第10章 实验误差和数据处理

主要内容

- 准确度和精密度，误差的种类和减免；
- 偏差指标：标准偏差、相对标准偏差、平均偏差；
- *t*分布和置信区间；
- 一元线性回归方程及其相关系数、标准化残差检测异常值；
- *Q*检验法。

10.1 误差的产生和减免

10.1.1 精密度和准确度

- 准确度 **accuracy** ——指观测值与真实值接近的程度。
- 精密度 **precision**——是指一组平行的观测值互相接近的程度。
- 两者都好测量才有高的可靠性。

10.1.2 误差和偏差

误差（准确度）

绝对误差——个别观测值x_i（或平均值$\bar{x}$）与真实值μ之差。

$$E_i = x_i - \mu$$

相对误差：$E_r = E / \mu$ “相对”更有意义

偏差（精密度）

绝对偏差——个别观测值x_i与平均值$\bar{x}$之差。

$$d_i = x_i - \bar{x}$$

相对偏差：$d_r = d_i / \bar{x}$

真值

(1) 理论真值：三角形的三个内角之和180°;

(2) 约定真值：米原器和千克原器;

(3) 相对真值：有限次重复测量值的算术平均值；高级（准确度）测量器具所测得的值作为低级测量器具测量值的真值。

10.1.3 系统误差和随机误差

- 系统误差（可测误差）——规律性（误差符号、大小有规律）、系统性（重现性），可针对性校正，常常仅影响准确度。
  - 方法误差——不适当的实验设计或所选方法不恰当所引起的误差（如指示剂不当或方法导致沉淀溶解）。
  - 仪器和试剂误差——由于仪器不良或试剂不合规格所引起的误差（如砝码不准、试剂不纯）。

– 操作误差——操作习惯性不符合规程（如不当吹液）。但不包括偶然性过失误差，如溶液飞溅、沉淀穿滤、读错记错等，这些要尽量避免。

– 个人误差——个人不良心理（接近）、生理（色盲）

- 随机误差（偶然误差）——由于偶然的原因引起，如环境温度、湿度难以预测的涨落等。
- 大小和正负不固定，但足够多次的测量服从正态分布（中央高两边低的“山峰”）统计规律，因此可用多次平行测量降低。在日常分析中，一般平行测定3~4次，较高要求5~9次，最多10~12次。
- 随机误差小则精密度高，而随机误差大则一般同时影响精密度、准确度。

10.1.4 系统误差的检测方式（对照试验）

- 与（组成接近的）标准试样对照——测定样值—标准样值
- 与标准方法对照——无适合标准样，现有方法—标准方法
- 用回收试验进行对照——两份试样，一份加已知量的待测组分，检测是否能定量回收

10.1.5 系统误差的减免

- 分析方法的选择与完善——选择时，如重量法和滴定法准确度高，但检出限也高（能检测的最低浓度、含量等）；仪分相反（不是越高档越好）。完善方面，如重量法应尽可能减少杂质。
- 减小测量误差——如为了控制相对误差，对测量的质量（天平）、体积（滴定管、移液管）的最小量提出要求。
- 进行仪器的校对——如砝码、滴定管等。
- 进行空白实验——不加试样测定，然后扣除空白值，排除试剂、蒸馏水、器皿中的杂质等影响。

10.2 精密度表达——偏差

- 标准偏差 (SD，standard deviation)

$$S=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{\sum_{i=1}^{n}d_i^2}{n-1}}=\sqrt{\frac{\sum x_i^2-\frac{1}{n}(\sum x_i)^2}{n-1}}$$

标准偏差越大，数据越分散，精密度越小。所以某对象的均值、*S*和测量次数一般是表达测量结果所必须的。

- 相对标准偏差(RSD，relative standard deviation)，又称变异系数

$$RSD = \frac{S}{\bar{x}}$$

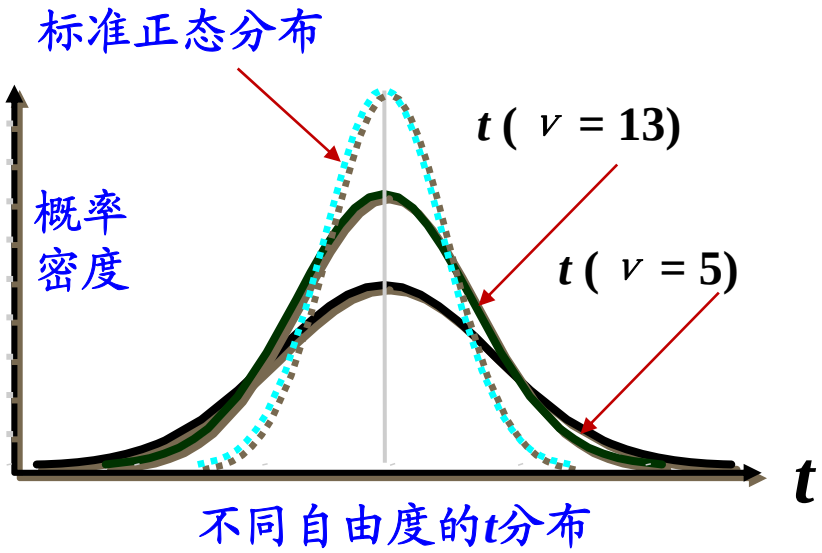
- 平均偏差（均差）

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

不会正负抵消

不如标准偏差对大偏差敏感（标准偏差中平方加大了大偏差的负面作用）

10.3 *t*分布



不同自由度的*t*分布

- 引入——如果测量次数*n*较少，随机误差导致的数据分布相对理想的正态分布有所偏移，可以用*t*分布描述。
- 特点——*t*分布曲线和正态分布曲线相似，只是变得与*n*有关（自由度 $\nu = n-1$），而且更加平坦和分散。当*n*趋向∞，*t*分布趋向（标准）正态分布。

- 置信度和置信区间——可以通过*t*分布估计某一组测量的可靠程度，即所谓置信度（置信水平），用概率*P*表示（常用95% ）。具体表达方式是在均值的基础上上下浮动一个数值，构成所谓置信区间，认为置信区间有置信度相应的概率包含了真值。

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$$

- 具体的一定置信度、自由度下对应的*t*临界值已经制表，见P397 表10.6。 P397例10.3要求的置信度越大，置信区间也就越宽。

10.4 回归Regression分析和一元线性回归

10.4.1 回归分析的内容

- 关联——从一组样本数据出发，确定变量之间的数学关系式；
- 评价——对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响显著，哪些不显著；

- 预测和控制——利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度。

10.4.2 一元线性回归方程

- 涉及一个自变量的回归
- 因变量*y*与自变量*x*之间为线性关系（或可变换为线性关系）

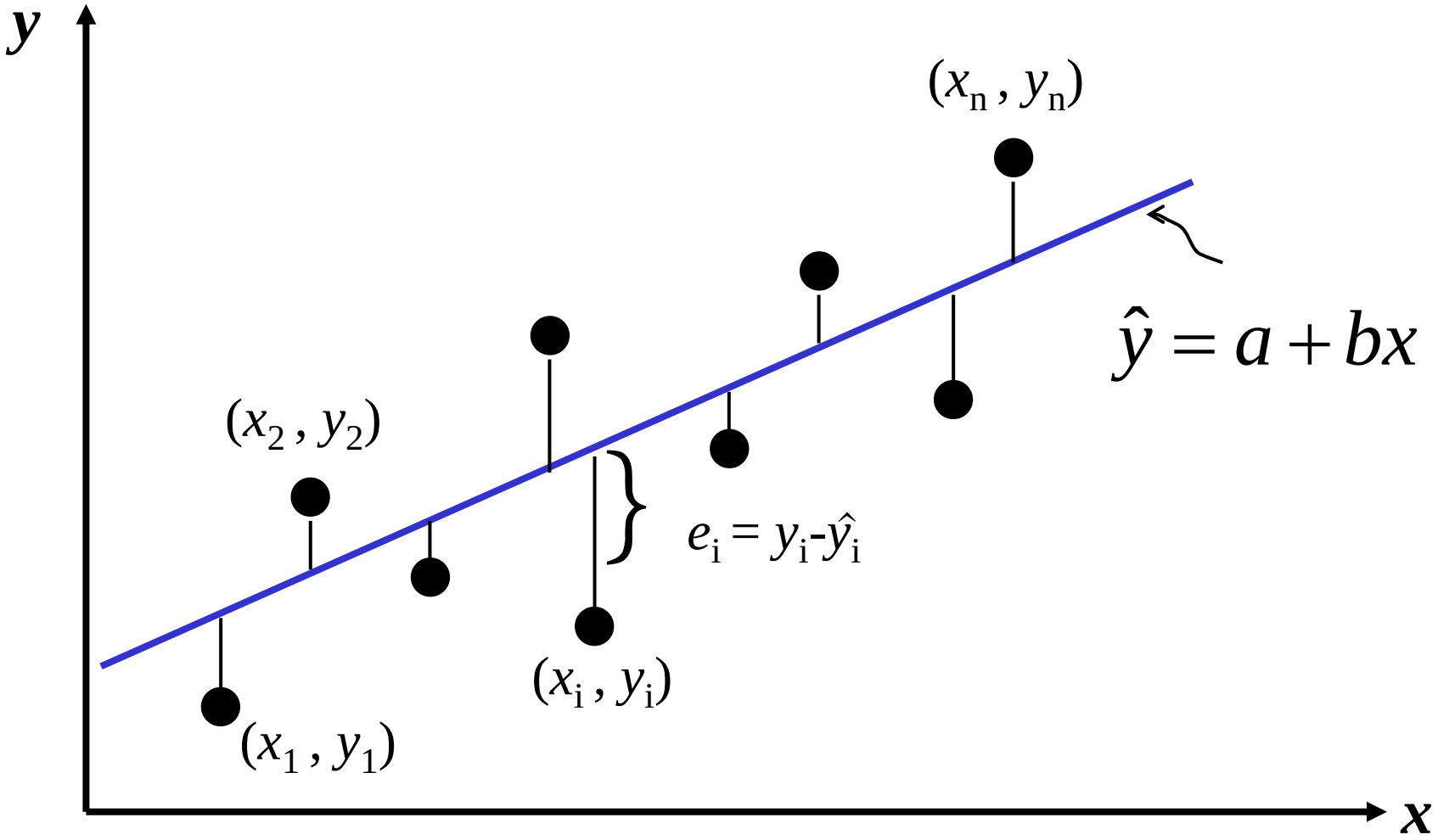
$$\hat{y} = a + bx$$

- 最小二乘法（依据）——使因变量的观察值与估计值之差的平方和达到最小来求得*a*和*b*的方法。即

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2=\text{最小}$$

可求得（P402 10.25有误）

$$\begin{cases} b=\dfrac{n\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2} \\ a=\bar{y}-b\bar{x} \end{cases}$$



最小二乘估计 (图示)

10.4.3 （线性）相关系数 *r*

- 描述变量之间密切程度和方向的指标

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$r=\pm 1$，存在线性关系，无实验误差；

$r=0$，毫无线性关系（但可能有非线性关系）；

$0<|r|<1$时，大於某临界值时，相关性显著，回归方程有意义，但不说明因果关系。

相关系数 *r* 的临界值（检验是否显著相关）

f=*n*-2 / *r* / 置信度	1	2	3	4	5	6	7	8	9	10
置 90%	0.988	0.900	0.805	0.729	0.669	0.622	0.582	0.549	0.521	0.497
信 95%	0.997	0.950	0.878	0.811	0.755	0.707	0.666	0.632	0.602	0.576
度 99%	0.999	0.990	0.959	0.917	0.875	0.834	0.798	0.765	0.735	0.708

例：分光光度法测定酚的数据如下

酚含量 x	0.005	0.010	0.020	0.030	0.040	0.050
吸光度 y	0.020	0.046	0.100	0.120	0.140	0.180

求回归方程表示含量与吸光度的关系，并检查方程是否有意义(置信度 95%)?

解：计算略。

回归方程为： $y = 0.013 + 3.40\,x$

计算相关系数，得： $r = 0.996$

查上表，当 $f = 6 - 2 = 4$ 时，选置信度 95%，$r_{临} = 0.811$

$$r_{计} > r_{临}$$

表明方程是有意义的。

- 常用还有决定系数R^2（相关系数的平方）。
- 数据处理软件的广泛应用给工作带来方便，现在比较简单常用的是Excel、Origin等。

10.5 标准化残差检验异常值

标准化残差 standardized residual

- 残差e_i除以它的标准差S_{ei}后得到的数值

$$z_{e_i} = \frac{e_i}{s_{e_i}} = \frac{y_i - \hat{y}_i}{s_{e_i}}$$

$$s_{e_i} = s_y\sqrt{1-h_i} = s_y\sqrt{1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}}$$

拟合的标准偏差

异常值 (outlier)

- 定义——如果某一个点与其他点所呈现的趋势不相吻合，这个点就有可能是异常点
- 判据——一般情况下，当一个观测值所对应的标准化残差的绝对值大于+2时，就可以将其视为异常值.

- 处理
  - 如果异常值是一个错误的数据，比如记录错误造成的，可设法修正乃至排除该数据，以便改善回归的效果;
  - 如果是由于模型的假定不合理，使得标准化残差偏大，应该考虑采用其他形式的模型，比如非线性模型。

10.6 可疑数据的取舍——Q检验法

- 同一数据的测量中，有的数据明显偏大或偏小，成为“可疑值”。可以采用Q检验法等方法决定取舍。

Q 检验法

- 步骤:

（1）数据从小到大排列 x_1，x_2 …… x_n

（2）求全距（极差） $x_n - x_1$

（3）求可疑数据与相邻数据之差（最大及最小）

$x_n - x_{n-1}$ 及 $x_2 - x_1$

（4）计算:

$$Q = \frac{x_n - x_{n-1}}{x_n - x_1} \quad 及 \quad Q = \frac{x_2 - x_1}{x_n - x_1}$$

（5）根据测定次数和要求的置信度查表

不同置信度下，舍弃可疑数据的Q值表

测定次数	Q_{90}	Q_{95}	Q_{99}
3	0.94	0.98	0.99
4	0.76	0.85	0.93
5	0.64	0.73	0.82
6	0.56	0.64	0.74
7	0.51	0.59	0.68
8	0.47	0.54	0.63
9	0.44	0.51	0.60
10	0.41	0.48	0.57

（6）将Q与Q_x（如Q_{90}）相比

若$Q > Q_x$可舍弃该数据，

若$Q \leqslant Q_x$该保留该数据。

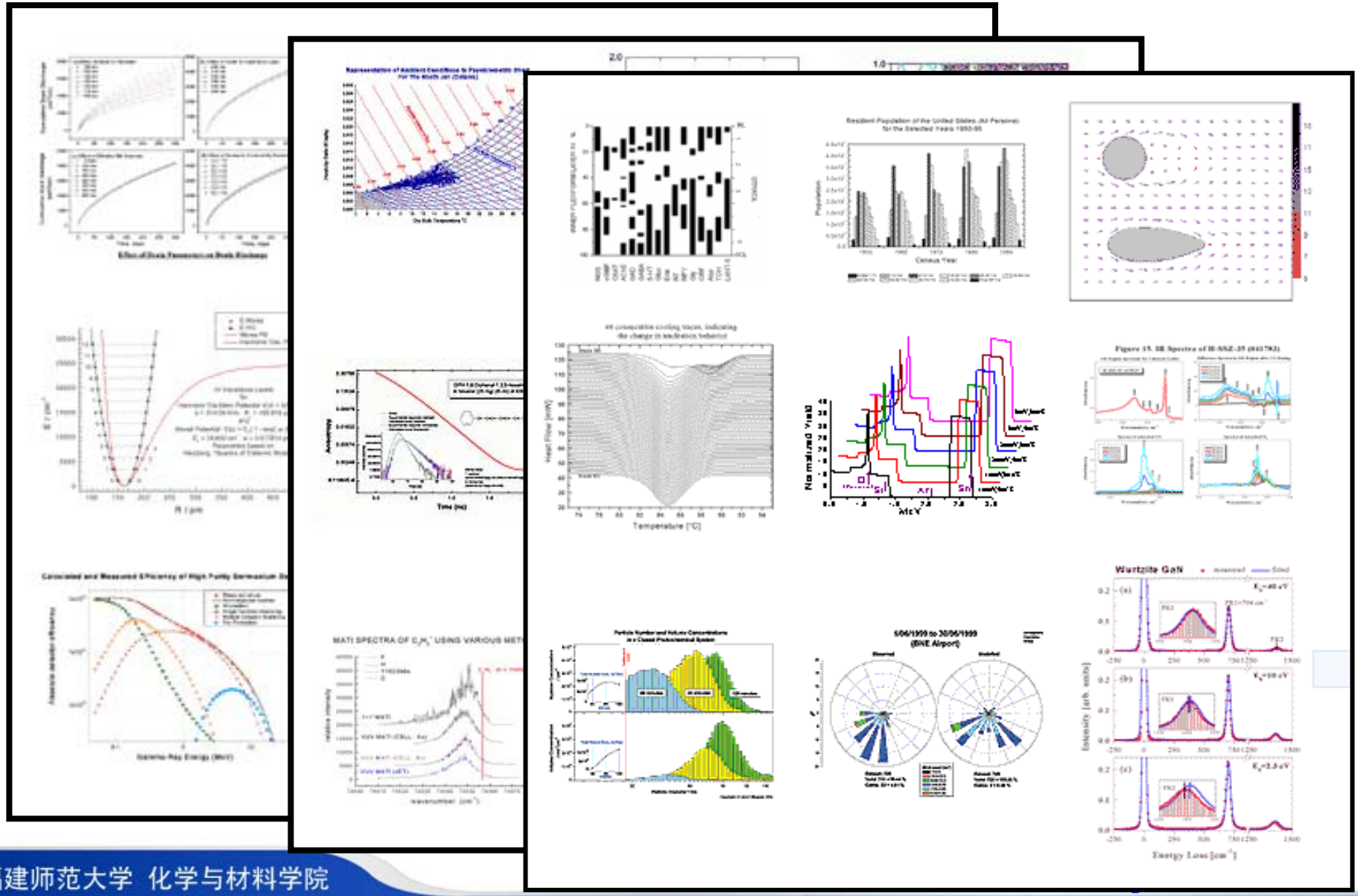
当数据较少时舍去一个后，应补加一个数据

- **P404 例10.7**

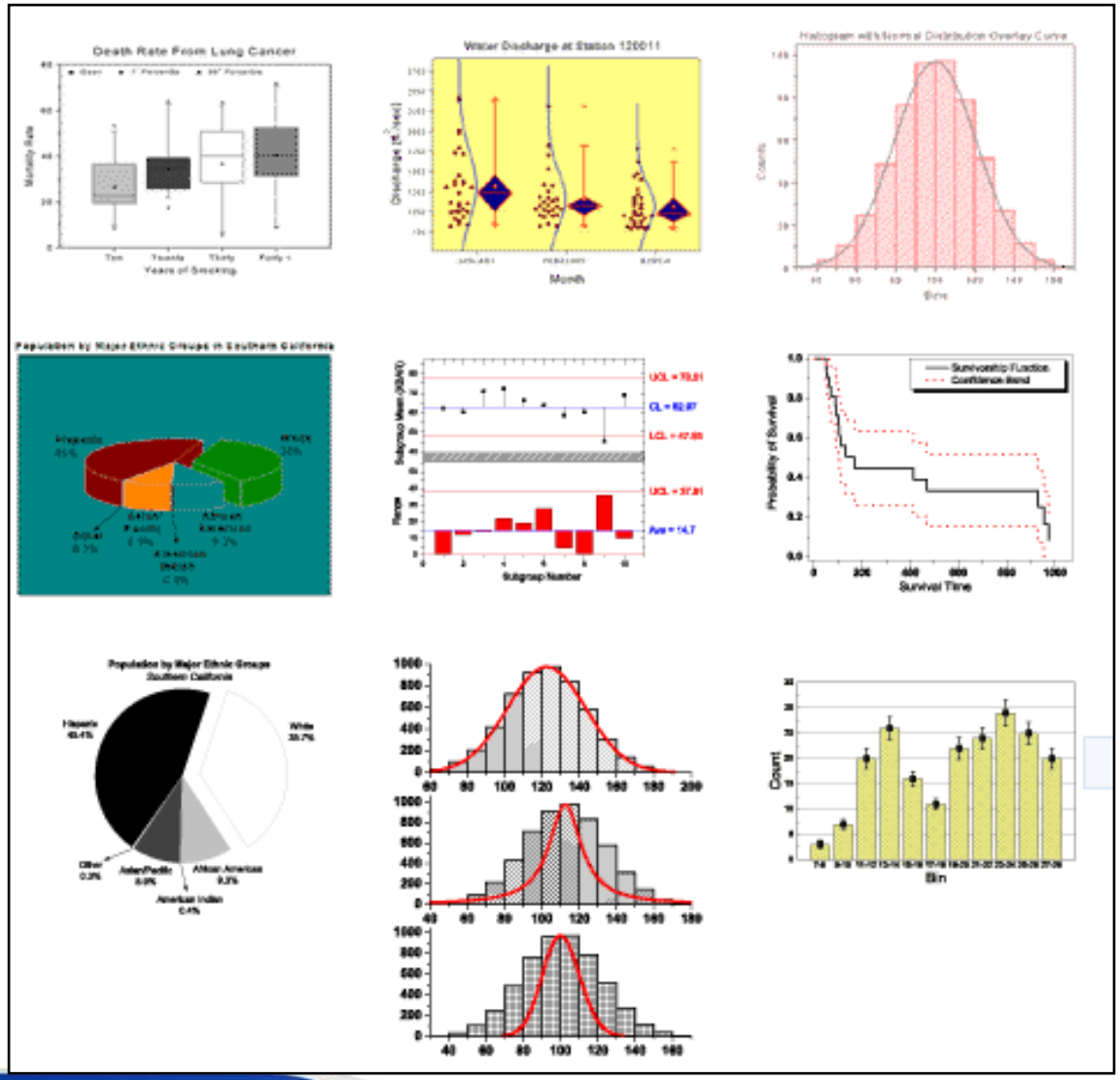
10.7 Origin 简介（了解）

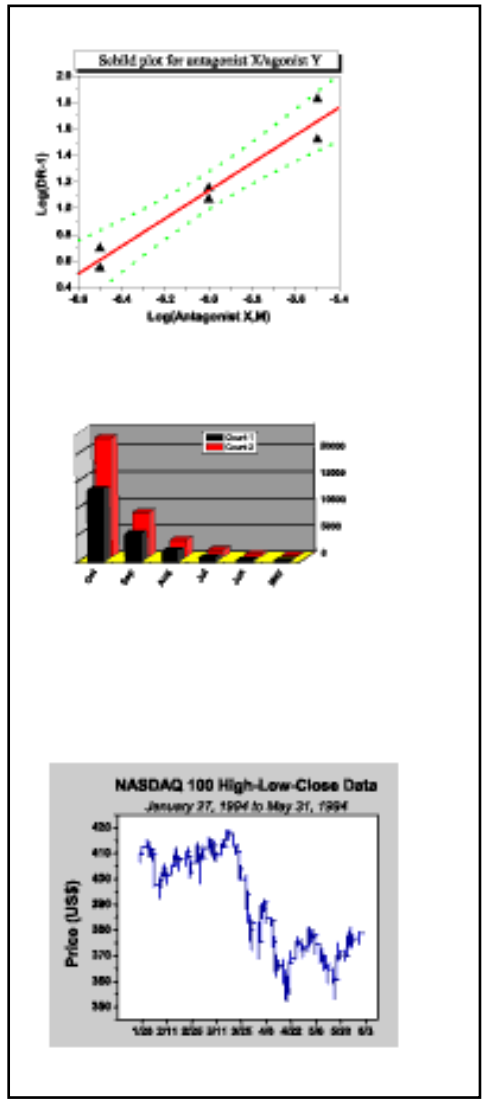
- **OriginLab公司的产品**
- 通用的科技绘图和数据分析软件
- 定位于基础级和专业级之间
- 国际科技出版界公认的标准作图软件
- 科学和工程研究人员的必备软件之一

Origin 绘制的2D图形

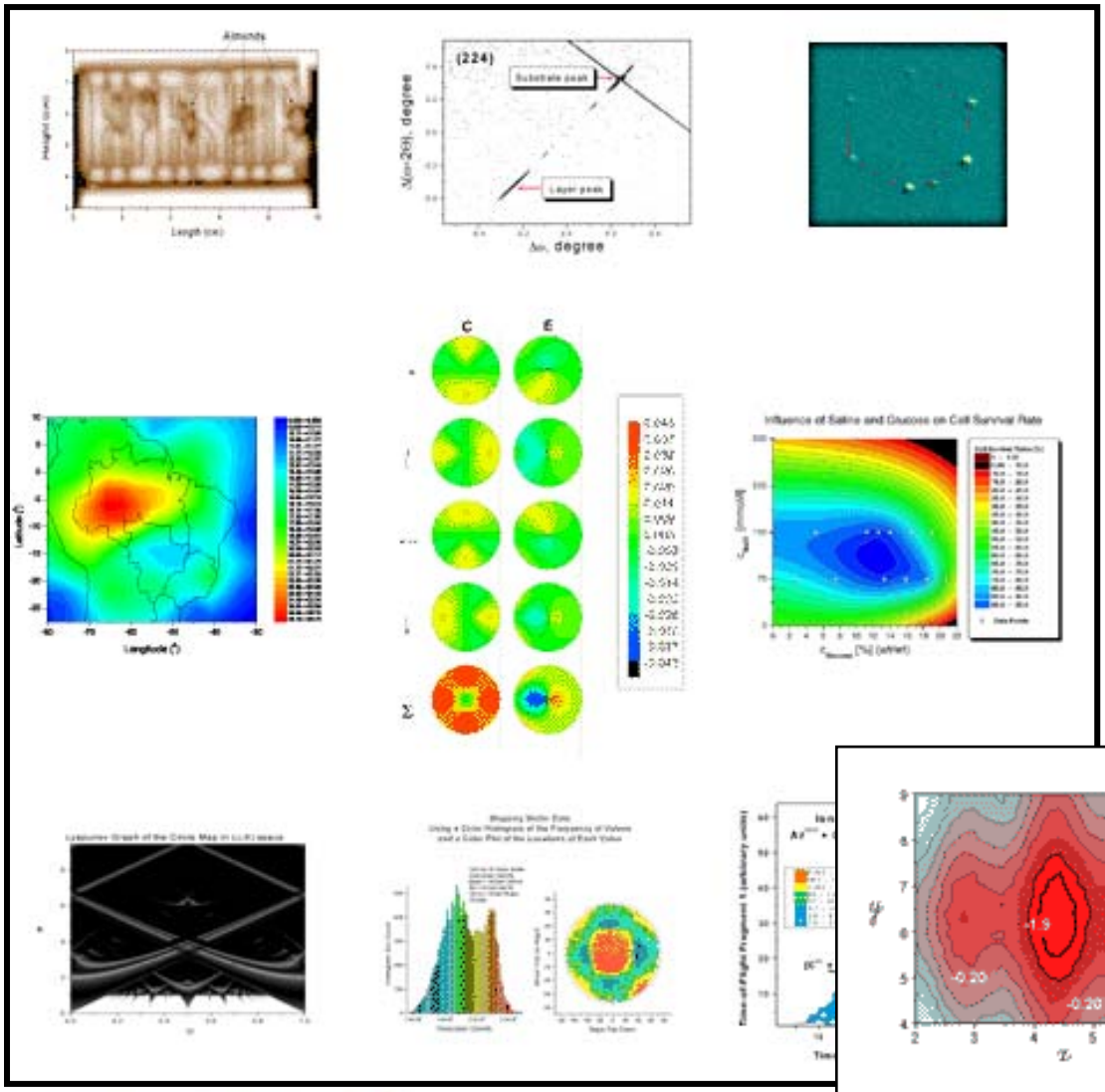


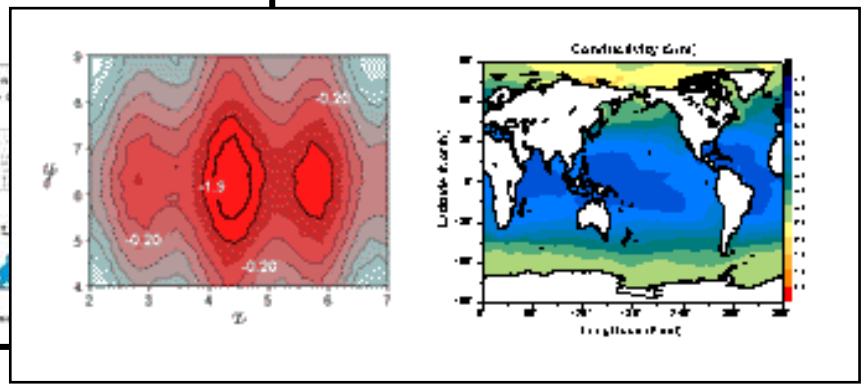
Origin 绘制的统计图形



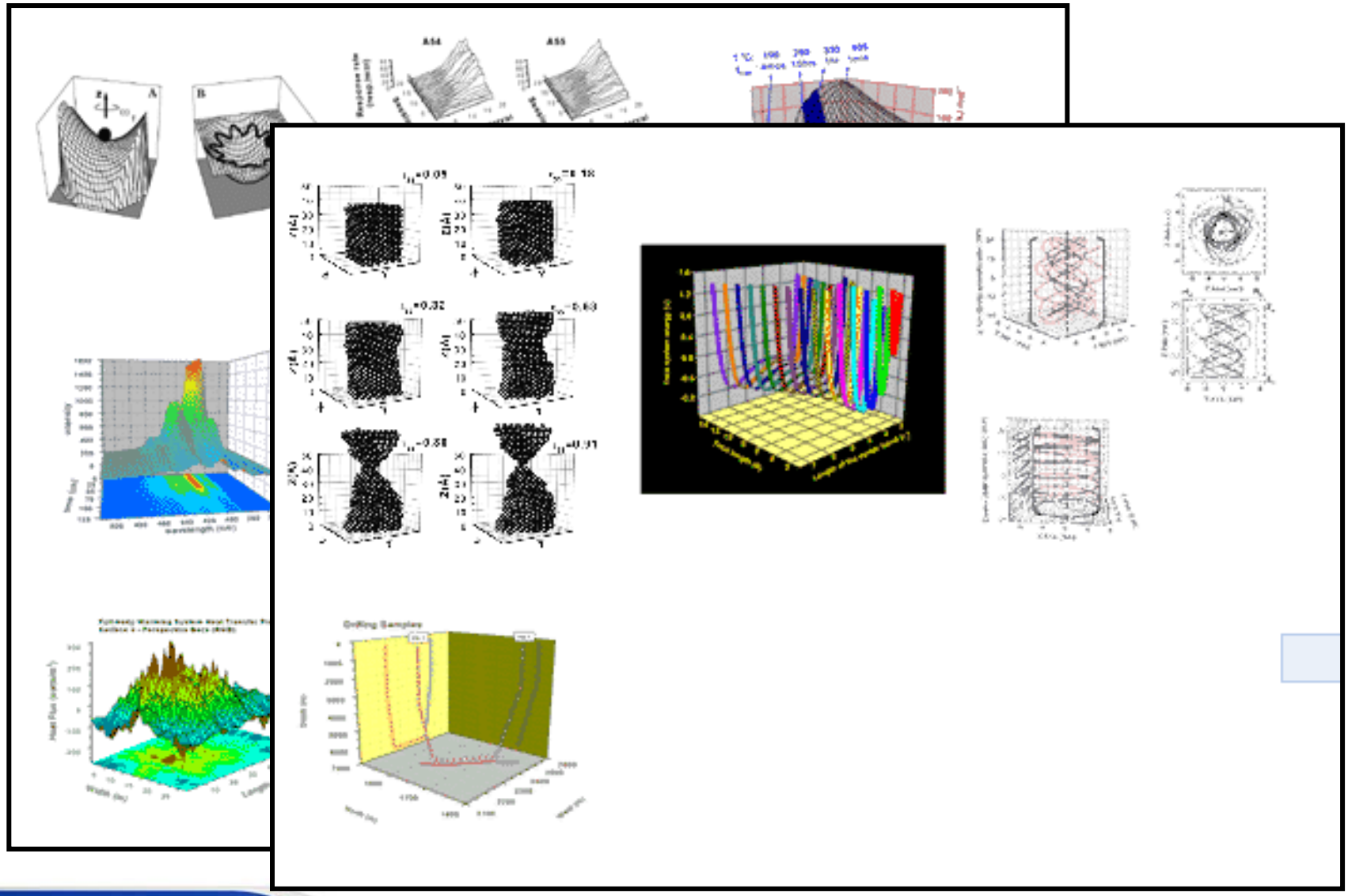


Origin 绘制的等高线图形





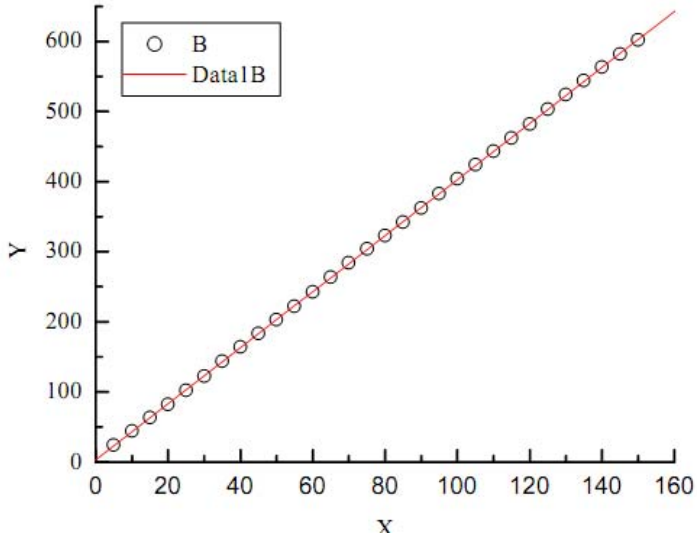
Origin 绘制的3D图形



[2000-7-20 16:20 "/Graph1" (2451745)]
Linear Regression for Data1_B:
Y = A + B * X

Parameter	Value	Error
A	3.1781	0.27652
B	3.99782	0.00312

R	SD	N	P
0.99999	0.73843	30	<0.0001



- **A: Intercept value and its standard error.** 截距值及它的标准误差
- **B: Slope value and its standard error.** 斜率值及它的标准误差
- **R: Correlation coefficient.** 相关系数
- **p: value - Probability (that R is zero).** R=0的概率
- **N: Number of data points.** 数据点个数
- **SD: Standard deviation of the fit.** 拟合的标准偏差

习题

- **P405：2、3**