

# 问题：骰子的公平性

## 1.0 Motivation example – Weldon's dice

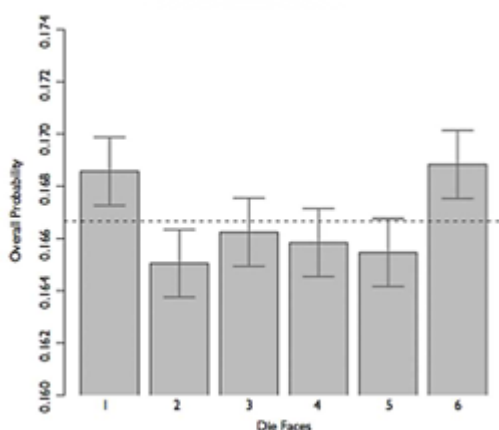
3

如何证明一个六面色子均匀公平（即每一面出现的概率相同）？

- 拉斐尔·威尔顿 (1860–1906), 英国著名生物学家, 统计学家, 在1894年投掷12个色子共计26,306次;
- 威尔顿记录下5或6发生的次数, 发现5或6**观测到的频次** > 预期的**期望频次**。



- 2009年, 芝加哥大学医学物理学博士生Zacariah Labby, 自制了一台机器, 并重复了威尔顿的试验。



↑柱状图

虚线: 预期次数 (越接近越公平)

## 拟合优度实验

检验**分类数据**的**观测频数**是否符合预先认定的分布

当提及卡方检验而没有特别指明类型时, 通常即指皮尔森卡方检验

### setting hypotheses

给定分布, 可以得到 **期望**的频数

观测频数O 期望频数E

- $H_0$ : There is no inconsistency between the observed and the expected counts. 观测符合期望
- $H_a$ : There is an inconsistency between the observed and the expected counts. 不符合 (出现点数的概率 不全等于  $1/6$ )

# Chi-square statistic

如何衡量O和E的差距

**Recall:** The general form of the test statistics we've seen this far is

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the **chi-square statistic**.

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}, \quad k = \text{类别的个数} \quad \text{“卡方值”}$$

Observed      Expected

$$\chi^2 = \sum_{j=1}^k \frac{(n_j - np_j)^2}{np_j}$$

参数性检验的检验统计量的范式

Null value: H0下的参数

非参数型检验:

根据具体问题

卡方检验也是一种非参数性检验（没有用到点估计的分布），而是要判断两组数据（OE）的匹配程度  
**注意分母不平方**

$p_i$ 是预期的概率（under H0）  $np_i$ : 期望观测数

## 计算

### 1.3 Chi-square test – back to the example

9

Outcome	Observed	Expected	$\frac{(O-E)^2}{E}$
1	53,222	52,612	$\frac{(53,222-52,612)^2}{52,612} = 7.07$
2	52,118	52,612	$\frac{(52,118-52,612)^2}{52,612} = 4.64$
3	52,465	52,612	$\frac{(52,465-52,612)^2}{52,612} = 0.41$
4	52,338	52,612	$\frac{(52,338-52,612)^2}{52,612} = 1.43$
5	52,244	52,612	$\frac{(52,244-52,612)^2}{52,612} = 2.57$
6	53,285	52,612	$\frac{(53,285-52,612)^2}{52,612} = 8.61$
Total	315,672	315,672	24.73

#### Why squaring the difference?

- Any standardized difference that is squared will now be **positive**.
- Differences that already looked unusual will become much larger after being squared.

#### Have we seen this before?

## 统计量的分布

- $\chi^2$  statistic follows  $\chi^2$  distribution
- $df=k-1$

## 1.4 Chi-square $\chi^2$ distribution

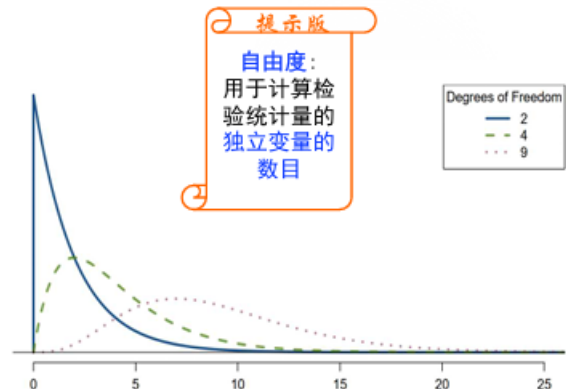
10

For goodness of fit test, the  $\chi^2$  statistic follows  $\chi^2$  distribution with the degrees of freedom is the number of categories - 1 ( $df = k - 1$ ).

The  $\chi^2$  distribution has just 1 parameter - **degrees of freedom (df)**, which influences the shape, center, and spread of the distribution.

As the df  $\uparrow$ :

- ✓ the center of the  $\chi^2$  distribution  $\uparrow$
- ✓ the variability of the  $\chi^2$  distribution  $\uparrow$



- $df \uparrow$  峰值右移  
O-E差异小  $\rightarrow$  统计量小  $\rightarrow$  统计量越大, O和E的差距越大  $\rightarrow$  单边检验

## p-value

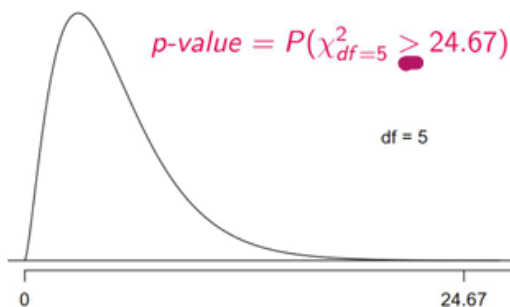
### 1.4 P-value

11

p-value for a chi-square test = tail area **above** the calculated test statistic  
(the test statistic is always positive, and a higher test statistic means a higher deviation from the  $H_0$  hypothesis. )

In our example, the chi-square statistic for this test is 24.67,  $df=5$ . What is p-value?

At 5% significance level, what is the conclusion of the hypothesis test?



In R: `pchisq(q = 24.67, df = 5, lower.tail = FALSE)`

P-value = 0.0001613338

Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df 1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52

p-value < 0.001



清华大学统计学研究中心

极端：偏离的方向（尾部）

注意 $df=5$ （一共六个独立变量）















## 总结

- 独立同分布eg抛骰子
- 样本量足够大  $n \geq 40$ , 否则无法收敛到卡方分布
- 期望频数  $np_i \geq 5$

## eg2

## 1.7 Example - Mendel's pea

14

Traits	Shape of seeds	Colour of seeds	Colour of pods	Shape of pods	Plant height	Position of flowers	Flower colour
Dominant trait	Round 	Yellow 	Green 	Full 	Tall 	At leaf junction 	Purple 
Recessive trait	Wrinkled 	Green 	Yellow 	Flat, constricted 	Short 	At tips of branches 	White 

Seven pairs of contrasting traits in pea plant

Type	Round and yellow	Angular and yellow	Round and green	Angular and green
Observed number	315	101	108	32
Mendel's Theoretical ratio	9	3	3	1

For any pea, let  $X = i$  denote it belongs to type  $A_i, i = 1, 2, 3, 4$

$$H_0: P(X = 1) = \frac{9}{16}, P(X = 2) = \frac{3}{16},$$









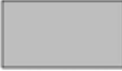
$$P(X = 3) = \frac{3}{16}, P(X = 4) = \frac{1}{16}$$

df=3

## 问题：两个变量是否独立

**Example.** In the dataset popular, students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them. A two-way table separating the students by grade and by choice of most important factor is shown below. Do these data provide evidence to suggest that **goals vary by grade?**

	Grades	Popular	Sports
4 <sup>th</sup>	63	31	25
5 <sup>th</sup>	88	55	33
6 <sup>th</sup>	96	55	32

	4 <sup>th</sup>	5 <sup>th</sup>	6 <sup>th</sup>
Grades			
Popular			
Sports			

↑棘状图

## 独立性检验

### hypotheses

- $H_0$ : 独立
- $H_a$ : 不独立

## 条件

重要

- 独立 (每一个条件内)
- 样本量 (每个格子 $\geq 5$ )

## 统计量

- The hypotheses are:

$H_0$ : Grade and goals are independent. Goals do not vary by grade.

$H_a$ : Grade and goals are dependent. Goals vary by grade.

- Conditions for the chi-square test of independence

Independence: Each case that contributes a count to the table must be independent of all the other cases in the table.

Sample size: Each particular scenario (i.e. cell) must have at least 5 *expected* counts.

- The test statistic

$$\chi^2_{df} = \sum_{i=1}^k \frac{(O - E)^2}{E} \quad \text{where} \quad df = (R - 1) \times (C - 1),$$

where  $k=R \times C$  is the number of cells,  $R$  is the number of rows, and  $C$  is the number of columns.

约束条件: 总和=1 (最后一个格子不算自由度)

如果有独立性,  $p_{ij}=p_i \times p_j$  ( $(R-1) \times (C-1)$ )

据此得expected count的计算方法:

**Expected Count = (row total)  $\times$  (column total)/table total**

	Grades	Popular	Sports	Total
4 <sup>th</sup>	63	31	25	119
5 <sup>th</sup>	88	55	33	176
6 <sup>th</sup>	96	55	32	183
Total	247	141	90	478

$$E_{\text{row 1, col 1}} = \frac{119 \times 247}{478} = 61$$

$$E_{\text{row 1, col 2}} = \frac{119 \times 141}{478} = 35$$

E=期望

code

```
> popular<-matrix(c(63,31,25,88,55,33,96,55,32),nrow=3)
> popular
      [,1] [,2] [,3]
[1,]   63   88   96
[2,]   31   55   55
[3,]   25   33   32
> chisq.test(popular, correct=FALSE)
```

Pearson's Chi-squared test

```
data: popular
X-squared = 1.3121, df = 4, p-value = 0.8593
```

Chi-square test works **well** when sample size is large enough, i.e., each (expected) cell count > 5.

When at least one cell in the contingency table has an expected frequency less than 5, **Yate's Continuity Correction** should be applied.

$$\chi_c^2 = \sum_{i=1}^k \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

In R, correct = TRUE (by default)



(注意前提条件)

如果有格子小于5→ **校正**

## summary

### Summary - Chi-square test

H0/ Ha:

检验统计量T:

抽样分布:

P-值和结论:

使用条件:

## categorical data analysis

### contingency tables列联表

- nominal无序型

- ordinal有序型

Tables representing all combinations of levels of **explanatory variable** 解释变量 and **response variable** 响应变量

		Response Variable				
		1	2	...	c	
Explanatory Variable	1	$n_{11}$	$n_{12}$	$\cdots$	$n_{1c}$	$n_{1.}$
	2	$n_{21}$	$n_{22}$	$\cdots$	$n_{2c}$	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	r	$n_{r1}$	$n_{r2}$	$\cdots$	$n_{rc}$	$n_{r.}$
		$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	

- 解释变量 取值 r
- 响应变量 呈现的不同值 c
- 落进格子的个数  $n_{ij}$ 
  - Numbers in table represent **counts** of the number of cases in each cell
  - Row and column **totals** are called **marginal counts**
  - Recall: categorical variables can be nominal or ordinal

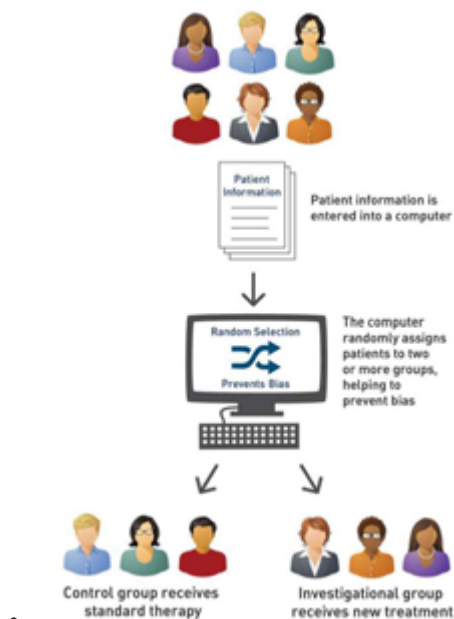
收入 \ 满意否	满意	不满意	合计
高	53	38	91
中	434	108	542
低	111	48	159
合计	598	194	792

清华大学统计学研究中心

- 边际 (总数)

## 常见研究类型

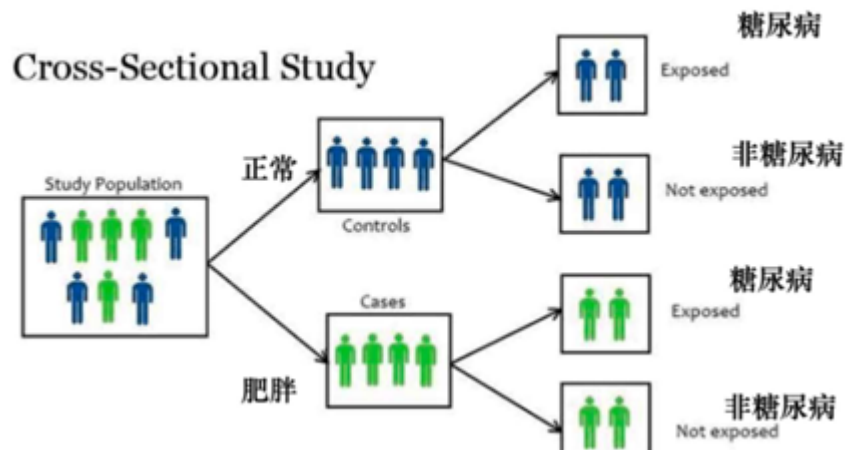
- 随机化临床试验
  - treatment/placebo (不一定是安慰剂, 也可能是已经确定效果的药物 control)
  - cure or not (结果称为响应变量)
  - 只能进行观察性研究, 没法预先分组 (跟踪一群人) eg 流行病 (不可能自行引发)



- epidemiological study/observational study①
  - prevalence survey 现况调查 (研究种类叫 cross-sectional study 横断面研究) 要研究人群, 直接把整个人群 (总数一定) 分组, 总会分到其中一组里, **每一个人服从多项分布**
  - 实验, 事先安排了很多组



- 响应变量：disease state
- 解释变量=研究问题：暴露因素risk factor



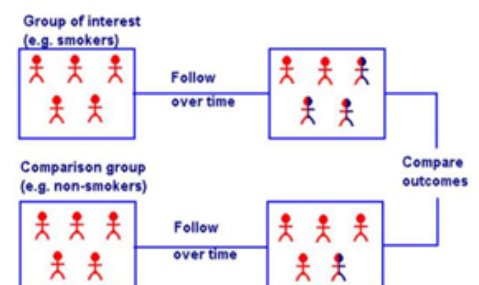
- epidemiological study/observational study②
  - cohort study队列研究 prospective cohort study前瞻性研究
  - 事先根据暴露因素（已知）分组，然后跟踪（和横断面研究的区别是队列研究会分别follow一段时间，再来看结果）eg不/吸烟的人一段时间后会不会得癌症
  - 在做研究的时候不能预先知道最后的结果是什么

**Epidemiological study/observational study:** disease state is observed (response variable), as well as exposure to risk factor (暴露因素, explanatory variable).

✓ Cohort study 队列研究

- 以暴露因素来分组，比如是否吃药，比如性别
- 前瞻性研究（prospective cohort study）

## Cohort Studies



- 每个人服从 二项分布or多项分布（看有几项）
- 有多项的时候是乘积，叫做product binominal/product multinominal
- epidemiological study/observational study③
  - case-control study结果导向，回过头寻找某些感兴趣的因素（暴露因素）是否存在，结果回溯研究 retrospective study回溯性研究



## 1.2 Common study types

7

- **Epidemiological study/observational study:** disease state is observed (response variable), as well as exposure to risk factor (暴露因素, explanatory variable).

- ✓ **Case-control study** 病例对照研究

- 以**结果（是否患有某病）**来分组，即**病例组-健康对照组**，计算有某因素特征（暴露）的率而进行比较的，是由结果寻原因
- **回顾性研究 (retrospective study)**



- 
- 从**时间**上来看暴露因素在前，病在后
- **注意：**case和control和临床试验中的意义不同，case指的是得病的组，control指的是没病的组；临床试验中，treatment和control相比，control还叫对照组，但意义不同（没有施加方法/施加非研究对象的方法）

随机化临床试验有很多考虑，比如每个人的饮食 学习时间 运动时间都要一样；不能一组全是女生，不然没法知道肥胖是性别相关还是巧克力相关（混杂因素），需要完全打乱；不能一组吃巧克力一组不吃，不然个人都知道自己在哪个组

“巧克力是否会引发肥胖”可以按吃/不吃巧克力分组，也可以按一周吃一次/两次/七次巧克力分组  
最开始选取的研究对象可能需要是健康的人

## 抽样方法

### common study types and sampling schemes

不同抽样机制可能会抽出完全一样的列联表，换言之列联表中看不出来怎么抽的  
需要从上下文中甄别

- 完全随机抽样
  - 每一个类别中 $n_{ij}$
  - 对于**eg横截面研究**， $n_{++}$ （总体，比如全班同学）是固定的，但每个格子里的数值不确定，因为落在每一个类别里的个体是随机数（比如随便抽了一个班做调研 男/女 食堂喜欢/不喜欢， $n_{++}$ 是**唯一已知的**，男女人数随机，喜欢人数也随机，每个格子都随机）

		Serum cholesterol (mg/100 cc)			
		0-199	200-219	220-259	260+
CHD		12	8	31	41
no CHD		307	246	439	245
total		319	254	470	286
		total			
		1329			

- 
- 每个格子里有多少人完全取决于当下截面的情况

- 分组随机抽样

- ① **eg 前瞻性队列研究**，每组人数固定，比如第一个暴露因素的组就想要200人，第二个也要200人，但这些人落入哪个列是不一定的（二项/多项分布）

- **fixed row sum**

OC-use group	MI status over 3 years	
	Yes	No
Current OC users	6.7	4993.3
→ Non-OC users	13.3	9986.7
Total	20	14,980

- 
- 在第一排这个组里和第二排的组里有各自的分布

- ② **eg case-control study**

- **fixed column sum**

	Lung cancer	No lung cancer	In total
Smoking	39	15	54
→ Non-smoking	21	25	46
In total	60	40	100

- 
- 也是二项/多项分布，但是按列分的

## 2x2列联表

Exposure State	Disease State		Total
	$D$ (Present)	$\bar{D}$ (Absent)	
	$E$ (Present)	$\bar{E}$ (Absent)	
	$n_{11}$	$n_{12}$	$n_{1.}$
	$n_{21}$	$n_{22}$	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n$

Treatment	Myocardial infarction		
	Yes	No	
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

We would like to investigate the effect of Exposure on Disease.

不影响-X Y独立

## proportion in 2x2 tables

- 两个条件概率  $\pi_1 = P[D|E]$      $\pi_2 = P[D|\bar{E}]$     条件概率
- 条件存在的情况下疾病出现的概率/条件不存在的情况下疾病出现的概率  
two sample proportion difference

## 2.1 Compare Proportions in 2x2 Tables

11

		Disease State		
		$D$ (Present)	$\bar{D}$ (Absent)	Total
Exposure State	$E$ (Present)	$n_{11}$	$n_{12}$	$n_{1.}$
	$\bar{E}$ (Absent)	$n_{21}$	$n_{22}$	$n_{2.}$
Total		$n_{.1}$	$n_{.2}$	$n$

Let  $\pi_1 = P[D|E]$   $\pi_2 = P[D|\bar{E}]$  条件概率

Case-control study 不可做此比较



Through proportion difference •  $E \perp D \Leftrightarrow \pi_1 - \pi_2 = 0$

Step 1 Estimate of  $\pi_1 - \pi_2$ :  $p_1 - p_2 = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}$

Step 2  $SE(p_1 - p_2) = \sqrt{p_1(1 - p_1)/n_{1+} + p_2(1 - p_2)/n_{2+}}$

Step 3 Large-sample  $(1 - \alpha)$  CI for  $\pi_1 - \pi_2$ :  $p_1 - p_2 \pm z_{\alpha/2} SE(p_1 - p_2)$ .

If this CI does not contain 0, we can reject  $H_0: E \perp D$  at significance level  $\alpha$ .

CAUTION: case-control study都没有发生率可言，竖着看是有意义的，横着看是没意义的

## relative risk in 2x2 tables (rare event)

问题：太小了，作差不好检测

- When (**rare event**) both  $\pi_1$  and  $\pi_2$  are close to zero, the difference  $\pi_1 - \pi_2$  may NOT be meaningful.
- For rare events, a more relevant measure for difference is **the relative risk**:  $RR = \pi_1/\pi_2$

- Properties of the relative risk (RR):

➤  $0 < RR$

➤  $RR > 1$ :  $\pi_1 > \pi_2$ , exposed group have a **higher** probability of contracting disease than unexposed group.

$RR < 1$ :  $\pi_1 < \pi_2$ , exposed group has a **lower** chance of contracting disease than unexposed group.

$RR = 1$ :  $\pi_1 = \pi_2$ , the risk of disease is the same in both exposure groups, exposure factor does not affect the disease status.

ratio 相对发生率

注意case-control study不能用

## 2.2 Relative risk in 2x2 table

- Given the 2x2 table from **cohort study** or **randomized clinical trials**, RR can be estimated by sample relative risk =  $p_1/p_2 = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$ .
- We can **NOT** calculate incidences or prevalence in a **case-control study**, and thus RR is not available.
- The distribution of RR is heavily skewed and it makes sense to **transform the data using natural log**.

$$\text{Var}(\log RR) = \frac{1-p_1}{n_{1+}p_1} + \frac{1-p_2}{n_{2+}p_2}$$

- Large sample CI:

$$\exp \left[ \log\left(\frac{p_1}{p_2}\right) \pm z_{\alpha/2} \sqrt{\frac{1-p_1}{n_{1+}p_1} + \frac{1-p_2}{n_{2+}p_2}} \right]$$

$= SE(\log(p_1/p_2))$

Note:  $n_{1+}p_1 = n_{11}$   
 $n_{2+}p_2 = n_{21}$

检验统计量RR:  $p_1/p_2$

检验统计量的分布: heavily skewed, 所以可以通过取对数 (之后比较像正态分布), 然后可以算CI (的端点), 然后再算指数换到原单位

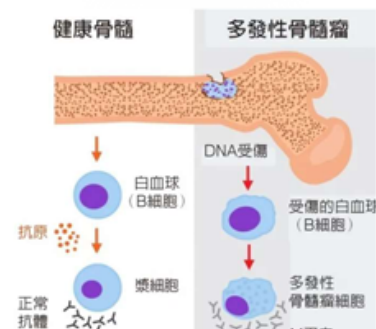
eg

**Example.** An efficacy study was conducted for the drug pamidronate in patients with stage III multiple myeloma and at least one lytic lesion 多发性骨髓瘤和至少一个溶血性病变.

In this randomized clinical trial, patients were assigned at random to receive either pamidronate ( $E$ ) or placebo ( $\bar{E}$ ). One endpoint reported was the occurrence ( $D$ ) of any skeletal events 骨骼事件 after 9 cycles of treatment or non-occurrence ( $\bar{D}$ ).

The results are given in Table. We will use the data to compute a 95% confidence interval for the relative risk of suffering skeletal events (in a time period of this length) for patients on pamidronate relative to patients not on the drug.

Treatment Group		Occurrence of Skeletal Event		
		Yes ( $D$ )	No ( $\bar{D}$ )	
Pamidronate ( $E$ )		47	149	196
Placebo ( $\bar{E}$ )		74	107	181
		121	256	377



暴露因素: 药物E

## 2.2 Relative risk in 2x2 table

17

$$\hat{\pi}_E = \frac{n_{11}}{n_{1.}} = \frac{47}{196} = 0.240 \quad \hat{\pi}_{\bar{E}} = \frac{n_{21}}{n_{2.}} = \frac{74}{181} = 0.409$$

$$RR = \frac{\hat{\pi}_E}{\hat{\pi}_{\bar{E}}} = \frac{.240}{.409} = 0.587$$

Treatment Group	Pamidronate (E) Placebo ( $\bar{E}$ )	Occurrence of Skeletal Event		
		Yes (D)	No ( $\bar{D}$ )	
		47	149	196
		74	107	181
		121	256	377

$$Var(\log RR) = \frac{(1 - \hat{\pi}_E)}{n_{11}} + \frac{(1 - \hat{\pi}_{\bar{E}})}{n_{21}} = \frac{(1 - .240)}{47} + \frac{(1 - .409)}{74} = .016 + .008 = .024$$

$$(0.587e^{-1.96\sqrt{.024}}, 0.587e^{1.96\sqrt{.024}}) = (0.587(0.738), 0.587(1.355)) = (0.433, 0.795)$$

We can be confident that the relative risk of suffering a skeletal event (in this time period) for patients on pamidronate (relative to patients not on pamidronate) is between 0.433 and 0.795.

Since this entire interval is below 1.0, we can conclude that pamidronate is effective at reducing the risk of skeletal events.

算对数的方差→算CI→算指数→结果和1相比较