统计学实训

《Excel 在统计中的应用》

实训指导书

经济与管理学院

刘红红

Excel 在统计中的应用

经济与管理各专业(本科)均开设统计学，总学时 48：理论课时 40、实训课时 8。实训内容:

1. Excel 在数据整理中的应用

2. 用“图表向导”工具绘制统计图表举例

3. Excel 在描述统计中的应用

4. Excel 在抽样推断中的应用

5. 用 **Excel** 进行相关与回归分析

6. 用 **Excel** 计算各种动态分析指标

7. 用 **Excel** 进行时间序列分析

1. Excel 在数据整理中的应用

数据的整理与显示

1. 要弄清所面对的数据类型
    - 不同类型的数据，采取不同的处理方式和方法
2. 对分类数据和顺序数据主要是作分类整理
3. 对数值型数据则主要是作分组整理
4. 适合于低层次数据的整理和显示方法也适合于高层次的数据；但适合于高层次数据的整理和显示方法并不适合于低层次的数据

1.1 品质数据的整理与显示

1.1.1 分类数据的整理与图示

分类数据整理的基本过程：

1. 列出各类别
2. 计算各类别的频数
3. 制作频数分布表
4. 用图形显示数据

分类数据的整理可计算的统计量：

1. 频数(frequency) ：落在各类别中的数据个数.
2. 比例(proportion) ：某一类别数据占全部数据的比值.
3. 百分比(percentage) ：将对比的基数作为 100 而计算的比值.
4. 比率(ratio) ：不同类别数值的比值.

1.1.2 顺序数据的整理与图示

顺序数据的整理可计算的统计量与图示

1.**累积频数(cumulative frequencies)：**各类别频数的逐级.

2.**累积频率(cumulative percentages)：**各类别频率(百分比)的逐级累加.

3.**顺序数据的图示**---环形图(doughnut chart):环形图中间有一个“空洞”，样本或总体中的每一部分数据用环中的一段表示.

1.2 数值型数据的整理与显示

1.2.1 数据分组

1.2.2 数值型数据的图示

分组数据—直方图和折线图(histogram)和(frequency polygon)

Excel：点击“工具——数据分析——直方图”，选择原始数据为输入区域，再接受区域中输入分组栏，即各组的上限-1，即可得到频数分布表和直方图。

未分组数据—茎叶图和箱线图(stem-and-leaf display)和(box plot)

时间序列数据—线图(line plot)

2. 用“图表向导”工具绘制统计图表举例

图形功能举例：

1. 利用图表向导作图：

①条形图 ②饼图 ③环形图 ④直方图 ⑤茎叶图 ⑥箱线图 ⑦趋势线

2. 在图形上添加曲线等附加信息。

表格功能举例：

1. 公式复制时的相对地址与绝对地址

2. 报表汇总—分两种情况

3. 作数据透视表

数据透视表 Excel 制作：

第 1 步：在 Excel 工作表中建立数据清单

第 2 步：选中数据清单中的任意单元格，并选择【数据】菜单中的【数据透视表和数据透视图】

第 3 步：确定数据源区域

第 4 步：在【向导—3 步骤之 3】中选择数据透视表的输出位置，然后选择【布局】

第 5 步：在【向导—布局】对话框中，依次将“分类变量”拖至左边的“行”区域，上边的“列”区域，将需要汇总的“变量” 拖至“数据区域”

第 6 步：然后单击【确定】，自动返回【向导—3 步骤之 3】对话框。然后单击【完成】 ，即可输出数据透视表

3. Excel 在描述统计中的应用

3.1 描述统计量

3.1.1 反映集中趋势的描述统计量

常用的反映集中趋势的描述统计量有三个：均值、中位数和众数。前一个平均数是根据所有标志值计算的，又被称为数值平均数，后两个平均数是根据与其所处位置有关的部分标志值计算的，又被称为位置平均数。

3.1.2 反映离中趋势的描述统计量

常用的反映离中趋势的描述统计量（简称离中指标）有三个：全距（极差）、平均差和标准差（方差）。当对两组数据的差异程度进行相对比较时，往往要计算离散系数，包括全距（极差）系数、平均差系数和标准差系数，它等于相应的离中指标除以均值，这样可以消除由于平均数的不同或单位

的差异而造成的影响。

3.1.2 反映分布趋势的描述统计量

常用的反映分布趋势的描述统计量有两个：偏斜度和峰值

偏斜度:反映以平均值为中心的分布的不对称程度 。

峰度:反映与正态分布相比某一分布的尖锐度或平坦度。

3.2 用 Excel 计算描述统计量

将已知数据输入到 Excel 工作表中，然后按下列步骤操作:

第 1 步：选择【工具】下拉菜单

第 2 步：选择【数据分析】选项

第 3 步：在分析工具中选择【描述统计】，然后选择【确定】

第 4 步：当对话框出现时，在【输入区域】方框内键入数据区域、在【输出选项】中选择输出区域、选择【汇总统计】、选择【确定】

4. Excel 在抽样推断中的应用

4.1 简单随机抽样(用 Excel 对分类数据随机抽样) (以 30 个学生为例):

第 1 步：将 30 个学生的名单录入到 Excel 工作表中的一列

第 2 步：给每个学生一个**数字代码**，分别为 1，2…，30，顺序排列，将代码录入到 Excel 工作表中的一列，与学生名单相对应

第 3 步：选择【**工具**】下拉菜单，并选择【**数据分析**】选项，然后在【数据分析】选项中选择【**抽样**】

第 4 步：在【抽样】对话框中的【输入区域】中输入学生代码区域，在【抽样方法】中单击【**随机**】 。在【样本数】中输入需要抽样的学生个数。在【**输出区域**】中选择抽样结果放置的区域。【确定】后即得到

要抽取的样本

简单随机抽样(用 Excel 对数值型数据随机抽样)

第 1 步：将原始数据录入到 Excel 工作表中的一列

第 2 步：选择【**工具**】下拉菜单，并选择【**数据分析**】选项 ，然后在【数据分析】选项中选择【**抽样**】

第 3 步：在【抽样】对话框中的【输入区域】中输入原始数据区域，在【抽样方法】中单击【**随机**】。在【样本数】中输入需要抽样的数据个数。在【输出区域】中选择抽样结果放置的区域。【确定】后即得到要抽取的样本数据

4.2 总体均值区间估计

设: 是总体 X 的一个样本，X~N(μ，σ^2)，求总体均值μ的置信区间。

1．正态总体、方差σ^2已知，或非正态总体、大样本，求μ的置信区间

构造总体均值μ的置信区间为：$\left\{\bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\bar{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\}$

2．正态总体、方差σ^2未知、小样本，求μ的置信区间

构造均值μ的置信区间为：$\left\{\bar{x}-t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\bar{x}+t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right\}$

例 从某班男生中随机抽取 10 名学生，测得其身高（cm）分别为 170、175、172、168、165、178、180、176、177、164，以 95%的置信度估计本班男生的平均身高。

在 95%的置信度下，本班男生身高的置信区间为（168.5063658，176.4936342）。计算结果如下图所示

	A	B	C
1	学生身高	抽样单位数	10
2	170	样本均值	172.5
3	175	标准差	5.582711408
4	172	标准误差	1.765408357
5	168	置信度	95%
6	165	t值	2.262158887
7	178	极限误差	3.993634204
8	180	估计下限	168.5063658
9	176	估计上限	176.4936342
10	177		
11	164		

总体均值置信区间的计算

4.2 总体比例区间估计

- 样本比例抽样分布的数量特征如下： $\mu_{p_i} = \pi$
- 样本比例抽样分布的标准差为 $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$
- 标准正态分布，确定围绕 π 值的置信区间是：

$$\left\{ p - z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}, p + z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right\}$$

5. 用 Excel 进行相关与回归分析

5.1 相关分析

1.相关分析要解决的问题

- 变量之间是否存在关系？
- 如果存在关系，它们之间是什么样的关系？
- 变量之间的关系强度如何？

- 样本所反映的变量之间的关系能否代表总体变量之间的关系？

2.为解决这些问题，在进行相关分析时，对总体有以下两个主要假定

- 两个变量之间是线性关系
- 两个变量都是随机变量

3. 相关系数 (计算公式)

- 样本相关系数的计算公式 $r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\cdot\sum(y-\bar{y})^2}}$

或化简为 $r=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\cdot\sqrt{n\sum y^2-\left(\sum y\right)^2}}$

5.2 一元线性回归

1. 涉及一个自变量的回归

2. 因变量 *y* 与自变量 *x* 之间为线性关系

- 被预测或被解释的变量称为因变量(dependent variable)，用 *y* 表示
- 用来预测或用来解释因变量的一个或多个变量称为自变量(independent variable)，用 *x* 表示

3. 因变量与自变量之间的关系用一个线性方程来表示

5.2.1 一元线性回归模型 (基本假定)

1. 因变量 *y* 与自变量 *x* 之间具有线性关系

2. 在重复抽样中，自变量 *x* 的取值是固定的，即假定 *x* 是非随机的

3. 误差项 ε 是一个期望值为 0 的随机变量，即 $E(\varepsilon)=0$。对于一个给

定的 x 值，y 的期望值为 $E(y)=\beta_0+\beta_1 x$

4. 对于所有的 x 值， ε 的方差 σ^2 都相同

5. 误差项 ε 是一个服从正态分布的随机变量，且相互独立。

即 $\varepsilon \sim N(0, \sigma^2)$

- 独立性意味着对于一个特定的 x 值，它所对应的 ε 与其他 x 值所对应的 ε 不相关
- 对于一个特定的 x 值，它所对应的 y 值与其他 x 所对应的 y 值也不相关。

5.2.2 回归方程(regression equation)

一元线性回归方程的形式如下:

$$E(y)=\beta_0+\beta_1 x$$

- 方程的图示是一条直线，也称为直线回归方程
- β_0 是回归直线在 y 轴上的截距，是当 x=0 时 y 的期望值
- β_1 是直线的斜率，称为回归系数，表示当 x 每变动一个单位时，y 的平均变动值

5.2.3 估计的回归方程(estimated regression equation)

1. 总体回归参数 β_0 和 β_1 是未知的，必须利用样本数据去估计
2. 用样本统计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 代替回归方程中的未知参数 β_0 和 β_1，就得到了估计的回归方程
3. 一元线性回归中估计的回归方程为: $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$

5.2.4 用 Excel 进行回归分析

第 1 步：选择**“工具”**下拉菜单

第 2 步：选择【**数据分析**】选项

第 3 步：在分析工具中选择【**回归**】，选择【**确定**】

第 4 步：当对话框出现时，在【**Y 值输入区域**】设置框内键入 Y 的数据区域；在【**X 值输入区域**】设置框内键入 X 的数据区域；在【**置信度**】选项中给出所需的数值；在【**输出选项**】中选择输出区域；在【**残差**】分析选项中选择所需的选项。

5.3 相关与回归的显著性检验

■ 线性关系的检验 (检验的步骤)

1.提出假设: H_0：β_1=0　　线性关系不显著

2.计算检验统计量 *F*

$$F=\frac{SSR/1}{SSE/n-2}=\frac{MSR}{MSE}\sim F(1,n-2)$$

3.确定显著性水平 α，并根据分子自由度 1 和分母自由度 *n*-2 找出临界值 F_α

4.作出决策：若 $F>F_\alpha$,拒绝 H_0；若 $F<F_\alpha$,不拒绝 H_0

■ 回归系数的检验 (检验步骤)

1. 提出假设

- H_0: β_1 = 0 (没有线性关系)
- H_1: β_1 ≠ 0 (有线性关系)

2. 计算检验的统计量

$$t=\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\sim t(n-2)$$

3. 确定显著性水平 α，并进行决策

■ $|t|>t_{\alpha/2}$，拒绝H_0；$|t|<t_{\alpha/2}$，不拒绝H_0

6. 用 Excel 计算各种动态分析指标

在 Excel 表中直接输入公式计算:

6.1.增长率(growth rate):也称增长速度.报告期观察值与基期观察值之比减 1，用百分比表示.由于对比的基期不同，增长率可以分为环比增长率和定基增长率.

环比增长率:报告期水平与前一期水平之比减 1.

定基增长率:报告期水平与某一固定时期水平之　　比减 1

6.2.平均增长率(average rate of increase):描述现象在整个观察期内平均增长变化的程度.

$$\overline{G}=\sqrt[n]{\frac{Y_1}{Y_0}\times\frac{Y_2}{Y_1}\times\Lambda\times\frac{Y_n}{Y_{n-1}}}-1=\sqrt[n]{\prod\frac{Y_i}{Y_{i-1}}}-1$$

$$=\sqrt[n]{\frac{Y_n}{Y_0}}-1 \qquad (i=1,2,\Lambda\ ,n)$$

6.3 增长率分析中应注意的问题(增长 1%绝对值):

1. 增长率每增长一个百分点而增加的绝对量
2. 用于弥补增长率分析中的局限性
3. 计算公式为：$\text{增长1\%绝对值}=\frac{\text{前期水平}}{100}$

7. 用 Excel 进行时间序列分析

7.1 时间序列的构成与分解

影响时间序列变动的因素主要有 4 种：

1. 趋势(trend)：持续向上或持续下降的状态或规律

2.季节性(seasonality)：也称季节变动(seasonal fluctuation)

- 时间序列在一年内重复出现的周期性波动

3. 周期性(cyclity)：也称循环波动(cyclical fluctuation)

- 围绕长期趋势的一种波浪形或振荡式变动

4.随机性(random)：也称不规则波动(irregular variations)

- 除去趋势、周期性和季节性之后的偶然性波动

时间序列模型：

1. 乘法模型： $\boldsymbol{Yi=Ti\times Si\times Ci\times Ii}$

2. 加法模型： $\boldsymbol{Yi=Ti+Si+Ci+Ii}$

7.2 移动平均法(moving verage)

简单移动平均法:

1.将每个观察值都给予相同的权数

2.只使用最近期的数据，在每次计算移动平均值时，移动的间隔都为 k

3.主要适合对较为平稳的序列进行预测

4. t 期的**移动平均值***即* t+1 期的**预测值**,*公式*为:

$$F_{t+1}=\overline{Y}_t=\frac{Y_{t-k+1}+Y_{t-k+2}+\Lambda\ +Y_{t-1}+Y_t}{k}$$

移动平均预测误差:

1.有了第 t+1 期的实际值，便可计算出预测误差为:

$$e_{t+1}=Y_{t+1}-F_{t+1}$$

2.预测误差用均方误差(*MSE*) 来衡量

$$MSE=\frac{\text{误差平方和}}{\text{误差个数}}=\frac{\sum_{i=1}^{n}(Y_i-F_i)^2}{n}$$

加权移动平均法(weighted moving average)

1. 对近期的观察值和远期的观察值赋予不同的权数后再进行预测

- 当序列的波动较大时，最近期的观察值应赋予最大的权数，较远的时期的观察值赋予的权数依次递减
- 当序列的波动不是很大时，对各期的观察值应赋予近似相等的权数
- 所选择的各期的权数之和必须等于 1。

2. 对移动间隔(步长)和权数的选择，也应以预测精度来评定，即用均方误差来测度预测精度，选择一个均方误差最小的移动间隔和权数的组合

移动平均法 Excel 操作：

（1）在“工具”菜单中选择“数据分析”选项，在弹出的“数据分析”对话框中选中“移动平均”选项，并单击“确定”按钮，此时将出现“移动平均”对话框。

（2）假定作三项移动:在输入区域中框定 A 列原始数据，间隔设为 3，在输出区域中输入选定位置，即输出区域的左上角的绝对引用。选择“图表输出” 单击“确定”按钮。

7.3 趋势模型分析

1. 趋势(trend)

- 持续向上或持续下降的状态或规律

2. 有线性趋势和非线性趋势

3. 方法主要有

- 线性趋势预测
- 非线性趋势预测

- 自回归模型预测

7.3.1 线性模型法(线性趋势方程)

线性方程的形式为: $\hat{Y}_t = a + bt$

- $\hat{Y}_t$—时间序列的预测值
- t—时间标号
- a—趋势线在 Y 轴上的截距
- b—趋势线的斜率，表示时间 t 变动一个单位时观察值的平均变动数量

7.4 季节变动分析

计算季节指数(seasonal index)

1. 刻画序列在一个年度内各月或季的典型季节特征
2. 以其平均数等于 100%为条件而构成
3. 反映某一月份或季度的数值占全年平均数值的大小
4. 如果现象的发展没有季节变动，则各期的季节指数应等于 100%
5. 季节变动的程度是根据各季节指数与其平均数(100%)的偏差程度来测定

- 如果某一月份或季度有明显的季节变化，则各期的季节指数应大于或小于 100%

存在趋势时的季节指数(计算步骤)：

1. 计算移动平均值(季度数据采用 4 项移动平均，月份数据采用 12 项移动平均)，并将其结果进行“中心化”处理

- 将移动平均的结果再进行一次二项的移动平均，即得出“中心

化移动平均值”(*CMA*)

2. 计算移动平均的比值，也成为季节比率

- 将序列的各观察值除以相应的中心化移动平均值，然后再计算出各比值的季度(或月份)平均值，即季节指数

3. 季节指数调整

- 各季节指数的平均数应等于 1 或 100%，若根据第 2 步计算的季节比率的平均值不等于 1 时，则需要进行调整

具体方法是：将第 2 步计算的每个季节比率的平均值除以它们的总平均值

分离季节因素后的线性趋势预测：

1. 将原时间序列除以相应的季节指数

$$\frac{Y}{S}=\frac{T\times S\times I}{S}=T\times I$$

2. 季节因素分离后的序列反映了在没有季节因素影响的情况下时间序列的变化形态

3. 根据分离季节性因素的序列确定线性趋势方程

4. 根据趋势方程进行预测

- 该预测值不含季节性因素，即在没有季节因素影响情况下的预测值

5. 计算最终的预测值

- 将回归预测值乘以相应的季节指数

7.5 循环变动(剩余法)

1. 先消去季节变动，求得无季节性资料

$$\text{无季节性资料} = \frac{T \times S \times C \times I}{S} = T \times C \times I$$

2. 将结果除以由分离季节性因素后的数据计算得到的趋势值，求得含有周期性及随机波动的序列

$$\text{周期与随机波动} = \frac{T \times C \times I}{T} = C \times I$$

3. 将结果进行移动平均(*MA*)，以消除不规则波动，即得循环波动值

■ $C = MA(C \times I)$