第十课：均数间的比较--t检验

@文彤老师

均数间的比较

- 假设检验原理入门
- 单样本t检验
- 两样本t检验
- 配对t检验

分析前的准备工作

- 运用统计知识根据研究设计和资料的性质正确选择分析过程。
- 初步的统计描述（集中趋势、离散趋势）和统计分析（资料的正态性、方差齐性）。
  - Descriptive statistics 菜单
  - Compare Means→means

连续性变量统计描述的常见指标

- 集中趋势
  - 均数
  - P50
- 离散趋势
  - 标准差/方差
  - 四分位数间距
- 分布特征
- 异常值及其他

假设检验原理入门

为什么要做检验

- 通过获得随机样本来实施抽样研究的例子很多，但此时研究中直接获取的只是样本的情况，而研究者关心的并不仅仅是样本，更希望了解相应的总体特征。
  - 参数估计：推估样本所在的总体特征
  - 假设检验：对提出的一些总体假设进行分析判断，做出统计决策。

假设检验原理

- 分析实例
  - 某产品的口味测试中，历史数据表明满意度均数如果低于7.4分，则该产品基本无市场价值（可近似认为7.4分是总体均数），现有新产品进行了30例样本的测试，满意度均数为6.8，标准差为0.21，是否需要进一步测试？
- 现有的样本均数和已知总体均数不同，其差别可能有两个方面的原因造成。
  - 样本来自已知总体，现有差别为抽样误差
  - 样本所来自的总体与已知总体不同，存在本质差异
- 为识别这两种可能，应当对其做假设检验

生活中隐含的假设检验

- 掷骰子，猜到点数为胜
  - 其实大家都明白如果筛子没问题，则六个点的出现概率应当相等（均为1/6，这就是一个事先假设），我们只是看每次具体的试验中谁的运气好
- 今天一共下了600次注，竟然一共只猜中了一次
  - 虽然平均应当赢约100次，但今天忘了查皇历，不宜搏彩，运气实在太差
  - 骰子有鬼，掷骰子的人可以人为控制结局，从而利用这种能力使自己得到了更多的收益。
  - 虽然第一种解释是可能的，但我们认为在筛子公平的前提假设下出现如此结果实在是太不可能了（概率小到不应当被我们一次就碰上），因此我们认为骰子实际上不均匀

假设检验原理

- 基础：小概率原理，即一般认为小概率事件在一次随机抽样中不会发生。
  - 最经典的小概率事件：瞎猫碰到死耗子
- 基本思想：先建立一个关于样本所属总体的假设，考察在假设条件下随机样本的特征信息是否属小概率事件，若为小概率事件，则怀疑假设成立有悖于该样本所提供特征信息，因此拒绝假设
- 事实上，小概率事件在随机抽样中还是可能发生的，只是发生的概率很小。若正好碰上了，则假设检验的结论就是错误的。当然，犯这种错误的概率很小

假设检验的基本步骤：建立假设

- 根据统计推断的目的而提出的对总体特征的假设。统计学中的假设有两方面的内容：
  - 一是检验假设(hypothesis to be tested),亦称原假设或无效假设(null hypothesis)，记为$H0$；
  - 二是与$H0$相对立的备择假设(alternative hypothesis)，记为$H1$。后者的意义在于当$H0$被拒绝时供采用。两者是互斥的，非此即彼。
  - $H0$：$\mu=\mu0$，　　$H1$：$\mu\neq\mu0$；
  - $H0$：$\mu=7.4$，　　$H1$：$\mu\neq7.4$。

假设检验的基本步骤：确定检验水准

- 实际上就是确定拒绝*H*0时的最大允许误差的概率。检验水准(size of test)，常用α表示，是指检验假设*H*0本来是成立的，而根据样本信息拒绝*H*0的可能性大小的度量，换言之，α是拒绝了实际上成立的*H*0的概率。
  - 常用的检验水准为$\alpha = 0.05$，其意义是：在所设H0的总体中随机抽得一个样本，其均数比手头样本均数更偏离总体均数的概率不超过5%

假设检验的基本步骤：计算检验统计量和P值

- 实际上在此之前还有一步叫做进行试验，所需的样本数据即从此得来
- 统计量只是工具，概率值才是目的，它可以客观衡量样本对假设总体偏离程度
  - 从H0假设的总体中抽出现有样本（及更极端情况）的概率，即P值
  - 例如600次赢100次是H0假设的情况，只赢1次就是现有样本情况，更极端的情况就是连一次也没有赢

假设检验的基本步骤：计算检验统计量和P值

- 检验统计量的特点
  - 该统计量应当服从某种已知分布，从而可以计算出P值
  - 各种检验方法所利用的分布及计算原理不同，从而检验统计量也不同
  - 初学者往往本末倒置，很认真地在学工具，却忘记了统计学的本质是思维方式

假设检验的基本步骤：得出推断结论

- 按照事先确定的检验水准α界定上面得到的P值，并按小概率原理认定对H0的取舍，作出推断结论

- 若P ≤α
  - 基于H0假设的总体情况出现了小概率事件
  - 则拒绝H0，接受H1，可以认为样本与总体的差别不仅仅是抽样误差造成的，可能存在本质上的差别，属“非偶然的(significant)”，因此，可以认为两者的差别有统计学意义。
  - 进一步根据样本信息引申，得出实用性的结论

假设检验的基本步骤：得出推断结论

- 若P＞α
  - 基于H0出现了很常见的事件
  - 则样本与总体间的差别尚不能排除纯粹由抽样误差造成，可能的确属“偶然的(non-significant)”，故尚不能拒绝H0
  - 因此，认为两者的差别无统计学意义，但这并不意味着可以接受H0。

关于掷筛子的假设检验

- 建立假设
  - *H*0：筛子均匀，pi=1/6　　*H*1：筛子不均匀
- 确定检验水准
  - α=0.05
- 进行试验，计算检验统计量和P值
  - 相应的试验结果在H0下对应的概率为1/600略多一点
- 得出推断结论
  - 基于H0出现了小概率事件，结果有非常非常显著的统计学意义，你出老千！

假设检验应注意的问题

- 结论不能绝对化
  - 本身就保留了犯错误的可能性
  - 样本量导致的检验效能问题
    - 样本量太小，导致检验效能不足，从而无法检出可能存在的差异
    - 样本量太大，得出的有统计学意义的结论可能根本就没有实际意义

单样本t检验

统计理论复习

- 推断样本是否来自某已知总体，即要检验样本所在总体的均数是否等于已知的总体均数
- 为了回答该问题，统计学上采用了小概率反证法的原理：我们有如下两种假设：
  - H0：样本均数与总体均数的差异完全是抽样误差造成
  - H1：样本均数与总体均数的差异除由抽样误差造成外，也反映了两个总体均数确实存在的差异

统计理论复习

$$t = \frac{\bar{x} - \mu_0}{S_{\bar{x}}}$$

- 先假设H0成立，即一切都是抽样误差造成的。在这个前提下，我们的样本是从已知均数的大总体中抽出来的。
- 显然，样本均数和假设总体均数之差就代表了偏离假设的程度
- 但此差异所对应的概率究竟是大还是小？仅看这一个数字很难做出判断。因为这还和数据的离散程度有关，为此我们需要找到某种方式对这一差值进行标准化

统计理论复习

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

- 显然，标准化的基本方式就是将差值除以表示样本均数离散程度的指标
  - 在单样本的情况下，样本的均数服从t分布
- 这个被标化的差值，就是本次检验中所谓的统计量
  - 由于该统计量服从t分布，可利用该分布得到相应的概率值，故而此处的方法被称为为单样本t检验。
- 最终求得的P值表示从假设总体中抽出当前样本均数（及更极端情况）的概率总和

统计理论复习

- 如果该P值太小，成为了我们所定义的小概率事件（小于等于α水准），则我们怀疑所做的假设不成立，从而拒绝H0。
  - 基本信念：小概率事件在一次实验中不可能发生
- 反之，我们就不能拒绝H0，但一般也不太好说去接受他。

分析实例

- 消费者信心指数以100作为基准值，现希望比较2007年12月的总消费者信心指数是否与基准值有差异

方法的适用条件

- 因为有中心极限定理，一般均数的抽样分布都不会有问题，真正会限制该方法使用的是均数是否能够代表相应数据的集中趋势。
- 也就是说，只要数据分布不是强烈的偏态，一般而言单样本t检验都是适用的。
- 基于计算统计学的新工具：Boostrap抽样

完全随机的两样本t检验

完全随机的两样本t检验

- 目的：
  - 推断两个样本是否来自相同的总体，更具体地说，是要检验两样本所代表的总体均数是否相等。
- 检验假设
  - 无效假设 H0：$\mu_1 = \mu_2$
  - 备择假设 H1：$\mu_1 \neq \mu_2$
  - 检验水准 α =0.05

完全随机的两样本t检验

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}$$

- 统计理论复习
  - 和上面单样本的t检验的原理相同，我们也采用了小概率反证法，首先假设H0：两样本来自同一总体。当该总体服从正态分布时，我们就可以采用两样本t检验来计算从该总体中抽得这样两个样本（及更加极端情况）的概率为多少，从而做出统计推断。

完全随机的两样本t检验

- 统计理论复习
  - 由于H0假设的是两样本来自同一总体，分析目的只涉及到均值，因此两样本t检验在推导过程中除了要求总体服从正态分布外，还要求两样本各自所在总体方差相同。
  - 应用条件不被满足
    - 情况较轻时可以采用校正t检验的结果
    - 否则应使用变量变换使之满足条件
    - 或采用非参数检验过程

分析实例

- 现希望评价2007年4月第一次调查时不同收入人群的消费者信心指数是否存在差异
- 分析：数据为定量资料，设计为成组设计，目的是两样本均数的比较。
  - 正态性： 可作直方图等。
  - 方差齐性：系统在 t 检验结果中自动给出。

分析实例

	方差齐性检验		两样本均数的t检验					
	F值	P值	t值	自由度	P值	均数差	合并标准误	可信区间 下限 上限
方差齐 方差不齐								

- 结论
  - 方差齐性检验的结果
  - T检验的结果

适用条件

- 独立性：对结果的影响较大，但一般没问题
- 正态性：有一定的耐受能力，可以通过直方图等进行观察，偏的不厉害就行
  - 注意应当要分组考察
- 方差齐性：相对而言对结论的影响较大，需要进行方差齐性检验

配对t检验

统计理论复习

- 配对设计的两种情况
  - 对同一个受试对象处理前后的比较
  - 将受试对象按情况相近者配对（或者自身进行配对），分别给予两种处理，以观察两种处理效果有无差别。
- 配对设计的特点
  - 在配对设计得到的样本数据中，每对数据之间都有一定的相关，如果采用成组的t检验就无法利用这种关系，浪费了大量统计信息
  - 对于这种情况，统计学上的解决办法是求出每对的差值，通过检验该差值总体均数是否为0，就可以得知两种处理有无差异。

基本思路

- H0:两总体均值无显著差异，差值序列均值u0=0

$$t = \frac{\overline{D}}{S/\sqrt{n}}$$

- 构造统计量：同单样本均值检验
  - D=X- u0　S为差值序列的标准差
  - 实质是先求出每对测量值的差值;然后检验差值序列的均值是否与0有显著差异.

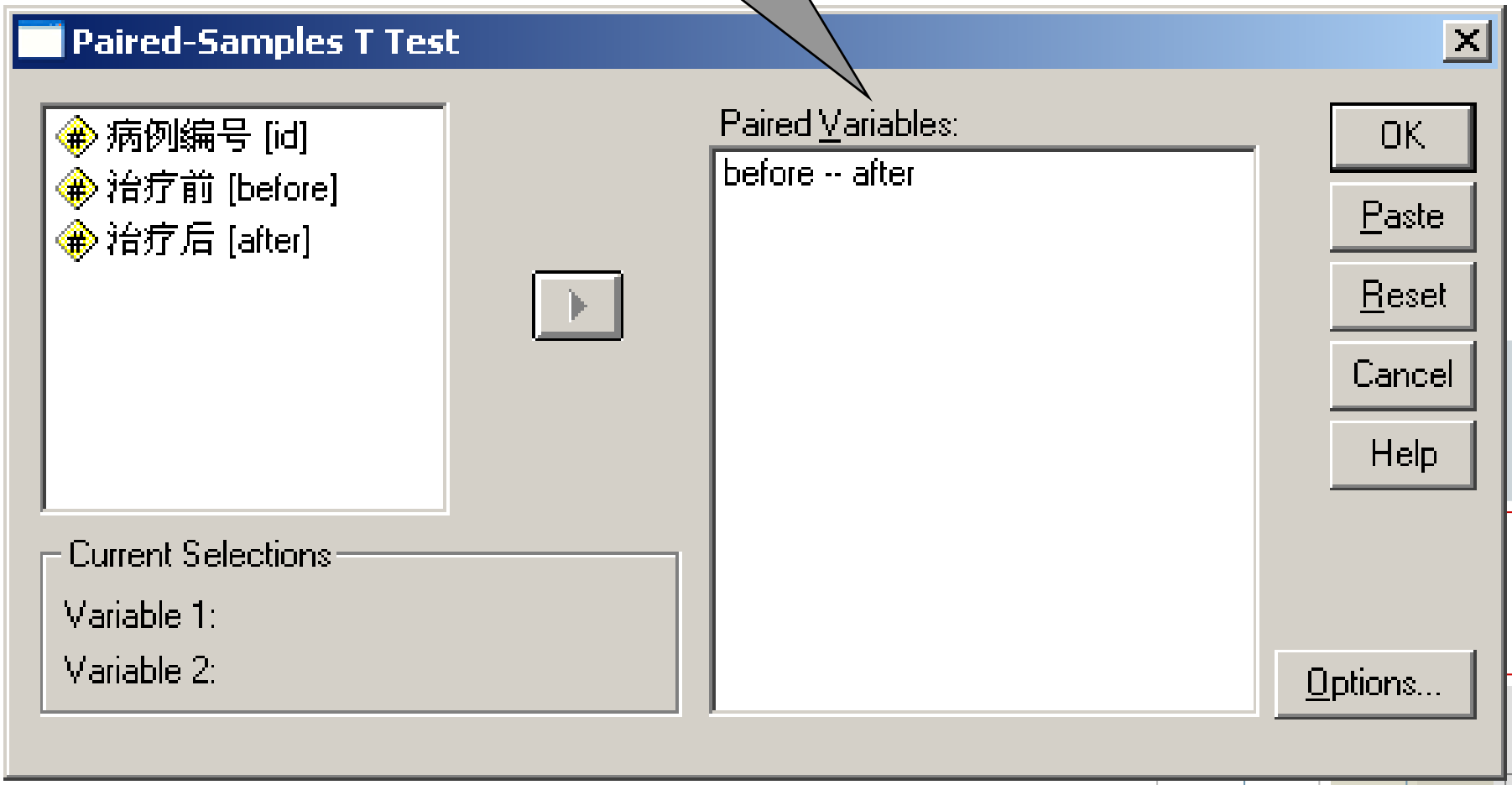
统计理论复习

- 如果差值的均值与0有显著差异,则认为两总体均值存在显著差异；否则，与0无显著差异,则认为两总体均值不存在显著差异
- 功能实际上和单样本t检验重复，但数据输入格式不同
- 和方差分析结果等价

分析实例

- 用某药治疗10名高血压病人，对每一病人治疗前、后的舒张压（mmHg）进行了测量，结果如下，问该药有无降压作用？数据见文件pairedt.sav。
  - 这是一个典型的个体自身治疗前后的配对设计，应当采用配对设计差值的t检验来进行分析。
  - 按照配对t检验对数据格式的要求，这里在输入数据时应当每个变量（一列）代表一个组，而每条记录（一行）代表一对数据。

- 两个变量同时被选中后输送到变量框



Paired Samples Test

		Paired Differences				
					95% Confidence Interval of the Difference	
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper
Pair 1	治疗前 -治疗后	10.00000	11.95361	3.78006	1.44890	18.55110

Paired Samples Test

		t	df	Sig. (2-tailed)
Pair 1	治疗前 -治疗后	2.645	9	.027

	对子间的差异						
	均数	标准差	标准误	可信区间 下限 上限	T值	自由度	P值

第十一课：单因素方差分析

@文彤老师

单因素方差分析

- 分析目的
  - 检验某一个影响因素的差异是否会给观察变量带来显著影响
- 例如：
  - 不同肥料对某农作物亩产量是否有显著差异
  - 不同学历是否对工资收入产生显著影响
  - 不同的推销策略是否对推销额产生显著影响
  - CCSS案例中提供了2007年4月，以及2007、2008、2009年12月四个时间点的消费者信心监测数据，现希望考察这四个时间点的消费者信心指数平均水平是否存在差异。

单因素方差分析

- 可以做两两检验吗
  - 比如共有4组的均数需要比较。如果用*t*检验进行两两比较，共要进行6次*t*检验。
  - 如果每次*t*检验犯第一类错误的概率为0.05，则不犯第一类错误的概率为0.95，6次都不犯第一类错误的概率为0.95的6次方，即0.7351，因此在6次*t*检验中至少有一次犯第一类错误的概率为0.2649。
  - 由此可见用两两检验的方式进行多组间的比较会增大犯第一类错误的概率。

单因素方差分析

- 基本分析原理
  - 方差分析是基于变异分解的原理进行的，在单因素方差分析中，整个样本的变异可以看成由如下两个部份构成：
    - 总变异=随机变异+处理因素导致的变异
    - 信心指数总变异＝不同月份导致的变异＋不同受访者间的随机变异
    - 处理因素导致的变异就是要研究的对象，我们希望能够证明他是否大于0

单因素方差分析

观测变量

控制因素

每个人具体的数值	访问月份
Xxx,xxx,xxx,xxx Xxx,xxx,xxx,xxx	200704
Xxx,xxx,xxx Xxx,xxx,xxx,xxx	200712
Xxx,xxx,xxx,xxx Xxx,xxx	200812
Xxx,xxx,xxx Xxx,xxx,xxx,xxx	200912

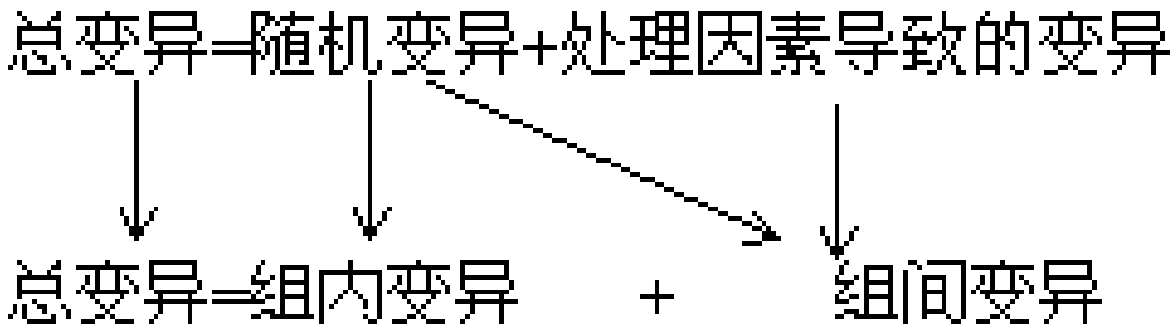
四个水平

单因素方差分析

- 实际数据的变异分解
  - 各组内部的变异（组内变异）只反映个体差异（随机变异）的大小
  - 各组均数的差异（组间变异）反映了个体差异（随机效应）的影响与可能存在的处理因素的影响之和
  - 总变异=组内变异+组间变异

单因素方差分析

- 统计量的含义
  - 采用合适的指标表示组内变异和组间变异的大小，将两者相比较，便可得知组内变异中是否真正包含了处理因素所导致的影响



单因素方差分析

- 统计量的含义
  - F>>1：组间变异远大于组内变异，处理因素有影响
  - F＝1：认为处理因素实际上无影响

$$F = \frac{SSA/(k-1)}{SSE/(n-k)} = \frac{MSA}{MSE}$$

  - 究竟F值要多大才算非常大？采用P值来确定。

单因素方差分析

- 方差分析本身是完美的，但在解决实际问题的时候，我们往往仍需要回答多个均数间究竟是哪些和哪些存在差异，这样问题又回到了两两比较上。
- 和t检验时的情况类似，方差分析也要求各样本来自正态总体，且各总体方差相等。如这些条件不满足，则应进行变量变换，或放弃使用该方法。

单因素方差分析

- 常用的两两比较方法
  - LSD法：实际上就是t检验的变形，只是在变异和自由度的计算上利用了整个样本信息，因此仍然存在放大一类错误的问题
  - S-N-K法：是运用最广泛的一种两两比较方法。它采用Student Range 分布进行所有各组均值间的配对比较。该方法保证在H0真正成立时总的α水准等于实际设定值，即控制了一类错误。

单因素方差分析

- 其它两两比较方法
  - Scheffe法：当各组人数不相等，或者想进行复杂的比较时，用此法较为稳妥。但它相对比较保守
  - 方差不齐时的两两比较方法：一般认为是Games-Howell法稍好一些，但最好直接使用非参数检验方法

单因素方差分析

- 两两比较方法
  - 以前国内外都以SNK法最为常用，但根据研究，当两两比较的次数极多时，该方法的假阳性非常之高，最终可以达到100%！因此比较次数较多时，包括SPSS和SAS在内的权威统计软件都不再推荐使用此法
  - 一般可以参照如下标准：如果存在明确的对照组，要进行的是验证性研究，即计划好的某两个或几个组间（和对照组）的比较，宜用Bonferroni（LSD）法；若需要进行的是多个均数间的两两比较（探索性研究），且各组人数相等，适宜用Tukey法；其它情况宜用Scheffe法。

SNK法的结果解释

- Student-Newman-Keuls两两比较结果：纵向按均数大小排序，横向为分组。只告诉P值小于预定的界值（默认0.05），而不显示具体P值。
- 进一步分析
  - 两两比较：LSD法

第十二课：卡方检验

@文彤老师

统计学回顾

- χ2检验
  - 是用途很广的一种假设检验方法，主要用于分类资料统计推断，包括：两个率或两个构成比比较的卡方检验；多个率或多个构成比比较的卡方检验以及分类资料的相关分析等
  - 它最基本的无效假设是：
    - H0：观察频数与期望频数没有差别
  - 其原理为考察基于H0的理论频数分布和实际频数分布间的差异大小，据此求出相应的P值。

Crosstabs过程

- 分析实例
  - 在CCSS的分析报告中，所有受访家庭会按照家庭年收入被分为低收入家庭和中高收入家庭两类，现希望考察不同收入级别的家庭其轿车拥有率是否相同。

方法原理

$$T_{RC} = \frac{n_R n_C}{n}$$

- 理论频数
  - 基于H0成立，两样本所在总体无差别的前提下计算出各单元格的理论频数来

			01. 是否拥有家用轿车		
			有	没有	合计
家庭收入2级	Below 48,000	计数	32	303	335
		家庭收入2级 中的 %	9.6%	90.4%	100.0%
	Over 48,000	计数	225	429	654
		家庭收入2级 中的 %	34.4%	65.6%	100.0%
合计		计数	257	732	989
		家庭收入2级 中的 %	26.0%	74.0%	100.0%

方法原理

- 残差
  - 设A代表某个类别的观察频数，E代表基于H0计算出的期望频数，A与E之差被称为残差
- 残差可以表示某一个类别观察值和理论值的偏离程度，但残差有正有负，相加后会彼此抵消，总和仍然为0。为此可以将残差平方后求和，以表示样本总的偏离无效假设的程度

方法原理

- 另一方面，残差大小是一个相对的概念，相对于期望频数为10时，20的残差非常大；可相对于期望频数为1000时20就很小了。因此又将残差平方除以期望频数再求和，以标准化观察频数与期望频数的差别。
  - 这就是我们所说的卡方统计量，在1900年由英国统计学家Pearson首次提出，其公式为：

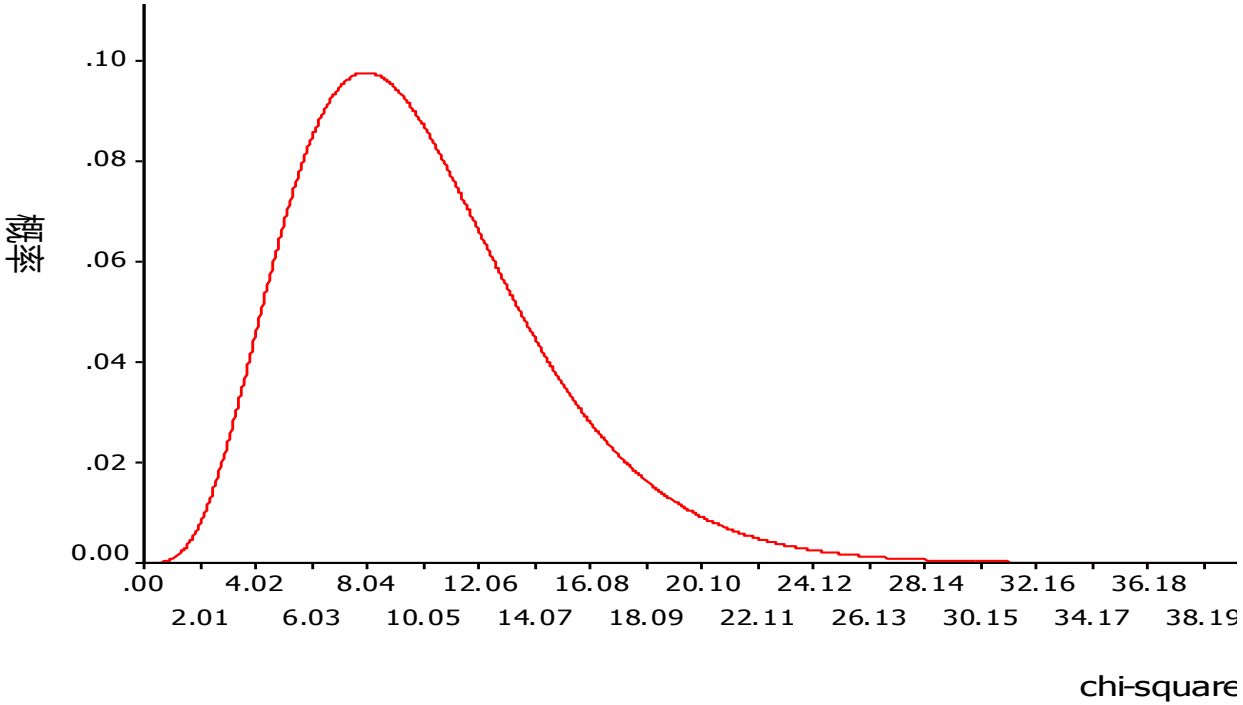
$$\chi^2 = \sum \frac{(A-E)^2}{E} = \sum_{i=1}^{k} \frac{(A_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(A_i - np_i)^2}{np_i} \qquad (i=1,2,3,\cdots k)$$

方法原理

- 从卡方的计算公式可见，当观察频数与期望频数完全一致时，卡方值为0；
- 观察频数与期望频数越接近，两者之间的差异越小，卡方值越小；
- 反之，观察频数与期望频数差别越大，两者之间的差异越大，卡方值越大。
- 当然，卡方值的大小也和自由度有关

方法原理

- 卡方分布
  - 显然，卡方值的大小不仅与A、E之差有关，还与单元格数（自由度）有关



结果解释

- 列出的检验结果

	值	df	渐进 Sig. (双侧)	精确 Sig.(双侧)	精确 Sig.(单侧)
Pearson 卡方	71.134[a]	1	.000		
连续校正[b]	69.848	1	.000		
似然比	80.146	1	.000		
Fisher 的精确检验				.000	.000
线性和线性组合	71.062	1	.000		
有效案例中的 N	989				

a. 0 单元格(.0%) 的期望计数少于 5。最小期望计数为 87.05。

b. 仅对 2x2 表计算

四格表χ2值的校正

- 英国统计学家Yates认为，χ2分布是一种连续型分布，而四格表资料是分类资料，属离散型分布，由此计算的χ2值的抽样分布也应当是不连续的，当样本量较小时，两者间的差异不可忽略，应进行连续性校正（在每个单元格的残差中都减去0.5）
  - 若n > 40 ，此时有 1< T <5时，需计算Yates连续性校正χ2值
  - T <1，或n<40时，应改用Fisher确切概率法直接计算概率

Crosstabs 过程

- 如何阅读卡方检验结果
  - 教科书的看法
    - 当n≥40且所有T≥5时，用普通的卡方检验，若所得P约等于Alpha，改用确切概率法；
    - 当n≥40但有1≤T<5时，用校正的卡方检验；
    - 当n<40或有T<1时，不能用卡方检验，改用确切概率法。
  - 实际的做法
    - 一律向下看齐

配对卡方检验

- 分析实例
  - 某公司期望扩展业务，增开几家分店，但对开店地址不太确定。于是选了20个地址，请两位资深顾问分别对20个地址作了一个评价，把它们评为好、中、差三个等级，以便确定应对哪些地址进行更进一步调查，那么这两位资深顾问的评价结果是否一致？

		顾问二的评价			
		差	中	好	合计
顾问一的评价	差	6	0	0	6
	中	5	2	2	9
	好	1	0	4	5
合计		12	2	6	20

配对卡方检验

- 方法原理
  - 显然，本例对同一个个体有两次不同的测量，从设计的角度上讲可以被理解为自身配对设计
  - 按照配对设计的思路进行分析，则首先应当求出各对的差值，然后考察样本中差值的分布是否按照H0假设的情况对称分布
  - 按此分析思路，最终可整理出如前所列的配对交叉表

方法原理

- 注意
  - 主对角线上两种检验方法的结论相同，对问题的解答不会有任何贡献
  - 非主对角线上的单元格才代表了检验方法间的差异
- 假设检验步骤如下（以四格表为例）：
  - $H0$：B = C
  - $H1$：B ≠ C

方法原理

根据H_0得b、c两格的理论数均为$T_b = T_c = (b+c)/2$，对应的配对检验统计量为：

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

一般在 $b + c < 40$ 时，需用确切概率法进行检验，或者进行校正。

分层卡方检验

- 进一步控制城市的影响，在控制城市影响的前提下得到更准确的家庭收入分级和轿车拥有情况的关联程度测量指标。
  - 层间差异的检验
  - 条件独立性的检验

第十三课：相关分析与回归分析

@文彤老师

内容安排

- 相关分析
- 线性回归模型简介
- 关于线性回归的高级话题

相关分析

- 常用术语
  - 直线相关
    - 两变量呈线性共同增大
    - 呈线性一增一减
  - 曲线相关
    - 两变量存在相关趋势
    - 并非线性，而是呈各种可能的曲线趋势
  - 正相关与负相关
  - 完全相关

相关分析

- 分析过程介绍
  - Bivariate过程
    - 进行两个/多个变量间的参数/非参数相关分析
    - 如果是多个变量，则给出两两相关的分析结果
  - Partial过程
    - 对其他变量进行控制
    - 输出控制其他变量影响后的相关系数

相关分析

- 分析过程介绍
  - Distances过程
    - 对同一变量内部各观察单位间的数值或各个不同变量间进行相似性或不相似性（距离）分析
    - 前者可用于检测观测值的接近程度
    - 后者则常用于考察各变量的内在联系和结构。
    - 一般不单独使用，而是作为MDS的预分析过程。
  - 典型相关分析

相关分析

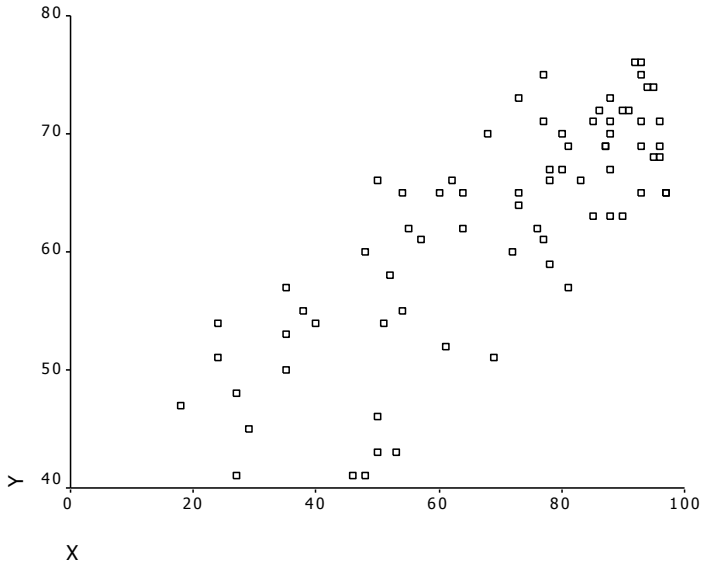
- Bivariate过程
  - 案例：考察信心指数值和年龄的相关性
    - 散点图
    - 非参数相关系数

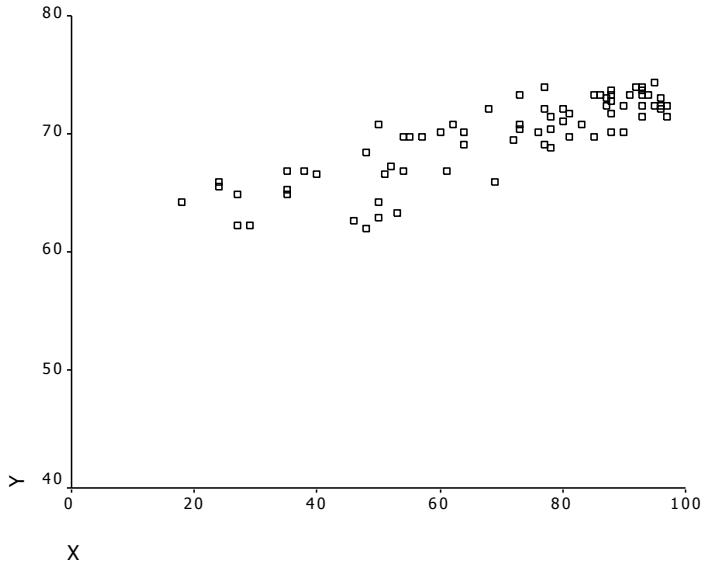
相关分析

- Partial过程
  - 在控制家庭收入QS9对总信心指数影响的前提下，考察总信心指数值和年龄的相关性。

线性回归模型简介

- 概述





线性回归模型简介

- 回归模型的分类
  - 线性回归
  - 非线性回归
  - 针对应变量为分类资料的回归方法
  - 其它回归过程

线性回归模型简介

- 基本模型
  - 希望研究月销售额与广告投入量、销售人员数量间的关系，并建立相应的多元线性回归方程，则实际上拟合的模型如下：

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

$$y_i = \hat{y} + e_i = a + b_1 x_{1i} + b_2 x_{2i} + e_i$$

线性回归模型简介

- 模型适用条件
  - 线性趋势
  - 独立性
  - 正态性
  - 方差齐性
    - 如果只是探讨自变量与因变量间的关系，则后两个条件可以适当放宽
  - 样本量
    - 根据经验，记录数应当在希望分析的自变量数的20倍以上为宜。

线性回归模型简介

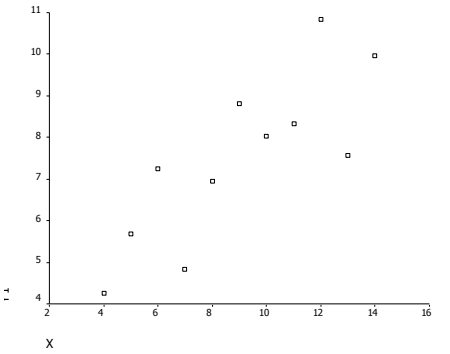
- 常用指标
  - 偏回归系数
    - 相应的自变量上升一个单位时，应变量取值的变动情况，即自变量对应变量的影响程度。
  - 标化偏回归系数：量纲问题
  - 决定系数
    - 相应的相关系数的平方，用R2表示，它反映应变量y的全部变异中能够通过回归关系被自变量解释的比例。

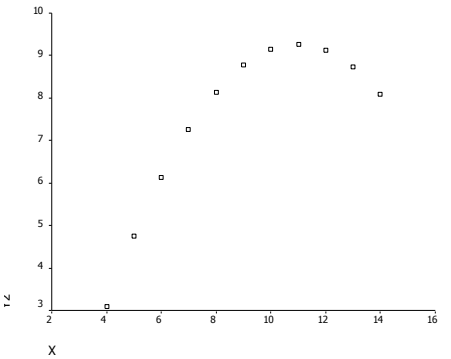
线性回归模型简介

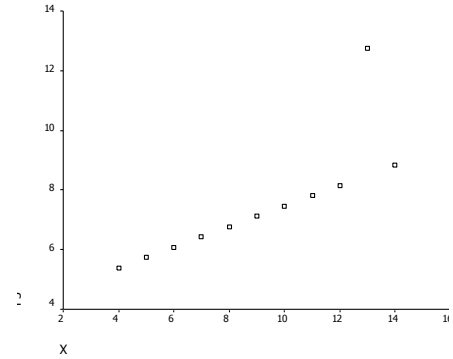
- 简单分析实例
  - 建立用年龄预测总信心指数值的回归方程
    - 使用方差分析模型拟合
    - 残差分析
    - 对残差的图形化分析
    - 绘制个体参考值范围以及均数值的可信区间

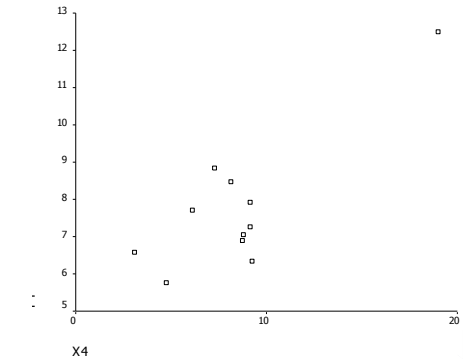
线性回归模型简介

- 分析步骤
  - 做出散点图，观察变量间的趋势







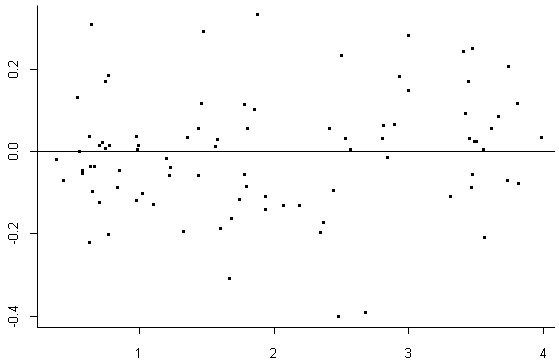


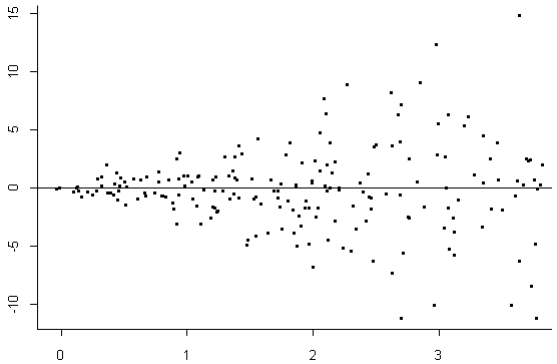
线性回归模型简介

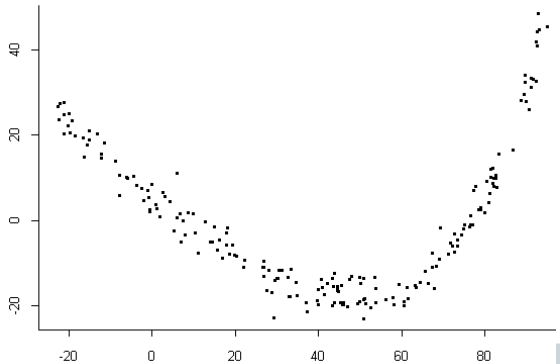
- 分析步骤
  - 考察数据的分布，进行必要的预处理。即分析变量的正态性、方差齐等问题
  - 进行直线回归分析
  - 残差分析
    - 残差间是否独立（Durbin-Watson检验）
    - 残差分布是否为正态（图形或统计量）

线性回归模型简介

- 分析步骤
  - 残差分析

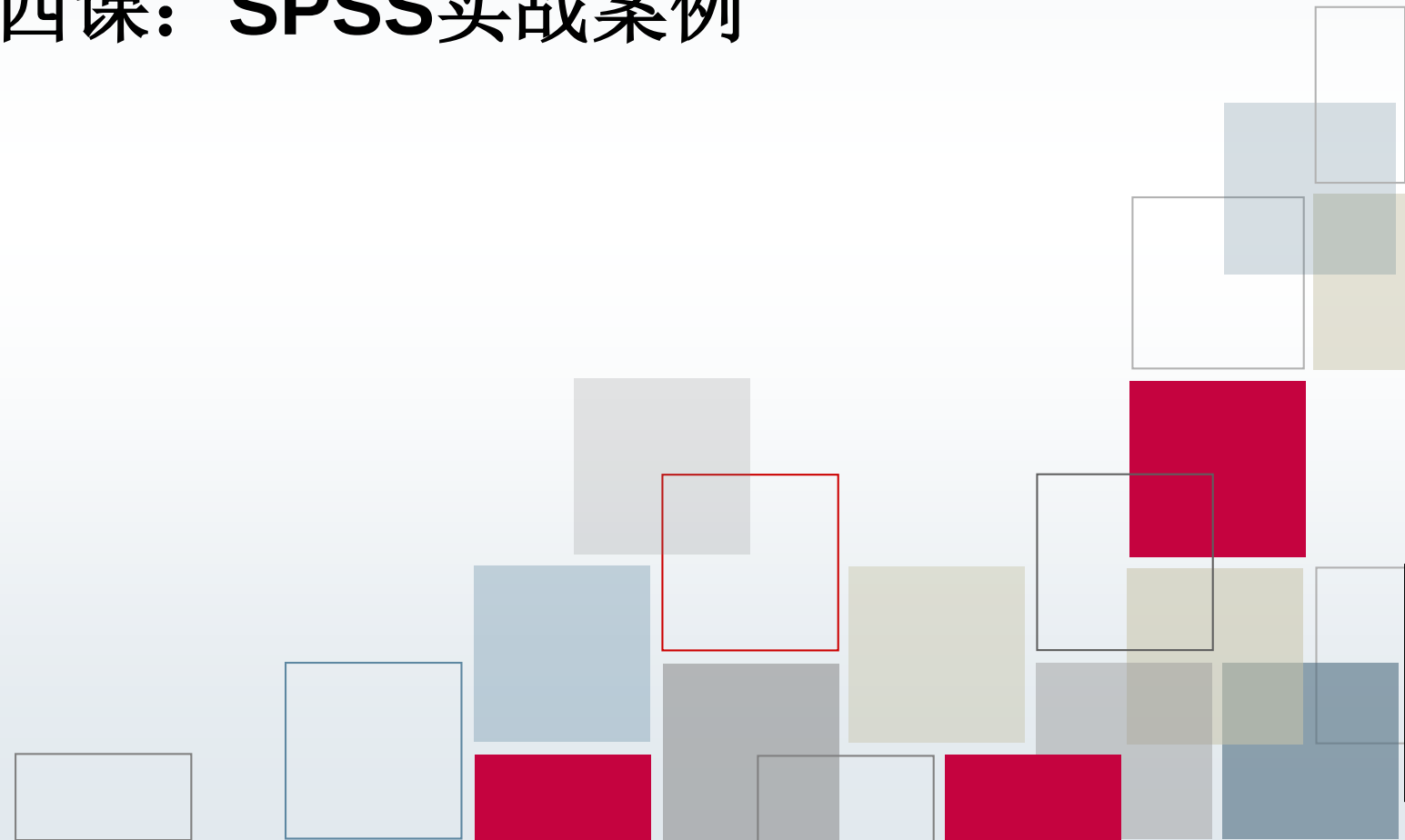






  - 强影响点的诊断
  - 多重共线性问题的判断
    - 这两个步骤和残差分析往往混在一起，难以完全分出先后

第十四课：SPSS实战案例



咖啡屋需求调查案例

研究目的

- 2003年中，受毕业校友的委托，北大的几位在读研究生在校内进行了一次关于北大师生对咖啡屋及类似休闲场所的需求调查，以便对这些校友的创业决策（在北大校内开设一家咖啡屋）提供数据支持。
  - 了解北大校内咖啡消费人群的基本背景状况；
  - 了解该消费人群的咖啡消费习惯，包括频次、额度、消费原因等；
  - 了解该消费人群可能存在，但目前尚未被满足的潜在需求。

问卷结构

- 第一部分：甄别问卷
- 第二部分：主体问卷
  - 最常去的
  - 最喜欢的，喜欢的原因
  - 消费的主要目的、项目、金额、时间
  - 信息来源
  - 预期位置
- 第三部分：个人信息

预分析

- 发现男性偏多
- 女性去过咖啡消费场所的比例要更高一些
- 注意收入、学历的分布
- 最终可以得到如下线索：
  - 整个研究接触到的核心人群应当就是本科/硕士在读学生，在抽样合理的情况下，这也应当是主要的咖啡消费人群。
  - 需要注意性别间可能存在的差异。

受访者对现有酒吧的U&A

- 对光顾频次和咖啡店偏好情况的交叉分析。
  - 为什么师生缘消费频繁程度明显高于其受欢迎程度的表现？
  - 雕刻时光的受欢迎程度为什么无法转换为其实际消费行为？
- 对多选题Q3进行分析
  - 受访者去咖啡吧最看重的就是情调和环境
  - 距离实际上也是重要因素

受访者在酒吧消费的情况

- 咖啡的消费比例在星巴克非常高；
- 类似于仙踪林则是以奶茶、果汁、冰激凌的消费为主，看来这两样比较适合于和恋人同行时饮用；
- 师生缘又一次走了中庸路线，没有发现他的消费人群更偏向于消费哪种饮料/食品；
- 西门外酒吧饮用啤酒和碳酸饮料的比例很高，这应当是一个很合理的结果。

酒吧/咖啡吧相关的信息来源

- 受访者对此类场所的了解还是通过路遇/朋友介绍/海报等传统方式为主，比较新的网上广告/BBS所占比例并不高；
- 对于恋爱人群和酒吧人群而言，校内BBS是一个可能有价值的推广渠道；
- 无论过去是否去过酒吧，其信息来源渠道是非常接近的。

加入背景资料进行结果验证

- 数据显示，相对而言，研究生/博士生更喜欢雕刻时光和师生缘，显然前者占领的是恋爱人群，后者占领的则是大众人群。而MBA、留学生更倾向于去星巴克这类“正宗”的地方消费。
- 相比之下，女性虽然去过咖啡吧的比例更高，但消费额度却比男性更低，更加集中在20-39元之间。

研究结论

- 校园咖啡吧的消费人群应当以本科/硕士学历，月收入1000元以下的人群为主；
- 校园咖啡吧大致有两种设计思路：便捷性为主（如师生缘），或者有突出的特色（如雕刻时光或者西门外酒吧），但是前者显然更贴合消费人群的需求；
- 主要消费人群的消费额度是人均30-60元的范围，相对而言，咖啡吧所提供的食品种类及特色并不重要，控制总价范围，或者说提供视频之外的消费选择可能更为重要；
- 咖啡吧选址应当尽量考虑宿舍区这种便捷性场所；
- 如果不是特殊定位，那么网络、BBS不是特别重要的宣传渠道；
- 在开业初期，女性群体可以作为首批推广的主要对象；

牙膏新品购买倾向研究案例

研究背景

- 在2003年，受客户的委托，我们对某牙膏新品的市场潜力进行了一次研究，其研究目的非常明确：
  - 考察该牙膏新品的市场欢迎程度是否达到预期；
  - 受访者对该牙膏新品的评价是否能超过现有市场品牌；
  - 受访者对该新品的评价受到哪些因素的影响，是否存在比较合适进入的细分市场；

研究设计

- 核心评判指标：本研究中的核心评价指标为对未来购买该新品的倾向性评分，按照1～10分设定，10分表示一定会购买，1分则表示一定不会购买；
- 考虑人口背景变量的影响：不同受访者的所在城市、性别、年龄、收入等人口背景资料显然应当被纳入分析范围，对其可能的影响进行分析；
- 考虑卫生习惯的影响：受访者的日常卫生习惯（如每日刷牙次数等）也应当是重要的潜在影响因素；
- 现在使用牙膏品牌的影响：不同品牌的牙膏实际上占据的是不同的细分市场，而新上市产品的细分市场定位是否正确，将会直接影响其上市是否成功。因此受访者目前最常使用的牙膏品牌也将是重要的潜在影响因素。

分析思路

- 购买倾向性评分，可直接按照连续性变量加以分析；
- 研究中的基本观察单位就是受访者；
- 本案例中所需要考虑的潜在影响因素有两个
  - 最常用的牙膏品牌，共分为六个水平（ANOVA）
  - 年龄（相关/回归）
- 如果同时考虑上述两个影响因素的作用，则必须要建立一个多变量分析模型
- 在实际分析中，我们没有必要直接去建立多变量模型，而应当先逐个进行变量的筛选和数据理解，在了解到足够的信息之后再来建立复杂模型。

统计描述（预分析）

- 因变量
  - 直方图
- 自变量
  - 频数表
  - 探索分析

数据建模

- 年龄对上市后指数影响程度的分析
  - 散点图
  - 相关分析
- 对品牌的作用进行总体检验
  - 方差分析
  - 两两比较

分析结论

- 基本结论
  - 在六种现有牙膏品牌的使用人群大致可被分为三组，最常使用黑*、其他品牌者对新产品表现出了较大兴趣，平均购买分值在7分以上；
  - 高*、佳*的使用者表现出了较高的忠诚度，上市后购买指数的平均分值仅在5.5分左右；
  - 中*、蓝*的适用者情况介于两者之间，评分居中。
- 营销建议
  - 建议该新产品在上市后应当主攻黑*、其他品牌的定位人群，相对而言成功进入该细分市场的可能性较大，应当会有较好的收益。