数据分析与预测方法

数据分析方法

原书作者：庄贵军

PPT汇报：何恺

官方网站：何恺说

数据分析方法

进行数据分析时，应该根据数据的特性，选择适当的分析方法，而不是为方法而方法。

目录 数据分析方法

1 单变量数据分析

描述性分析

- 平均值
- 中位数
- 众数
- 标准差
- 四方位差
- 频数

推断性分析

- 区间估计
- 假设检验

2 双变量数据分析

简单相关分析

简单回归分析

方差分析

与之相关的假设检验

3 多变量数据分析

多元相关分析

多元回归分析

因子分析

判别分析

聚类分析

联合分析

结构方程建模分析

与之相关的假设检验

单变量

描述性分析

描述性分析的目的在于对样本所有元素在某一方面（变量）的观察值进行概括性的描述。可以从描述样本数据的中心趋势和描述样本数据的离散程度两方面进行。

单变量

描述性分析——中心趋势

1 平均数

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

式中：$\bar{x}$表示样本均值； n表示样本单位数或样本容量。

2 中位数

中位数是数据按大小顺序排序之后，位置在最中间的数值。

中位数性质：数据值与中位数之差的绝对值之和最小，即

$$\sum_{i=1}^{n} |x_i - M_e| = min$$

3 众数

众数是将数据按大小顺序排序形成次数分配后，在统计分布中具有明显集中趋势点的数值，是数据一般水平代表性的一种。

单变量

描述性分析——离散程度

1 标准差

样本的标准差：

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

式中：S表示样本标准差；$\bar{x}$表示样本均值；n表示样本容量；$(\mathrm{n}-1)$称为自由度。

当样本单位数较大时(n≥30)，自由度变为n。

补充：样本标准差的自由度为什么是(n-1)?

答：计算样本标准差S的目的除了分析样本数据外，还要估计总体标准差σ。

数学计算可以证明，以自由度为(n-1)，计算的标准差S的数学期望就是σ，因而是总体标准差σ的无偏估计。

单变量

描述性分析——离散程度

2 四分位差

四分位差是顺序数据中处于75%位置上的数据与处于25%位置上的数值之差。

四分位差基本不受极端值的影响。

3 频率

在相同条件下，进行了n次试验，在这n次试验中，事件A发生的次数n_A成为事件A发生的频数；比值$\frac{n_A}{n}$称为事件A发生的频率。

单变量

推断性分析——单个正态总体参数的区间估计

基本原理

总体参数区间估计的基本原理是，根据给定概率保证程度的要求，利用实际抽样资料，指出总体估计值的上限和下限，即指出总体参数可能存在的区间范围。

基本方法

设 $X\sim N(\mu,\sigma^2)$，$X_1,X_2,\cdots,X_n$ 为总体的一个样本。

1. 均值 μ 的区间估计

由于 $\overline{X}$ 是 μ 最小方差无偏估计量，因此在没有其他信息的情况下，μ 应该在 $\overline{X}$ 附近，

μ 的置信区间应为 $(\overline{X}-C,\overline{X}+C)$，即 $P(\overline{X}-C<\mu<\overline{X}+C)=1-\alpha$，区间估计转为确定常数 C.

单变量

推断性分析——单个正态总体参数的区间估计

基本方法

(1) 当 σ^2 已知时，$u=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$，根据标准正态分布的密度函数关于 y 轴对称，密度函数与 x 轴围成的面积为 1，左右两边各去掉 $\dfrac{\alpha}{2}$ 的面积，$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ 的密度函数在区间 $(-u_{1-\frac{\alpha}{2}},u_{1-\frac{\alpha}{2}})$ 内与 x 轴的面积为 $1-\alpha$(见图 2.1)，即

$$P\left(-u_{1-\frac{\alpha}{2}}<\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}<u_{1-\frac{\alpha}{2}}\right)=1-\alpha,$$

由此得到

$$P\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}<\mu<\overline{X}+\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right)=1-\alpha,$$

μ 的置信区间为

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}},\overline{X}+\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right).$$

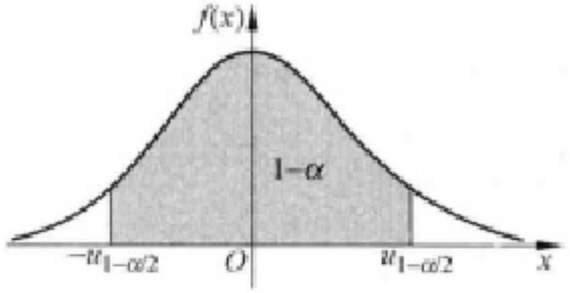

图 2.1

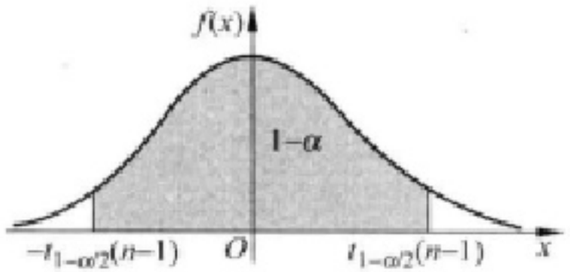

图 2.2

单变量

推断性分析——单个正态总体参数的区间估计

基本方法

(2) 当 σ^2 未知时，我们应该选择 $T=\dfrac{\overline{X}-\mu}{\dfrac{S}{\sqrt{n}}}\sim t(n-1)$，$T$ 分布的密度函数也关于 y 轴对称，与 x 轴围成的面积为 1，左右两边各去掉 $\dfrac{\alpha}{2}$ 的面积，$\dfrac{\overline{X}-\mu}{\dfrac{S}{\sqrt{n}}}$ 的密度函数在区间 $(-t_{1-\frac{\alpha}{2}}(n-1),t_{1-\frac{\alpha}{2}}(n-1))$ 上与 x 轴围成的面积为 $1-\alpha$，见图 2.2，即

$$P\left\{-t_{1-\frac{\alpha}{2}}(n-1)<\frac{\overline{X}-\mu}{\dfrac{S}{\sqrt{n}}}<t_{1-\frac{\alpha}{2}}(n-1)\right\}=1-\alpha,$$

$$P\left(\overline{X}-\frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)<\mu<\overline{X}+\frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\right)=1-\alpha,$$

所以 μ 的置信区间为 $\left(\overline{X}-\dfrac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1),\overline{X}+\dfrac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\right)$.

单变量

推断性分析——单个正态总体参数的区间估计

示例

某市常住居民的人均年食糖需要量的区间估计

某市常住居民为70万人。现采用简单随机方法抽样，对该市常住居民人均年食糖需要量进行调查。已知人均年食糖需要量为5.6公斤，共抽取1400人进行调查，样本方差为40.46。如果允许人均年食糖需要量误差为0.34公斤，请问该市常住居民年食糖需要量的置信区间和置信概率各多少？

解：

$$5.6-0.34\le\mu\le5.6+0.34$$

$$5.26\le\mu\le5.94$$

$$5.26\times700000\le C\le5.94\times700000$$

$$3682000\le C\le4158000$$

$$\sigma_{\bar{x}}=\frac{S}{\sqrt{n}}=\sqrt{\frac{40.46}{1400}}=0.17$$

$Z=\frac{\bar{X}-\mu}{\sigma_{\bar{x}}}=\frac{0.34}{0.17}=2$，查Z值表可得置信概率为95.45%。

答：该市居民年食糖需要量不低于3682000公斤，不高于4158000公斤的把握程度为95.45%。

单变量

推断性分析——假设检验(小概率事件)

原理

通过假设检验可以知道，在多大的把握程度下，我们应该拒绝原假设，接受备择假设。假设检验包括Z检验、t检验和卡方检验。

示例

一个批发企业定向供给一些工厂某种原料。原来每个工厂每月的平均购买量为950吨，该批发企业为了鼓励各厂增加购买量，采用批量作价的价格策略推销原料，即每次购买的批量越大，享受越高的数量折扣。半年以后，企业市场部随机抽出64家工厂作为样本进行调查，结果发现64家工厂平均购买量增加到了1000吨，标准差为200吨。现在该批发企业想知道：平均购买量的增加是由价格策略的改变引起的，还是一种随机现象？

单变量

推断性分析——假设检验

单变量

推断性分析——假设检验

1. $H_0: \mu \leq 950$

   $H_1: \mu \geq 950$

2. 选择Z检验

3. α =0.05（双尾），Z_α=1.96的一种。

4. $Z=\frac{\bar{X}-\mu}{\sigma_{\bar{x}}}=\frac{1000-950}{\frac{200}{\sqrt{64}}}=2$

5. Z=2>Z_α=1.96，拒绝H_0，接收H_1

双变量

简单相关分析

本质　简单相关分析用于描述两个变量之间的相关程度。

举例　有一种产品的价格为x，销量为y，那么两者有怎样的关系？

双变量

简单回归分析

本质　简单回归分析的目的在于找出两个变量之间的相关关系。在进行简单回归分析时，变量之间暗含着因果关系。

举例　考察国家经济发展与企业商品销售的关系时，实际上假设国家经济发展为自变量，企业商品销售为因变量。

双变量

方差分析

本质

方差分析(ANOVA)，一般用于检验两组或两组以上调查对象在某一变量均值上的差异。

举例

一家企业想了解其某种品牌产品的使用者之间在态度上是否存在差异。

根据市场调查得到的数据，它先将使用者分为频繁使用者、普通使用者、少量使用者和未使用者。

然后，采用方差分析，就可以观察不同组别的使用者对这一品牌产品的态度偏好是否存在差异。

多变量

多元相关分析

本质　多元相关分析适用于描述两个以上变量之间的相关程度。

举例　一个企业认为它的某种产品的销售额与该产品的价格、广告支出和推销人员的数量有关。为了确定这些变量之间是否两两相关以及它们之间两两相关的程度，就需要使用多元相关分析，并计算偏相关系数。

多变量数据分析：

用于确定两个以上变量之间的关系。

多变量

多元回归分析

举例　　对销售量进行预测，相关的解释变量就有广告费用、销售代理人的数量、产品价格和季节等因素。

多变量

因子分析

本质

因子分析的主要目的是简化数据、用少量的概括性指标（即因子）来反映包含在许多测量项目（问卷题目）中的信息。因子分析提出的因子，每一个都是一组相关变量根据各项目对因子变化的贡献来加权而得到的加权组合。

多变量

判别分析

本质　判别分析适用于因变量为类别数据的情况。

举例　因变量是对汽车品牌a b或c的选择，自变量为消费者对这些品牌在多种属性上的评价。

判别分析可用于回答以下问题：

1. 对于某品牌忠诚的顾客与其他顾客在人口特征方面有何差异？
2. 价格敏感的顾客与价格不敏感的顾客在心理特征上有哪些差异？
3. 不同的细分市场在媒体接触的习惯上有什么差异？

多变量

聚类分析

本质　聚类分析是一种将研究对象聚合归类的统计分析工具。聚类分析不要求事先知道研究对象的组别。

举例　进行市场细分，可以用聚类分析确定同质的消费者，然后分别研究不同消费群的购买行为。

对品牌和产品进行聚类分析，可以识别市场中哪些品牌或产品更相近，借以识别竞争对手和市场机会。

多变量

联合分析

本质　联合分析用于估计产品、品牌、服务或商店的不同属性对消费者的相对重要性，以及消费者对不同属性水平及其组合的偏好。

举例　在收集数据时，向调查对象展示由不同属性水平组成的选项，然后让他对其渴望的程度进行评价。联合分析就可以为研究对象找到一个比较合适的属性组合。

多变量

结构方程建模分析

本质

结构方程建模的最大优点在于能够对模型中的变量做因果关系的推测。结构方程建模分析主要由测量模型和路径模型组成。前者用于对指标或变量的测量结果进行评价，后者用于对指标或变量的因果关系进行分析。

多有不足　欢迎交流