第 16 章 总复习 ·

一、 概率分布的关联

在概率分布的关联图(图 2)，我们从柏努利分布 Bern(p) 开始说明:

1. 第 1 次成功前的「失败的次数」是几何分布 G(p)

2. 第 k 次成功前的「失败的次数」是负二项分布 NB(k,p)

3. n 次柏努利试验次数中，「成功的次数」是二项分布 B(n,p)

4. 卜松分布 Pois(λ)是单位时间内成功的次数，则 t 时间内「成功的次数」，为卜松分布 Pois(λt)

5. 卜松分布的第 1 次出现成功的「时间」是指数分布 Exp(λ)

6. 卜松分布的第 k 次出现成功的「时间」是伽玛分布 Gamma(k,λ)

7. 柏努利分布、几何分布与负二项分布的三角关系，相当于卜松分布、指数分布与伽玛分布的三角关系 *Bern(p)∴ G(p)∴ NB(k,p) = Pois(λ)∴ Exp(λ)∴ Gamma(k,λ)*

8. 伽玛分布 Gamma(k,λ)，当 $n = 2k$，$\lambda = \frac{1}{2}$ ，则为卡方分布 χ_n^2

9. 正态分布 $N(\mu,\sigma^2)$，利用 $\frac{X-\mu}{\sigma}$，转换为标准正态分布 *N(0,1)*

10. 抽样检验有关的主要概率分布之四角关系:

    标准正态分布 N(0,1)、 t 分布 t_n、卡方分布 χ_n^2、F 分布 $F_{n,m}$

11. k 个独立标准正态分布 $Z_i \sim N(0,1)$ 的平方和 $\sum_{i=1}^{k} Z_i^2$，为卡方分布 χ_k^2

12. t 分布 t_n 的平方 $(t_n)^2$ 是 F 分布 $F_{1,n}$

13. t 分布 t_k，当自由度 k 相当大时，近似标准正态分布 N(0,1)

14. 若 $X \sim \chi_n^2$，$Y \sim \chi_m^2$，则 $\frac{X/n}{Y/m}$ 是 F 分布 $F_{n,m}$

15. 若 $X \sim \chi_n^2$， $Z \sim N(0,1)$，则 $\frac{Z}{\sqrt{X/n}}$ 是 t 分布 t_n

16. $X_1, X_2, \cdots, X_n$ iid ~ U(0,1)，独立相同连续均匀分布的第 j 排序统计(order statistics)

    $X_{(j)} \sim Beta(j, n-j+1)$

17. 超几何分布 $HG(N,Np,n)$，当 N 很大时，则近似二项分布 $B(n,p)$

18. 二项分布 B(n,p)，当 np > 5 且 n(1-p) 时，则近似正态分布 $N(np,npq), q=1-p$

19. 二项分布 B(n,p)，当 $n \to \infty$ 和 $p \to 0$ 时，则近似卜松分布 Pois(np)

20. 卜松分布 Pois(λ)，当 λ > 5 时，则近似正态分布 N(λ,λ)

21. 若 $X \sim \chi_n^2$，当 $n \to \infty$，则 $\sqrt{2X} \to N(\sqrt{2n-1},1)$

22. 当 $m \to \infty$，则 $F_{n,m} \to F_{n,\infty} = \dfrac{\chi_n^2}{n}$

23. 对数正态分布 $Y \sim LN(\mu,\sigma^2)$ 的对数 $log\, Y$，是正态分布 $N(\mu,\sigma^2)$

24. 下列分布有相加性： X, Y 是独立的

$X \sim B(n_1,p),\ Y \sim B(n_2,p) \Rightarrow X+Y \sim B(n_1+n_2,p)$

$X \sim NB(n_1,p),\ Y \sim NB(n_2,p) \Rightarrow X+Y \sim NB(n_1+n_2,p)$

$X \sim Pois(\lambda_1),\ Y \sim Pois(\lambda_2) \Rightarrow X+Y \sim Pois(\lambda_1+\lambda_2)$

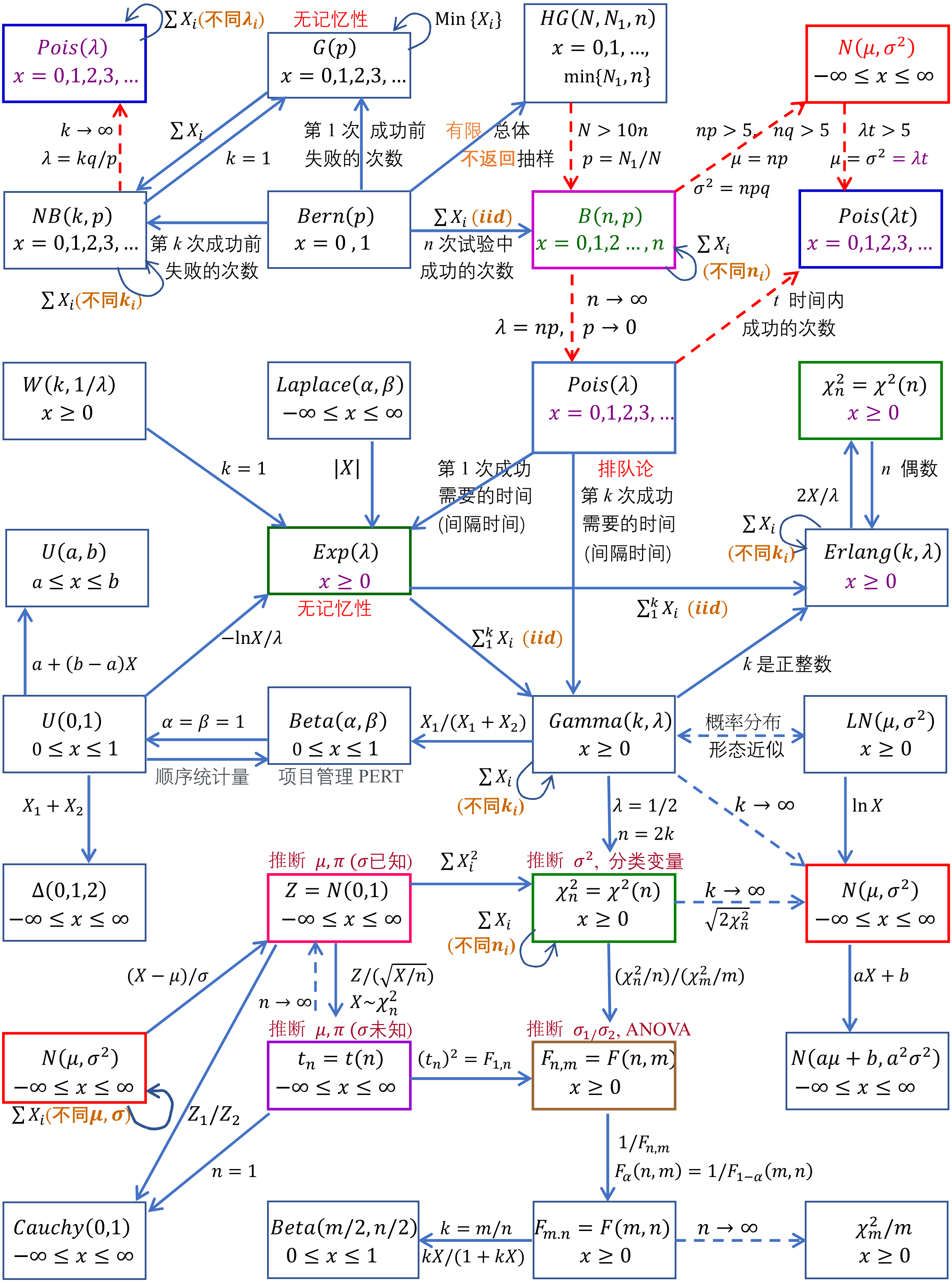
$X \sim N(\mu_1,\sigma_1^2),\ Y \sim N(\mu_2,\sigma_2^2) \Rightarrow X+Y \sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$

$X \sim \chi_n^2,\ Y \sim \chi_m^2 \Rightarrow X+Y \sim \chi_{n+m}^2$

$X \sim Gamma(\alpha_1,\beta),\ Y \sim Gamma(\alpha_2,\beta) \Rightarrow X+Y \sim Gamma(\alpha_1+\alpha_2,\beta)$

$X \sim F_{n_1,m},\ Y \sim F_{n_2,m} \Rightarrow X+Y \sim F_{n_1+n_2,m}$

$X \sim U(0,1),\ Y \sim U(0,1) \Rightarrow X+Y \sim \Delta(0,1,2)$



概率分布关联图　陈文贤 著作 可供教学，请勿引用出版。

概率符号定义 请见 《大话统计学》， 表示 独立随机变量之线性组合仍为 相同分布。

二、 概率分布汇总表

概率分布的定义、参数、期望值、变异数、与动差母函数如下表:

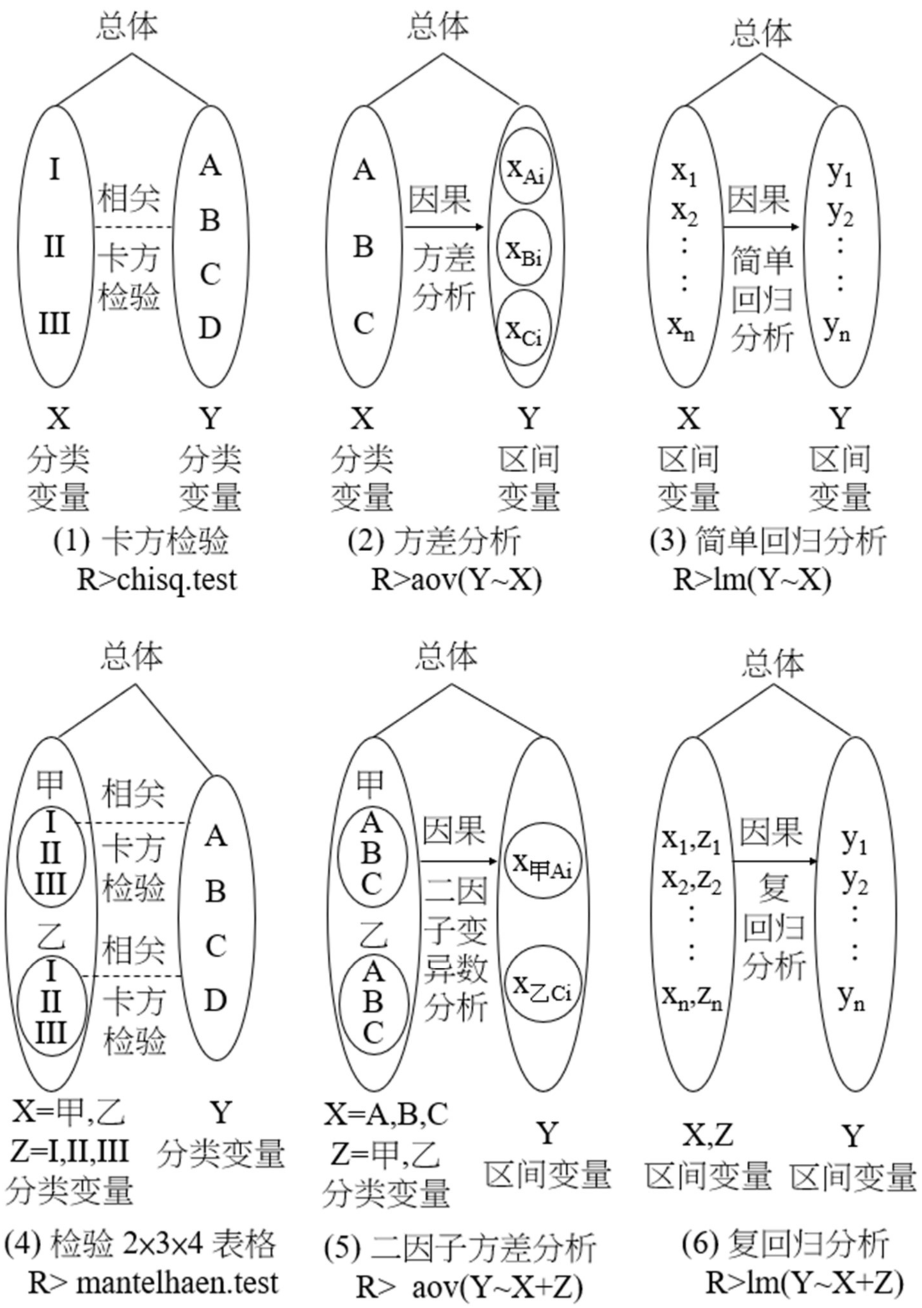
機率函數	機率密度	定義域	參數	期望值	變異數	動差母函數
Disc. Unif.	$\frac{1}{N}$	$x = 1, 2, \ldots, N$	N	$\frac{N+1}{2}$	$\frac{N^2-1}{12}$	$\frac{e^t(1-e^{Nt})}{N(1-e^t)}$
Bernoulli	$p^x q^{1-x}$	$x = 0, 1$	$p \in [0, 1]$	p	pq	$q + pe^t$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	$x = 0, 1, \ldots, n$	$p \in [0, 1]$, n	np	npq	$(q + pe^t)^n$
Poisson	$\frac{\lambda^x e^{-\lambda}}{x!}$	$x = 0, 1, 2, \ldots$	$\lambda > 0$	λ	λ	$e^{\lambda(e^t-1)}$
Geometric	$p(1-p)^x$	$x = 0, 1, 2, \ldots$	$p \in [0, 1]$	$\frac{q}{p}$	$\frac{q}{p^2}$	$\frac{p}{1-qe^t}$
Hypergeom.	$\frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}}$	$0, \ldots, \min\{n, N_1\}$	N, N_1, n	$\frac{nN_1}{N}$	§ 6.1.6	不適用
Negat. Bino.	$\binom{x+k-1}{x} p^k (1-p)^x$	$x = 0, 1, 2, \ldots$	$p \in [0, 1]$, k	$\frac{kq}{p}$	$\frac{kq}{p^2}$	$\left(\frac{p}{1-qe^t}\right)^k$
Uniform	$\frac{1}{b-a}$	$a \le x \le b$	a, $b \in R$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$-\infty \le x \le \infty$	$\mu \in R$, $\sigma > 0$	μ	σ^2	$e^{\mu t + \frac{1}{2}\sigma^2 t^2}$
Lognormal	$\frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$	$x \ge 0$	$\mu > 0$, $\sigma > 0$	$e^{\mu+\frac{\sigma^2}{2}}$	§ 6.2.3	$\mu_r' = e^{r\mu + \frac{1}{2}r^2\sigma^2}$
Exponential	$\lambda e^{-\lambda x}$	$x \ge 0$	$\lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$
Gamma	$\frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$	$x \ge 0$	$\alpha > 0$, $\lambda > 0$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-t}\right)^\alpha$
χ_n^2	$\frac{x^{\frac{(n-2)}{2}} e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}$	$x \ge 0$	n 正整數	n	$2n$	$\left(\frac{1}{1-2t}\right)^{\frac{n}{2}}$
t_n	$\frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} (1+\frac{x^2}{n})^{-\frac{n+1}{2}}$	$-\infty \le x \le \infty$	n 正整數	0	$\frac{n}{n-2}$	不存在
$F_{n,m}$	$\frac{\Gamma(\frac{n+m}{2}) n^{\frac{n}{2}} m^{\frac{m}{2}} x^{\frac{n-2}{2}}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})(m+nx)^{\frac{n+m}{2}}}$	$x \ge 0$	n, m 正整數	$\frac{m}{m-2}$	§ 6.2.7	不存在
Beta	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$0 \le x \le 1$	$\alpha > 0$, $\beta > 0$	$\frac{\alpha}{\alpha+\beta}$	§ 6.2.9	$\mu_r' = \frac{\Gamma(r+\alpha+\beta)\Gamma(\alpha)}{\Gamma(r+\alpha)\Gamma(\alpha+\beta)}$
Weibull	$\alpha\beta^{-\alpha} x^{\alpha-1} e^{-(\frac{x}{\beta})^\alpha}$	$x \ge 0$	$\alpha > 0$, $\beta > 0$	§ 6.2.11	§ 6.2.11	$\mu_r' = \beta^{-r}\Gamma(1+\frac{r}{\alpha})$
Laplace	$\frac{1}{2\beta} e^{-\frac{\lvert x-\alpha\rvert}{\beta}}$	$-\infty \le x \le \infty$	$\beta > 0$, $\alpha \in R$	α	$2\beta^2$	$\frac{e^{\alpha t}}{1-(\beta t)^2}$
Cauchy	$\frac{1}{\pi\beta[1+(\frac{x-\alpha}{\beta})^2]}$	$-\infty \le x \le \infty$	$\beta > 0$, $\alpha \in R$	不存在	不存在	不存在
Logistic	$\frac{e^{-\frac{x-\alpha}{\beta}}}{\beta[1+e^{-\frac{x-\alpha}{\beta}}]^2}$	$-\infty \le x \le \infty$	$\beta > 0$, $\alpha \in R$	α	$\frac{\beta^2\pi^2}{3}$	$e^{\alpha t}\pi\beta t \csc(\pi\beta t)$

三、因果与关系

我们在第 1 章提到，本书第 9 章到第 12 章，还有第 13 章两个以上总体的检验，都有因果或相关(有关系)。通常虚无假设是：没有差异、没有显着影响、或没有相关。而拒绝虚无假设表示：显着、有差异存在、有影响、有关系。如果检验结果是：不显着、没有关系，就还要再去找新的关系，所以说：没关系，就(要找)有关系。

在统计学：相关可以检验，因果是无法检验。

卡方独立性检验没有假设因果，只有相关检验，但是没有相关系数或正负相关，在 2×2 列联表有胜算率(odds ratio)表示相关的程度。变异数分析检验：分类变动是否影响反应变量的平均数。分开卡方检验 2×3×4 表格，是将列联表根据另一个分类变量拆开。



以下是一些统计的因果或相关的例子：

1. 抽烟会导致肺癌

2. 每个国家(城市)枪枝管制和枪击案的数目有相关

3. 每年冰淇淋的销售量和每年犯罪人数有相关 (r = 0.5)

4. 每年离婚率和每年出国人数有相关 (r = 0.9225)

5. 每个人的收入和他的血压有相关 (r = 0.667)

6. 每个国家的人均电视机(手机或 PC)数目和人均寿命有相关

7. 企业的服务质量和顾客忠诚度有相关

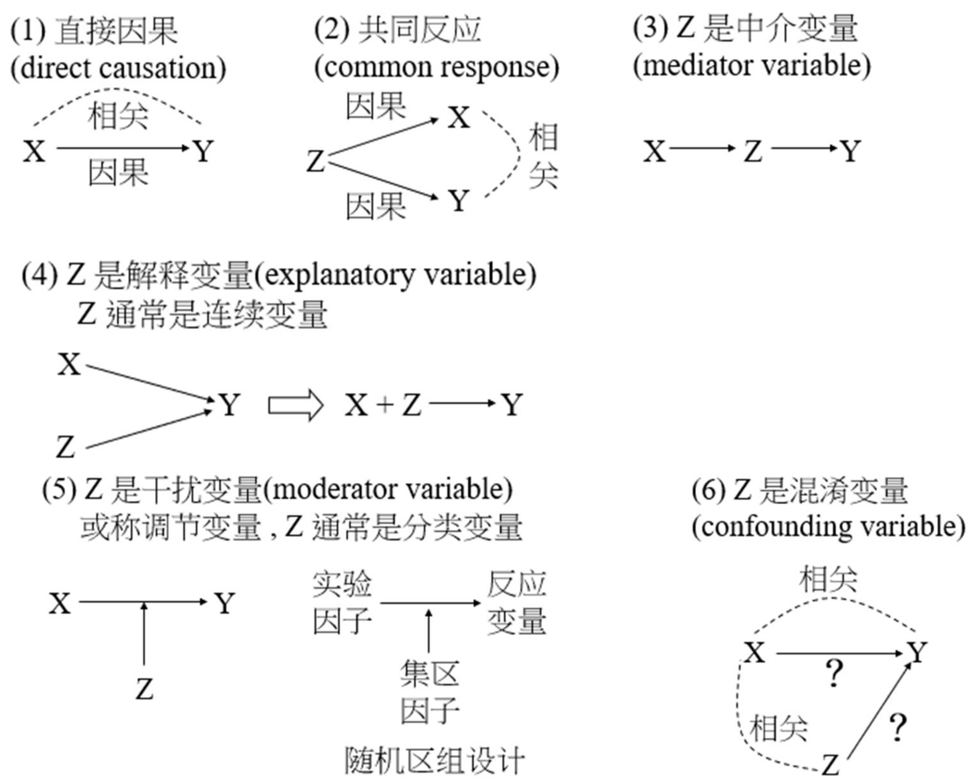
8. 在 60 年代美国，玩飞盘(Frisbee)的人和得到性病有相关

9. 买啤酒的人和买尿片有相关

10. 军人或平民的体力有差异

11. 母女关系(遗传因素)和体重(BMI)有相关 (r = 0.506)

以上例子其实多数有另一个变量，影响其因果或相关，有下列 6 种情况：



混淆变数是：两个变量对反应变量的影响是混在一起，无法区分，因此不能判定因果的相关性。例如下例：遗传、环境和体重的关系。

简单回归分析是直接因果，复回归分析是加入解释变量，分开卡方检验和随机集区设计是加入干扰变数。以下是一些统计的因果或相关的例子：

1. 抽烟 X 与肺癌 Y：直接因果

2. 枪枝管制 X 和枪击案 Y：直接因果

3. 冰淇淋 X 与犯罪人数 Y：共同反应，天气变数 Z，天气热易犯罪

4. 离婚率 X 和出国人数 Y：共同反应，经济成长 Z，可共患难，不能共富贵

5. 收入 X 和血压 Y：共同反应，年纪 Z 大，收入和血压高；或住城市 Z，收入和血压高

6. 电视机数目 X 和寿命 Y：共同反应，国民所得 Z

7. 服务质量 X 和顾客忠诚度 Y：中介变数，顾客免满意度 Z

8. 玩飞盘 X 和得到性病 Y：干扰变数，早期玩飞盘多为抽大麻年青人 Z

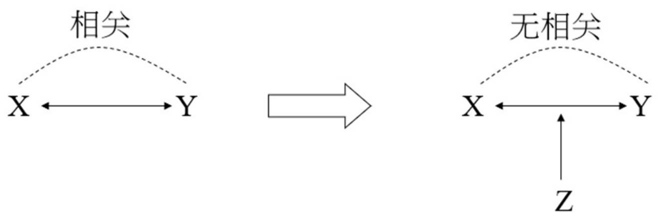
9. 买啤酒 X 和买尿片 Y：干扰变数，家有婴儿 Z

10. 军人或平民 X 和体力 Y：干扰变数，性别 Z、或年龄 Z，军人多为青年男性

11. 遗传 X 和体重 Y：混淆变量，环境因素 Z，母女有相同的生活环境和喜好(如吃住)

四、辛普森悖论(Simpson's paradox)

一般说来，加入解释变量或干扰变量 Z，会使原来「没关系」的 X,Y 变量，变成「有关系」或「判定系数更大」，例如：加入集区的随机集区设计或加入新解释变量的复回归分析。但是，辛普森矛盾(辛普森诡论)是对于上述结果是相反的：加入干扰变量 Z，使原来「有关系」的 X,Y 变量，变成「无关系」或「负关系」。辛普森矛盾的例子，都是利用列联表的卡方检验，以下例题修改自墨尔(2003)。



例题：某大学只有两个学院：文学院和工学院。下表是今年全校招生录取人数：

大学	男性	女性
录取	310	170
不录取	330	470
总和	640	640

利用卡方列联表独立检验：卡方检验值 $= 65.33$，临界值 $= 3.84$， p 值 $= 0$

结论：录取率和性别是有「显着相关」。

2×2 列联表胜算率(odds ratio, OR)的计算:

$$OR = \frac{a \times d}{b \times c}$$

		变量 Y	
		甲	乙
变量 X	阳性	a	c
	阴性	b	d

若 OR = 1，则变量 X 与 Y 独立无相关。

若 OR > 1，则「甲且阳性」发生的可能性较大。

若 OR < 1，则「甲且阳性」发生的可能性较小。

OR 的越大或越接近 0 ，表示变量 X 与 Y (正或负)相关性越强。

该大学录取表的胜算率 OR = (310×470) / (330×170) = 2.597

所以，男性的录取率显着较高。

因此，性别平等委员会抗议：录取率偏向男性。

但是，大学教务处提出文学院和工学院的录取资料如下：

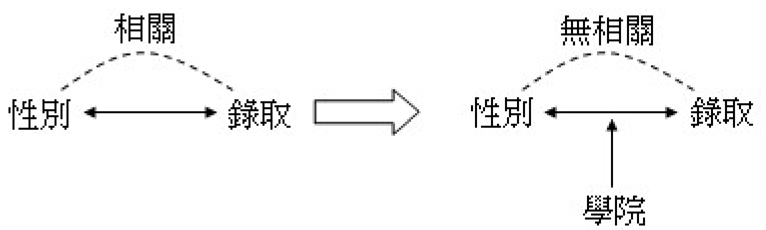
工学院	男性	女性
录取	300	20
不录取	300	20
总和	600	40

文学院	男性	女性
录取	10	150
不录取	30	450
总和	40	600

两个列联表的独立卡方检验：卡方检验值 = 0，临界值 = 3.84， p 值 = 1

OR = 1

结论：两个学院的录取率和性别是「没有相关」。



问题：计算下列列联表的卡方独立性检验、胜算率、及男女性的录取率。

大学	男性	女性
录取	50	25
不录取	70	95
总和	120	120

工学院	男性	女性
录取	49	15
不录取	51	5
总和	100	20

文学院	男性	女性
录取	1	10
不录取	19	90
总和	20	100