卡方检验与 Fisher 精确检验：拟合优度检验、独立性检验、同质性检验和 Fisher 精确检验

目录

1 卡方拟合优度检验

1.1 目的

卡方拟合优度检验用于判断观测的分类数据是否符合某个特定的理论分布或预期比例，评估观测频数与期望频数之间的差异是否具有统计学显著性。

1.2 应用场景

- 检验骰子或硬币是否公平。
- 验证样本中的分类频数是否符合已知的总体分布。
- 检查市场份额是否与预期比例一致。

1.3 数据要求

- 单个分类变量的观测频数。
- 理论分布或预期比例已知。

1.4 假设

- 观测值相互独立。
- 每个类别的期望频数不少于 5。

1.5 HATPC 步骤

1.5.1 假设（Hypotheses）

- 原假设（H_0）：观测数据与理论分布无显著差异，符合预期分布。
- 备择假设（H_1）：观测数据与理论分布存在显著差异。

1.5.2 假设条件（Assumptions）

- 观测值是独立的随机样本。
- 期望频数 $E_i \geq 5$。

1.5.3 检验统计量（Test Statistic）

卡方统计量计算公式为：

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

其中：

- O_i：第 i 类别的观测频数。
- E_i：第 i 类别的期望频数，$E_i = N \times P_i$。
- N：总样本量。
- P_i：第 i 类别的理论概率或预期比例。
- k：类别总数。

1.5.4 p 值计算（P-value）

- 自由度：$df = k - 1 - m$，其中 m 为估计的参数数。
- 方法一：使用卡方分布查表或计算机函数直接获得 p 值。
- 方法二：使用检验统计量和自由度，通过累积分布函数计算：

$$p\text{-value} = 1 - \text{pchisq}(\chi^2, df)$$

1.5.5 结论（Conclusion）

- 若 $p \leq \alpha$（显著性水平），拒绝原假设，认为观测数据与理论分布存在显著差异。
- 若 $p > \alpha$，不拒绝原假设，认为观测数据符合理论分布。

1.6 R 代码示例

Listing 1: 卡方拟合优度检验的 R 代码示例

```r
# 观测频数
observed <- c(15, 22, 18, 20, 25, 20)

# 理论概率（均匀分布）
expected_prob <- rep(1/6, 6)

# 卡方拟合优度检验，方法一
test_result <- chisq.test(x = observed, p = expected_prob)

# 显示结果
print(test_result)

# 方法二：手动计算检验统计量和 p 值
chisq_stat <- sum((observed - expected_prob * sum(observed))^2 / (expected_prob *
    sum(observed)))
df <- length(observed) - 1
p_value <- 1 - pchisq(chisq_stat, df, lower.tail = FALSE)

# 显示手动计算的结果
cat("Chi-squared statistic:", chisq_stat, "\n")
cat("Degrees of freedom:", df, "\n")
cat("p-value:", p_value, "\n")
```

1.7 说明

上述两种方法都会得到相同的检验统计量和 p 值。第一种方法使用 `chisq.test()` 函数直接计算，第二种方法手动计算检验统计量，然后使用 `pchisq()` 计算 p 值。

1.8 示例

假设一位研究者想检验一颗骰子是否公平，进行了 120 次投掷，结果如下：

点数	1	2	3	4	5	6
观测频数 O_i	15	22	18	20	25	20

- 计算期望频数：$E_i = 120 \times \frac{1}{6} = 20$
- 计算卡方统计量：

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - 20)^2}{20} = \frac{(15-20)^2}{20} + \frac{(22-20)^2}{20} + \cdots + \frac{(20-20)^2}{20} = 4.5$$

- 自由度：$df = 6 - 1 = 5$

- 计算 p 值：

$$p\text{-value} = 1 - \text{pchisq}(4.5, 5, lower.tail = FALSE) \approx 0.4795$$

- 结论：由于 $p > 0.05$，不拒绝原假设，认为骰子可能是公平的。

2 卡方独立性检验

2.1 目的

卡方独立性检验用于评估两个分类变量之间是否存在统计学上的关联，即判断它们是否相互独立。

2.2 应用场景

- 分析性别与购买偏好之间的关系。
- 评估治疗方法与康复情况之间的关联。
- 研究教育水平与就业状况的关系。

2.3 数据要求

- 同一随机样本中两个分类变量的观测频数。
- 构建二维列联表。

2.4 假设

- 观测值相互独立。
- 期望频数 $E_{ij} \geq 5$。

2.5 HATPC 步骤

2.5.1 假设（Hypotheses）

- 原假设（H_0）：两个分类变量相互独立，没有关联。
- 备择假设（H_1）：两个分类变量不独立，存在关联。

2.5.2 假设条件（Assumptions）

- 观测值是独立的随机样本。
- 期望频数 $E_{ij} \geq 5$。

2.5.3 检验统计量（Test Statistic）

卡方统计量计算公式为：

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

其中：

- O_{ij}：第 i 行，第 j 列单元格的观测频数。

- E_{ij}：第 i 行，第 j 列单元格的期望频数，计算公式为：

$$E_{ij} = \frac{(\text{行总计}_i \times \text{列总计}_j)}{\text{总样本量}}$$

- r：行数（变量 A 的类别数）。
- c：列数（变量 B 的类别数）。

2.5.4 p 值计算（P-value）

- 自由度：$df = (r-1) \times (c-1)$
- 方法一：使用卡方分布查表或计算机函数直接获得 p 值。
- 方法二：使用检验统计量和自由度，通过累积分布函数计算：

$$p\text{-value} = 1 - \text{pchisq}(\chi^2, df, lower.tail = FALSE)$$

2.5.5 结论（Conclusion）

- 若 $p \leq \alpha$，拒绝原假设，认为两个变量之间存在显著关联。
- 若 $p > \alpha$，不拒绝原假设，认为两个变量相互独立。

2.6 R 代码示例

Listing 2: 卡方独立性检验的 R 代码示例

```r
# 创建列联表
data <- matrix(c(70, 50, 40, 40), nrow = 2, byrow = TRUE)
dimnames(data) <- list(
  Gender = c("Male", "Female"),
  Training = c("Yes", "No")
)

# 卡方独立性检验，方法一
test_result <- chisq.test(data)

# 显示结果
print(test_result)

# 方法二：手动计算检验统计量和 p 值
chisq_stat <- sum((data - chisq.test(data)$expected)^2 / chisq.test(data)$expected)
df <- (nrow(data) - 1) * (ncol(data) - 1)
p_value <- 1 - pchisq(chisq_stat, df, lower.tail = FALSE)

# 显示手动计算的结果
cat("Chi-squared statistic:", chisq_stat, "\n")
cat("Degrees of freedom:", df, "\n")
cat("p-value:", p_value, "\n")
```

2.7 说明

两种方法的检验统计量和 p 值相同。第一种方法使用 `chisq.test()` 函数直接计算，第二种方法手动计算检验统计量，然后使用 `pchisq()` 计算 p 值。

2.8 示例

某公司调查了 200 名员工的性别和是否接受过培训，结果如下：

	接受培训	未接受培训	行总计
男	70	50	120
女	40	40	80
列总计	110	90	200

- 计算期望频数：

$$E_{\text{男，接受培训}} = \frac{120 \times 110}{200} = 66$$

其他单元格同理计算。

- 计算卡方统计量：

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(70-66)^2}{66} + \frac{(50-54)^2}{54} + \frac{(40-44)^2}{44} + \frac{(40-36)^2}{36} \approx 1.212$$

- 自由度：$df = (2-1) \times (2-1) = 1$
- 计算 p 值：

$$p\text{-value} = 1 - \text{pchisq}(1.212, 1, lower.tail = FALSE) \approx 0.271$$

- 结论：由于 $p > 0.05$，不拒绝原假设，认为性别与是否接受培训之间没有显著关联。

3 卡方同质性检验

3.1 目的

卡方同质性检验用于比较不同群体在同一分类变量上的分布是否一致，评估群体间的同质性。

3.2 应用场景

- 比较不同地区消费者对某产品的偏好分布是否一致。
- 评估不同教学方法下学生成绩分布是否相同。
- 分析不同年龄段人群对某事件的反应是否一致。

3.3 数据要求

- 多个独立样本，每个样本对应一个群体。
- 同一分类变量的观测频数。

3.4 假设

- 各群体的样本是独立的随机样本。
- 观测值相互独立。
- 期望频数 $E_{ij} \geq 5$。

3.5 HATPC 步骤

3.5.1 假设（Hypotheses）

- 原假设（H_0）：不同群体在分类变量上的分布相同，具有同质性。
- 备择假设（H_1）：至少有一个群体的分布与其他群体不同。

3.5.2 假设条件（Assumptions）

- 各群体的样本是独立的随机样本。
- 观测值相互独立。
- 期望频数 $E_{ij} \geq 5$。

3.5.3 检验统计量（Test Statistic）

卡方统计量计算公式为：

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

其中：

- O_{ij}：第 i 群体，第 j 类别的观测频数。

- E_{ij}：期望频数，计算公式为：

$$E_{ij} = \frac{(\text{行总计}_i \times \text{列总计}_j)}{\text{总样本量}}$$

- r：群体数（行数）。
- c：分类变量的类别数（列数）。

3.5.4 p 值计算（P-value）

- 自由度：$df = (r-1) \times (c-1)$
- 方法一：使用卡方分布查表或计算机函数直接获得 p 值。
- 方法二：使用检验统计量和自由度，通过累积分布函数计算：

$$p\text{-value} = 1 - \text{pchisq}(\chi^2, df)$$

3.5.5 结论（Conclusion）

- 若 $p \leq \alpha$，拒绝原假设，认为群体间分布存在显著差异。
- 若 $p > \alpha$，不拒绝原假设，认为群体在分类变量上的分布一致。

3.6 R 代码示例

Listing 3: 卡方同质性检验的 R 代码示例

```r
# 创建数据框
data <- matrix(c(50, 30, 20,
                 60, 25, 15,
                 40, 35, 25),
               nrow = 3, byrow = TRUE)
dimnames(data) <- list(
  Region = c("A", "B", "C"),
  Satisfaction = c("Satisfied", "Neutral", "Dissatisfied")
)

# 卡方同质性检验，方法一
test_result <- chisq.test(data)

# 显示结果
print(test_result)

# 方法二：手动计算检验统计量和 p 值
chisq_stat <- sum((data - chisq.test(data)$expected)^2 / chisq.test(data)$expected)
df <- (nrow(data) - 1) * (ncol(data) - 1)
p_value <- 1 - pchisq(chisq_stat, df, lower.tail = FALSE)

# 显示手动计算的结果
cat("Chi-squared statistic:", chisq_stat, "\n")
```

```r
cat("Degrees of freedom:", df, "\n")
cat("p-value:", p_value, "\n")
```

3.7 说明

上述两种方法得到的检验统计量和 p 值一致，第一种方法使用 `chisq.test()` 函数直接计算，第二种方法手动计算检验统计量并使用 `pchisq()` 计算 p 值。

3.8 示例

研究人员想比较三个地区（A、B、C）消费者对某产品的满意度分布是否一致，数据如下：

地区	满意	中立	不满意	行总计
A	50	30	20	100
B	60	25	15	100
C	40	35	25	100
列总计	150	90	60	300

- 计算期望频数：

$$E_{\text{A，满意}} = \frac{100 \times 150}{300} = 50$$

其他单元格同理计算。

- 计算卡方统计量：

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(50-50)^2}{50} + \frac{(30-30)^2}{30} + \cdots = 10$$

- 自由度：$df = (3-1) \times (3-1) = 4$
- 计算 p 值：

$$p\text{-value} = 1 - \text{pchisq}(10, 4, lower.tail = FALSE) \approx 0.0404$$

- 结论：由于 $p \leq 0.05$，拒绝原假设，认为三个地区的满意度分布存在显著差异。

4 Fisher 精确检验

4.1 目的

Fisher 精确检验用于评估两个分类变量之间的关联性，特别适用于样本量较小或期望频数较低的 2x2 列联表。

4.2 应用场景

- 当数据中的某些期望频数小于 5，无法满足卡方检验的最低频数要求时。
- 适用于小样本的 2x2 列联表分析。
- 分析罕见事件的发生是否与某因素有关联。

4.3 数据要求

- 2x2 列联表的数据。
- 观测值相互独立。

4.4 假设

- 原假设（H_0）：两个分类变量相互独立，没有关联。
- 备择假设（H_1）：两个分类变量不独立，存在关联。

4.5 检验统计量

Fisher 精确检验基于超几何分布，计算精确的 p 值。对于如下的 2x2 列联表：

	特征存在	特征不存在
组别 1	a	b
组别 2	c	d

p 值的计算公式为：

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

其中 $n = a + b + c + d$。

4.6 结论

- 若 $p \leq \alpha$，拒绝原假设，认为两个变量之间存在显著关联。
- 若 $p > \alpha$，不拒绝原假设，认为两个变量相互独立。

4.7 R 代码示例

Listing 4: Fisher 精确检验的 R 代码示例

```r
# 创建2x2列联表
data <- matrix(c(8, 2, 1, 5), nrow = 2)
dimnames(data) <- list(
  Treatment = c("New Drug", "Placebo"),
  Outcome = c("Recovered", "Not Recovered")
)

# Fisher 精确检验
test_result <- fisher.test(data)

# 显示结果
print(test_result)
```

4.8 示例

某研究调查了 16 名患者，对比新药与安慰剂的疗效，结果如下：

	痊愈	未痊愈
新药组	8	2
安慰剂组	1	5

- 由于样本量较小，且某些期望频数小于 5，采用 Fisher 精确检验。
- 使用 R 计算：

```r
data <- matrix(c(8, 2, 1, 5), nrow = 2)
fisher.test(data)
```

- 得到 p 值 $p \approx 0.0347$。
- 结论：由于 $p \leq 0.05$，拒绝原假设，认为新药的疗效与安慰剂存在显著差异。