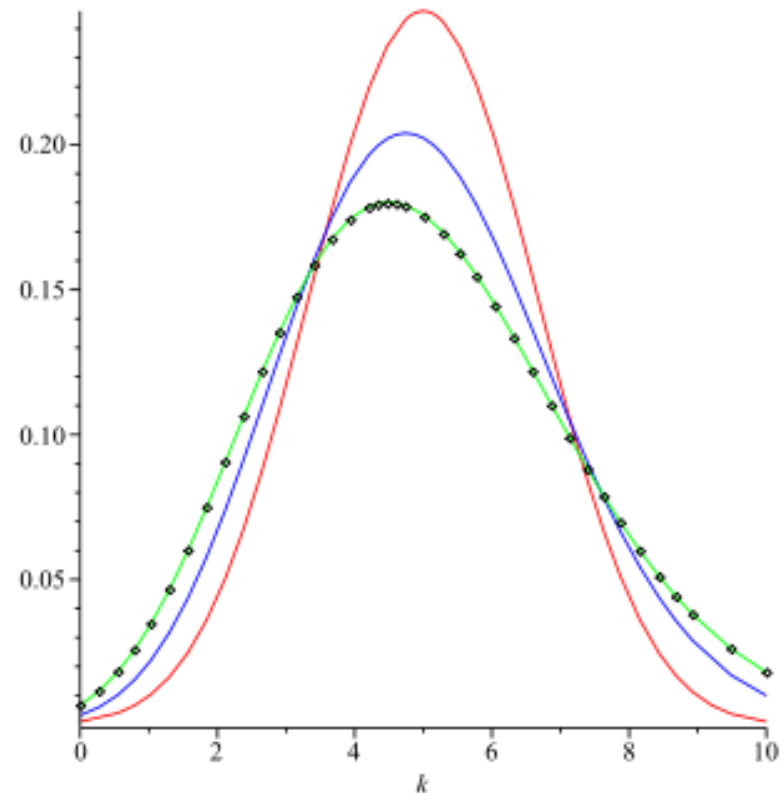
The classic Poisson example is the data set of von Bortkiewicz (1898), for the chance of a Prussian cavalryman being killed by the kick of a horse.

See:

<http://www.umass.edu/wsp/statistics/lessons/poisson/index.html>

<http://mathworld.wolfram.com/PoissonDistribution.html>

Comparison of the Poisson distribution (black dots) and the [binomial distribution](http://en.wikipedia.org/wiki/Binomial_distribution) with $n=10$ (red line), $n=20$ (blue line), $n=1000$ (green line). All distributions have a mean of 5. The horizontal axis shows the number of events k .



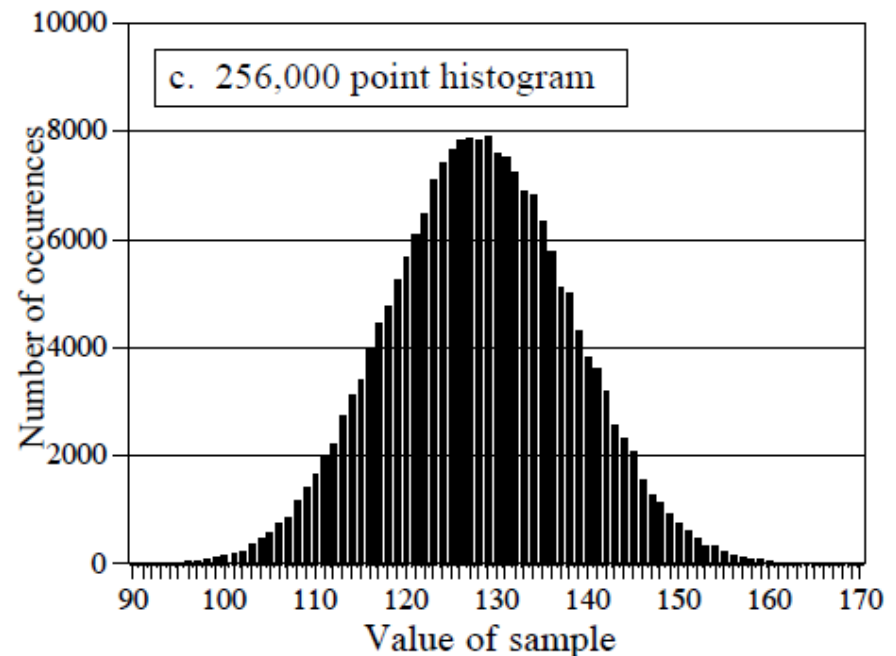
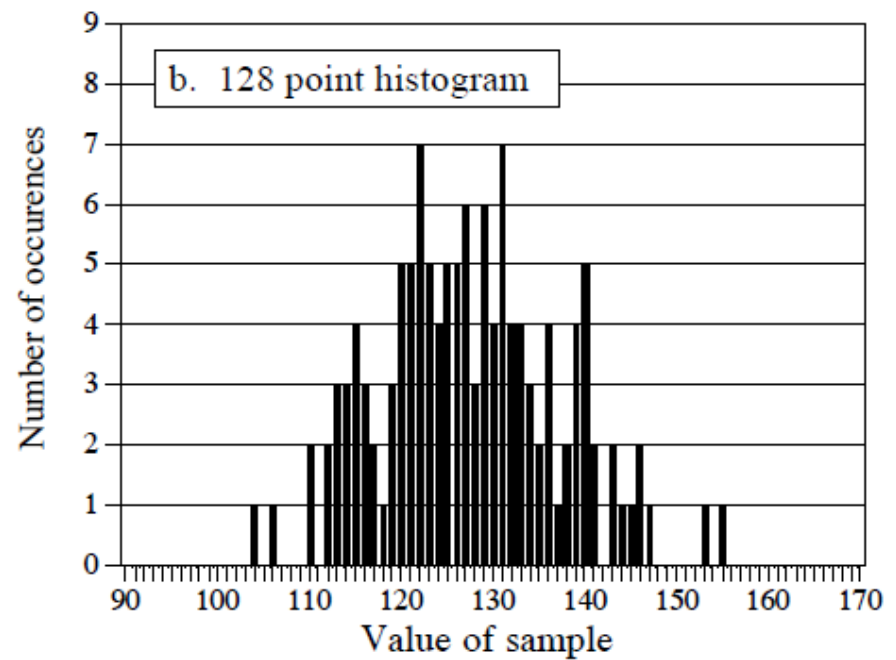
http://en.wikipedia.org/wiki/Poisson_distribution

Histograms

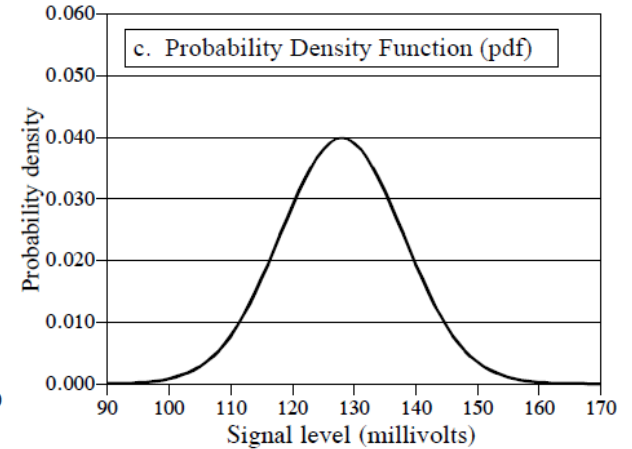
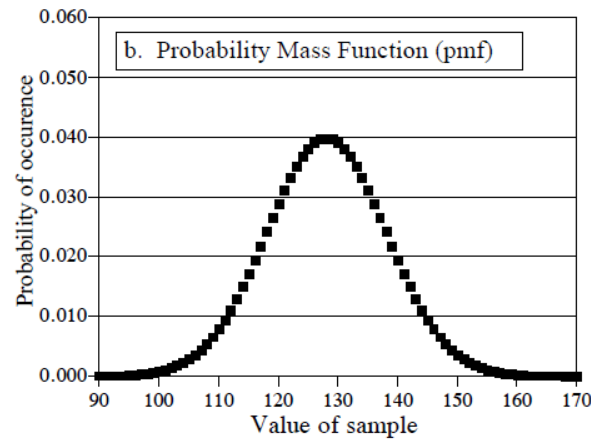
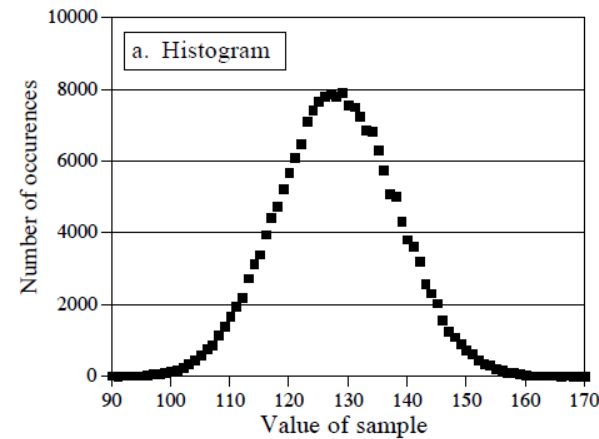
$$N = \sum_{i=0}^{M-1} H_i$$

$$\mu = \frac{1}{N} \sum_{i=0}^{M-1} i H_i$$

$$\sigma^2 = \frac{1}{N-1} \sum_{i=0}^{M-1} (i - \mu)^2 H_i$$



(a) the histogram, (b) the probability mass function (pmf) and (c) the probability density function (pdf)



The amplitude of these three curves is determined by: (a) the sum of the values in the histogram being equal to the number of samples in the signal; (b) the sum of the values in the pmf being equal to one, and (c) the area under the pdf curve being equal to one.

Examples of probability density functions.

