准备R软件

1. 安装R：http://cran.r-project.org/
2. 简介: http://staff.ustc.edu.cn/~ynyang/lm2020/lab/aboutR.htm
3. R 入门书籍：R in action by Kabacoff （/books/6.pdf）

回归分析 (01714601)
主讲: 杨亚宁 ynyang@ustc.edu.cn
助教: 赵明华 zmh07@mail.ustc.edu.cn
曾正浩 zzh98052@mail.ustc.edu.cn

课程主页：http://staff.ustc.edu.cn/~ynyang/lm2020

第二讲 案例

2020.2.21

内容

- 三个案例
- **Karl Pearson**与相关系数

回忆：两个二项分布的概率相等性检验

$$a \sim B(n_1, p_1),\quad c \sim B(n_0, p_0);\qquad H_0 : p_1 = p_0$$

$$\hat{p}_1 = a/n_1,\qquad \hat{p}_0 = c/n_0,\qquad \hat{p} = (a+c)/(n_1+n_0),$$

两样本 z - 检验：

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_0)\,\hat{p}(1-\hat{p})}},$$

H_0下近似地，$z \sim N(0,1)$

如果$|z| \geq z_{\alpha/2}$，在α水平下拒绝原假设，p值$= P(|Z^*| > |z|\,||z), Z^* \sim N(0,1)$

2x2列联表的Pearson卡方检验

记$b=n_1-a$,$d=n_0-c$,将数据以列联表展示:

	1	0	总计
组1	a	b	n_1
组2	c	d	n_0
总计	m_1	m_0	n

$$n=n_1+n_0=m_1+m_0$$

Pearson卡方统计量: $X^2=\dfrac{n(ad-bc)^2}{n_1n_0m_1m_0}$, H_0下近似地$X^2\sim\chi_1^2$.

$X^2\geq\chi_1^2(\alpha)$时在α水平下拒绝零假设, $p=P(T^*\geq X^2\mid X^2), T^*\sim\chi_1^2$

命题1. $X^2=z^2$, 因此2×2表格的Pearson卡方的等价于上页的z检验。

案例1：霍乱传播（Freedman第一章）

19世纪中期，人类对霍乱(cholera)传染几乎一无所知，细菌学说只是众多理论中的一种。 1855年伦敦医生John Snow 利用通过精巧的分析发现霍乱是一种通过饮用水传染的疾病。

- 1848年，伦敦爆发了霍乱。Snow找出了第一个病例，他是刚从霍乱流行的汉堡乘船到伦敦的海员。并且发现第二例病人住过第一个病例住过的房间。

  这表明霍乱可能具有传染性

- 然后发现附近的两个公寓，一个公寓发生了霍乱，另一个没有。前者的饮水系统被污染了，而后者没有。

  这表明霍乱可能是通过饮用水传染的。

- 1854年伦敦又爆发了霍乱。Snow在地图上标识了疾病发生区域。很多病例集中在Broad Street的供水泵附近。
  - Snow发现其它地区的零星病例，大多与Broad Street有关
  - 另外，Broad Street也有病例很少的地方：比如一个酿酒厂，该厂的工人习惯于喝麦芽酒，而且该厂有自己的供水系统；救济院也是如此。

所有证据都表明，疫情与供水系统有关。

- **Snow注意到伦敦有两大供水公司，并收集了用户数据**
  - **S公司（Southwark and Vauxhall）：水源在泰晤士河下游，污染较重，**
  - **L公司（ Lambeth ）：水源在上游，污染不严重。**

Table 2. Death rate from cholera by source of water. Rate per 10,000 houses. London. Epidemic of 1854. Snow's table IX.

	No. of Houses	Cholera Deaths	Rate per 10,000
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

S公司的客户死亡率显著高于L 公司

应用两个二项分布的正态检验，得p值=0，显著不同，S公司的客户死亡率显著高于L公司。 所以霍乱死亡率与供水公司有关。

但这不能说明两家公司水质不同导致了死亡率不同，因为这是一个观察研究：我们并不知道两家公司及其客户是否存在其它差异。

■ **Snow进一步研究发现：**

两个公司在伦敦的某些地区的供水并没分开，而是混在了一起，比如同一个房子两侧的两家可能选用不同的公司。

两个公司的价格、服务各方面差异不大，用户一般不知道两个公司水源差异。

各家选用供水公司几乎是随机的：不依赖于贫富、房子大小、房主的职业、所处位置等。

所以，虽然这是观察研究数据，但 **Nature在该地区做了一个随机试验（称为天然试验）**。两个公司客户的发病率差异显著只能归因于公司的不同，而公司的不同主要是水质。

基本可以断言：饮用水传播霍乱

案例2：乳腺癌研究

乳腺癌在北美妇女中较为常见. Mammograph 是一种X光早期筛查方法。为了检验Mammograph的有效性， 1960’s在纽约进行了一个大型随机化控制试验。

- HIP (Health Insurance Plan)是一种集体医疗保险，有700,000个成员。其中62,000个年龄在40-64的女成员被随机地分为处理组(treatment)和对照组(control)：
  - 处理：邀请参加一年4次的Mammograph筛查，另外也参加一般临床检查。 1/3被邀请的人拒绝参加筛查。
  - 对照：只参加一般临床检查，但不接受Mammograph筛查 .

5年跟踪结果：

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

H0： Mammograph无效

1. 下面我们检验 Screened组和对照组的死亡率是否相同：

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

$$\hat{p}_1 = 23/20200 = 0.0011, \quad n_1 = 20200$$
$$\hat{p}_2 = 63/31000 = 0.002, \quad n_2 = 31000$$
$$\hat{p} = (23+63)/(20200+31000) = 0.00168$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\,\hat{p}(1-\hat{p})}} = \frac{0.0011 - 0.002}{\sqrt{\left(\frac{1}{20200} + \frac{1}{31000}\right) \times 0.00168 \times (1-0.00168)}} = -2.431$$

p值$= 0.015$, 在0.05显著性水平下拒绝原假设。这是否说明筛查有效？

由于成员自己决定是否接受筛查，所以同意接受筛查的Screened组是处理组的一个有偏样本（Screened组数据是一组观察研究数据！）。

事实上，Screened组与Refused组有系统性差异：
富裕和教育程度高的人更倾向于接受邀请，但这些人乳腺癌发病率比其他人高 （因为生育率低）。

作业：你能否从Table1数据中提供Screened 组与Refused组存在差异的证据？

因此，以 Screened 作为处理组会导致检验结果出现偏差：
即使筛查没有任何作用（原假设），Screened组与Control组相比，乳腺癌发病率也比较高。这导致检验的 I型错误率偏大.

2. 正确的分析方法是比较Treatment组(含所有被邀请的人,不论是否接受筛查!) 和 Control 组。

	Group size	Breast cancer No.	Breast cancer Rate	All other No.	All other Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

$$\hat{p}_1 = 0.0013,\ \hat{p}_2 = 0.002,$$
$$\hat{p} = (39+63)/62000 = 0.00165$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\,\hat{p}(1-\hat{p})}} = \frac{0.0013 - 0.002}{\sqrt{\left(\frac{1}{31000} + \frac{1}{31000}\right) \times 0.00165 \times (1-0.00165)}} = -2.378$$

$pvalue = 0.017$, 所以Mammograph筛查能显著地降低死亡率。

案例3：贫困的原因

在十九世纪的英国，穷人（pauper）生存主要依赖救济院（poor-houses）的救济，但被要求必须在救济院工作。此外政府也逐渐增加救济院外的政府救济(out-relief)。 英国统计学家Yule (1899) 研究了院外救济是否会增加贫困人口比例。

观察研究。

需要收集什么数据？

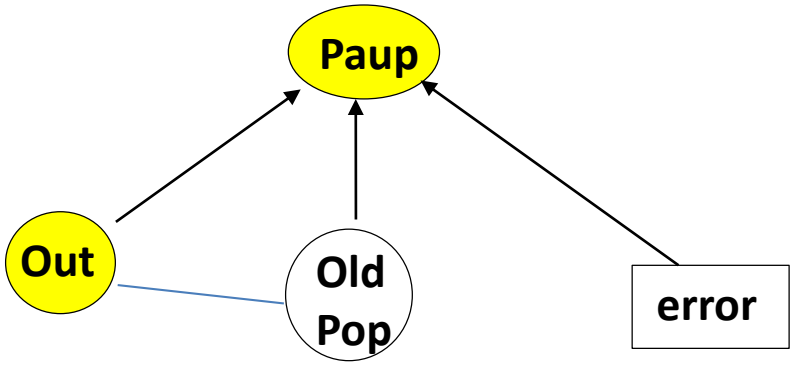
Yule收集了1871，1881，1891年若干区会的数据。Yule希望知道变量Out是否影响Paup：

- Paup：穷人占总人口的比例
- Out=N/D: 救济院外和救济院内救济人数之比，
  其中N= # out-relief； D=# inside poor-houses

Yule注意到老年人比例和人口总数可能既与Paup有关，也与Out有关，是干扰因素，需要加以控制

- Old：65岁以上老人站总人口的比例
- Pop ：总人口数

变量之间的关系如下表示（路径模型）

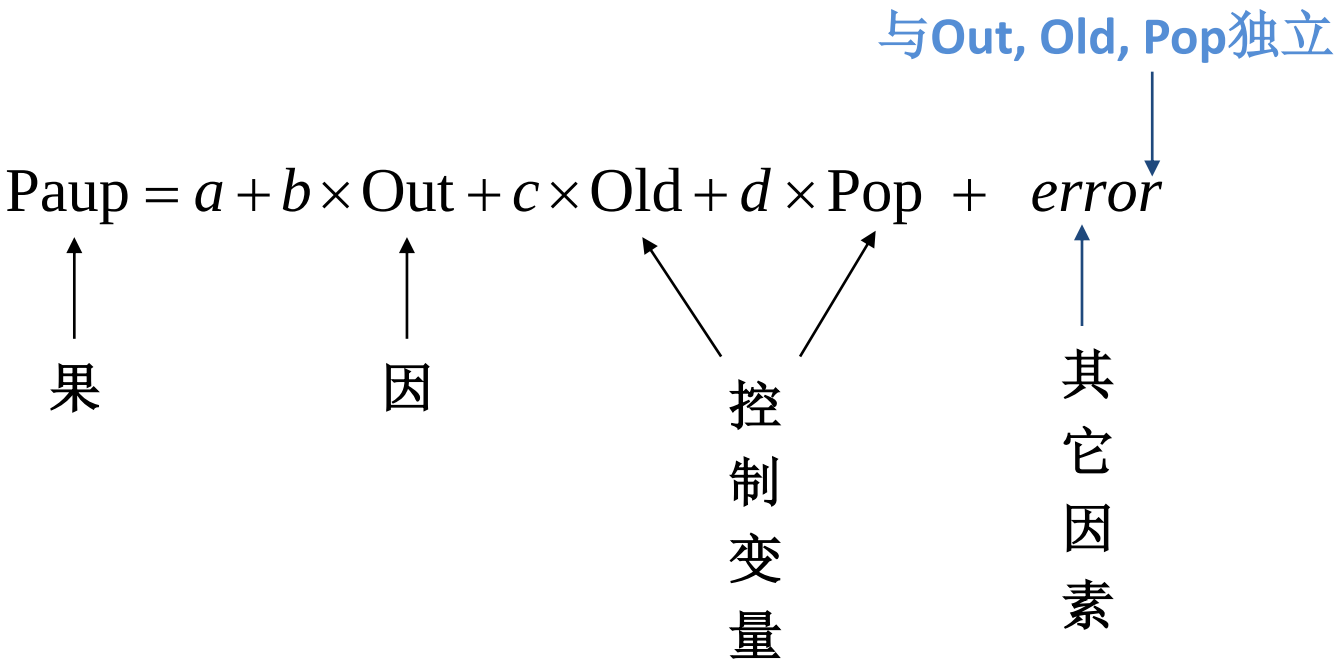


error代表了与Paup有关的其它因素。

假设error与Out、Old、Pop无关(error与它们无连线)。

上述模型的数学表达 - 线性模型：

$$\text{Paup} = a + b \times \text{Out} + c \times \text{Old} + d \times \text{Pop} + \ error$$



如果模型是正确的（主要是指“error与Out，Old，Pop独立），那么我们有因果关系：其它变量不变时，out增加一个单位，Paup增加b个单位。

但模型很可能是不正确的，可能还存在其它干扰因素，比如社会经济状况，政府管理的效率等。

注：实际上Yule使用的是两个年份的相对变化率数据，

对任一变量x_t，定义相对变化率$\Delta x_t = \frac{x_t - x_{t-10}}{x_{t-10}}, t = 1881, 1891,$

前面我们使用的是简化的模型，Yule的模型：

$$\Delta \text{Paup} = a + b \times \Delta \text{Out} + c \times \Delta \text{Old} + d \times \Delta \text{Pop} + error$$

分层（stratification）：

而且*Yule*认为不同的地区(农村，小城市，大城市，城乡结合地区）具有不同的模型关系,且认为1881年和1891年的模型不同，所以他实际上建立了$4 \times 2 = 8$个线性模型。

卡尔.皮尔逊 (Karl Pearson,1857-1936)

卡尔.皮尔逊，英国数学家，现代统计的创始人(以1900年的皮尔逊卡方检验为标志)。他是高尔顿的门徒和传记作者。

1901年他和Galton, Weldon 一起创办了第一份统计杂志Biometrika, 1925年创办了优生学/遗传学杂志*Annals of Eugenics* (*Annals of Human Genetics*)。1911年在伦敦大学学院建立了世界上第一个(生物)统计系。

主要贡献：相关系数，矩方法，p值，Pearson卡方检验，主成分分析，Pearson分布族，...

Pearson相关系数

随机变量之间的关联性通常以Pearson相关系数度量，它实际上度量的是线性关联程度。相关系数的概念和初始定义由Galton提出，但深入的研究和推广属于K. Pearson。

记 $\mathbf{x}=(x_1,....,x_n)^{\mathsf{T}}, \mathbf{y}=(y_1,....,y_n)^{\mathsf{T}}.$

*Galton*定义相关系数 $g=\dfrac{\mathbf{x}^{\mathsf{T}}\mathbf{y}}{\|\mathbf{x}\|\cdot\|\mathbf{y}\|}=\dfrac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}=\cos(\theta_{\mathbf{xy}}).$

缺点是平移会改变g.

样本$(x_1,y_1),...,(x_n,y_n)$ 的 Pearson 样本相关系数：

$$r_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}.$$