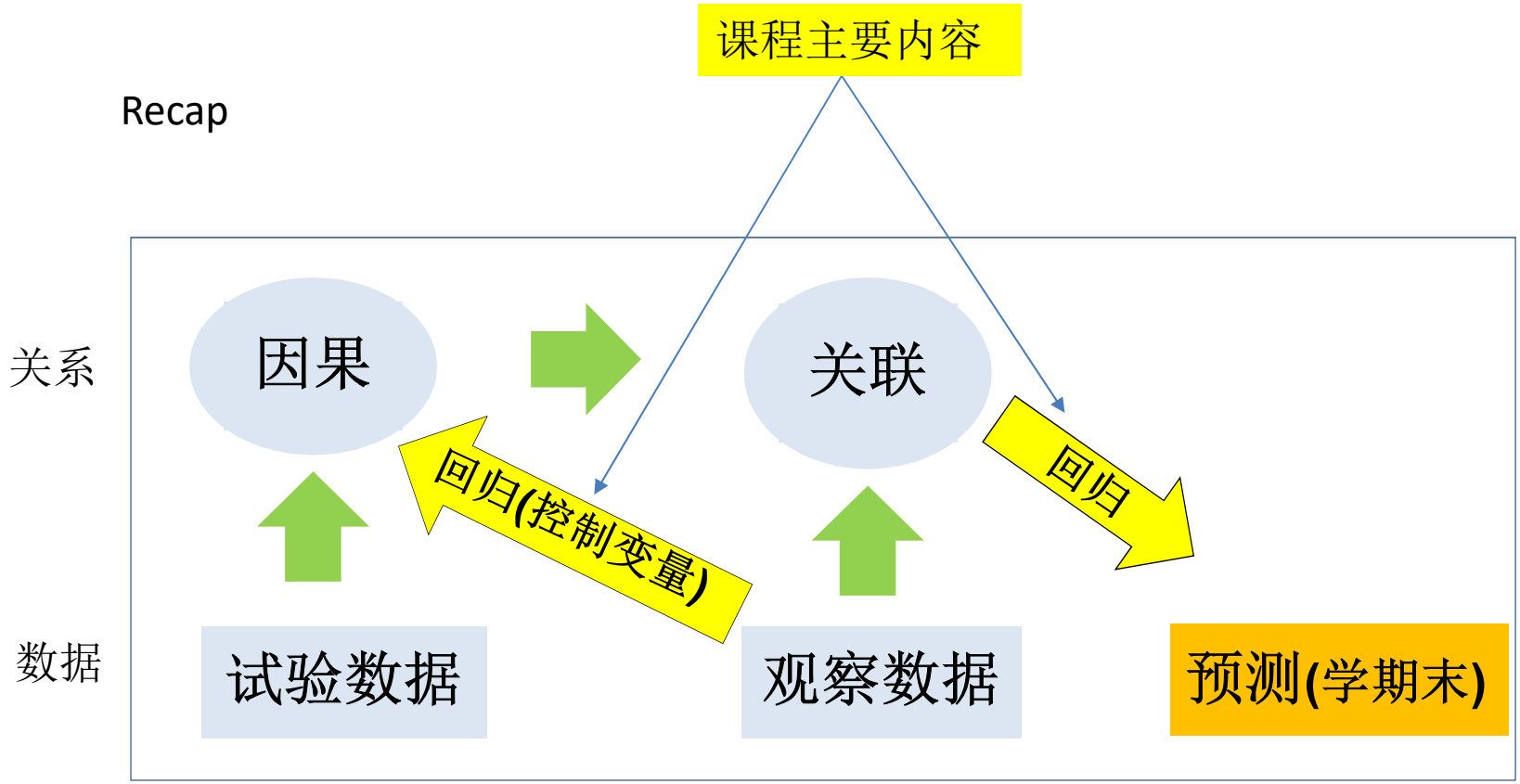
课程主页：http://staff.ustc.edu.cn/~ynyang/2022

第二讲 相关系数与偏相关系数

2022.9.9

问题：样本相关系数$r_{xy} = 0.3$是否足够大，是否表明x和y相关？ 考察$|\sqrt{n}r|$是否大于1.96

Recap



研究两个随机变量x, y的关系，有两种常见策略

- 相关系数度量：无模型，两个变量地位相同、对称。最常用的是Pearson相关系数，还有非参数的Kendall's tau, Spearman's rho.
- 简单线性回归模型：$y = a + bx + \varepsilon$, b与相关系数成正比， x, y地位不对称。

其中第一种方式给出一个数字度量，简单明了，但缺少细节；第二种方式给出变量关系的显式表达。

研究给定/控制其它变量$\boldsymbol{z}$的条件下，两个随机变量x, y的关系：

- 偏相关系数度量：无模型， x, y地位不对称。类似于Pearon相关系数（未见过Kendall's tau, Spearman's rho对应版本）.
- 多重线性回归模型：$y = a + bx + \mathbf{c}^{\mathrm{T}}\boldsymbol{z} + \varepsilon$, b 与偏相关系数成正比。y为主（响应），$x, \mathbf{z}$ 是解释变量/自变量/预测变量。

今天回顾相关系数，并介绍三元情形下的偏相关系数

Pearson相关系数

Pearson相关系数，它度量的是两个随机变量之间的线性相关程度。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 Pearson。

相关系数

随机变量x，y的(总体)Pearson相关系数：$\rho_{xy}=\dfrac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)}\sqrt{\operatorname{var}(y)}}$

样本相关系数

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

$$r_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}\hat{=}\frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}},$$

其中$s_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})$，$s_{xx}=\sum(x_i-\bar{x})^2$.

性质：$|\rho_{xy}|\le 1,|r_{xy}|\le 1$ (Cacuhy-Schwartz不等式).

约定：
小写字母：随机变量
小写黑体：向量
大写字母：矩阵

记$\mathbf{x}=(x_1,....,x_n)^{\mathsf{T}}$，$\mathbf{y}=(y_1,....,y_n)^{\mathsf{T}}$，它们的中心化记为：

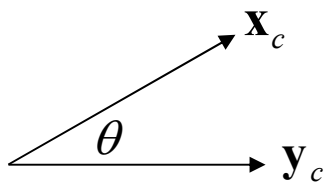
$$\mathbf{x}_c=(x_1-\bar{x},....,x_n-\bar{x})^{\mathsf{T}}=\mathbf{x}-\mathbf{1}\bar{x},$$

$$\mathbf{y}_c=(y_1-\bar{y},....,y_n-\bar{y})^{\mathsf{T}}=\mathbf{y}-\mathbf{1}\bar{y}.$$

样本相关系数度量相似性

Pearson相关系数度量了中心化向量之间的相似性/角度：

$$r_{xy}=\frac{\mathbf{x}_c^{\mathsf{T}}\mathbf{y}_c}{\|\mathbf{x}_c\|\cdot\|\mathbf{y}_c\|}=\left(\frac{\mathbf{x}_c}{\|\mathbf{x}_c\|}\right)^{\mathsf{T}}\left(\frac{\mathbf{y}_c}{\|\mathbf{y}_c\|}\right)=\cos(\theta_{\mathbf{x}_c\mathbf{y}_c}).$$



高尔顿定义的相关系数：$g=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$.

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。现代统计学的奠基人。

- 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*（后改名为*Annals of Human Genetics)*。
- 1911年在伦敦大学学院（UCL）建立了世界上第一个(生物)统计系。

生物统计：biostatistics, 与人有关的统计，可能翻译为医学统计更恰当

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.

虚线框内容不要求掌握

样本均值的期望、方差、分布相对容易获得，但样本方差、相关系数等二阶统计量的期望、方差、分布通常较难计算，即使假设正态总体也是如此。

样本均值、样本方差的期望和方差

总体均值μ，方差σ^2,样本均值$\bar{x}$，样本方差s^2，样本量n, 则

$$\mathrm{E}(\bar{x}) = \mu, \mathrm{var}(\bar{x}) = \sigma^2 / n;$$

$$E(s^2) = \sigma^2, \mathrm{var}(s^2) = \frac{1}{n}\left(E(x-\mu)^4 - \frac{n-3}{n-1}\sigma^4 \right)$$

如果总体是正态分布$N(\mu, \sigma^2)$,则$\mathrm{var}(s^2) = 2\sigma^4/(n-1)$.

样本相关系数的期望和方差

如果总体是二元正态，则基于简单随机样本(样本量n)的样本相关系数的均值和方差为

$$E(r) = \rho\left[1 - \frac{(1-\rho^2)}{2n} + O\left(\frac{1}{n^2}\right)\right], \quad \mathrm{var}(r) = \frac{(1-\rho^2)^2}{n-2} + O\left(\frac{1}{n^2}\right),$$

由此得到ρ的偏差更小的估计 $r^* = r + \frac{r(1-r^2)}{2n}$; r的方差的估计$\widehat{\mathrm{var}}(r) = \frac{(1-r^2)^2}{n-2}$.

(参见Lehmann《点估计理论》)

相关性的精确检验

为了求 r 的分布，我们需如下引理

引理1. 假设随机向量 $\mathbf{x}$的分量$x_1,...,x_n$ $iid \sim N(0,1)$, 记作$\mathbf{x} \sim N_n(0, I_n)$，则

(1) 对任何常数向量 $\mathbf{a} \in R^n, \| \mathbf{a} \| = 1$, 则$\mathbf{a}^{\mathsf{T}}\mathbf{x} \sim N(0,1)$, $\| \mathbf{x} \|^2 - (\mathbf{a}^{\mathsf{T}}\mathbf{x})^2 \sim \chi^2_{n-1}$, 两者独立，且

$$\sqrt{n-1}\frac{\mathbf{a}^{\mathsf{T}}\mathbf{x}}{\sqrt{\| \mathbf{x} \|^2 - (\mathbf{a}^{\mathsf{T}}\mathbf{x})^2}} \sim t_{n-1}$$

(2) $\| \mathbf{a} \| = \| \mathbf{b} \| = 1, \mathbf{a}^{\mathsf{T}}\mathbf{b} = 0$, 则 $\sqrt{n-2}\frac{\mathbf{a}^{\mathsf{T}}\mathbf{x}}{\sqrt{\| \mathbf{x} \|^2 - (\mathbf{a}^{\mathsf{T}}\mathbf{x})^2 - (\mathbf{b}^{\mathsf{T}}\mathbf{x})^2}} \sim t_{n-2}$

证明（经典方法）：(1)构造$n\times n$正交矩阵A, 第一行为$\mathbf{a}^{\mathsf{T}}$（其它行随意）,

令$\mathbf{y}=A\mathbf{x}=\begin{pmatrix}\mathbf{a}^{\mathsf{T}}\mathbf{x}\\ *\end{pmatrix}\overset{\text{记作}}{=}\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}$，可以证明$y_1, y_2, ..., y_n$ $iid\sim N(0,1)$，即$\mathbf{y}\sim N_n(0, I_n)$。

注意

(a) $y_1=\mathbf{a}^{\mathsf{T}}\mathbf{x}\sim N(0,1)$,

(b) 由$\|\mathbf{x}\|^2=\|\mathbf{y}\|^2\Rightarrow\|\mathbf{x}\|^2-(\mathbf{a}^{\mathsf{T}}\mathbf{x})^2=\|\mathbf{y}\|^2-y_1^2=y_2^2+...+y_n^2\sim\chi^2_{n-1}$,

且$\|\mathbf{x}\|^2-(\mathbf{a}^{\mathsf{T}}\mathbf{x})^2$只与$y_2,...,\ y_n$独立，因而与$\mathbf{a}^{\mathsf{T}}\mathbf{x}=y_1$独立

(c) 由t分布的定义知

$$\sqrt{n-1}\,\frac{\mathbf{a}^{\mathsf{T}}\mathbf{x}}{\sqrt{\|\mathbf{x}\|^2-(\mathbf{a}^{\mathsf{T}}\mathbf{x})^2}}=\frac{y_1}{\sqrt{(y_2^2+...+y_n^2)/(n-1)}}\sim t_{n-1}.$$

(2)的证明类似。

正态假设下相关系数的分布

命题1(正态总体). 假设 $(x_1, y_1),...,(x_n, y_n)$ iid 服从二元正态分布，则当 $\rho_{xy} = 0$ 时，有

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} \sim t_{n-2}$$

注： 我们用引理1进行（经典的）证明。

以后将在线性模型框架 下更简单地证明命题 1

命题1的证明：当 $\rho_{xy} = 0$ 时，y_i 与 x_i 独立，且 $y_i \sim$ 一元正态,因为 r 平移刻度变换下不变，我们 不妨设 $y_i \sim N(0,1)$，即 $\mathbf{y} = (y_1,..., y_n)^{\mathsf{T}} \sim N(0, I_n)$。

因为 $r = \dfrac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$，所以

$$\frac{r}{\sqrt{1-r^2}} = \frac{s_{xy}/\sqrt{s_{xx}}}{\sqrt{s_{yy} - s_{xy}^2/s_{xx}}} = \frac{s_{xy}/\sqrt{s_{xx}}}{\sqrt{\sum y_i^{\,2} - n\bar{y}^2 - \left(s_{xy}/\sqrt{s_{xx}}\right)^2}},$$

下面我们在给定 $x_1..., x_n$ 的条件下计算 t 的分布。

其中，分子项

$$s_{xy}/\sqrt{s_{xx}}=\sum(x_i-\bar{x})(y_i-\bar{y})/\sqrt{s_{xx}}=\sum(x_i-\bar{x})y_i/\sqrt{s_{xx}}=\mathbf{a}^{\mathsf{T}}\mathbf{y},$$

其中$\mathbf{a}=(x_1-\bar{x},....,x_n-\bar{x})^{\mathsf{T}}/\sqrt{s_{xx}}$，$\|\mathbf{a}\|=1$。

另外，$\sqrt{n}\bar{y}=\mathbf{b}^{\mathsf{T}}\mathbf{y}$，其中$\mathbf{b}=(1,....,1)^{\mathsf{T}}/\sqrt{n}$，$\|\mathbf{b}\|=1$，且$\mathbf{a}\perp\mathbf{b}$.

给定$x_1...,x_n$条件下$\mathbf{a}$，$\mathbf{b}$皆为常数向量，由引理1,

$$\sqrt{n-2}\frac{r}{\sqrt{1-r^2}}\Big|_{x_1...,x_n}=\sqrt{n-2}\frac{\mathbf{a}^{\mathsf{T}}\mathbf{y}}{\sqrt{\sum y_i^{\,2}-\left(\mathbf{b}^{\mathsf{T}}\mathbf{y}\right)^2-\left(\mathbf{a}^{\mathsf{T}}\mathbf{y}\right)^2}}\Big|_{x_1...,x_n}\sim t_{n-2}$$

该分布与条件无关，所以$\sqrt{n-2}\frac{r}{\sqrt{1-r^2}}\sim t_{n-2}$。

下面考虑相关系数的检 验。一般假设检验或显 著性检验的大致步骤：

- 选一个待检验参数的估 计量，
- 适当变换 / 标准化构建检验统计量 ，使得它在原假设下服 从一个标准的分布（与数 据无关、与参数无关）
- 若实际计算得到的检验 统计量在该标准分布的 尾端（小概率），拒绝 原假设。

相关性检验原假设$H_0:\rho_{xy}=0$。当$|r|$较大（$\Leftrightarrow|t|$较大）时，拒绝原假设。我们以t作为检验统计量。是否 可用r作为检验统计量？

> 可以，但其原假设下的分布不常见，而$t\sim t_{n-2}$分布，是一个常见的标准分布（历史原因，Fisher和Student没有定义r的分布）。

假设实际的$|t|$值较大，而从t_{n-2}分布得到一个处于尾端 的值的概率较小（比如<0.05)。因为小概率事件难以 发生，所以“H_0下$t\sim t_{n-2}$分布”这一结论很可能是错的，从 而原假设很可能不成立 。

相关性的精确检验

若$(x_1,y_1),...,(x_n,y_n)$ iid,假设$y_i \mid x_i \sim N(a+bx_i,\sigma^2)$。

原假设为$H_0:\rho_{xy}=0$，检验统计量 $t=\sqrt{n-2}\dfrac{r}{\sqrt{1-r^2}}$，

- 当$|t| \geq t_{n-2}(\alpha/2)$时,在$\alpha$水平下否定原假设.
- 对于给定的检验统计量t的数值，p值$=P(|T| \geq |t|)$, $T \sim t_{n-2}$

注：t检验的表达式说明，在评价r的大小的时候，也要考虑样本量n.

例1. 样本相关系数$r_{xy}=0.3, H_0:\rho_{xy}=0$，样本相关系数是否足 够大到拒绝原假设？ $n=20,\ 50$两种情形下分别检验。

假设数据来自于二元正 态分布，应用精确检验，

若$n=20, t=1.334$，$p=P(|t_{18}| \geq 1.33)=0.199$，不显著；

若$n=50, t=2.179, p=P(|t_{48}| \geq 2.179)=0.034$，显著。

相关性的大样本检验

相关系数的渐近分布

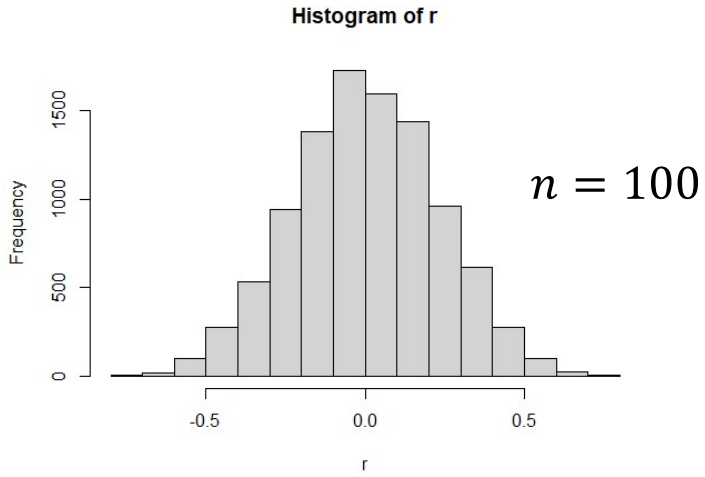
命题2. 若$(x_1,y_1),\ldots,(x_n,y_n)$服从二元正态，总体相关系数为$\rho=\rho_{xy}$，样本相关系数为$r$,则有渐近分布

$$\sqrt{n}\,(r-\rho)\xrightarrow{d}N\left(0,(1-\rho^2)^2\right),\text{当}n\to\infty.$$

证明：略（不要求）。

注：当n足够大时，近似地$r\sim N\left(\rho,(1-\rho^2)^2/n\right)$,

$\rho=0$时，近似地$r\sim N(0,1/n)$.



相关性的大样本检验

二元正态总体假设下， 零假设$H_0:\rho_{xy}=0$的检验统计量取为

$$z=\sqrt{n}\times r,$$

- 检验准则为： 当$|z|\geq z_{\alpha/2}$时在α水平下拒绝原假设 .
- 给定z， p值$\approx P(|X|>|z|)=2(1-\Phi(|z|),X\sim N(0,1).$

注1: 以r与样本量n的综合，即$\sqrt{n}r$， 判定相关与否。

注2: 等价地，可用卡方检验： $z^2=nr^2$，原假设下近似服从χ_1^2

例1(续). 大样本检验

若$n=20,z=\sqrt{n}r=1.34,p=P(|N(0,1)|\geq 1.34)=0.179$， 不显著

若$n=50,z=2.12,p=P(|N(0,1)|\geq 2.12)=0.034$， 显著。

结果与精确检验类似， 事实上，除非样本量n很小，通常大样本检验和精确检验的结果相差 不大。

相关性的置换检验(精确检验)

在总体分布不是正态的情况下，难以求得零假设下t或z的精确分布或大样本渐近分布。相关性的置换检验是一种替代方法，它对数据模型不做任何假设，应用广泛。缺点是需要大量置换，计算耗时，且功效有时候偏低。

置换检验

原始数据：$(x_1, y_1), \ldots, (x_n, y_n) \Rightarrow$ 相关系数 r

随机置换：$(x_{i_1}, y_1), \ldots, (x_{i_n}, y_n) \Rightarrow$ 相关系数 $r^{(per)}$，

其中 $(i_1, i_2, \ldots, i_n)$ 是 $(1,2,\ldots,n)$ 的一个随机置换

反复置换 N 次，得到置换数据的相关系数 $r_k^{(per)}, k = 1, \ldots, N$

p 值 $= \#\{k : |r_k^{(per)}| \geq |r|\} / N.$

置换方法中相关性度量不限于Pearson相关系数，其它可选的相关性统计量包括Pearson卡方，非参数的Kenadall's tau 和 Spearman's rho等。

相关系数的置信区间

虚线框内的内容可忽略

命题2说明，当n较大时，r近似地服从正态分布：$r \sim N(\rho,\ (1-\rho^2)^2/n)$，其方差与均值$\rho$有关，我们称这种现象为“方差不稳定”。基于该分布，我们可以构建ρ的置信水平为$1-\alpha$置信区间：

$$\left\{\rho:\ \left|\frac{r-\rho}{(1-\rho^2)/\sqrt{n}}\right| \le z_{\alpha/2}\right\} \qquad (1)$$

该区间可能是两个区间的并集，其覆盖率不够精确。

Fisher’s z-变换是r的变换，其渐近分布是方差稳定的正态分布，且比r的分布更接近正态（更快地收敛到正态）

Fisher变换

样本相关系数 r的Fisher' s z - 变换(方差稳定化变换):

$$z = \operatorname{atanh}(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right),\ \text{反双曲正切函数}$$

双曲正切函数：

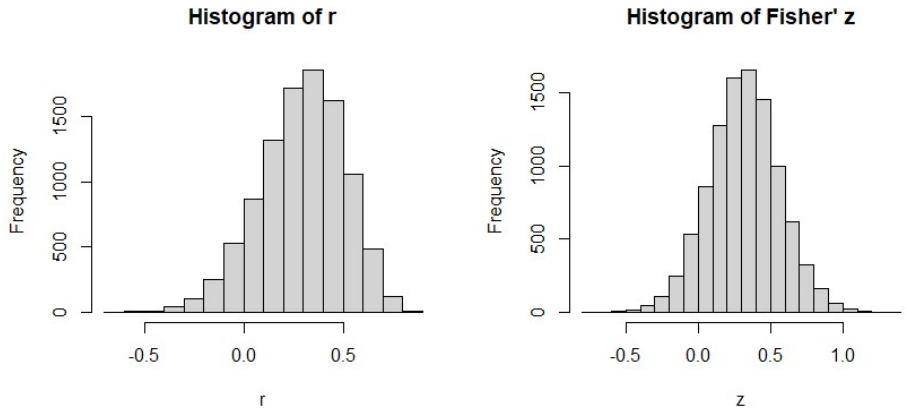
$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

命题3. 假设 $(x_1, y_1), ..., (x_n, y_n)$ iid ~ 二元正态，

设 $\rho = \rho_{xy}$ 为总体相关系数， r 为样本相关系数，则

$$\sqrt{n}\left[\text{atanh}(r) - \text{atanh}(\rho)\right] \xrightarrow{d} N(0,1)，\text{当} n \to \infty$$

证明：基于命题2的结果，应用 *delta* 方法。



Fisher's z 的分布比r更快地收敛到正态分布.

模拟实验：对比r, z的分布

$n = 20$。从二元正态分布($\rho = 0.3$)反复抽样N=10000次, 每次都计算r和Fisher变换 z。N个r，z的直方图如上页所示。代码如下：

```
n=20
rho=0.3
N=10000
r=NULL
 for (i in 1:N){
  x=rnorm(n)
  y=rho*x+sqrt(1-rho^2)*rnorm(n)
  corr=cor(x,y)
  r=c(r,corr)
 }
z=(log(1+r)-log(1-r))/2
par(mfrow=c(1,2))
hist(r, main="Histogram of r",sub="")
hist(z, main="Histogram of Fisher' z",sub="")
```

基于Fisher's z的置信区间

基于Fisher变换的ρ的置信水平为$1-\alpha$的置信区间：

$$\left\{\rho:\ \sqrt{n}\left|\ \operatorname{atanh}(r)-\operatorname{atanh}(\rho)\ \right|\le z_{\alpha/2}\ \right\},\qquad (2)$$

其中$z_{\alpha/2}=\Phi^{-1}(1-\alpha/2),0<\alpha<1.$

注1：基于Fisher变换构造的置信区间(2)比基于r的区间(2)表现更好（更精确的覆盖率或更短的区间长度）。

注2：你可通过模拟实验，对比基于r和基于z的渐近分布构造的置信区间.

注3：基于Fisher变换的检验也比基于r的z检验表现更好。

两样本t检验和Pearson卡方检验都与相关系数有关

两样本t检验

假设第一组 $y_1,..,y_{n_0} iid \sim N(\mu_0,\sigma^2)$, 第二组 $y_{n_0+1},..,y_{n_0+n_1} iid \sim N(\mu_1,\sigma^2)$,

记 $s^2 = \frac{1}{n-2}\left(\sum_{i=1}^{n_0}(y_i-\bar{y}_0)^2 + \sum_{i=n_0+1}^{n_0+n_1}(y_i-\bar{y}_1)^2\right)$, $n = n_0+n_1$,

$H_0:\mu_1=\mu_0$，两样本 t 检验：$t_{\text{twosample}} = \frac{\bar{y}_1-\bar{y}_0}{\sqrt{(1/n_1+1/n_0)s^2}} \overset{H_0}{\sim} t_{n_0+n_1-2}$.

事实(作业)：以x_i表示第i个样本的组号，比如第一组的$x_i=0$，第二组$x_i=1$，记r为$(x_i,y_i), i=1,2,\ldots,n$ 的样本相关系数，则两样本t检验统计量$t_{\text{twosample}} = \sqrt{n-2}r/\sqrt{1-r^2}$

2x2列联表的卡方检验

2×2列联表：

		y 1	y 0	总计
x	1	a	b	n_1
	0	c	d	n_0
	总计	m_1	m_0	n

$Pearson$卡方：$X^2 = \dfrac{n(ad-bc)^2}{n_1 n_0 m_1 m_0}$,

原假设下$(x, y$独立$)$,近似地$X^2 \sim \chi_1^2$.

事实（作业）：假设二元随机样本 $(x_i, y_i), i = 1,2, \ldots, n$ 中 x_i, y_i都是0-1伯努利变量， r为样本相关系数，则Pearson卡方$X^2 = nr^2$.

置换检验、上页作业以及hw1第6题表明：对于二元正态相独立性检验，两正态均值齐一性检验，以及二元伯努利独立性检验，两正态概率齐一性检验等问题，精确检验都可统一为置换检验，大样本检验都可统一为卡方检验$z^2 = nr^2 \sim_{H_0} \chi_1^2$

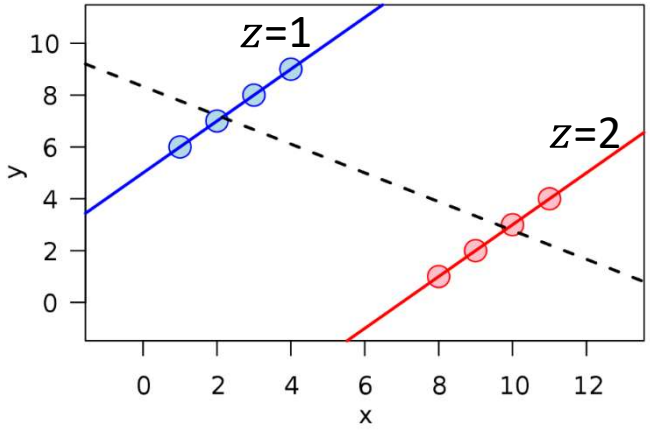
	二元正态	两正态	两二项 (齐一性检验)	二元伯努利 (独立性检验)
数据,模型	$(x_i, y_i), i = 1, \dots, n \sim$ 二元正态，相关系数ρ	$y_1, \dots, y_{n_1} \sim N(\mu_1, \sigma^2)$ $y_{n_1+1}, \dots, y_n \sim N(\mu_2, \sigma^2)$ $n = n_1 + n_2$ 两组的标号$x_i = 1$或0	$y_1, \dots, y_{n_1} \sim B(1, p_1)$ $y_{n_1+1}, \dots, y_n \sim B(1, p_2)$ $n = n_1 + n_2$ 两组的标号$x_i = 1$或0	$(x_i, y_i), i = 1, \dots, n$ 服从二元伯努利， 相关系数ρ
原假设	H_0: ρ=0 $(x, y$独立)	H_0: μ_1=μ_2	H_0: p_1=p_2	H_0: ρ=0 $(x, y$独立)
精确检验 /置换检验	下述精确检验（基本上）等价于置换检验			
	$t = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}} \sim t_{n-2}$	$t = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^2}}$ $= \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}} \sim t_{n-2}$	Fisher's exact test	Fisher's exact test
大样本检验 /卡方检验	原假设下近似地，$z^2 = nr^2 \sim \chi_1^2$			
	$t \approx z$ $= \sqrt{n} r \sim N(0,1)$	$t \approx z = \sqrt{n} r \sim N(0,1)$	$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) p(1-p)}}$ $= \sqrt{n} r \sim N(0,1)$	Pearson卡方 $X^2 = nr^2 \sim \chi_1^2$

Pearson卡方检验

本课程将上述检验拓广到控制干扰因素的情形。

辛普森悖论(Simpson's paradox)

辛普森悖论是一种组内趋势与各组合并后趋势不同或相反的现象。这里组别是干扰因素，分组意味着控制干扰因素。如下图，组内负相关（实线），所有数据呈负相关（虚线）：分组变量z既与x有关，也与y有关，是个干扰因素。 https://en.wikipedia.org/wiki/Simpson%27s_paradox



不控制z (虚线)： $y = a + bx + \varepsilon$

控制z (实线)： $y = a + bx + cz + \varepsilon$

回归分析中辛普森悖论描述的是这样的现象：在研究y, x 的关系时，控制干扰因素 z与不控制z时结果不一致。

例2. 加州大学Berkeley分校1973年研究生招生的女生录取率 显著低于男生的录取率 44%。校方担心被控告存在女性歧视。但分别考察各个系的录取率，反而大部分是女生录取率高。这里我们构造一个简单例子演示这种现象（分母为申请人数，分子为录取人数）：

1系录取率：$4/10$ （男）$< 1/2$ （女）
2系录取率：$2/10$ （男）$< 5/20$ （女）
合并两个系：$(4+2)/(10+10)$ （男）$> (1+5)/(2+20)$ （女）

各个系录取过程中有均衡男女人数的趋势：如果女生申请人数较少，则倾向于多录取女生，录取率就会偏高。

参考：

1. Bickel, Hammel and O'Connell (1975) Sex Bias in Graduate Admissions: Data From Berkeley". *Science*. 187 (4175): 398–404.
2. Freedman,Pisani,Purves (2007) *Statistics*, Norton P17. 第一章下载英文、中文版: x=2.5, 2.6

Berkeley实际数据如下：女生录取率 35%，男生录取率 44%（p值=0）。

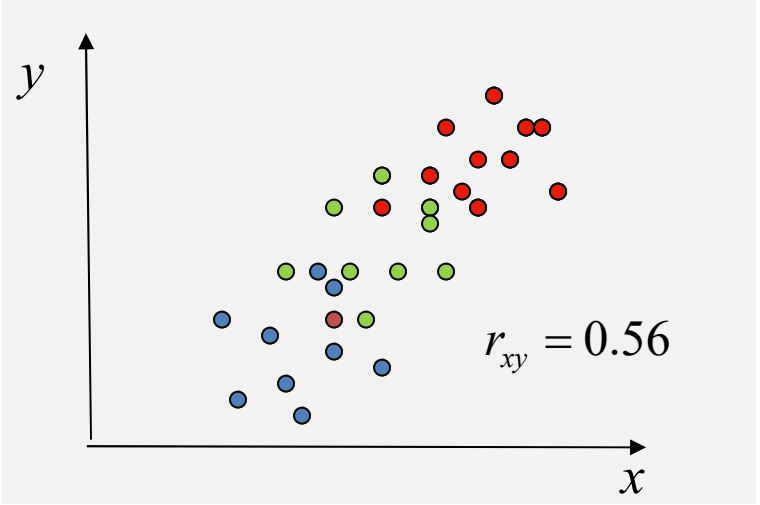
	所有人		男生		女生	
	申请人数	录取率	申请人数	录取率	申请人数	录取率
总计	12,763	41%	8442	**44%**	4321	**35%**

但从各个系的数据来看，并没有发现明显的歧视证据，例如6个最大的系的录取情况如下，多数系女生的录取率高于男生（红色）。

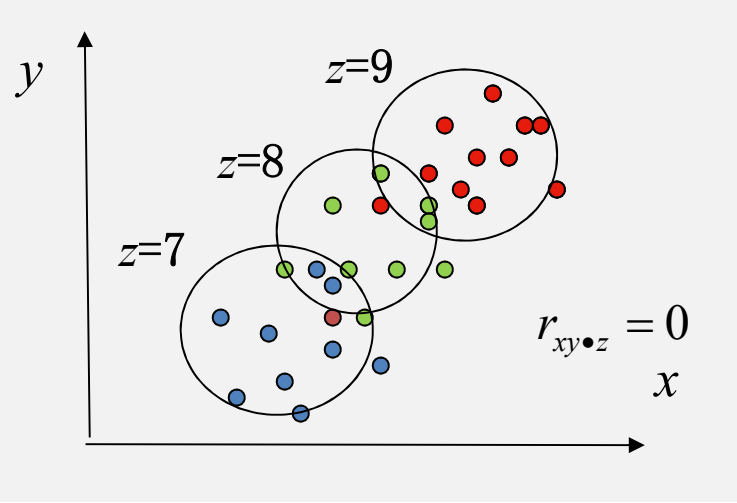
系	所有人		男生		女生	
	申请人数	录取率	申请人数	录取率	申请人数	录取率
A	933	64%	*825*	62%	108	**82%**
B	585	63%	*560*	63%	25	**68%**
C	918	35%	325	**37%**	*593*	34%
D	792	34%	417	33%	375	**35%**
E	584	25%	191	**28%**	*393*	24%
F	714	6%	373	6%	341	**7%**

Bickel et al. (1975)认为多数女生申请了竞争比较激烈的专业（C,D,E,F），拉低了女生的总录取率。女生录取率很高的两个系（A,B,工程、计算机）女生申请人数偏少，对女生总体录取率影响不大。

例3. 调查100名 7-9岁儿童(z=7, 8, 9)，发现阅读能力y与身高x正相关，相关系数 r_{xy} = 0.56 （右图）.



但如果考虑年龄组,z = 7,8,9, 则发现各个组内x, y不相关 （下图）



x, y 都随年龄增加而增加，使得x, y 呈现出正相关现象。年龄z是干扰因素。

控制年龄，即给定年龄 z 时，x 与 y 条件独立： $P(x, y|z) = P(x|z)P(y|z)$

如何控制z?

(1) 我们在线性回归模型中添加z项

(2) 计算控制 $\boldsymbol{z}$ 条件下 x, y 的偏相关系数。

两者等价。

偏相关系数(partial correlation coefficient)

三元情形下的偏相关系数

当x,y,z都是一维随机变量时,相关系数矩阵$R=\begin{pmatrix}1 & \rho_{xy} & \rho_{xz}\\ \rho_{xy} & 1 & \rho_{yz}\\ \rho_{xz} & \rho_{yz} & 1\end{pmatrix}$,

其中ρ_{ab}代表随机变量a,b的Pearson相关系数，则偏相关系数$\rho_{xy\bullet z}$定义为

$$\rho_{xy\bullet z}=\frac{\rho_{xy}-\rho_{xz}\rho_{yz}}{\sqrt{1-\rho_{xz}^2}\sqrt{1-\rho_{yz}^2}}$$

注1: $|\rho_{xy\bullet z}|\le 1$

这是因为$\det(R)\ge 0\Rightarrow 1-\rho_{xy}^2-\rho_{yz}^2-\rho_{zx}^2+2\rho_{xy}\rho_{xz}\rho_{yz}\ge 0\Leftrightarrow|\rho_{xy\bullet z}|\le 1$

注2: 后面我们将会定义一般情况下($\mathbf{z}$为向量)的偏相关系数。

例3（续）．阅读能力(y)、身高(x)和年龄(z)的相关系数矩阵如下，求偏相关系数 $r_{xy\bullet z}$，检验阅读与身高是否相关（n=100）。

	x身高	y阅读	z年龄
x	1	0.56	0.8
y	0.56	1	0.7
z	0.8	0.7	1

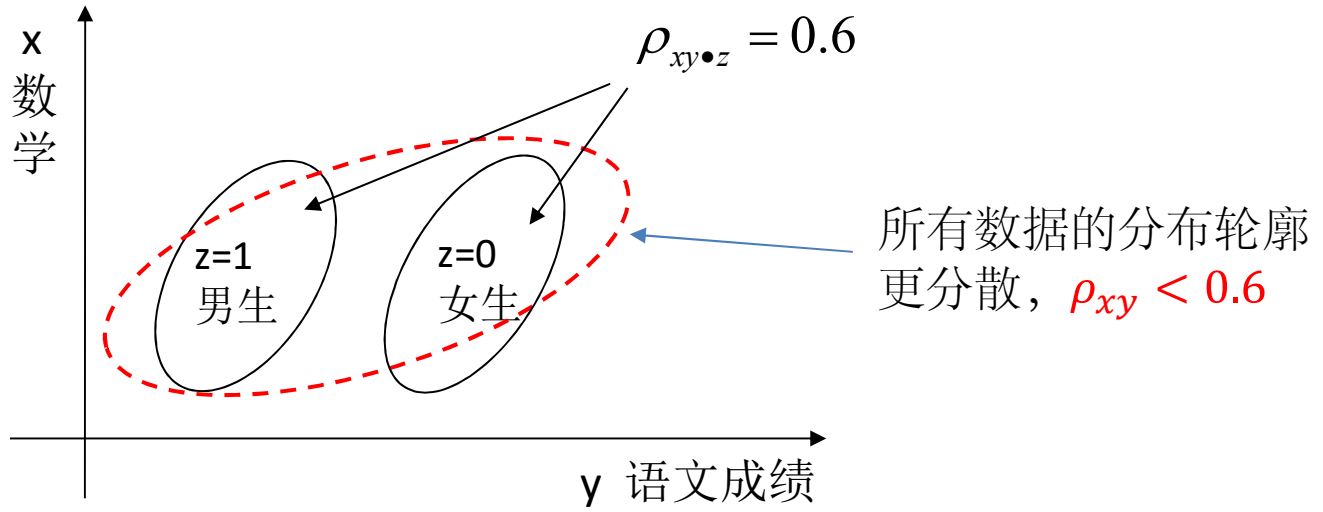
$$r_{xy\bullet z}=\frac{r_{xy}-r_{xz}r_{yz}}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{yz}^2}}=\frac{0.56-0.8\times 0.7}{\sqrt{1-0.8^2}\sqrt{1-0.7^2}}=0$$

即，控制年龄之后，阅读能力和身高相关系数为0。

例4. 高中男、女生的数学成绩 x与语文成绩 y 的相关系数都为0.6，假设男女生数学成绩没有差异，但女生的语文成绩要好于男生。如果男女生混合在一起，得到的相关系数是大于、等于还是小于0.6？

以$z=1$、0分别代表男、女。已知$\rho_{xy\bullet z}=0.6$，$\rho_{xz}=0$ (数学与性别不相关)

故$0.6=\rho_{xy\bullet z}=\dfrac{\rho_{xy}}{\sqrt{1-\rho_{yz}^2}}\geq\rho_{xy}$



总结：

- 考察样本相关系数大小的时候也要考虑样本量。具体地，
  - $|t| = \left|\frac{\sqrt{n-2}r}{\sqrt{1-r^2}}\right| > t_{n-2}(\alpha/2)$?
  - $|z| = |\sqrt{n}r| > z\,(\alpha/2)$?
- 偏相关系数是控制干扰因素条件下的相关系数 。
- 相关系数≈ 简单回归（仅一个自变量）；
  偏相关系数≈多重回归（多个自变量。