

第二讲 相关系数

2023.9.15

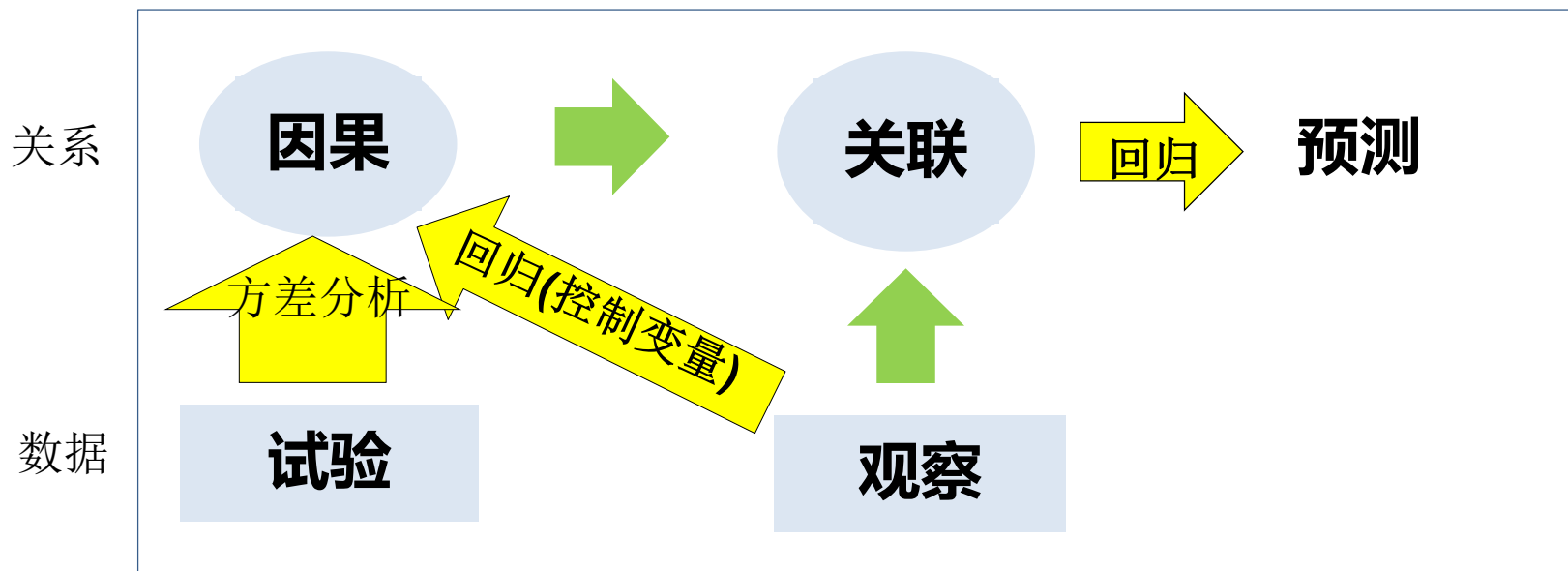
问题：样本相关系数 $r_{xy} = 0.2$ 否足够大，是否表明 x 和 y 相关？

Recap

随机化控制试验推断因果：

- 原因 x 变化：有1（处理组）且有0（对照组）才能比较。
- 原因 x 随机地变化： x 随机地取值，保证被比较的对象“相同”。

课程主要内容



研究变量之间的关系的两种不同但有内在联系的方法：

- 相关分析：无方程模型，对称，简洁。
 - 两个变量的相关系数；
 - 多个变量情形下的偏相关系数；
- 基于模型的回归分析：区分响应变量和自变量（位居方程两端）。
 - 两个变量的简单回归模型；
 - 多个自变量的多重回归模型。

课程内容将大致按上述顺序进行。

Pearson相关系数

Pearson相关系数度量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:

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}$$

记号: $s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x})y_i$,

$$s_{xx} = \sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x})x_i = \sum x_i^2 - n\bar{x}^2.$$

样本相关系数

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

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

性质: $|\rho_{xy}| \leq 1, |r_{xy}| \leq 1$ (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 $\mathbf{x} = (x_1, \dots, x_n)^\top$, $\mathbf{y} = (y_1, \dots, y_n)^\top$, 它们的中心化记为:

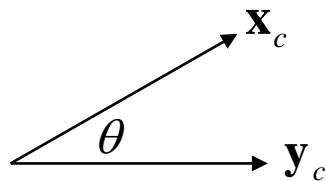
$$\mathbf{x}_c = (x_1 - \bar{x}, \dots, x_n - \bar{x})^\top = \mathbf{x} - \mathbf{1}\bar{x},$$

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

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:

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



高尔顿定义的相关系数: $g = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}} = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$

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.

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

假设总体均值 μ ，方差 σ^2 ，样本均值 \bar{x} ，样本方差 s^2 ，样本量 n ，我们熟知：

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

样本方差的方差不太常见：

$$\text{var}(s^2) = \frac{1}{n} \left(E(x - \mu)^4 - \frac{n-3}{n-1} \sigma^4 \right) + O(1/n^2)$$

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

虚线框内容
仅供参考

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

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

(参见 Lehmann 《点估计理论》)

独立性假设的精确检验

为了求 r 的分布，我们需要如下引理

引理1: 若随机向量 $\mathbf{y} \sim N_n(0, I_n)$, 即 $y_1, y_2, \dots, y_n \text{ iid} \sim N(0,1)$, 假设 O 是任一正交矩阵, 则 $O\mathbf{y} \sim N_n(0, I_n)$ 。

证明: \mathbf{y} 的密度 $f(\mathbf{y}) = 1/(2\pi)^{n/2} \exp(-\|\mathbf{y}\|^2 / 2)$ 。

引理2. 假设随机向量 \mathbf{x} 的分量 $x_1, \dots, x_n \text{ iid} \sim N(0,1)$, 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 $\mathbf{a}^\top \mathbf{x} \sim N(0,1)$, $\|\mathbf{x}\|^2 - (\mathbf{a}^\top \mathbf{x})^2 \sim \chi_{n-1}^2$,

两者独立, 且 $\sqrt{n-1} \frac{\mathbf{a}^\top \mathbf{x}}{\sqrt{\|\mathbf{x}\|^2 - (\mathbf{a}^\top \mathbf{x})^2}} \sim t_{n-1}$

(2) $\|\mathbf{a}\|=\|\mathbf{b}\|=1, \mathbf{a}^\top \mathbf{b}=0$, 则 $\mathbf{a}^\top \mathbf{x} \sim N(0,1)$, $\|\mathbf{x}\|^2 - (\mathbf{a}^\top \mathbf{x})^2 - (\mathbf{b}^\top \mathbf{x})^2 \sim \chi_{n-2}^2$,

两者独立且 $\sqrt{n-2} \frac{\mathbf{a}^\top \mathbf{x}}{\sqrt{\|\mathbf{x}\|^2 - (\mathbf{a}^\top \mathbf{x})^2 - (\mathbf{b}^\top \mathbf{x})^2}} \sim t_{n-2}$

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

$$\text{令 } \mathbf{y} = A\mathbf{x} = \begin{pmatrix} \mathbf{a}^\top \mathbf{x} \\ * \\ y_n \end{pmatrix} \stackrel{\text{记作}}{=} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \text{ 则 } \mathbf{y} \sim N_n(\mathbf{0}, I_n). \text{ 注意}$$

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

(b) 由 $\|\mathbf{x}\|^2 = \|\mathbf{y}\|^2 \Rightarrow \|\mathbf{x}\|^2 - (\mathbf{a}^\top \mathbf{x})^2 = \|\mathbf{y}\|^2 - y_1^2 = y_2^2 + \dots + y_n^2 \sim \chi_{n-1}^2$,

因为 $\|\mathbf{x}\|^2 - (\mathbf{a}^\top \mathbf{x})^2$ 只与 y_2, \dots, y_n 独立，所以它与 $\mathbf{a}^\top \mathbf{x} = y_1$ 独立

(c) 由t分布的定义知

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

(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 $(x_1, y_1), \dots, (x_n, y_n)$ 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: 我们应用引理2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 $\mathbf{y} = (y_1, \dots, y_n)^\top \sim N(0, I_n)$ 。

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

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

下面我们在给定 x_1, \dots, x_n 的条件下计算 t 的分布。

其中，分子项

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

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

另外, $\sqrt{n}\bar{y} = \mathbf{b}^\top \mathbf{y}$, 其中 $\mathbf{b} = (1, \dots, 1)^\top / \sqrt{n}$, $\|\mathbf{b}\| = 1$, 且 $\mathbf{a} \perp \mathbf{b}$ 。

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

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

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

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，即 p 值较小（小概率），则拒绝原假设。

独立性原假设 / 零假设：

$$H_0 : x, y \text{ 独立,}$$

当总体为正态时， x, y 独立 $\Leftrightarrow \rho_{xy} = 0$ 。我们基于 r 构建检验统计量：

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

即使总体不是正态，我们也将基于 r 构造检验： $z = \sqrt{nr}$ 。

相关性的 精确检验

若 $(x_1, y_1), \dots, (x_n, y_n)$ iid ~ 二元正态, 原假设为 $H_0: \rho_{xy} = 0$ 。

取检验统计量 $t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}$,

- 当 $|t| \geq t_{n-2}(\alpha/2)$ 时, 在 α 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。

两样本t检验可表示为相关性检验

两样本t检验:

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

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

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

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 $(x_i, y_i), i = 1, 2, \dots, n$ 的样本相关系数, 则两样本t检验统计量 $t_{\text{twosample}} = \sqrt{n-2}r/\sqrt{1-r^2}$

补充：渐近分布

依分布收敛不是传统意义上的收敛， $x_n \xrightarrow{d} x$ 并不表示 $n \rightarrow \infty$ 时 x_n 无限接近 x ，故依分布收敛也称为弱收敛。渐近正态或渐近卡方理论的证明一般基于

中心极限定理 \oplus Slusky引理（大数律或切比雪夫型不等式）

或delta方法（渐近正态随机变量的函数依然渐近正态）。

具体地，为了证明 $u_n \xrightarrow{d} N(0,1)$ ，可将 u_n 表示成（比如Taylor展开）

$$u_n = v_n + w_n, \text{ 其中 } v_n = \text{独立和}, \text{ 而 } w_n \xrightarrow{P} 0,$$

前者应用中心极限定理，依分布收敛到正态，后者由大数律或者切比雪夫型不等式依概率收敛到0，不影响极限分布(Slusky)。

中心极限定理：假设 x_1, x_2, \dots, x_n 独立，记 $E_n = E(\sum_{i=1}^n x_i)$, $V_n = \text{var}(\sum_{i=1}^n x_i)$,

则在较弱条件下，
$$\frac{\sum_{i=1}^n x_i - E_n}{\sqrt{V_n}} \xrightarrow{d} N(0,1).$$

条件：

(1) iid 或

(2) 高阶矩存在，或

(2) $n \rightarrow \infty$ 时, $\max_{1 \leq i \leq n} \text{var}(x_i)/V_n \rightarrow 0$ (必要条件 $V_n \rightarrow \infty$)

Slusky引理：假设随机变量或向量 $x_n \xrightarrow{d} x$, $y_n \xrightarrow{P} c$ (常数), 则

$$x_n + y_n \xrightarrow{d} x + c, \quad x_n y_n \xrightarrow{d} xc, \quad x_n / y_n \xrightarrow{d} x / c$$

例2. 假设 x_1, \dots, x_n iid, $E(x_1) = \mu, \text{var}(x_1) = \sigma^2, s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, 则

$$(1) \sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2),$$

$$(2) \sqrt{n}(s^2 - \sigma^2) \xrightarrow{d} N(0, E(x - \mu)^4 - \sigma^4).$$

证: (1) 由中心极限定理 $\frac{\sum_{i=1}^n x_i - n\mu}{\sqrt{n\sigma^2}} = \sqrt{n}(\bar{x} - \mu) / \sigma \xrightarrow{d} N(0, 1)$.

$$(2) \sqrt{n}(s^2 - \sigma^2) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (x_i - \mu)^2 - \sqrt{n}(\bar{x} - \mu)^2,$$

由中心极限定理 $\frac{\sum_{i=1}^n (x_i - \mu)^2 - n\sigma^2}{\sqrt{n(E(x - \mu)^4 - \sigma^4)}} \xrightarrow{d} N(0, 1)$.

由大数定律 $\bar{x} - \mu \xrightarrow{P} 0$, 而 $\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2)$, 由Slusky引理

$\sqrt{n}(\bar{x} - \mu)^2 = [\sqrt{n}(\bar{x} - \mu)](\bar{x} - \mu) \xrightarrow{P} 0$, 再由Slusky引理即得。

独立性假设的大样本检验

命题2. 若 $(x_1, y_1), \dots, (x_n, y_n)$ iid, 存在有限二阶矩, 样本相关系数为 r 。
假设 x_i, y_i 独立 (\Rightarrow 总体相关系数 $\rho = 0$), 则当 $n \rightarrow \infty$ 时,

$$\sqrt{n} r \xrightarrow{d} N(0,1).$$

命题2的证明: 首先考虑独立和 $\sum (x_i - \mu_x)(y_i - \mu_y)$ 的极限分布, 然后再将总体均值替换成样本均值 (Slusky引理)。给定 \mathbf{x} 条件下,

$$E\left(\sum (x_i - \mu_x)(y_i - \mu_y) \mid \mathbf{x}\right) = 0, \quad \text{var}\left(\sum (x_i - \mu_x)(y_i - \mu_y) \mid \mathbf{x}\right) = \sum (x_i - \mu_x)^2 \sigma_y^2.$$

由中心极限定理, 给定所有 \mathbf{x} 条件下

$$\frac{\sum (x_i - \mu_x)(y_i - \mu_y) - 0}{\sqrt{\sum (x_i - \mu_x)^2 \sigma_y^2}} \xrightarrow{d} N(0,1),$$

极限分布与 \mathbf{x} 无关, 所以无条件上下式也成立。因为

$$\bar{x} \xrightarrow{P} \mu_x, \quad \bar{y} \xrightarrow{P} \mu_y, \quad \sum (y_i - \bar{y})^2 / n \xrightarrow{P} \sigma_y^2,$$

可将上式 μ_x, μ_y, σ_y^2 分别替换为 $\bar{x}, \bar{y}, \sum (y_i - \bar{y})^2 / n$ (Slusky's lemma)

$$\frac{\sum (x_i - \bar{x})(y_i - \mu_y)}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2 / n}} = \sqrt{n} r \xrightarrow{d} N(0,1)$$

由命题2，独立性假设“ $H_0: x_i$ 与 y_i 独立”的检验可构造如下

“ $H_0: x_i$ 与 y_i 独立”的大样本检验：

当 $|\sqrt{nr}| > z_{\alpha/2}$ 时，在水平 α 下拒绝原假设。

注1: 以 r 与样本量 n 的综合，即 \sqrt{nr} ，判定相关系数大小。

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). 大样本检验结果与精确检验类似

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

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

2x2列联表的卡方检验

假设 $(x_i, y_i), i = 1, 2, \dots, n$ iid, x_i, y_i 都是二值(0-1)变量, 数据统计成如下 2×2 列联表

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

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

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

作业: 假设二元随机样本 $(x_i, y_i), i = 1, 2, \dots, n$ 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 $X^2 = nr^2$.

独立性假设的置换检验(精确检验)

当零假设下检验统计量的精确分布或大样本渐近分布难以求得的时候，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。缺点是计算耗时且功效有时候偏低。

置换检验

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

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

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

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

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

命题2给出了独立条件下相关系数的渐近分布，为了构造相关系数的置信区间，需要总体相关系数 $\rho \neq 0$ 情形下的分布。应用中心极限定理和delta方法，可以证明如下一般结论（参见Ferguson: A course in large sample theory, 1996）“

定理 A1.1 假设 $(x_i, y_i), i = 1, \dots, n$ iid 服从均值为0、相关系数为 ρ 且四阶矩存在的二元分布，则当 $n \rightarrow \infty$ 时

$$\sqrt{n}(r - \rho) \xrightarrow{d} N(0, \gamma^2).$$

$$\text{其中 } \gamma^2 = \frac{1}{4}\rho^2 c_1 - \rho c_2 + c_3, \quad c_1 = \frac{\text{var}(x^2)}{\sigma_x^4} + 2 \frac{\text{COV}(x^2, y^2)}{\sigma_x^2 \sigma_y^2} + \frac{\text{var}(y^2)}{\sigma_y^4}, \quad c_2 = \frac{\text{COV}(x^2, xy)}{\sigma_x^3 \sigma_y} + \frac{\text{COV}(y^2, xy)}{\sigma_x \sigma_y^3}, \quad c_3 = \frac{\text{COV}(xy, xy)}{\sigma_x^2 \sigma_y^2}.$$

注1：当 x_i, y_i 独立时,可验证 $\gamma = 1$, 即命题2.

注2：理论上可以依据该分布构造相关系数的置信区间，但过于复杂。下面仅考虑总体为二元正态的情形。

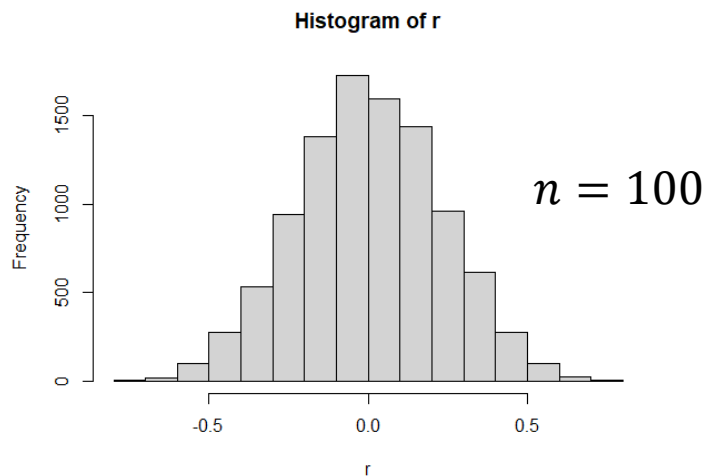
正态总体情形

命题3. 若 $(x_1, y_1), \dots, (x_n, y_n)$ 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

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

证明: 定理A1.1的推论。

注: 当 n 足够大时, 近似地 $r \sim N(\rho, (1 - \rho^2)^2 / n)$,
 $\rho = 0$ 时, 近似地 $r \sim N(0, 1/n)$.



相关系数的 置信区间

基于命题3, 我们可以构建 ρ 的置信水平为 $1-\alpha$ 置信区间:

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

该区间可能是两个区间的并集, 其覆盖率不够精确。

方差稳定

命题2说明, 当 n 较大时, r 近似地服从正态分布:

$$r \sim_{\text{近似}} N(\rho, (1 - \rho^2)^2 / n),$$

其方差与均值 ρ 有关, 我们称这种现象为“方差不稳定”。

一个好的正态分布, 均值和方差应该是两个独立的参数,

即均值/位置改变时, 方差/形状不受影响。

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):

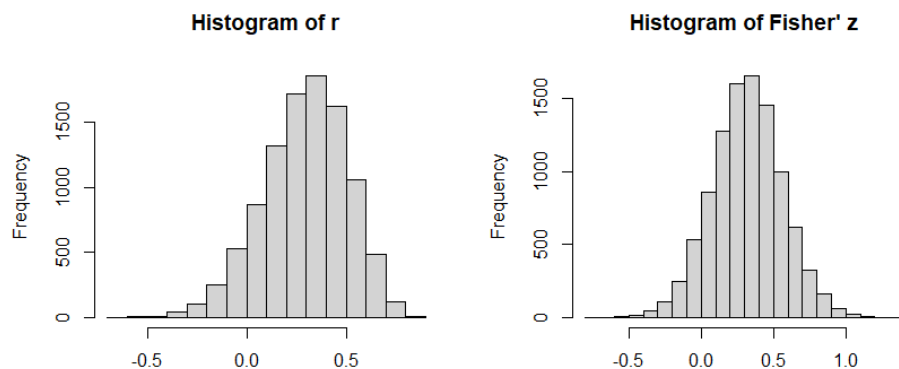
$$z = \operatorname{atanh}(r) = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right), \text{ 反双曲正切函数}$$

双曲正切函数:
$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

命题4. 假设 $(x_1, y_1), \dots, (x_n, y_n)$ iid \sim 二元正态,
设 $\rho = \rho_{xy}$ 为总体相关系数, r 为样本相关系数, 则

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

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。

基于Fisher's z 的置信区间

二元正态总体假设下，基于Fisher变换的 ρ 的置信水平为 $1-\alpha$ 的置信区间：

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

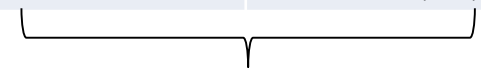
其中 $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2), 0 < \alpha < 1$.

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

独立性检验总结如下

(1) 精确检验: t-检验/置换检验; (2) 大样本检验: 卡方检验

	一般二元总体	二元正态	两正态	二元伯努利 (独立性检验)	两二项 (齐一性检验)
模型	$(x_i, y_i),$ $i = 1, \dots, n$	$(x_i, y_i), i =$ $1, \dots, n \sim$ 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	$(x_i, y_i), i = 1, \dots, n$ 服从二元伯努利, 相关系数 ρ	$y_1, \dots, y_{n_1} \sim B(1, p_1)$ $y_{n_1+1}, \dots, y_n \sim B(1, p_2)$ $n = n_1 + n_2$ 两组的标号 $x_i = 1$ 或 0
原假设	$H_0: x, y$ 独立	$H_0: \rho=0$ (x, y 独立)	$H_0: \mu_1=\mu_2$	$H_0: \rho=0$ (x, y 独立)	$H_0: p_1=p_2$
精确检 验	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 (\Leftrightarrow 置换检验)	Fisher's exact test (\Leftrightarrow 置换检验)
大样本 卡方检 验	$z = \sqrt{nr} \sim N(0,1)$ \Leftrightarrow $z^2 = nr^2 \sim \chi_1^2$			Pearson卡方 $X^2 = nr^2 \sim \chi_1^2$	$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) p(1-p)}}$ $= \sqrt{nr} \sim N(0,1)$



Pearson卡方检验

本课程将上述检验拓广到多组、存在干扰因素的情形。

除了 Pearson 相关系数，还有更稳健的关联度量：
Kenadall's tau 和 Spearman's rho，
是否还有其它非线性的关联度量？