Lecture 8 – Contingency Analysis 独立性检验

- Outline for today
  - Recall: testing on proportion & frequency data
  - Contingency Analysis 独立性检验
    - Estimate the association between two variables (关联的程度)
    - Test the association between two variables (关联的显著性)
  - Summary
  - R Lab & Discussion

生物统计学

李 勤

生态与环境科学学院

学习目标

- Estimate the association between two variables (关联的程度)
  - Relative Risk
  - Odds Ratio

- Test the association between two variables (关联的显著性)
  - Chi-square test with contingency table
  - Fisher’s exact test

1. Recall from L07 - Proportion & Frequency

- 有关频数和比例数据的假设检验

一维频数数据

- 两个类别的频数（二项分布 $\Pr[X\ success] = \binom{n}{X} p^X (1-p)^{n-X}$） → 二项检验
- 两个或两个以上类别的频数 → 卡方检验
- 单位时间/空间内的频数（泊松分布 $\Pr[X\ success] = \frac{e^{-\mu}\mu^{-X}}{X!}$） → 卡方检验

$$\chi^2 = \sum_i \frac{(Oi - Ei)^2}{E_i}$$

二维频数数据

- 两个分类变量间的独立性检验

1. Recall from L07 - Proportion & Frequency

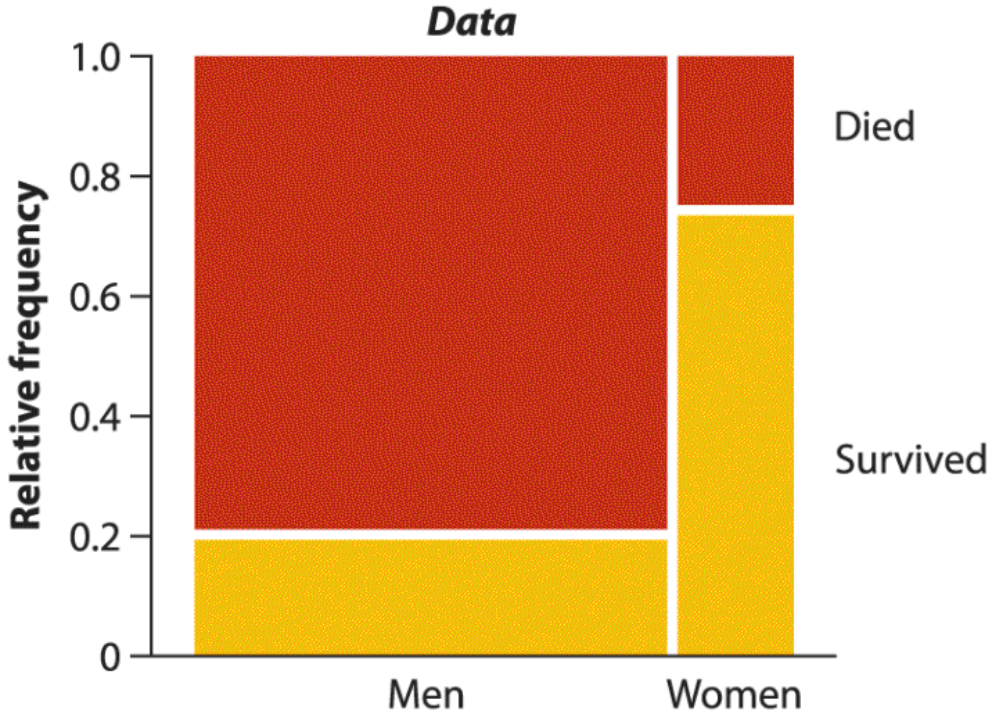
- 二维频数数据的例子:
  - 鲜艳和暗淡的蝴蝶被吃掉的概率不同吗?
  - 吸烟者喝酒的概率比不吸烟者高多少?
  - 每天服用阿司匹林的人心脏病发作的概率更低吗?

2. Contingency Analysis 独立性检验

- 独立性检验可以检验两个（及以上）分类变量间的**关联性**
  (the association between two or more categorical variables):
  - If one variable is “contingent” on the other 是否有关联/依赖性?
    - contingent: 取决于，依情况而言;
  - To what extent one variable is “contingent” on the other 关联性/依赖性有多强?
- 独立性检验的核心是研究变量的独立性 (independence)。

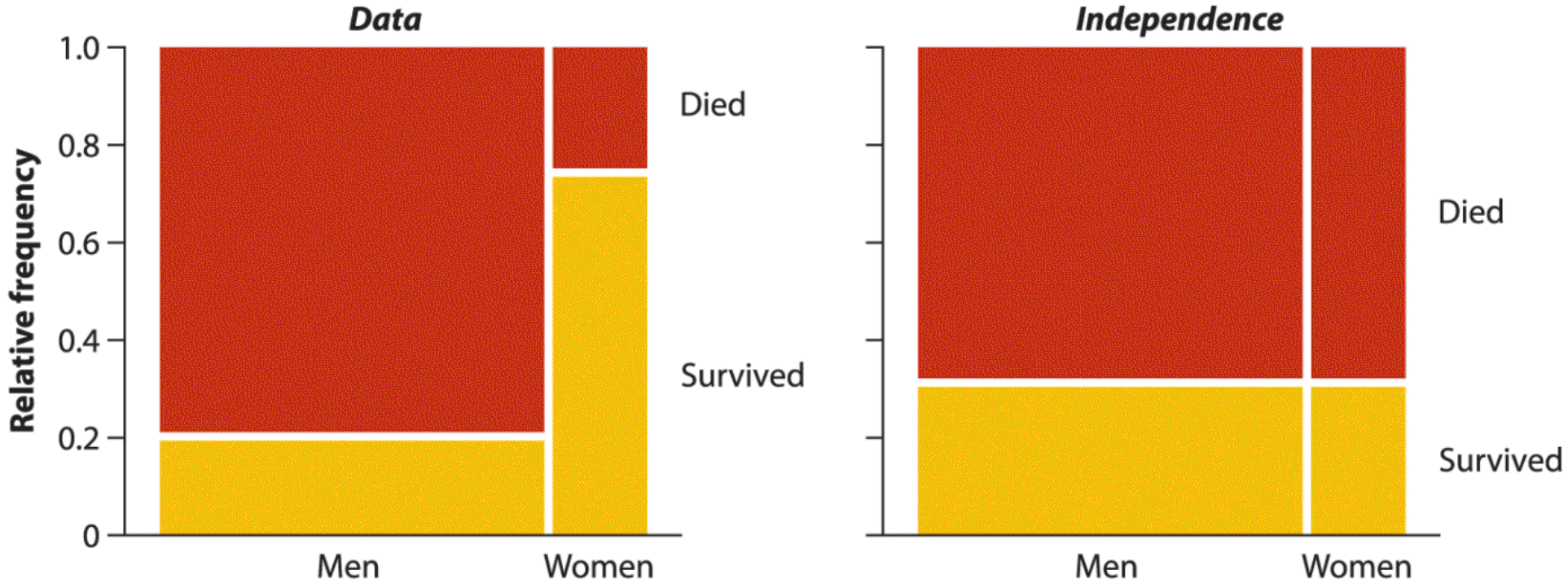
2.1 两个分类变量间的关联性

- 泰坦尼克号沉没事故（1912.4.14）
  - 幸存 vs. 死亡；男性 vs. 女性



2.1 两个分类变量间的关联性

- 泰坦尼克号沉没事故（1912.4.14）
  - 幸存 vs. 死亡；男性 vs. 女性



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.1 两个分类变量间的关联性

- 泰坦尼克号沉没事故（1912.4.14）
  - 幸存 vs. 死亡；男性 vs. 女性

- 哪两个变量？解释变量和响应变量分别是？
  - 鲜艳和暗淡的蝴蝶被吃掉的概率不同吗？
  - 吸烟者喝酒的概率比不吸烟者高多少？
  - 每天服用阿司匹林的人心脏病发作的概率更低吗？

2. Contingency Analysis 独立性检验

- 独立性检验可以检验两个（及以上）分类变量间的**关联性**
  (the association between two or more categorical variables):
  - If one variable is “contingent” on the other 是否有关联/依赖性?
    - contingent: 取决于，依情况而言;
  - To what extent one variable is “contingent” on the other 关联性/依赖性有多强?
- 独立性检验的核心是研究变量的独立性 (independence)。

2.2 估计2 × 2列联表中的关联性：相对风险

- Relative Risk (RR): 相对风险
  - 风险 (risk)：出现负面 (undesired) 结果的概率；
  - 相对风险是指处理组出现负面 (undesired) 结果的概率除以对照组出现相同结果的概率。
  - $RR = p_1/p_2$
- Association: 关联性、相关性

处理/解释变量

结果/响应变量	Treatment group (1)	Control group (2)
undesired outcome	p_1	p_2
(desired) outcome	$1 - p_1$	$1 - p_2$

2.2 估计2 × 2列联表中的关联性：相对风险

- Relative Risk (RR): 相对风险
  - 相对风险是指处理组出现负面 (undesired) 结果的概率除以对照组出现相同结果的概率。
  - $RR = p_1/p_2$
  - 取值范围：$0 < RR < +\infty$

- Reduction of relative risk: $1 - RR$
  - 医学中广泛用此表示疫苗效力。

2.2 估计2 × 2列联表中的关联性：相对风险

- Relative Risk (RR): 相对风险
  - 处理组出现负面结果的概率除以对照组出现相同结果的概率。
- Estimate 估计方法
  - $\hat{p}_1 = a/(a+c)$
  - $\hat{p}_2 = b/(b+d)$
  - $\widehat{RR} = \hat{p}_1/\hat{p}_2$

处理

结果	Treatment group (1)	Control group (2)
undesired outcome	*a*	*b*
(desired) outcome	*c*	*d*

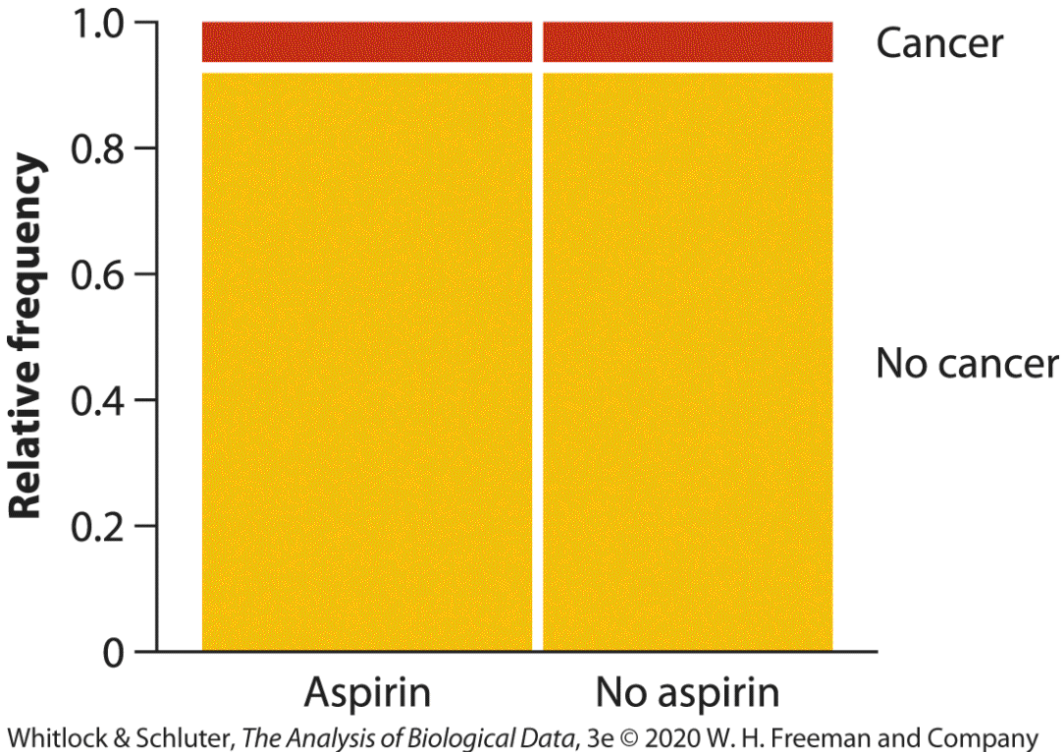
2.2 估计2 × 2列联表中的关联性：相对风险

- Relative Risk (RR): 相对风险
  - $\hat{p}_1 = \frac{a}{(a+c)} = \frac{1438}{1438+18496} = 0.0721$
  - $\hat{p}_2 = \frac{b}{(b+d)} = \frac{1427}{1427+18515} = 0.0716$
  - $\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{0.0721}{0.0716} = 1.007$

结果 \ 处理	Aspirin 阿司匹林	Placebo 安慰剂
Cancer 患癌症	*a* = 1438	*b* = 1427
No cancer 未患癌症	*c* = 18,496	*d* = 18,515

2.2 估计2 × 2列联表中的关联性：相对风险

- Relative Risk (RR): 相对风险
  - $\hat{p}_1 = \frac{a}{(a+c)} = \frac{1438}{1438+18496} = 0.0721$
  - $\hat{p}_2 = \frac{b}{(b+d)} = \frac{1427}{1427+18515} = 0.0716$
  - $\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{0.0721}{0.0716} = 1.007$
  - 这两组患癌的概率估计值非常接近，因此相对风险接近于 1。



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 估计2 × 2列联表中的关联性：比值比

- Odds Ratio (OR): 比值比、优势比
- Odds: 胜算、几率
  - 成功的几率是成功的概率除以失败的概率。
  - $O = \frac{p}{1-p}$
  - $O = 1$ (或1: 1)

处理

结果	Treatment group (1)	Control group (2)
"Success" outcome	p_1	p_2
"Failure" outcome	$1 - p_1$	$1 - p_2$

2.3 估计2 × 2列联表中的关联性：比值比

- Odds Ratio (OR): 比值比、优势比
  - 成功的几率是成功的概率除以失败的概率。
  - $OR = O_1/O_2$
  - $O_1 = p_1/(1-p_1)$
  - $O_2 = p_2/(1-p_2)$

处理

结果	Treatment group (1)	Control group (2)
"Success" outcome	p_1	p_2
"Failure" outcome	$1-p_1$	$1-p_2$

2.3 估计2 × 2列联表中的关联性：比值比

- Odds Ratio (OR): 比值比、优势比
  - $\widehat{O_1} = \frac{\widehat{p_1}}{1-\widehat{p_1}} = \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}$
  - $\widehat{O_2} = \frac{\widehat{p_2}}{1-\widehat{p_2}} = \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$
  - $\widehat{OR} = \frac{\widehat{O_1}}{\widehat{O_2}} = \frac{a/c}{b/d} = \frac{ad}{bc}$

结果 \ 处理	Treatment group (1)	Control group (2)
"Success" outcome	*a*	*b*
"Failure" outcome	*c*	*d*

2.3 估计2 × 2列联表中的关联性：比值比

- Odds Ratio (OR): 比值比、优势比
  - $\widehat{OR} = \frac{\widehat{O_1}}{\widehat{O_2}} = \frac{a/c}{b/d} = \frac{ad}{bc}$
  - $\widehat{OR} = \frac{1438 \times 18515}{1427 \times 18496} = 1.009$

处理

结果	Aspirin 阿司匹林	Placebo 安慰剂
Cancer 患癌症	*a* = 1438	*b* = 1427
No cancer 未患癌症	*c* = 18,496	*d* = 18,515

2.4 估计2 × 2列联表中的关联性

- Relative Risk (RR): 相对风险
  - $\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{a/(a+c)}{b/(b+d)}$
- Odds Ratio (OR): 比值比、优势比
  - $\widehat{OR} = \frac{\widehat{O_1}}{\widehat{O_2}} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{a/c}{b/d} = \frac{ad}{bc}$

结果 \ 处理	Treatment group (1)	Control group (2)
undesired/ success	a (p_1)	b (p_2)
desired/ failure	c $(1-p_1)$	d $(1-p_2)$

2.4 估计2 × 2列联表中的关联性

- Relative Risk (RR): 相对风险（两个比例的比）
  - 估计值: $\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{a/(a+c)}{b/(b+d)}$
  - 标准误: $\mathrm{SE}\left[\ln\left(\widehat{RR}\right)\right] = \sqrt{\frac{1}{a} + \frac{1}{b} - \frac{1}{a+c} - \frac{1}{b+d}}$
  - 95%置信区间:
    - $\ln\left(\widehat{RR}\right) - 1.96 \times \mathrm{SE}\left[\ln\left(\widehat{RR}\right)\right] < \ln(RR) < \ln\left(\widehat{RR}\right) + 1.96 \times \mathrm{SE}\left[\ln\left(\widehat{RR}\right)\right]$
- Odds Ratio (OR): 比值比、优势比（两个几率的比）
  - 估计值: $\widehat{OR} = \frac{\widehat{O_1}}{\widehat{O_2}} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{ad}{bc}$
  - 标准误: $\mathrm{SE}\left[\ln\left(\widehat{OR}\right)\right] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
  - 95%置信区间:
    - $\ln\left(\widehat{OR}\right) - 1.96 \times \mathrm{SE}\left[\ln\left(\widehat{OR}\right)\right] < \ln(OR) < \ln\left(\widehat{OR}\right) + 1.96 \times \mathrm{SE}\left[\ln\left(\widehat{OR}\right)\right]$

2.4 估计2 × 2列联表中的关联性

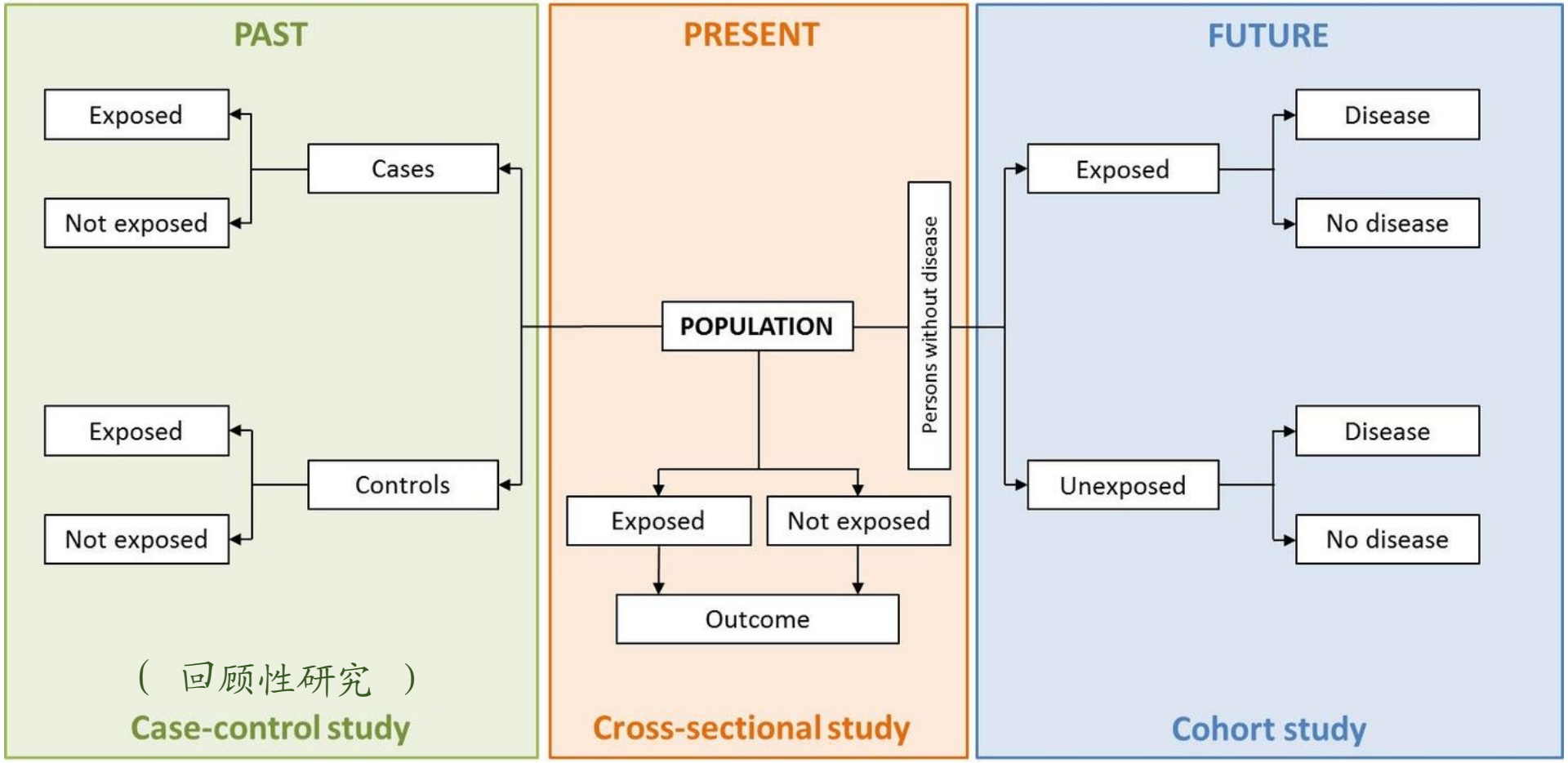
- 阿司匹林与癌症的相关性
  - 相对风险: $\widehat{RR} = 1.007; \quad 0.94 < RR < 1.08$
  - 比值比: $\widehat{OR} = 1.009; \quad 0.93 < OR < 1.09$
- 取值范围: $0 < OR/RR < +\infty$
  - $= 1$ → 表示风险或胜算与处理无关。
    - the risk (prob. of getting an undesired outcome) is not affected by the treatment.
    - the odds of success in the response variable are independent of treatment.
  - > 1 → 表示第一组（处理组）的风险或胜算更高（正相关）。
    - the risk is higher in the treatment group than in the control group.
    - the event has higher odds in the first group (treatment) than in the second group (control).
  - < 1 → 表示第一组（处理组）的风险或胜算更低（负相关）。
    - the risk is higher in the treatment group than in the control group.
    - the event has higher odds in the first group (treatment) than in the second group (control).

2.4 估计2 × 2列联表中的关联性

- 相对风险RR和比值比OR都是对关联性程度的估计。
- RR更具直觉性（两个比例的比）
  - 不同处理需要获得随机样本。
- OR 可被用于“病例对照研究”(case-control studies)
  - 病例对照研究是一种（回顾型）观察研究方法，将患有某种疾病的样本（或具有某种后果，“病例”）与不患有这种疾病但其他特征相似的样本（“对照”）进行比较;
  - 患有和未患有疾病的人数不一定与该疾病在人群中的发病率成正比;
  - 因此，我们无法估计风险；但可以用OR来探讨疾病（响应变量）和另一个变量（解释变量，如分析暴露风险因素）的相关性。

2.4 估计2 × 2列联表中的关联性

- OR 可被用于“病例对照研究”(case-control studies)



2.4 估计2 × 2列联表中的关联性

- OR 可被用于“病例对照研究”
  - OR最大的优势的是不受实验组和对照组比例（incidence rate）的影响;
  - 因为病例和对照组的总体比例会在比率中抵消； 理论上讲，
    - 不管实验组样本为多少例，a/c 是不变的;
    - 不管对照组样本量如何变化，c/d 的比例也是固定的;
  - $\widehat{O_1} = \frac{\widehat{p_1}}{1-\widehat{p_1}} = \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}$
  - $\widehat{O_2} = \frac{\widehat{p_2}}{1-\widehat{p_2}} = \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$
  - $\widehat{OR} = \frac{\widehat{O_1}}{\widehat{O_2}} = \frac{a/c}{b/d} = \frac{ad}{bc}$

处理

结果

	Treatment group (1)	Control group (2)
“Success” outcome	*a*	*b*
“Failure” outcome	*c*	*d*

2.4 估计2 × 2列联表中的关联性

- 相对风险RR和比值比OR都是对关联性程度的估计。
- OR 可被用于“病例对照研究”(case-control studies)
  - E.g., 弓形虫病可能与某些精神疾病有关，也可能与危险行为有关;
  - 研究目的：了解弓形虫感染是否会导致事故发生概率的变化。
    - 根据司机的情况来收集观察数据；而非实操控制实验;
    - 假定了发生事故司机的比例为0.5;
    - 解释变量是？响应变量是？
  - $\widehat{OR} = \frac{ad}{bc} = \frac{61\times169}{124\times16} = 5.20$

“处理” control

结果 case	弓形虫感染	无弓形虫感染	
发生事故的司机	a = 61	b = 124	(185)
未发生事故的司机	c = 16	d = 169	(185)

2.4 估计2 × 2列联表中的关联性

- 相对风险RR和比值比OR都是对关联性程度的估计。
- RR更具直觉性（两个比例的比）
- OR 可被用于“病例对照研究”(case-control studies)
- 当a和b足够小时（例如罕见病）
  - $a/(a+c) \approx a/c$
  - …?
  - $RR \approx OR$

处理

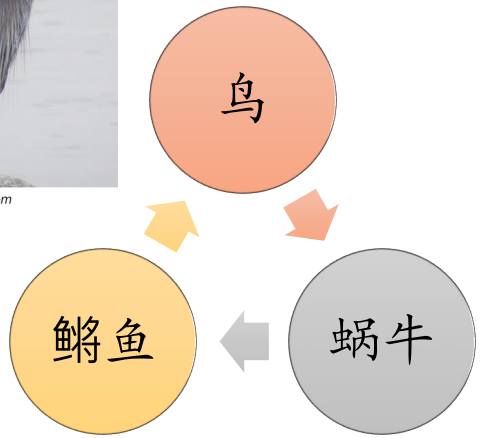
结果	Treatment group (1)	Control group (2)
undesired/ success	a (p_1)	b (p_2)
desired/ failure	c $(1-p_1)$	d $(1-p_2)$

3. Contingency Analysis 独立性检验

- 独立性检验可以检验两个（及以上）分类变量间的**关联性** (the association between two or more categorical variables):
  - To what extent one variable is “contingent” on the other 关联性/依赖性有多强？ → 看置信区间是否包含1
  - If one variable is “contingent” on the other 是否有关联/依赖性?
    - 显著性 → 多维分类变量的卡方检验
- 独立性检验的核心是研究变量的独立性 (independence)。

3.1 The χ^2 contingency test — $R \times C$列联表

- $R \times C$ contingency table
  - 一种吸虫的三个寄主
    - 其中受感染的鱼在水面附近停留的时间过长，可能更容易被鸟类捕食。



		鱼的处理（室外水箱）			总计
		Uninfected	Lightly infected	Highly infected	Total
结果	Eaten by birds	1	10	37	48
	Not eaten by birds	49	35	9	93
总计	Total	50	45	46	141

3.1 The χ^2 contingency test — $R \times C$列联表

- The χ^2 contingency test —— 基于列联表的卡方检验
  - 一种吸虫的三个寄主
    - 其中受感染的鱼在水面附近停留的时间过长，可能更容易被鸟类捕食。
- 假设检验
  - 1. 零假设和备择假设
    - H_0: 鱼被寄生虫感染与被鸟捕食之间是相互独立的。
    - H_A: 鱼被寄生虫感染与被鸟捕食之间不是相互独立的。
  - 2. 计算 χ^2 检验统计量

$$\chi^2 = \sum_i \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

3.1 The χ^2 contingency test — $R \times C$列联表

- 假设检验
  - 1. 零假设和备择假设
    - H_0: 鱼被寄生虫感染与被鸟捕食之间是相互独立的。
    - H_A: 鱼被寄生虫感染与被鸟捕食之间不是相互独立的。
  - 2. 计算χ^2检验统计量
    - H_0下的期望频数怎么计算?

		鱼的处理（室外水箱）			总计
		Uninfected	Lightly infected	Highly infected	Total
结果	Eaten by birds	?	?	?	48
	Not eaten by birds	?	?	?	93
总计	Total	50	45	46	141

3.1 The χ^2 contingency test — $R \times C$列联表

- 期望频数：H_0下两组变量之间是独立 (Pr[A & B] = Pr[A] × Pr[B])
  - Pr[uninfected & eaten by birds] = Pr[uninfected] × Pr[eaten by birds]

- Pr[uninfected] = 50 / 141 = 0.3546
- Pr[eaten by birds] = 48 / 141 = 0.3404
- Pr[uninfected & eaten by birds] =
  0.3546 × 0.3404 = **0.1207**
  - **0.1207** × 141 = 17.01

		鱼的处理（室外水箱）			总计
		Uninfected	Lightly infected	Highly infected	Total
结果	Eaten by birds	?	?	?	48
	Not eaten by birds	?	?	?	93
总计	Total	50	45	46	141

3.1 The χ^2 contingency test — $R \times C$列联表

- 期望频数：H_0下两组变量之间是独立
  - Pr[uninfected & eaten by birds] = Pr[uninfected] × Pr[eaten by birds]
  - 期望频数 = $Pr_{[r,c]} \times 141$；或简化算法：(50/144) × (48/144) × 144 = 58 ×48 / 144

- Pr[uninfected] = 50 / 141 = 0.3546
- Pr[lightly infected] = ?
- Pr[highly infected] = ?
- Pr[eaten by birds] = 48 / 141 = 0.3404
- Pr[not eaten by birds] = ?

		鱼的处理（室外水箱）			总计
		Uninfected	Lightly infected	Highly infected	Total
结果	Eaten by birds	***17.0***	***15.3***	***15.7***	48
	Not eaten by birds	***33.0***	***29.7***	***30.3***	93
总计	Total	50	45	46	141

3.1 The χ^2 contingency test — $R \times C$列联表

- 期望频数：H_0下两组变量之间是独立
  - Pr[鱼的处理 & 鸟的捕食] = Pr[鱼的处理] × Pr[鸟的捕食]
  - 期望频数 = $Pr_{[r,c]}$ × Total Freq. = Row total × Column total × Total Freq.
  - 观察频数（observed） / 期望频数（expected）

		鱼的处理（室外水箱）			总计
		Uninfected	Lightly infected	Highly infected	Total
结果	Eaten by birds	**1 / *17.0***	**10 / *15.3***	**37 / *15.7***	48
	Not eaten by birds	**49 / *33.0***	**35 / *29.7***	**9 / *30.3***	93
总计	Total	50	45	46	141

3.1 The χ^2 contingency test — $R \times C$列联表

- 假设检验
  - 2. 计算χ^2检验统计量
    - $\chi^2 = \frac{(1-17)^2}{17} + \frac{(10-15.3)^2}{15.3} + \cdots + \frac{(9-30.3)^2}{30.3} = 69.5$
    - 自由度?
      - $df = (r-1)(c-1) = 2$

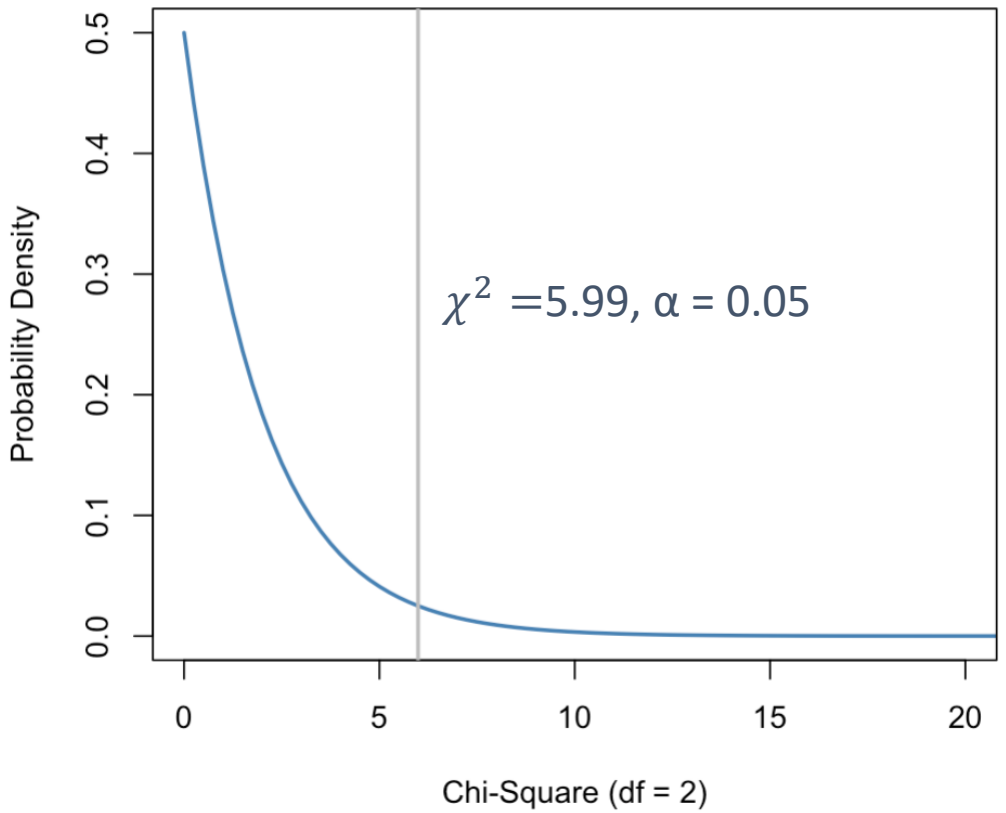
$$\chi^2 = \sum_{r,c} \frac{(Observed_{r,c} - Expected_{r,c})^2}{Expected_{r,c}}$$

		鱼的处理（室外水箱）			总计
		Uninfected	Lightly infected	Highly infected	Total
结果	Eaten by birds	**1 / *17.0***	**10 / *15.3***	**37 / *15.7***	48
	Not eaten by birds	**49 / *33.0***	**35 / *29.7***	**9 / *30.3***	93
总计	Total	50	45	46	141

3.1 The χ^2 contingency test — $R \times C$列联表

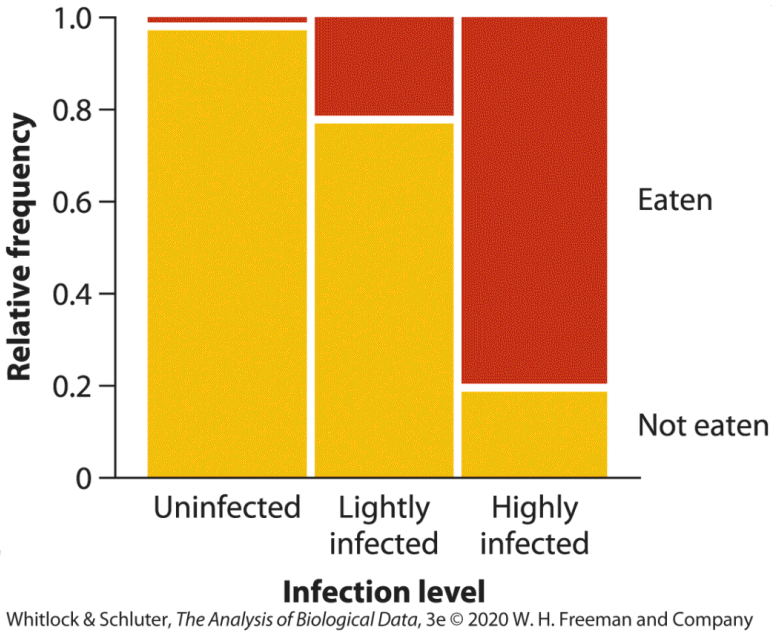
$$\chi^2 = \sum_{r,c} \frac{(Observed_{r,c} - Expected_{r,c})^2}{Expected_{r,c}}$$

- 假设检验
  - 2. 计算χ^2检验统计量
    - $\chi^2 = \frac{(1-17)^2}{17} + \frac{(10-15.3)^2}{15.3} + \cdots + \frac{(9-30.3)^2}{30.3} = 69.5$
    - 自由度
      - $df = (r-1)(c-1) = 2$
  - 3. 确定P值
    - P < 0.05（对比关键值）
    - P < 10^{-10}（软件直接计算）



3.1 The χ^2 contingency test — $R \times C$列联表

- 基于列联表的卡方检验 —— 吸虫对宿主的影响
  - 1. 零假设和备择假设
    - H_0: 鱼被寄生虫感染与被鸟捕食之间是相互独立的。
    - H_A: 鱼被寄生虫感染与被鸟捕食之间不是相互独立的。
  - 2. 计算χ^2检验统计量
    - $\chi^2 = \sum \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}} = 69.5$; 自由度为 $df = (r-1)(c-1) = 2$
  - 3. 确定P值： $P < 0.05$
  - 4. 结论
    - 统计学结论：拒绝H_0，说明两者之间不是相互独立的;
    - 生物学结论：（最可能的解释是）吸虫改变了鱼类的行为或阻碍了它们逃跑的能力，增加了它们被鸟类吃掉的概率，从而完成了生活史的最后一次寄主的转移。

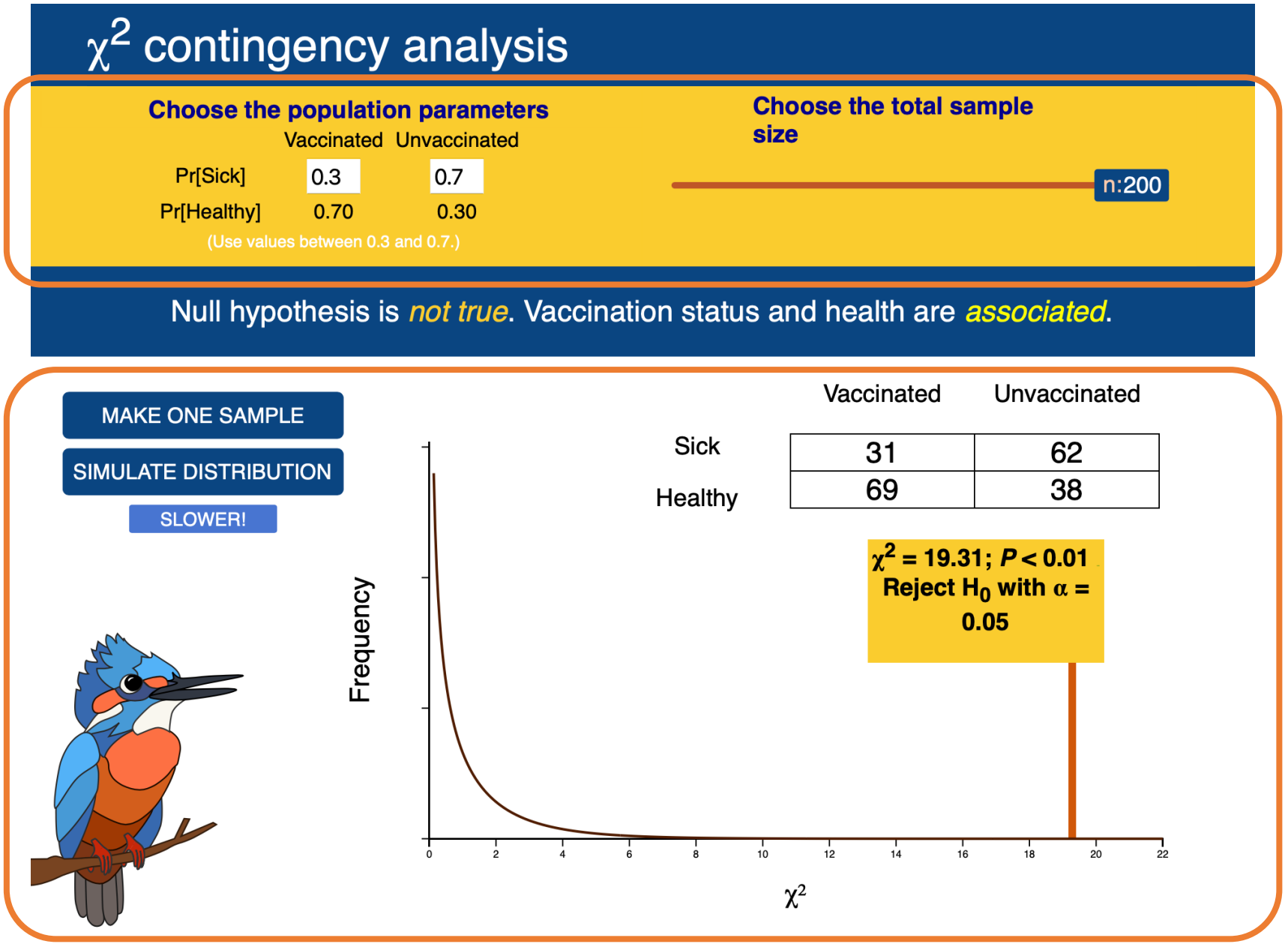


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

- 网页工具 https://www.zoology.ubc.ca/~whitlock/Kingfisher/ContingencyAnalysis.htm

设定

结果



3.2 The χ^2 contingency test — $R \times C$列联表的前提条件

- 卡方检验的特例： the χ^2 test with a $R \times C$ contingency table
- 前提条件 assumptions
  - 随机样本
  - 期望频数≥ 1
  - 期望频数<5的类别不超过20%

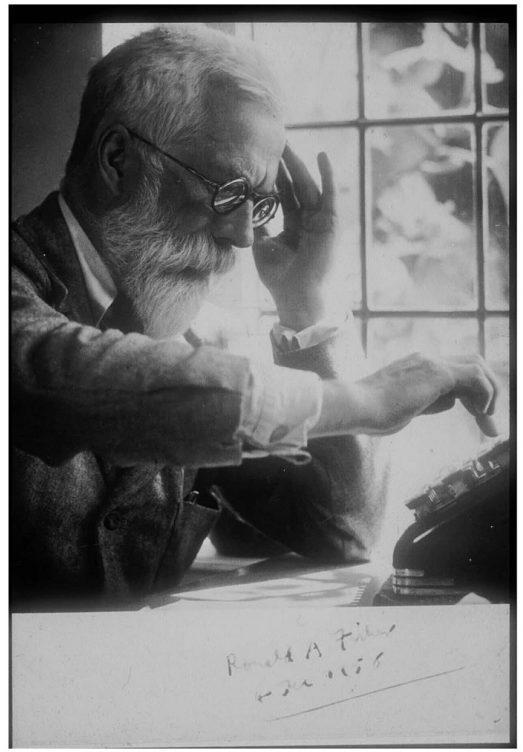
		处理			总计
		Group 1	Group 2	Group 3	Total
结果	Outcome 1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$n_{r=1}$
	Outcome 2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	$n_{r=2}$
总计	Total	$n_{c=1}$	$n_{c=2}$	$n_{c=3}$	$n_{grand\ total}$

3.2 The χ^2 contingency test — $R \times C$列联表的前提条件

- 卡方检验的特例： the χ^2 test with a $R \times C$ contingency table
- 前提条件 assumptions
  - 随机样本
  - 期望频数≥ 1
  - 期望频数<5的类别不超过20%
- 如若不满足前提条件
  - 1. 合并某些类别（注意自由度的改变）
  - 2. Fisher’s exact test
  - 3. Permutation test (Ch13)

4. Contingency Analysis - Fisher's exact test

- 独立性检验可以检验两个（及以上）分类变量间的**关联性**
  - 独立性检验的核心是研究变量的独立性
  - 关联性有多强？→ 检验RR或OR的置信区间是否包含1
  - 是否有关联性?
    - $R \times C$列联表的卡方检验 (the χ^2 contingency test)
    - **Fisher's exact test** (a 2 × 2 contingency table)
      - （Fisher精确检验）
      - 当不满足卡方检验前提时
      - 例如，当预期频数很小时



R.A. Fisher at his calculator in 1958 (courtesy of the Fisher Memorial Trust).

4.1 Fisher's exact test — 2 × 2 列联表

- 例如：哥斯达黎加的吸血蝙蝠以家畜的血液为食
  - 与公牛相比，蝙蝠更偏好母牛；
  - 表明蝙蝠可能对荷尔蒙信号做出了反应。

		处理		总计
		Cows in estrus	Cows not in estrus	Total
结果	Bitten by vampire bat	15	6	21
	Not bitten by vampire bat	7	322	329
总计	Total	22	328	350

观察频数

		处理		总计
		Cows in estrus	Cows not in estrus	Total
结果	Bitten by vampire bat	1.3	19.7	21
	Not bitten by vampire bat	20.7	308.3	329
总计	Total	22	328	350

期望频数（1/4=25%的类别频数<5）

4.1 Fisher's exact test — 2 × 2 列联表

- 1. 零假设和备择假设
  - H_0：发情状态和吸血蝙蝠攻击是独立的。
  - H_A：发情状态与吸血蝙蝠袭击无关。

		处理		总计
		Cows in estrus	Cows not in estrus	Total
结果	Bitten by vampire bat	15	6	21
	Not bitten by vampire bat	7	322	329
总计	Total	22	328	350

观察频数

- 2. 列出H_0成立下与观察结果一样或更极端的2 × 2 列联表
  - 行和列的总计都不变（H_0假设的独立性）
  - 改变其中一个格子的数值（更极端），并调整其它个字的值以匹配总计

16	5
6	323

17	4
5	324

18	3
4	325

19	2
3	326

20	1
2	327

21	0
1	328

分布一侧的更极端的列联表(one tail)

4.1 Fisher's exact test — 2 × 2 列联表

- 1. 零假设和备择假设
  - H_0：发情状态和吸血蝙蝠攻击是独立的。
  - H_A：发情状态与吸血蝙蝠袭击无关。
- 2. 列出H_0成立下与观察结果一样或更极端的2 × 2 列联表
  - 行和列的总计都不变（H_0假设的独立性）；
  - 改变其中一个格子的数值（更极端）， 并调整其它个字的值以匹配总计。
- 3. P 值

$$P = 2 \sum_{\substack{\text{all equally or more} \\ \text{extreme tables}}} \frac{R_1!R_2!C_1!C_2!}{a!b!c!d!n!}$$

  - 所有这些极端表格的概率之和，包括表格分布的另一侧同样或更多的极端表格。
  - 基于计算机统计软件，$P < 0.0001$
- 4. 结论
  - 拒绝独立性的零假设。吸血蝙蝠显然更偏好对发情期的母牛吸血（其原因尚不清楚）。

5. 小结 – 独立性检验 Contingency Analysis

- 两个分类变量间关联性的程度
  - 估计：RR或OR;
    - 检验其置信区间是否包含1;
  - 风险 (risk) 是发生不希望发生的事件的概率。
  - 相对风险 (RR) 是实验组的风险除以对照组的风险。
    - 如果相对风险小于1，则说明实验处理与降低风险有关。
  - 成功的几率 (odds) 是成功的概率除以失败的概率
    - 其中“成功”指感兴趣的结果。
  - 比值比 (OR) 是第一组（实验组）出现成功结果的几率除以第二组（对照组）出现该结果的几率。

5. 小结 – 独立性检验 Contingency Analysis

- 两个分类变量间关联性的显著性检验
  - $R \times C$列联表的 χ^2 检验 (the χ^2 contingent test)
    - H_0下 χ^2 统计量的抽样分布近似为 χ^2 分布
    - 自由度为 (r-1)(c-1)
  - 2 × 2 列联表的Fisher's exact test
    - 可计算精确 P 值（借住计算机软件）
    - 尤其适用于当不符合 χ^2 检验的前提条件

R commands summary

Contingency table

First categorical variable — Second categorical variable

```
sex_survive_table <- table(titanicData$sex, titanicData$survive)
```

Mosaic plot

```
mosaicplot(sex_survive_table)
```

Odds ratio

Estimate the odds ratio

```
fisher.test(sex_survive_table)$estimate
```

```
fisher.test(sex_survive_table)$conf.int
```

Give 95% confidence interval of odds ratio

χ^2 contingency analysis

Calculate the expected values

```
chisq.test(sex_survive_table)$expected
```

```
chisq.test(sex_survive_table, correct = FALSE)
```

Fisher's exact test

```
fisher.test(sex_survive_table)
```

Confidence interval for a proportion

Use the Agresti-Coull method

```
binom.confint(x = 30, n = 87, method = "ac")
```

Number of "successes" — Sample size

Binomial test

Proportion specified by null hypothesis

```
binom.test(x = 14, n = 18, p = 0.5)
```

Number of "successes" — Sample size

Frequency table

Name of frequency table — Categorical variable

```
MMtable <- table(MMlist$color)
```

χ^2 Goodness of fit test

Vector of expected proportions

```
chisq.test(MMtable, p = expected_proportions)
```

Poisson distribution

Number of successes — Mean number of successes

```
dpois(x = 3, lambda = 4.21)
```

P-value from χ^2

Observed χ^2 — Degrees of freedom

```
pchisq(q = 23.939, df = 6, lower.tail = FALSE)
```

6. 课堂练习Ch9 – Q1

- 1. 计算题： 相对风险RR。Wilson 等人（2011 年）对一组男性卫生专业人员进行了长达 20 年的跟踪调查。在所有参与研究的男性中，7890 人不喝咖啡，2492 人平均每天喝咖啡超过六杯。在 "不喝咖啡 "组中，有 122 人在研究期间患上了晚期前列腺癌，而在 "多喝咖啡 "组中，有 19 人患上了晚期前列腺癌。
- a. 依据这些数据创建一个列联表（按照惯例：解释变量在列，响应变量在行）。数据能说明什么关联？
- b. 高浓度咖啡组患晚期前列腺癌的概率估计值是多少？
- c. 不喝咖啡组患晚期前列腺癌的概率估计值是多少？
- d. 治疗组（高浓度咖啡）与对照组（不喝咖啡）相比，患晚期前列腺癌的相对风险是多少？

6. 课堂练习Ch9 – Q2

- 1. 计算题： 比值比OR。使用问题 1 中有关喝咖啡与前列腺癌的数据。
- a. 喝咖啡多的人患晚期前列腺癌的几率是多少？
- b. 不喝咖啡的人患晚期前列腺癌的几率是多少？
- c. 这两组人相比，患晚期前列腺癌的比值比是多少？
- d. 这两组人相比，比值比的对数值是多少？
- e. 在这种情况下，比值比的对数值的标准误差是多少？
- f. 比值比的对数值的 95% 置信区间是多少？ g. 比值比的95%置信区间呢?
- h. 解释比值比的置信区间。它是否与喝咖啡和罹患晚期前列腺癌是相互独立的这一可能性相一致？饮用咖啡是否倾向于增加或减少罹患晚期前列腺癌的概率？

6. 课堂练习Ch9

- Q3, Q6, Q8, Q10

- Next Week: Ch10 & Ch11
  - Normal distribution & t-test