《医学统计学》第四次课

正态分布与医学参考值范围

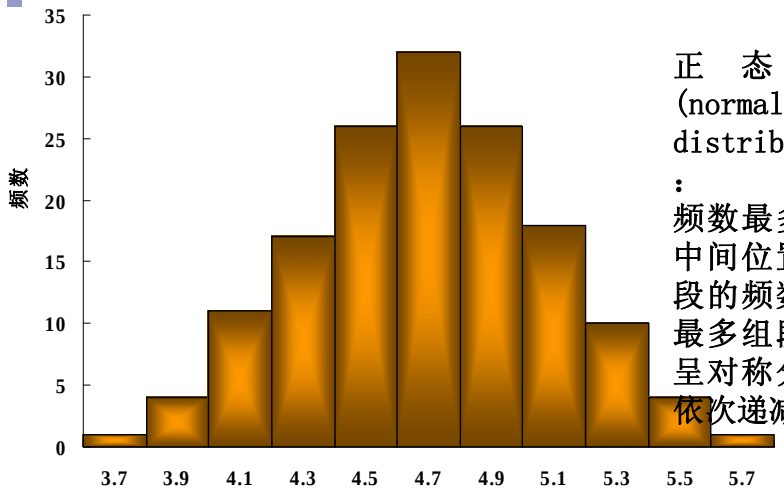
(Normal distribution and medical reference range)

Department of Health Statistics

第一节 正态分布

(Normal distribution)

- 正态分布的数学形式
- 正态分布的特征
- 标准正态分布
- 曲线下面积
- 正态分布的应用



正态分布(normal distribution)：
频数最多组段居中间位置，各组段的频数以频数最多组段为中心呈对称分布，并依次递减

图4-2 某地150名正常成年男子红细胞数(10^{12} / L)频数分布图

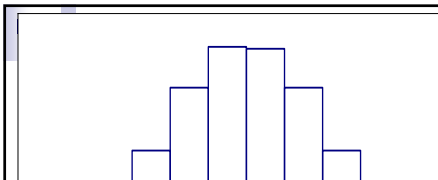


图6-1 频数分布图与正态分布曲线示意图

正态分布的数学形式

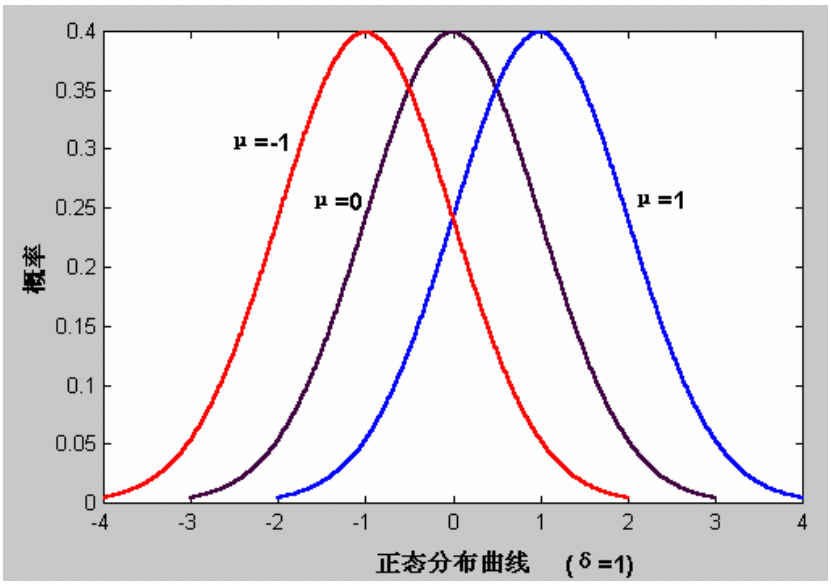
- 正态分布是描述连续型变量值分布的曲线，许多医学资料服从正态分布
- 正态分布的概率密度函数表达式:

$$f(X)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(X-\mu)^2}{2\sigma^2}},-\infty<X<+\infty$$

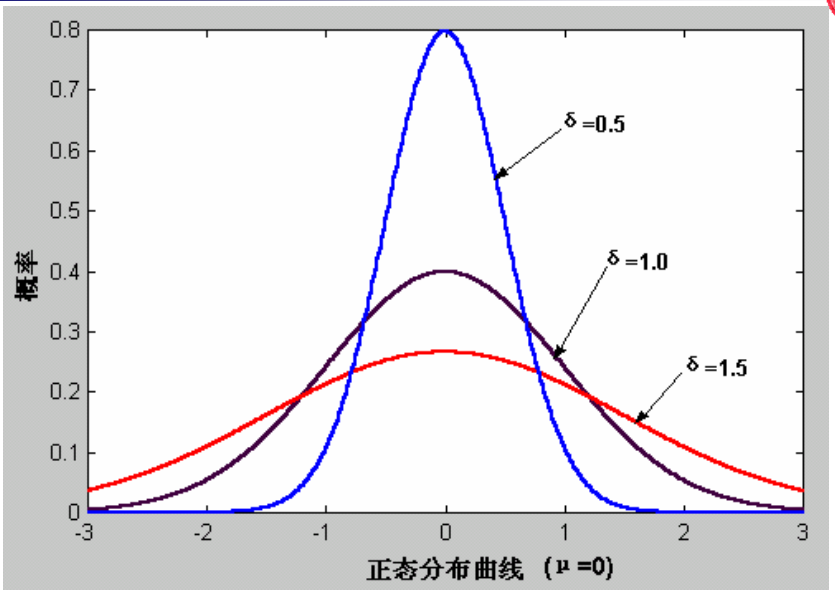
*X*为连续随机变量，μ为*X*值的总体均数，σ^2为总体方差，记为*X*~*N*（μ，σ^2）

正态分布的特征

- 1.在横轴上方呈钟形分布，两端与X轴永不相交，以*x*=μ为对称轴，左右完全对称
- 2.在*x*=μ处，*f*(*x*)取最大值，*x*离μ越远，*f*(*x*)值越小
- 3.正态分布有两个参数：位置参数—均数和形态参数—标准差
- 4.正态曲线下面积分布有一定的规律



正态分布位置变换示意图



正态分布形态变换示意图

标准正态分布

为应用方便，将X作变量变换，就图形来说就是把原点移到μ的位置上，即u变换，则将正态分布转换为标准正态分布，即 $u \sim N(0,1)$，u变换公式为：

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

- 标准正态分布的概率密度函数表达式：

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, -\infty < u < +\infty$$

正态曲线下面积

- 标准正态分布的分布函数

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{z^2}{2}} dz$$

根据标准正态分布的分布函数 $\phi(u)$，列出$u \leqslant 0$的所有 $\phi(u)$值，表示$X \leqslant u$的曲线下面积

$$P(u < a) = \phi(a)$$
$$P(a < u < b) = \phi(b) - \phi(a)$$
$$P(|u| > |b|) = 2 \times \phi(b)$$

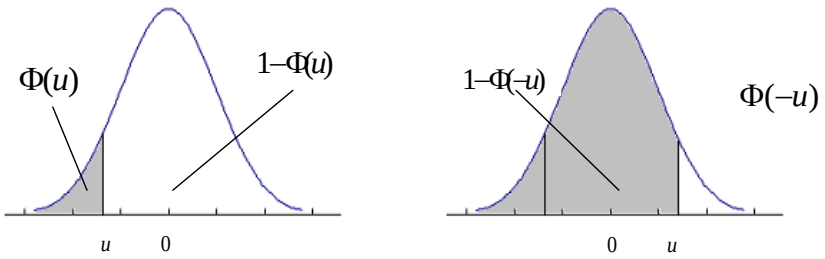
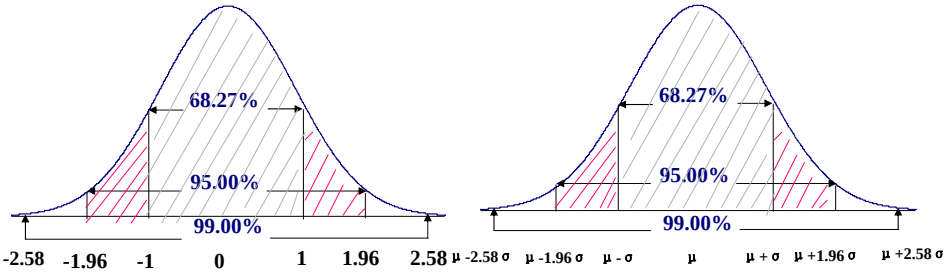


图 6-5　标准正态分布曲线下面积示意图

正态曲线下面积分布规律

- 横轴与正态曲线所夹面积恒等于1或100％



标准正态分布	正态分布	面积或概率
-1～1	μ ± σ	68.27%
-1.96～1.96	μ ± 1.96 σ	95.00%
-2.58～2.58	μ ± 2.58 σ	99.00%

计算正态曲线下面积实例

例6-1 某地108名正常成年女子的血清总蛋白(g/L)含量如表6-1，试估计该地正常女子血清总蛋白＜68.0(g/L)、＜78.0(g/L)和≥78.0(g/L)所占正常女子总人数的百分比

表 6-1　某地 108 名正常成年女子的血清总蛋白 (g/L) 含量

67.3	75.4	73.1	70.9	75.1	72.6	78.2	68.8	73.8	71.5	66.5	75.1
70.7	68.9	73.3	72.3	76.5	74.3	75.9	75.4	67.2	71.8	76.2	70.6
70.7	75.6	73.3	72.4	76.6	67.3	80.8	74.3	73.9	71.6	79.9	69.3
80.3	75.7	73.5	81.2	74.4	72.5	77.1	67.3	74.1	68.0	76.4	70.4
71.0	75.8	73.6	78.1	68.7	72.6	77.6	72.2	74.2	72.1	76.3	69.7
71.1	75.7	73.5	72.7	78.3	72.5	77.2	68.2	74.2	72.3	76.5	70.5
71.2	83.7	73.7	75.8	74.7	72.6	69.5	66.0	76.1	77.7	80.5	83.1
64.1	75.1	76.3	77.8	65.2	75.0	72.7	78.8	71.1	71.8	72.9	76.1
71.2	75.2	72.9	79.5	73.9	75.2	73.1	79.5	81.8	74.5	81.6	74.5

$$\bar{X} = \frac{7982.0}{108} = 73.9(g/L), S = 3.9(g/L)$$

进行***u*转换:**

$$u_1 = \frac{68.0 - 73.9}{3.9} = -1.51, u_2 = \frac{78.0 - 73.9}{3.9} = 1.05$$

查表:

$$\Phi(-1.51) = 0.0655$$

$$\Phi(1.05) = 1 - \Phi(-1.05) = 1 - 0.1469 = 0.8531$$

结论:

该地正常女子血清总蛋白含量＜68.0g/L者占总人数的6.55%，＜78.0g/L者占总人数的85.31%，≥78.0g/L者占总人数的14.69%

正态分布的应用

- 1.估计医学参考值范围
- 2.质量控制(quality control)
- 3.正态分布是许多统计方法的理论基础

第二节 医学参考值范围

(medical reference range)

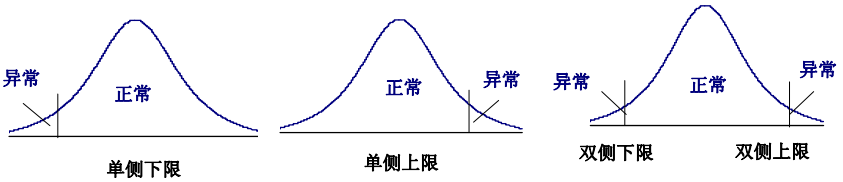
- 基本概念
- 正态分布法
- 百分位数法

基本概念

- 参考值(reference value)是指包括绝大多数正常人的人体形态、机能和代谢产物等各种生理及生化指标
- 由于个体的差异性，临床上常用的参考值一般不是一个常数，而是在一定范围内波动，故临床上采用医学参考值范围(medical reference range)作为判定正、异常的参考标准

- 正常人的概念：排除了影响所研究指标的疾病和有关因素的同质人群
- 单、双侧界值问题：根据专业知识确定

单侧下限---过低异常　单侧上限---过高异常　双侧---过高、过低均异常



肺活量

血清转氨酶
体内有害物质含量

红、白细胞计数
体温、脉搏

➢参考值范围：

指绝大多数正常人的某指标测量值都在一定的范围内

95%参考值范围指95%的正常人的某指标测量值在某范围内，而在该范围之外的还有5%

参考值范围的计算

➢正态分布法适用于正态分布或近似正态分布资料，偏态分布资料变量变换为正态分布后也可以采用，要求样本量较大

常用参考值范围的制定

参考值范围（%）	正态分布法		
	双侧	单侧	
		下限	上限
90	$\bar{X} \pm 1.64S$	$\bar{X} - 1.28S$	$\bar{X} + 1.28S$
95	$\bar{X} \pm 1.96S$	$\bar{X} - 1.64S$	$\bar{X} + 1.64S$
99	$\bar{X} \pm 2.58S$	$\bar{X} - 2.32S$	$\bar{X} + 2.32S$

例 6-3 估计例 6-1 某地 108 名正常成年女子血清总蛋白（$\bar{X}=73.8$ (g/L)，$S=3.8$ (g/L)）的95%参考值范围。

因血清总蛋白过多或过少均为异常，故按双侧估计正常成年女子血清总蛋白的95%参考值范围。

已知 $\bar{X}=73.9$ (g/L)， $S=3.9$ (g/L)， $u_{0.05/2}=1.96$，可得出

下限： $\bar{X}-u_{\alpha/2}S=73.9-1.96\times3.9=66.3$ (g/L)

上限： $\bar{X}+u_{\alpha/2}S=73.9+1.96\times3.9=81.5$ (g/L)

故该地正常成人血清总蛋白 95%的医学参考值范围为 66.3~81.5 (g/L) 。

➢百分位数法适用于偏态分布资料，所用的样本量较正态分布法要多

常用参考值范围的制定

参考值范围（%）	百分位数法		
	双侧	单侧	
		下限	上限
90	$P_5\sim P_{95}$	P_{10}	P_{90}
95	$P_{2.5}\sim P_{97.5}$	P_5	P_{95}
99	$P_{0.5}\sim P_{99.5}$	P_1	P_{99}

表 6-5　130 名正常人的血清肌红蛋白含量频数表

肌红蛋白含量 (μg/mL)	人数	累积频数	累积频率 (%)
0~	2	2	1.54
5~	3	5	3.85
10~	9	14	10.77
15~	12	26	20.00
20~	15	41	31.54
25~	27	68	52.31
30~	33	101	77.69
35~	18	119	91.54
40~	10	129	99.23
45~	1	130	100.00
合 计	130	—	—

$$P_{2.5} = L + \frac{i}{f_{2.5}}(n \times 2.5\% - \sum f_L) = 5 + \frac{5}{3}(130 \times 2.5\% - 2) = 7.1(\mu g/mL)$$

$$P_{97.5} = L + \frac{i}{f_{97.5}}(n \times 97.5\% - \sum f_L) = 40 + \frac{5}{10}(130 \times 97.5\% - 119) = 43.9(\mu g/mL)$$

参考值范围的计算

常用参考值范围的制定

参考值范围 (%)	正态分布法			百分位数法		
	双侧	单侧		双侧	单侧	
		下限	上限		下限	上限
90	$\overline{X} \pm 1.64S$	$\overline{X} - 1.28S$	$\overline{X} + 1.28S$	$P_5 \sim P_{95}$	P_{10}	P_{90}
95	$\overline{X} \pm 1.96S$	$\overline{X} - 1.64S$	$\overline{X} + 1.64S$	$P_{2.5} \sim P_{97.5}$	P_5	P_{95}
99	$\overline{X} \pm 2.58S$	$\overline{X} - 2.32S$	$\overline{X} + 2.32S$	$P_{0.5} \sim P_{99.5}$	P_1	P_{99}

第三节　　t 分布

$$N\left(\mu,\sigma^2\right)\xrightarrow{u=(X-\mu)/\sigma}N(0,1)$$

$$N\left(\mu,\sigma^2/n\right)\xrightarrow{u=\left(\overline{X}-\mu\right)/\left(\sigma/\sqrt{n}\right)}N(0,1)$$

$$N\left(\mu,\sigma^2/n\right)\xrightarrow{t=\left(\overline{X}-\mu\right)/\left(S/\sqrt{n}\right)}t\ \textit{distribution}$$

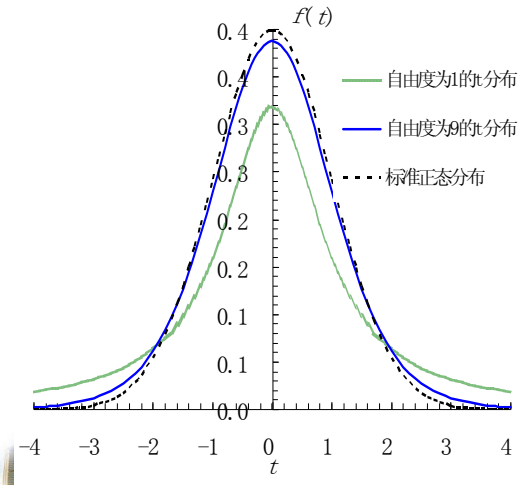
t分布的概率密度函数

$$f(t)=\frac{\Gamma\left[(\nu+1)/2\right]}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}(1+t^2/\nu)^{-(\nu+1)/2}$$

式中 $\Gamma(\bullet)$ 为伽玛函数；π 为圆周率；ν 为自由度 (degree of freedom)，是 t 分布的唯一参数；t 为随机变量。

以t为横轴，$f(t)$为纵轴,可绘制t分布曲线。

t分布曲线



*t*分布有如下性质：

①单峰分布，曲线在$t=0$处最高，并以$t=0$为中心左右对称

②与正态分布相比，曲线最高处较矮，两尾部翘得高（见绿线）

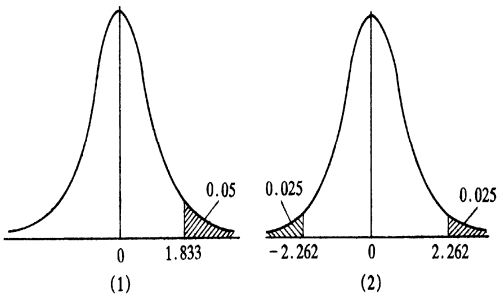
③ 随自由度增大，曲线逐渐接近正态分布；分布的极限为标准正态分布

t分布曲线下面积(附表2)

附表 2　*t* 界值表

自由度 υ	概率，P					
	单侧：0.25	0.05	0.025	0.01	0.005	0.0025
	双侧：0.50	0.10	0.05	0.02	0.010	0.0050
1	1.000	6.314	12.706	31.821	63.657	127.321
2	0.816	2.920	4.303	6.965	9.925	14.089
3	0.765	2.353	3.182	4.540	5.841	7.453
4	0.741	2.132	2.776	3.747	4.604	5.597
31	0.683	1.696	2.040	2.453	2.744	3.022
32	0.682	1.694	2.037	2.449	2.738	3.015
…	…	…	…	…	…	…
34	0.682	1.691	2.032	2.441	2.728	3.002
…	…	…	…	…	…	…
∞	0.6745	1.6449	1.9600	2.3263	2.5758	2.8070

t分布曲线下面积(附表2)



双侧$t_{0.05/2,\ 34}$=2.032

=单侧$t_{0.025,\ 34}$

单侧$t_{0.05,\ 34}$=1.691

双侧$t_{0.05/2,\ \infty}$=1.96

=单侧$t_{0.025,\ \infty}$

单侧$t_{0.05,\ \infty}$=1.64

SAS应用

- FREQ过程可以生成单向和多向的频数表和交叉表。
- MEANS过程用于对数值变量计算简单描述性统计量。
- UNIVARIATE过程可以计算的描述性统计量是最多的，而且还可以用图表的形式反映变量值的分布情况，并对变量进行正态性检验。

频数表的编制

```
/*例4.1*/
data prg4_1;
  input x @@;
  low=2.3;
  dis=0.3;
  z=x-mod(x-low,dis);
cards;
2.35 4.21 3.32 5.35 4.17 4.13 2.78 4.26 3.58 4.34 4.84 4.41
4.78 3.95 3.92 3.58 3.66 4.28 3.26 3.50 2.70 4.61 4.75 2.91
3.91 4.59 4.19 2.68 4.52 4.91 3.18 3.68 4.83 3.87 3.95 3.91
4.15 4.55 4.80 3.41 4.12 3.95 5.08 4.53 3.92 3.58 5.35 3.84
3.60 3.51 4.06 3.07 3.55 4.23 3.57 4.83 3.52 3.84 4.50 3.96
4.50 3.27 4.52 3.19 4.59 3.75 3.98 4.13 4.26 3.63 3.87 5.71
3.30 4.73 4.17 5.13 3.78 4.57 3.80 3.93 3.78 3.99 4.48 4.28
4.06 5.26 5.25 3.98 5.03 3.51 3.86 3.02 3.70 4.33 3.29 3.25
4.15 4.36 4.95 3.00 3.26
;
proc freq;
tables z;
run;
```

The SAS System 21:16 Sunday

The FREQ Procedure

z	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2.3	1	0.99	1	0.99
2.6	3	2.97	4	3.96
2.9	6	5.94	10	9.90
3.2	8	7.92	18	17.82
3.5	17	16.83	35	34.65
3.8	20	19.80	55	54.46
4.1	17	16.83	72	71.29
4.4	12	11.88	84	83.17
4.7	9	8.91	93	92.08
5	5	4.95	98	97.03
5.3	2	1.98	100	99.01
5.6	1	0.99	101	100.00

MEANS 过程

- 程序4-2

```
proc means data=prg4_1;
var x;
run;
```

运行结果：

The MEANS Procedure

Analysis Variable : x

N	Mean	Std Dev	Minimum	Maximum
101	4.0295050	0.6592183	2.3500000	5.7100000

Stderr：均数的标准差，即标准误；

Sum：和；

Variance：方差；

CV：变异系数；

Nmiss：缺失变量值的观测的例数；

Range：极差；

USS：平方和；

CSS：离均差平方和；

T：检验假设为总体均数为0的student t检验的检验统计量t值；

Probt：总体均数为0的检验假设中，t值所对应的概率值（P值）；

Sumweight：权重变量值的和；

Skewness：偏度系数；

Kurtosis：峰度系数；

CLM：双侧95%可信区间的下限（LCLM）和上限（UCLM）。

Median|P50：中位数或50%分位数。

P1：1%分位数。

P5：5%分位数。

P10：10%分位数。

Q1|P25：下四分位数或25%分位数。

Q3|P75：上四分位数或75%分位数。

P90：90%分位数。

P95：95%分位数。

P99：99%分位数。

Qrange：四分位数间距。

程序4-3

proc means n mean std stderr cv clm;
var x;

N	Mean	Std Dev	Std Error	Coeff of Variation	Lower 95% CL for Mean	Upper 95% CL for Mean
101	4.0295050	0.6592183	0.0655947	16.3597836	3.8993670	4.1596429

- 程序4-4

```
data prg4_4;
input x f @@;
cards;
2.45  1 2.75  3 3.05 6 3.35 8 3.65 17 3.95 20
4.25 17 4.55 12 4.85 9 5.15 5 5.45  2 5.75  1
;
proc means;
freq f;
var x;
run;
```

The MEANS Procedure

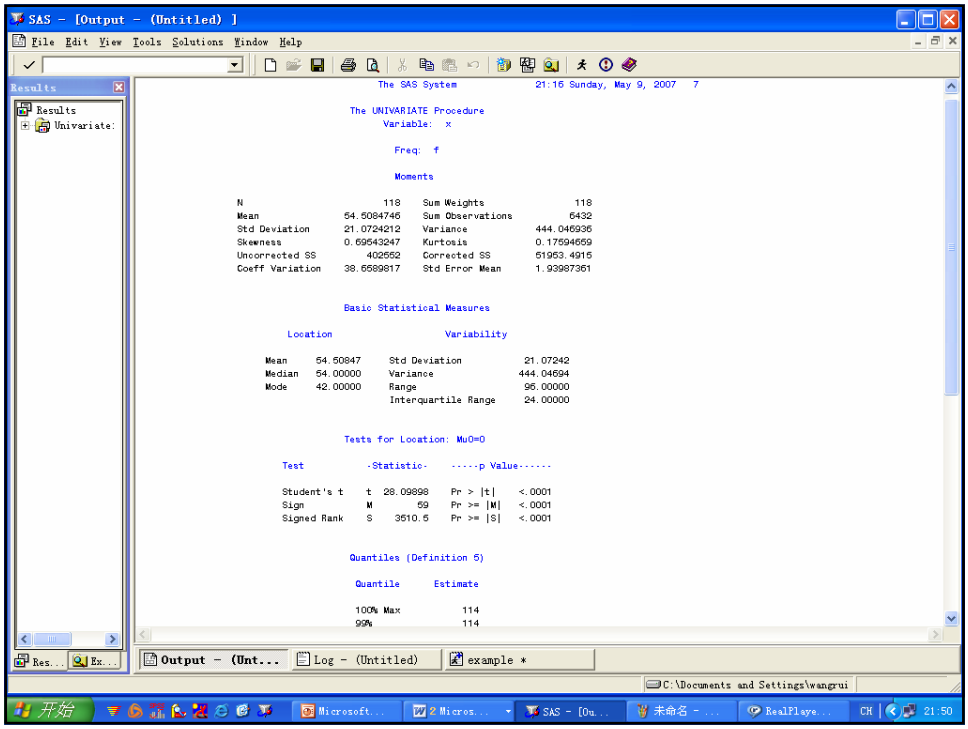
Analysis Variable : x

N	Mean	Std Dev	Minimum	Maximum
101	4.0569307	0.6539507	2.4500000	5.7500000

UNIVARIATE过程

- **UNIVARIATE过程能够给出的描述性统计量比较多，除了上述MEANS过程给出的统计量外，它还能输出符号统计量、正态性检验的统计量以及用户自己定义的百分位数，而且可以生成若干个描述变量分布的茎叶图、箱式图、正态概率图等统计图。**

```
/*例4.6*/
data prg4_6;
input x f @@;
cards;
18  4 30 17  42 32  54 24 66 18 78 12 90  5 102  4 114  2
;
proc univariate;
var x;
freq f;
run;
```



- **UNIVARIATE过程除了能够给出几个特定的百分位数，还能输出用户自己定义的百分位数。此时在过程中要使用output语句。**

程序4-7

```
proc univariate data=prg4_6;
var x;
freq f;
output out=pct pctlpre=p pctlpts=2.5 97.5;
run;
proc print data=pct;
run;
```

运行结果：

OBS	P2_5	P97_5
1	18	102

正态性检验

- **在“proc univariate”后面加上“normal”和“plot”选项，就能输出该组数据正态性检验的结果和茎叶图、箱式图及正态概率图。**

程序4-8

```
proc univariate normal plot;
var x;
run;
```

本章重点内容

1、正态分布曲线下面积的分布特征

2、标准误的计算及标准误与标准差的区别

3、参考值范围和可信区间的计算及区别