《医学统计学》实习指导

（适用于临床医学专业五年制，总学时 52；实习 14 学时）
(试用本)

华中科技大学同济医学院
公共卫生学院流行病与卫生统计学系
2006-3-25

目　录

实习一　SPSS简介　　返回

目的：让学生对 SPSS 有较全面了解，为后面的实习打基础。

要点

一、SPSS 主要窗口及其功能

SPSS 三大窗口：**数据编辑窗（Data Editor）**、**结果输出窗（Viewer）**和**程序编辑窗（Syntax Editor）**。

二、数据文件的建立与读入

SPSS 所处理的数据文件有两种来源：

（1）**在 SPSS 环境下建立数据文件**

（2）**调用已建立的数据文件**

SPSS 能调用 SPSS（*.sav），　Excel（*.xls），dBASE（*.dbf），ASCII（*.dat，*.txt）等数据文件，详细过程可参阅其它参考书。

（3）**数据存储**　SPSS 可将数据存为 SPSS（*.sav），Excel（*.xls），dBASE（*.dbf），ASCII（*.dat，*.txt）等数据文件形式。

（4）**SPSS 的文件类型与主要按钮**　文件类型主要有：数据文件，扩展名为“.sav”；结果文件，扩展名为“.spo”；图形文件，扩展名为“.cht”；程序文件，扩展名为“.sps”。

主要按钮功能：**OK**：执行已选择的操作；**Paste**：将语句命令粘贴到语句命令窗中；**Reset**：重新设置选项；**Cancel**：取消；**Help**：帮助。

三、数据文件的整理与转换

（一）数据文件的整理

（1）定义时间（Define Dates…）

（2）到某一记录（Go to Case…）

（3）插入变量（Insert Variable）与删除（Delete Variable）

（4）插入记录（Insert Case）与删除记录（Delete Case）

（5）观测值排序　（Sort Cases）

（6）选择观察单位（Select Cases）
（7）数据转置（Transpose）
（8）拆分文件（Split Files）
（9）合并文件（Merge Files）
（10）数据分类汇总（Aggregate Data）
（11）变量加权（Weight Cases）

（二）数据文件的转换

（1）计算产生变量（Compute…）
（2）随机数种子（Random Number Seed…）
（3)多个变量中定义观察值的计数统计（Count）数据转换中的 **Count** 功能可产生新的变量，以表示各观察单位中某一（些）观察值的个数。
（4）重新赋值（Recode）
（5）. 连续型变量转换为分类变量（Categorize Variables…）
（6）观察单位排秩（Rank Cases）
（7）自动重新赋值（Automatic Recode）
（8）产生时间序列变量（Create Time Series…）
（9）缺失值的替代（Replace Missing Values）

四、如何得到 SPSS 软件的帮助

（1）单击帮助菜单
（2）在不懂的地方单击右键
（3）上网 google、百度搜索

实习二　频数表、定量资料描述　　返回

一、目的要求

1. 掌握频数表的编制方法和用途。
2. 掌握平均数的意义,应用和计算方法。
3. 掌握几种常用的离散程度指标的意义和应用,并熟悉其计算。
4. 熟悉正态分布及 t 分布的特征和二者之间的关系,掌握正态分布曲线下面积分布的规律和应用。
5. 掌握置信区间的计算及意义。
6. 掌握不同设计类型 t 检验。

二、讨论题

1. P.326/2, 编制频数表：取几组,组段值怎样,组距多少?
2. 数值变量资料频数表的组段是否越细越好? 你接触过的频数表中,有组距不等的吗? 等距有何好处?
3. 对称分布资料中,尤其正态分布资料, 常用均数描述其平均水平, 而不用中位数, 是否因为中位数不能正确描述其平均水平? 如果不是这个原困, 应怎样解释这个选择?
4. 某医生用直接法和加权法计算一组数据的均数和标准差, 发现结果不相同, 检查没发现错误, 哪个结果好?
5. 非对称分布资料或无端界资料(开口资料), 通常用中位数描述其平均水平,能不能使用均数,为什么?
6. 常用几何均数描述一组对数正态分布资料的平均水平, 用中位数可以吗? 为什么? 用几何均数有何好处?
7. 你学过的用于描述离散趋势的指标有哪几个? 试作比较。
8. 制定正常值范围的方法有哪些? 如何选用?为什么有时用双侧正常值范围, 有时用单侧?
9. 置信区间与正常值范围的区别?

三、最佳选择题

1. 各观察值均加(或减)同一不为 0 的数后____

A 均数不变,标准差改变

B 均数改变,标准差不变

C 两者均不变　　D 两者均改变

2. 用均数与标准差可全面描述____资料的特征

A 正偏态分布　　B 负偏态分布

C 正态分布和近似正态分布

D 对称分布

3. 比较身高和体重两组数据变异程度大小宜采用____

A 变异数(CV)　　B 方差(S^2)

C 极差(R)　　D 标准差(S)

4. 描述一组偏态分布资料的变异度,以____指标较好

A 全距(R)　　B 变异系数(CV)

C 四分位间距($Q_{25\%}$～$Q_{75\%}$)　　D 标准差(S)

5. 正态分布曲线下,从均数（横轴）μ 到 u+1.96 倍标准差的面积为____

A 95%　　B 45%　　C 97.5%　　D 47.5%

6. 若正常成人血铅含量近似对数正态分布,拟用 300 名正常成人血铅值确定 99%参考值范围, 最好采用____

A $\overline{X}\pm 2.58S$　　B $lg^{-1}(\overline{x}_{lgX}+2.58S_{lgX})$

C $P_{99}=L+\frac{i}{f_x}(\frac{300\times 99}{100}-\Sigma f_L)$

D $lg^{-1}(\overline{X}_{lgX}+2.33S_{lgX})$

作业: P326: 第 2 题、第 4 题

要点

SPSS 操作：

1. Analyze __ Descriptive Statistics __ Frequencies...(在 Statistics...选项中选所需统计的指标，在 Charts...选项中选 Histograms 可作直方图)
2. Analyze __ Reports __ Case Summaries... (在 Statistics...选项中选 Geometric Mean 可获得几何均数)

1. 频数表的编制

Range(max-min) → length of the interval(10-15) → set groups → counting the numbers → Histogram

（以便观察资料的分布；容易估计集中趋势与变异性；发现异常值）

2. 描述分布的指标

1) average(平均数)：the position of the distribution or central tendency

mean(均数)　　geometric mean(几何均数)

median（中位数）　mode(众数)

[percentile(百分位数),sum(总和)]

2) variability(变异指标)：dispersion tendency (the shape of the distribution)

rang(全距)　standard deviation(SD.标准差)

variance(方差)　　interquartile(四分位间距　P75-P25)

coefficient of variation(变异系数)

[standard error(SE.标准误)]

3. Normal(Gauss) distribution(正态分布)

特点：均数位于曲线正中，两侧完全对称，两端永远不与横轴相交，呈钟型。总体参数为均数（μ）、方差（σ^2）。

曲线下面积的分布概率：

（±1 标准差＝68.27%，±1.96 标准差=95%，±2.58 标准差＝99%）

4 参考值范围

正态分布法

双侧(two side) $\overline{X} \pm u_{\alpha/2} S$；单侧 $\overline{X} - u_{\alpha} S$，$\overline{X} + u_{\alpha} S$

百分位数法（95%）

单侧（one side）　　双侧(two side)

>P5 或<P95　　P2.5 – P97.5

5 *t* 分布

样本均数服从正态分布，当总体标准差未知，以样本均数的标准差（即标准误）取代，正态变换就变成 *t* 变换。

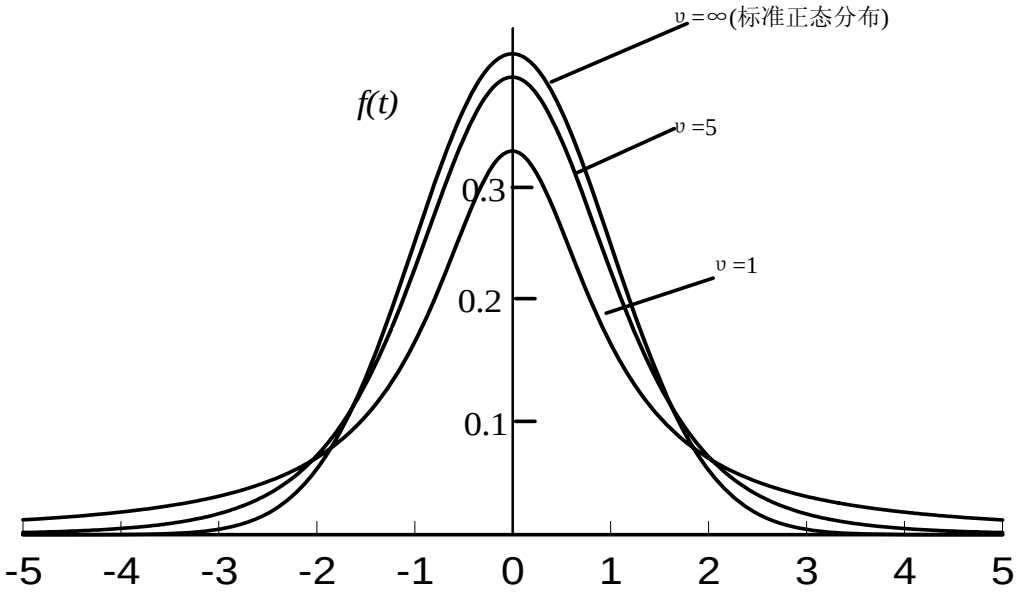


图　标准正态分布和 t 分布的图形

ν =∞时的 t 分布即标准正态分布

t 分布

参数只有自由度，自由度很大时，*t* 分布就逼近正态分布。

实习三　定量资料的统计推断　　返回

一、目的要求

1. 掌握正态分布与 t 分布的图形特征
2. 掌握正常值范围的确定
3. 掌握抽样误差的来源
4. 掌握可信区间的意义及可信区间估计

二、讨论内容

1. 抽样误差是如何产生的？
2. 正态分布的图形有何特征？
3. t 分布的图形有何特征？与正态分布有何关系？
4. 制定正常值范围的方法有哪些？如何选用?为什么有时用双侧，有时用单侧?
5. 标准差与标准误有何区别与联系?

三、多选题

(1)____小，表示用该样本均数估计总体均数的可靠性大.

A. CV　B. S　C. $\sigma_{\bar{X}}$　D. R

(2) 正态分布N(μ，σ^2)下，在μ~μ+1.96σ范围内曲线下面积为:

A. 0.95　B. 0.05　C. 0.45　D. 0.475

(3) 假定某人群的智商服从正态分布,均值为 100,标准差为 15,则此人群中

(A) 智商高于 129.4 的约占 2.5%

(B) 智商高于 100 的约占一半

(C) 智商低于 70.6 的约占 5%

(D) 智商介于 70.6 和 129.4 之间的约占 95%

(4)对于 t 分布，正确的是:

A. 变量取同一数值时，t 分布双侧尾部面积大于标准正态分布双侧尾部面积

B. t 分布在 0~1.96 范围内曲线下面积大于标准正态分布在 0~1.96 范围

内曲线下面积

C. 对于同一自由度，单侧尾部面积为 0.05 时所对应的 t 值小于双侧尾部面积为 0.05 的所对应的 t 值

D. t 分布的标准差小于 1

(5)用大量来自同一总体的独立样本对总体参数作估计时，关于 95%可信区间(CI)，正确的说法是:

A. 大约有 95%的样本的 CI 覆盖了总体参数

B. 对于每一个 CI 而言，总体参数约有 95%的可能性落在其内

C. 各个样本的 CI 是相同的

D. 对于每一个 CI 而言,有 95%的可能性覆盖总体参数

作业：P327　第 5 题、第 8 题、第 10 题

要点：　抽样实验

从总体中随机抽取若干份样本，其均数往往不等于总体均数，且均数之间也存在差异。这种由于个体的差异，由抽样造成的样本均数与总体均数的差别，称为样本均数的抽样误差(sampling error)。

1. 正态总体样本均数的分布

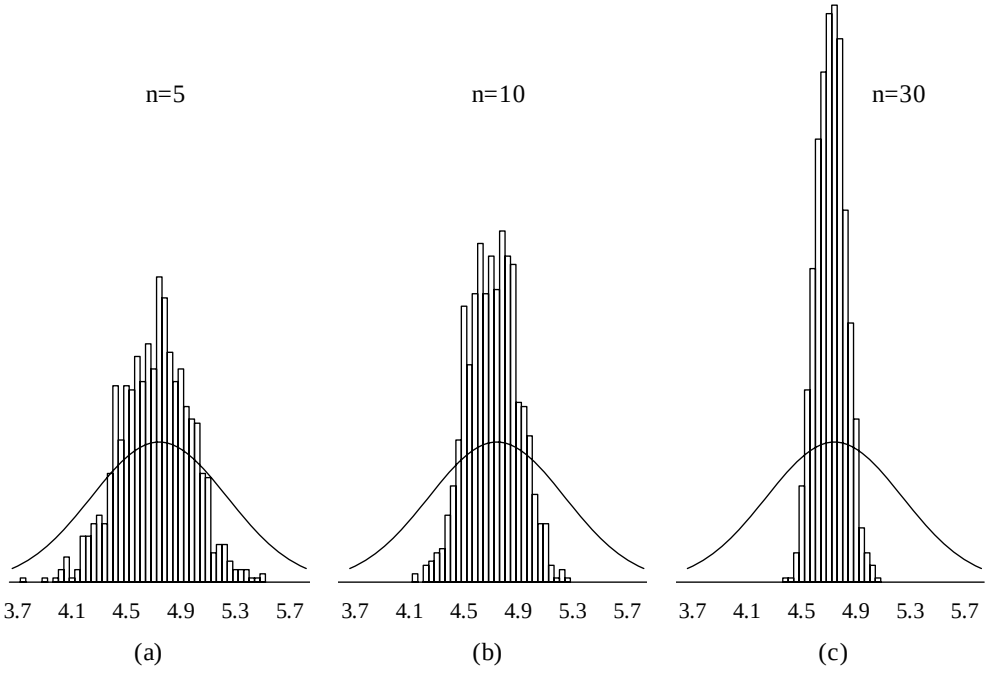


图 1　从正态分布总体抽样的实验结果

曲线是原正态总体曲线为$N(4.6602, 0.5746^2)$的分布密度；

直方图是从总体中抽取的样本均数的分布

实验 1　从正态分布总体抽样的实验　假定正常男子的红血球计数(10^{12}/L)服从图 1 所示的正态分布N(4.6602, 0.5746^2)，运用电脑从这个总体中随机抽取 1000 份样本，每份样本含n＝5（10、30 个个体）。

实验结论：

(1)样本均数的分布也为正态分布。

(2) 样本均数的**变异范围**较原变量的变异范围大大缩小。

(3) 随着样本量的增大（5→10→30），样本均数的变异范围逐渐缩小。

为与原变量的标准差相区别，样本均数的标准差习惯上又称为样本均数的标准误(standard error)，简称标准误。值得注意的是如下的普遍规律：

$$\text{样本均数的标准误} = \text{总体标准差} / \sqrt{n}$$

$$\text{或} \qquad \sigma_{\bar{x}} = \sigma / \sqrt{n}$$

实际应用中往往总体标准差σ未知，人们只能用样本标准差 S 代替σ，从而获得 $\sigma_{\bar{x}}$ 的估计值 $S_{\bar{x}}$，则有

$$S_{\bar{x}} = S / \sqrt{n}$$

表 1　从N(4.6602, 0.5746^2)中随机抽样，样本量为 5, 100 份独立样本的均数、标准差和总体均数的 95%置信区间(单位：10^{12} /L)

样本号	均数	标准差	95%置信区间	样本号	均数	标准差	95%置信区间
1	5.00	.5688	4.2939, 5.7062	51	4.48	.4006	3.9827, 4.9773
2	4.72	.3470	4.2891, 5.1509	52	4.32	.5487	3.6388, 5.0012
…	…	…	…	…	…	…	…
48	4.76	.5837	4.0354, 5.4846	98	4.36	.3368	3.9419, 4.7781
49*	4.04	.3595	3.5937, 4.4863	99	4.56	.6197	3.7907, 5.3293
50	4.52	.6094	3.7634, 5.2766	100	4.60	.4566	4.0331, 5.1669

*由这份样本估计的 95%置信区间实际上并未复盖总体均数。

2. 非正态总体样本均数的分布

实验 2　从正偏峰的分布总体抽样的实验　图 2(a)是一个正偏峰的分布，用电脑从中随机抽取 1000 份样本并计算样本均数。图 2(b), (c), (d)

和(e)分别是样本量为 5, 10, 20 和 30 时样本均数的直方图。可以看出：

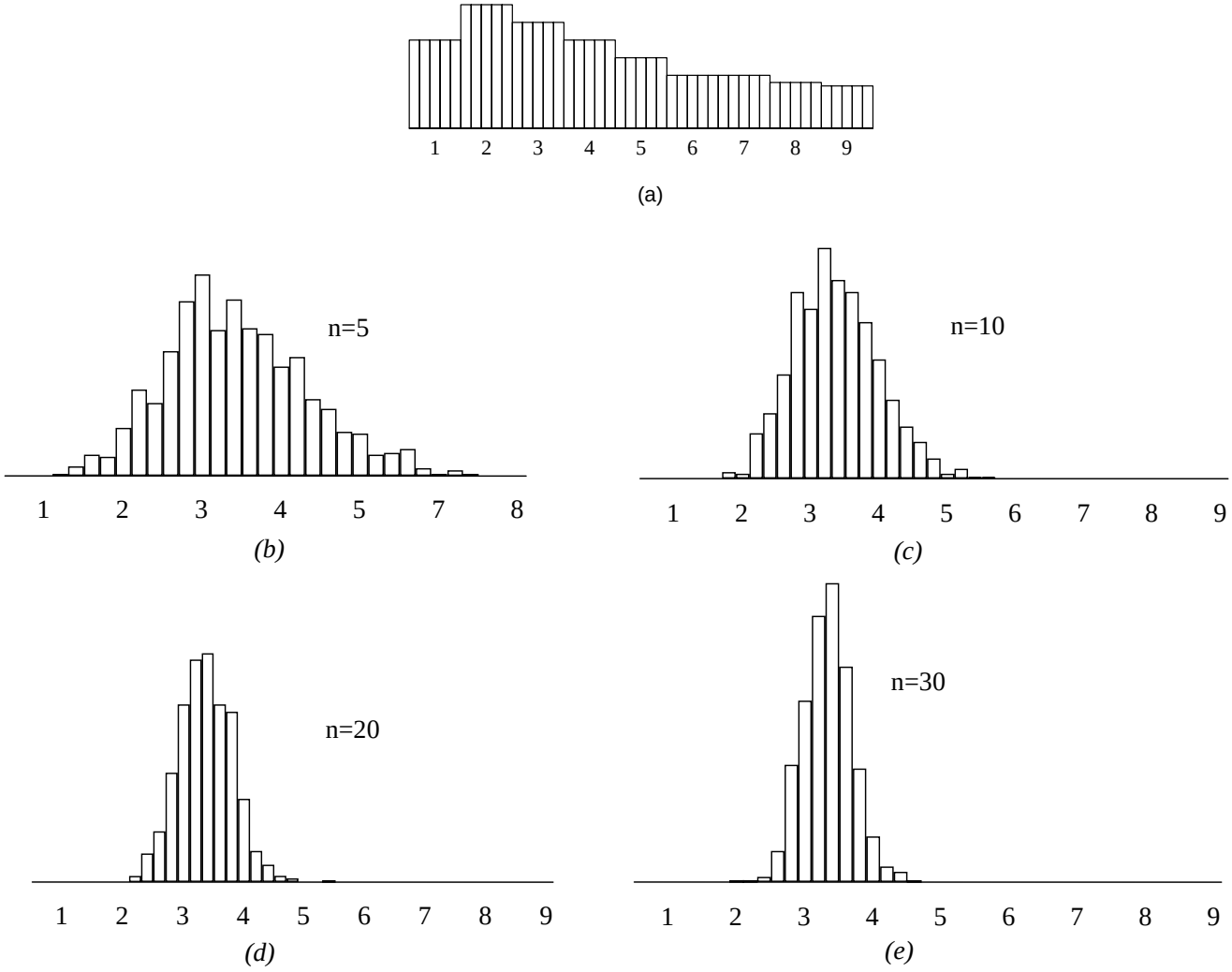


图 2　从正偏峰的分布总体分布抽样实验的结果

(a)是原分布,正偏峰；其它为不同样本含量时样本均数的直方图

实验结论

(1) 随着样本量的增大，样本均数分布的对称性逐渐改善，样本量为 30 时，样本均数的分布接近正态分布；

(2) 随着样本量的增大，样本均数的变异范围逐渐变窄。

实验 3　其它偏态分布（不对称钩形分布）的总体抽样的实验　图 3(a)是一个两边高、中间低、不对称的分布。

实验结论：

(1)样本均数的分布再不象个钩子，样本量很小时就象正态分布了；

(2)随着样本量的增大，样本均数的变异范围也逐渐变窄。

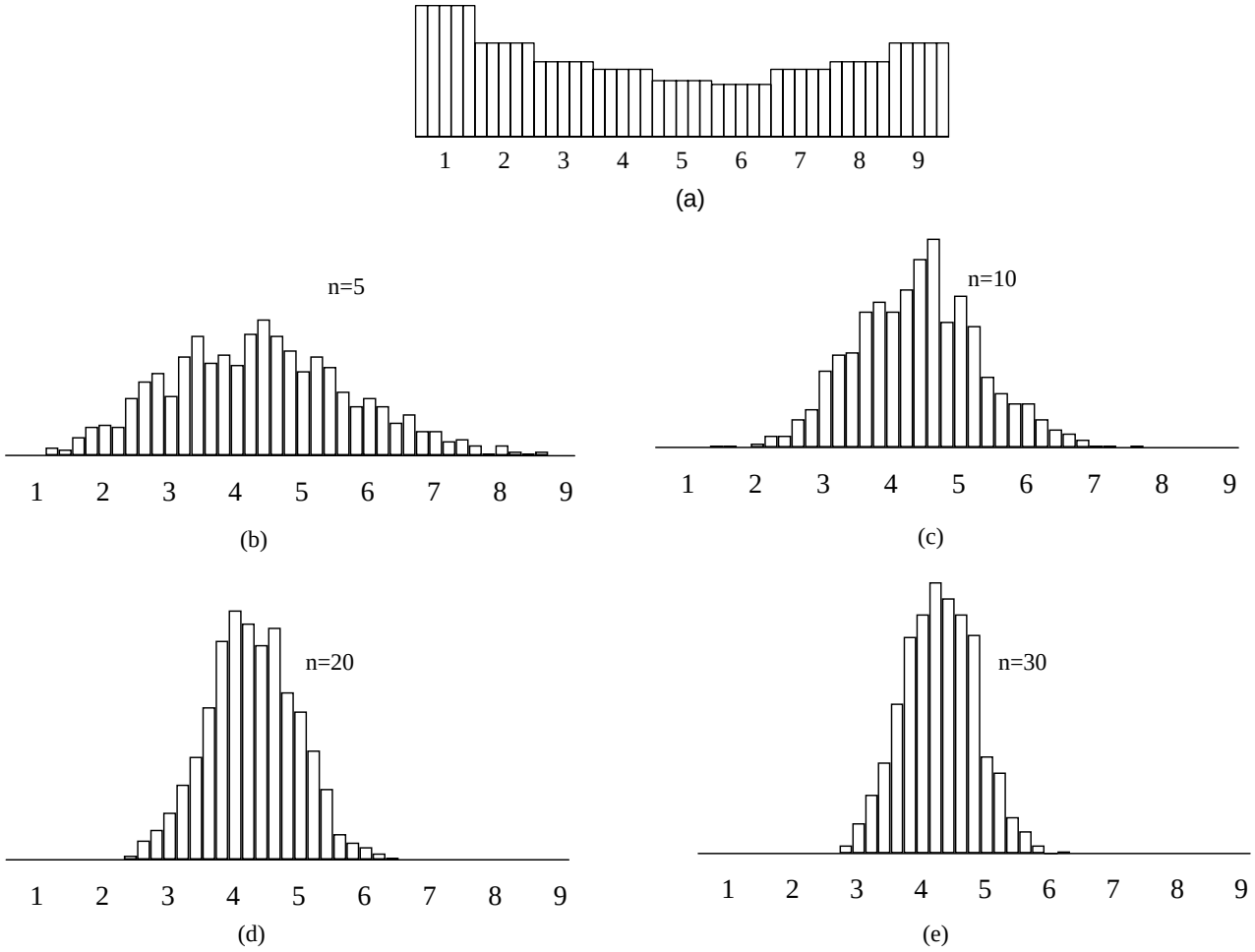


图 3　从不对称钩形分布总体抽样实验的结果
(a)是原分布;(b),(c),(d)和(e)是从总体(a)中抽取样本量为 5,10,20 和 30 的样本时图形

以上实验的结果具有普遍性。理论上可以证明，当样本量较小时，非正态总体样本均数的分布并不是正态分布；但当样本量较大时(例如，**n≥30)**，样本均数的分布接近正态分布（中心极限定理**central limit theorem**）；标准误仍然是原总体标准差的$1/\sqrt{n}$倍。

计量资料的统计推断(*t* 检验)

spss 操作

1. Analyze__Compare Means___ One－Sample T Test…

样本与总体均数比较的 t 检验

2. Analyze__Compare Means___Paired－Sample T Test…

配对 t 检验

3. Analyze__Compare Means__Independent － Sample T Test…

两个独立样本均数比较 t 检验

一、统计推断三步骤

1. 建立检验假设（H_0 、H_1），确定单双侧与检验水准（一般$\alpha = 0.05$）
2. 计算统计量（见 t 检验）
3. （查表确定 P 值，SPSS 等统计软件会直接给出）下结论。

注意：t 值越大，P 越小。

二、三种 t 检验方法：

1. 样本与总体均数的比较

（Analyze ⟶ Compare Means ⟶ One－Sample T Test…）

$$t = \frac{\overline{X} - \mu}{s_{\overline{x}}} = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

自由度＝n-1

2. 配对 t 检验

（Analyze ⟶ Compare Means ⟶ Paired－Sample T Test…）

$$t = \frac{\overline{d}}{s_{\overline{d}}} = \frac{\overline{d}}{s_d / \sqrt{n}}$$

自由度＝n-1（n 为对子数）

3. 两个独立样本均数 t 检验

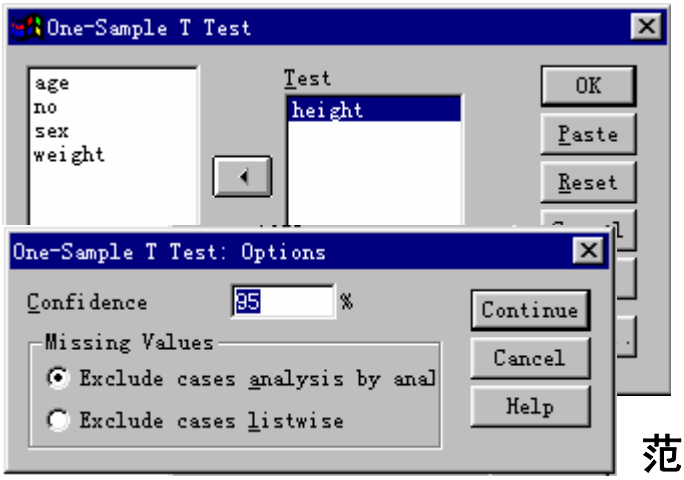
（Analyze ⟶ Compare Means ⟶ Independent－Sample T Test…）

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_c^2 (1/n_1 + 1/n_2)}}$$

$$s_c^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}, \qquad \nu = n_1 + n_2 - 2$$

三、用 spss 实现

1. One-Sample T Test 过程

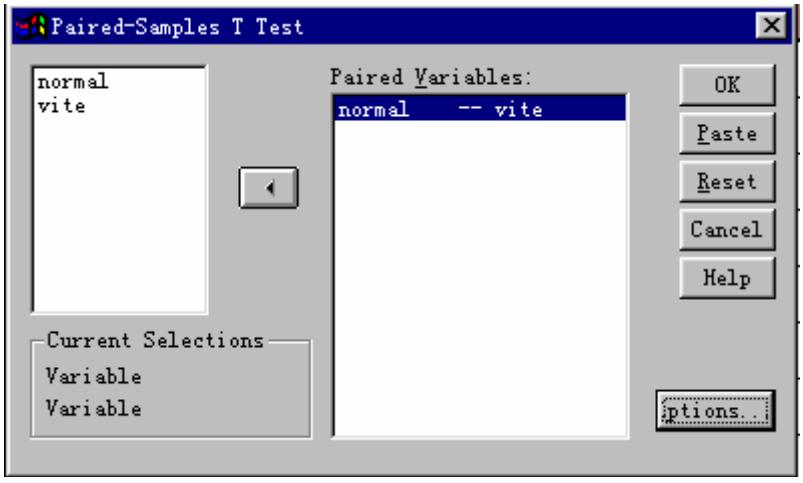


这过程主要做样本均数与已知的总体均数比较的假设检验，当然也可以做配对资料差值的均数与总体均数为零比较的假设检验。

当输好数据后，从菜单中选择该过程，再选择所需分析的变量及与之比较的总体均数。在选择项(Options)确定要估计的置信区间范围和对缺失值的处理方法。

normal	vite
3550	2450
2000	2400
3000	1800
3950	3200
3800	3250
3750	2700
3450	2500
3050	1750

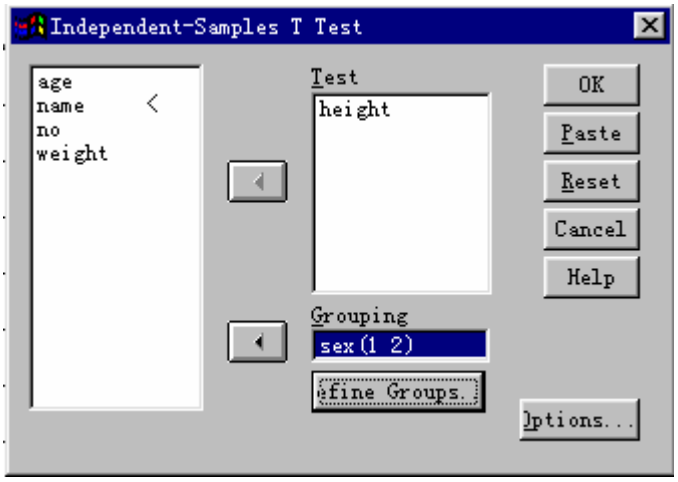
2、Paired-Sample T Test 过程



该过程做两样本配对设计资料的均数比较，用两变量代表配对的两组，两变量同时选入窗口。在选择项(Options)确定要估计的置信区间范围和对缺失值的处理方法。

3. Independent-Samples T Test 过程

该过程主要用于两独立样本资料的假设检验。当输好数据后，从菜单中选择该过程，再选择所需分析的测量变量及分组变量，定义分组的值。在选择项(Options)确定要估计的置信区间范围和对缺失值的处理方法。



实习四　定性资料的统计推断　　返回

一、目的要求

1. 理解χ^2检验的基本思想, 熟悉χ^2检验的用途.
2. 掌握几种常见资料的χ^2检验方法
3. 理解秩和检验的概念及其优点.
4. 掌握几种不同设计类型资料秩和检验的编秩方法和统计量的选用.

二、讨论内容

1. 简述χ^2值的意义.

2. 何时须作χ^2值的连续性校正?

3. 如何正确选用教材中的χ^2统计量计算公式?

4.作多个样本率比较的χ^2检验, 若结论是拒绝H_0, 应如何正确报告结果?

5. 某医生收治了100例临床确诊的小儿佝偻病患者, 入院时均分别作血生化检查与Ξ光片检查,欲了解此两法何者较敏感, 试设计一整理表,并指出宜作何统计处理?

6. 用两种方法检查已确诊的乳腺癌患者120名。甲法的检出率为60%，乙法检出率为50%，甲乙两法阳性一致的检出率为35%，问：两种方法何者为优?

		乙		
		+	−	合计
甲	+	42	30	72
	−	18	30	48
	合计	60	60	120

7. p330/21

8. 什么是秩和检验,有什么优缺点?

9. 两样本比较的假设检验,符合τ检验条件,与τ检验相比, 秩和检验:

A　更易拒绝

B　更不易拒绝

C　检验结果与　检验相同

D　检验结果与　检验不同

E　以上都不对

10. 两样本比较的秩和检验.

A　要求两样本大小相同

B　两样本混合编秩　相同值取平均秩

C　两样本混合编秩　零值舍去不要

D　样本较大时　计算　统计量　属参数检验

E　以上都不对

11. 两样本比较,你选用τ检验还是秩和检验,依据是什么?

12. p331/25.

问:(1)三种矽肺患者的肺门密度分级有无程度上的差别

(2)三种矽肺患者的肺门密度的分级构成比有无差别

要点：　定性资料的统计推断

相对数与率的抽样误差

SPSS 操作：复习 Tansform→Compute 使用

1. 由统计量获得 P 值（Cumulative Distribution Functions, CDF）

CDF.T(q,df); CDFNORM(zvalue) CDF.NORMAL(q,mean stddev)

默认：单侧（从左到右累积概率），

单侧 p=1-cdf.*(*);

双侧 p=（1-cdf.*(*)）*2;

2. 由 P 值或检验水准 α 获得临界值（Inverse Distribution Functions, IDF）

IDF.T(p,df)　IDF.NORMAL(p,meanstddev)

默认：单侧左尾临界值，

单侧 q=-idf.*(p,*);

双侧 q=-idf.*(p/2,*)

absolute number(绝对数)

relative number（相对数）:

rate(率)、proportion(构成比)、ratio(相对比)

Attention :

a. The denominator (分母) is required big enough.
b. Distinguish between rate and proportion.
c. The inner constituent of two populations should be similar.
d. Sampling error should be estimated.

Standardization of rate

county A and county B

Age	County A			County B			standard population proportion
	Population proportion	Mortality rate		Population proportion	Mortality rate		
0-	0.6555	7.4	5.0	0.6949	6.0	4.1	**0.6758**
30-	0.1150	132.0	15.7	0.1226	116.6	13.9	**0.1189**
40-	0.0859	242.9	19.7	0.0765	259.4	21.0	**0.0810**
50-	0.0618	285.2	16.6	0.0549	291.7	17.0	**0.0582**
60-	0.0431	323.9	11.9	0.0309	333.3	12.3	**0.0368**
70-	0.0387	172.8	5.0	0.0202	207.5	6.1	**0.0292**
Total	1.0000	79.2	74.0	1.0000	68.8	74.3	**1.0000**

SE of P

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}, s_p = \sqrt{\frac{p(1-p)}{n}}$$

confidence interval（正态近似法）

100(1-α)%CI= $p \pm Z_{\alpha} s_p$

95%CI= $p \pm 1.96 s_p$

99%CI= $p \pm 2.58 s_p$

Z 检验

1. Comparison of sample proportion and population proportion

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.316 - 0.2}{\sqrt{\frac{0.2(1-0.2)}{152}}} = 3.58$$

2. Comparison of two sample proportions

$$Z = \frac{p_1 - p_2}{s_{p_1 - p_2}} = \frac{p_1 - p_2}{\sqrt{p_c(1-p_c)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$= \frac{0.2875 - 0.1529}{\sqrt{0.2182(1-0.2182)(\frac{1}{80} + \frac{1}{85})}} = 2.09$$

以上设 $p_1 = X_1 / n_1,\ \ p_2 = X_2 / n_2$

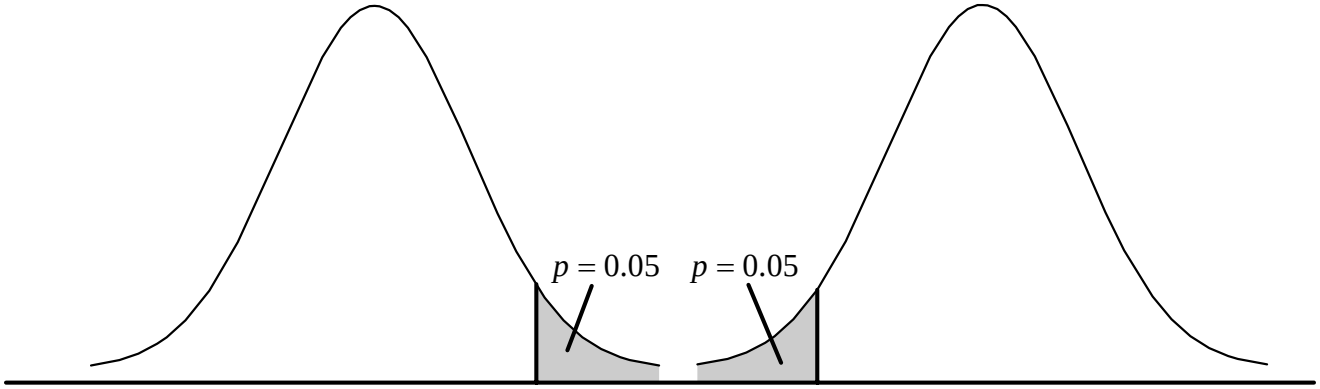
$$p_c = (X_1 + X_2)/(n_1 + n_2)$$

合并率（平均率）

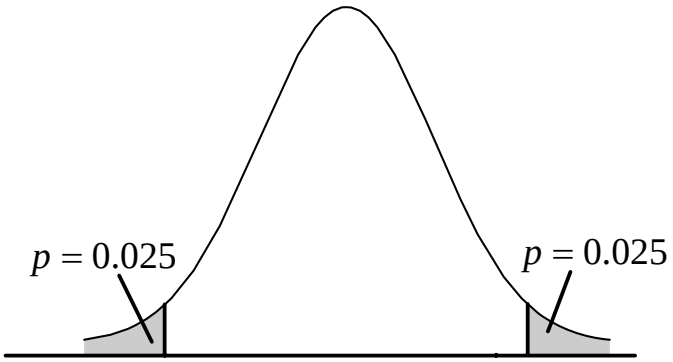
$$= (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$$

(以上总体率的可信区间与 U 检验均假设样本量很大，率服从标准正态分布，即

$n\pi > 5$ 且 $n(1-\pi) > 5$）



(a) 单侧检验 t 曲线下面积示意图



(b) 双侧检验 t 曲线下面积示意图

第八章 χ^2 检验

SPSS: Analyze__Descriptive Statistics__
Crosstabs(注意在 Statistics 中选 Chi-square)

用途：比较两个或多个总体率、构成比是否相等（通常假设某因素与疾病无关）。

一、 基本公式

$$Pearson\,Chi-square\quad \chi^2=\Sigma\frac{(A-T)^2}{T}$$

$$T=\frac{n_R n_C}{n},\quad \nu=(R-1)(C-1)$$

基本公式的连续性校正公式：

$$Continuity\ \ Correction\quad \chi^2=\Sigma\frac{(|A-T|-0.5)^2}{T}$$

二、 四格表专用公式

Pearson Chi-square

$$\chi^2=\frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)},\quad \nu=1$$

Continuity Correction

$$\chi^2=\frac{(|ad-bc|-n/2)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

方法选择：

1. 如果$n\geq 40$,且$T\geq 5$，则选非校正方法

2. 如果$n\geq 40$,且$1\leq T<5$，则选校正方法

3. 如果$n<40$,或$T<1$，则选Fisher's Exact Test

三、 配对四格表

(McNemar test)，(b+c<40 用矫正公式)

$$\chi^2=\frac{(b-c)^2}{b+c},\quad \nu=1,\ \chi^2=\frac{(|b-c|-1)^2}{b+c}$$

四、　列联表（contingency table）通用公式

$$\chi^2=n\left(\Sigma\frac{A^2}{n_R n_C}-1\right),\quad \nu=(R-1)(C-1)$$

注意：如 20%的格子理论频数小于 5，或 1 个格子小于 1，则应考虑合并、删除、增加样本量等问题。

作业：P330 第 21、22、23 题

第九章　秩和检验

（rank sum test）

SPSS 操作：

Analyze___Nonparametric Tests__

2 Related samples...:符号配对比较；

2 Independent samples...：两样本比较

K Independent samples...：多样本比较

应用范围与注意事项：

1. 总体分布不限（分布未知）
2. 具有一端或两端的值不确定
3. 探索性分析
4. 适合参数法的资料采用非参数法，将损失统计信息

一、符号秩和检验（sign rank test）（Two-Related-Samples Tests）

1. 建立假设，2.求差值，

3.编秩

A. 依差值的绝对值从小到大编秩，标明符号。B. 绝对值相同取平均秩次。C.差值为 0 弃除。

4.求秩和（+，-）

5. 以秩和绝对值较小 T 查表，T 大 P 大。

6. 对子数>25,采用正态近似 Z 检验计算 Z 值。

二、两独立样本 Wilcoxon 秩和检验

（Mann-Whitney U，Two-Independent-Samples Tests）

计算 T 或 Z（统一编秩，平均秩次，选样本量较小者 T）（在表所给界值范围内，P 大，否则…）或 Z 值。

三、多个独立样本 Kruskal—Wallis 法

（SPSS 称 Kruskal-Wallis H，

Tests for Several Independent Samples）

统计量为 H～χ^2，自由度＝处理组数－1，（统一编秩，平均秩次）。

注意备择假设为“分布位置不同或不全相同”

1. 计量资料

2. 频数资料（计数资料）

a. 根据分组资料的合计编秩次范围

b. 平均秩次＝（秩次范围上限+下限）/2

c. 求秩和，有每组的平均秩次乘各类疾病的实际频数。

四． 多个相关样本

（Friedman, Tests for Several Related Samples）

注意计算 *Z* 或 *H* 统计量时，对于相同秩次（tie）须进行矫正。

作业:P331 第 26、27、28 题

实习五　直线相关与回归　　返回

一、目的与要求

1. 掌握直线相关与回归分析的意义和用途.
2. 掌握直线相关与回归分析的统计分析方法.
3. 掌握直线相关与回归的区别与联系及应用注意事项.

二、讨论内容

1.简述直线相关与回归分析的意义和用途.

2.何种情况下可作直线相关与回归分析?

3.改变 X 和 Y 的单位,回归方程和相关系数是否也改变? （回归方程改变，相关系数不变）

4."最小二乘法"的含义是什么?

5.设小学生身高 Y(米)对年龄 X(岁)的回归方程是 $\hat{Y}$ =0.5+0.07X, 则初生婴儿的平均身高是 0.5 米,对吗?（错）

6.总体相关系数 ρ =0, 则总体回归系数 β =0, 对否?（对）

7.经检验认为回归方程有意义,表明两变量间存在因果关系, 对否? （错）

8.直线相关与回归分析中r、b、s_b、$s_{Y.X}$、s_r等的意义是什么?

9. 何种情况下需作等级相关分析?

三. 作业: P332　　第 31 题

要求：1. 绘制散点图。2. 计算相关系数 r,及检验。3.计算 Spearman 等级相关系数。4. 给出身高为 Y 的直线回归方程。

要点：

第十章　直线相关与回归

(linear correlation and regression)

SPSS 操作：

回归：Analyze——Regression——Linear…

【Dependent（Y 应变量）　Independent(s) （X 自变量）】

相关：Analyze——Correlate——Bivariate…

【一般：Pearson　等级相关：Kendall tau-b /Spearman 】

散点图：Graphs——Scatter…——Simple——Define

一、用途与意义

（一）直线回归：1. 反映应变量（Y）与预报因子（X，自变量）间的依存关系，2. 预测：由自变量 X 估算应变量 Y，3. 控制：Y 估算 X。

$$\hat{Y} = a + bX$$

b(regression coefficient 或 slope):X 每增（减）一个单位，Y 平均改变 b 个单位; b>0 表示 Y 随 X 增（减）而增（减），b<0 表示 Y 随 X 增（减）而减（增），b=0 表示直线与 X 轴平行。

采用最小二乘法（least square method）（即令 $\Sigma(Y-\hat{Y})^2$ 最小）求解 b，

$$b = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\Sigma(X-\overline{X})^2} = \frac{\Sigma XY-(\Sigma X)(\Sigma Y)/n}{\Sigma X^2-(\Sigma X)^2/n} = \frac{l_{XY}}{l_{XX}}$$

$$a = \overline{Y} - b\overline{X}$$

a 为截距项（intercept）或常数项（constant），表示 X＝0，对应地 Y 值。a=0 时直线经过原点。

（二）直线相关：用相关系数 r（correlation coefficient）反映一变量（Y）与另一变量（X）间的相互关系，

Pearson 相关系数

$$r=\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2\Sigma(Y-\bar{Y})^2}}=\frac{l_{XY}}{\sqrt{l_{XX}l_{YY}}}$$

Spearman's rho（等级，或非参数）相关系数

$$r_s=1-\frac{6\Sigma d^2}{n(n^2-1)}$$

相同秩次矫正公式$r_s'=\frac{[(n^3-n)/6]-(T_X-T_Y)\Sigma d^2}{\sqrt{[(n^3-n)/6-2T_X][(n^3-n)/6-2T_Y]}}$ Kendall's tau_b

r 没有单位，$-1\le r\le 1$，其绝对值大小反映了相互关系的密切程度，一般样本量较大时，$|r|\ge 0.7$ 为高度相关，$0.4\le|r|<0.7$ 中度相关，$|r|<0.4$ 为低度相关。符号与 b 相同，r>0 表示正相关【Y 随 X 增（减）而增（减）】，r<0 表示负相关【Y 随 X 增（减）而减（增）】，r=0 表示零相关。

二、使用条件

（一） 直线回归

1. X 为选定值，而 Y 为正态分布。

2. X、Y 服从双变量正态分布。

（二） 直线相关

1.Pearson 相关：X、Y 服从双变量正态分布

2.等级（非参数）相关：①不服从双变量正态分布，②总体分布未知，③原始数据用等级表示。

三、对总体回归系数（β）总体相关系数（ρ）作统计推断

	回归系数	相关系数
H_0	β =0	ρ =0
检验统计量	$t_b = \lvert b \rvert / s_b$	$t_r = \lvert r \rvert / s_r$
标准误	$s_b = \sqrt{\dfrac{\Sigma(Y-\hat{Y})^2}{n-2}} \Big/ \sqrt{l_{xx}}$	$s_r = \sqrt{\dfrac{(1-r^2)}{n-2}}$
自由度	**n-2**	**n-2**
备注	$\Sigma(Y-\hat{Y})^2 = \Sigma(Y-\bar{Y})^2 - \dfrac{[\Sigma(X-\bar{X})(Y-\bar{Y})]^2}{\Sigma(X-\bar{X})^2} = l_{yy} - \dfrac{l_{xy}^2}{l_{xx}}$	

四、相关与回归的区别与联系

	直线回归	直线相关
区别	**1.** 依存关系 **2.** 仅 **Y** 正态或 **X/Y** 双正态 **3.** ***b***（直线斜率）	**1.**相互关系 **2. X/Y** 双正态（等级相关不需要） **3. r(**散点接近直线的程度**)**
联系	（资料相同） **1. r/b** 正负号一致。**2.** $t_r = t_b$	

实习六　实验设计　　返回

一、目的要求

1. 掌握实验设计的三要素及基本原则
2. 掌握常用实验设计方法
3. 了解调查设计的内容及常用的调查方法

二、讨论内容

1. 实验设计的要素及基本原则有哪些?
2. 设立对照的意义是什么? 有何原则? 常用的对照形式有哪些?
3. 科学研究中, 是否样本含量越大, 则调查结果越可靠?
4. 估计样本含量的依据有哪些?（见 Excel 样本量计算试验）
5. 常用的实验设计方法有哪些?各有何特点?
6. 常用的调查方法有哪些?
7. 按实验设计的要求和原则,对下列各题加以评述:

(1).1976 年某单位报告了果胶驱铅的疗效观察, 30 名铅中毒者脱离现场后住院治疗,治疗前测得尿铅的均数为 0.116mg/L, 血铅均数为 1.81mg/L, 服用 20 天后再测, 尿铅均数降为 0.087mg/L, 血铅均数降为 0.73mg/L, 说明果胶有较好的驱铅作用.

(2).某单位研究菊花艾叶香预防感冒及空气消毒效果, 对象为某幼儿园分住三个楼的儿童,中楼是中班儿童(160 人), 东楼是小班(80 人),此两楼大燃香; 西楼是大班(160 人), 不燃香为对照组, 结论为该香无预防感冒效果, 但有空气消毒作用(肉汤平板上菌落数较少).

作业: P335/42; P336/46

要点:

第十二章　医学科研设计

SPSS 操作：（复习 Transform 与 Data）

利用函数 **RV.UNIFORM(0,1)** 产生一个新的随机数变量，结合利用 **Data __ Sort Cases…**排序，辅助进行完全随机设计、配对设计、配伍组设计。

实习七　统计表与统计图　　返回

一、目的要求

1. 掌握制作统计表和绘制统计图的基本要求。
2. 学会制作常用统计表和绘制常用统计图。
3. 掌握不同统计图的适用条件

二、讨论内容

1．表示事物内部构成比例大小可选用_____

①圆图　　②直方图

③构成比直条图　　④线图

2．某现象的数量随时间上的变化而变化的趋势，在绘制统计图时应选择

①构成图　　②直条图

③线图　　④直方图

3．统计表与统计计图的标题，根据惯例应该

①放在图、表的上方

②放在图、表的下方

③图的标题在图的下方，表的标题在表的上方

④图的标题在图的上方，表的标题在表的下方

⑤可以任意安排

4．a. 散点图　　b. 条图

c. 百分条图或圆图　　d. 线图

e. 半对数线图　　f. 直方图

g. 统计地图

(1) 绘制某地 1985-1995 年肝癌死亡率的变动趋势,应绘制(d)

(2) 分析胎儿不同出生体重(kg)和围产儿死亡率的关系,宜绘制(a)

(3) 比较甲、乙、丙三地某两种传染病的发病率时，宜绘制(b)

(4) 描述某市、某年、各区、县肝炎患病率的分布，宜绘制(g)

(5) 比较某地 10 年间结核和白喉两病死亡率下降速度，宜绘制(e)

(6) 描述某地某年 210 名健康成人发汞含量的分布，宜绘制(f)

5．统计表、图在表达资料中有何特殊作用？
6．列表的原则和基本要求是什么？
7．为什么半对数线图可描述发展速度的变化？
8．P333-335 的第 34～40。

要点：

第十一章　统计表与统计图

SPSS10.0 操作

统计图的绘制：Graphs__ Bar…(条图、构成条图)、Pie…（圆图）、Line…（线图，半对数线图）、Scatter…(散点图)、Histogram…（直方图）

统计表	统计图
定义： 将统计分析的事物及指标用表格列出。 特点： 1.避免长篇文字叙述，便于阅读和对比分析。 2.数据具体。	定义： 用点的位置，线段的升降，直条的长短或面积的大小等 形式表达统计资料。 特点： 直观、醒目，常给人以深刻印象。

统计表 统计图

标题	简明扼要给出图表的基本内容	
	表上端中央	图下方正中
标目	如有度量单位，应标注	
横	一般为叙述事物的分组或动态变量	
纵	一般为叙述事物的统计指标	
线条	顶、低线，或标目线或合计线	
数字	阿拉伯数字（不用文字）	
	小数位数一致、位次对齐	
	不能有空	
备注	必要时	
图例		放在图右上角或标题的上方

表 各种统计图适用的资料类型与分析目的

图类型	资料性质	分析目的
1. 条图(0)	横轴为间断独立的分组	直条长短表达统计指标大小
2. 构成条图	构成比	长条各段长度(面积)表达构成
3. 圆图	构成比	圆的扇形面积表达构成
4. 线图	横轴年龄,时间等动态变量	统计指标随某一变量(时间)的变化趋势
5. 半对数线图	横轴同上，纵轴取对数	统计指标随某一变量(时间)的发展速度
6. 直方图(0)	横轴为分组，纵轴为频数	直条矩形面积表达各组段的频数(率)
7. 多边图	直方图的变形	同上
8. 散点图	两个变量	用点的密集程度和趋势反映两变量关系
9. 统计地图	地域性资料	点的疏密、颜色的深浅等说明疾病地域分布

宇传华整理
2006-3-25