R 语言编程：基于 tidyverse

第 20 讲　假设检验

张敬信

2022 年 12 月 6 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，

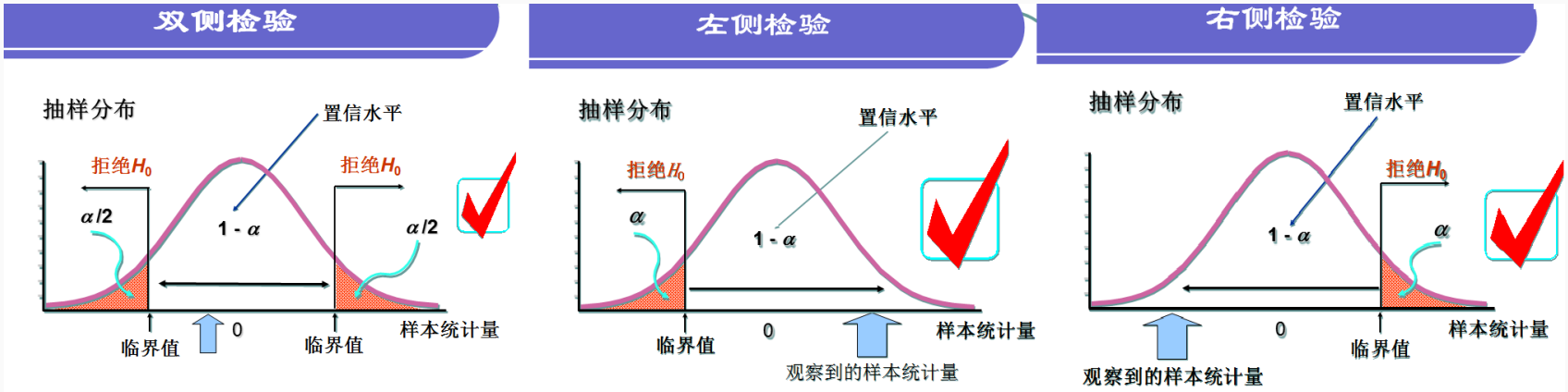


图 1：双侧、左侧、右侧假设检验原理示意图

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。

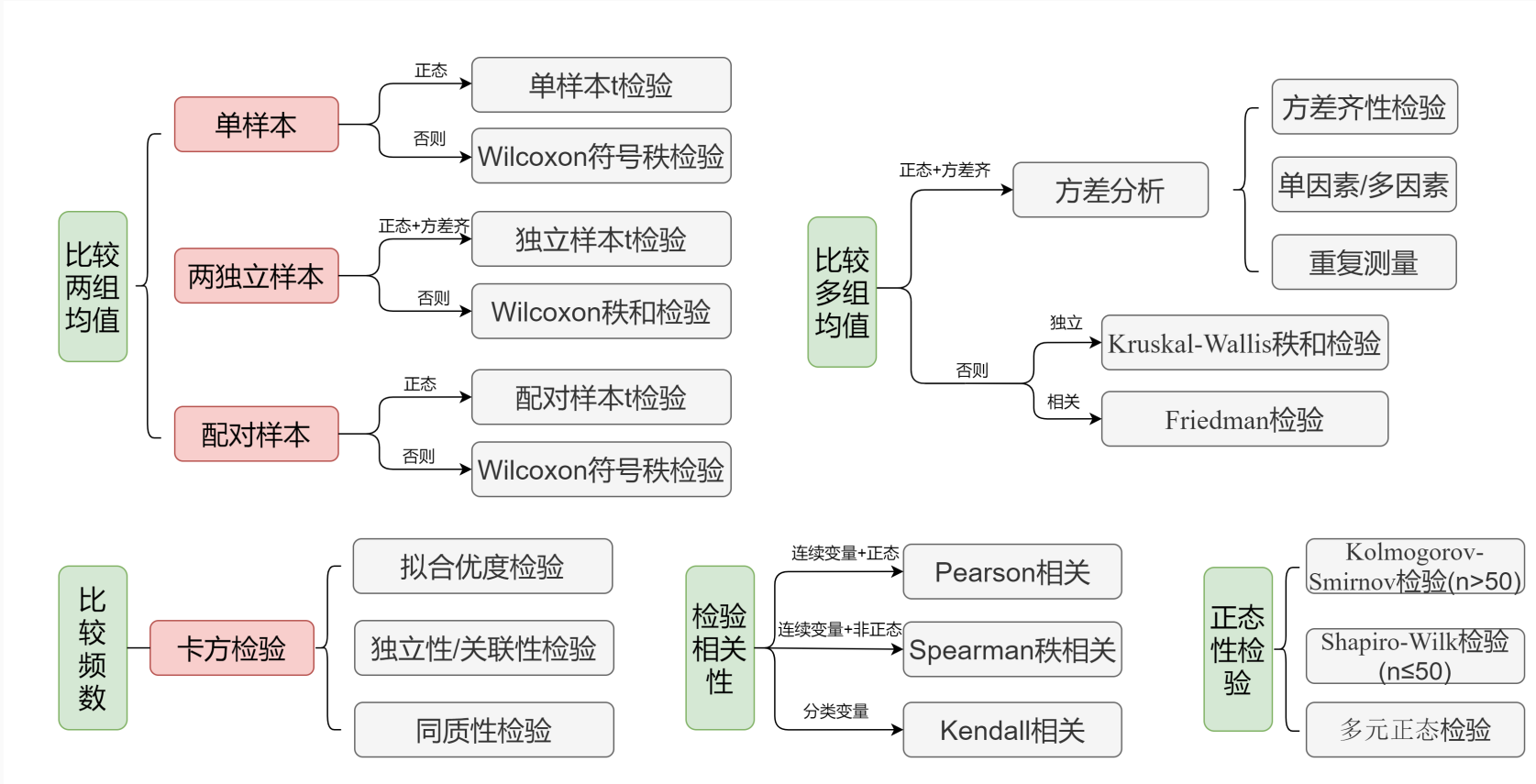


图 2：常用的假设检验汇总

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。

1. 方差分析

方差分析，是针对连续变量的参数检验，检验多个分组的均值有无差异，其中分组是按影响因素的不同水平值组合进行划分的。它是对总变异进行分解，看总变异是由哪些部分组成的，这些部分间的关系如何。

方差分析对数据的要求：满足**正态性**（来自同一正态总体）和**方差齐性**（各组方差相等），在这两个条件下，若各组有差异，则只可能是来自影响因素的不同水平。

方差分析可用于：

- 完全随机设计（单因素）、随机区组设计（双因素）、析因设计、拉丁方设计和正交设计等资料；
- 可对两因素间交互作用差异进行显著性检验；
- 进行方差齐性检验。

方差分析假定每一个观测值都由若干部分累加而成，即总的效应可分解为若干部分，每一部分都有特定含义，称为**效应的可加性**。根据效应的可加性，将总的离均差平方和分解成若干部分，每一部分都与某一种效应相对应，总自由度也被分成相应的各个部分，各部分的离均差平方除以各自的自由度得出各部分的均方（Mean Square)，两个均方之比服从 F 分布。

以焦虑症的治疗疗效为例，一个因素是治疗方案，有 2 种治疗方案，即该因素有 2 个水平；(治疗方案称为**组间因子**，因为每个患者只能被分配到一个组别中，没有患者同时接受两种治疗)；再考虑一个因素治疗时间，也有两个水平：治疗 5 周和治疗 6 个月，同一患者在 5 周和 6 个月不止一次地被测量（两次)，称为**重复测量**（治疗时间称为 * **组内因子**，因为每个患者在所有水平下都进行了测量)。

建立方差分析模型时，既要考虑两个因素治疗方案和治疗时间（主效应），又要考虑治疗方案和时间的交互影响（交互效应），称为**两因素混合模型方差分析。**

当某个因素的各个水平下的因变量的均值呈现统计显著性差异时，必要时可作两两水平间的比较，称为**均值间的两两比较。**

```
library(rstatix)
df = ToothGrowth %>%
  mutate(dose = factor(dose))
head(df, 4)
#>    len supp dose
#> 1  4.2   VC  0.5
#> 2 11.5   VC  0.5
#> 3  7.3   VC  0.5
#> 4  5.8   VC  0.5
```

牙齿长度（len）为因变量，关于喂食方法（supp）和剂量（dose）做两因素混合模型方差分析，其模型分解公式为：

$$\text{总差异 } Y_{ijk} = \text{平均差异 } \mu + \text{因素 1 差异 } \alpha_i + \text{因素 2 差异 } \beta_j + \text{因素 1,2 交互作用差异 } \gamma_{ij} + \text{随机差异 } \varepsilon_{ijk}$$

```
# 正态性检验 (H0: 正态)
shapiro_test(df, len)
#> # A tibble: 1 x 3
#>   variable statistic     p
#>   <chr>        <dbl> <dbl>
#> 1 len          0.967 0.109
```

```r
# 检验方差齐性 (H0: 方差齐)
levene_test(df, len ~ supp)
#> # A tibble: 1 x 4
#>     df1   df2 statistic     p
#>   <int> <int>     <dbl> <dbl>
#> 1     1    58      1.21 0.275
levene_test(df, len ~ dose)
#> # A tibble: 1 x 4
#>     df1   df2 statistic     p
#>   <int> <int>     <dbl> <dbl>
#> 1     2    57     0.646 0.528
```

```
# 两因素混合模型方差分析
anova_test(df, len ~ supp * dose)
#> ANOVA Table (type II tests)
#>
#>       Effect DFn DFd     F         p p<.05   ges
#> 1       supp   1  54 15.57 2.31e-04     * 0.224
#> 2       dose   2  54 92.00 4.05e-18     * 0.773
#> 3 supp:dose   2  54  4.11 2.20e-02     * 0.132
```

- len ~ supp * dose 是设定模型公式，遵从 R 的 formula 语法，~ 左边是因变量，右边是自变量公式，supp * dose 是 supp + dose + supp:dose 的简写，supp:dose 表示这两个变量的交互项。

方差分析结果的主效应 supp 和 dose 都非常显著（P 值都远小于 0.05），交互效应也显著（P 值 =0.022<0.05），表明 supp 和 dose 的协同变化下的各组均值显著不同[4]。

若要做 Tukey'HSD 组间的两两比较（多重比较）：

```
tukey_hsd(df, len ~ supp * dose)
#> # A tibble: 19 x 9
#>   term  group1 group2 null.value estimate conf.low conf
#> * <chr> <chr>  <chr>       <dbl>    <dbl>    <dbl>
#> 1 supp  OJ     VC              0    -3.70    -5.58
#> 2 dose  0.5    1               0     9.13     6.36
#> 3 dose  0.5    2               0    15.5     12.7
#> # ... with 16 more rows, and abbreviated variable name
```

[4]若交互作用不显著，可以只做去掉交互效应的方差分析.

重复测量方差分析

方差分析要求观测之间相互独立，而重复测量数据是在分组因素之外，分别在组内不同的时间点上重复测量同一个体获得因变量的观测值，或者是通过重复测量同一个体的不同部位获得因变量的观测值。这就不再具有相互独立性，需要专门方法来处理，称为**重复测量方差分析**。

重复测量数据，常用来分析因变量在不同时间点上的变化性。分析前需要对重复测量数据之间是否存在相关性进行球形检验，若 P 值 <0.05 则说明存在相关性，应该做重复测量方差分析。

重复测量方差分析的模型公式一般形式为：

$$Y \sim B * W + \mathrm{Error}(\mathrm{Subject/W})$$

其中，B 为组间因子，W 为组内因子，Subject 为个体标记。

```r
df %>%   # 相当于 10 只豚鼠，每只重复测量 6 次
  mutate(ID = rep(1:10, 6)) %>%
  anova_test(len ~ supp * dose + Error(ID / (supp * dose)))
#> ANOVA Table (type III tests)
#>
#> $ANOVA
#>       Effect DFn DFd      F        p p<.05   ges
#> 1       supp   1   9  34.87 2.28e-04     * 0.224
#> 2       dose   2  18 106.47 1.06e-10     * 0.773
#> 3 supp:dose   2  18   2.53 1.07e-01       0.132
#>
#> $`Mauchly's Test for Sphericity`
#>       Effect     W     p p<.05
#> 1       dose 0.807 0.425
#> 2 supp:dose 0.934 0.761
#>
#> $`Sphericity Corrections`
```

球形检验结果表明，重复测量数据存在相关性，两个主效应都很显著，交互效应不显著。

注：重复测量方差分析也要求满足方差齐性，若不满足可以考虑用 `lme4::lmer()` 拟合混合效应模型。

另外，`bruceR` 包整合了丰富的方差分析和结果的格式化文档输出。

卡方检验

卡方检验，是针对无序分类变量的非参数检验，其理论依据是：实际观察频数 f_0 与理论频数 f_e（又称期望频数）之差的平方再除以理论频数所得的统计量，近似服从 χ^2 分布。

卡方检验的一般是用来检验无序分类变量的实际观察频数和理论频数分布之间是否存在显著差异，要求：

- 分类变量相互排斥，互不包容；
- 观测相互独立；
- 样本容量不宜太小，理论频数 ≥ 5，否则需要进行校正（合并单元格或校正卡方值）

卡方检验常用于：

- **拟合优度检验**：检验某连续变量的数据是否服从某种分布，检验某分类变量各类的出现概率是否等于指定概率；
- **独立性/关联性检验**：检验两个分类变量是否相互独立；
- **同质性检验**：检验两组频数是否来自同一总体，若是，则每一类出现的概率应该是差不多的；检验两种方法的结果是否一致，例如两种方法对同一批人进行诊断，其结果是否一致。

以检验 `Titanic` 船舱等级与是否生存之间是否有相互独立为例：

$$H_0：\text{相互独立} \quad H_1：\text{不相互独立}$$

```r
titanic = read_rds("data/titanic.rds")
tbl = titanic %>%
  janitor::tabyl(Survived, Pclass)
tbl
#>  Survived   1  2   3
#>        No  80 97 372
#>       Yes 136 87 119
rstatix::chisq_test(titanic$Survived, titanic$Pclass)
#> # A tibble: 1 x 6
#>       n statistic         p    df method           p.sign
#> * <int>     <dbl>     <dbl> <int> <chr>            <chr>
#> 1   891      103. 4.55e-23     2 Chi-square test ****
```

P 值为 0，拒绝原假设，故结论是船舱等级与是否生存之间有关联。

若要进一步比较各等级的船舱之间生存率是否有差异:

```
pairwise_prop_test(as.matrix(tbl[,-1]))
#> # A tibble: 3 x 5
#>   group1 group2        p    p.adj p.adj.signif
#> * <chr>  <chr>     <dbl>    <dbl> <chr>
#> 1 1      2      2.32e- 3 2.32e- 3 **
#> 2 1      3      1.21e-22 3.64e-22 ****
#> 3 2      3      1.22e- 8 2.45e- 8 ****
```

三. 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

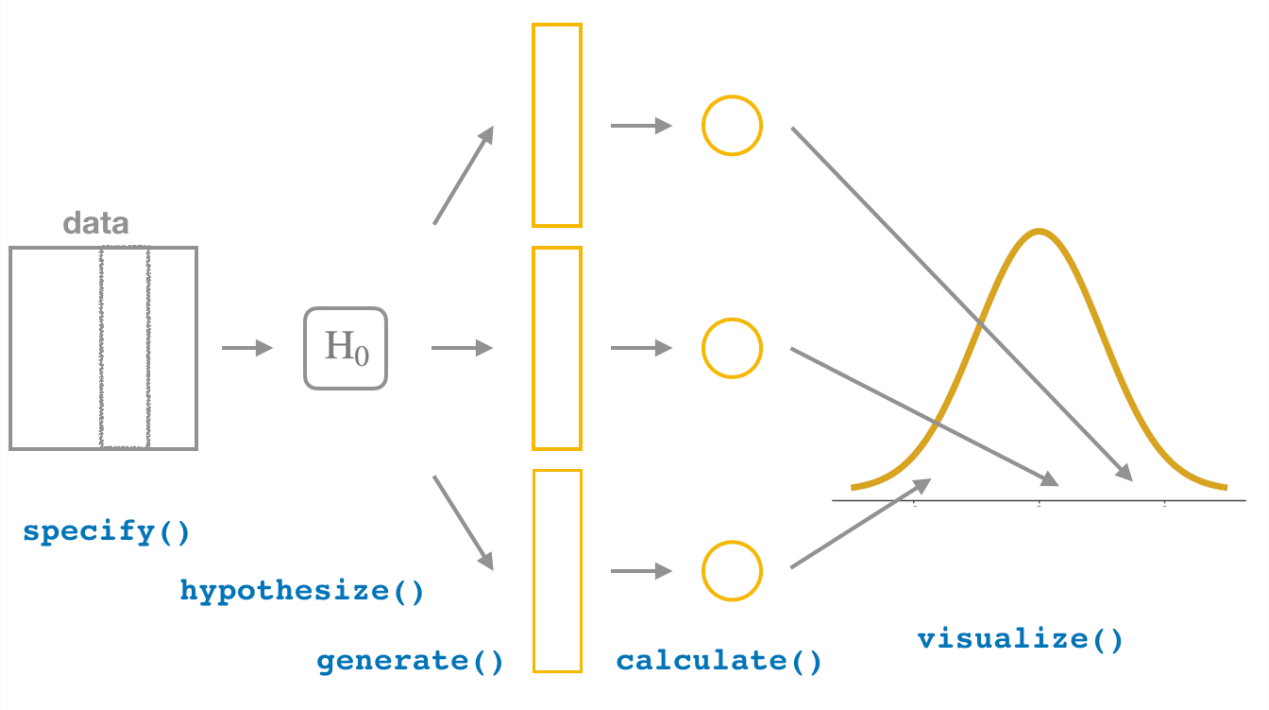


图 3: 用 infer 包实现重排假设检验的一般流程

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。

以检验电影爱情片与动作片评分差异为例：

```
load("data/movies_sample.rda")
movies_sample
#> # 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_sample %>%
  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?**

先构造假设检验：

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

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[5]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。

[5]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：

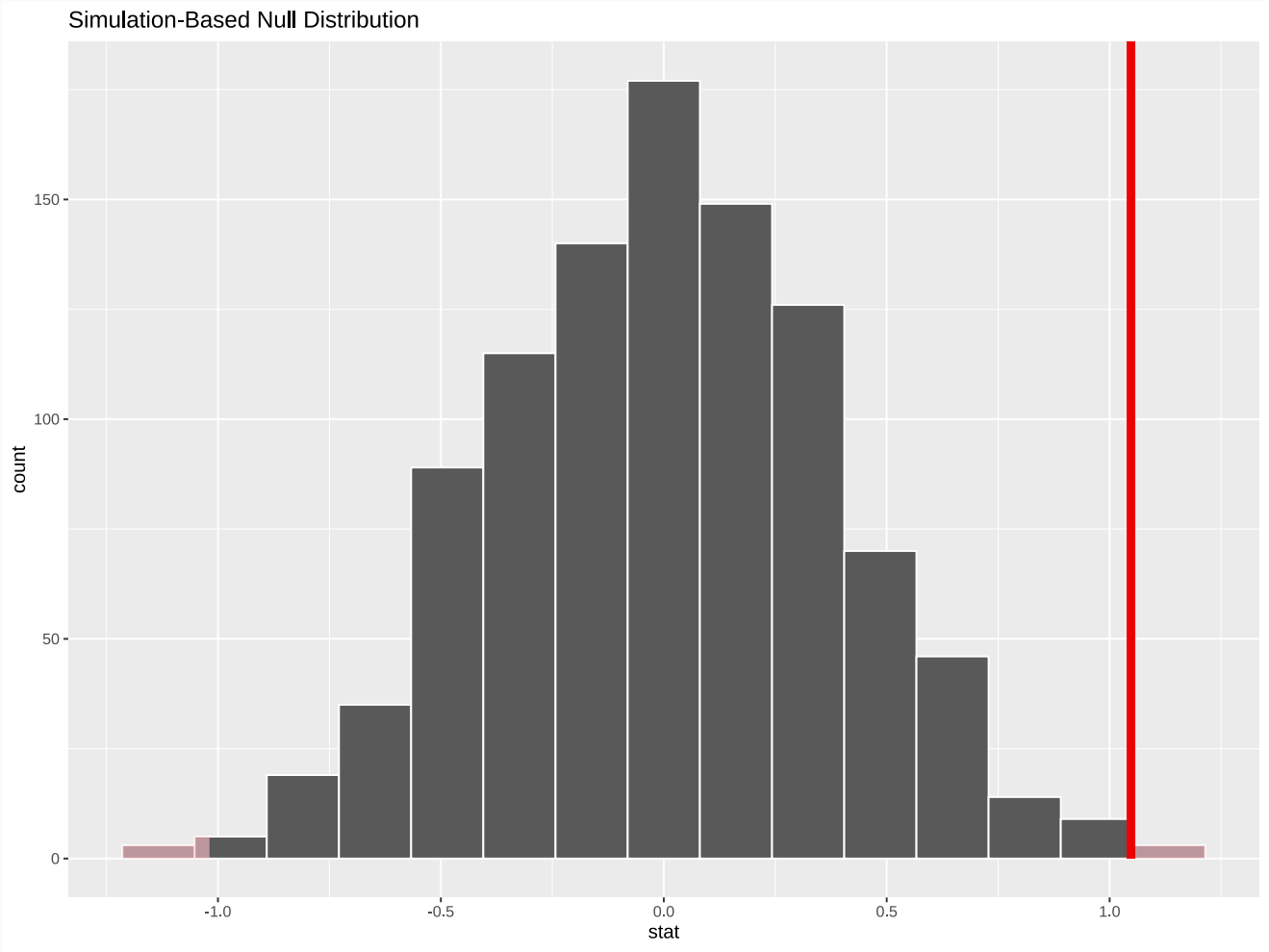
```
library(infer)
null_distribution = movies_sample %>%
  specify(formula = rating ~ genre) %>% # 响应变量 ~ 解释变
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means",
            order = c("Romance", "Action"))
```

```
null_distribution
#> Response: rating (numeric)
#> Explanatory: genre (factor)
#> Null Hypothesis: independence
#> # A tibble: 1,000 x 2
#>   replicate    stat
#>       <int>   <dbl>
#> 1         1 -0.488
#> 2         2 -0.0802
#> 3         3 -0.228
#> # ... 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   0.006
```

- 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")
```



注：更多案例可参阅 `infer` 包 `Vignettes`.

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*.