大数据挖掘软件系列课之：
R语言:假设检验
案例：方差分析和列联表分析

肖韬

对两种变量之间关系的探索

用图形来探索

		自变量	
		连续型	类别型
反应变量/因变量	连续型	散点图	箱线图，概率密度图（分类别画）
	类别型	脊髓图/条件密度图	脊柱图

对两种变量之间关系的探索

用假设检验来探索

		自变量	
反应变量/因变量		连续型	类别型
	连续型		方差分析
	类别型		列联表分析

对两种变量之间关系的探索

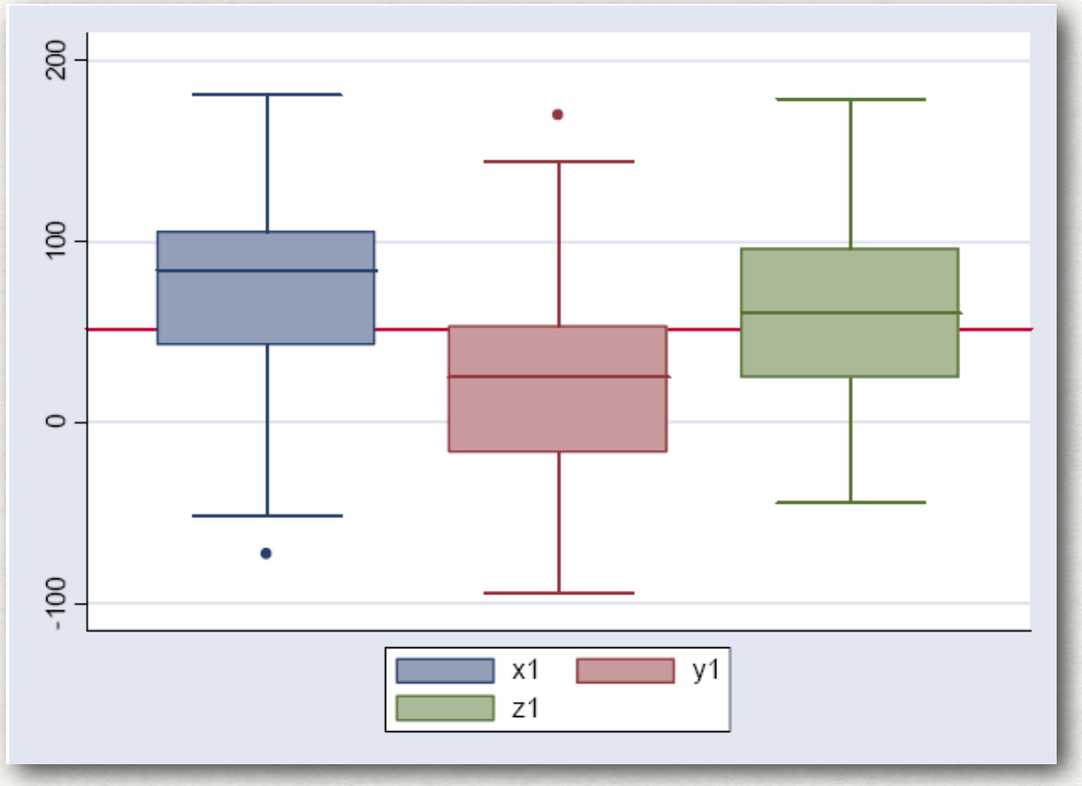
用假设检验探索

		自变量	
		连续型	类别型
反应变量/因变量	连续型		方差分析
	类别型		列联表分析

两种变化幅度：组间变化和组内变化

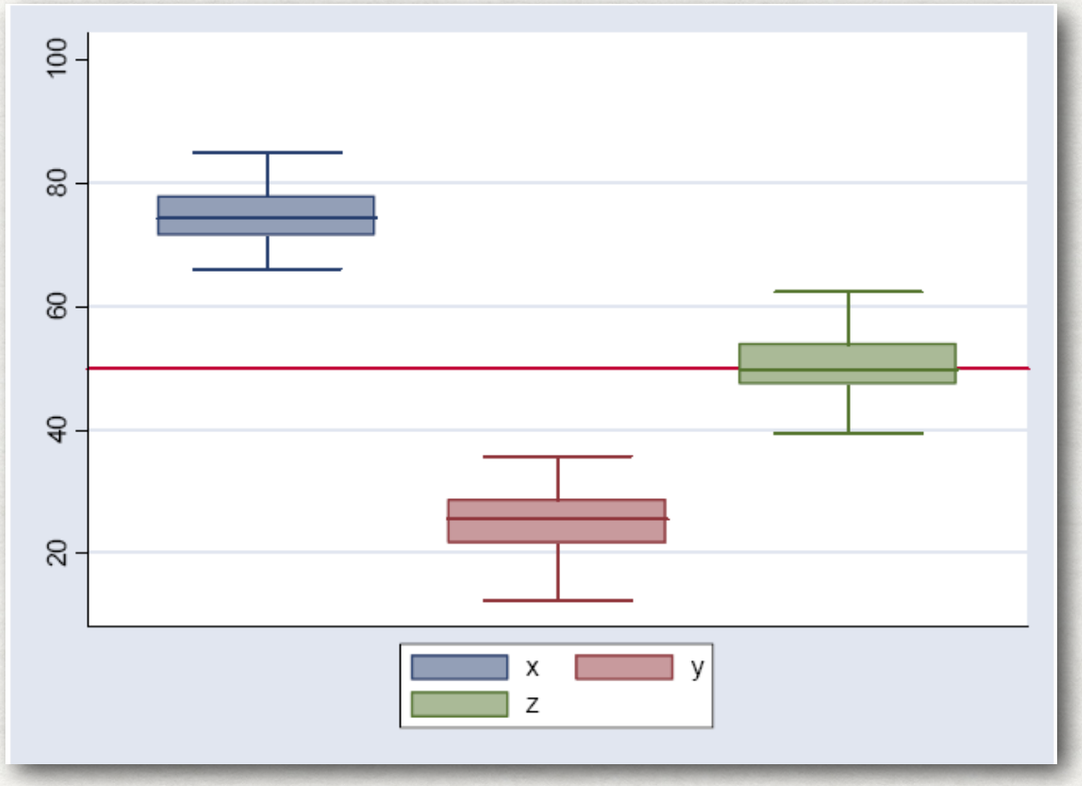
情况1: 组内变化幅度大，组间变化幅度和组内变化幅度的比值小

X1, Y1, Z1三个组的均值分别是75, 25, 50



情况2: 组内变化幅度小, 组间变化幅度和组内变化幅度的比值大

X, Y, Z三个组的均值分别是75, 25, 50



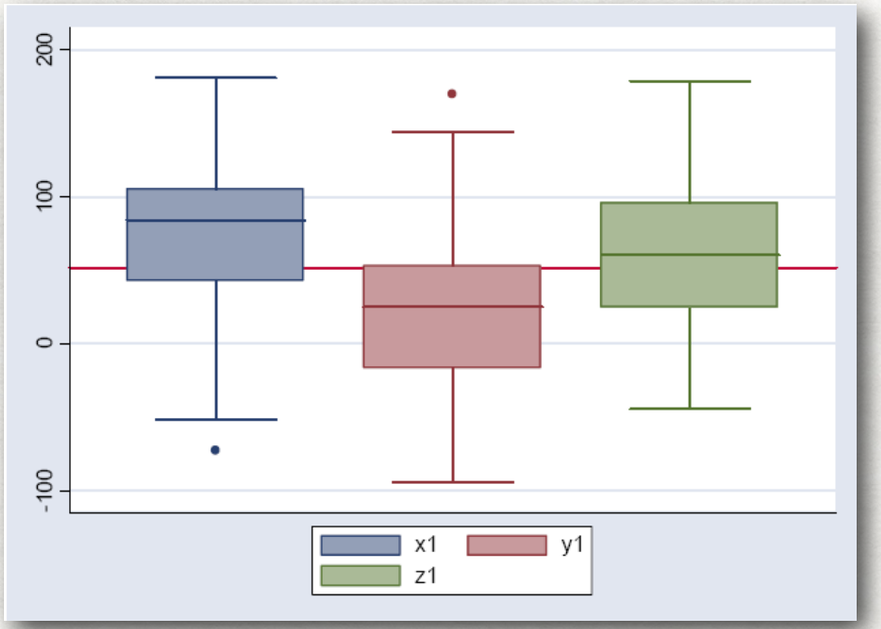
两种变化幅度：

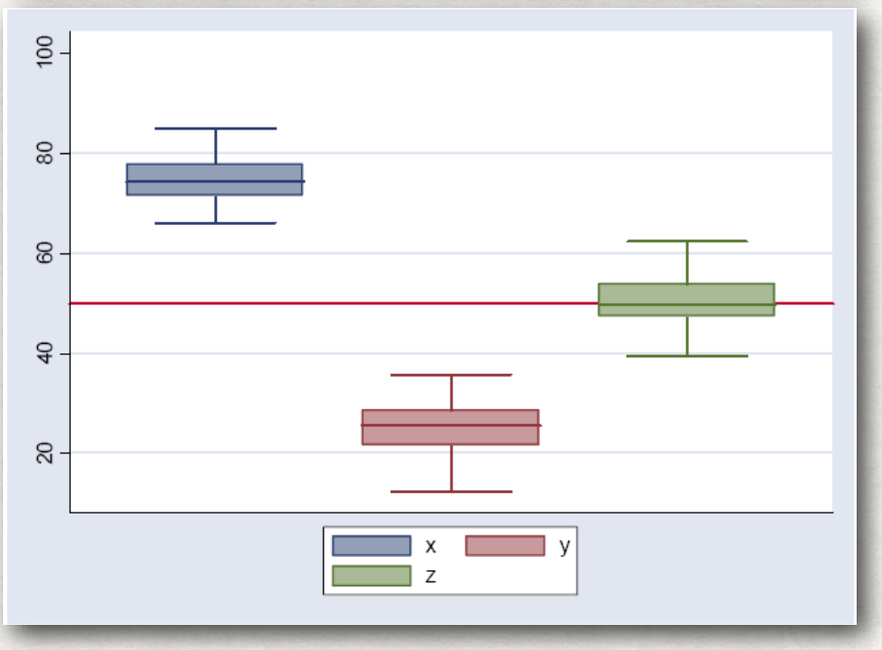
1. 组内变化（*Within-group variation*）： 每组内的观测值在组平均值的上下进行波动变化的幅度。

2. 组间变化（*Between-group variation*）: *K*个组各组的平均值在*K*个组总体的总平均值的上下进行波动变化的幅度。

- 注意，如果*K*个组的各组的组平均值很接近，那么我们就会得到一个很小的“组间变化/组内变化”的比值。因此，这个比值的大小可以对原假设的可靠性进行评估。

单因素方差分析





单因素方差分析 (one-way ANOVA).

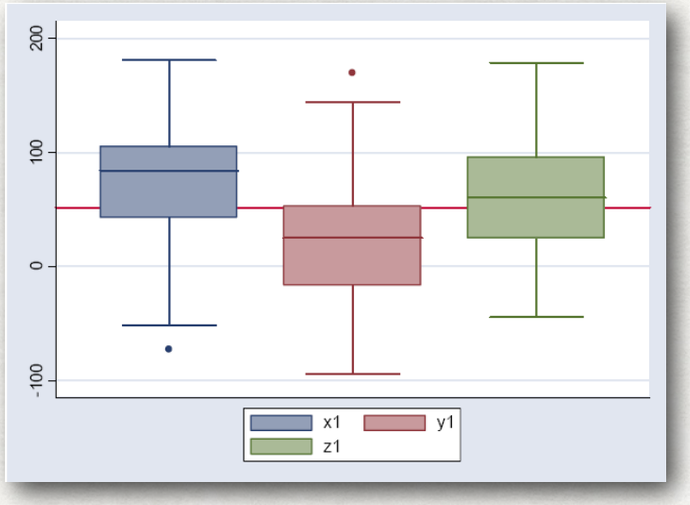
. K个类别的人群总体：

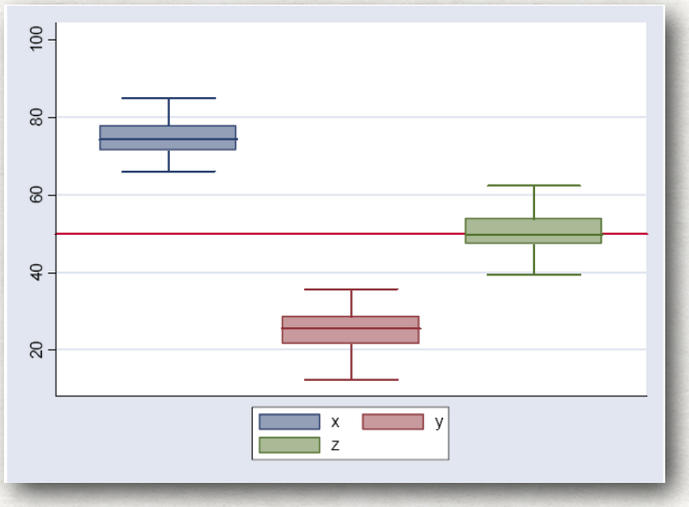
		类别（组）1	类别（组）2	…	类别（组）K
人群总体	均值	μ_1	μ_2	…	μ_K
	方差	σ_1^2	σ_2^2	…	σ_K^2
样本	样本均值	$\bar{x}_1$	$\bar{x}_2$	…	$\bar{x}_K$
	样本方差	s_1^2	s_2^2	…	s_K^2
	样本量	n_1	n_2	…	n_K

原假设: $H_0 : \mu_1 = \mu_2 = ... = \mu_K$ （即K个组的均值全相等）

替代假设: K 个组的均值不全相等

构建方差分析（ANOVA）的假设检验统计量





组内变化：

$$SSW = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots (n_K - 1)s_K^2$$

组间变化：

$$SSB = \sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2$$

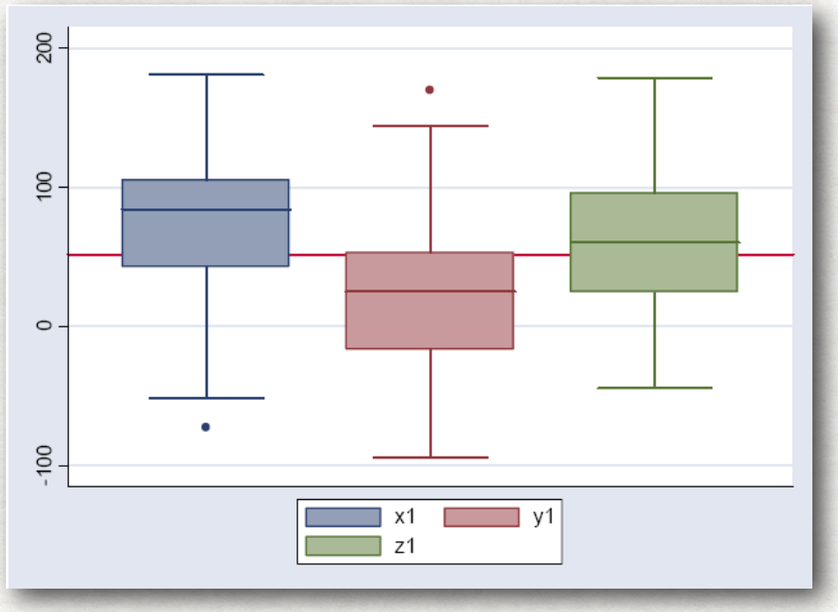
总变化：

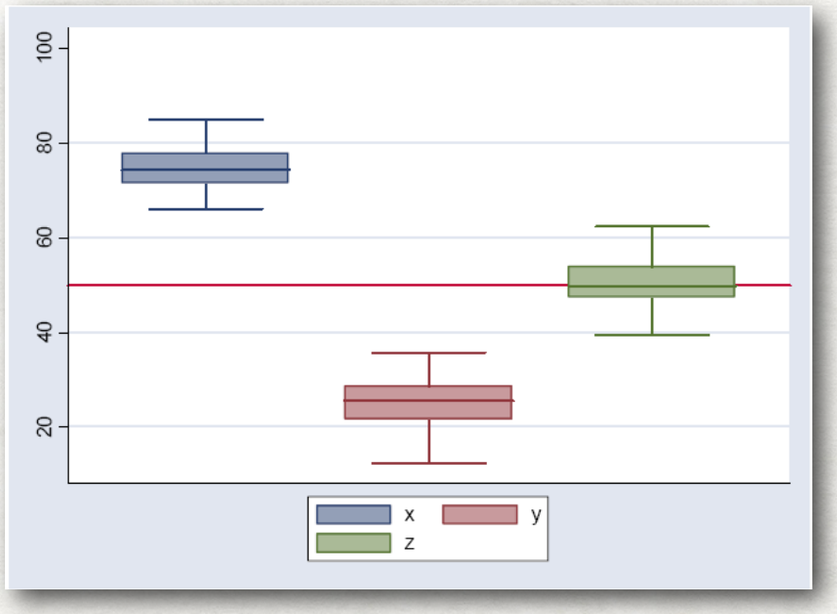
$$SST = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$$

注意，经推导我们会得到：SST= SSW + SSB

如果“K个组的均值全相等”的原假设是对的，那么我们就期望从观测到的数据中会看到“组间变化/组内变化（SSB/SSW）的比值小”的结果。如何来衡量这个结果是不是“够小”呢？我们使用“概率”——“比我们从观测到的数据中计算出的SSB/SSW比值更反常的值的概率”“更反常”是相对于原假设来说的，对于“K个组的均值全相等”的原假设，“反常”指SSB/SSW比值很大，因为这一现象和这个原假设相抵触。如果我们能知道这个SSB/SSW比值的概率分布函数，那么我们就可以从这个概率分布函数算出这个概率（这个概率被称为*p*值）。如果这个概率很小（比如小于0.05），那么我们就可以推翻原假设（比如我们可以说：我们在0.05的显著性水平上推翻原假设）。

作出决策





如果我们把 SSB/SSW 比值做一点如下的调整，那么这个调整后的比值就真的具有一个概率分布函数（可证明得出）：

$$F = \frac{SSB/(K-1)}{SSW/(n-K)} \sim F_{K-1,n-K}$$

我们把以上这个比值叫做 F 统计量（记住：统计量都是随机数，因此有概率分布函数），它的概率分布函数是一个有两个参数（K-1 和 n-K）的概率分布函数，记为：$F_{K-1,n-K}$。我们称这两个参数 K-1 和 n-K 为 F 概率分布函数的自由度。

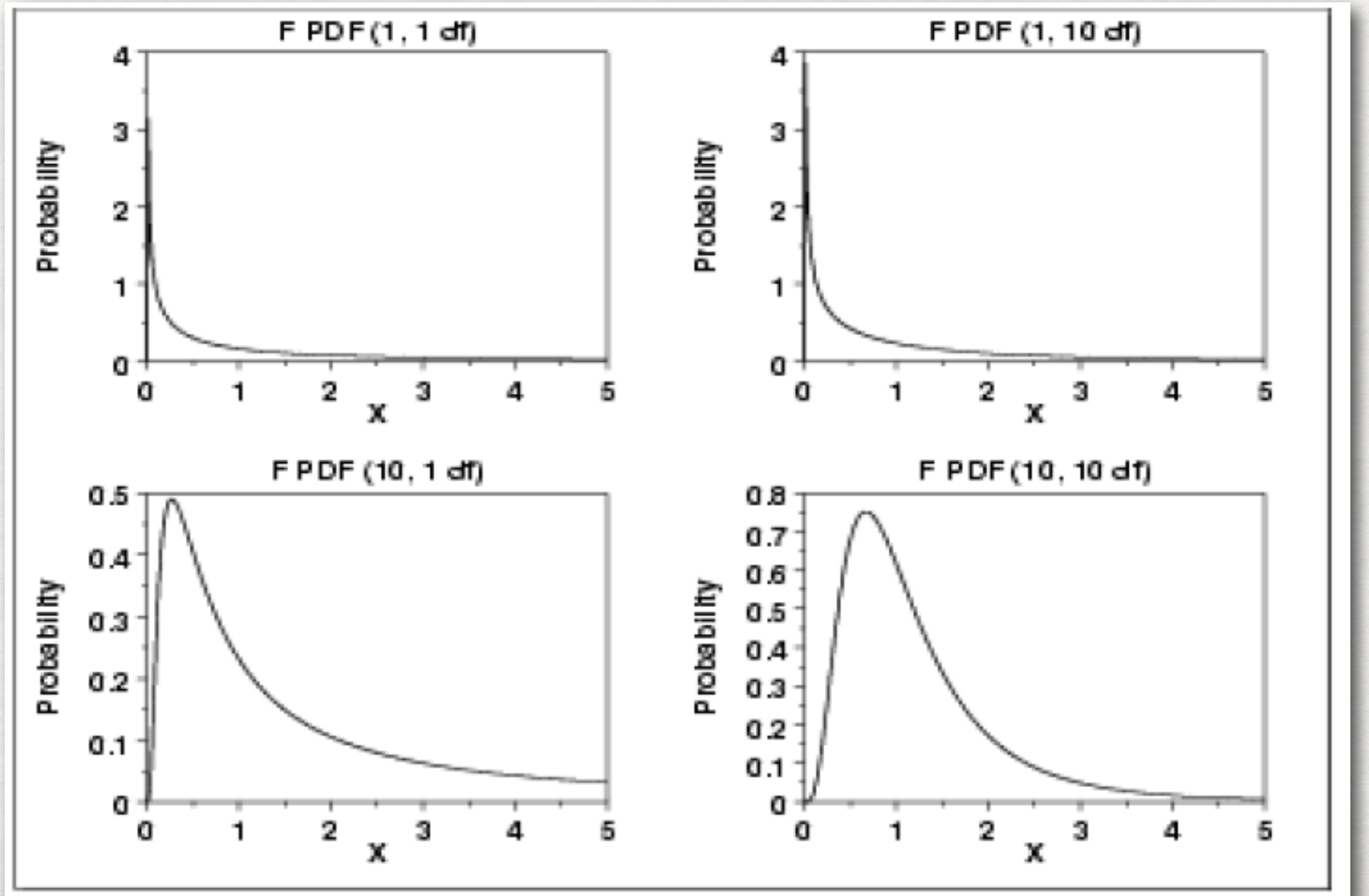
当原假设H_0是正确的时，从观测到的值算出的组间变化幅度和组内变化幅度的比值（SSB/SSW）不出意外会较小，因此算出的F的值也会较小。

当原假设H_0是错误的时，从观测到的值算出的组间变化幅度会超过组内变化幅度，因此算出的F的值会较大（比1大很多）。

什么时候作出推翻原假设的决策：如果我们从观测到的值中计算出的F统计量的值很大，大到比这个值更大的值出现的概率小于0.05，那么我们就可以在0.05的显著性水平上推翻原假设。

F统计量的概率分布函数（即F分布函数）的特点

- F统计量的可能出现值(即F分布函数的定义域)都是正数
- F分布函数是不对称的，呈右偏斜形态
- F分布函数的形状由两个自由度参数来决定。



进行单因素方差分析假设检验的过程

变化类型	SS	d.f.	MS	F ratio
组间变化	SSB	K-1	MSB	MSB/MSW
组内变化	SSW	n-K	MSW	
总变化	SST	n-1		

统计量为观测到的 F 比值MSB/MSW，这个比值是个符合F(K-1,n-K)分布的随机数，其值会随着观测值的变化而变化。

- 如果通过已得到的观测值计算出的F 比值MSB/MSW对应的*p*值小于0.05（即F(K-1,n-K)分布的随机数比这个计算出的F 比值MSB/MSW更大的概率小于0.05），则推翻原假设H_0；
- 如果*p*值大于或等于0.05，则不推翻原假设H_0

一个例子：血压收缩压(SYSTOLIC BLOOD PRESSURE)增量数据

Blackboard的电子教案模块的Dataset文件夹的sbp.csv文件

变量名	标签
drug	所用药的类别（1-4）
disease	病人的病的类别（1-3）
systolic	病人血压收缩压的增量

```
> sbp = read.csv("Datasets/sbp.csv")
> sbp$drug = as.factor(sbp$drug)
> fit1 = aov(systolic~drug,data=sbp)
> sbp$disease = as.factor(sbp$disease)
> fit2 = aov(systolic~disease,data=sbp)
> summary(fit2)
             Df Sum Sq Mean Sq F value Pr(>F)
disease       2    489   244.3   1.518  0.228
Residuals    55   8852   160.9
> 1-pf(1.518,2,55)
[1] 0.2281884
> sbp = read.csv("Datasets/sbp.csv")
> sbp$drug = as.factor(sbp$drug)
> fit1 = aov(systolic~drug,data=sbp)
> summary(fit1)
             Df Sum Sq Mean Sq F value   Pr(>F)
drug          3   3133  1044.4   9.086 5.75e-05 ***
Residuals    54   6207   114.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 1-pf(9.086,3,54)
[1] 5.751618e-05
> sbp$disease = as.factor(sbp$disease)
> fit2 = aov(systolic~disease,data=sbp)
> summary(fit2)
             Df Sum Sq Mean Sq F value Pr(>F)
disease       2    489   244.3   1.518  0.228
Residuals    55   8852   160.9
> 1-pf(1.518,2,55)
[1] 0.2281884
```

对两种变量之间关系的探索

用假设检验来探索

		自变量	
反应变量/因变量		连续型	类别型
	连续型		方差分析
	类别型		列联表分析

列联表：显示两个类别型变量的相关关系

一个例子：乳腺癌

乳腺癌	有乳腺癌家族史		总数
	是	否	
是	400	100	500
否	9600	4900	14500
总数	10000	5000	15000

列联表的假设检验：自行车骑行安全的例子

在一个关于戴头盔对防止自行车骑行跌倒导致的头部损伤的效果的研究中，研究人员搜集了793个自行车事故的当事人的数据：

头部受损伤	戴了头盔		总数
	是	否	
是	17	218	235
否	130	428	558
总数	147	646	793

原假设（H_0）：在戴了头盔和没戴头盔的这两类自行车事故的当事人人群中，头部受损伤的人的比例是一样的（或者说，戴头盔和头部受损伤两个事件是独立不相关的两个事件）

替代假设（H_A）：以上两类人群中头部受损伤的比例是不一样的（或者说，戴头盔和头部受损伤两个事件是相关的两个事件）

列联表的假设检验：自行车骑行安全的例子

头部受损伤	戴了头盔		总数
	是	否	
是			235
否			558
总数	147	646	793

头部受损伤	戴了头盔		总数
	是	否	
是	43.6	191.4	235
否	103.4	454.6	558
总数	147	646	793

如果原假设是对的 (即在戴了头盔和没戴头盔的这两类自行车事故的当事人人群中，头部受损伤的人的比例是一样的)

那么可以从数据中估计所有事故当事人中头部受伤的人数比例为:

$$\frac{235}{793} = 29.6\%$$

估计所有事故当事人中头部没有受伤的人数比例为:

$$\frac{558}{793} = 70.4\%$$

$147(29.6\%) = 43.6$

$646(29.6\%) = 191.4$

$147(70.4\%) = 103.4$

$646(70.4\%) = 454.6$

列联表的假设检验：自行车骑行安全的例子

头部受损伤	戴了头盔		总数
	是	否	
是	44 (43.6)	191 (191.4)	235
否	103 (103.4)	455 (454.6)	558
Total	147	646	793

一个极端的情况:

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

$$+ \frac{(O_3 - E_{3i})^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

$$= \frac{(44 - 43.6)^2}{43.6} + \frac{(191 - 191.4)^2}{191.4}$$

$$+ \frac{(103 - 103.4)^2}{103.4} + \frac{(455 - 454.6)^2}{454.6}$$

$$= 0.0037 + 0.0008 + 0.0015 + 0.0004$$

$$= 0.0064$$

头部受损伤	戴了头盔		总数
	是	否	
是	0 (43.6)	235 (191.4)	235
否	147 (103.4)	411 (454.6)	558
总数	147	646	793

另一个极端的情况:

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

$$+ \frac{(O_3 - E_{3i})^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

$$= \frac{(0 - 43.6)^2}{43.6} + \frac{(235 - 191.4)^2}{191.4}$$

$$+ \frac{(147 - 103.4)^2}{103.4} + \frac{(411 - 454.6)^2}{454.6}$$

$$= 43.60 + 9.93 + 18.38 + 4.18$$

$$= 76.09$$

卡方分布（χ^2分布）

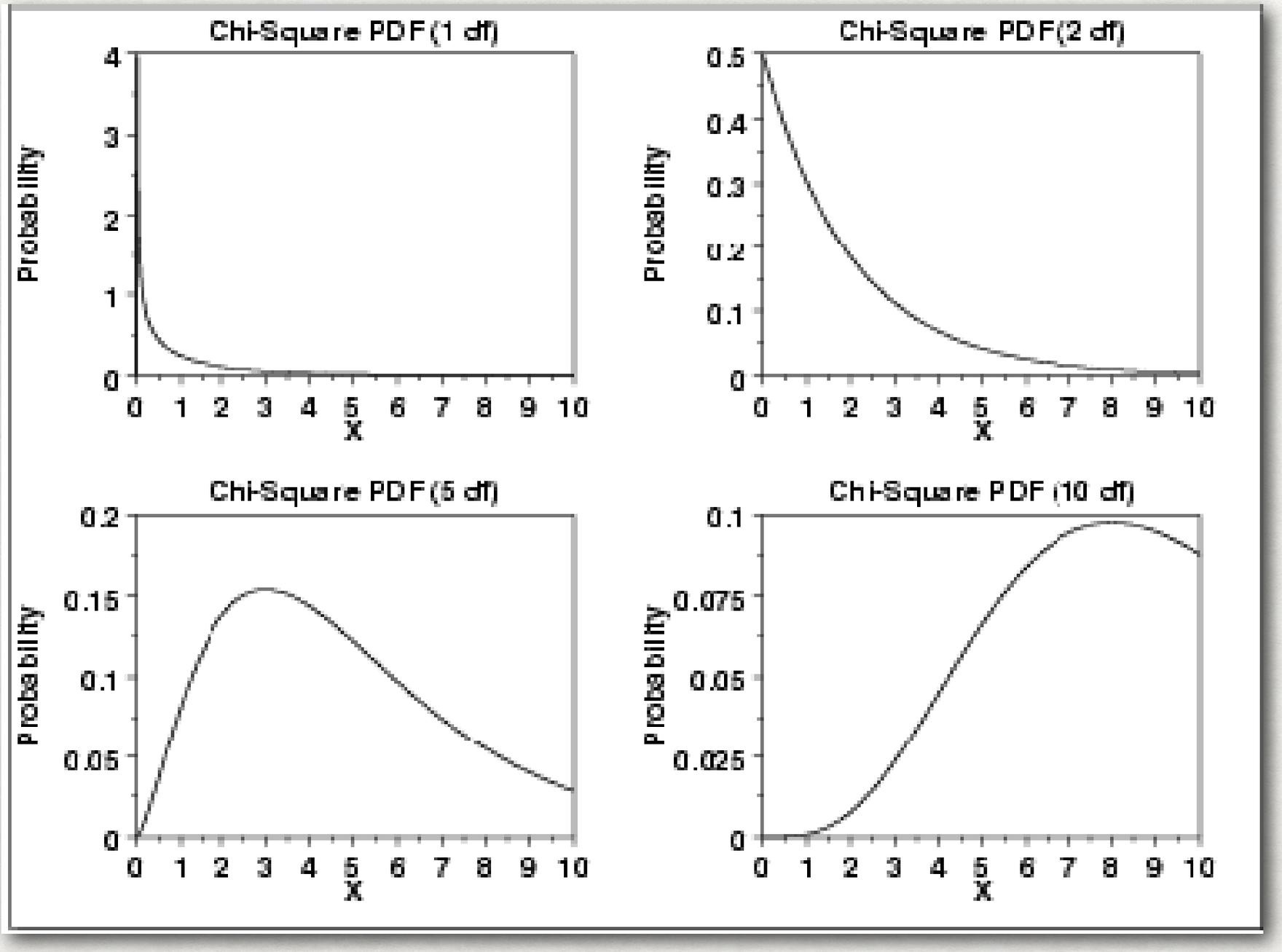
$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

当样本量足够大时，以上这个从2x2型列联表计算出的统计量满足自由度参数为1的卡方分布（Chi-square分布，记为χ^2分布，其也是一种概率分布函数，这个分布函数只有一个参数，就是自由度参数）。

- 当从观测到的数据计算得出的统计量的值大到一定程度时，我们推翻原假设。但是，要大到什么程度呢？如何决定？

卡方分布（χ^2分布）的特点

- 卡方分布函数的定义域都是正数；
- 卡方分布函数是不对称的；
- 卡方分布函数的形状由其自由度参数来决定。



列联表的假设检验：自行车骑行安全的例子

头部受损伤	戴了头盔		总数
	是	否	
是	17 43.6	218 191.4	235
否	130 103.4	428 454.6	558
总数	147	646	793

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

$$+ \frac{(O_3 - E_{3i})^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

$$= \frac{(17-43.6)^2}{43.6} + \frac{(218-191.4)^2}{191.4}$$

$$+ \frac{(130-103.4)^2}{103.4} + \frac{(428-454.6)^2}{454.6}$$

$$= 16.23 + 3.70 + 6.84 + 1.56$$

$$= 28.33$$

$$d.f. = 1$$

列联表的假设检验：自行车骑行安全的例子

头部受损伤	戴了头盔		总数
	是	否	
是	17 43.6	218 191.4	235
否	130 103.4	428 454.6	558
总数	147	646	793

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

$$+ \frac{(O_3 - E_{3i})^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

$$= \frac{(17-43.6)^2}{43.6} + \frac{(218-191.4)^2}{191.4}$$

$$+ \frac{(130-103.4)^2}{103.4} + \frac{(428-454.6)^2}{454.6}$$

$$= 16.23 + 3.70 + 6.84 + 1.56$$

$$= 28.3$$

$$d.f. = 1$$

从观测到的数据计算出的χ^2统计量的值约为28.255，其对应的p值为1.03896e-07<0.05，因此，我们在0.05的显著性水平上推翻原假设，我们的结论是：基于我们的数据，骑行事故当事人的头部是否受伤跟戴头盔的因素有关联（或者说，戴头盔的骑行事故当事人和不戴头盔的骑行事故当事人的头部受伤人数的比例是不一样的）。

```
> 1-pchisq(28.255,1)
[1] 1.063396e-07
> #helmet <- as.table(rbind(c(17, 218), c(130, 428)))
> #helmet <- rbind(c(17, 218), c(130, 428))
> helmet <- matrix(c(17, 218, 130, 428),nrow=2, byrow = TRUE)
> chisq.test(helmet,correct=FALSE)

        Pearson's Chi-squared test

data:  helmet
X-squared = 28.255, df = 1, p-value = 1.063e-07

> 1-pchisq(28.255,1)
[1] 1.063396e-07
```