

χ^2 (Chi-square) 检验的效应量通常不直接计算为单一数值，但可以通过一些方式来描述或量化效应的大小。对于 χ^2 检验，特别是用于分类数据的独立性检验（如列联表分析）时，一个常用的效应量指标是 Cramer's V（也称为 Phi 系数或 Φ 系数），它表示了两个分类变量之间关系的强度。

Cramer's V 的计算公式依赖于 χ^2 统计量和样本大小，并且它的取值范围在 0 到 1 之间，其中 0 表示没有关联，1 表示完全关联。但是，需要注意的是，当样本量很大时，即使是很小的实际效应也可能导致 χ^2 统计量很大，从而可能高估效应量。

```
library(rstatix)
cramer_v <- cramer_v(as.table(as.matrix(Two_categories)))
```

vcd 包中的 `assocstats()` 函数也可以用来计算 Cramer's V 系数。这个函数主要用于计算二维列联表的 phi 系数、列联系数和 Cramer's V 系数，以度量两个分类变量之间的相关性。

```
library(vcd)

# 使用 assocstats() 函数计算统计量
assoc_stats <- assocstats(my_table)
```

```
# 查看结果，包括 Cramer's V 系数
print(assoc_stats)
```

请注意，你需要将 `my_table` 替换为你实际使用的列联表。`assocstats()` 函数将返回一个包含多个统计量的对象，你可以通过打印该对象来查看所有结果，包括 Cramer's V 系数。

`dchisq` 给出密度，`pchisq` 给出分布函数，`qchisq` 给出分位数函数，`rchisq` 生成随机偏差。

```
pchisq(q = 4, df = 3, lower.tail = F)
```

```
q =  $\chi^2$ 
```

If `lower.tail = T` (default) probabilities are $P[X \leq x^2]$, otherwise, $P[X > x^2]$.

```
p-value = pchisq(q =  $\chi^2$ , df = n-1, lower.tail = F)
```

`n` = number of categories (e.g. 4 season preferences, `df` = 3)

With $\alpha = 0.05$, with can reject H_0 . There is a difference between the observed and the expected season preferences.

Example1: Season Preferences: Test for Goodness-of-Fit

Difference between observed and expected distribution

```
> chisq.test(Poll_seasons, correct = FALSE, p = rep(1/4, 4))
```

Chi-squared test for given probabilities

data: Poll_seasons

X-squared = 8.0968, df = 3, p-value = 0.04405

One-way table → degrees of freedom = #columns - 1
d.f. = c - 1

Example2: Two Categorical Variables: Chi-square test for homogeneity

If two or more variables (row variable) have different distributions of a categorical variable (column variable)

	Spring	Summer	Fall	Winter
Severe allergies	5	1	1	9
Mild allergies	8	5	2	5
Sporadic allergies	9	8	3	9
Never allergic	18	16	12	5

Test for Homogeneity

Question: Is there a difference between the distribution of allergic reactions in the different seasons?

H₀: The distribution of allergic reactions is the same for the people who preferred different seasons

H₁: The distribution of allergic reactions is **not** the same for the people who preferred different seasons

如何算 expected? 对应行的 sum 乘对应列的 sum 除以 total sum

	Spring	Summer	Fall	Winter	Total
Severe allergies	5	1	1	9	16
Mild allergies	8	5	2	5	20
Sporadic allergies	9	8	3	9	29
Never allergic	18	16	12	5	51
Total	40	30	18	28	116

= 5.52

Expected Frequencies

	Spring	Summer	Fall	Winter
Severe allergies	5.52 (40 x 16 / 116)	4.14	2.48	3.86
Mild allergies	6.90	5.17	3.10	4.83
Sporadic allergies	10	7.5	4.5	7
Never allergic	17.59	13.19	7.91	12.31

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right)$$

$$= \left(\frac{(5 - 5.52)^2}{5.52} \right) + \left(\frac{(1 - 4.14)^2}{4.14} \right) + \left(\frac{(1 - 2.48)^2}{2.48} \right) + \left(\frac{(9 - 3.86)^2}{3.86} \right) + \dots$$

```
> Severe <- data.frame(Spring = 5, Summer = 1, Fall = 1, Winter = 9)
> Mild <- data.frame(Spring = 8, Summer = 5, Fall = 2, Winter = 5)
> Sporadic <- data.frame(Spring = 9, Summer = 8, Fall = 3, Winter = 9)
> Never <- data.frame(Spring = 18, Summer = 16, Fall = 12, Winter = 5)
> Two_categories <- rbind(Severe, Mild, Sporadic, Never)
> chisq.test(Two_categories)
```

Pearson's Chi-squared test

data: Two_categories
X-squared = 18.994, df = 9, p-value = 0.02524

With $\alpha = 0.05$

Degrees of Freedom

(#rows - 1) x (#columns - 1)

d.f. = $r \times c - 1 - (r-1) - (c-1) = (r-1)(c-1)$

Example 3: chi-square test for independency

If two or more response variables are independent on an explanatory variable

Question: We need to analyse the survival data of a geneX knockout mice at 1 year. Does geneX affect lifespan of mice?

~~H₀: The survival of mice is independent on geneX~~

H₁: The survival of mice is dependent on geneX

The lifespan is dependent on geneX

		Explanatory variable		
Response variable		WT	KO	Total
	Alive	7	2	9
	Dead	3	7	10
	Total	10	9	19

Expected frequencies		WT	KO
	Alive	4.7	4.3
	Dead	5.3	4.7

$\chi^2 = 4.3372, d.f. = 1, p = 0.037$
With $\alpha = 0.05$

3-Way Sample: Three Categorical Variables (rxcxl)

For data from one sample, does the three categorical variables are independent of each other?

Example: We need to analyse the survival data of a geneX knockout mice at 1 year.
Is geneX, sex and lifespan independent of each other?

	WT		KO	
	Male	Female	Male	Female
Alive	40	34	20	25
Dead	9	7	15	20

Chi-square test

Total mice: 170

Total male vs female: 84:86 (49% vs 51%)

Total alive vs dead: 119:51 (70% vs 30%)

Total WT vs KO: 90:80 (53% vs 47%)

Expected:

Male, alive, WT = $170 \times 49\% \times 70\% \times 53\% = 31$

Male, alive, KO

Male, dead, WT

Male, dead, KO

Female, alive, WT

Female, alive, KO

Female, dead, WT

Female, dead, KO = $170 \times 51\% \times 30\% \times 47\% = 12.1$

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right)$$

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right) = 15.765, \quad p = 0.0034, \quad d.f. = 4 \quad \text{Degrees of Freedom}$$

With $\alpha = 0.05$

$$d.f. = r \times c \times 1 - 1 - (r-1) - (c-1) - (1-1) \\ = 2 \times 2 \times 2 - 1 - (2-1) - (2-1) - (2-1) \\ = 4$$

There is interdependency among geneX, sex and lifespan in mice at 1 year

Apply the chi-square test. Note: You cannot use an 3-dimensional object in the `chisq.test()`. Convert the array to a table object and find the chi-square test result in the summary of the object.

```
data2 = array(data = 0, dim = c(2, 2, 2))
mice = as.table(data2)
mice
summary(mice)
```

NOTE:

Ordinal Variable is a categorical variable with an associated set order or scale (ratings, pain levels, age groups, allergy severity)

There is no standardized interval scale of measurement.

Severe allergies
Mild allergies
Sporadic allergies
Never allergic

It can measure qualitative traits

Numeric operations cannot be used.

Ordinal variable only has median but not a mean!

Statistic analysis: compare the median values across samples

Wilcoxon test:

1- or 2-sample test, especially useful for paired samples

下面这个例子我觉得应该是 one sample 的 test 你说呢是吧

品质: best good medium bad (4 3 2 1)

个数: 10 15 20 5

Kruskal-Wallis 1-way test:

3- or more sample test, non-parametric alternative to the 1-way ANOVA

比如说之前我们做过的 attendance 状况, previous grade 状况, recent grade 状况??? 是这样吗, 因为之前做的 recent grade 是数字

这个可能需要的是 attendance, previous grade, recent grade 的状况然后对应不同的人数

e. g.

Kruskal-Wallis 检验是一种非参数检验, 用于比较多个独立样本 (通常来自三个或更多组) 的中位数是否相同。这个检验可以看作是多个独立样本的 Mann-Whitney U 检验 (也就是 Wilcoxon rank sum test) 的推广。

要在 R 中使用 Kruskal-Wallis 检验, 你可以使用 `kruskal.test()` 函数。

假设你有三个向量, 分别代表三个不同的组

```
group1 <- c(1, 2, 3, 4, 5, 6, 7, 8, 9)
```

```

group2 <- c(3, 4, 5, 6, 7, 8, 9, 10, 11)
group3 <- c(6, 7, 8, 9, 10, 11, 12, 13, 14)

# 将这些向量合并到一个列表中，并创建一个分组向量来指示每个值属于哪个组
data <- c(group1, group2, group3)
group <- factor(rep(c("A", "B", "C"), each = length(group1)))

# 使用 kruskal.test() 函数进行 Kruskal-Wallis 检验
result <- kruskal.test(data ~ group)

# 打印结果
print(result)

```

在上面的代码中，我们首先创建了三个向量 group1、group2 和 group3，它们分别代表三个不同组的数据。然后，我们将这些数据合并到一个向量 data 中，并创建了一个分组向量 group，它指示了 data 中的每个值属于哪个组。

接下来，我们使用 kruskal.test() 函数进行 Kruskal-Wallis 检验。在这个函数中，我们使用公式 data ~ group 来指定我们的数据和分组信息。这个公式告诉 R 我们想要比较 data 中的值，这些值根据 group 进行分组。

最后，我们打印出结果。这个结果将包含一个统计量（通常是卡方统计量）、自由度、p 值和一个描述性标签。如果 p 值小于你选择的显著性水平（通常是 0.05），那么你将拒绝原假设（即所有组的中位数都相同），并认为至少有一个组的中位数与其他组不同。

我感觉和上面 test for homogeneity 的区别就是，

上面的 test for homogeneity 是一个 response variable，可以有多个 explanatory variable，而 test for independency 是一个 explanatory variable，可以有多个 response variable，但是计算过程的本质是一样的，而且列联表的行和列可以转置，计算结果是一样的。

然而上面的理解并不正确！

下面是文心一言的回答：

1. 独立性检验 (Test of Independence)：

主要用于判断两个分类变量（通常是两个响应变量或观察变量，意思是响应变量和分类变量分别是行变量和列变量，一般是两行两列，但是可以是多行多列）之间是否独立。相当是一个样本里我有好多分类变量！

在列联表中，两个分类变量分别对应行和列。

例如，我们可能想要知道性别（男/女）和投票倾向（支持/反对）之间是否独立。

我们不会有多于两个的响应变量，但可能会有多个类别水平（比如，性别可能有男、女、其他等多个水平；投票倾向可能有支持、反对、中立等多个水平）。

2. 同质性检验 (Test for Homogeneity)：

主要用于判断多个独立样本在某一分类变量上的分布是否相同（即它们是否具有同质性）。

这通常用于检验来自不同总体（如不同群体、不同地区、不同时间点的样本）的观测值在某一分类变量上的分布是否一致。

在这种情况下，有一个或多个解释变量（通常是组别、时间等），但响应变量是共同的，并用于比较不同组的分布。

列联表中的行通常代表组别（e. g. 国家），列代表分类变量的不同水平（e. g. season preference）。

关于您的理解：

1. 您提到“一个 response variable 可以有多个 explanatory variable”在卡方检验中并不准确。在卡方检验中，我们通常不谈论“解释变量”和“响应变量”的区分，而是讨论分类变量。
2. 您提到“test for independency 是一个 explanatory variable，可以有多个 response variable”也是不正确的。独立性检验涉及两个分类变量，没有解释和响应之分。
3. 关于列联表的行和列可以转置，确实如此，但这不会改变检验的本质。转置只是从另一个角度看待相同的数据，检验结果应该相同（或至少在统计显著性上等价）。

Appropriate use of chi-square test

1. The appropriate use of the chi-square to approximate the distribution of the good-of-fit test statistic depends on both the sample size (i. e. observation number) and the number of cells.
2. Pearson's chi-square is an approximation that requires large sample sizes.
3. 列联表 (contingency table) 中，每个单元格 (cell) 的预期频数 (expected frequency) 都不应该小于 1。
4. 在列联表的单元格中，预期频数小于 5 的单元格的比例不应超过 20%。
5. Fisher's Exact Test: Alternative for contingency tables when sample sizes are small

	C1	C2	Row Total
R1	a	b	a+b
R2	c	d	c+d
Column Total	a+c	b+d	a+b+c+d=n

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

The Fisher's exact test can also be used on contingency tables larger than 2x2

We need to analyse the survival data of a geneX knockout mice at 1 year. Does geneX affect survival of mice?

	WT	KO	Total
Alive	7	2	9
Dead	3	7	10
Total	10	9	19

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} = \frac{9!10!10!9!}{7!2!3!7!19!} = 0.069 \quad \text{accept } H_0$$

With $\alpha = 0.05$

The survival of mice is independent on geneX!

Statistical power of chi-square test

DescTools::power.chisq.test 和 pwr::pwr.chisq.test 是两个 R 包中的函数，分别用于计算卡方检验 (Chi-squared test) 的功率 (Power)。

1. DescTools::power.chisq.test

DescTools::power.chisq.test 是 DescTools 包中的一个函数，用于计算卡方检验的功率。

```
# 安装 DescTools 包（如果尚未安装）
install.packages("DescTools")

# 加载 DescTools 包
library(DescTools)

# 使用 power.chisq.test 函数
# 示例：计算 df=1 的卡方检验在 alpha=0.05，效应大小 (w) 为 0.3，样本大小为 100 时的功率
power_result <- power.chisq.test(w = 0.3, N = 100, df = 1, sig.level = 0.05)
print(power_result)
```

在这个示例中，w 是效应大小（通常表示为非中心参数），N 是样本大小（对于 2x2 表格，这是两个组的总和），df 是自由度（对于 2x2 表格，这通常是 1），sig.level 是显著性水平（通常是 0.05）。

在 DescTools::power.chisq.test 中，你通常会设定一个效应大小（比如两个类别间相对频数差异的预期），然后使用其他参数（如样本大小、显著性水平和自由度）来计算所需的功率。但是，DescTools 包并没有直接提供一个函数来计算这个非中心参数 w，因为它通常是根据研究背景和假设来设定的。

2. pwr::pwr.chisq.test

pwr::pwr.chisq.test 是 pwr 包中的一个函数，也用于计算卡方检验的功率。

```
# 安装 pwr 包（如果尚未安装）
install.packages("pwr")

# 加载 pwr 包
library(pwr)

# 使用 pwr.chisq.test 函数
# 示例：计算 df=1 的卡方检验在 alpha=0.05，效应大小 (w) 为 0.3，样本大小为 100 时的功率
# 注意：pwr.chisq.test 需要的是每组的样本大小，而不是总样本大小
power_result <- pwr.chisq.test(w = 0.3, df = 1, sig.level = 0.05, power = NULL, n = 50) # 假设每个组有 50 个样本
print(power_result)
```

在这个示例中，n 是每个组的样本大小（不是总样本大小）。注意，pwr.chisq.test 允许你指定你想要的功率（power 参数），然后它会返回满足该功率所需的样本大小。如果你不提供 power 参数，它会返回在当前参数下的功率。

在统计学的卡方检验中，w 通常代表效应大小 (effect size) 或称为非中心参数 (noncentrality parameter)。这个参数描述了观察到的效应相对于随机误差的大小。

效应大小 (w) 的计算

在 `pwr.chisq.test` 函数中, `w` 的计算并不是直接通过观测数据得出的, 而是基于研究者的预期或假设。但是, 对于 2x2 的列联表, 可以使用一些方法估算效应大小。

1. 四格表的效应大小估算:

使用 `ES.w1(P0, P1)` 公式, 其中 `P0` 是零假设下的概率分布 (通常是每个单元格概率相等的情况, 例如对于 2x2 表, `P0 = [0.25, 0.25, 0.25, 0.25]`), `P1` 是备择假设下的预期概率分布。

例如, 如果预计实验组的有效率为 45%, 对照组的有效率为 30%, 则 `P1` 可以是 `[0.3, 0.2, 0.45, 0.05]` (按照行和列的顺序)。

```
chisq.test(Poll_seasons, p = rep(1/4, 4))
```

```
P0 = rep(0.25, 4)
```

```
P1 = Poll_seasons / sum(Poll_seasons)
```

```
library(pwr)
```

```
pwr.chisq.test(w = ES.w1(P0, P1), N = sum(Poll_seasons), df = 4-1)
```

2. R*C 表的效应大小估算:

对于更大的列联表 (`R` 行 `C` 列), 可以使用 `ES.w2(prob)` 公式, 其中 `prob` 是一个包含 `R*C` 个概率值的矩阵或向量。

结合 `ljz` 的笔记看:

`chisq.test()` 函数中, `correct` 参数是用于控制是否应用连续性校正 (continuity correction) 的。连续性校正是一个小的调整, 旨在改进在 小样本或稀疏表格 时卡方检验的近似性。

具体来说, 当样本量较小, 特别是当预期的频数低于 5 时, 卡方分布的近似性可能不那么好。为了改进这种近似性, 统计学家开发了连续性校正。这种校正通过从每个观察频数中减去 0.5 (或相应的分数, 如果观察频数不是整数), 并相应地调整期望频数来实现。

在 `chisq.test()` 函数中, `correct` 参数的默认值是 `TRUE`, 这意味着默认情况下会应用连续性校正。然而, 在某些情况下, 特别是当样本量较大或预期的频数较高时, 连续性校正可能不是必要的, 甚至可能导致检验的保守性 (即, 更可能拒绝原假设)。在这些情况下, 你可以将 `correct` 参数设置为 `FALSE` 来禁用连续性校正。

```
# 创建一个 2x2 的表格
```

```
my_table <- matrix(c(10, 20, 30, 40), nrow = 2)
```

```
dimnames(my_table) <- list(Group = c("A", "B"), Outcome = c("Yes", "No"))
```

```
# 使用连续性校正进行卡方检验
```

```
chi_test_with_correction <- chisq.test(my_table)
```

```
print(chi_test_with_correction)
```

```
# 不使用连续性校正进行卡方检验
```

```
chi_test_without_correction <- chisq.test(my_table, correct = FALSE)
```

```
print(chi_test_without_correction)
```

在这个示例中, 我们首先创建了一个 2x2 的表格, 并使用 `chisq.test()` 函数对其进行了两次卡方检验, 一次使用了连续性校正 (默认), 另一次禁用了连续性校正 (通过设置 `correct = FALSE`)。然后, 我们打印了两次检验的结果, 以便比较它们之间的差异。

`ljz` 给出的例子:

Q: 现在有四个品质 level, 分别为 best, good, median, bad, 对应的商品个数为 10,15,20,5。然后现在由于品质变量是 categorical variable 嘛, 所以我想验证品质 variable 对商品个数 variable 是否有影响, 因为品质 variable 是 ordinal 的, 所以应该用属于 one-sample wilcoxon test 是吗? 所以应该如何用 R 实现呢?

A: 我觉得下面文心一言的回答不一定对。因为这个情况实际上我感觉等价于 chi-square test for goodness-of-fit, 但是这个 p-value 一个拒绝一个不拒绝。但是现在 sample size 这么小可能 chi-square test 也不合适。不知道现在是一个什么情况。

对于您所描述的情况, 由于品质变量 (best, good, median, bad) 是有序的 (ordinal), 而您想验证这个有序变量是否对商品个数 (一个数值变量) 有影响, 您实际上不应该使用单样本 Wilcoxon 秩和检验 (one-sample Wilcoxon test)。单样本 Wilcoxon 秩和检验是用于检验一个样本的数据分布是否显著不同于某个指定的理论分布 (例如中位数是否为某个特定值), 而不是用于检验两个或多个独立样本 (特别是当这些样本是有序的) 之间是否有差异。

对于您的情况, 您应该使用非参数检验中的秩和检验 (Mann-Whitney U 检验的推广, 即 Kruskal-Wallis 秩和检验), 因为您有多个独立的组 (即不同的品质级别)。Kruskal-Wallis 检验能够检验来自多个独立总体的样本是否具有相同的分布, 而不需要这些分布是正态的。

在 R 中, 您可以使用 `kruskal.test()` 函数来进行 Kruskal-Wallis 检验。以下是如何在 R 中实现它的示例:

r 复制代码

```
# 假设品质级别和对应的商品个数
quality_levels <- factor(c("best", "good", "median", "bad"), ordered = TRUE, levels = c("best", "good", "median", "bad"))
item_counts <- c(10, 15, 20, 5)

# 进行 Kruskal-Wallis 秩和检验
kw_test <- kruskal.test(item_counts ~ quality_levels)

# 打印结果
print(kw_test)
```

在这个例子中, `quality_levels` 是一个有序因子 (ordered factor), 它表示不同的品质级别。`item_counts` 是一个数值向量, 表示每个品质级别对应的商品个数。`kruskal.test()` 函数将检验这些商品个数的分布是否因品质级别的不同而有所不同。

如果 Kruskal-Wallis 检验的结果显示 p 值小于您选择的显著性水平 (例如 0.05), 那么您将拒绝原假设 (即所有品质级别的商品个数分布相同), 并认为至少有一个品质级别的商品个数与其他品质级别有显著差异。如果 p 值大于显著性水平, 则您不能拒绝原假设, 即认为所有品质级别的商品个数分布相同。

Tips: c jy 和我都觉得这个是对的, 因为没有四组数据, 写不出来我之前给的 Kruskal-Wallis 的示例

A2: by c jy

```
c=rep(c('bad','median','good','best'),
      c(5,20,15,10)) %>%
  factor(levels=c('bad','median','good','best'))
table(c)
as.numeric(c)
wilcox.test(as.numeric(c),mu=3,alternative='less')
```

测的是是否平均品质都低于 good

样品个数分布是否 Independent of 品质, 只能用 chi-square test for goodness-of-fit

所以 wilcoxon test 是检测一组的 median 是否等于给定的 mu, 或者两组 median 是否相等
而 Kruskal-Wallis test 相当于能比较三组及以上的组的 median 是否相等, 相当于就是
non-parametric alternative to 1-way anova 了, 相当于就是只有 one factor (即来自不同的
组别)

当检查 goodness-of-fit 的时候

1.

```
chisq.test(Poll_seasons)
```

```
chisq.test(Poll_seasons, simulate.p.value = T)
```

文心一言说当样本量较小的时候, 比如说这个时候只有 16, 小于 40

可以使用 chisq.test 函数的 simulate.p.value 参数来进行基于模拟的 p 值计算

2.

wangchaochen 课件里面也是这么说的

The Fisher's exact test can also be used on contingency tables larger than 2x2

我去网上查了一下对于大于 2*2 的列联表, 就是那个矩阵形式表达的公式极其复杂看不懂

```
data5 <- matrix(c(5,20,15,10,12.5,12.5,12.5,12.5), nrow = 2, byrow = T)
```

```
fisher.test(data5)
```

但是估计我这个方法应该是有道理的 (cjy 虽然说我是对的, 但是我估计他不清楚原理)