Corrections for Chapter Two

- **P35, line 8 from the bottom, "confidence level" should be "significant level".**
- **P35, line , adding ", called confidence level" after "1-α"**
- **P36, line below equation 2-13, "1-a" should be "(1- α)x100%"**

Chapter 2 Data Process and Analysis

数据处理与分析

Why Statistical Analysis of Analytical Results is Necessary and Important?

- **What is the maximum error can I tolerate?**
- **Lower? Identical? Higher?**

Alcohol in blood sample (%): 0.084, 0.089, 0.079 (critical value 0.090)

- **Is the new method can be adopted?**

Glucose in serum (mg/L)

Method	# 1	#2	#3	#4	#5	#6
Recognized Method	1044	720	845	800	957	650
New Method	1028	711	820	795	935	639

Objectives 目的

- Minimize errors减小误差
- Estimate the size of errors with acceptable accuracy 在合理的准确度要求下估计误差
- Estimate the “true” value估计真值
  - Define limits within which the true value of a measured quantity lies with a given level of probability可能性
    - What maximum error can be tolerated in the result?
    - How many determinations are necessary, then?
- Compared your results with the “known value” or standard material.与已知值或者标准物质比较实验结果

Terminologies

- **Replicates重复测定次数:**
  - **Samples of about the same size are carried through analysis in exactly the same way.**
- **Mean (average)平均值**
- **Outlier离群值:**
  - **A result that differs significantly from others in the same data set.**

Terminologies

Mean (average)样本平均值: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Median 中位数 $\tilde{x}$

Deviation偏差

绝对偏差$d_i = x_i - \bar{x}$

相对偏差$R_{d_i} = (d_i / \bar{x}) \times 100\%$

平均偏差$\bar{d} = \sum_{i=1}^{n} |d_i| \Big/ n$

相对平均偏差$R_{\bar{d}} = (\bar{d} / \bar{x}) \times 100\%$

Range极差 $R = x_{max} - x_{min}$

What is the advantage about Median?

Accuracy and Precision

Accuracy准确度

The closeness of a measured value to the "True" or accepted value. Denoted as Error which indicates systematic error of a method.

测定结果与"真值"接近的程度, 以误差表示

通常反映测定方法的系统误差

Absolute error绝对误差 $E_a = \bar{x} - T$

Relative error相对误差 $E_r = \frac{E_a}{T} \times 100\%$ ☺

Accuracy and Precision
准确度与精密度

Precision精密度

Degree of agreement between replicate measurements of the same quantity. It describes the reproducibility of the measurements. Denoted as random error.

平行测定的结果互相靠近的程度. 以偏差表示

- Standard Deviation标准偏差
- Coefficient of Variance 变异系数

Corrections

- **P29,** $$y = \phi(u)\cancel{du} = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \cancel{du} \qquad (2.4)$$

- **P42, line 2**
  - **From Table 2.5, $F_{crit}(n_2\text{-}1, n_1\text{-}1) = F_{0.05}(4,3) = 0.12$ 改为**
  - **From Table 2.5, $F_{crit}(n_2\text{-}1, n_1\text{-}1) = F_{0.05}(3,4) = 6.59$**

- **P49 表格下面第一行 删除 " and 99%"**

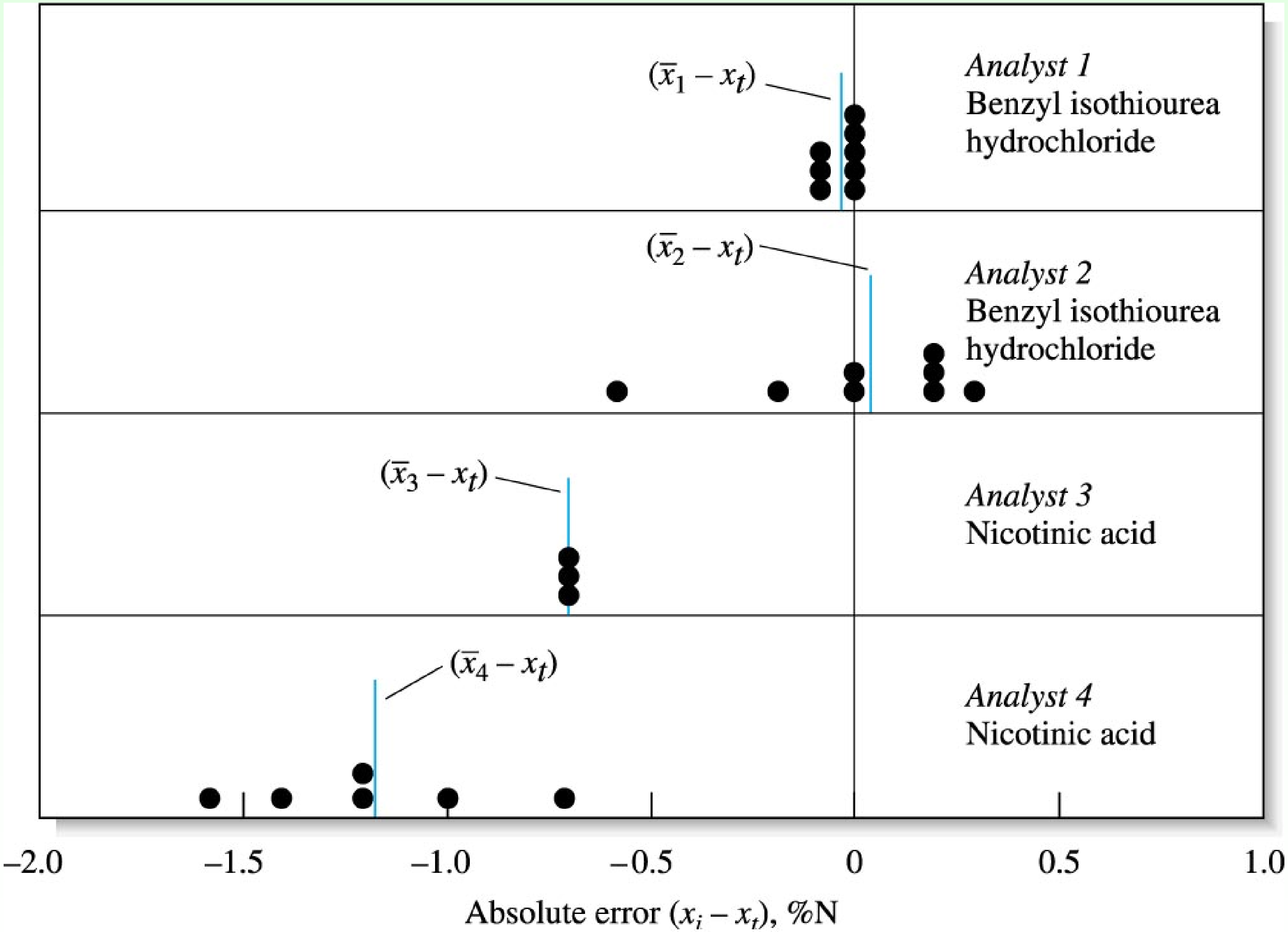
Homework

- Chapter 2: 2.5, 2.6

Homework

- Chapter 2: 2.7, 2.9~12

Accuracy versus Precision
准确度与精密度的关系



Skoog p94

Why Statistical Analysis of Analytical Results is Necessary and Important?

- **What is the maximum error can I tolerate?**
- **Lower? Identical? Higher?**

  Alcohol in blood sample (%): 0.084, 0.089, 0.079 (critical value 0.090)

- **Is the new method can be adopted?**

  Glucose in serum (mg/L)

Method	# 1	#2	#3	#4	#5	#6
Recognized Method	1044	720	845	800	957	650
New Method	1028	711	820	795	935	639

Classification of Errors, Source and Way to Eliminate or Reduce Errors

误差的分类、产生的原因及减免办法

- Systematic/Determinate Error系统误差

  Affect accuracy of results

- Random/Indeterminate Error随机误差

  Affect measurement precision

- Gross Error过失误差

  Lead to outliers 离群值

Systematic/Determinate Error系统误差

具方向性(bears a sign)和重现性，为可测误差

➢Method error方法误差

—Difficult to determine

—Comes from non-ideal chemical or physical behavior of the reagents(试剂) and reactions

—Slowness, incompleteness of reaction

—Solubility溶解度

—Instability of some species

—Non-specificity 非特异性of most reagents

—Side-reaction副反应

—Calibrating by other methods or Analysis of standard samples

Systematic Error

- **Instrument Errors仪器误差**

  —Uncalibrated Instrument 仪器未校准

  — Uncalibrated/ worn weights砝码磨损

  — Uncalibrated glassware刻度不准

  —Calibrating is the way to do it 校准 (绝对、相对)

- **Reagent Errors试剂不纯引起的误差**

  Running reagent blank

- **Personal Errors 个人（操作）误差**

  Prejudice or bias

  Color observation颜色观察

Gross (Error)过失(误差)
Personal Errors

- Lead to outliers 离群值
- **It is mistake!!!!!!!**

Random/Indeterminate Error随机误差

Uncertainty不确定性, accidental, can not be avoided

– Follows mathematical laws of probability

服从数学概率法则 (统计规律)

– Random errors in analytical results follow a Normal Distribution/Gaussian Curve

分析结果随机误差一般服从正态分布或 高斯分布

Karl Friedrich Gauss, b.1777

Studied math in University of Gottingen, 1795-1798

Terminologies

- **Population总体:**
  - **The collection of all measurements of interest to the experimenter (finite and real/hypothetical假定的 or conceptual 概念上的in nature).**
  - **All members of a system**
- **Sample样本:**
  - **A subset子集 of measurements selected from the population**
- **Population mean (μ)总体平均值 and sample mean样本平均值 $\overline{x}$**
- **Population standard deviation (σ) 总体标准偏差 sample standard deviation (s)样本标准偏差**

Statistical Treatment of Random Error

Population Standard Deviation

总体标准偏差： $\sigma = \sqrt{\dfrac{\sum (x_i - \mu)^2}{n}}$

Sample Standard Deviation

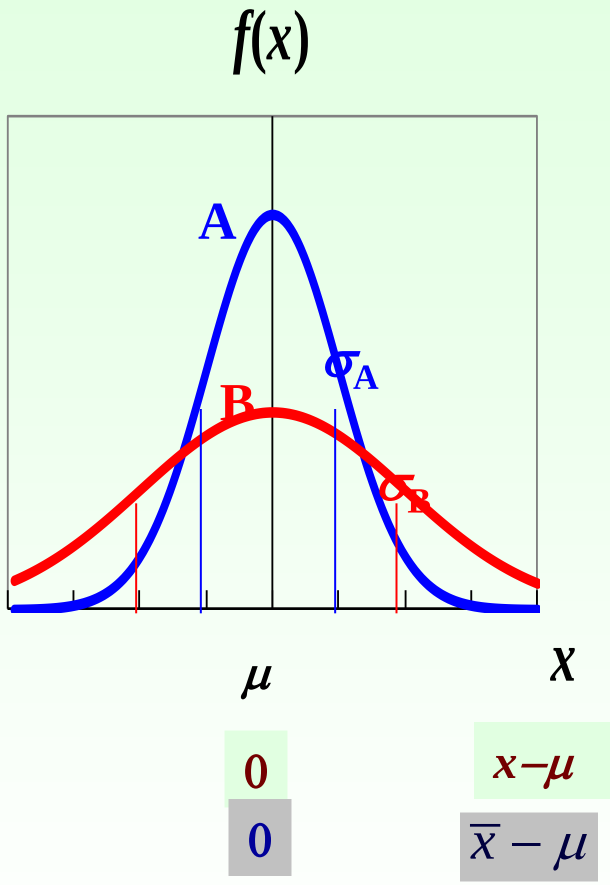
样本标准偏差： $s = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}}$

(n-1)为自由度，用f (degree of freedom)表示

Relative Standard Deviation相对标准差 $\boldsymbol{RSD} = \dfrac{s}{\bar{x}}$

Coefficient of variation变异系数， **CV=($s/\bar{x}$)×100%**

The Normal (Gaussian) Distribution正态分布曲线



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

***y*: relative frequency频率密度**

***x*: data测量值**

***μ*: population mean总体平均值**

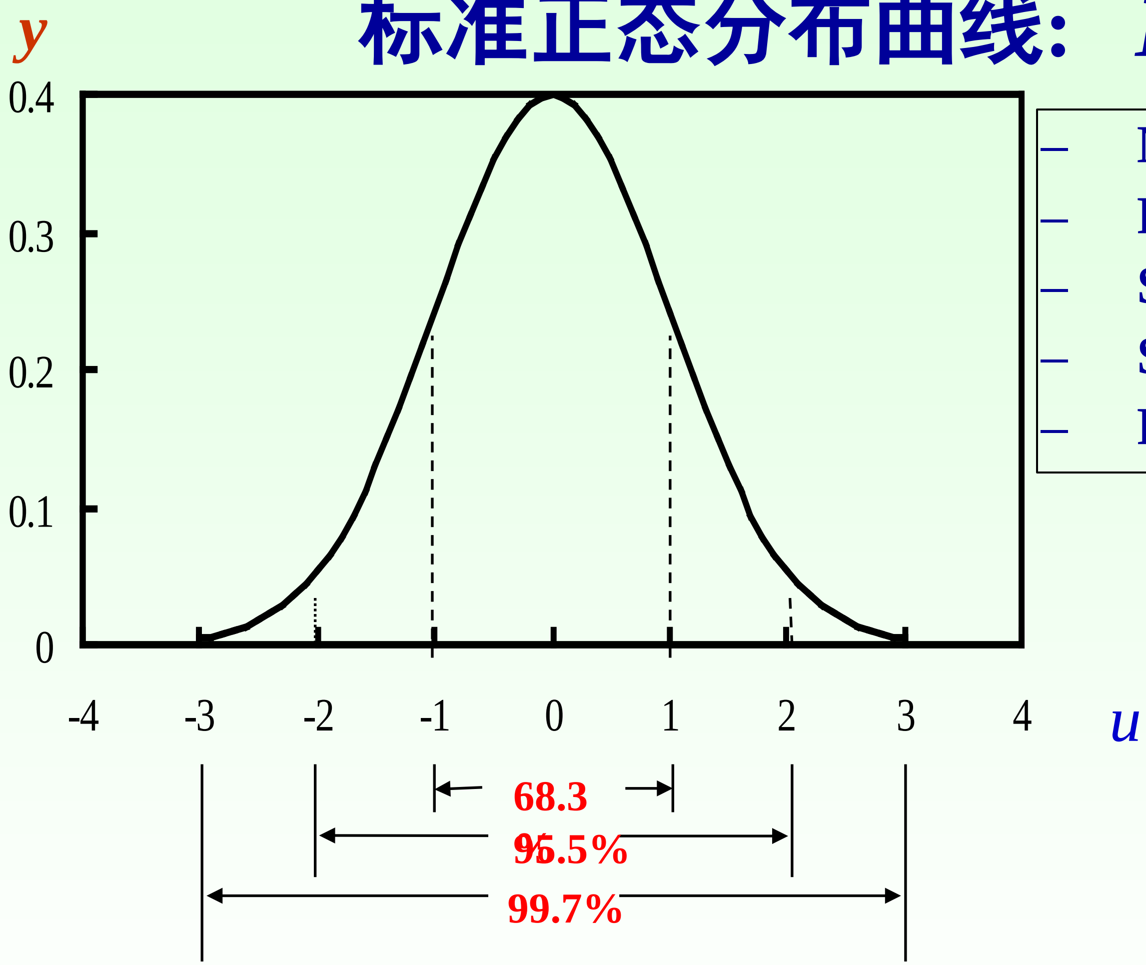
***x-μ*: deviation from mean随机误差**

***σ* : population standard deviation 总体标准差**

- **Maximum value at *x*=*μ*.**
- **Reflection point at *x*=*μ*±*σ*.**
- **Symmetrical according to *x*=*μ*.**
- **Profiles varies with σ**
- **Smaller error, bigger frequency**

Standard Normal (z) Distribution

标准正态分布曲线：N（0,1）



- Maximum value at $x=\mu$.
- Reflection point at $x=\mu \pm \sigma$.
- Symmetrical to $x=\mu$.
- Smaller error, bigger frequency
- Profile does not vary with σ

$$u = \frac{x-\mu}{\sigma}$$

***u*：the deviation of a result from the population Mean relative to the standard deviation.**

Normal Error Curve Properties(几点结论，记住)

- **The mean occurs at the central point of maximum frequency** 频率最大的点对应平均值
- **Small uncertainties are observed much more often than very large ones**

  小误差出现的概率大, 大误差出现的概率小
- **There is a symmetrical distribution of positive and negative deviations about the maximum**

  正、负误差出现的概率相等
- **Area under the curve gives the probability of a measured value occurring between those limits**

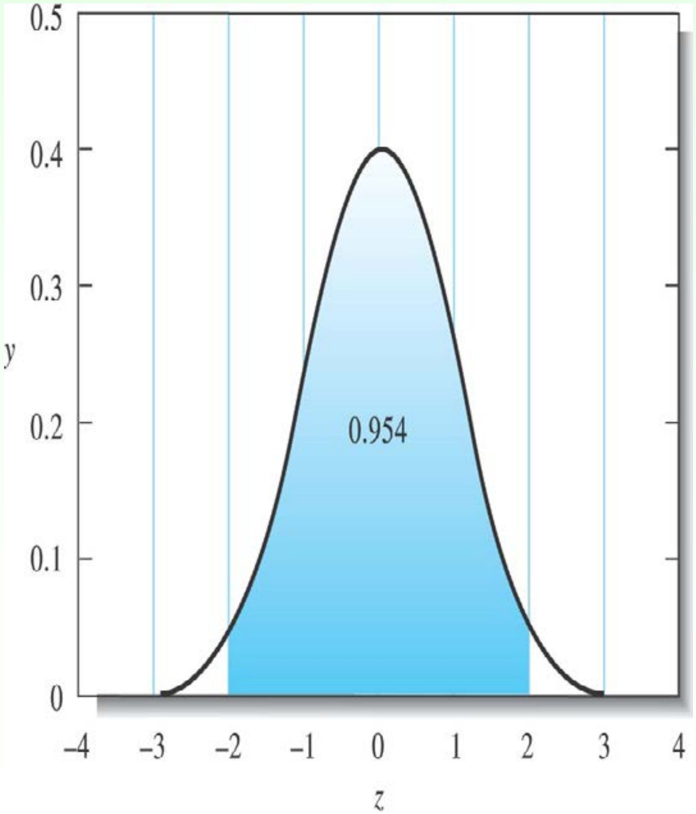
  曲线下的面积表示一个测定值出现在一定范围内的概率

Area under the Curve 曲线下面积

the probability of the population mean occurring between those limits

$$area = \frac{1}{\sqrt{2\pi}}\int_{-u}^{u} e^{\frac{-u^2}{2}}du$$

正态分布概率积分表



$\|u\|$	area	概率p
0.674	0.500	50.0%
1.000	0.683	68.3%
1.645	0.900	90.0%
1.960	0.950	95.0%
2.000	0.954	95.4%
2.576	0.990	99.0%
3.000	0.997	99.7%
∞	1.000	100.0%

Our Task Left

- Predict the true value by experimental data
- Compare experimental data with true value
- Compare two group experimental data
- Exclusion of Outlier离群值

The Confidence Interval (CI)
（总体均值的）置信区间

The true value falls within the confidence limit with a certain probability.

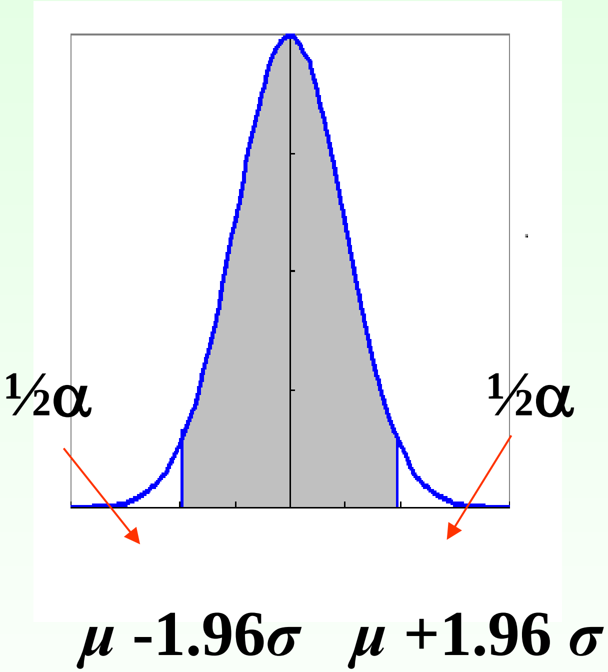
在一定的置信度下(把握性), 估计总体均值可能存在的区间, 称置信区间。

Confidence Interval (CI) & Confidence Level
置信区间与置信水平

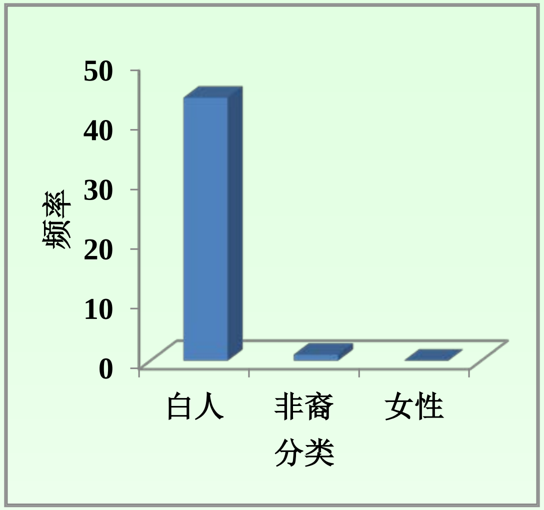
- **(x -1.96 σ, x +1.96 σ) has 95% probability containing the true population mean μ.**
- **The probability that the true mean lies within a confidence interval**
- **Significance Level-α value显著性水平**

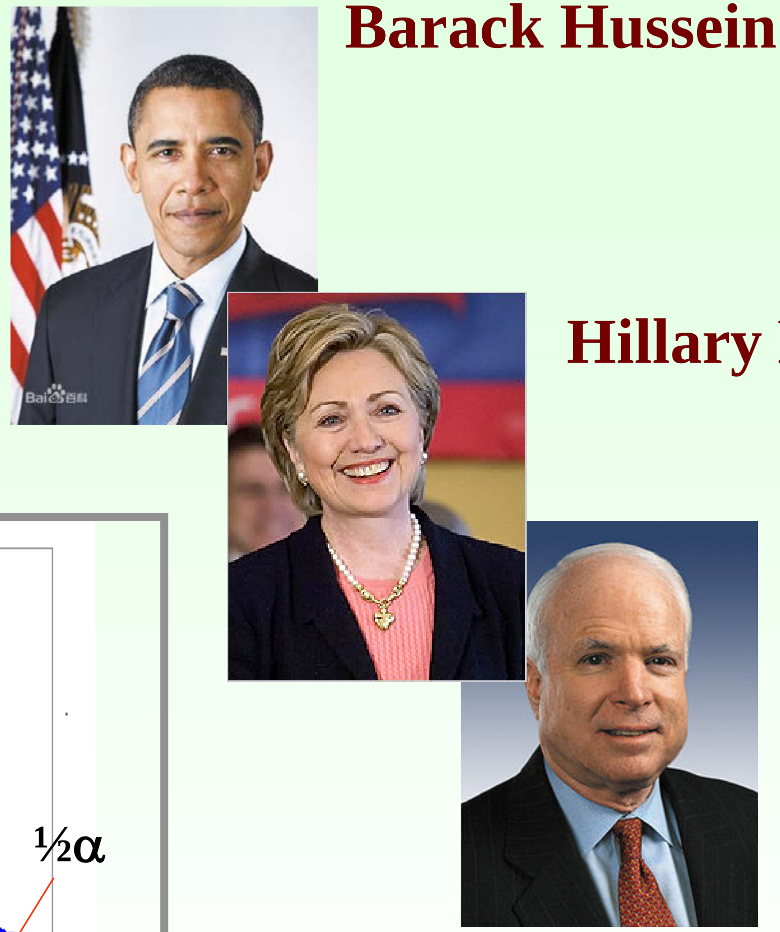
  The probability that a result is *outside* the confidence interval

  小概率事件*The Small Probability Event*



$$\pm u = \frac{x-\mu}{\sigma} \rightarrow x = \mu \pm u\sigma \rightarrow \mu = x \pm u\sigma$$

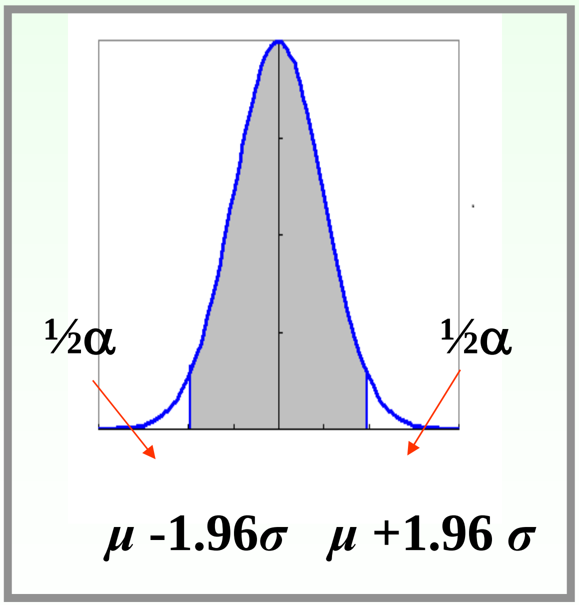




Barack Hussein Obama Jr.

Hillary Rodham Clinton

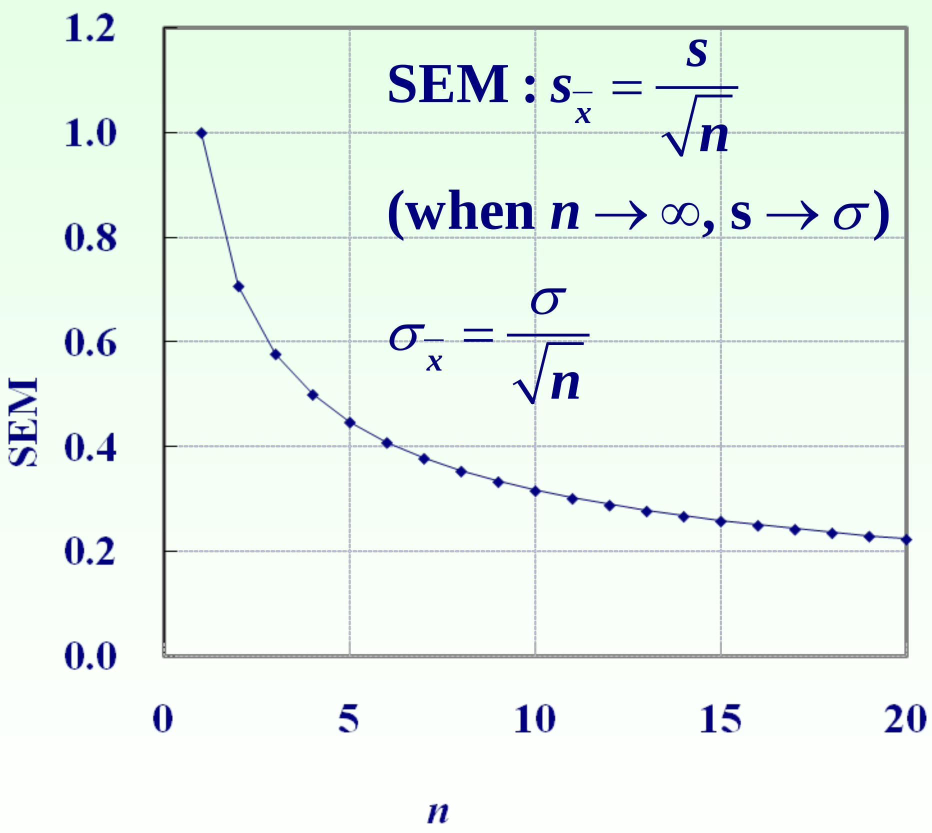
John McCain

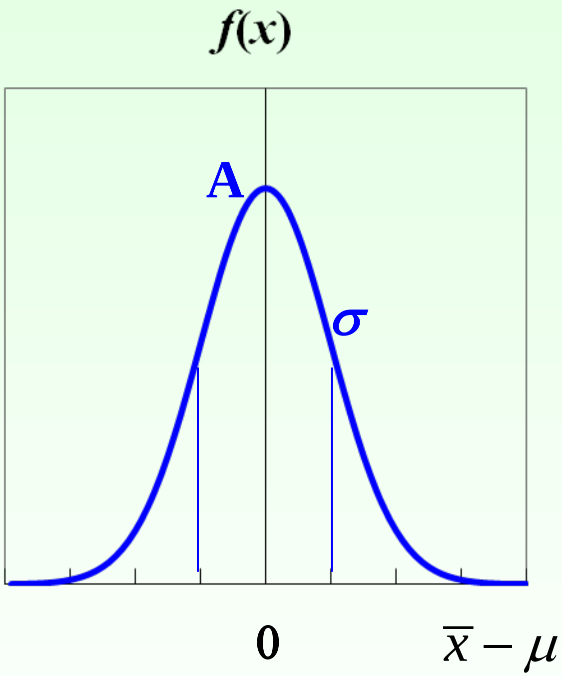


A female President? A black president? A Caucasian President?

Standard Deviation of the Mean (SEM)

平均值的标准差





SEM is the precision of the mean.

Confidence Interval (CI)

- 若**置信度**(把握)为95%, 则u=1.96, μ的置信区间为 (x -1.96σ, x+1.96σ)

  It is 95% probable that the true population mean lies in the f (x -1.96σ, x+1.96σ) (single determination)

- 若平行测定n次, μ 的**置信区间**为

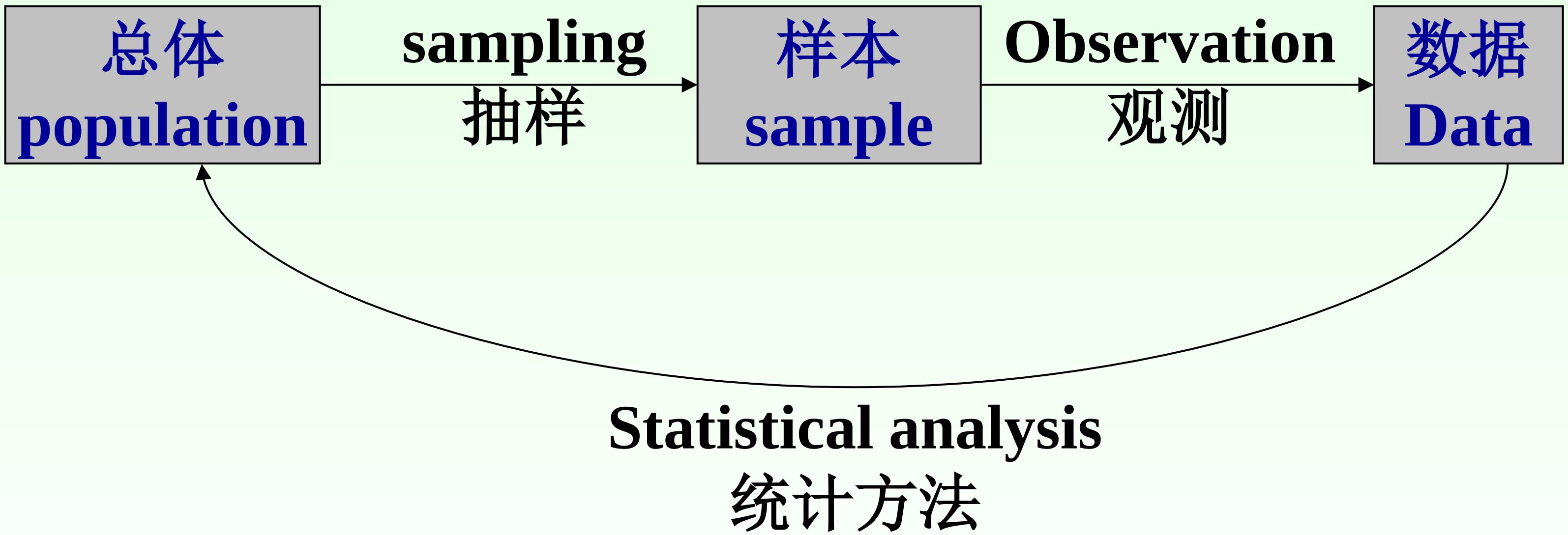
  For n replicate measurements, CI at α

$$(\bar{x} - u_{\alpha}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + u_{\alpha}\frac{\sigma}{\sqrt{n}})$$

$$(\bar{x} - u_{\alpha}\mathrm{SEM}, \quad \bar{x} + u_{\alpha}\mathrm{SEM})$$

Statistical Data Treatment and Evaluation Based on Sampling

有限数据的统计处理



样本容量*n*：样本所含的个体数.

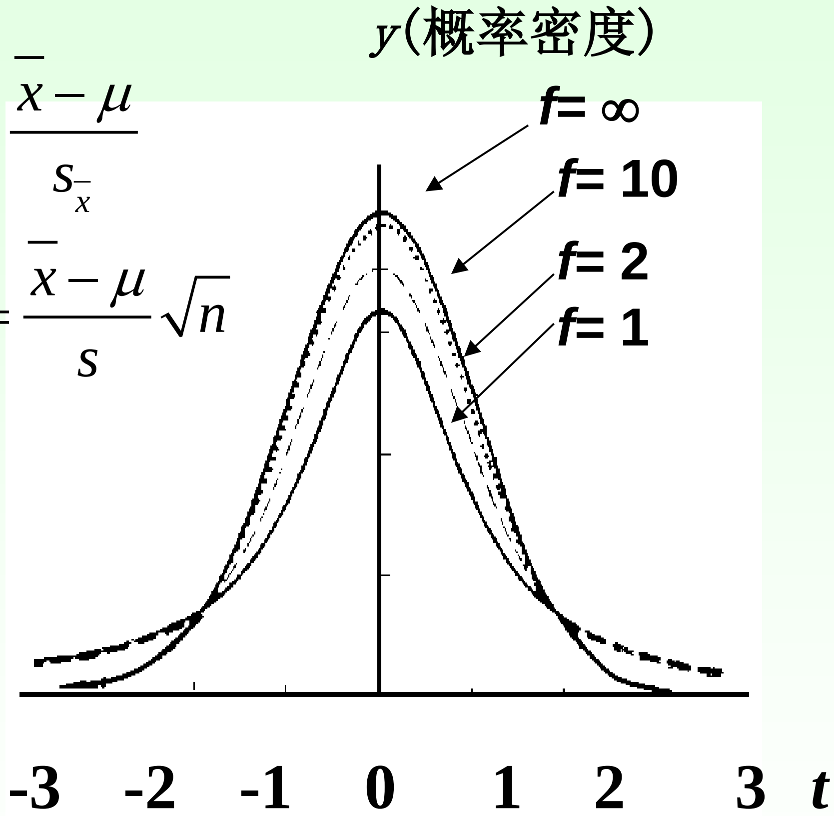
The number of measurements in the sample set

The number of total number of measurements in the population

t-distribution *t* 分布

-for limited number of determinations

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s}\sqrt{n}$$



William Gosset, 1876-1937, England
A Chemist & a Mathematician

***n*-1**

Gosset invented the t-test to handle small samples for quality control in brewing. He wrote under the name "Student".

CI for Limited Number of Determination

$$\bar{x},\ n,\ s$$

Confidence interval for μ

$$\left(\bar{x} - t_\alpha(f)\frac{s}{\sqrt{n}},\ \ \bar{x} + t_\alpha(f)\frac{s}{\sqrt{n}}\right),$$

that is $(\bar{x} - t_\alpha(f)$**SEM**, $\bar{x} + t_\alpha(f)$**SEM**$)$

t **is determined by 1-**α **(Confidence Level)**

and f **(Degree of Freedom自由度).**

Values of *t* at Levels of Probability as Indicated

*t*分布值表

$t_\alpha(f)$ / f	显 著 水 平 α			
	0.50	0.10	0.05	0.01
1	1.00	6.31	12.71	63.66
2	0.82	2.92	4.30	9.93
3	0.77	2.35	3.18	5.84
4	0.74	2.13	2.78	4.60
5	0.73	2.02	2.57	4.03
6	0.72	1.94	2.45	3.71
20	0.69	1.73	2.09	2.85
∞	0.67	1.64	1.96	2.58

Finding the CI 置信区间的确定

1. s is a good estimate of σ (σ已知),

(1 - α)% CI for μ is (置信度为 (1 - α)时，μ的CI为):

$$(\bar{x} - u_\alpha \frac{\sigma}{\sqrt{n}}, \bar{x} + u_\alpha \frac{\sigma}{\sqrt{n}}), \text{ i.e., } (\bar{x} - u_\alpha \text{SEM}, \bar{x} + u_\alpha \text{SEM})$$

2. σ is unknown (σ未知)

置信度为(1 - α)时，μ的置信区间为:

$$(\bar{x} - t_\alpha(f)\frac{s}{\sqrt{n}}, \ \bar{x} + t_\alpha(f)\frac{s}{\sqrt{n}}),$$

i.e., $(\bar{x} - t_\alpha(f)\text{SEM}, \ \bar{x} + t_\alpha(f)\text{SEM})$

CI *versus n*

Description of Analytical Results
分析结果的正确表达

$$\bar{x},\ s,\ n$$

有了这三个数据就可以对μ和σ值进行统计推断——真值处于的区间。

Test of Significance 显著性检验

- Comparing an experimental mean with a known value测定值与标准值比较
  - **Known σ (*s* is a good estimate of σ) —*u(z)*检验**
  - **Unknown σ---*t*检验**
- **Comparison of two experimental means 两组测定值相比较**

Null hypothesis零假设/原假设 (H_0)

—A null hypothesis postulates that two or more observed quantities are the same.

—进行统计检验时预先建立的假设。零假设成立时，有关统计量应服从已知的某种概率分布。当统计量的计算值落入否定域时，可知发生了小概率事件，应否定原假设。

Test of Significance 显著性检验

$u(z)$-Test ($s \to \sigma$)

1. **Null Hypothesis H_0: $\mu=\mu_0$**
2. **Calculate u**

$$u_{cal} = \frac{\bar{x}-\mu_0}{\sigma / \sqrt{n}} = \frac{\bar{x}-\mu_0}{SEM}$$

3. **Reject H_0 if $|u_{cal}| > u_\alpha$ at a given α.**

   That is, $\mu=\mu_0$, there is significant difference between the estimated μ and μ_0.

   即在给定显著性水平下，μ与μ_0有显著差异, 测定存在系统误差.

T- Test (σ unknown)

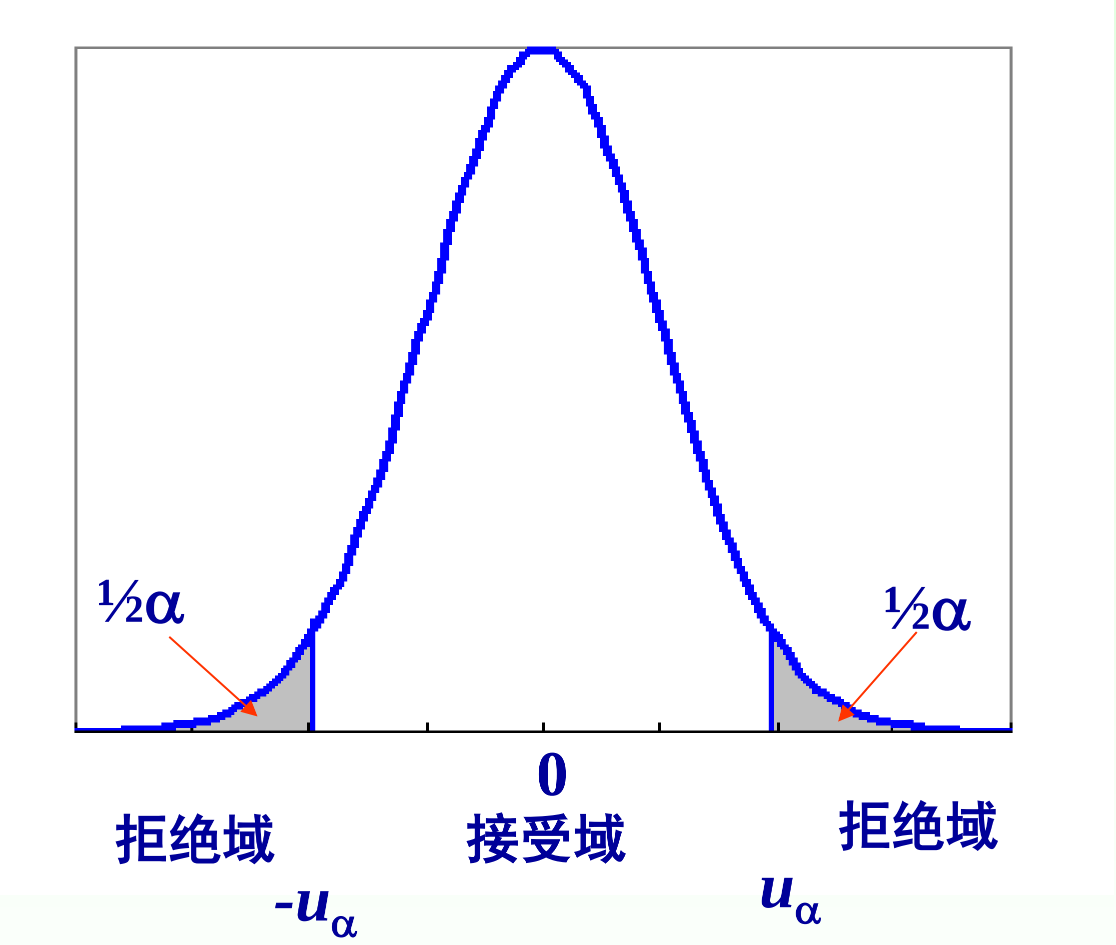
1. **Null Hypothesis H_0: $\mu=\mu_0$**
2. **Calculate u**

$$t_{cal} = \frac{\bar{x}-\mu_0}{s / \sqrt{n}} = \frac{\bar{x}-\mu_0}{SEM}$$

3. **Reject H_0 if $|t_{cal}| > t_\alpha$ at a given α.**

   That is, $\mu=\mu_0$, there is significant difference between the estimated μ and μ_0

   即在给定显著性水平下，μ与μ_0有显著差异, 测定存在系统误差.



Rejection Region and Acceptance Region

拒绝域和接受域

Example 2.5

For molten Fe, w(C)=4.55%(μ_0), σ =0.08%. There is suspect of change of Fe content. Samples from 5 furnaces are (%): 4.28, 4.40, 4.42, 4.35, and 4.37. Estimate if there is any change of Fe content at 95% confidence level?(α=0.05)

Null hypothesis $\mu=\mu_0$ =4.55%, $\bar{x}$ =4.36%

$$u_{\mathrm{cal}} = \frac{|\bar{x} - \mu_0|}{\sigma / \sqrt{n}} = \frac{|4.36\% - 4.55\%|}{0.08\% / \sqrt{5}} = 5.31$$

$u_{0.05}$=1.96,　　u_{cal} =5.31>1.96

Null hypothesis is rejected at the 95% confidence level.

That is the iron content not the same as before.

拒绝假设, 即平均含碳量与原来相比有变化.

Example 2.6

For a calcium sample with known μ_0, w(CaO)=30.43% . Estimate whether there are systematic errors of one student's results (n=6, $\bar{x}$ =30.51%, s=0.05%) at 95% confidence level.

Null hypothesis: $\mu=\mu_0$ =30.43%

$$t_{\text{cal}} = \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} = \frac{|30.51\% - 30.43\%|}{0.05\%/\sqrt{6}} = 3.9$$

$t_{0.05}(5)$=2.57, $t_{\text{cal}} > t_{\text{crit}}$

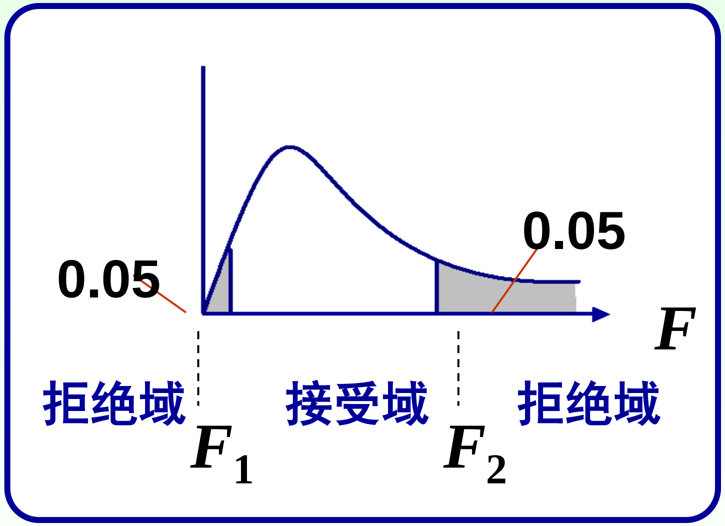
Null hypothesis is rejected. That is, there are systematic errors.拒绝假设, 此测定存在系统误差.

Comparison of Two Experimental Means
两组测量结果比较

Step 1: Comparison of precision – *F* test

第一步: ***F* 检验**—比较两组的精密度

(1) $H_o : \sigma_1 = \sigma_2$



(2) $F_{cal} = \dfrac{s_{big}^2}{s_{small}^2}$

(3) Reject H_o if $F_{cal} > F_{\alpha/2}(f_1, f_2)$

不对称的分布函数来源于两个自由度 $y = \dfrac{cF^{(n_1-2)/2}}{(n_2 + n_1 F)^{(n_1+n_2)/2}}$

Critical Values of *F* at the 5% Probability Level (95% confidence Level)显著水平 为0.05的*F*分布值表

No. of Degrees of Freedom		分　子 f_1 (bigger *s*)						
		2	3	4	5	6	7	∞
分母 f_2	2	19.00	19.16	19.25	19.30	19.33	19.36	19.50
	3	9.55	9.28	9.12	9.01	8.94	8.88	8.53
	4	6.94	6.59	6.39	6.26	6.16	6.09	5.63
	5	5.79	5.41	5.19	5.05	4.95	4.88	4.36
	6	5.14	4.76	4.53	4.39	4.28	4.21	3.67
	9	4.26	3.86	3.63	3.48	3.37	3.29	2.71
	∞	3.00	2.60	2.37	2.21	2.10	2.01	1.00

Step 2: t test to compare two means ($\sigma_1 = \sigma_2$)

第二步：t检验—比较 $\bar{x}_1$与 $\bar{x}_2$ $\sigma_1 = \sigma_2$

(1) $H_o: \mu_1 = \mu_2$

(2) $t_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{Sp}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$

pooled standard deviation 合并标准差 **:**

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(3) Reject H_o if $|t_{cal}| > t_\alpha(n_1 + n_2 - 2)$

Example 2.7

Compare the results from two methods for determination of *w*(Na_2CO_3) at 95% confidence level.

Method 1

n_1 = 5

$\bar{x}_1$ = 42.34%

s_1 = 0.10%

Method 2

n_2 = 4

$\bar{x}_2$ = 42.44%,

s_2 = 0.12%

1. *F* test ($\alpha = 0.10$)

$$F_{\mathrm{cal}} = \frac{s_{\mathrm{big}}^2}{s_{\mathrm{small}}^2} = 0.12^2/0.10^2 = 1.44 < F_{0.05}(3,4) = 6.59$$

σ_1 and σ_2 are statistically identical. That is, there is no significant difference between σ_1 and σ_2;

2. *t* test ($\alpha = 0.05$)

$$t_{\mathrm{cal}} = \frac{\overline{x_1} - \overline{x_2}}{s_p}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = 1.36 < t_{0.05}(7) = 2.37$$

There is no systematic error at 95% confidence level.

两种方法不存在系统误差。

Why Statistical Analysis of Analytical Results is Necessary and Important?

- **What is the maximum error can I tolerate?**
- **Lower? Identical? Higher?**

Alcohol in blood sample (%): 0.084, 0.089, 0.079 (critical value 0.090)

- **Is the new method can be adopted?**

Glucose in serum (mg/L)

Method	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6
A	994	670	795	750	907	600
B	1028	711	820	795	935	639

2. *t* test for paired data ($\alpha = 0.05$)

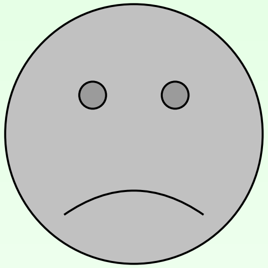
$$t_{\mathrm{cal}} = \frac{\bar{d} - \Delta_0}{s_{\mathrm{d}}\sqrt{N}} = \frac{\bar{d} - 0}{s_{\mathrm{d}}\sqrt{N}} = 4.628 > t_{0.05}(5) = 2.57,$$

$$\bar{d} = \frac{\sum\left|d_i\right|}{N} = 14.67,\ s_{\mathrm{d}} = \sqrt{\frac{\sum d_i^2 - \frac{\sum\left|d_i^2\right|}{N}}{N-1}} = 7.76,\ N = 6$$

We reject the null hypothesis and conclud that these two methods give different results. 两种方法存在系统误差。

2. *t* test for difference in means($\alpha = 0.05$)

$$t_{\text{cal}} = \frac{\overline{x_1} - \overline{x_2}}{s_p}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = 0.35 < t_{0.05}(10) = 2.23$$



The null hypothesis would be accepted.

The large patient-ti-patient variability masks the method differences that are of interest. Pairing allow us to focus on the differences.

Method	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Mean	SD
A	**994**	**670**	**795**	**750**	**907**	**600**	**786**	**146.5**
B	**1028**	**711**	**820**	**795**	**935**	**639**	**821.3**	**142.7**

$$S_p = 144.6$$

Detection of Gross Errors

异常值(Outlier)的检验—*Q* test

$$Q_{cal} = \frac{|x_q - x_n|}{x_{max} - x_{min}}$$

if $Q_{cal} > Q_{crit}$, then x_q is an outlier and can be rejected with the indicated degree of confidence.

Critical values for the Rejection Quotient, *Q*值表

Number of observations 测量次数*n*	3	4	5	6	7	8	9	10
$Q_{0.90}$	0.94	0.76	0.64	0.56	0.51	0.47	0.44	0.41
$Q_{0.95}$	0.97	0.84	0.73	0.64	0.59	0.54	0.51	0.49

2.4 测定方法的选择与测定准确度的提高

(Will also be lectured at the end of the semester!)

1.选择合适的分析方法：根据待测组分的含量、性质、试样的组成及对准确度的要求；

2.减小测量误差：取样量、滴定剂体积等；

3.平行测定4-6次，使平均值更接近真值；

4.消除系统误差：

(1) 显著性检验确定有无系统误差存在,

(2) 找出原因, 对症解决。

Regulations几项规定

1. 数字前0不计,数字后计入：0.02450

2. 数字后的0含义不清楚时, 最好用指数形式表示：1000（ 1.0×10^3， 1.00×10^3 ,1.000×10^3)

3. 自然数可看成具有无限多位数(如倍数关系、分数关系)；常数亦可看成具有无限多位数，如π, e

4. 数据的第一位数大于等于8的,可看作多一位有效数字，如 9.45×10^4, 95.2%, 8.65

5. 对数与指数的有效数字位数按尾数计，

   如 $10^{-2.34}$; pH=11.02, 则$[H^+]=9.5\times10^{-12}$

6. 误差只需保留1～2位；

7. 化学平衡计算中,结果一般为两位有效数字(由于K值一般为两位有效数字)；

8. 常量分析法一般为4位有效数字(E_r≈0.1%），微量分析为2~3位。

Rule of thumb for rounding data

四舍六入 五成双