四、數據誤差處理

實驗過程中或是撰寫結報時，同學常常遇到幾個問題：問題(1)不知該如何正確的分析與處理手邊的數據；問題(2)為什麼要學習誤差的處理；問題(3)誤差處理對科學研究的意義為何。

關於問題(1)，我們在以下的文章內，會做詳細的介紹。而問題(2)，答案是我們希望培養同學們用科學方法處理數據。至於問題(3)，簡單的說，我們希望以世界通用的數據處理方法，達到科學交流上正確且有效率的溝通。希望以下介紹，能對同學有所幫助。

何謂誤差？

$$\boxed{\text{誤差} = \text{測量值} - \text{真值}} \qquad (1)$$

上面的式子，是實驗數據處理會談到的誤差的定義。先想一想，為什麼我們要從事測量？如果我已經知道待測物理量的真值，我為什麼還要去測它？難道就為了要知道測量的誤差嗎？我們先定義三個名詞：測量值、理論值和真值。

- ✓ 「測量值」就是我們做實驗去測量到的物理量。
- ✓ 「理論值」就是我們依照既有的物理理論模型及公式推導歸納出來的物理量，是一個推論出來的真值。
- ✓ 「真值」就是真正的物理量，也有人廣義的解釋為我們做無限多次實驗，測量到的平均物理量。

那麼誤差和以上介紹的三個名詞，又有什麼關係呢？實驗數據的處理與分析，是想運用統計的方法，讓我們從多次的測量數據中，估算出最接近真值的數據，真值是我們想要獲得的測量結果。但是我們實驗做不到無限多次的測量，對吧？所以有時候可以用理論值代替真值。我們也需要藉由誤差分析，讓我們瞭解我們所做量測與理論模型估算之間，有多大的差距，並探討實驗誤差的可能來源。請找一個舒適的閱讀環境，靜下心來，閱讀以下的文章。

(1). 「系統誤差(systematic error)」與「隨機誤差(random error)」

系統誤差：

所謂測量，就是拿一個標準的測量工具，在此測量工具（例如尺）上含有刻度。將測量工具和待測物相互比較，我們可判得測量值。

- ■ 直接測量的物理量，可能的系統誤差來源為：
  - ◆ 測量工具本身所顯示的刻度，因為校正時疏忽，造成不正確。
  - ◆ 因為環境的因素（例如溫度、壓力等），使得刻度的數值產生變化。
  - ◆ 人為不正確的操作或使用錯誤的觀測方法。
- ■ 非直接測量的物理量：
  - ◆ 有可能因為實驗設計錯誤，或實驗設計不滿足理論原理的要求，這種情況也會造成系統誤差，不過這種情況常被很多人忽略。

通常系統誤差會使得所有測量值，都有過高或過低的偏差，且偏差量大致相同，不含機率分佈的因素。

隨機誤差：

實驗的基本方法，總是希望控制所有影響的變因，且一次只讓一種變因發生變化。為了實驗簡便，往往忽略對實驗影響較微小的因素，但實際操作時，不見得盡如人意。這些不易控制（有時候無法控制）的小變因，便會使測量值產生隨機分佈的誤差。

(2).降低系統誤差的方法：

- ✓ 儀器造成的→設法改良儀器。
- ✓ 環境造成的→設法控制實驗環境。
- ✓ 操作不良的→加強訓練自己。

理論上或許可能將儀器誤差完全消除，但是前兩項的改善，並不需要做到最完美的情形。因為不是儀器越精良、環境越穩定，實驗結果就越好。要減少系統誤差，我們必須考慮測量值所要求的「精密度」、實驗環境與經費。所以改善時，應該考慮主要誤差的來源。如果把所有經費都拿去買最精密的儀器，且假設儀器本身的測量精準度是 0.01%，而實驗室環境雖然已經改善至最好，但是在此環境的影響下，我們無論如何也只能使誤差達到 1%，則再精密的儀器，也改善不了我們的總誤差，那麼，買高精密的儀器，不過是花冤枉錢罷了！

(3).降低隨機誤差的方法：

藉由統計的方法，增加測量次數，能最有效率的改善隨機誤差。以下介紹兩個名詞。

- ■ 精密度：當多次重複測量時，不同測量值彼此間偏差量的大小。如果多次測量時，彼此間結果皆很接近，則稱為精密度較高。
- ■ 準確度：測量值與真值（或公認值）的偏差程度。公認值通常由廠商提供，那是使用已知較準確且精密度高的實驗儀器，在優良訓練的實驗人員重複操作下，所得出精密度相當高的實驗結果。但實驗時不見得有所謂公認值存在。

(4).統計分析方法

- ■ 母分佈：

  每一個待測物理量，我們可以假想存在一個「真值」。假設只有隨機誤差而完全沒有系統誤差，那麼我們增加同一物理量的測量次數，使隨機誤差大於真值與小於真值的機率分佈一樣，則所有測量值的平均值，將隨著測量次數增加而越接近真值。當測量次數等於「無窮多次」時，測量值的分佈稱為母分佈。而「有限次」的測量屬於母分佈的部份樣本，稱為「樣本分佈」。於是有限次數的算數平均值是我們對於真值所能給（猜）的最好的估計值。

- ■ 算數平均值(mean) $\bar{x}$ ：

$$\bar{x} \equiv \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (2)$$

■ 偏差（deviation)：

每一個數據與平均值的差值，稱為偏差。

$$d_1 = x_1 - \bar{x}, d_2 = x_2 - \bar{x}, \cdots, d_n = x_n - \bar{x} \quad (3)$$

偏差值有正有負，且所有偏差值的總和必為零。$d_1 + d_2 + \cdots + d_n = \sum_{i=1}^{n} d_i = \sum x_i - n\bar{x} = 0$ (4)

為了想量化實驗數據的精密度，且解決偏差值總和必為零的情形。我們可以將偏差值平方後相加，而定義出

■ 偏差平方的平均值（variance），或稱為變異數：

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (5)$$

當然將偏差值取絕對值後相加，也可以顯示實驗的精密度，但是數學計算上採用變異數，比較方便。

$$\begin{aligned} \sigma^2 &= \frac{1}{n}\sum (x_i - \bar{x})^2 \\ &= \frac{1}{n}(\sum x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2) \\ &= \frac{1}{n}(\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2) \\ &= \frac{1}{n}\sum x_i^2 - \bar{x}^2 \end{aligned} \quad (6)$$

變異數在計算時，可簡化為「平方的平均值減去平均值的平方」。比直接用公式計算，簡單多了。

■ 標準偏差：

對於母分佈而言（n→∞）時，取偏差平方的平均值的平方根（與測量值相同單位）定義母分佈的標準偏差（代表實驗數據分佈的精密度）

$$\sigma_n = \sqrt{\frac{d_1^2 + d_2^2 + \cdots + d_n^2}{n}} = \sqrt{\frac{\sum d_i^2}{n}} \quad (7)$$

σ_n 稱為「方均根」。方均根英文為 root（根）mean（均）square（方）。

如果直接利用上面的定義來處理有限次數的測量數據時，會發生矛盾的情形。例如：對於某一物理待測量，只有測量一個數據，則平均值等於唯一測量值，因此偏差為零。當然偏差的方均根值必為零。也就是有最良好的精密度。那豈不是所有測量皆測一次就夠了？

問題出在哪兒呢？因為計算 n 個數據的個別偏差時，需先計算平均值。當有平均值時，只要有 n-1 個數據便可以算出所有的偏差量。也就是計算偏差平方的平均值時，數據中的獨立變數僅有 n-1 個，因此計算平均值時，分母若改為 n-1 較為合理。因此樣本分佈（有限次數）數據的**標準差**(Standard Deviation)定義為

$$\sigma = \sigma_{n-1} = \sqrt{\frac{\sum d_i^2}{n-1}} \qquad (8)$$

如此一來只測量一次時，上式中分子分母皆為零，也就是無法確定標準差。當（n→∞）時，則分母為 n 或 n-1 已經沒有差別了。工程用計算機上有 σ_n 與 σ_{n-1} 差別便在於分母。以上定義的標準差，代表所有測量數據與平均值之間平均的偏差量（也就是每一測量數據的精密度的平均值）。

可是通常我們也關心所計算出平均值的可信度是多少？也就是實驗結果的精密度有多高？平均值的精密度應該要高於個別測量數據的精密度。我們先寫下依據統計理論所得出的結果。

算數平均值 $\bar{x}$ 的統計標準差（standard error of the mean）

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sum d_i^2}{n(n-1)}} \qquad (9)$$

多次實驗測量結果寫為

$$\bar{x} \pm \sigma_{\bar{x}} \qquad (10)$$

也就是測量平均值加上所對應的統計標準差（俗稱測量之不準度：uncertainty）。

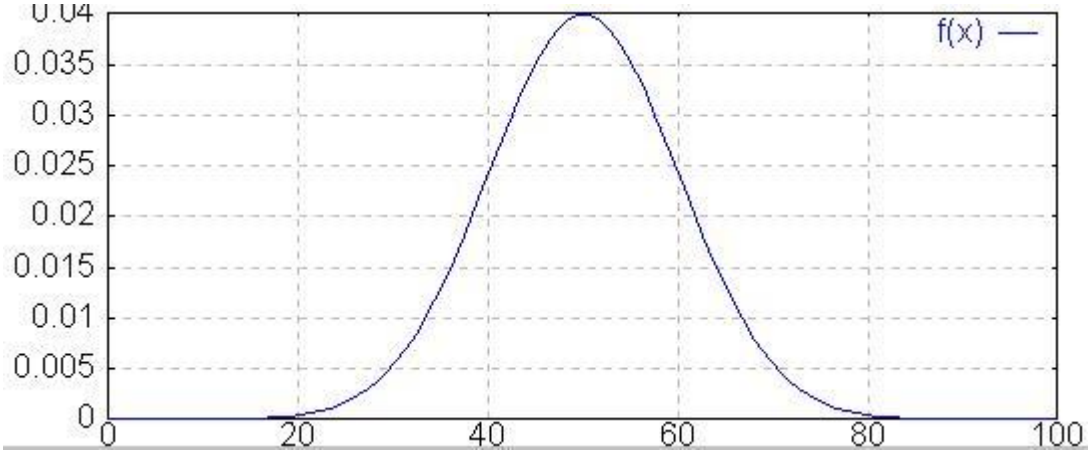
（請注意：實驗結果不見得一定都是平均值。）

■ 標準偏差所代表的意義與運用：

通常當測量次數多時，測量數據的隨機分佈滿足「常態分佈 (normal distribution)」或稱「高斯分布(gaussian distribution)」：

$$P = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x-\bar{x})^2}{2\sigma^2}] \qquad (11)$$

P 是測量值為 x 的機率。(次數少時為二項式分佈）。如下圖，平均值為 50，標準差為 10 的常態分佈。



測量值出現在

$\bar{x}+\sigma > x > \bar{x}-\sigma$　　範圍內的機率為 68.3%。

$\bar{x}+2\sigma > x > \bar{x}-2\sigma$　　範圍內的機率為 95.4%。　　（12）

$\bar{x}+3\sigma > x > \bar{x}-3\sigma$　　範圍內的機率為 99.7%。

$\bar{x}+4\sigma > x > \bar{x}-4\sigma$　　範圍內的機率為 99.994%。

做多次測量時，有時候某些數據與平均值相差較多，若懷疑是因為測量時不小心的觀測錯誤，怎樣判斷該不該捨去那些數據呢？例如：測量某物體長度 100 次，計算出平均值與標準差後，發現有 3 組數據落在 3 倍標準差外，4 組落在 2 倍與 3 倍之間，其餘皆在平均值與標準差之間。

若採用常態分佈來看這 100 次的測量結果，由於數據落在 2 倍標準差外的機率有 4.6%。因此那四組數據的出現是符合常態分佈的。但是數據落在 3 倍標準差外的機率應小於千分之三，所以那 3 組落在 3 倍標準差外的數據，通常是測量錯誤造成的，可以捨去並重新計算剩餘數據的平均值與標準差，得出所要的測量結果。

■　平均值的標準差的意義：

每次(組）的多次實驗所得平均值都不會相同。這些平均值也會形成一種分佈。平均值的標準差便是代表這些不同的平均值的可能差異性（精密度）。綜合說來，實驗數據的標準差(standard deviation)顯示單一個測量值與平均值間可能偏差的程度。重複（增加實驗次數）並不會減少其數值。（單一測量的精密度）

平均值的標準差(standard error of the mean)顯示所得平均值的可重覆性程度（結果的精密度）。如果多組重覆測量所計算出平均值的標準差。其數值可以藉由增加測量次數而減少，與$\sqrt{n}$成反比。因此 10000 次測量平均值的標準差為 100 次測量的 1/10。為了增加一位有效位數，測量次數必須由 100 增加到 10000。

■　誤差傳遞

經常一個物理量是經由測量數個物理量，再藉由之間的關係式計算而得出。例如：動量是由測量質量與速度相乘而得（速度又由測量位移與時間而得）。當測量時，質量、位移與時間的個別誤差將影響最後結果的誤差。假設 X 代表某一個物理量，由$u,v,\cdots$等測量值所決定。即$X = f(u,v,\cdots)$，而以$\bar{u},\bar{v},\cdots$分別代表$u,v,\cdots$等分量樣本分佈的平均值。則平均值$\bar{X} = f(\bar{u},\bar{v},\cdots)$。

對於某一組測量樣本數據，可以表示為$X_i = f(u_i,v_i,\cdots)$則

$$
\begin{aligned}
X_i - \bar{X} &= f(u,v,\cdots) - f(\bar{u},\bar{v},\cdots) \\
&= [f(\bar{u},\bar{v},\cdots) + (\frac{\partial f}{\partial u})_{\bar{u}}(u_i-\bar{u}) + (\frac{\partial f}{\partial v})_{\bar{v}}(v_i-\bar{v}) + \cdots] - f(\bar{u},\bar{v},\cdots) \\
&= (\frac{\partial f}{\partial u})_{\bar{u}}(u_i-\bar{u}) + (\frac{\partial f}{\partial v})_{\bar{v}}(v_i-\bar{v}) + \cdots \\
&= (\frac{\partial X}{\partial u})_{\bar{u}}(u_i-\bar{u}) + (\frac{\partial X}{\partial v})_{\bar{v}}(v_i-\bar{v}) + \cdots
\end{aligned}
\qquad (13)
$$

算數平均值的標準差

$$\sigma_X^2 = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$= \lim_{n\to\infty}\frac{1}{n}\sum[(\frac{\partial X}{\partial u})_{\bar{u}}(u_i - \bar{u}) + (\frac{\partial X}{\partial v})_{\bar{v}}(v_i - \bar{v}) + \cdots]^2$$

$$= \lim_{n\to\infty}\frac{1}{n}\sum[(\frac{\partial X}{\partial u})_{\bar{u}}^2(u_i - \bar{u})^2 + (\frac{\partial X}{\partial v})_{\bar{v}}^2(v_i - \bar{v})^2 + 2(\frac{\partial X}{\partial u})_{\bar{u}}(\frac{\partial X}{\partial v})_{\bar{v}}(u_i - \bar{u})(v_i - \bar{v})\cdots]$$

$$= \sigma_u^2(\frac{\partial X}{\partial u})_{\bar{u}}^2 + \sigma_v^2(\frac{\partial X}{\partial v})_{\bar{v}}^2 + 2\sigma_{uv}^2(\frac{\partial X}{\partial u})_{\bar{u}}(\frac{\partial X}{\partial v})_{\bar{v}} + \cdots \qquad (14)$$

其中

$$\sigma_u^2 = \lim_{n\to\infty}\frac{1}{n}\sum_i(u_i - \bar{u})^2,$$

$$\sigma_v^2 = \lim_{n\to\infty}\frac{1}{n}\sum_i(v_i - \bar{v})^2,$$

$$\sigma_{uv}^2 = \lim_{n\to\infty}\frac{1}{n}\sum_i(u_i - \bar{u})(v_i - \bar{v}), \qquad \text{稱為協方差(covariance)} \qquad (15)$$

如果 u 和 v（測量物理量）彼此不相關，則協方差為零。

通常測量時的個別參數間是互不相干的，於是方差可以簡化為

$$\sigma_X^2 \approx \sigma_u^2(\frac{\partial X}{\partial u})_{\bar{u}}^2 + \sigma_v^2(\frac{\partial X}{\partial v})_{\bar{v}}^2 + \cdots \qquad (16)$$

當測量物體密度時，質量與體積的測量通常不相干，因此可用上式計算質量與體積的誤差所造成密度測量的誤差。但是體積測量誤差的計算，若體積是由長、寬、高等測量值相乘而得。當長、寬、高都是用同一測量工具且同樣方式測量時，往往彼此間的誤差是相關的。尤其當測量工具的系統誤差大於隨機誤差時，例如校正失誤所造成誤差將造成長、寬、高的系統誤差。則體積的百分誤差將直接等於長、寬、高百分誤差之和。（而非長、寬、高百分誤差平方之和開根號）。當使用誤差傳遞時要辨別測量值間是否彼此相關。

讓我們運用上式計算算數平均值的標準差。

$$\bar{X} = \bar{X}(X_1, X_2, \cdots X_i, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^{n}X_i$$

平均值是由各測量值取平均而得到（視為以各測量值為獨立變數的函數）。

$$\sigma_{\bar{X}}^2 = \sum_{i=1}^{n}[\sigma_i^2(\frac{\partial \bar{X}}{\partial X_i})^2] \qquad (17)$$

$$\frac{\partial \bar{X}}{\partial X_i} = \frac{\partial}{\partial X_i}[\frac{1}{n}\sum_{i=1}^{n}X_i] = \frac{1}{n} \qquad (18)$$

$$\sigma_{\bar{X}}^2 = \sum_{i=1}^{n}(\sigma_i^2\frac{1}{n^2}) \qquad (19)$$

若各測量值的標準差皆相同時，$\sigma_i = \sigma$，則上式可以簡化為

$$\sigma_{\bar{X}}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \qquad (20)$$

於是平均值的標準差

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

讓我們做幾個例題：

(1) $X = au \pm bv$

$$\frac{\partial X}{\partial u} = a$$

$$\frac{\partial X}{\partial v} = \pm b$$

$$\sigma_X^2 = a^2\sigma_u^2 + b^2\sigma_v^2 \pm 2ab\sigma_{uv}^2$$

例如：(3.1257 ± 0.0138) - (1.892 ± 0.0095)

= (3.1257 - 1.892) ± $(0.0138^2 + 0.0095^2)^{1/2}$

= 1.234 ± 0.017

注意：誤差並非 0.0138 + 0.0095 。為什麼呢？

3.1257 ± 0.0138 表示測量值在 3.1257-0.0138 與 3.1257+0.0138 之間，多次測量時應該越接近 3.1257 的數值越多，離開越遠的機率越少（滿足常態分佈）。因為隨機分佈的關係，大於平均與小於平均的機率皆相等。當兩測量值相加時，兩者偏差皆為最大正偏差或皆為最大負偏差的機率，應該很小，經統計分析以平方相加開根號為較適當

(2) $X = \pm auv$

$$\frac{\partial X}{\partial u} = \pm av$$

$$\frac{\partial X}{\partial v} = \pm au$$

$$\sigma_X^2 = a^2v^2\sigma_u^2 + a^2u^2\sigma_v^2 \pm 2a^2uv\sigma_{uv}^2$$

$$\frac{\sigma_X^2}{X^2} = \frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2} + 2\frac{\sigma_{uv}^2}{uv}$$

若協方差為零時，則結果的百分誤差的平方等於個別參數的百分誤差的平方和。參數間為相除的情形時，也有相同結果，請你自己試一試。以上皆討論獨立變數間的誤差皆互不相干，彼此不受影響。

若是討論包含系統誤差的情形，或是變數間相互影響時，就必須考慮協方差。例如：體積是由三個測量值長、寬、高相乘而得，假使測量的尺因為溫度的變化而收縮。使用同一把尺測量時，則長、寬、高誤差皆會有相同趨勢（同時過大或過小）。則百分誤差不再是平方後相加再開根號，而是直接相加。

■ 有效位數的說明：

當使用測量工具從事測量時，工具的最小刻度限制了測量值的有效位數。通常我們以儀器最小能讀到的刻度值外加一位估計值作為記錄的結果。但是由於科技的進步，現代很多儀表顯示時都已經數位化（直接顯示數值），在正常的情形下，最後一位顯示的數值，已經包含了儀器幫你估計的成分。

但是：並非數位化的儀器所顯示的數值，完全都是必須記錄的。儀器顯示的最小刻度值，應該要配合儀器的精密度。但是儀器商生產不同精密度的儀器時，為了成本問題很可能使用相同的顯示元件。因此某些儀器顯示的數值，可能多於實際的精密度。另外一種情形是，儀器也的確夠精密，但是你所測量的環境本身造成的影響，超過儀器精密度的範圍。

例如：使用 6 位數的精密電表去量電阻。結果數值後幾位連續不斷的跳動。（也就是選用太過精密的儀器）多記了後面一直變動的數值，是沒有用的。（這也是一般學生常犯的毛病，所有數值皆記下來）

基本原則：

實驗紀錄所顯示的最小刻度值，也應該要配合測量的精密度。否則只是增加自己計算的負擔和增加記錄的負擔。

數據處理時：

反正用計算機在計算，可能計算完畢，還多了好多位有效位數呢！用 10 位顯示的計算機，實驗結果變成 10 位有效位數。如果用 12 位顯示的計算機，實驗結果變成 12 位有效位數。好像實驗的精密度取決於計算機的功能！？？？這不是笑話！這是現代很多學生的毛病，且已經變成一種壞習慣，請彼此互相提醒，不要犯這種錯誤。

舉一個實例：如下表（ n=5 ）

測量序號	長度 L（cm）	寬度 W（cm）
1	10.78	8.21
2	10.80	8.20
3	10.75	8.22
4	10.73	8.21
5	10.78	8.22

	長度	寬度
平均值 $\bar{x} \equiv \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$	10.768→10.77	8.212→8.21→8.212
標準差 $\sigma = \sigma_{n-1} = \sqrt{\frac{\sum d_i^2}{n-1}}$	0.0278→0.03 ±0.03	0.008→0.01→0.008 ±0.01→±0.008
平均值的統計標準差 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sum d_i^2}{n(n-1)}}$	0.0124→0.01 ±0.01	0.0038→0.004 ±0.004
結果 $\bar{x} \pm \sigma_{\bar{x}}$	10.77±0.01	8.21±0.004→8.212±0.004

從以上的例子，可以看出記錄值的有效位數都是小數點以下兩位，因為尺的最小刻度為公釐，加上一位估計值。則算出來的平均值、標準差、平均值的標準差，也都該是兩位。先算出小數點以下三位，再利用四捨、五入的原則，得到小數點以下兩位的結果。

實驗量測使用的尺的最小刻度為公厘，測量時我們最多只能擁有一位估計值，也就是取到小數點以下第二位，經過計算後，結果記錄為 8.21±0.004 cm，但是用同一把尺量測到的算術平均數和平均值的統計標準差有效數字怎麼會不同？問題出在哪裡呢？如何決定平均值的有效數字呢？

最正確的方法是算出數據的平均值的統計標準差，並藉此決定平均值的有效數字應該取到哪一位。**如寬度的量測中；首先取平均值為 8.21 cm，算出平均值的統計標準差為 0.004 cm。但因平均值的統計標準差在小數點第三位，所以平均值的有效數字也應該在小數點第三位。重新取平均值為 8.212 cm，算出平均值的統計標準差為 0.004 cm。最後結果為 8.212±0.004 cm。**

如果最後的結果是利用好幾層的關係式，計算而得到的，是否每計算一次就要將數據取至適當的有效位數，再繼續算下去。還是，用計算機一直算，最後再取有效位數呢？原則是這樣的：

當數據計算時，運算的數字來源是由於數學推導的常數或物理常數，則最後再取有效位數便可。（視常數完全有效）但是若遇到測量值，則必須運算完後，馬上取至適當的有效位數。例如：面積等於長乘寬，算出後馬上要決定適當的有效位數，再繼續運算下去。做加，減，乘，除等運算時，有效位數以最不準確的因子的有效位數為基準。

(5).補充說明：

■ 有限次數的平均值是我們對於真值所能給（猜）的最好的估計值。由於偏差平方的平均值代表著數據的偏差量，對於一組數據而言，此偏差量越小越好。問題改成：採用怎樣的平均值計算方式會有較小的偏差平方的平均值？取偏差平方的平均值對平均值（偏）微分等於零的結果如下：

$$\delta = \sigma^2 = \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - 2\bar{x}\sum_i x_i + n\bar{x}^2$$

$$\frac{\partial \delta}{\partial \bar{x}} = \frac{\partial}{\partial \bar{x}}(\sum_i (x_i - \bar{x})^2) = -2\sum_i x_i + 2n\bar{x} = 0 \qquad (21)$$

$$\bar{x} = \frac{\sum_i x_i}{n}$$

所以採用算數平均值的計算方式時，偏差平方的平均值有最小值。

■ 最小平方作圖法：

實驗時，我們常會需要測量某物理量（應變數）隨物理參數（自變數）變化時，彼此間的關係。例如：電阻（縱軸）隨溫度（橫軸）的變化。最小平方曲線作圖法便是在所繪出數據圖中（電阻--溫度圖），描繪出一條曲線，使所有數據點到曲線距離平方總和(偏差平方的平均值)為最小。用 $f(x_i, y_i)$ 表示數據點，我們希望找出 $y = f(x)$（最小偏差平方的平均值曲線），使得 $\delta = \sum (f(x_i) - y_i)^2$ 有最小值。以上假設自變量沒有誤差（或相對很小）：

以下我們以常見的線性關係 $f(x) = ax + b$ 為例，希望找出 a, b 使得 $\delta = \sum (f(x_i) - y_i)^2 = \sum (ax_i + b - y_i)^2$ 有極小值。也就是找出最能代表測量數據線性關係的直線。欲使偏差平方的平均值有最小值，

$$\frac{\partial \delta}{\partial a} = 0$$
$$\Rightarrow \sum (ax_i + b - y_i)x_i = a\sum x_i^2 + b\sum x_i - \sum x_i y = 0 \qquad (22)$$
$$\frac{\partial \delta}{\partial b} = 0$$
$$\Rightarrow \sum (ax_i + b - y_i) = a\sum x_i + nb - \sum y_i = 0$$

聯立解上兩個方程式，可得到

$$a = \frac{\sum x_i \sum y_i - n\sum x_i y_i}{(\sum x_i)^2 - n\sum x_i^2}$$
$$b = \frac{\sum x_i y_i \sum x_i - \sum y_i \sum x_i^2}{(\sum x_i)^2 - n\sum x_i^2} \qquad (23)$$

上式中，a 為直線斜率，b 為其截距。

經常所測量物理量之間的關係式並非如 $f(x) = ax + b$ 如此簡單的關係，可以仿造上面計算最小方差的方式，找出各係數的值。但是大多數情況，皆可以利用變數變換的方式，將關係式轉換成簡單線性關係。

例如：電容放電時，電容電壓隨時間變化的關係，$V_0(t) = V_0 \cdot e^{-t/RC}$。實驗時測得電壓 V 隨時間 t 變化的數值，欲求得 V_0 以及放電時間 RC 值。可將所測得電壓取對數

$\ln V_0(t) = \ln V_0 - t/RC$

令 $y = V_0(t)$，$x = t$ 則有 $y = ax + b$ 的關係。利用上面最小平方法求得斜率 $a = -1/RC$，截距 $b = \ln V_0$。

接下來的問題是：

1. 這樣計算出來的直線，用來代表原有數據的關係好不好呢？

提示：偏差平方的平均值 $\sigma^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ax_i - b)^2$ 越小越好。[除以 n-2 是個關鍵，因為兩個點決定一條直線，所以至少要有三個點以上，偏差平方的平均值才有意義。]

2. 所計算出來的直線斜率 a 和截距 b 的誤差又是多少呢？

提示：利用誤差傳遞的計算法去計算。將 a，b 視為 x_i 以及 y_i 的函數，但是上面的計算中皆假設 x_i 沒有誤差。因此只需要計算由於 y_i 的誤差所傳遞給 a，b 係數的誤差。

令 $\Delta = (\sum x_i)^2 - n\sum x_i^2$，　　（24）

則 $\frac{\partial a}{\partial y_i} = \frac{1}{\Delta}(\sum x_i - nx_i)$ 且 $\frac{\partial b}{\partial y_i} = \frac{1}{\Delta}(x_i\sum x_i - \sum x_i^2)$。於是得到　　（25）

$$\begin{aligned}
\sigma_a^2 &= \sum_{i=1}^{a}\frac{\sigma^2}{\Delta^2}[(\sum x_i)^2 - 2nx_i\sum x_i + n^2x_i^2] \\
&= \frac{\sigma^2}{\Delta^2}[n(\sum x_a)^2 - 2n(\sum x_i)^2 + n^2\sum x_i^2] \\
&= \frac{n\sigma^2}{\Delta^2}[n\sum x_i^2 - (\sum x_a)^2] \\
&= -\frac{n\sigma^2}{\Delta}
\end{aligned} \qquad (26)$$

$$\begin{aligned}
\sigma_b^2 &= \sum_{i=1}^{a}\frac{\sigma^2}{\Delta^2}[x_i^2(\sum x_i)^2 - 2x_i\sum x_i\sum x_i^2 + (\sum x_i^2)^2] \\
&= \frac{\sigma^2}{\Delta^2}[\sum x_i^2(\sum x_a)^2 - 2(\sum x_i)^2\sum x_i^2 + n(\sum x_i^2)^2] \\
&= \frac{\sigma^2}{\Delta^2}\sum x_i^2[n\sum x_i^2 - (\sum x_a)^2] \\
&= -\frac{\sigma^2}{\Delta}\sum x_i^2
\end{aligned} \qquad (27)$$

若是所有測量數據標準差相同 $\sigma_i = \sigma$，我們又可將原點平移（任選原點）使得 $\sum x_i = 0$。於是上面結果可以簡化為 $\sigma_a = \frac{\sigma}{\sqrt{\sum x_i^2}}$，$\sigma_b = \frac{\sigma}{\sqrt{n}}$。

對於任何數據我們皆可以代入上面最小平方法找出一條直線 $f(x) = ax + b$。

可是數據 x，y 之間，是否真的適合用線性關係描述呢？我們用這樣的想法來評斷：若兩者之間真的滿足 $y = ax + b$，則若是我們改用 $x = a'y + b'$ 去描述，應該也可以得到適當的曲線。理想情況應當滿足 $a = \frac{1}{a'}$。我們可以檢驗—用以上兩種直線方式所得出之斜率—其相乘積若越接近於 1，表示 x，y 間越相關。我們於是定義(linear-correlation coefficient)

$$\gamma \equiv \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \qquad (28)$$

若是 γ 值越接近於 1，則表示 x，y 數據間越適合用上述線性關係描述。

參考資料：

國立台灣師範大學物理系 普通物理實驗手冊