Fitting and Statistical description of data

Biased and unbiased estimates : Typically we deal with data consisting of several sets of values and we make a basic assumption that values in each set tend to cluster around some particular value. We say they have central tendency and characterize the data set with some *measures of central tendency* which are called *moments*. In physics we are familiar with two such moments – first moment is the center of mass and the second one is moment of inertia. Mathematically the k-th moment about origin (or raw moment) is defined as,

$$E\left(x^k\right) = \frac{1}{N}\sum_{i=1}^{N} w_i\, X_i^k \tag{1}$$

$$E\left(x^1\right) = \frac{1}{N}\sum_{i=1}^{N} w_i\, X_i \equiv \mu = \text{mean} \tag{2}$$

$$E\left(x^2\right) = \frac{1}{N}\sum_{i=1}^{N} w_i\, X_i^2 \tag{3}$$

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

$$\langle f(x)\rangle = \sum_x f(x)\, w_x \quad \text{and} \quad \int_x f(x)\, w(x)\, dx \tag{4}$$

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

$$E\left((x-\mu)^k\right) = \frac{1}{N}\sum_{i=1}^{N} w_i\, (X_i-\mu)^k \tag{5}$$

$$E\left((x-\mu)^1\right) = \frac{1}{N}\sum_{i=1}^{N} w_i\, (X_i-\mu) = 0 \tag{6}$$

$$E\left((x-\mu)^2\right) = \frac{1}{N}\sum_{i=1}^{N} w_i\, (X_i-\mu)^2 = \sigma^2 = \text{variance} \tag{7}$$

etc. The next two higher central moments are known as *skewness* and *kurtosis*. It is generally considered that the higher moments are statistically less *robust*. Besides, $n \geq 2$ moments may or may not exist meaning they 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 to note is the existence of weight function (or frequency) w_i which corresponds to probability distribution from which data is drawn. The k-th central moment of real-valued continuous function $f(x)$ is written as

$$E\left((x-\alpha)^k\right) = \int (x-\alpha)^k\, f(x)\, dx \tag{8}$$

The point to note is that mean is not the only first moment estimator of the data (and not necessarily the best), the other available estimators are *median* and *mode*. However, they are not widely used in physics and so we will not discuss here.

Consider a population of size N of discrete random variables x_i ($i = 1, 2, \ldots, N$). The variables are independent, but they being uncorrelated is also sufficient [1]. The population average is then given by,

$$\mu \equiv \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{9}$$

and the population variance by,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{10}$$

In many practical situations, the true variance (and also mean) of a population is not known a priori but requires to be computed somehow which becomes almost impossible if the population size is infinite. A common method is estimating the variance of large (finite or infinite) populations from a sample. We take a sample $\{y_j\}$ of size n ($j = 1, 2, \ldots, n$) and estimate the mean and the variance on the basis of this sample. The estimator of mean is,

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j \tag{11}$$

and the two estimators of variance are,

$$s_n^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \mu)^2 \quad \text{and} \quad s_{n-1}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2 \tag{12}$$

The estimated mean (11) and variance (12) depend on the data size n but μ (9) and σ^2 (10) are not. Therefore, difference between the expected value of the estimators's $\langle\theta\rangle$ 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*. The sample average is an unbiased estimator of population mean μ,

$$\langle\overline{y}\rangle = \frac{1}{N}\sum_{i=1}^{N}\overline{y}_i \;=\; \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n} y_{j,i} \;=\; \frac{1}{n}\sum_{j=1}^{n}\left[\frac{1}{N}\sum_{i=1}^{N} y_i\right]_j \;=\; \frac{1}{n}\sum_{j=1}^{n}\mu \;=\; \mu \tag{13}$$

Of the two, sample variance s_{n-1}^2 is an unbiased estimator of population variance σ^2,

$$\begin{aligned}
\langle s_{n-1}^2\rangle &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2\right]_i = \frac{1}{N(n-1)}\sum_{i=1}^{N}\sum_{j=1}^{n}\Big[(y_j-\mu)_i - (\overline{y}-\mu)_i\Big]^2 \\
&= \frac{1}{n-1}\sum_{j=1}^{n}\left[\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)_j^2 - \frac{2}{n}\sum_{k=1}^{n}\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)(y_i-\mu)_{jk} + \frac{1}{n^2}\sum_{k=1}^{n}\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)_k^2\right] \\
&= \frac{1}{n-1}\sum_{j=1}^{n}\left[\sigma_j^2 - \frac{2}{n}\sum_{k=1}^{n}\sigma_{jk}^2 + \frac{1}{n^2}\sum_{k=1}^{n}\sigma_k^2\right] \\
&= \frac{1}{n-1}\sum_{j=1}^{n}\left[\sigma_j^2 - \frac{2}{n}\sigma_j^2 + \frac{1}{n}\sigma^2\right] = \frac{1}{n-1}\sum_{j=1}^{n}\left[\sigma^2 - \frac{2}{n}\sigma^2 + \frac{1}{n}\sigma^2\right] \\
\langle s_{n-1}^2\rangle &= \frac{1}{n-1}\sum_{j=1}^{n}\frac{n-1}{n}\sigma^2 = \sigma^2
\end{aligned} \tag{14}$$

[1] Two variables A and B are independent if $\Pr(A \cap B) = \Pr(A)\Pr(B)$ and they are uncorrelated if $\mathrm{Cov}(A, B) = 0$. If two variables are independent then they are also uncorrelated but not vice-versa.

A similar calculation with the sample variance s_n^2 can be showed to be biased,

$$\langle s_n^2 \rangle = \frac{n-1}{n}\sigma^2. \tag{15}$$

In the derivation above, we have used $\mu = \sum_j \mu/n$, μ and σ being same for all samples *i.e.* independent of indices j or k and, since y_i's uncorrelated, $\mathrm{Var}(\sum_j y_j) = \sum_j \mathrm{Var}(y_j)$.

We will not bother ourselves too much about the use of $n-1$ and n for defining the two variances. It is sufficient to say that use s_n^2 when you know apriori the μ but if you estimate $\bar{y}$ from data then use s_{n-1}^2. However, none of these matters when n is large. One important property we use in the above calculation is that the data points y_i are uncorrelated *i.e.*

$$\mathrm{var}\left(\sum_i y_i\right) = \sum_i \mathrm{var}\ (y_i) \tag{16}$$

The next step is to consider the *standard error i.e.* square of uncertainty in the sample average $\bar{y}$ from the population average,

$$\begin{aligned}
\sigma_{\bar{y}} &= \bar{y} - \mu & (17)\\
\langle \sigma_{\bar{y}}^2 \rangle &= \frac{1}{N}\sum_{i=1}^{N}(\overline{y}_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n^2}\sum_{k=1}^{n}(y_k-\mu)_i^2 = \frac{1}{Nn^2}\sum_{i=1}^{N}\sum_{k=1}^{n}(y_k-\mu)_i^2 \\
&= \frac{1}{n^2}\sum_{k=1}^{n}\left[\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2\right]_k = \frac{1}{n^2}\sum_{k=1}^{n}\sigma_k^2 = \frac{\sigma^2}{n} \\
\therefore\ \langle \sigma_{\bar{y}}^2 \rangle &= \langle \overline{y}^2 \rangle - \mu^2 = \frac{\sigma^2}{n} & (18)
\end{aligned}$$

Hence, the *error* is,

$$\text{error} \equiv \sqrt{\langle \sigma_{\bar{y}}^2 \rangle} = \frac{\sigma}{\sqrt{n}} \tag{19}$$

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, the sampling distribution of the sample mean can be approximated by a normal density with mean $= \mu$ and standard deviation $= \sigma/\sqrt{n}$.* Therefore, the physical result is quoted as

$$y = \overline{y} \pm \Delta y, \tag{20}$$

then $\overline{y}$ is given by eqn. (11) and the Δy by eqn. (19) *i.e.*,

$$\Delta y = \sqrt{\frac{s_{n-1}^2}{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{j=1}^{n}(y_j - \overline{y})^2}. \tag{21}$$

From the Central Limit Theorem we know that the distribution of the sample mean is approximately normal, and so the probabilistic interpretation of the result (20) is,

$$\begin{aligned}
P\Big[\overline{y} - \Delta y \le y \le \overline{y} + \Delta y\Big] &\simeq 68.3\% \\
P\Big[\overline{y} - 2\Delta y \le y \le \overline{y} + 2\Delta y\Big] &\simeq 95.4\% \\
P\Big[\overline{y} - 3\Delta y \le y \le \overline{y} + 3\Delta y\Big] &\simeq 99.7\% \qquad (22)
\end{aligned}$$

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

Estimation of average function : In order to estimate some function f of population average $f(\mu)$. There can be two ways to estimate it,

$$(a)\ \overline{f} = \frac{1}{N}\sum_{j=1}^{n} f(y_j) \quad \text{or} \quad (b)\ f(\overline{y}) = f\left(\frac{1}{N}\sum_{j=1}^{N} y_j\right) \tag{23}$$

Assuming the sample points y_j close to the population average μ, Taylor expansion of the function $f(y_j)$ about μ can be performed and,

$$\begin{aligned} f(y_j) &= f(\mu) + (y_j-\mu)f'(\mu) + \frac{1}{2!}(y_j-\mu)^2 f''(\mu) + \cdots \qquad (24)\\ \overline{f} - f(\mu) &= \frac{1}{n}\sum_{j=1}^{n} f(y_j) - f(\mu) \\ &= f'(\mu)\left[\frac{1}{n}\sum_{j=1}^{n}(y_j-\mu)\right] + \frac{1}{2}f''(\mu)\left[\frac{1}{n}\sum_{j=1}^{n}(y_j-\mu)^2\right] + \cdots \\ &= f'(\mu)(\overline{y}-\mu) + \frac{1}{2}f''(\mu)\left[\frac{1}{n}\sum_{j=1}^{n}(y_j-\mu)^2\right] + \cdots \qquad (25) \end{aligned}$$

Therefore, $\overline{f}$ is not an unbiased estimator of $f(\mu)$ because $\langle\overline{f}\rangle \neq f(\mu)$, unless $f''(\mu)$ and higher order derivatives vanish,

$$\begin{aligned} \langle\overline{f} - f(\mu)\rangle &= f'(\mu)\frac{1}{N}\sum_{i=1}^{N}(\overline{y}-\mu)_i + \frac{1}{2}f''(\mu)\frac{1}{Nn}\sum_{j=1}^{n}\sum_{i=1}^{N}(y_j-\mu)_i^2 \\ &= f'(\mu)(\mu-\mu) + \frac{1}{2}f''(\mu)\frac{1}{n}\sum_{j=1}^{n}\sigma_j^2 + \cdots \\ &= 0 + \frac{1}{2}f''(\mu)\sigma^2 + \cdots \neq 0. \qquad (26) \end{aligned}$$

A relatively better *i.e.* less biased estimator for $f(\mu)$ is $f(\overline{y})$, which can be shown similarly following the steps in equation (24) through equation (26),

$$\begin{aligned} f(\overline{y}) - f(\mu) &= f(\mu) + f'(\mu)(\overline{y}-\mu) + \frac{1}{2!}f''(\mu)(\overline{y}-\mu)^2 + \cdots - f(\mu) \\ &= f'(\mu)(\overline{y}-\mu) + \frac{1}{2}f''(\mu)\frac{1}{n^2}\sum_{j=1}^{n}(y_j-\mu)^2 + \cdots \qquad (27)\\ \langle f(\overline{y}) - f(\mu)\rangle &= 0 + \frac{1}{2n}f''(\mu)\sigma^2 + \cdots \qquad (28) \end{aligned}$$

Thus we can see, the bias for the estimator $f(\overline{y})$ is of order $1/n$ less than that of $\overline{f}$. But then, it raises a problem in defining the variance of $f(\overline{y})$. The quantity $\overline{f^2} - \overline{f}^2$ is the variance of $f(y_j)$ and not that

of $f(\overline{y})$. This is where **jackknife** and **bootstrap** come in. A brief but good discussion can be found at [1].

Jackknife : Jackknife is a statistical method for estimating and compensating for bias and for deriving estimates of standard errors and confidence intervals. Jackknife statistics are created by systematically dropping out subsets of data one at a time and assessing the resulting variation in the studied parameter. Starting from a sample of n values, the jackknife begins by throwing out the first value, leaving a jackknife sample set of $n-1$, called the *resampled* values. In the next step it drops the second, then third and so on. Analysing such reduced sample set, suppose we measure jackknife sample average,

$$\overline{y}_k = \frac{n\overline{y} - y_k}{n-1} = \frac{1}{n-1}\sum_{j\neq k} y_j \tag{29}$$

This process results in a set of parameter values, the *deleted averages* $\{\overline{y}_k,\ k = 1, 2, \ldots, n\}$, and we define jackknife average as,

$$\overline{y}_{\rm JK} = \frac{1}{n}\sum_{k=1}^{n} \overline{y}_k \tag{30}$$

However, the jackknifing does not change the average, $\overline{y}_{\rm JK}$ is still equal to the sample average $\overline{y}$,

$$\overline{y}_{\rm JK} \;=\; \frac{1}{n}\sum_{j=1}^{n} \overline{y}_j \;=\; \frac{1}{n}\sum_{j=1}^{n} \frac{n\overline{y} - y_j}{n-1} \;=\; \frac{n\overline{y} - \frac{1}{n}\sum_{j=1}^{n} y_j}{n-1} \;=\; \frac{n\overline{y} - \overline{y}}{n-1} \;=\; \overline{y} \tag{31}$$

The jackknife estimator of standard error $\sigma^2_{\overline{y}}$ (17) is,

$$\begin{aligned}
\sum_{j=1}^{n} \left(\overline{y}_j - \overline{y}_{\rm JK}\right)^2 \;&=\; \sum_{j=1}^{n} \left(\frac{n\overline{y} - y_j}{n-1} - \overline{y}\right)^2 = \frac{1}{n-1}\sum_{j=1}^{n} \frac{1}{n-1}\left(y_j - \overline{y}\right)^2 = \frac{s^2_{n-1}}{n-1} \\
\sigma^2_{\overline{y}} \;&=\; \frac{s^2_{n-1}}{n} = \frac{n-1}{n}\sum_{j=1}^{n} \left(\overline{y}_j - \overline{y}_{\rm JK}\right)^2 = (n-1)\left(\overline{y^2_j} - \overline{y}^2_{\rm JK}\right) = (n-1)\sigma^2_{\rm JK}
\end{aligned} \tag{32}$$

Define $f^{\rm JK}$ as a function of jackknife variables $\overline{y}_j$, defined in equation (29), to estimate $f(\mu)$,

$$f^{\rm JK}_j \equiv f\left(\overline{y}_j\right) \quad \Rightarrow \quad \overline{f}_{\rm JK} = \frac{1}{n}\sum_{j=1}^{n} f^{\rm JK}_j \tag{33}$$

A few lines of calculation gives the standard error in this estimate,

$$\sigma^2_{f^{\rm JK}} = \frac{n-1}{n}\sum_{j=1}^{n} \left(f^{\rm JK} - \overline{f}^{\rm JK}\right)^2 = (n-1)\left(\overline{\overline{f}^2}_{\rm JK} - \overline{f}^2_{\rm JK}\right) \tag{34}$$

The bias in estimating $f(\mu)$ using jackknife $\overline{f}^{\rm JK}$ can be calculated by expanding $f^{\rm JK}$ in Taylor expansion, similar to what done in (24), yielding

$$\left\langle \overline{f}^{\rm JK} - f(\mu) \right\rangle = \frac{1}{2(n-1)} f''(\mu)\sigma^2_y + \cdots \tag{35}$$

which is almost identical to (28). However, the bias can further be reduced to $\mathcal{O}(1/n^2)$ by using the estimate,

$$F^{\rm JK} = n\, f(\overline{y}) - (n-1)\,\overline{f}^{\rm JK} \tag{36}$$

Bootstrap : The bootstrap method involves resampling of the data points (of unknown distribution) by choosing a set of data points randomly with replacement from the sampled / experimental data. In contrast, jackknife, resampled $\{n\}$ set of data, each of size $n-1$, by systematically omitting one data point at a time. The method we are going to discuss here is known as *empirical bootstrap*. Let us suppose the bootstrap generates $\mathcal{B}$ data sets each containing ν data points, randomly choosing from the original data set $\{x_j\}$ with replacement,

$$\begin{array}{lcll} \underline{\text{Bootstrap samples}} & & \underline{\text{average}} & \underline{\text{variance}} \\ \left\{x_1^{(1)}, x_2^{(1)}, \ldots, x_\nu^{(1)}\right\} & \rightarrow & \overline{x}^{(1)} & s_{\nu-1}^{(1)\,2} \\ \left\{x_1^{(2)}, x_2^{(2)}, \ldots, x_\nu^{(2)}\right\} & \rightarrow & \overline{x}^{(2)} & s_{\nu-1}^{(2)\,2} \\ \vdots & & \vdots & \vdots \\ \left\{x_1^{(\mathcal{B})}, x_2^{(\mathcal{B})}, \ldots, x_\nu^{(\mathcal{B})}\right\} & \rightarrow & \overline{x}^{(\mathcal{B})} & s_{\nu-1}^{(\mathcal{B})\,2} \end{array} \tag{37}$$

An empirical bootstrap sample is a resample of the same size of the original *i.e.* $\nu = n$. Since the 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 bootstrap sample set. The probability of x_i to be chosen is $1/n \equiv p$ for each drawing. Suppose x_i appears in a bootstrap sample n_i times, then

$$\sum_{i=1}^{n} n_i = n \tag{38}$$

Therefore, the probability that x_i occurring n_i times is a binomial distribution, either occurring p or not $q = 1-p$,

$$P(n_i) = \binom{n}{n_i} p^{n_i}(1-p)^{n-n_i} \tag{39}$$

$$\text{mean } \overline{n}_i = \lambda = np = 1 \tag{40}$$

$$\text{variance } \sigma_{n_i}^2 = \overline{n_i^2} - \overline{n}_i^2 = np(1-p) = 1 - \frac{1}{n} \tag{41}$$

$$\Rightarrow \overline{n_i^2} = 2 - \frac{1}{n} \tag{42}$$

$$\text{cov}(n_i, n_j) = \overline{n_i n_j} - \overline{n}_i \overline{n}_j = -\frac{1}{n} \quad \text{for } i \neq j \tag{43}$$

$$\Rightarrow \overline{n_i n_j} = 1 - \frac{1}{n} \tag{44}$$

When $n \to \infty$, keeping $\lambda = np$ fixed, the binomial distribution goes over to the Poisson distribution, with mean and variance λ which is equal to 1 in the present case. Please note, that because of the constraint (38), n_i and n_j are not independent and $\text{cov}(n_i, n_j) = -1/n$ where $i \neq j$.

The averages of the bootstrap samples in (37) are

$$\overline{x}^{(\alpha)} = \frac{1}{n}\sum_{j=1}^{n} x_i^{(\alpha)} = \frac{1}{n}\sum_{j=1}^{n} n_i^{(\alpha)} x_i \tag{45}$$

The bootstrap mean of the averages of all the bootstrap data sets can thus be calculated as

$$\begin{aligned}
\overline{x}^{\mathcal{B}} &= \frac{1}{\mathcal{B}}\sum_{\alpha=1}^{\mathcal{B}}\overline{x}^{(\alpha)} = \frac{1}{n}\frac{1}{\mathcal{B}}\sum_{\alpha=1}^{\mathcal{B}}\sum_{i=1}^{n} n_i^{(\alpha)} x_i \\
&= \frac{1}{n}\sum_{1=1}^{n} x_i \,\frac{1}{\mathcal{B}}\sum_{\alpha=1}^{\mathcal{B}} n_i^{(\alpha)} = \frac{1}{n}\sum_{i=1}^{\nu} x_i \overline{n}_i = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \qquad (46) \\
\therefore\ \left\langle \overline{x}^{\mathcal{B}} \right\rangle &= \frac{1}{n}\sum_{j=1}^{n}\frac{1}{n}\sum_{i=1}^{n} x_{i,j} = \langle \overline{x} \rangle = \mu \qquad (47) \\
\sigma^2_{\mathcal{B}_{\overline{x}}} &= \overline{(x^{\mathcal{B}})^2} - \left(\overline{x}^{\mathcal{B}}\right)^2 = \frac{1}{\mathcal{B}}\sum_{\alpha=1}^{\mathcal{B}}\left(\overline{x}^{(\alpha)} - \overline{x}^{\mathcal{B}}\right)^2 \qquad (48)
\end{aligned}$$

Hence, the bootstrap average is an unbiased estimator of the population mean. This actually follows from the Central Limit Theorem which states that the means from a large number of independent random samples will be approximately normally distributed regardless of the underlying distribution. The equation (48) defines the bootstrap standard error of the bootstrap mean. A few lines of nasty but straight forward calculations using the equations (41), (43), (45) and (46) give the following and hence the standard error

$$\begin{aligned}
\sigma_{\mathcal{B}^2_{\overline{x}}} &= \frac{n-1}{n^3}\sum_i x_i^2 - \frac{1}{n^3}\sum_{i\neq j} x_j x_j \qquad (49) \\
\left\langle \sigma^2_{\mathcal{B}_{\overline{x}}} \right\rangle &= \frac{n-1}{n}\sigma^2 \qquad (50)
\end{aligned}$$

For bootstrap standard error of variance, first the average of variances $s_{n-1}^{(k)\,2}$ in (37) is defined, followed by estimator of the variance of the original sample variances $S^2_{\mathcal{B}-1}$,

$$\overline{s}^2_{\mathcal{B}} = \frac{1}{\mathcal{B}}\sum_{\alpha=1}^{\mathcal{B}} s_{n-1}^{(\alpha)\,2} \quad \text{and} \quad S^2_{\mathcal{B}-1} = \frac{1}{\mathcal{B}-1}\sum_{\alpha=1}^{\mathcal{B}}\left(s_{n-1}^{(\alpha)\,2} - \overline{s}^2_{\mathcal{B}}\right)^2 \qquad (51)$$

The standard error square for variance is thus

$$\sigma^2_{\mathcal{B}} = \frac{1}{\mathcal{B}}\sum_{\alpha=1}^{\mathcal{B}}\left(s_{n-1}^{(\alpha)\,2} - \overline{s^2_{\mathcal{B}}}\right)^2 \qquad (52)$$

The calculation of average function is similar to jackknife (33),

$$f^{\mathcal{B}}_{\alpha} \equiv f(\overline{x}^{(\alpha)}) \quad \Rightarrow \quad \overline{f}^{\mathcal{B}} = \frac{1}{\mathcal{B}}\sum_{\alpha=1}^{\mathcal{B}} f^{\mathcal{B}}_{\alpha} \qquad (53)$$

and likewise the variance is similar to the jackknife case, as usual with a bias of $1/2n$. This bias too can be reduced to $\mathcal{O}(1/n^2)$ by defining,

$$F^{\mathcal{B}} = 2f(\overline{x}) - \overline{f}^{\mathcal{B}} \qquad (54)$$

Comparing means and variances : It often happens in Physics, an observable, particularly the mean value along with its standard deviation or error, is measured (say Higgs boson mass) with certain

amount of statistics and then measure that again after an upgrade (for instance after upgrading LHC with higher luminosity and larger statistics). The standard deviation or error will be different no doubt, but are the means statistically same or different? A difference of means can be very small compared to s^2_{n-1} and yet significant if the statistics is large and vice-versa. The significance of difference in means uses the standard error (19) and is assessed by *Student's t-test*. If our data $\{y_i\}$ of size n is obtained from a normally distributed population $\mathcal{N}(\mu, \sigma^2)$ then we have the following distribution of the random variables,

$$t = \frac{\overline{y} - \mu}{\sigma/\sqrt{n}} \quad \rightarrow \quad \text{normal distribution} \tag{55}$$

$$t = \frac{\overline{y} - \mu}{s^2_{n-1}/\sqrt{n}} \quad \rightarrow \quad \text{Student's distribution with } n-1 \text{ degrees of freedom} \tag{56}$$

$$\tag{57}$$

Student's distribution $A(t|\nu)$ is the probability for ν degrees of freedom that a statistic t (measuring the observed difference of means) would be smaller than the observed value if the means are the same. The null hypothesis is

$$H_0: \text{ the means of two different data set are equal} \tag{58}$$

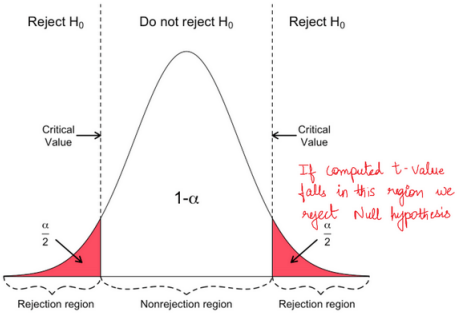
In other words, two means are significantly different if $A(t|\nu) > 0.99$ or $1 - A(t|\nu)$ is the significance level at which the hypothesis H_0 in (58) that the means are equal is disproved. The Student's t-distribution has the form,

$$A(t|\nu) = \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \int_{-t}^{t} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} dx = \frac{1}{\sqrt{\nu}\, B(1/2, \nu/2)} \int_{-t}^{t} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} dx \tag{59}$$

where Γ and B are respectively gamma and beta functions. The limiting values are

$$A(0|\nu) = 0 \quad \text{and} \quad A(\infty|\nu) = 1 \quad \text{where, significance} = 1 - A(t|\nu) \tag{60}$$

The significance is, therefore, a number between 0 and 1 and gives the probability that $|t|$ can be this large or larger just by chance for distributions with equal means. Hence, a small value of significance 0.05 or 0.01 implies that the observed difference is quite significant.



Our *null hypothesis* H_0 is that the *two means are the same and any measured difference is due to chance only*. The test t is calculated and it is compared with the critical value t_{crit} (which depends on significance level of the test *i.e.* the probability of erroneously rejecting null hypothesis). For example, the calculated $t = 1.92$ for $n = 25$ and hence $\nu = n - 1 = 24$, the level of significance is $\alpha = 0.05$ then

$|t_{\text{crit}}| = 2.064$ for $\nu = 24$ (refer to a t-table). Therefore, $|t| < |t_{\text{crit}}|$ meaning H_0 cannot be rejected with $95\% = 100 \times (1 - \alpha)$ confidence. What if $\alpha = 0.1$? Then from the t-table we get $|t_{\text{crit}}| = 1.711$, implying we reject the null hypothesis with $90\% = 100 \times (1 - \alpha)$ confidence. These two conclusions may sound strange that at 90% we are *rejecting* a hypothesis which we *accept* at 95% confidence. A more user-friendly way to interpret this is – at $\alpha = 0.05$ the possibility of erroneously *rejecting* a correct hypothesis is $100 \times \alpha = 5\%$, whereas for $\alpha = 0.10$ that error is $100 \times \alpha = 10\%$.

We can as well compare the variances where our null hypothesis is

$$\begin{aligned} H_0 &: \text{two observed samples have the same variance } \sigma_1^2 = \sigma_2^2 \\ H_A &: \text{of the two observed samples, 1 has larger variance than 2 } \sigma_1^2 > \sigma_2^2 \text{ or } \sigma_1^2 < \sigma_2^2 \text{ or } \sigma_1^2 \neq \sigma_2^2 \end{aligned} \tag{61}$$

The variance can differ because of differences in data size. This test is done by the F-test. The statistics F is the ratio of one variance to the other

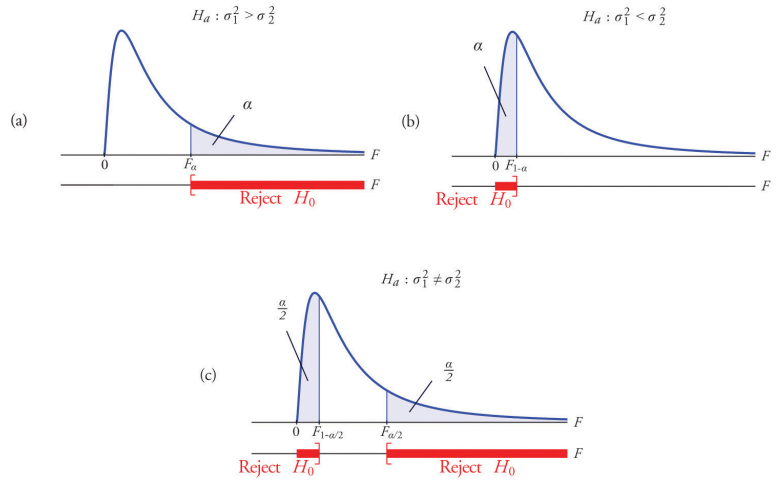
$$F = s_1^2 \,/\, s_2^2 \quad \text{where} \quad s_1^2 \geq s_2^2 \tag{62}$$

The basic assumptions are data must be drawn from (approximately) normal distribution and the samples are independent events. The condition in equation (62) simply refers to the practice that the larger variance should always go in the numerator to get a *right-tailed test* since it is easier to calculate. The F-distribution is denoted by $Q(F|\nu_1, \nu_2)$ and is evaluated by using *incomplete beta function* $I_x(a, b)$,

$$Q(F|\nu_1, \nu_2) = I_{\frac{\nu_2}{\nu_2 + \nu_1 F}}\left(\frac{\nu_2}{2}, \frac{\nu_1}{2}\right) \quad \text{where} \quad I_x(a, b) = \frac{1}{B(a, b)} \int_0^x t^{a-1}(1-t)^{b-1}\, dt \tag{63}$$

where $a, b > 0$ and $B(a, b)$ is the beta function. The limiting values of $Q(F|\nu_1, \nu_2)$ is

$$Q(0|\nu_1, \nu_2) = 1 \quad \text{and} \quad Q(\infty|\nu_1, \nu_2) = 0 \tag{64}$$



Suppose same experiment is performed at two labs A and B and the data collected yields

$$\begin{aligned} \text{Lab } A &: n_A = 16,\ s_A^2 = 2.09 \\ \text{Lab } B &: n_B = 21,\ s_B^2 = 1.10 \end{aligned}$$

The test hypothesis is

$$\begin{aligned} H_0 &: \sigma_A^2 = \sigma_B^2 \quad \text{versus} \\ H_1 &: \sigma_A^2 > \sigma_B^2 \quad @\ \alpha = 0.10 \end{aligned}$$

i.e. at 10% level of significance. The degrees of freedom is $\nu_A = 16-1 = 15$ and $\nu_B = 21-1 = 20$ and the F statistics is $F = s_A^2/s_B^2 = 2.09/1.10 = 1.90$. Then from F-table, $Q(F|\nu_1, \nu_2) = Q(1.90|15, 20) = 1.80$ at $\alpha = 0.10$. Hence, the H_0 rejection region $[1.84, \infty]$. Hence we can conclude that at 10% level of significance, the data provide evidence to reject H_0 *i.e.* the variances are not equal, rather $\sigma_A^2 > \sigma_B^2$.

Goodness of fit : Given two sets of data, the questions on sameness of mean and variance can be combined into a single query (null hypothesis)

$$H_0 : \text{Two data sets are drawn from the same population distribution} \tag{65}$$

For instance, are the data obtain on top quark at LHC and Fermilab come from the same distribution *i.e.* comparable? Or, say, the distribution of the brightness / luminosity of stars in Andromeda is same as that of Milky Way, both being spiral galaxies and of approximately same age.

Often our query can involve a single data set, for which null hypothesis can be

$$H_0 : \text{Data are drawn from a known distribution} \tag{66}$$

whose mean and variance may or may not be known. For example, is the covid infection across age following some distribution? Or, does the distribution of marks in NEST exam are normally distributed? In the first step the data, discrete or continuous, are divided into k bins and the test statistics is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{67}$$

where O_i is the observed frequency or number of events in the i-th bin and E_i is what is expected according to some known distribution. The O_i are in general integers but E_i may not necessarily be. A large χ^2 value indicates that the null hypothesis in (66) is rather unlikely. The χ^2 probability function $P(\chi^2|\nu)$ determined the outcome of the test. The $P(\chi^2|\nu)$ is defined as the probability that the observed χ^2-statistics for accepting a model should be less than $\chi^2_{\rm crit}$. However, similar to interpreting t-test, more user-friendly would be to define $Q(\chi^2|\nu)$, the probability that the observed χ^2-statistics will exceed the value by chance even for a correct model. In both case, ν is the degrees of freedom. The form of the P, and hence $Q = 1 - P$ is *incomplete Gamma function* and their limiting values are,

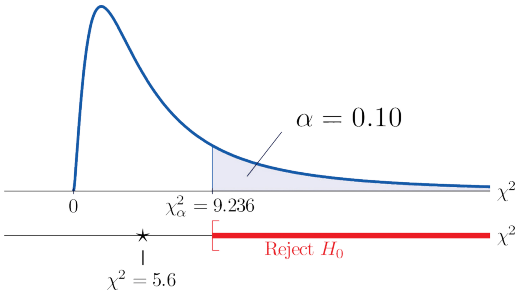
$$P(\chi^2|\nu) = P\left(\frac{\nu}{2}, \frac{\chi^2}{2}\right) \quad \text{and} \quad Q(\chi^2|\nu) = Q\left(\frac{\nu}{2}, \frac{\chi^2}{2}\right) \quad \text{where,} \tag{68}$$

$$\begin{aligned} P(a,x) &= \frac{1}{\Gamma(a)} \int_0^x e^{-t}\, t^{a-1}\, dt \quad \text{and} \quad P(0|\nu) = 0,\ \ P(\infty|\nu) = 1 \\ Q(a,x) &= 1 - P(a,x) \quad \text{and} \quad Q(0|\nu) = 1,\ \ Q(\infty|\nu) = 0 \end{aligned} \tag{69}$$

Consider the example of a dice, rolled $N = 60$ times (to keep calculation simple, but in practice can be few hundreds) and the number of times each face landed up is given in the table below,

Face value	Unbiased distribution $f(x)$	Expected frequency $E = N \times f(x)$	Observed frequency O
1	1/6	10	9
2	1/6	10	15
3	1/6	10	9
4	1/6	10	8
5	1/6	10	6
6	1/6	10	13

The null hypothesis is H_0 : *the dice is fair / unbiased*. The calculation of χ^2 statistics (67) is simple, $\chi^2 = 5.6$ for $\nu = 5$ degrees of freedom. If we decide to test at 90% level of significance $1-\alpha = 0.90$, then the critical value defining the rejection region can be read off from a χ^2 table, which is $\chi^2_{\rm crit} = 9.236$. Since $\chi^2 < \chi^2_{\rm crit}$ implies we do not have enough evidence to reject the null hypothesis *i.e.* we cannot say the dice is biased.



However, if the need is to compare two data set and whether they are drawn from the same population then the corresponding null hypothesis is given in (65) and the χ^2 statistics is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - A_i)^2}{O_i + A_i} \tag{70}$$

The rest of the testing follows exactly the same route as above.

Modeling of data : The χ^2 test can be put in a slightly different context. If we are planning to fit N data points (x_i, y_i, σ_i), where σ_i is the standard deviation of y_i, to a model with M parameters, then *maximum likelihood estimator* of the model parameters is obtained by minimizing the quantity with respect to the parameters,,

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i; a_1, a_2, \ldots, a_M)}{\sigma_i} \right)^2, \quad \text{dof } \nu = N - M \tag{71}$$

$$0 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i)}{\sigma_i^2} \right) \left(\frac{\partial y(x_i; \ldots, a_k, \ldots)}{\partial a_k} \right), \quad k = 1, 2, \ldots, M \tag{72}$$

Upon minimization, a rule of thumb is that a value of χ^2/ν for a moderately good fit is ≈ 1. Technically the distribution of the RHS in (71) is χ^2 only if the y_i's are drawn from a Gaussian distribution. When they are not, it is more commonly known as just *least square method*. Before getting into fitting of data, here is a quick example of the use of (71). Consider the measurement of the mass M_Z of the Z^0 boson at CERN made by four different detectors – L3, OPAL, Aleph and Delphi,

Detector	Mass (GeV/c^2)
L3	91.161 ± 0.013
OPAL	91.174 ± 0.011
Aleph	91.186 ± 0.013
Delphi	91.188 ± 0.013

The *weighted* average of the four measurements is $\overline{M}_Z = 91.177$. Our null hypothesis is $H_0 =$ *it is justified to say 91.177 GeV/c^2 is the global average of the Z^0 mass.* Then the χ^2 statistics, according to the definition above (71)

$$\chi^2_{\nu=3} = \sum_{i=1}^{4} \frac{\left(M_i - \overline{M}_Z\right)^2}{\sigma_i^2} \approx 2.78 \quad \text{or} \quad \chi^2/3 \approx 2.78/3 = 0.926 \tag{73}$$

If we test our null hypothesis at 90% level of significance *i.e.* $\alpha = 0.10$, then $\chi^2_{\rm crit}/3 = 1.547 > \chi^2/3$, therefore we have no good reason to reject H_0. There is a lesson in this example – if $\chi^2/\nu \approx 1$ then a model can be accepted at a 90% level of significance.

Linear regression : It is essentially linear fitting *i.e.* fitting data to a straight line. We model a set of N data points (x_i, y_i, σ_i), where σ_i's are the errorbars on y_i's, to a straight line model,

$$y(x) \equiv y(x; a, b) = a + bx \quad \Rightarrow \quad \chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2 \tag{74}$$

The minimization of χ^2 as in (72), with respect to a and b yields,

$$\begin{aligned} \frac{\partial \chi^2}{\partial a} &= -2 \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2} = 0 \\ \frac{\partial \chi^2}{\partial b} &= -2 \sum_{i=1}^{N} \frac{x_i \left(y_i - a - bx_i\right)}{\sigma_i^2} = 0 \end{aligned} \tag{75}$$

Let us introduce a convenient notation, as given in Numerical Recipes, using $\sum_i$ for $\sum_{i=1}^{N}$

$$\begin{aligned} S &= \sum_i \frac{1}{\sigma_i^2}, \qquad S_x = \sum_i \frac{x_i}{\sigma_i^2}, \qquad S_y = \sum_i \frac{y_i}{\sigma_i^2} \\ S_{xx} &= \sum_i \frac{x_i^2}{\sigma_i^2}, \qquad S_{xy} = \sum_i \frac{x_i y_i}{\sigma_i^2} \end{aligned} \tag{76}$$

$\therefore$ leading to,

$$\begin{aligned} S_y &= aS + bS_x \\ S_{xy} &= aS_x + bS_{xx} \end{aligned} \tag{77}$$

The parameters a and b are obtained by solving the above equations (77),

$$\begin{aligned} a &= \frac{S_{xx}S_y - S_x S_{xy}}{\Delta} \qquad b = \frac{S S_{xy} - S_x S_y}{\Delta} \\ \text{where, } \Delta &= S S_{xx} - (S_x)^2 \end{aligned} \tag{78}$$

The next step is to estimate the errors on the parameters a and b and how are they related. Using the propagation of error, we get

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

$$\frac{\partial b}{\partial y_i} = \frac{S\, x_i - S_x}{\Delta\ \sigma_i^2} \tag{79}$$

$$\therefore \qquad \sigma_a^2 = S_{xx}/\Delta \quad \text{and} \quad \sigma_b^2 = S/\Delta \tag{80}$$

$$\mathrm{Cov}(a,b) = -S_x/\Delta \tag{81}$$

$$r^2 = \frac{S_{xy}}{S_{xx} S_{yy}} \tag{82}$$

where r is the Pearson's r ($\leq r \leq 1$) which gives an estimate of *quality* of fit. As $r^2 \to 1$ better is the fit. But is the straight line model itself good to fit the data? This brings us to the final step where χ^2 is calculated by plugging in the values of (a, b) in (74) with degrees of freedom $\nu = N - 2$. and thus determining the goodness of fit as described before where the rule of thumb is the fit is *good* if $\chi^2/\nu \leq 1$.

The 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.

$$\text{exponential :} \qquad f(x) = a\, e^{b\,x} \quad \to \quad \log f(x) = \log a + bx \tag{83}$$

$$\text{logarithm :} \qquad f(x) = a + b\, \log x \tag{84}$$

$$\text{powe law :} \qquad f(x) = a\, x^b \quad \to \quad \log f(x) = \log a + b\, \log x \tag{85}$$

A polynomial model, $f(x) = \sum_{k=0}^{n} a_k x^k$ can also be subjected to linear fitting,

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdot + a_n x^n \tag{86}$$

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

$$= -2 \sum_i \frac{x_i^k}{\sigma_i^2} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \cdots - a_k x_i^k \cdots - a_n x_i^n\right) \tag{87}$$

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

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

$$\vdots$$

$$a_0 \sum_i \frac{x_i^n}{\sigma_i^2} + a_1 \sum_i \frac{x_i^{n+1}}{\sigma_i^2} + a_2 \sum_i \frac{x_i^{n+2}}{\sigma_i^2} + \cdots + a_n \sum_i \frac{x_i^{2n}}{\sigma_i^2} = \sum_i x_i^n \frac{y_i}{\sigma_i^2} \tag{88}$$

Solving for $a_0, a_1, \ldots, a_n$ from the set of equations (88) is a problem of matrix inversion in linear

algebra,

$$\begin{pmatrix} \sum_i \frac{1}{\sigma_i^2} & \sum_i \frac{x_i}{\sigma_i^2} & \sum_i \frac{x_i^2}{\sigma_i^2} & \cdots & \sum_i \frac{x^n}{\sigma_i^2} \\ \sum_i \frac{x_i}{\sigma_i^2} & \sum_i \frac{x_i^2}{\sigma_i^2} & \sum_i \frac{x_i^3}{\sigma_i^2} & \cdots & \sum_i \frac{x_i^{n+1}}{\sigma_i^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_i \frac{x_i^n}{\sigma_i^2} & \sum_i \frac{x_i^{n+1}}{\sigma_i^2} & \sum_i \frac{x_i^{n+2}}{\sigma_i^2} & \cdots & \sum_i \frac{x_i^{2n}}{\sigma_i^2} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum_i \frac{y_i}{\sigma_i^2} \\ \sum_i \frac{x_i y_i}{\sigma_i^2} \\ \vdots \\ \sum_i \frac{x_i^n y_i}{\sigma_i^2} \end{pmatrix} \tag{89}$$

You are free to use your favorite inverter to solve for the least square values of the parameters. In the subsequent steps, calculate the errors for the parameters and the χ^2/ν, where $\nu = N-n$, to determine the goodness of fit. The calculations of variances of the fitted parameters a_k and the covariance matrix can be made more transparent if we write all the above expressions in matrix element form,

$$\sum_{k=0}^{n} A_{jk}\, a_k = b_j, \quad \text{where } b_j = \sum_{i=1}^{N} \frac{x_i^j y_i}{\sigma_i^2} \quad \text{and } A_{jk} = \sum_{i=1}^{N} \frac{x_i^j\, x_i^k}{\sigma_i^2}, \quad j = 0, 1, \ldots, n \tag{90}$$

The $\mathbf{A}$ here is obviously an $(n+1) \times (n+1)$ matrix and $\mathbf{b}$ is a vector of length $(n+1)$. Hence the solution *i.e.* the estimate of the parameters and their variances are,

$$a_j = \sum_{k=0}^{n} A_{jk}^{-1}\, b_k = \sum_{k=0}^{n} C_{jk} \left(\sum_{i=1}^{N} \frac{x_i^k\, y_i}{\sigma_i^2} \right) \tag{91}$$

$$\sigma^2(a_j) = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a_j}{\partial y_i} \right)^2 \tag{92}$$

Since A_{jk}'s are independent of y_j, from equation (91 we have

$$\begin{aligned} \frac{\partial a_j}{\partial y_i} &= \sum_{k=0}^{n} C_{jk} \frac{x_i^k}{\sigma_i^2} \\ \sigma^2(a_j) &= \sum_{k,l=0}^{n} C_{jk} C_{jl} \left(\frac{x_i^k\, x_i^l}{\sigma_i^2} \right) = \sum_{k,l=0}^{n} C_{jk} C_{jl}\, A_{kl} \\ &= \sum_{k,l=0}^{n} C_{jk} C_{jl} C_{lk}^{-1} = \sum_{k,l=0}^{n} C_{jk} \delta_{jk} = C_{jj} \end{aligned} \tag{93}$$

Therefore, in other words, the diagonal elements of $A_{jk}^{-1} = C_{jk}$ are the variances of the fitted parameters a_j and the non-diagonal elements are covarianvce between a_j and a_k. Since the dtermination of the parameters involve matrix inversion there can arise certain issues with small or near zero diagonal elements of $\mathbf{A}$, which may not be known a priori. One of the symptoms of existence of such problem(s) is obtaining extreme values of a_j differing by order(s) of magnitude. Since, the determination of a_j's are minimization problem in multi-dimension parameter space, this can happen when one lands up in some arcane local minimum. This is where *Singular Value Decomposition* or SVD comes in. Here we will discuss SVD in the context of square matrix only. This discussion is heavily borrowed from Numerical Recipes.

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

$$\mathbf{A} = \mathbf{U}\,\mathbf{W}\,\mathbf{V}^T \quad \text{where } \sum_{j=1}^{n} U_{ij} U_{jk} = \delta_{ik} \text{ and } \sum_{j=1}^{n} V_{ij} V_{jk} = \delta_{ik} \tag{94}$$

The point here is that the decomposition in (94) can always be done irrespective of how singular $\mathbf{A}$. The decomposition is also unique modulo permutation or linear combinations of columns. The $\mathbf{U}$ and $\mathbf{V}$ being orthogonal, their inverses are equal to their transpose and $\mathbf{W}$ being diagonal, its inverse is just the reciprocal of the diagonal elements. Hence,

$$\mathbf{A}^{-1} = \mathbf{V}\left(\mathrm{diag}\,(1/W_{ii})\right)\mathbf{U}^T \tag{95}$$

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

$$\mathbf{A}\cdot\mathbf{x} = \mathbf{b} \tag{96}$$

In terms of mapping, the equation (96) defines $\mathbf{A}$ as a linear mapping from the vector subspace $\mathbf{x}$ to the vector subspace $\mathbf{b}$. If $\mathbf{A}$ is singular then there is some subspace of $\mathbf{x}$ (called *nullspace*) which is mapped to $\mathbf{A}\cdot\mathbf{x} = 0$. There is also some subspace of $\mathbf{b}$ (called *range* of $\mathbf{A}$), where $\mathbf{A}$ can map $\mathbf{x}$ into. SVD explicitly constructs orthonormal bases for the nullspace and range of a matrix.

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

Hence, any column of $\mathbf{V}$ whose corresponding $W_{ii} = 0$ yields a solution. If $\mathbf{b} \neq 0$ and also $\mathbf{b}$ is in the range of $\mathbf{A}$ then we can we have a solution $\mathbf{x}$. To find it, say when $|\mathbf{x}|^2$ is smallest, simply *replace* $1/W_{ii}$ *by zero if* $W_{ii} = 0$, and then compute

$$\mathbf{x} = \mathbf{V}\cdot\left(\mathrm{diag}\,(1/W_{ii})\right)\cdot\left(\mathbf{U}^T\cdot\mathbf{b}\right) \tag{97}$$

where the columns of $\mathbf{V}$ are in nullspace. If $\mathbf{b}$ is not in the range of $\mathbf{A}$, then we cannot exactly solve $\mathbf{A}\mathbf{x} = \mathbf{b}$, but we can find a $\mathbf{x}$ which minimizes the *residual* $r = |\mathbf{A}\mathbf{x} - \mathbf{b}|$. The same problem can be viewed in yet another way, consider the following,

$$\left.\begin{array}{rcl} \mathbf{A}\mathbf{A}^T &=& \mathbf{U}\mathbf{W}\mathbf{V}^T\cdot\mathbf{V}\mathbf{W}^T\mathbf{U}^T \\ \mathbf{A}^T\mathbf{A} &=& \mathbf{V}\mathbf{W}^T\mathbf{U}\cdot\mathbf{U}\mathbf{W}\mathbf{V}^T \end{array}\right\} \;\Rightarrow\; \left\{\begin{array}{rcl} \mathbf{A}\mathbf{A}^T\mathbf{U} &=& \mathbf{U}\mathbf{W}^2 \\ \mathbf{A}^T\mathbf{A}\mathbf{V} &=& \mathbf{V}\mathbf{W}^2 \end{array}\right. \tag{98}$$

Since $\mathbf{W}$ is diagonal, the $\mathbf{U}$ and $\mathbf{V}$ contain the eigenvectors of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$. Hence, it is essentially an eigenvalue problem. The eigenvectors can be found by power iteration method. For an exact implementation of construction of orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ see Numerical Recipes.

Getting back to the present context of χ^2 fitting, we want to minimize the χ^2 for an $\mathbf{a}$, the vector of the parameters,

$$\chi^2 = |\mathbf{A}\cdot\mathbf{a} - \mathbf{b}| \tag{99}$$

for possible *ill-conditioned i.e.* near singular matrix $\mathbf{A}$. This is exactly what SVD does and without going into the details or proofs, the solution of least square problem

$$\mathbf{a} = \sum_{k=0}^{n}\left(\frac{\mathbf{U}_i\cdot\mathbf{b}}{W_{ii}}\right)\mathbf{V}_k, \quad \sigma^2(a_j) = \sum_{k=0}^{n}\left(\frac{V_{ji}}{W_{ii}}\right)^2 \quad \text{and} \quad \mathrm{Cov}\,(a_i, a_j) = \sum_{k=0}^{n}\frac{V_{ik}V_{kj}}{W_{ii}^2} \tag{100}$$

Non-linear least square : The problem in non-linear least square is similar to linear least square (note we are not talking about χ^2) *i.e.* fitting to an N data set (x_i, y_i) but with non-linear model / function having n parameters. The route remains the same, minimize the squared differences,

$$S = \frac{1}{2}\sum_{i=1}^{N}\Big(y_i - f(x_i, \{a_k\})\Big)^2 \equiv \sum_{i=1}^{N} r_i^2 \quad \text{where} \;\; k = 1, 2, \ldots n \tag{101}$$

The r_i is called the *residuals* and the parameters are $a \equiv \{a_k\}$. Obviously, the minimum of S occurs when its change due to change in a_k is zero,

$$\begin{aligned} a_k^{\alpha+1} &\approx a_k^\alpha + \Delta_k & (102)\\ \frac{\partial S}{\partial \Delta_k} &= 0 & (103)\end{aligned}$$

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

$$S(a_k^{\alpha+1}) = S(a_k^\alpha + \Delta_k) \approx S(a_k^\alpha) + \Delta_k \frac{\partial S}{\partial a_k^\alpha} + \frac{\Delta_k \Delta_l}{2} \frac{\partial^2 S}{\partial a_l^\alpha \partial a_k^\alpha} + \cdots \quad (104)$$

The various derivatives of S with respect to a_k are calculated as (omitting the sum_i before each term),

$$\begin{aligned} \frac{\partial S}{\partial a_k} &= \Big(y_i - f(x_i, a_k)\Big) \frac{\partial (y_i - f(x_i, a_k))}{\partial a_k} = -r_i \frac{\partial r_i}{\partial a_k} = -r_i\, J_{ik} & (105)\\ \frac{\partial^2 S}{\partial a_l \partial a_k} &= = -\frac{\partial r_i}{\partial a_l}\frac{\partial r_i}{\partial a_k} - \frac{\partial^2 r_i}{\partial a_k \partial a_l} \approx -\frac{\partial r_i}{\partial a_l}\frac{\partial r_i}{\partial a_k} = -J_{ik}\, J_{il} = H_{kl} & (106)\end{aligned}$$

where J_{ik}'s are the Jacobians. In the above definition of Jacobian we have dropped the index for iteration step α. So the Jacobians change from one iteration to another. The H_{kl} is called the *Hessian* of S. The iterate $a_k^{\alpha+1}$ is defined so as to minimize S with respect to Δ. Thus from (103) and (104),

$$\frac{\partial S}{\partial \Delta_k} = 0 = -r_i\, J_{ik} - \Delta_l\, J_{ik} J_{il} \quad \rightarrow \quad \Delta_l = -\left(J_{ik} J_{il}\right)^{-1} J_{ik} i\, r_i \quad (107)$$

where we have used $\partial S(a_k^\alpha)/\partial \Delta_k = 0$. Hence the update equation (102) and its corresponding matrix equation become,

$$\begin{aligned} a_k^{\alpha+1} &= a_k^\alpha - \left(J_{ik} J_{il}\right)^{-1} J_{il}\, r_i \\ \mathbf{a}^{\alpha+1} &= \mathbf{a}^\alpha - \Big(\mathbf{J}^T \mathbf{J}\Big)^{-1} \mathbf{J}^T \mathbf{r} & (108)\end{aligned}$$

This is the Gauss-Newton method. In general, Gauss-Newton method does not converge quadratically but does so as we approach a minimum. However, the convergence is not guarranted but usually it does. An advantage of this method is, one does not need to compute the second order partial derivatives. One problem of Gauss-Newton is, since the convergence is not guarranted, the residue may not get reduced upon iterations. This is where Levenberg-Marquardt algorithm comes. In the Levenber-Marquardt method, the Δ in (107) is replaced with

$$\Delta = -\Big(\mathbf{J}^T \mathbf{J}\Big) + \lambda\, \mathbf{D})^{-1} \mathbf{J}^T \mathbf{r} \quad (109)$$

where $\mathbf{D}$ is a positive diagonal matrix and λ is the so-called Marquardt paramters. For details, please see Numerical Recipes.

References

[1] `https://young.physics.ucsc.edu/jackboot.pdf`