统计学与 R 语言

第 17 讲　t 检验与 Wilcoxon 检验

张敬信

2022 年 4 月 25 日

哈尔滨商业大学

一. 假设检验原理

实际中，只能得到所抽取样本（部分）的统计结果，要进一步推断总体（全部）的特征，但是这种推断必然有可能犯错，犯错的概率为多少时应该接受这种推断呢?

为此，统计学家基于**小概率反证法思想**开发了**假设检验**这一统计方法进行统计检验。

假设检验的基本逻辑是：**如果原假设是真的，则检验统计量（样本数据的函数）将服从某概率分布。**

具体来说，

- 先提出原假设（也称为零假设），接着在原假设为真的前提下，基于样本数据计算出检验统计量值，与统计学家建立的这些统计量应服从的概率分布进行对比，就可以知道在百分之多少（P 值[1]）的机遇下会得到目前的结果。
- 若经比较后发现，出现该结果的概率（P 值）很小，就是说是基本不会发生的小概率事件；则可以有把握地说：这不是巧合，拒绝原假设是具有统计学上的意义的；否则就是不能拒绝原假设。

[1]假设检验的 P 值，是在 H_0 为真时根据检验统计量服从的理论概率分布计算的，衡量的是在原假设 H_0 下出现当前观测结果可能性的大小。

原假设与备择假设：

- 原假设（H_0）：研究者想收集证据予以反对的假设；
- 备择假设（H_1）：研究者想收集证据予以支持的假设；

假设检验判断方法有：P 值法和临界值法。

以 t 检验为例，

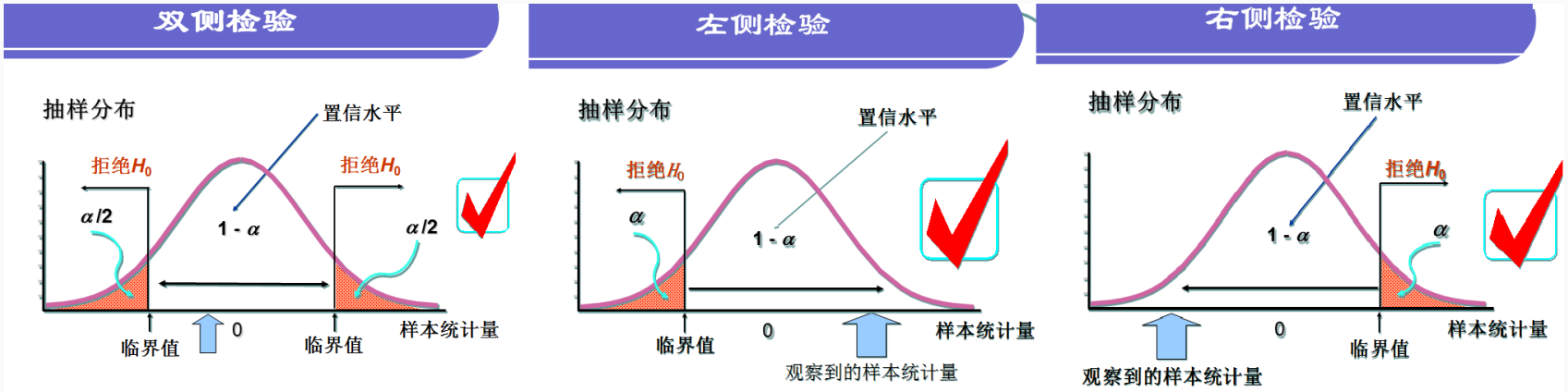


图 1: 双侧、左侧、右侧假设检验原理示意图

双侧检验：$H_0:\ \mu = \mu_0, \mu \neq \mu_0$

- 在原假设 H_0 下，根据样本数据计算出 t 统计量值 t_0
- P值 $= P\{|t| \geq t_0\}$, 表示 t_0 的双侧尾部的面积
- 若 $P < 0.05$ (在双尾部分)，则在 0.05 显著水平下拒绝原假设 H_0.

临界值法，是以显著水平处的统计量值为界限，中间白色区域是接受域，两侧阴影部分是拒绝域，看统计量值 t_0 是落在哪部分而下结论。

左侧检验: $H_0:\ \mu \geq \mu_0,\ H_1:\ \mu < \mu_0$

- 在原假设 H_0 下根据样本数据计算出 t 统计量值 t_0
- P值 $= P\{t \leq t_0\}$, 表示 t_0 的左侧尾部的面积
- 若 $P < 0.05$ (在左尾部分)，则在 0.05 显著水平下拒绝原假设 H_0.

右侧检验: $H_0:\ \mu \leq \mu_0,\ H_1:\ \mu > \mu_0$.

- 在原假设 H_0 下，根据样本数据计算出 t 统计量值 t_0
- P值 $= P\{t \geq t_0\}$, 表示 t_0 的右侧尾部的面积
- 若 $P < 0.05$ (在右尾部分)，则在 0.05 显著水平下拒绝原假设 H_0.

假设检验的两类错误

I **型错误**：在原假设 H_0 为真时，仍然有可能得到检验统计量的 P 值很小，因此拒绝了 H_0 就犯了 I 型错误，用 α 表示（一般设为 0.05)。显然，犯 I 型错误的概率等于显著水平[2]，若要减小它，只需要减小显著水平，比如 0.01.

II **型错误**：在备择假设为真时，但由于种种原因（抽样运气不好、样本量不够等）并没有拒绝原假设，这就犯了 II 型错误，用 β 表示（一般设为 0.2)。

[2]假设检验的显著水平可理解为：若原假设为真，拒绝原假设的概率.

假设检验的功效

在备择假设为真时，拒绝原假设的概率，称为假设检验的功效（Power, 等于 $1-\beta$），它反映了你的研究结果的把握度。

备择假设为真，拒绝原假设的概率应该是 100%, 故该功效越大越好，通常要求不低于 80%.

提高假设检验功效的一种可行办法是，增大样本量。一旦设定了显著水平（如 0.05）和功效（如 0.8），根据检验统计量就可以科学地计算样本量。

pwr 包可以计算常用统计检验的功效或要达到某功效需要的样本量。

以右侧 t 检验为例：

```
library(pwr)
# 每组样本量 50, Cohen 效应量 0.5, 显著水平 0.05, 计算功效
pwr.t.test(n = 50, d = 0.5, sig.level = 0.05,
           alternative = "greater")
# Cohen 效应量 0.5, 显著水平 0.05, 功效 0.8, 计算每组样本量
pwr.t.test(power = 0.8, d = 0.5, sig.level = 0.05,
           alternative = "greater")
```

注：若不用研究就知道差异应该很大，Cohen 效应量应设大一些，比如 0.8。

二. 基于理论的假设检验

基于理论的假设检验，可分为两类：

- 参数检验：要求样本来自的总体分布已知，对总体参数进行估计；优点是对数据信息充分利用，统计分析效率高；缺点是对数据要求高、适用范围有限。
- 非参数检验：不依赖数据的总体分布，也不对总体参数进行推断；优点是不受总体分布限制，适用范围广，对数据要求不高；缺点是检验功效相对较低，不能充分利用数据信息。

选择原则：首先考察是否满足参数检验的条件，若满足首选参数检验，若不满足只能采用非参数检验。

对于定量数据和定性数据适用的假设检验方法是不同的。

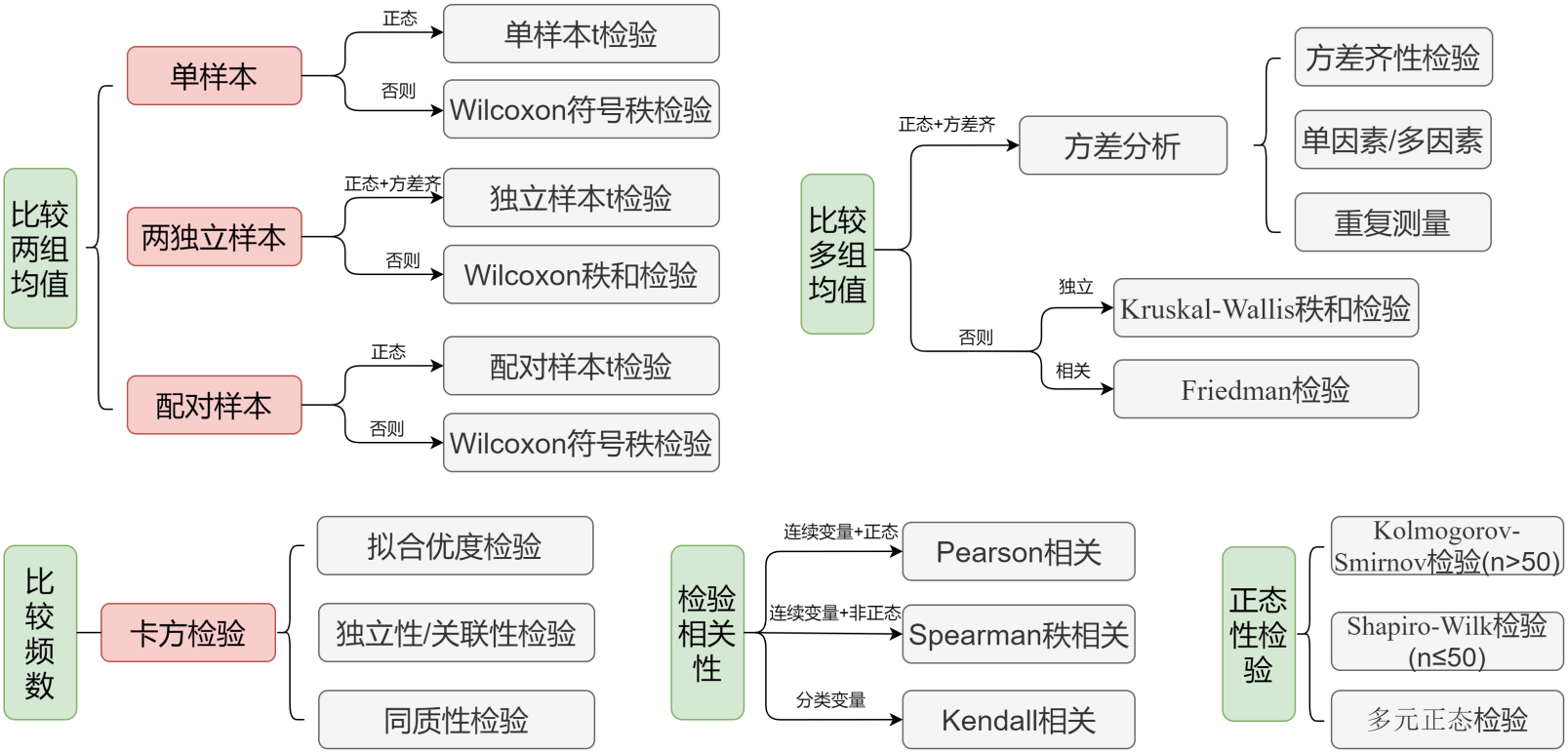


图 2: 常用的假设检验汇总

rstatix 包，提供了一个与「tidyverse」设计哲学一致的简单且直观的管道友好框架用于执行上述经典统计检验，支持结合 `group_by()` 做分组检验，且将检验结果转化为整洁的数据框输出。

- 比较均值：
  - `t_test()`: 单样本、两独立样本、配对 t 检验
  - `wilcox_test()`: 单样本、两独立样本、配对 Wilcoxon 检验
  - `sign_test()`: 单样本、两样本符号秩检验
  - `anova_test()`: 独立测量、重复测量、混合方差分析
  - `kruskal_test()`: Kruskal-Wallis 秩和检验
  - `friedman_test()`: Friedman 检验

- 比较比例
  - `prop_test()`: 单样本、两样本比例的 z 检验
  - `fisher_test()`: Fisher 精确检验，适用于单元格频数 <5
  - `chisq_test()`: 拟合优度、同质性、独立性卡方检验
  - `binom_test()/multinom_test()`: 精确二项/多项检验
  - `mcnemar_test()/cochran_qtest()`: McNemar 卡方检验，对比两对或多对比例
  - `prop_trend_test()`: 趋势卡方检验
- 正态性检验[3]：`shapiro_test()/mshapiro_test()`
- 方差齐性检验：`levene_test()`
- 相关性检验：`cor_test()`

[3] Kolmogorov-Smirnov 正态性检验可用 `ks.test(x, "pnorm", mean=mean(x), sd=sd(x))`.

使用软件做假设检验的简单步骤：

- 首先，要明确其原假设和备择假设是什么；
- 然后，调用相应函数得到检验结果；
- 最后，解读结果，根据 P 值得到结论：若 $P < 0.05$, 则拒绝原假设，否则不能拒绝原假设。

二. t 检验

t **检验**，是针对连续变量的参数检验，可用来检验 “单样本均值与已知均值（单样本 t 检验)、两独立样本均值（独立样本 t 检验)、配对设计资料的均值（配对样本 t 检验）” 是否存在差异。

t 检验适用于小样本量（比如 $n < 60$，大样本数据可以用 U 检验)，要求数据满足：正态性和方差齐性，若不满足可尝试变换数据，或用 Wilcoxon 符号秩/秩和检验。

1. 案例 1：检验电影爱情片与动作片评分差异

```
load("datas/movies.rda")
movies
#> # A tibble: 68 x 4
#>   title         year rating genre
#>   <chr>        <int>  <dbl> <chr>
#> 1 Underworld    1985    3.1 Action
#> 2 Love Affair   1932    6.3 Romance
#> 3 Junglee       1961    6.8 Romance
#> # ... with 65 more rows
```

```
movies %>%
  group_by(genre) %>%
  summarise(n = n(), avg_rat = mean(rating),
            sd_rat = sd(rating))
#> # A tibble: 2 x 4
#>   genre       n avg_rat sd_rat
#>   <chr>   <int>   <dbl>  <dbl>
#> 1 Action     32    5.28   1.36
#> 2 Romance    36    6.32   1.61
```

对于该样本，平均评分爱情片为 6.32，动作片为 5.28，二者之差为 1.04，这是真实差异的点估计。那么，**该差异能否用来推断总体（所有电影），还是只是随机抽样的偶然因素造成的?**

- 检验各组正态性[4]

```
library(rstatix)
movies %>%
  group_by(genre) %>%
  shapiro_test(rating)
#> # A tibble: 2 x 4
#>   genre   variable statistic     p
#>   <chr>   <chr>        <dbl> <dbl>
#> 1 Action  rating       0.958 0.246
#> 2 Romance rating       0.963 0.262
```

[4]大样本（$n > 50$）适合用 `ks.test()` 做 Kolmogorov-Smirnov (K-S) 检验，多元正态性检验用 `mshapiro_test()`.

- 检验各组方差齐性

```
movies %>%
  levene_test(rating ~ genre)
#> # A tibble: 1 x 4
#>     df1   df2 statistic     p
#>   <int> <int>     <dbl> <dbl>
#> 1     1    66     0.279 0.599
```

先构造原假设和备择假设：

$$H_0：\mu_r - \mu_a = 0 \qquad H_1：\mu_r - \mu_a \neq 0$$

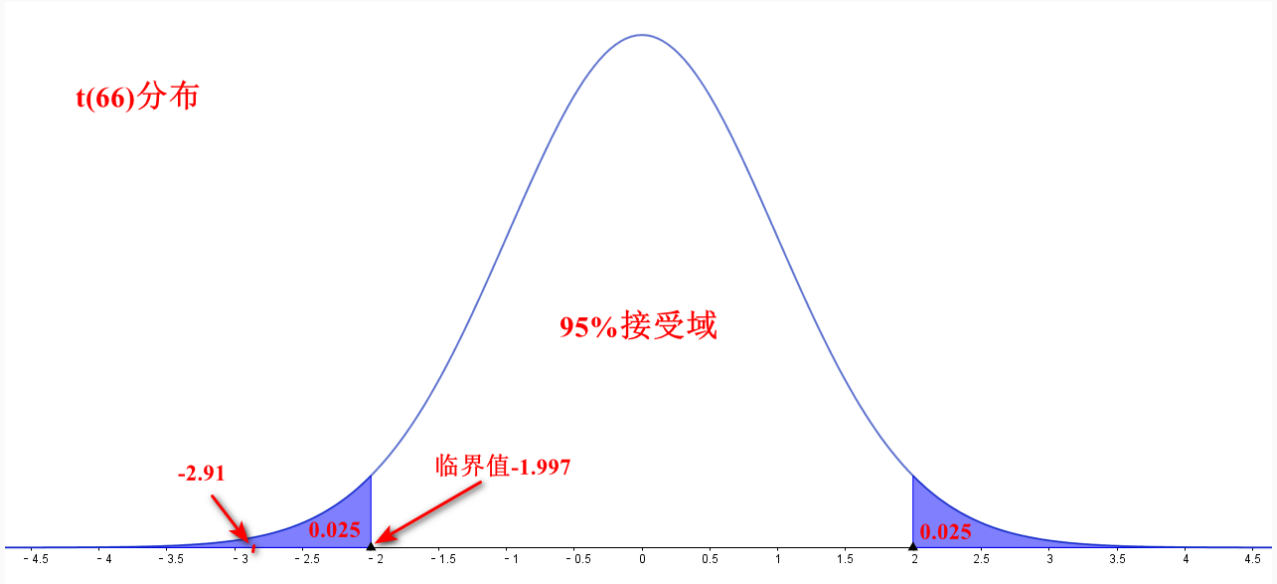
在正态性、方差齐性条件下，t 统计量为

$$t = \frac{\bar{x}_a - \bar{x}_r}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_r^2}{n_r}}}$$

服从 $t(n_a + n_r - 2)$ 分布。

(1) 手动计算

```
(t = (5.28 - 6.32) / sqrt(1.36^2 / 32 + 1.61^2 / 36)) # t 值
#> [1] -2.89
qt(0.05/2, 32 + 36 - 2)                    # 临界值
#> [1] -2
pt(t, 32 + 36 - 2) * 2                     # p 值（双侧）
#> [1] 0.00526
```



(2) 用现成函数

```
t_test(
  data,
  formula,
  comparisons = NULL,
  ref.group = NULL,
  p.adjust.method = "holm",
  paired = FALSE,
  var.equal = FALSE,
  alternative = "two.sided",
  mu = 0,
  conf.level = 0.95,
  detailed = FALSE
)
```

```
movies %>%
  t_test(rating ~ genre)
#> # A tibble: 1 x 8
#>   .y.    group1 group2     n1    n2 statistic    df
#> * <chr>  <chr>  <chr>   <int> <int>     <dbl> <dbl>   <dbl
#> 1 rating Action Romance    32    36     -2.91  65.8 0.0049
```

2. 单样本 t 检验

- 检验电影平均评分是否等于 6 分

```
movies %>%
  t_test(rating ~ 1, mu = 6)
#> # A tibble: 1 x 7
#>   .y.    group1 group2          n statistic    df     p
#> * <chr>  <chr>  <chr>       <int>     <dbl> <dbl> <dbl>
#> 1 rating 1      null model    68    -0.892    67 0.376
```

3. 配对样本 t 检验

配对设计实验的数据：

- 同一受试对象处理前后的数据；
- 同一受试对象两个部位的数据；
- 同一样品用两种方法/仪器检验的结果；
- 配对的两个受试对象分布接受进行两种处理后的数据。

配对数据之间有一定的相关性，配对样本 t 检验就考虑到了这种相关性，其基本原理是为每对数据求差值，若无差异则差值的总体均值为 0.

案例 2：某降压药物是否有降压作用

```
df = tibble(id = 1:10,
  before = c(120,127,141,107,110,114,115,138,127,122),
  after = c(123,108,120,107,100,98,102,152,104,107))
df
#> # A tibble: 10 x 3
#>       id before after
#>    <int>  <dbl> <dbl>
#> 1      1    120   123
#> 2      2    127   108
#> 3      3    141   120
#> # ... with 7 more rows
```

```r
df %>%
  pivot_longer(-id, names_to = "Type", values_to = "Hyper") %
  pairwise_t_test(Hyper ~ Type, paired = TRUE)
#> # A tibble: 1 x 10
#>   .y.   group1 group2    n1    n2 statistic    df      p p.
#> * <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl> <dbl> <d
#> 1 Hyper after  before    10    10     -2.65     9 0.027 0.
```

三. Wilcoxon 秩和检验

若两独立样本不满足正态性或方差齐性[5]，可以用 Wilcoxon 秩和检验。

Wilcoxon 秩和检验是先将两样本看成是单一样本（混合样本）然后由小到大排列观测值统一编秩。

- 若"H0：两个独立样本来自相同的总体" 为真，则小的、中等的、大的秩值大约均匀分布在两个样本中。
- 若"H1：两个独立样本来自不相同的总体" 为真，则其中一个样本有更多的小秩值，这样就会得到一个较小的秩和；另一个样本将会有更多的大秩值，会得到一个较大的秩和。

基于秩和构造的统计量服从 Mann-Whitney-Wilcoxon 分布。

Wilcoxon 秩和检验，也分针对单样本、两独立样本、配对样本。

[5]方差不齐也可以用 Welch 法的 t 检验，以两组各自标准差计算统计量近似服从 t 分布.

案例 3: 比较铅作业与非铅作业的血铅值

```
df = tibble(
  nonPb = c(24,26,29,34,43,58,63,72,87,101),
  Pb = c(82,87,97,121,164,208,213,NA,NA,NA))
df
#> # A tibble: 10 x 2
#>    nonPb    Pb
#>    <dbl> <dbl>
#> 1     24    82
#> 2     26    87
#> 3     29    97
#> # ... with 7 more rows
```

```r
df = df %>%
  pivot_longer(1:2, names_to = "Type",
               values_to = "val",
               values_drop_na = TRUE)
df %>%
  group_by(Type) %>%
  shapiro_test(val)
#> # A tibble: 2 x 4
#>   Type  variable statistic     p
#>   <chr> <chr>        <dbl> <dbl>
#> 1 nonPb val          0.920 0.358
#> 2 Pb    val          0.863 0.162
```

```
df %>%
  wilcox_test(val ~ Type, exact = TRUE,
              alternative = "less")
#> # A tibble: 1 x 7
#>   .y.   group1 group2    n1    n2 statistic       p
#> * <chr> <chr>  <chr>  <int> <int>     <dbl>   <dbl>
#> 1 val   nonPb  Pb        10     7       4.5 0.0017
```

四. 基于重排的假设检验

与 Bootstrap 法估计置信区间的区别是:

- 多了一步用 hypothesize() 设定原假设
- 重复生成数据的方法不是 Bootstrap 而是 permute

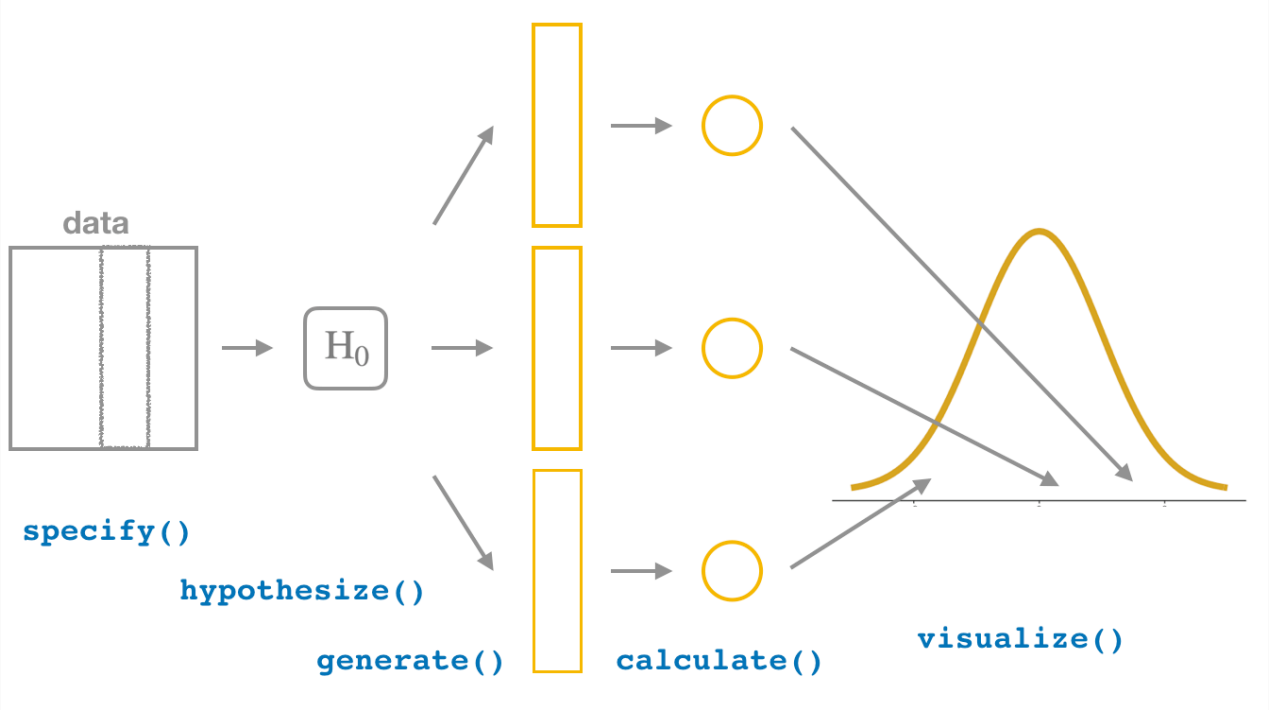


图 3: 用 infer 包实现重排假设检验的一般流程

仍以检验爱情片和动作片评分差异为例：

$$H_0: \mu_r - \mu_a = 0 \qquad H_1: \mu_r - \mu_a \neq 0$$

在原假设 H_0 下，即假设爱情片与动作片的平均评分没有差别，用重排法[6]生成 1000 个原样本的重抽样数据。因为假设爱情片与动作片的平均评分没有差别，那就将 genre 列随机重排（shuffled），让每个电影评分随机地对应这些爱情片或动作片。

[6]重排法是不重复抽样，原数据是 68 个样本，每个重抽样数据仍是不重复的 68 个样本.

然后，对每个重排样本分别计算检验统计量，这里是均值差 $\hat{\mu}_r - \hat{\mu}_a$. 这 1000 个统计量值就是 H_0（随机抽样的偶然因素）下，产生的均值差异的分布，也称为**零分布**。

那么，这 1000 个随机的统计量（均值差）中，有多少个会比点估计值 1.04 更大呢？其占比不就是假设检验的 P 值吗？即在 H_0 假设下，有多大的概率会出现当前观测结果。

若该 P 值小于置信水平 0.05，则表明由随机抽样的偶然因素造成这样大的均值差异 1.04，是很罕见的，因此有理由拒绝相应的原假设。

- 用参数 `null` 设定零假设，可选"point"（单样本）和"independence"（两样本）；用重排法生成 1000 个模拟样本；用参数 `stat` 指定要计算的检验统计量，参数 `order` 设定均值差是谁减谁：

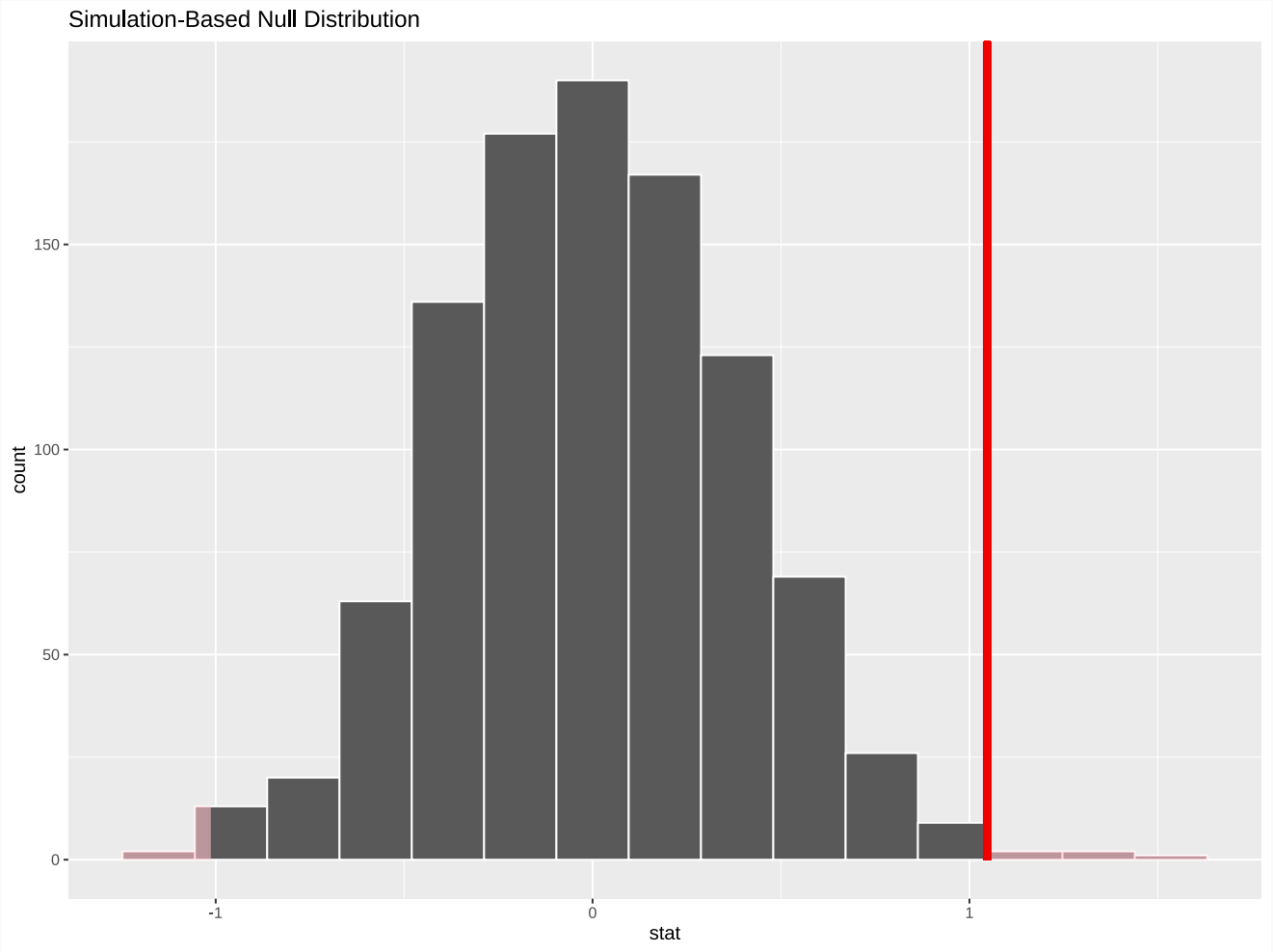
```
library(infer)
null_distribution = movies %>%
  specify(formula = rating ~ genre) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means",
            order = c("Romance", "Action"))
```

```
null_distribution
#> Response: rating (numeric)
#> Explanatory: genre (factor)
#> Null Hypothesis: independence
#> # A tibble: 1,000 x 2
#>   replicate    stat
#>       <int>   <dbl>
#> 1         1 -0.0212
#> 2         2  0.847
#> 3         3 -0.0330
#> # ... with 997 more rows
```

```r
null_distribution %>%   # 获取 P 值
  get_p_value(obs_stat = tibble(stat = 1.047),
              direction = "both")
#> # A tibble: 1 x 1
#>   p_value
#>     <dbl>
#> 1    0.01
```

- 可视化零分布数据，并标记点估计竖线及 P 值对应区域：

```r
visualize(null_distribution, bins = 15) +
  shade_p_value(obs_stat = tibble(stat = 1.047),
                direction = "both")
```


Simulation-Based Null Distribution
count
150
100
50
0
-1
0
1
stat

本篇主要参阅 (张敬信, 2022), (Chester Ismay, 2018), (Mine Çetinkaya Rundel, 2021), 以及包文档，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

Chester Ismay, A. Y. K. (2018). *Statistical Inference via Data Science A ModernDive into R and the Tidyverse*. CRC.

Mine Çetinkaya Rundel, J. H. (2021). *Introduction to Modern Statistics*. CRC, 1 edition.

张敬信 (2022). *R* 语言编程：基于 *tidyverse*. 人民邮电出版社, 北京.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

黄湘云 (2021). *Github: R-Markdown-Template*.