Statistical description of data

Moments

Data consisting of several sets of values is assumed to cluster around some particular value $\Rightarrow$ central tendency.

Measures of central tendency – moments

k-th moments about origin is defined as

$$E(x^k) = \frac{1}{N}\sum_{i=1}^{N} w_i X_i^k \quad \Rightarrow \quad \begin{cases} E(x^1) &= \frac{1}{N}\sum_{i=1}^{N} w_i X_i \equiv \mu = \text{mean} \\ E(x^2) &= \frac{1}{N}\sum_{i=1}^{N} w_i X_i^2 \\ \cdots \end{cases}$$

w_i are the weight function. For real-valued function $f(x)$, mean for discrete and continuous distribution of x

$$\langle f(x) \rangle = \sum_x f(x)\, w_x \quad \text{and} \quad \int_x f(x)\, w(x)\, dx$$

Moments can also be defined with respect to non-zero origin, most popularly about mean μ, called central moment,

$$E\left((x-\mu)^k\right) = \frac{1}{N}\sum_{i=1}^{N} w_i\,(X_i-\mu)^k \Rightarrow \begin{cases} E\left((x-\mu)^1\right) = \frac{1}{N}\sum_i w_i\,(X_i-\mu) = 0 \\ E\left((x-\mu)^2\right) = \frac{1}{N}\sum_i w_i\,(X_i-\mu)^2 = \sigma^2 \\ \qquad\qquad = \text{variance etc.} \end{cases}$$

Next two higher central moments are skewness and kurtosis.

Generally higher moments are statistically less robust, $k \geq 2$ moments may or may not exist, may not converge with increasing N nor show any central tendency.

For example variance or standard deviation may not decrease with increasing data points.

Another important point Weight function (or frequency) w_i corresponds to probability distribution from which data is drawn. The k-th central moment of real-valued continuous function $f(x)$ is

$$E\left((x-\alpha)^k\right) = \int (x-\alpha)^k f(x)\, dx$$

N.B. mean is not the only first moment estimator of the data and not necessarily the best. Other available estimators are median and mode. However, they are not widely used in physics and so will not be discussed.

Unbiased estimator

Consider a population of N discrete random variables $x_i, \quad i = 1, 2, \ldots, N$.

Variables are independent *i.e.* $\Pr(A \cap B) = \Pr(A)\Pr(B)$, but they being uncorrelated *i.e.* $\mathrm{Cov}(A, B) = 0$ is also sufficient.

Population average and variance are obtained from

$$\mu \equiv \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

In practice, true mean and variance of population is not known a priori, possibly because $N \to$ large.

Estimate μ, σ^2 of population from finite sample(s) : $\{y_j\}$ for $j = 1, 2, \ldots, n$ where $n < N$. The estimator of mean and variance (two of them) are,

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

$$s_n^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \mu)^2 \quad \text{and} \quad s_{n-1}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2$$

Estimators $\overline{y}$, s^2 depend on data size but true μ, σ^2 don't!

Difference between expected value of the estimator $\langle\theta\rangle$ and the true population value Θ of the parameter being estimated is called the bias

$$\langle\theta\rangle - \Theta$$

An estimator having zero bias is said to be unbiased.

Sample average is an unbiased estimator of population mean μ

$$\langle\overline{y}\rangle = \frac{1}{N}\sum_{i=1}^{N}\overline{y}_i = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}y_{j,i} = \frac{1}{n}\sum_{j=1}^{n}\left[\frac{1}{N}\sum_{i=1}^{N}y_i\right]_j = \frac{1}{n}\sum_{j=1}^{n}\mu = \mu$$

It is tedious but straight forward to show that s^2_{n-1} is an unbiased estimator of σ^2 but s^2_n is not

$$\langle s^2_{n-1}\rangle = \sigma^2, \quad \text{but} \quad \langle s^2_n\rangle = \frac{n-1}{n}\sigma^2$$

Do not bother too much about use of $n-1$ and n for defining two variances. It is sufficient to say that use s^2_n when μ is known apriori, but use s^2_{n-1} if mean $\bar{y}$ is estimated from data. However, none of these matters when n is large.

When we talk about error we mean standard error *i.e.* square of uncertainty in the sample average $\overline{y}$ from the population average μ

$$\sigma_{\bar{y}} \equiv \bar{y} - \mu$$

$$\langle \sigma_{\overline{y}}^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} (\overline{y}_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n^2} \sum_{k=1}^{n} (y_k - \mu)_i^2 = \frac{1}{Nn^2} \sum_{i=1}^{N} \sum_{k=1}^{n} (y_k - \mu)_i^2$$

$$= \frac{1}{n^2} \sum_{k=1}^{n} \left[\frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2 \right]_k = \frac{1}{n^2} \sum_{k=1}^{n} \sigma_k^2 = \frac{\sigma^2}{n}$$

$$\therefore \quad \langle \sigma_{\overline{y}}^2 \rangle = \langle \overline{y}^2 \rangle - \mu^2 = \frac{\sigma^2}{n}$$

Hence, the standard error is

$$\text{error} \equiv \sqrt{\langle \sigma_{\overline{y}}^2 \rangle} = \frac{\sigma}{\sqrt{n}}$$

This is precisely the statement of Central Limit Theorem:

If random samples of n observations $y_1, y_2, \ldots, y_n$ are drawn from a population of finite mean μ and variance σ^2, then when n is sufficiently large, sampling distribution of the sample mean can be approximated by a normal density with mean μ and standard deviation $= \sigma/\sqrt{n}$.

When a physical result is quoted, it is generally given in the form

$$y = \overline{y} \pm \Delta y \quad \text{where} \quad \Delta y = \sqrt{\frac{s^2_{n-1}}{n}} = \sqrt{\frac{1}{n(n-1)} \sum_{j=1}^{n} (y_j - \overline{y})^2}$$

From Central Limit Theorem it follows that the distribution of the sample mean is approximately normal, and so the probabilistic interpretation of the result is,

$$P\left[\overline{y} - \Delta y \leq y \leq \overline{y} + \Delta y\right] \simeq 68.3\%$$

$$P\left[\overline{y} - 2\Delta y \leq y \leq \overline{y} + 2\Delta y\right] \simeq 95.4\%$$

$$P\left[\overline{y} - 3\Delta y \leq y \leq \overline{y} + 3\Delta y\right] \simeq 99.7\%$$

Even if the form of underlying distribution of y is unknown, Central Limit Theorem enables us to make an approximate quantitative statement about probability of y lying within a specified range.

Data and Functions

Consider determining Young's modulus from a somewhat unusual load – depression data consisting of only one load m.

load	depression					average	Young's modulus
m	d_1	d_2	d_3	$\cdots$	d_n	$\overline{d} \pm \Delta d$	$Y(m, \overline{d}) \pm \Delta Y$
	Y_1	Y_2	Y_3	$\cdots$	Y_n		$\overline{Y} \pm \Delta Y$

For this data, usual approach by fitting load-depression straight line from data points (m_i vs $\overline{d}_i \pm \Delta d_i$) will not work, since for that we need multiple loads m_i and their corresponding average depressions and errorbars $\overline{d}_i \pm \Delta d_i$. The fit yields $Y \pm \Delta Y$.

First approach could be to calculate $\overline{d}$ and then $Y(m, \overline{d})$ and its error ΔY from Δd by using binomial theorem as in the first line of the table above.

Secondly, Young's modulus measured in 2nd line of the above table.

$$\overline{Y} = \sum_{i=1}^{n} Y(m, d_i) \quad \text{and} \quad \Delta Y = \sqrt{\frac{1}{n} \sum_i \left(Y_i - \overline{Y} \right)^2}$$

Which of one these is a better *i.e.* unbiased estimator of Young's modulus both in terms of average and variance.

To estimate some function of population average $f(\mu)$, we just saw, two estimators can be defined

$$\overline{f} = \frac{1}{n}\sum_{j=1}^{n} f_j(y_j) \quad \text{and} \quad f(\overline{y}) = f\left(\frac{1}{n}\sum_{j=1}^{n} y_j\right)$$

Assuming the sampled points $\{y_j\}$ are clustered close to μ, Taylor expansion of the function $f(y_j)$ about μ can be performed

$$f_j(y_j) = f(\mu) + (y_j - \mu)f'(\mu) + \frac{1}{2!}(y_j - \mu)^2 f''(\mu) + \cdots$$

$$\begin{aligned}
\overline{f} - f(\mu) &= \frac{1}{n}\sum_{j=1}^{n} f_j(y_j) - f(\mu) \\
&= f'(\mu)\left[\frac{1}{n}\sum_{j=1}^{n}(y_j - \mu)\right] + \frac{1}{2}f''(\mu)\left[\frac{1}{n}\sum_{j=1}^{n}(y_j - \mu)^2\right] + \cdots \\
&= f'(\mu)(\overline{y} - \mu) + \frac{1}{2}f''(\mu)\left[\frac{1}{n}\sum_{j=1}^{n}(y_j - \mu)^2\right] + \cdots
\end{aligned}$$

$$\langle \overline{f} - f(\mu) \rangle = f'(\mu)\frac{1}{N}\sum_{i=1}^{N}(\overline{y}-\mu)_i + \frac{1}{2}f''(\mu)\frac{1}{Nn}\sum_{j=1}^{n}\sum_{i=1}^{N}(y_j-\mu)_i^2$$

$$= f'(\mu)(\mu-\mu) + \frac{1}{2}f''(\mu)\frac{1}{n}\sum_{j=1}^{n}\sigma_j^2 + \cdots$$

$$= 0 + \frac{1}{2}f''(\mu)\sigma^2 + \cdots \neq 0$$

Since $\langle \overline{f} \rangle \neq f(\mu)$, the $\overline{f}$ is not an unbiased estimator unless $f''(\mu)$ and higher order derivatives vanish. Hence, Young's modulus estimator in the 2nd line is biased!

For $f(\overline{y})$, perform Taylor expansion as before

$$f(\overline{y}) - f(\mu) = f(\mu) + f'(\mu)(\overline{y}-\mu) + \frac{1}{2!}f''(\mu)(\overline{y}-\mu)^2 + \cdots - f(\mu)$$

$$= f'(\mu)(\overline{y}-\mu) + \frac{1}{2}f''(\mu)\frac{1}{n^2}\sum_{j=1}^{n}(y_j-\mu)^2 + \cdots$$

$$\langle f(\overline{y}) - f(\mu) \rangle = 0 + \frac{1}{2n}f''(\mu)\sigma^2 + \cdots$$

This implies estimator $f(\overline{y})$ (in the 1st line) is of order $1/n$ less biased than that of $\overline{f}$.

Variance of functions

If $Y(m, \overline{d})$ is less biased then what about its variance σ_Y^2? It is certainly not $(\Delta Y)^2$ since use of binomial theorem cannot lead to analysis of bias.

However, the quantity $(\Delta Y)^2 = \overline{f^2} - \overline{f}^2$ is the correct variance of $\overline{f}$ but obviously not of $f(\overline{y})$.

Two ways to do it (see young.physics.ucsc.edu/jackboot.pdf) :

- Bootstrap (relatively smaller sample size)
- Jackknife (relatively larger sample size)

Bootstrap

Resampling of data points (of unknown distribution) by choosing data points randomly with replacement from sampled / experimental data.

load	depressions					bootstrap samples
m	d_1	d_2	d_3	$\cdots$	d_n	
			sample 1			$\{d_1, d_4, d_{10}, d_3\}$
			sample 2			$\{d_2, d_7, d_9, d_4\}$
			$\cdots$			$\{d_8, d_6, d_6, d_2\}$
			sample 25			$\{d_{10}, d_8, d_4, d_6\}$

In generic notation, each bootstrap sample of size ν is drawn from a data set of size n generating a total of $\mathcal{B}$ bootstrap sample set,

Bootstrap samples		average	variance
$\left\{x_1^{(1)}, x_2^{(1)}, \ldots, x_\nu^{(1)}\right\}$	$\rightarrow$	$\overline{x}^{(1)}$	$s_{\nu-1}^{(1)\,2}$
$\left\{x_1^{(2)}, x_2^{(2)}, \ldots, x_\nu^{(2)}\right\}$	$\rightarrow$	$\overline{x}^{(2)}$	$s_{\nu-1}^{(2)\,2}$
$\ldots$		$\ldots$	$\ldots$
$\left\{x_1^{(\mathcal{B})}, x_2^{(\mathcal{B})}, \ldots, x_\nu^{(\mathcal{B})}\right\}$	$\rightarrow$	$\overline{x}^{(\mathcal{B})}$	$s_{\nu-1}^{(\mathcal{B})\,2}$

Since data points are chosen at random with replacement, a particular data point x_i can appear multiple number of times in a bootstrap sample or never at all in any of the sample set.

Probability of x_i to be chosen is $p(x_i) = 1/n \equiv p$ for each drawing.

Suppose x_i appears in a bootstrap sample n_i times, then

$\sum_{i=1}^{n} n_i = n \quad \Rightarrow \quad$ probability distribution is binomial,

$$P(n_i) = \binom{n}{n_i} p^{n_i} (1-p)^{n-n_i}$$

$$\text{mean} \quad \overline{n}_i = \lambda = np = 1$$

$$\text{variance} \quad \sigma^2_{n_i} = \overline{n_i^2} - \overline{n}_i^2 = np(1-p) = 1 - \frac{1}{n}$$

$$\Rightarrow \overline{n_i^2} = 2 - \frac{1}{n}$$

$$\text{cov}(n_i, n_j) = \overline{n_i n_j} - \overline{n}_i \overline{n}_j = -\frac{1}{n} \quad \text{for} \quad i \neq j$$

$$\Rightarrow \overline{n_i n_j} = 1 - \frac{1}{n}$$

When $n \to \infty$, keeping $\lambda = np$ fixed, binomial distribution goes over to Poisson, with both mean and variance $\lambda = 1$.

Averages of bootstrap samples

$$\overline{x}^{(\alpha)} = \frac{1}{n} \sum_{j=1}^{n} x_i^{(\alpha)} = \frac{1}{n} \sum_{j=1}^{n} n_i^{(\alpha)} x_i$$

Bootstrap average is an unbiased estimator of population mean μ, while standard error for bootstrap samples is a biased estimator of σ^2

$$\overline{x}^{\mathcal{B}} = \frac{1}{\mathcal{B}} \sum_{\alpha=1}^{\mathcal{B}} \overline{x}^{(\alpha)} = \frac{1}{n} \frac{1}{\mathcal{B}} \sum_{\alpha=1}^{\mathcal{B}} \sum_{i=1}^{n} n_i^{(\alpha)} x_i = \overline{x} \quad \Rightarrow \quad \langle \overline{x}^{\mathcal{B}} \rangle = \mu$$

$$\sigma^2_{\mathcal{B}_{\bar{x}}} = \overline{(x^{\mathcal{B}})^2} - \left(\overline{x}^{\mathcal{B}}\right)^2 = \frac{1}{\mathcal{B}} \sum_{\alpha=1}^{\mathcal{B}} \left(\overline{x}^{(\alpha)} - \overline{x}^{\mathcal{B}}\right)^2 \quad \Rightarrow \quad \left\langle \sigma^2_{\mathcal{B}_{\bar{x}}} \right\rangle = \frac{n-1}{n} \sigma^2$$

Define average function in bootstrap scheme as

$$f_\alpha^{\mathcal{B}} \equiv f\left(\overline{x}^{(\alpha)}\right)$$

The average, standard error and bias of the bootstrap function are

$$\overline{f}^{\mathcal{B}} = \frac{1}{\mathcal{B}} \sum_{\alpha=1}^{\mathcal{B}} f_\alpha^{\mathcal{B}}$$

$$\sigma^2_{f^{\mathcal{B}}} = \frac{n-1}{n} \sum_{j=1}^{n} \left(f^{\mathcal{B}} - \overline{f}^{\mathcal{B}}\right)^2 = (n-1)\left(\overline{f^{\mathcal{B}\,2}} - \overline{f}^{\mathcal{B}\,2}\right)$$

$$\langle \overline{f}^{\mathcal{B}} - f(\mu) \rangle = \frac{1}{2(n-1)} f''(\mu) \sigma_y^2 + \cdots$$

Jackknife

Statistics are created by systematically dropping out subsets of data one at a time and assessing the resulting variations in the studied parameter.

From a sample of n values, jackknife begins by throwing away the first value or first subset of r values resulting a jackknife sample set of $n-1$ or $n-r$ data. Subsequently it drops the second, then third and so on.

load	depressions	jackknife samples	deleted average	variance
m	$\{d_1, d_2, d_3, \cdots, d_n\}$			
	sample 1	$\{d_2, d_3, d_4, \cdots, d_n\}$	$\overline{y}_1$	$s^{(1)\,2}_{n-2}$
	sample 2	$\{d_1, d_3, d_4, \cdots, d_n\}$	$\overline{y}_2$	$s^{(2)\,2}_{n-2}$
	sample k	$\{d_1, d_2, d_3, \cdots, d_n\}$	$\overline{y}_k$	$s^{(k)\,2}_{n-2}$
	sample n	$\{d_1, d_2, d_3, \cdots, d_{n-1}\}$	$\overline{y}_n$	$s^{(n)\,2}_{n-2}$

Jackknife sample average or deleted average for $r = 1$ is

$$\overline{y}_k = \frac{n\overline{y} - y_k}{n-1} = \frac{1}{n-1}\sum_{j\neq k} y_j$$

This process results in a set of parameter values $\{\overline{y}_k,\ k = 1, 2, \ldots, n\}$.

Jackknife average is then defined by

$$\overline{y}_{\mathrm{JK}} = \frac{1}{n}\sum_{k=1}^{n}\overline{y}_k$$

Jackknifing does not change the data average,

$$\overline{y}_{\mathrm{JK}} = \frac{1}{n}\sum_{k=1}^{n}\overline{y}_k = \frac{1}{n}\sum_{k=1}^{n}\frac{n\overline{y}-y_k}{n-1} = \frac{n\overline{y}-\sum_{k=1}y_k/n}{n-1} = \frac{n\overline{y}-\overline{y}}{n-1} = \overline{y}$$

Jackknife estimate of standard error σ^2_{JK} is

$$\sum_{k=1}^{n}(\overline{y}_k-\overline{y}_{\mathrm{JK}})^2 = \sum_{k=1}^{n}\left(\frac{n\overline{y}-y_k}{n-1}-\overline{y}\right)^2 = \frac{s^2_{n-2}}{n-1}$$

$$\sigma^2_{\mathrm{JK}} = \frac{n-1}{n}\sum_{k=1}^{n}(\overline{y}_k-\overline{y}_{\mathrm{JK}})^2$$

Define f^{JK} as a function of jackknife variables $\overline{y}_k$,

$$\overline{f}^{\mathrm{JK}} = \frac{1}{n}\sum_{k=1}^{n}f_k^{\mathrm{JK}}$$

$$\sigma^2_{f^{\mathrm{JK}}} = \frac{n-1}{n}\sum_{k=1}^{n}\left(f^{\mathrm{JK}}-\overline{f}^{\mathrm{JK}}\right)^2$$

Confidence level

Goal of statistical tests is to make statement how well observed data is in agreement with predicted / expected / previous result : Hypothesis.

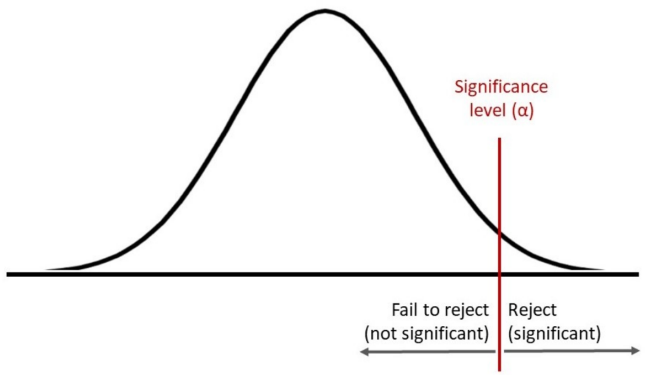
Null hypothesis H_0 represents default assumption that no significant difference or, contrastingly, relationship exists.

In statistical term – either reject H_0 or fail to reject H_0

Significance level α is the probability cut-off or threshold of rejecting H_0 when it is true! If t is test statistics with p.d.f. $g(t|H_0)$ then

$$\alpha = \int_{t_{\text{cut}}}^{\infty} g(t|H_0)\, dt$$

This is expressed in terms of fraction or percentage. If $\alpha = 0.05$, there is a 5% probability of rejecting H_0 even when it is true!



Condfidence level $\gamma = 1 - \alpha$ (significance level)

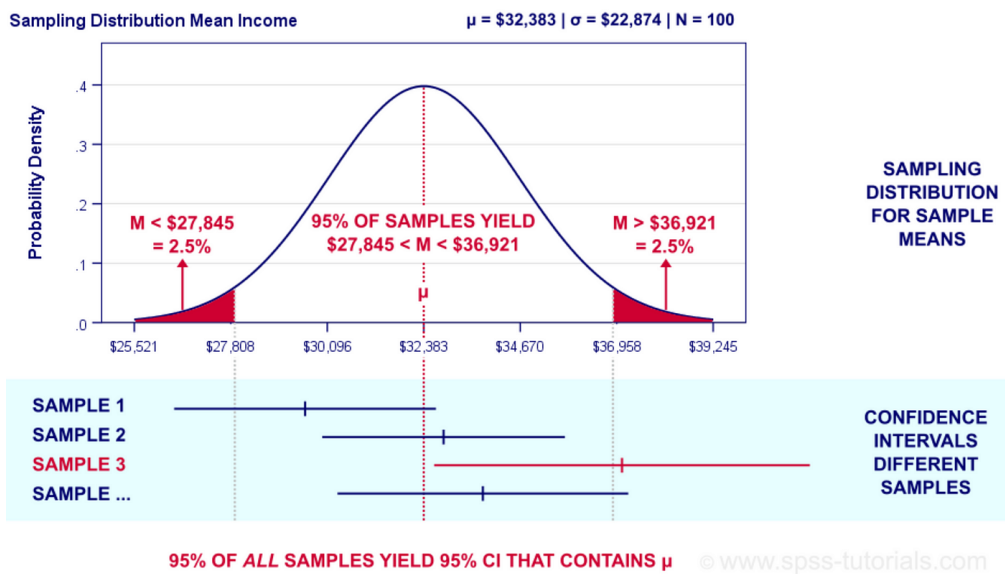
Typical Confidence levels considered in statistical studies are 90%, 95% and 99% meaning H_0 would be accepted that many percentage of time if test is performed ad infinitum.

Confidence interval is a related concept that helps define H_0. It is an interval where the observation or statistics will land γ% of time repeat test.

$$\text{Prob}\left(a(\hat{t}) < t < b(\hat{t})\right) = \gamma$$

where $\hat{t}$ are the measured values and $a(\hat{t})$, $b(\hat{t})$ are evaluated using values from the tests. Confidence interval can be constructed assuming normal distribution of observed mean (ref. central limit theorem),

$$[a,\ b] \equiv \overline{X} \pm \gamma \frac{s_{n-1}}{\sqrt{n}}$$



Comparing means

Often the need is to compare the mean and its standard error of a single observable across experiments or upgrades with different statistics. Say, for instance Higgs boson mass before and after upgrading LHC with higher luminosity and larger statistics.

σ^2 will be different certainly, but are means statistically same or different?

Difference of means can be small compared to s^2_{n-1} and yet significant if statistics is large and vice-versa. The significance of difference in means uses standard error and is assessed by Student's t-test.

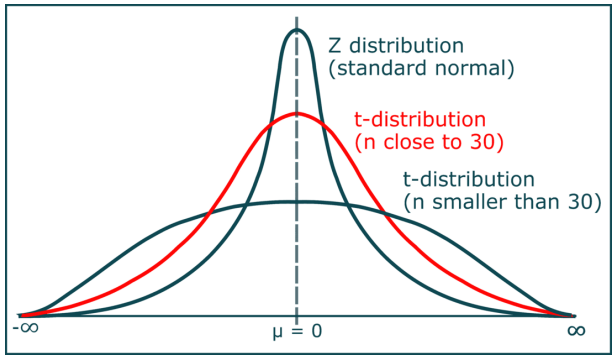
If data $\{y_i\}$ of size n is obtained from normally distributed population $\mathcal{N}(\mu, \sigma^2)$ then the distribution of variables are,

$$t = \frac{\overline{y}_1 - \overline{y}_2}{\sigma/\sqrt{n}} \rightarrow \text{ normal distribution}$$

$$t = \frac{\overline{y}_1 - \overline{y}_2}{s^2_{n-1}/\sqrt{n}} \rightarrow \text{ Student's distribution with } \nu = n - 1 \text{ degrees of freedom}$$

$\overline{y}_2$ can either be μ from theoretical calculation (prediction / expectation) or previous result with different sample size or different sample of same sample size. Typical sample size here is ≤ 30.

Student's distribution $A(t|\nu)$ with ν d.o.f is the probability that generalizes normal distribution, symmetric around zero and bell-shaped.



For $\nu \to \infty$ it becomes $\mathcal{N}(0,1)$. This distribution is used when $\nu < 30$.

The H_0 rejection threshhold t_{cut} or t_{crit} (*i.e.* confidence limit) at a given confidence level γ for ν d.o.f. is read-off from t-table.

one-tailed α	0.10	0.05	0.025	0.01	0.005	0.0005
two-tailed α	0.20	0.10	0.05	0.02	0.01	0.001
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725

Example : Let expected value of an observable is $\mu = 4$. Total experimental or numerical observation is $n = 21 \Rightarrow \nu = n - 1 = 20$ and the obtained mean $\overline{x} = 4.52$ and varaince $s^2_{n-1} = 1.2$.

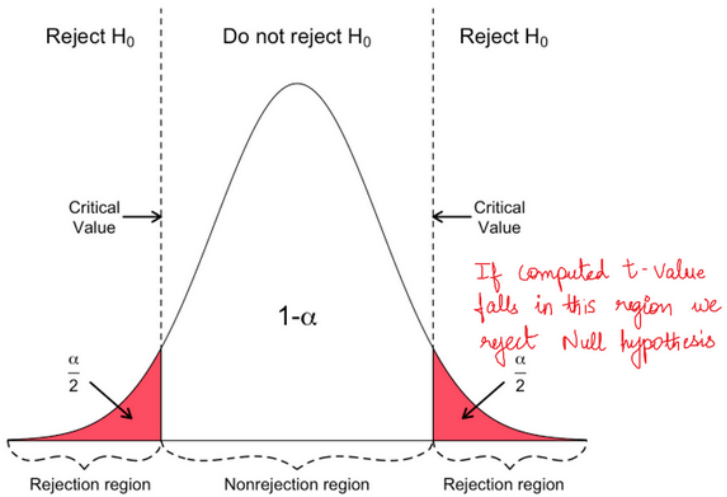
$$H_0 : \text{no significant difference between } \overline{x} \text{ and } \mu$$

Confidence limits are

Confidence level 95% : $t_{\text{crit}} = 2.086$, Confidence interval [3.45, 4.55]

Confidence level 90% : $t_{\text{crit}} = 1.725$, Confidence interval [3.55, 4.45]

Conclusion – at 95% confidence level H_0 cannot be rejected but at 90% it is rejected. It sounds strange, but when read in terms of significance level it says 5% possibility of errorneously rejecting null hypothesis in the first case and 10% in the second.



Comparing variance

Hypothesis for comparing variances, the difference because of different data size or sample set.

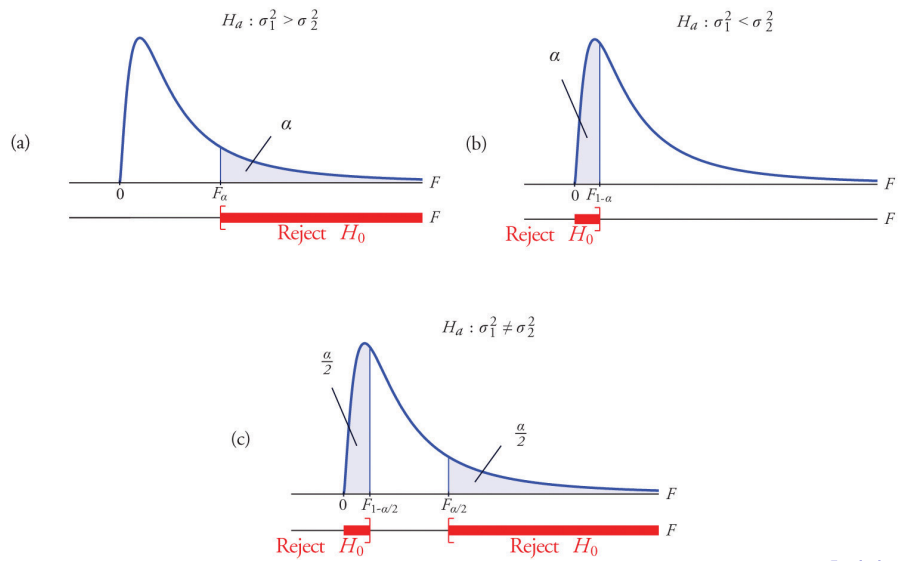
H_0 : Two observed samples have the same variances $s_1^2 = s_2^2$

H' : $s_1^2 > s_2^2,\ \ s_1^2 < s_2^2,\ \ s_1^2 \neq s_2^2$

Test is done by F-test, the F-distribution being $Q(F|\nu_1, \nu_2)$

$$F = s_1^2/s_2^2 \quad \text{where} \quad s_1^2 \geq s_2^2$$

Basic assumptions are data must be drawn from (approximate) normal distribution and samples are independent events. The condition refers to the practice that larger variance always go in the numerator to get a right-tailed test since it is easier to calculate.



Consider same experiment performed at two different labs A and B

Lab A : $n_A = 16 \Rightarrow \nu_A = 15, \quad s_A^2 = 2.09$

Lab B : $n_B = 21 \Rightarrow \nu_B = 20, \quad s_B^2 = 1.10$

The F-statistics is

$$F = s_A^2/s_B^2 = 2.09/1.10 = 1.90$$

$$Q(F|\nu_A, \nu_B) = Q(1.90|15, 20) = 1.845 \text{ at } \alpha = 0.10$$

Hence, H_0 rejection region is $[1.845, \infty]$. Concluding that at 90% confidence level, data rejects H_0 *i.e.* $s_A^1 \neq s_B^2$, in fact $s_A^2 > s_B^2$.

DF1 $\alpha = 0.10$

	1	2	3	4	5	6	7	8	9	10	12	15	20	24
1	39.863	49.5	53.593	55.833	57.24	58.204	58.906	59.439	59.858	60.195	60.705	61.22	61.74	62.002
2	8.5263	9	9.1618	9.2434	9.2926	9.3255	9.3491	9.3668	9.3805	9.3916	9.4081	9.4247	9.4413	9.4496
3	5.5383	5.4624	5.3908	5.3426	5.3092	5.2847	5.2662	5.2517	5.24	5.2304	5.2156	5.2003	5.1845	5.1764
4	4.5448	4.3246	4.1909	4.1073	4.0506	4.0098	3.979	3.9549	3.9357	3.9199	3.8955	3.8704	3.8443	3.831
5	4.0604	3.7797	3.6195	3.5202	3.453	3.4045	3.3679	3.3393	3.3163	3.2974	3.2682	3.238	3.2067	3.1905
6	3.776	3.4633	3.2888	3.1808	3.1075	3.0546	3.0145	2.983	2.9577	2.9369	2.9047	2.8712	2.8363	2.8183
7	3.5894	3.2574	3.0741	2.9605	2.8833	2.8274	2.7849	2.7516	2.7247	2.7025	2.6681	2.6322	2.5947	2.5753
8	3.4579	3.1131	2.9238	2.8064	2.7265	2.6683	2.6241	2.5894	2.5612	2.538	2.502	2.4642	2.4246	2.4041
9	3.3603	3.0065	2.8129	2.6927	2.6106	2.5509	2.5053	2.4694	2.4403	2.4163	2.3789	2.3396	2.2983	2.2768
10	3.285	2.9245	2.7277	2.6053	2.5216	2.4606	2.414	2.3772	2.3473	2.3226	2.2841	2.2435	2.2007	2.1784
11	3.2252	2.8595	2.6602	2.5362	2.4512	2.3891	2.3416	2.304	2.2735	2.2482	2.2087	2.1671	2.1231	2.1
12	3.1766	2.8068	2.6055	2.4801	2.394	2.331	2.2828	2.2446	2.2135	2.1878	2.1474	2.1049	2.0597	2.036
13	3.1362	2.7632	2.5603	2.4337	2.3467	2.283	2.2341	2.1954	2.1638	2.1376	2.0966	2.0532	2.007	1.9827
14	3.1022	2.7265	2.5222	2.3947	2.3069	2.2426	2.1931	2.1539	2.122	2.0954	2.0537	2.0095	1.9625	1.9377
15	3.0732	2.6952	2.4898	2.3614	2.273	2.2081	2.1582	2.1185	2.0862	2.0593	2.0171	1.9722	1.9243	1.899
16	3.0481	2.6682	2.4618	2.3327	2.2438	2.1783	2.128	2.088	2.0553	2.0282	1.9854	1.9399	1.8913	1.8656
17	3.0262	2.6446	2.4374	2.3078	2.2183	2.1524	2.1017	2.0613	2.0284	2.0009	1.9577	1.9117	1.8624	1.8362
18	3.007	2.624	2.416	2.2858	2.1958	2.1296	2.0785	2.0379	2.0047	1.977	1.9333	1.8868	1.8369	1.8104
19	2.9899	2.6056	2.397	2.2663	2.176	2.1094	2.058	2.0171	1.9836	1.9557	1.9117	1.8647	1.8142	1.7873
20	2.9747	2.5893	2.3801	2.2489	2.1582	2.0913	2.0397	1.9985	1.9649	1.9367	1.8924	1.8449	1.7938	1.7667
21	2.961	2.5746	2.3649	2.2333	2.1423	2.0751	2.0233	1.9819	1.948	1.9197	1.875	1.8272	1.7756	1.7481

Goodness of fit

Given two sets of data, questions on sameness of mean and variance can be combined into a single query

H_0 : Two data sets are drawn from same population distribution

Examples of questions that can be asked

- Are data on top quark from LHC and Fermilab comparable?
- Are distribution of brightness of stars in Andromeda and Milky Way galaxy same, both being spiral and of approximately same age?
- Does COVID infection across age follow same distribution?
- Are distribution of marks in NEST exam normally distributed?

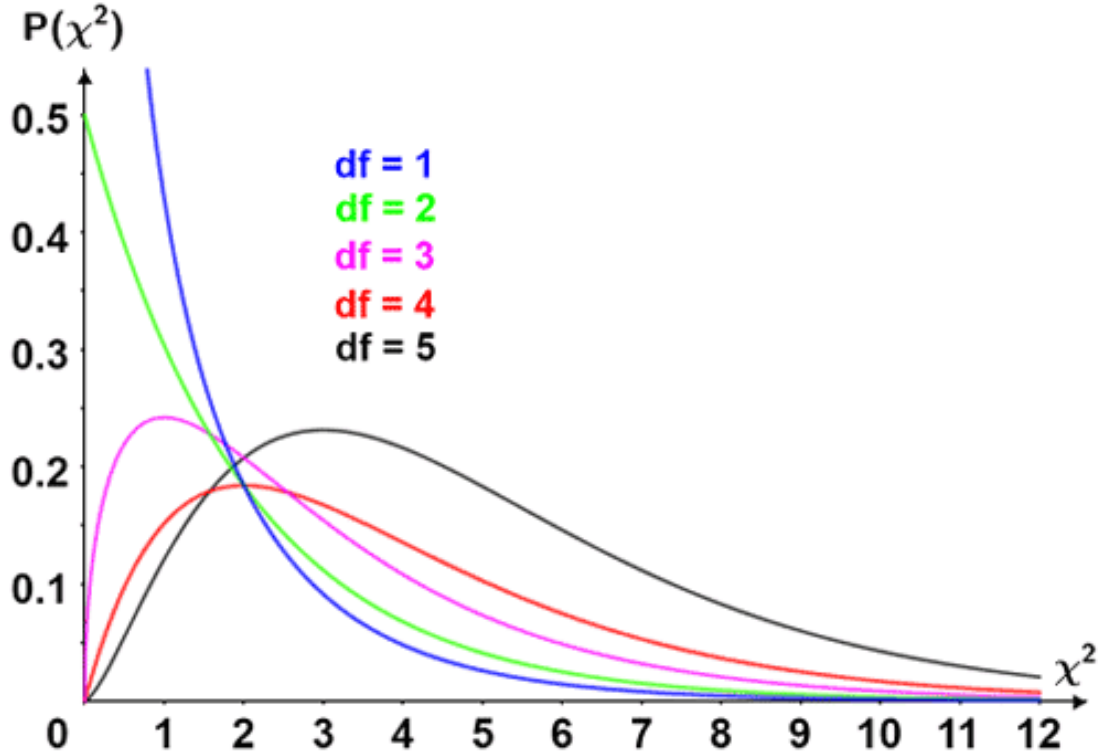
Most of the time, the data (discrete or otherwise) are divided into bins of size k and the test statistics is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is observed frequency or number of events in the i-th bin and E_i is what expected according to some known distribution.

χ^2 probability function $P(\chi^2|\nu)$ determined the outcome of the test. It is defined as probability that the observed χ^2-statistics for accepting a model should be less than χ^2_{crit}.

Instead of $P(\chi^2|\nu)$ a more user friendly distribution is $Q(\chi^2|\nu)$, probability that observed χ^2-statistics will exceed the value by chance even for a correct model. Obviously, $Q(\chi^2|\nu) = 1 - P(\chi^2|\nu)$.



Consider an example with dice, rolled $N = 60$ times (in actual practice can be few hundreds) and number of times each face landed up

Face value	Unbiased distribution $f(x)$	Expected frequency $E = N \times f(x)$	Observed frequency O	$(O-E)^2$	$(O-E)^2/E$
1	1/6	10	9	1	0.10
2	1/6	10	15	25	2.50
3	1/6	10	9	1	0.10
4	1/6	10	8	4	0.40
5	1/6	10	6	16	1.60
6	1/6	10	13	9	0.90
$\nu = 5$		$N = 60$			$\chi^2 = 5.6$

H_0 : The dice is fair / unbiased.

Significance level (α)

Degrees of freedom (*df*)	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666

At 10% level of significance, $\chi^2_{\text{crit}} = 9.236 > 5.6 \Rightarrow$ we do not have enough evidence to say the dice is biased!

χ^2-test can also be used in modelling of data, defining as

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - \overline{y}_M}{\sigma_i} \right)^2 \quad \text{for } \nu = N - M$$

where the average $\overline{y}_M$ is calculated over M observations.

Consider measurement of mass M_Z of the Z^0 boson at CERN made by four different detectors – L3, OPAL, Aleph and Delphi. The weighted average of these four measurements is claimed to be $\overline{M}_Z = 91.177$

Detector	M_Z (GeV/c^2)	$(M_Z - \overline{M}_Z)/\sigma$
L3	91.161 ± 0.013	-1.231
OPAL	91.174 ± 0.011	-0.273
Aleph	91.186 ± 0.013	0.692
Delphi	91.188 ± 0.013	0.846

The χ^2-statistics for $\nu = 4 - 1 = 3$ gives

$$\chi^2_{\nu=3} = \sum_{i=1}^{4} \frac{\left(M_i - \overline{M}_Z \right)^2}{\sigma_i^2} \approx 2.964$$

At 10% level of significance, $\chi^2_{\text{crit}} = 6.25 > 2.964$ implying no good reason to reject H_0, accept $\overline{M}_Z = 91.177$ as global average.

There is one lesson in the above example – if $\chi^2/\nu \approx 1$ then a model can be accepted at 90% level of confidence.

Modelling of data

If we are planning to fit N data points (x_i, y_i, σ_i) to a model with M parameters, then maximum likelihood estimator of model parameters is obtained by minimizing χ^2 with respect to the parameters,

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i; a_1, a_2, \ldots, a_M)}{\sigma_i} \right)^2, \quad \text{dof } \nu = N - M$$

$$\frac{\partial \chi^2}{\partial a_k} = 0 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i)}{\sigma_i^2} \right) \left(\frac{\partial y(x_i; \ldots, a_k, \ldots)}{\partial a_k} \right), \quad k = 1, 2, \ldots, M$$

Linear regression

It is essentially linear fitting *i.e.* fitting data to a straight line.

$$y(x) \equiv y(x; a, b) = a + bx \quad \Rightarrow \quad \chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

Minimization of χ^2 w.r.t a and b yields,

$$\frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2} = 0 \qquad \Rightarrow \qquad S_y = aS + bS_x$$

$$\frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^{N} \frac{x_i \left(y_i - a - bx_i \right)}{\sigma_i^2} = 0 \qquad \Rightarrow \qquad S_{xy} = aS_x + bS_{xx}$$

Above notation is borrowed from Nuermical Recipes, where

$$S = \sum_i \frac{1}{\sigma_i^2} \qquad S_x = \sum_i \frac{x_i}{\sigma_i^2} \qquad S_y = \sum_i \frac{y_i}{\sigma_i^2}$$

$$S_{xx} = \sum_i \frac{x_i^2}{\sigma_i^2} \qquad S_{xy} = \sum_i \frac{x_i y_i}{\sigma_i^2}$$

Solving for parameters a, b, we get

$$a = \frac{S_{xx} S_y - S_x S_{xy}}{\Delta}, \quad b = \frac{S S_{xy} - S_x S_y}{\Delta} \quad \text{where } \Delta = S S_{xx} - (S_x)^2$$

Errors on parameters a, b are estimated using propagation of error,

$$\sigma_p^2 = \sum_i \sigma_i^2 \left(\frac{\partial p}{\partial y_i}\right)^2 \Rightarrow \frac{\partial a}{\partial y_i} = \frac{S_{xx} - S_x x_i}{\Delta \sigma_i^2}, \quad \frac{\partial b}{\partial y_i} = \frac{S x_i - S_x}{\Delta \sigma_i^2}$$

$$\Rightarrow \sigma_a^2 = S_{xx}/\Delta, \quad \sigma_b^2 = S/\Delta, \quad \mathrm{Cov}(a, b) = -S_x/\Delta$$

$$r^2 \equiv \frac{S_{xy}}{S_{xx} S_{yy}} \quad \text{Pearson's } r$$

Pearson r gives an estimate of quality of fit – $r \to 1$ better is the fit. The final question : is the straight line model itself good to fit the data?

Plug in a, b in the expression for $\chi^2_{\nu=N-2}$ and perform χ^2-test.

Straight line model is generic enough for use in a few other models which can be reduced to straight line form $y = a + bx$ usually by taking logs,

$$\text{exponential} : f(x) = a\, e^{b\,x} \quad \rightarrow \quad \log f(x) = \log a + bx$$

$$\text{logarithm} : f(x) = a + b\, \log x$$

$$\text{power law} : f(x) = a\, x^{b} \quad \rightarrow \quad \log f(x) = \log a + b\, \log x$$

Polynomial model $f(x) = \sum_{i=0}^{n} a_i x^i$ can also be subjected to linear fitting.

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdot + a_n x^n$$

$$\frac{\partial \chi^2}{\partial a_k} = 0 = \frac{\partial}{\partial a_k} \sum_{i=i}^{N} \frac{\left(y_i - \cdots - a_k x_i^k - \cdots\right)^2}{\sigma_i^2}, \quad \text{where } k = 0, 1, \ldots, n$$

$$= -2 \sum_i \frac{x_i^k}{\sigma_i^2} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \cdots - a_k x_i^k \cdots - a_n x_i^n\right)$$

$$\Rightarrow a_0 \sum_i \frac{1}{\sigma_i^2} + a_1 \sum_i \frac{x_i}{\sigma_i^2} + a_2 \sum_i \frac{x_i^2}{\sigma_i^2} + \cdots + a_n \sum_i \frac{x_i^n}{\sigma_i^2} = \sum_i \frac{y_i}{\sigma_i^2}$$

$$a_0 \sum_i \frac{x_i}{\sigma_i^2} + a_1 \sum_i \frac{x_i^2}{\sigma_i^2} + a_2 \sum_i \frac{x_i^3}{\sigma_i^2} + \cdots + a_n \sum_i \frac{x_i^{n+1}}{\sigma_i^2} = \sum_i \frac{x_i y_i}{\sigma_i^2}$$

$$\cdots$$

$$a_0 \sum_i \frac{x_i^n}{\sigma_i^2} + a_1 \sum_i \frac{x_i^{n+1}}{\sigma_i^2} + a_2 \sum_i \frac{x_i^{n+2}}{\sigma_i^2} + \cdots + a_n \sum_i \frac{x_i^{2n}}{\sigma_i^2} = \sum_i x_i^n \frac{y_i}{\sigma_i^2}$$

Solving for $a_0, a_1, \ldots, a_n$ is a problem of matrix inversion,

$$\begin{pmatrix} \sum_i \frac{1}{\sigma_i^2} & \sum_i \frac{x_i}{\sigma_i^2} & \sum_i \frac{x_i^2}{\sigma_i^2} & \cdots & \sum_i \frac{x^n}{\sigma_i^2} \\ \sum_i \frac{x_i}{\sigma_i^2} & \sum_i \frac{x_i^2}{\sigma_i^2} & \sum_i \frac{x_i^3}{\sigma_i^2} & \cdots & \sum_i \frac{x_i^{n+1}}{\sigma_i^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_i \frac{x_i^n}{\sigma_i^2} & \sum_i \frac{x_i^{n+1}}{\sigma_i^2} & \sum_i \frac{x_i^{n+2}}{\sigma_i^2} & \cdots & \sum_i \frac{x_i^{2n}}{\sigma_i^2} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum_i \frac{y_i}{\sigma_i^2} \\ \sum_i \frac{x_i y_i}{\sigma_i^2} \\ \vdots \\ \sum_i \frac{x_i^n y_i}{\sigma_i^2} \end{pmatrix}$$

Subsequently, calculate the errors in parameter estimations and χ^2_{N-n} to determine the goodness of fit.

Re-writing the above equation in matrix element form,

$$\sum_{k=0}^{n} A_{jk}\, a_k = b_j, \quad \text{where } b_j = \sum_{i=1}^{N} \frac{x_i^j y_i}{\sigma_i^2} \quad \text{and } A_{jk} = \sum_{i=1}^{N} \frac{x_i^j\, x_i^k}{\sigma_i^2}, \quad j = 0, 1, \ldots, n$$

Hence, estimate of the parameters and their variances are,

$$a_j = \sum_{k=0}^{n} A_{jk}^{-1}\, b_k = \sum_{k=0}^{n} C_{jk} \left(\sum_{i=1}^{N} \frac{x_i^k\, y_i}{\sigma_i^2} \right) \quad \text{where, } A_{jk}^{-1} = C_{jk}$$

$$\sigma^2\left(a_j\right) = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a_j}{\partial y_i} \right)^2$$

Since A_{jk}'s are independent of y_j, the derivative $\partial C_{jk}/\partial y_i = 0$,

$$\frac{\partial a_j}{\partial y_i} = \sum_{k=0}^{n} C_{jk} \frac{x_i^k}{\sigma_i^2}$$

$$\sigma^2(a_j) = \sum_{k,l=0}^{n} C_{jk} C_{jl} \left(\frac{x_i^k x_i^l}{\sigma_i^2} \right) = \sum_{k,l=0}^{n} C_{jk} C_{jl} A_{kl}$$

$$= \sum_{k,l=0}^{n} C_{jk} C_{jl} C_{lk}^{-1} = \sum_{k,l=0}^{n} C_{jk} \delta_{jk} = C_{jj}$$

Diagonal elements of $A_{jk}^{-1} = C_{jk}$ are variances of the fitted parameters a_j and non-diagonal elements are the covarianvce between a_j and a_k.

Determination of parameters involve matrix inversion. Problems can arise with small or near zero diagonal elements of **A**, which may not be known a priori. One of the symptoms of such problem is extreme values of a_j differing by order(s) of magnitude. This is where Singular Value Decomposition or SVD comes in.

In what follows, we restrict to square matrix only and the discussion is heavily borrowed from Numerical Recipes.

Singular Value Decomposition

Decompose matrix **A** as product of two orthogonal matrices **U**, $\mathbf{V}^T$ and a diagonal matrix **W** with positive or zero (singular values) elements

$$\mathbf{A} = \mathbf{U}\,\mathbf{W}\,\mathbf{V}^T \quad \text{where} \quad \sum_{j=1}^{n} U_{ij}\,U_{jk} = \delta_{ik} \quad \text{and} \quad \sum_{j=1}^{n} V_{ij}\,V_{jk} = \delta_{ik}$$

This decomposition can always be done irrespective of how singular **A** is.

Decomposition is also unique, modulo permutation or linear combinations of columns.

U, **V** being orthogonal, their inverses are their transpose. **W** being diagonal, its inverse is reciprocal of its diagonal elements,

$$\mathbf{A}^{-1} = \mathbf{V}\left(\text{diag}\,(1/W_{ii})\right)\mathbf{U}^T$$

if any of $W_{ii} = 0$ then SVD would flag the occurance. Consider

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$$

It defines **A** as a linear mapping from vector subspace **x** to vector subspace **b**.

If $\mathbf{A}$ is singular then there is some subspace of $\mathbf{x}$, called nullspace which is mapped to $\mathbf{A} \cdot \mathbf{x} = 0$ and some subspace of $\mathbf{b}$, called range of $\mathbf{A}$ where it can map into. SVD explicitly constructs orthonormal bases for nullspace and range of a matrix.

1. Columns of $\mathbf{V}$ whose same-numbered elements $W_{ii} = 0$ are orthonormal basis for nullspace.
2. Columns of $\mathbf{U}$ whose same-numbered elements $W_{ii} \neq 0$ are orthonormal set of basis vectors that span the range.

Solving for $\mathbf{x}$,

$$\mathbf{x} = \mathbf{V} \cdot \left(\text{diag}\left(1/W_{ii}\right) \right) \cdot \left(\mathbf{U}^T \cdot \mathbf{b} \right)$$

When $|\mathbf{x}|^2$ is smallest, replace $1/W_{ii} = 0$ by zero if $W_{ii} = 0$. Columns of $\mathbf{V}$ are in nullspace. If $\mathbf{b}$ is not in the range of $\mathbf{A}$, then we cannot exactly solve $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$, but can find a $\mathbf{x}$ which minimizes the residual $r = |\mathbf{A}\mathbf{x} - \mathbf{b}|$.

In context of fitting, χ^2 is minimised for parameter vector $\mathbf{a}$ for possible singular / near singular $\mathbf{A}$ $\Rightarrow \chi^2 = |\mathbf{A} \cdot \mathbf{a} - \mathbf{b}|$

$$\mathbf{a} = \sum_{k=0}^{n} \left(\frac{\mathbf{U}_i \cdot \mathbf{b}}{W_{ii}} \right) \mathbf{V}_k$$

$$\sigma^2(a_j) = \sum_{k=0}^{n} \left(\frac{V_{ji}}{W_{ii}} \right)^2 \quad \text{and} \quad \text{Cov}\left(a_i, a_j\right) = \sum_{k=0}^{n} \frac{V_{ik} V_{kj}}{W_{ii}^2}$$

Nonlinear least square

Fitting to an N data set $\{x_i, y_i\}$ with a non-linear model or function having n parameters. Instead of minimising χ^2, minimised the squared differences

$$S = \frac{1}{2}\sum_{i=1}^{N}\Big(y_i - f(x_i, \{a_k\})\Big)^2 \equiv \sum_{i=1}^{N} r_i^2 \quad \text{where} \quad k = 1, 2, \ldots n$$

where r_i is called residuals. Minimum of S occurs when its change due to changes in $\{a_k\}$ is zero

$$a_k^{\alpha+1} \approx a_k^{\alpha} + \Delta_k \quad \rightarrow \quad \frac{\partial S}{\partial \Delta_k} = 0$$

where Δ_k is change in a_k^{α} in α-th iteration. It is generally difficult to solve, hence is done approximately. The first step is to linearize S by Taylor expanding about Δ_k

$$S(a_k^{\alpha+1}) = S(a_k^{\alpha} + \Delta_k) \approx S(a_k^{\alpha}) + \Delta_k \frac{\partial S}{\partial a_k^{\alpha}} + \frac{\Delta_k \Delta_l}{2} \frac{\partial^2 S}{\partial a_l^{\alpha} \partial a_k^{\alpha}} + \cdots$$

Various derivatives of S w.r.t. a_k are (without sum and iteration symbol),

$$\frac{\partial S}{\partial a_k} = \Big(y_i - f(x_i, a_k)\Big) \frac{\partial (y_i - f(x_i, a_k))}{\partial a_k} = -r_i \frac{\partial r_i}{\partial a_k} = -r_i J_{ik}$$

$$\frac{\partial^2 S}{\partial a_l \partial a_k} = -\frac{\partial r_i}{\partial a_l}\frac{\partial r_i}{\partial a_k} - \frac{\partial^2 r_i}{\partial a_k \partial a_l} \approx -\frac{\partial r_i}{\partial a_l}\frac{\partial r_i}{\partial a_k} = -J_{ik} J_{il} = H_{kl}$$

where J_{ik}'s are Jacobians and H_{kl} is called Hessian.

The iterate $a_k^{\alpha+1}$ is defined so as to minimize S w.r.t. Δ,

$$\frac{\partial S}{\partial \Delta_k} = 0 = -r_i J_{ik} - \Delta_l J_{ik} J_{il} \quad \rightarrow \quad \Delta_l = -\left(J_{ik} J_{il}\right)^{-1} J_{ik} r_i$$

The updating (iterative) equation becomes,

$$a_k^{\alpha+1} = a_k^{\alpha} - \left(J_{ik} J_{il}\right)^{-1} J_{il} r_i \quad \Rightarrow \quad \mathbf{a}^{\alpha+1} = \mathbf{a}^{\alpha} - \Big(\mathbf{J}^T \mathbf{J}\Big)^{-1} \mathbf{J}^T \mathbf{r}$$

This is Gauss-Newton method. In general it does not converge quadratically but does so as minimum is approached. Even its convergence is not guarranted but usually it does.

Levenberg-Marquardt algorithm addresses convergence problem by introducing λ, called Marquardt parameter, and a positive diagonal matrix $\mathbf{D}$, thus rewriting Δ as

$$\Delta = -\Big(\mathbf{J}^T \mathbf{J}\Big) + \lambda \mathbf{D})^{-1} \mathbf{J}^T \mathbf{r}$$