概 率 论 与 数 理 统 计

Probability and Statistics

南京大学 张绍群 & 徐科

更新： December 15, 2025

Part XI – Ch11: 假设检验

Ch11: 假设检验

Hypothesis Testing

December 15, 2025

提纲：假设检验

- 假设检验的基本流程
  - 建立假设
  - 选择检验统计量
  - 确定显著水平
  - 给出拒绝域
- 检验的两类错误
- 假设检验的应用案例
  - 假设检验与正态分布
  - 非参假设检验
  - 假设检验与机器学习
  - 假设检验与置信区间的关系
- 列联表的独立性检验

假设检验

根据样本信息来检验关于总体的某个假设 (猜测) 是否正确. 此类问题称为 **假设检验问题**, 可分为两类:

- 参数检验问题: 总体分布已知, 检验某未知参数的假设
- 非参数检验问题: 总体分布未知, 检验两个总体的差异

假设检验方法 (反证法):

- 先假设所做的假设 H_0 成立
- 然后从总体中取样, 根据样本来判断是否有 “不合理” 的现象出现
  - 可能会用到参数估计的方法
- 最后做出接受或者拒绝所做假设的决定. “不合理” 的现象是指小概率事件在一次事件中几乎不会发生

假设检验：例 0.1

例 0.1 某产品出厂检验规定次品率 $p \leq 0.04$ 才能出厂, 现从 10000 件产品中任抽取 12 件

- 若抽样结果有 3 件次品, 问该批产品是否该出厂?
- 若抽样结果有 1 件次品, 问该批产品是否该出厂?

解答：例 0.1

题目: 如上所述.

解答:

- 首先做出假设 $H_0 : p \leq 0.04$, 设随机变量 $X \sim B(12, p)$, 若假设 H_0 成立,

$$\Pr[X = 3] = \binom{12}{3} p^3 (1-p)^9 \leq 0.0097 \, .$$

  这是一个小概率事件, 即 $p \leq 0.04$ 时在 12 个样本中观测到 3 个次品的可能性极小, 这是一种 “不合理” 的现象. 因此, 应该拒绝原假设 $H_0 : p \leq 0.04$, 即 $p > 0.04$, 该产品不该出厂.

- 若 $X = 1$, 则

$$\Pr[X = 1] = \binom{12}{1} p (1-p)^1 \leq 0.4608$$

  这不是一个小概率事件, 故没有理由拒绝原假设 H_0, 产品可以出厂.

- 注意：当 $X = 1$ 情况下, 若直接利用参数估计计算 $p = 1/12 = 0.083 > 0.04$, 则不能出厂, 因此参数估计与假设检验是两回事.

假设检验的流程

建立假设

定义 0.1 设来自某个参数分布 $F\{(x,\theta)|\theta \in \Theta\}$ 的样本 $X_1, X_2, \ldots, X_n$, 其中 Θ 是参数空间, 设 $\Theta_0 \subset \Theta$, 且 $\Theta_0 \neq \emptyset$, 则命题 $H_0 : \theta \in \Theta_0$ 称为原假设或零假设 (null hypothesis). 若有另一个 $\Theta_1(\Theta_1 \subset \Theta, \Theta_0\Theta_1 = \emptyset)$, 则命题 $H_1 : \theta \in \Theta_1$ 称为对立假设或备择假设 (alternative hypothesis). 记为

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1$$

Remarks:

- 原假设 $H_0 : \mu = \mu_0$ 和备选假设 $H_1 : \mu \neq \mu_0$, 称为双边假设检验
- 原假设 $H_0 : \mu \leq \mu_0$ 和备选假设 $H_1 : \mu > \mu_0$, 称为单边 (右边) 检验
- 原假设 $H_0 : \mu \geq \mu_0$ 和备选假设 $H_1 : \mu < \mu_0$, 称为单边 (左边) 检验

选择检验统计量

由样本对原假设进行检验通常可以通过一个统计量完成, 该统计量称为 **检验统计量**.

- 总体均值的检验统计量可选为样本均值 $\bar{X}$
- 总体方差的检验统计量可选为无偏样本方差 S^2
- 事件 A 发生的概率的检验统计量可选为事件 A 出现的频率
- ...

确定显著水平

在假设检验中, 需要对 “不合理” 的事件给出一个定性描述: 即给出一上界 α, 当一事件发生的概率小于 α 时, 称为小概率事件. 通常取 $\alpha = 0.1,\ 0.05,\ 0.01$, 其具体取值根据实际问题而定.

- 在假定 H_0 成立下, 根据样本提供的信息判断出不合理的现象 (即, 概率小于 α 的事件发生了), 则认为假设 H_0 不显著, α 被称为显著水平.

- 但是不否定假设 H_0 并不代表假设 H_0 一定成立, 而只能说试验结果与假设 H_0 之间的差异不够显著, **没达到否定的程度**, 所以假设检验也被称为 “显著性检验”.

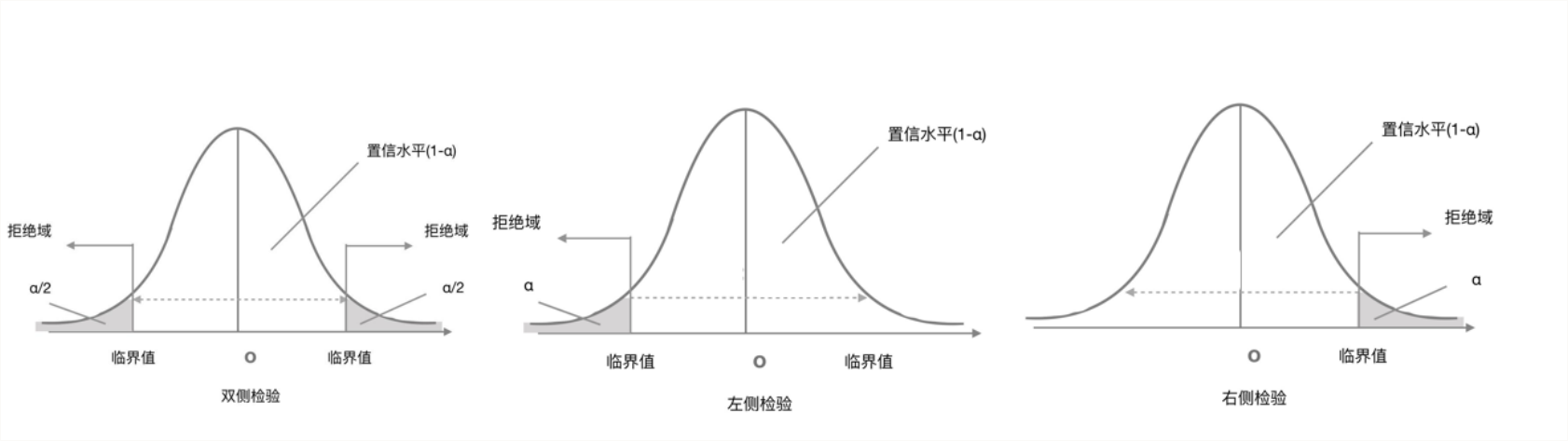
给出拒绝域, 由样本统计量做判断

给定显著性水平 α 后, 查表就可以得到具体临界值, 拒绝域是由显著性水平围成的区域, 拒绝域通常记为 W.

- 拒绝域的功能主要用来判断假设检验是否拒绝原假设的.
  - 拒绝域的作用. 确定统计量及其分布、显著性水平 α 之后, 原假设是否成立的问题可以转换为样本统计量是否属于拒绝域的问题.
- 由样本观测值 $X_1, X_2, \ldots, X_n$ 计算出来的检验统计量 $T(X_1, X_2, \ldots, X_n)$ 判断是否拒绝原假设:
  - 如果 $T(X_1, X_2, \ldots, X_n) \in W$, 则拒绝 H_0.
  - 如果 $T(X_1, X_2, \ldots, X_n) \in \overline{W}$, 则接受 H_0.

双边假设与双侧检验

- 原假设 $H_0 : \mu = \mu_0$ 和备选假设 $H_1 : \mu \neq \mu_0$, 称为双边假设检验
- 原假设 $H_0 : \mu \leq \mu_0$ 和备选假设 $H_1 : \mu > \mu_0$, 称为单边 (右边) 检验
- 原假设 $H_0 : \mu \geq \mu_0$ 和备选假设 $H_1 : \mu < \mu_0$, 称为单边 (左边) 检验



假设检验：例 0.2

例 0.2 假设某产品的重量服从 $\mathcal{N}(500, 16)$, 随机取出 5 件产品, 测得重量为 509, 507, 498, 502, 508, 问产品的期望是否正常? (显著性水平 $\alpha = 0.05$)

解答：例 0.2

解答:

- 建立建设: $H_0 : \mu = 500 \quad \text{vs} \quad H_1 : \mu \neq 500$.
- 设计检验统计量: 在原假设 H_0 成立下的条件求出其分布. 令样本均值 $\bar{X} = \sum_{i=1}^{5} X_i/5 = 504.8$, 设检验统计量为

$$Z = \frac{\bar{X} - 500}{\sqrt{16/5}} \sim \mathcal{N}(0, 1)$$

- 给定显著性水平 $\alpha = 0.05$, 查表得到临界值 $\mu_{0.025} = 1.96$, 使得

$$\Pr[|Z| > 1.96] = 0.05 \qquad \text{注意绝对值}$$

  成为一个小事件, 从而得到拒绝域 $\{Z : |Z| > 1.96\}$.
- 将样本值代入计算统计量 Z 的实测值

$$Z = \frac{|\bar{X} - 500|}{\sqrt{16/5}} = \frac{4.8}{4/\sqrt{5}} = 1.2 \times \sqrt{5} = 2.68 > 1.96$$

  由于实测值落入拒绝域, 因此判断为拒绝原假设 H_0 .

检验的两类错误

我们通过样本数据来判断总体参数的假设是否成立, 但样本是随机的, 因而有可能出现小概率的错误. 这种错误是由于采样的随机性导致的错误, 该错误分两种, 一种是弃真错误, 另一种是取伪错误.

- 第 I 类错误: “弃真”, 即当 H_0 为真时, 仍可能拒绝 H_0. 设犯第 I 类错误的概率为 α, 即显著性水平, 则有:

$$\alpha = \Pr[\text{拒绝} H_0 \mid H_0 \text{为真}]$$

- 第 II 类错误: “存伪”, 即当 H_0 不成立时, 仍可能接受 H_0. 设犯第 II 类错误的概率为 β, 即显著性水平, 则有:

$$\beta = \Pr[\text{接受} H_0 \mid H_0 \text{为假}]$$

检验的两类错误 – – Remarks

- 第 I 类错误与第 II 类错误互相关联, 当样本容量固定时, 一类错误概率的减少导致另一类错误概率的增加. 既然我们不可能同时控制一个检验犯第 I 类错误与第 II 类错误的概率, 则采取折中方案 (即 Neymam-Pearson 原则): 在控制第 I 类错误的前提下, 尽可能减小第 II 类错误的概率.

- 由于 Neymam-Pearson 原则提出了要控制犯第 I 类错误的概率 α, 因此在假设检验中, 通常将不宜轻易拒绝的假设作为原假设.
  - 给出设置 H_0 和 H_1 的原则, 因为 H_0 和 H_1 是相对的.

假设检验：例 0.3

例 0.3 设 (X_1, X_2, X_3, X_4) 是取自正态分布 $\mathcal{N}(\mu, 1)$ 的一个样本, 检验假设

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu = 1 \, ,$$

拒绝域为 $W = \{\bar{X} \geq 0.98\}$, 求此检验的两类错误概率.

解答：例 0.3

题目: 设 (X_1, X_2, X_3, X_4) 是取自正态分布 $\mathcal{N}(\mu, 1)$ 的一个样本, 检验假设 $H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu = 1$, 拒绝域为 $W = \{\bar{X} \geq 0.98\}$, 求此检验的两类错误概率.

解答:

- 第 I 类错误是指原假设 H_0 成立时, 但由于样本落入拒绝域而做出了拒绝原假设的情况, 当原假设 H_0 成立时, $\bar{X} \sim \mathcal{N}(0, 1/4)$, 因此犯第 I 类错误的概率为

$$P[\bar{X} \geq 0.98 \mid H_0] = 1 - \Phi\left(\frac{0.98 - 0}{\sqrt{1/4}}\right) = 1 - \Phi(1.96) = 0.025 \, .$$

- 第 II 类错误是指原假设 H_0 不成立而接受备择假设 H_1 时, 但由于样本落入接受域而做出了不拒绝原假设的情况, 当原假设 H_1 成立时, $\bar{X} \sim \mathcal{N}(1, 1/4)$, 因此犯第 II 类错误的概率为

$$P[\bar{X} < 0.98 \mid H_1] = 1 - \Phi\left(\frac{1 - 0.98}{\sqrt{1/4}}\right) = 1 - \Phi(0.04) = 0.4840 \, .$$

单个正态总体均值的检验

设 $X_1, X_2, \ldots, X_n$ 是来自正态分布 $\mathcal{N}(\mu, \sigma^2)$ 的样本, 考虑如下三种关于 μ 的检验问题:

$$\text{I} \quad H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

$$\text{II} \quad H_0 : \mu \geq \mu_0 \quad \text{vs} \quad H_1 : \mu < \mu_0$$

$$\text{III} \quad H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

其中, μ_0 是已知常数.

由于正态总体含两个参数, 总体方差 σ^2 已知与否对检验有影响. 下面我们分 σ 已知和未知两种情况讨论.

单个正态总体均值的检验

下面我们分 σ 已知和未知两种情况讨论.

- σ 已知: 由于 μ 的点估计是样本均值 $\bar{x}$, 根据正态分布的性质选择检验量

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \ .$$

给定显著性水平 α 后可得拒绝域, 这种检验方法称为 Z 检验法.

- σ 未知: 由于 σ 的点估计是无偏样本方差 S, 根据正态分布的性质选择检验量

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1) \ .$$

给定显著性水平 α 后可得拒绝域, 这种检验方法称为 t 检验法.

单个正态总体均值的检验

条件	H_0	H_1	检验统计量	拒绝域
	$\mu \le \mu_0$	$\mu > \mu_0$		$\{Z \ge \mu_\alpha\}$
σ 已知	$\mu \ge \mu_0$	$\mu < \mu_0$	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$	$\{Z \le -\mu_\alpha\}$
	$\mu = \mu_0$	$\mu \ne \mu_0$		$\{\lvert Z\rvert \ge \mu_{\alpha/2}\}$
	$\mu \le \mu_0$	$\mu > \mu_0$		$\{t \ge t_\alpha(n-1)\}$
σ 未知	$\mu \ge \mu_0$	$\mu < \mu_0$	$t = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$	$\{t \le -t_\alpha(n-1)\}$
	$\mu = \mu_0$	$\mu \ne \mu_0$		$\{\lvert t\rvert \ge t_{\alpha/2}(n-1)\}$

假设检验：例 0.4

例 0.4 已知某产品的重量 $X \sim \mathcal{N}(4.55, 0.108^2)$, 现随机抽取 5 个产品, 其质量分别为 4.28, 4.40, 4.42, 4.35, 4.27. 问产品的期望在 $\alpha = 0.05$ 下有无显著性变化. ($\mu_{0.025} = 1.96$)

解答：例 0.4

题目: 已知某产品的重量 $X \sim \mathcal{N}(4.55, 0.108^2)$, 现随机抽取 5 个产品, 其质量分别为 4.28, 4.40, 4.42, 4.35, 4.27. 问产品的期望在 $\alpha = 0.05$ 下有无显著性变化. ($\mu_{0.025} = 1.96$)

解答:

- 提出假设: $H_0 : \mu = 4.55 \quad \text{vs} \quad H_1 : \mu \neq 4.55.$
- 若 H_0 成立, 选择检验量

$$Z = \frac{\bar{X} - 4.55}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \ ,$$

  求得拒绝域为 $|Z| \geq \mu_{\alpha/2} = 1.96.$
- 计算样本均值可知 $\bar{X} = 4.364$, 于是有

$$\frac{|\bar{X} - 4.55|}{0.108/\sqrt{5}} = 3.851 > 1.96 \ ,$$

  由此可拒绝 H_0, 说明有显著变化.

假设检验：例 0.5

例 0.5 某灯泡平均寿命要求不低于 1000 小时被称为合格, 已知灯泡的寿命 $X \sim \mathcal{N}(\mu, 100^2)$, 现在随机抽取 25 件, 其样本均值为 $\bar{X} = 960$. 在显著性水平 $\alpha = 0.05$ 的情况下, 检验这批灯泡是否合格. ($\mu_{0.05} = 1.645$)

解答：例 0.5

题目: 某灯泡平均寿命要求不低于 1000 小时被称为合格, 已知灯泡的寿命 $X \sim \mathcal{N}(\mu, 100^2)$, 现在随机抽取 25 件, 其样本均值为 $\bar{X} = 960$. 在显著性水平 $\alpha = 0.05$ 的情况下, 检验这批灯泡是否合格. ($\mu_{0.05} = 1.645$)

解答:

- 提出假设: $H_0 : \mu \geq 1000 \quad \text{vs} \quad H_1 : \mu < 1000.$
- 若 H_0 成立, 选择检验量

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

  求得拒绝域为 $Z < -\mu_\alpha = -1.645$.
- 计算样本均值可知 $\bar{X} = 960$, 于是有

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = -2 < -1.645$$

  由此可拒绝 H_0, 说明这批灯泡不合格.

假设检验：例 0.6

例 0.6 某厂生产的铝材长度服从正态分布, 假设其均值为 240 cm. 现从一批铝材中随机抽取 5 件产品, 测得其长度 (单位: cm) 为 239.7, 239.6, 239, 240, 239.2, 试判断这批铝材的长度是否满足设定要求?

解答：例 0.6

题目: 某厂生产的铝材长度服从正态分布, 假设其均值为 240 cm. 现从一批铝材中随机抽取 5 件产品, 测得其长度 (单位: cm) 为 239.7, 239.6, 239, 240, 239.2, 试判断这批铝材的长度是否满足设定要求?

解答:

- 提出假设: $H_0 : \mu = 240 \quad \text{vs} \quad H_1 : \mu \neq 240.$
- 若 H_0 成立, 选择检验量
$$t = \frac{\bar{X} - 240}{S/\sqrt{n}} \sim t(4) \, ,$$
求得拒绝域为 $|t| \geq t_{\alpha/2}(n-1)$.
- 若取 $\alpha = 0.05$, 则查表得 $t_{0.025}(4) = -t_{0.975}(4) = 2.7764$. 计算样本均值可知 $\bar{X} = 239.5$, $S = 0.4$, 故
$$t = \frac{239.5 - 240}{0.4/\sqrt{5}} = -2.795 < -2.7764 \, .$$
由此可拒绝 H_0, 说明这批铝材的长度不满足设定要求.

两个正态总体均值差的检验

设 $X_1, X_2, \ldots, X_n$ 是来自正态分布 $\mathcal{N}(\mu_1, \sigma_1^2)$ 的样本, $Y_1, Y_2, \ldots, Y_m$ 是来自正态分布 $\mathcal{N}(\mu_2, \sigma_2^2)$ 的样本, 两个样本相互独立. 考虑如下三种检验问题:

$$\text{I} \quad H_0 : \mu_1 - \mu_2 \leq 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 > 0$$

$$\text{II} \quad H_0 : \mu_1 - \mu_2 \geq 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 < 0$$

$$\text{III} \quad H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

下面我们分 σ_1 和 σ_2 已知和相等但未知两种情况讨论.

两个正态总体均值差的检验

下面我们分 σ_1 和 σ_2 已知和相等但未知两种情况讨论.

- σ_1 和 σ_2 已知: 由于 $\mu_1 - \mu_2$ 的点估计 $\bar{X} - \bar{Y}$ 的分布已知

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right) ,$$

  根据正态分布的性质选择检验量

$$U = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0, 1) .$$

  给定显著性水平 α 后可得拒绝域.

两个正态总体均值差的检验

- $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知: 有

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right), \quad \frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2} \sim \chi^2(m+n-2),$$

故可以选择检验量

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_W \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2),$$

其中

$$S_W = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2},$$

给定显著性水平 α 后可得拒绝域.

两个正态总体均值差的检验

条件	H_0	H_1	检验统计量	拒绝域
σ_1 和 σ_2 已知	$\mu_1-\mu_2\le 0$	$\mu_1-\mu_2>0$	$U=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}$	$\{U\ge\mu_\alpha\}$
	$\mu_1-\mu_2\ge 0$	$\mu_1-\mu_2<0$		$\{U\le-\mu_\alpha\}$
	$\mu_1-\mu_2=0$	$\mu_1-\mu_2\ne\mu_0$		$\{\lvert U\rvert\ge\mu_{\alpha/2}\}$
$\sigma_1^2=\sigma_2^2=\sigma^2$ 未知	$\mu_1-\mu_2\le 0$	$\mu_1-\mu_2>0$	$t=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n}+\frac{1}{m}}}$	$\{t\ge t_\alpha(m+n-2)\}$
	$\mu_1-\mu_2\ge 0$	$\mu_1-\mu_2<0$		$\{t\le-t_\alpha(m+n-2)\}$
	$\mu_1-\mu_2=0$	$\mu_1-\mu_2\ne\mu_0$		$\{\lvert t\rvert\ge t_{\alpha/2}(m+n-2)\}$

假设检验：例 0.7

例 0.7 某厂铸造车间为提高零件的耐磨性, 试制了一种镍合金零件以取代铜合金零件, 为此从两种零件中各抽取容量分别为 8 和 9 的样本, 测得其耐磨性为下表. 假设两类零件的耐磨性服从正态分布, 且方差相等, 在显著性水平 $\alpha = 0.05$ 的情况下, 检验判断镍合金的耐磨性是否有明显提高.

镍合金	76.43	76.21	73.58	69.69	65.29	70.83	82.75	72.34	
铜合金	73.66	64.27	69.34	71.37	69.77	68.12	67.27	68.07	62.61

解答：例 0.7

解答:

- 用 X 表示镍合金的耐磨性, Y 表示铜合金的耐磨性, 且 $X \sim \mathcal{N}(\mu_1, \sigma^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma^2)$. 提出假设: $H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$.
- 由于两者的方差相等但未知, 选择 t 检验量

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_W \sqrt{\frac{1}{n} + \frac{1}{m}}} ,$$

  经计算有 $\bar{X} = 73.39$, $\bar{Y} = 68.2756$, $\sum_{i=1}^{8}(X_i - \bar{X})^2 = 191.7958$, $\sum_{i=1}^{9}(Y_i - \bar{Y})^2 = 91.1548$. 从而, $S_W = \sqrt{\frac{1}{8+9-2}(191.7958 + 91.1548)} = 4.3432$,

$$t = \frac{73.39 - 68.2756}{4.3432 \times \sqrt{\frac{1}{8} + \frac{1}{9}}} = 2.4234 .$$

- 查表可知 $t_{0.975}(15) = 2.1314 < 2.4234$, 故拒绝原假设, 即判断镍合金耐磨性有所提高.

解答：例 0.7

解答:

- 该问题可以采用另一种假设方式. 提出假设: $H_0 : \mu_1 \le \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \ge \mu_2$.
- 由于两者的方差相等但未知, 选择 t 检验量

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_W \sqrt{\frac{1}{n} + \frac{1}{m}}} ,$$

  经计算有 $\bar{X} = 73.39$, $\bar{Y} = 68.2756$, $\sum_{i=1}^{8}(X_i - \bar{X})^2 = 191.7958$, $\sum_{i=1}^{9}(Y_i - \bar{Y})^2 = 91.1548$. 从而, $S_W = \sqrt{\frac{1}{8+9-2}(191.7958 + 91.1548)} = 4.3432$,

$$t = \frac{73.39 - 68.2756}{4.3432 \times \sqrt{\frac{1}{8} + \frac{1}{9}}} = 2.4234 .$$

- 查表可知 $t_{0.95}(15) = 1.7531 < 2.4234$, 故拒绝原假设, 即判断镍合金耐磨性有所提高.

成对数据检验

在很多实际应用中, 为了比较两个总体之间的差异, 往往会得到一批成对的数据, 然后基于观察的数据分析判断两个总体之间是否有显著的区别, 这种方法称为 成对 (pairwise) 比较法.

假设观察到 n 对相互独立的随机变量 $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, 其中 $X_1, X_2, \ldots, X_n$ 和 $Y_1, Y_2, \ldots, Y_n$ 分别是总体 X 和 Y 的两个样本, 检验这两种方法是否性能相同, 即检验总体 X 和 Y 的期望是否相等. 因为对相同的数据集 i 而言, X_i 和 Y_i 不能被认为相互独立. 由此假设

$$Z = X - Y \sim \mathcal{N}(\mu, \sigma^2)$$

并建立假设 $H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu \neq 0$, 方差 σ^2 未知考虑 t 检验量

$$t = \frac{\bar{Z}}{S/\sqrt{n}} \sim t(n-1)$$

在显著性水平 α 下得到拒绝域为: $|t| > t_{\alpha/2}(n-1)$.

假设检验：例 0.8

例 0.8 假设有两种学习方法 A 和 B, 在 9 个数据集上取得的效果如下表. 问这两种方法在 $\alpha = 0.05$ 下是否有显著区别?

数据集	1	2	3	4	5	6	7	8	9
方法 A	0.6	0.9	0.8	0.7	0.6	0.9	0.8	0.9	0.7
方法 B	0.7	0.95	0.7	0.6	0.7	0.9	0.9	0.8	0.6

解答：例 0.8

题目: 假设有两种学习方法 A 和 B, 在 9 个数据集上取得的效果如下表. 问这两种方法在 $\alpha = 0.05$ 下是否有显著区别?

数据集	1	2	3	4	5	6	7	8	9
方法 A	0.6	0.9	0.8	0.7	0.6	0.9	0.8	0.9	0.7
方法 B	0.7	0.95	0.7	0.6	0.7	0.9	0.9	0.8	0.6

解答:

- 设 $Z_i = X_i - Y_i$ $(i \in [10])$, 可得样本均值 $\bar{Z} = 0.0056$ 和方差 $S^2 = 0.009$, 由此可得观察值

$$|t| = \frac{|\bar{Z}|}{S/\sqrt{n}} = \frac{0.0056}{0.03} \approx 0.176 < t_{0.025}(8) = 2.3060$$

  由此说明这两种方法没有显著性区别.

单个正态总体方差的检验

设 $X_1, X_2, \ldots, X_n$ 是来自正态分布 $\mathcal{N}(\mu, \sigma^2)$ 的样本. 假定 μ 未知, 采用 χ^2 检验统计量

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

在显著性水平为 α 时求解拒绝域, 这种检验方法称为 χ^2 检验法.

- 假设 $H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 \neq \sigma_0^2$ 的拒绝域为

$$\{\chi^2 \geq \chi^2_{\alpha/2}(n-1)\} \cup \{\chi^2 \leq \chi^2_{1-\alpha/2}(n-1)\}$$

- 假设 $H_0 : \sigma^2 \geq \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 < \sigma_0^2$ 的拒绝域为

$$\{\chi^2 \leq \chi^2_{1-\alpha}(n-1)\}$$

- 假设 $H_0 : \sigma^2 \leq \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 > \sigma_0^2$ 的拒绝域为

$$\{\chi^2 \geq \chi^2_{\alpha}(n-1)\}$$

假设检验：例 0.9

例 0.9 某类钢板每块的重量 X 服从正态分布, 质量指标要求钢板重量的方差不得超过 0.016. 现从某批钢板中随机抽取 25 块, 得其样本方差 $S^2 = 0.025$, 问该批钢板的重量是否满足指标要求?

解答：例 0.9

题目: 某类钢板每块的重量 X 服从正态分布, 质量指标要求钢板重量的方差不得超过 0.016. 现从某批钢板中随机抽取 25 块, 得其样本方差 $S^2 = 0.025$, 问该批钢板的重量是否满足指标要求?

解答:

- 提出假设: $H_0 : \sigma^2 \leq 0.016 \quad \text{vs} \quad H_1 : \sigma^2 > 0.016.$
- 查表得 $\chi^2_{0.05}(24) = 36.415$, 若 H_0 成立, 选择检验量

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{24 \times 0.025}{0.016} = 37.5 > 36.415.$$

由此可拒绝 H_0, 说明批钢板的重量不满足要求.

两个正态总体方差比的检验

设 $X_1, X_2, \ldots, X_n$ 是来自正态分布 $\mathcal{N}(\mu_1, \sigma_1^2)$ 的样本, $Y_1, Y_2, \ldots, Y_m$ 是来自正态分布 $\mathcal{N}(\mu_2, \sigma_2^2)$ 的样本. 假定 μ_1, μ_2 未知, 无偏样本方差分别为 S_x^2, S_y^2, 由此可建立如下检验统计量:

$$F = \frac{S_x^2}{S_y^2} \sim F(n-1, m-1)\,, \quad \text{当 } \sigma_1^2 = \sigma_2^2 \text{ 时}$$

在显著性水平为 α 时求解拒绝域:

- 假设 $H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$ 的拒绝域为
  $\{F \leq F_{1-\alpha/2}(n-1, m-1)\} \cup \{F \geq F_{\alpha/2}(n-1, m-1)\}$
- 假设 $H_0 : \sigma_1^2 \geq \sigma_2^2 \quad \text{vs} \quad H_1 : \sigma_1^2 < \sigma_2^2$ 的拒绝域为
  $\{F \leq F_{1-\alpha}(n-1, m-1)\}$
- 假设 $H_0 : \sigma_1^2 \leq \sigma_2^2 \quad \text{vs} \quad H_1 : \sigma_1^2 > \sigma_2^2$ 的拒绝域为
  $\{F \geq F_{\alpha}(n-1, m-1)\}$

假设检验：例 0.10

例 0.10 甲乙两台机床加工某种零件，零件的直径服从正态分布，总体方差反映了加工精度，为比较两台机床的加工精度有无差别，现从各自加工的零件中分别抽取 7 件产品和 8 件产品，测得其直径为

甲机床	16.2	16.8	15.8	15.5	16.7	15.6	15.8
乙机床	15.9	16	16.4	16.1	16.5	15.8	15.7	15

解答：例 0.10

题目: 甲乙两台机床加工某种零件, 零件的直径服从正态分布, 总体方差反映了加工精度, 为比较两台机床的加工精度有无差别, 现从各自加工的零件中分别抽取 7 件产品和 8 件产品, 测得其直径为

甲机床	16.2	16.8	15.8	15.5	16.7	15.6	15.8
乙机床	15.9	16	16.4	16.1	16.5	15.8	15.7	15

解答:

- 提出假设: $H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$.
- 经计算有 $S_x^2 = 0.2729$, $S_y^2 = 0.2164$, 于是有 $F = 0.2729/0.2164 = 1.261$. 查表得 $F_{0.025}(6,7) = 5.12$, $F_{0.975}(6,7) = 1/F_{0.025}(7,6) = 0.175$. 其拒绝域为

$$W = \{F \leq 0.175\} \cup \{F \geq 5.12\}$$

  样本未落入拒绝域, 可认为两台机床的加工精度无差别.

假设检验：补充例 0.11

例 0.11 某农场对甜瓜的培育引入了新方法，声称他们培育出来的甜瓜平均含糖量达到了 6g/100g。从该农场一批成熟的甜瓜中随机抽取了 25 个进行含糖量测试，测得

$$\bar{x} = 5.7, \quad s = 1.2.$$

设甜瓜含糖量服从正态分布 $N(\mu, \sigma^2)$，且 μ, σ^2 未知，问这种培育是否有效？

(1) 如果你是农场主，要求第一类错误不超过 5%；

(2) 如果你是消费者，要求第一类错误不超过 5%。

解答：例 0.11

题目: 如上所述.

解答:

- (1) 农场主角度：根据题意可设检验问题为

$$H_0 : \mu \geq 6 \Longleftrightarrow H_1 : \mu < 6,$$

只有这样定义，才能保证第一类错误概率，即当含糖量达到了 $6g/100g$ 而错误地否认说没有达到 $6g/100g$ 这个错误概率不超过 5%。这是一个单正态总体关于均值的左侧检验，且方差未知，因此拒绝域为

$$W = \left\{ \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} < -t_{1-\alpha}(n-1) \right\}.$$

经计算可得

$$\frac{\sqrt{n}(\bar{x} - \mu_0)}{s} = \frac{5 \times (5.7 - 6)}{1.2} = -1.25 > -t_{0.05}(24) = -1.711,$$

说明样本观测值不落在拒绝域内，因此不能拒绝原假设，即认为这种培育是有效的。

- (2) 消费者角度：根据题意可设检验问题为

$$H_0 : \mu \leq 6 \Longleftrightarrow H_1 : \mu > 6,$$

这样的原假设，控制了当含糖量没有明显达到 $6g/100g$ 而错误地否认说已经达到了 $6g/100g$ 这个错误概率不超过 5%。这是一个单正态总体关于均值的右侧检验，且方差未知，因此拒绝域为

$$W = \left\{ \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} > t_{1-\alpha}(n-1) \right\}.$$

经计算可得

$$\frac{\sqrt{n}(\bar{x} - \mu_0)}{s} = \frac{5 \times (5.7 - 6)}{1.2} = -1.25 < t_{0.05}(24) = 1.711,$$

因此不能拒绝原假设，即认为这种培育是无效的。

非参假设检验

前面介绍的各种检验都是在总体服从正态分布前提下, 对参数进行假设检验. 实际中可能遇到这样的情形: 总体服从何种理论分布并不知道, 要求我们直接对总体分布提出一个假设.

设 $X_1, X_2, \ldots, X_n$ 是来自总体 $F(x)$ 的样本, 检验原假设

$$H_0 : F(x) = F_0(x)$$

其中 $F_0(x)$ 是形式已知但含有若干个未知参数的分布函数. 这个分布检验问题就是检验观测数据是否与理论分布相符. 这一类检验问题统称为 **分布的拟合检验**, 它们是一类非参数检验问题.

- 若总体 X 为离散随机变量: $H_0 : \Pr[X = X_i] = p_i, (i = 1, 2, \ldots)$.
- 若总体 X 为连续随机变量: $H_0 : X$ 的密度函数 $p(x) = p_0(x)$.
- 若 p_i 或 $p_0(x)$ 包含未知参数, 应先用最大似然估计/矩估计估计参数.

分布的 χ^2 拟合优度检验

当样本容量较大时, 分布的拟合检验可以用 χ^2 拟合优度检验来解决. 下面介绍 χ^2 拟合优度检验法.

- 将随机试验结果全体 Ω 分为 k 个互不相容的事件 $A_1, A_2, \ldots, A_k$, 并使得落入每个 A_i 的样本个数不小于 5, 且 $\cup_{i=1}^{k} A_i = \Omega$.
- 根据假设 $H_0 : F(x) = F_0(x)$ 计算概率 $p_i = \Pr(A_i)$.
- 对样本 $X_1, X_2, \ldots, X_n$, 事件 A_i 出现的频率为 n_i/n.
- 当 H_0 为真时, 频率 n_i/n 与概率 p_i 差异不应太大.
- 基于这种思想, 皮尔逊构造了检验量:

$$W = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

称为皮尔逊 χ^2 统计量.

分布的 χ^2 拟合优度检验

定理 0.1 若分布 $F_0(x)$ 不包含未知参数, 当 H_0 为真时 (无论 H_0 中的分布属于什么分布), 统计量

$$W = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2(k-1)$$

给定显著性水平 α, 若 $W > \chi^2_\alpha(k-1)$ 则拒绝 H_0.

定理 0.2 当 $n \to +\infty$ 时, 有 $W \xrightarrow{d} \chi^2(k-r-1)$ 成立. 这里 r 指的是模型/测试分布的未知参数个数.

假设检验：例 0.15

例 0.15 某试验有四种不同的结果 $\{A, B, C, D\}$. 现进行如下实验: 独立重复实验直到结果 A 发生为止. 试验 200 次, 记录抛掷的次数结果如下表, 试问该试验是否为均匀分布?

重复次数	1	2	3	4	≥5
频数	56	48	32	28	36

解答：例 0.15

题目: 如上所述.

解答:

- 提出假设: H_0 : 均匀分布, 用随机变量 X 表示试验结果 A 发生时重复的次数, 有

$$p_1 = P(X=1) = \frac{1}{4} \quad p_2 = P(X=2) = \frac{3}{4} \times \frac{1}{4} \quad p_3 = P(X=3) = \left(\frac{3}{4}\right)^2 \cdot \frac{1}{4}$$

$$p_4 = P(X=4) = \left(\frac{3}{4}\right)^3 \cdot \frac{1}{4} \quad p_5 = P(X=5) = 1 - \frac{1}{4} - \frac{3}{16} - \left(\frac{3}{4}\right)^3 \cdot \frac{1}{4}$$

- 计算统计检验量

$$W = \sum_{i=1}^{5} \frac{(n_i - np_i)^2}{np_i} = 18.21$$

均匀分布的假设中无任何参数, 根据统计量实值 $W > \chi^2_{0.05}(5-1) = 9.488$ 则拒绝 H_0, 该试验不服从均匀分布.

假设检验：例 0.16

例 0.16 1911 年著名物理学家卢瑟福等人为探索原子的内部结构进行了一项实验, 即用一束带正电的、质量比电子大得多的高速运动的 α 粒子轰击金箔, 证明了正电荷集中在原子中心. 考察下表中卢瑟福实验的数据, 是以 7.5s 为时间单位所做的 2608 次观察所得的数据, 观测的是一枚放射性 α 物质在单位时间内放射的质点数. 试问 7.5s 中放射出的 α 质点数是否服从泊松分布 $P(\lambda)$?

质点数 k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
观察数 n_k	57	203	383	525	532	408	273	139	45	27	10	4	2	0	0

解答：例 0.16

解答:

- 提出假设: H_0 : 7.5s 中放射出的 α 质点数服从泊松分布 $P(\lambda)$, 又泊松分布参数 λ 的极大似然估计为 $\hat{\lambda} = \bar{X}$, 即

$$\hat{\lambda} = \frac{1}{n}\sum_{k=1}^{n} X_i = \frac{\sum_{k=0}^{14} k n_k}{\sum_{k=0}^{14} n_k} = 3.87$$

- 计算泊松分布的概率估计值

$$\hat{p}_k = \frac{\hat{\lambda}^k}{k!} e^{-\hat{\lambda}}, \quad k = 1, 2, \ldots$$

为了满足每一类出现的样本观测次数不小于 5, 我们把 $k \geq 11$ 作为一类, 记为第 12 类 (4+2>5), 可以得到检验统计量的值为

$$W = \sum_{i=1}^{12} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} = 12.8967\ .$$

此处分布自由度为 $12 - 1 - 1 = 10$. 对 $\alpha = 0.05$, 查表得 $W \leq \chi^2_{0.95}(10) = 18.3070$,

因此不能拒绝原假设 H_0, 可以认为该放射物质在 7.5 秒时间内放射的 α 质点数与泊松分布吻合.

- 附: 分布拟合检验计算过程:

序号	质点数	观测数 n_i	概率估计 $\hat{p}_i$	$(n_i - n\hat{p}_i)^2/n\hat{p}_i$
1	0	57	0.0209	0.1147
2	1	203	0.0807	0.2672
3	2	383	0.1562	1.4614
4	3	525	0.2015	0.0005
5	4	532	0.195	1.0766
...	...	...	...	...
11	10	10	0.0043	0.1286
12	≥ 11	6	0.0022	0.0158

假设检验：例 0.17

例 0.17 某工厂生产一种滚珠, 现随机抽取 50 件产品, 测得其直径 (单位: mm) 为

15.0	15.8	15.2	15.1	15.9	14.7	14.8	15.5	15.6	15.3
15.0	15.6	15.7	15.8	14.5	15.1	15.3	14.9	14.9	15.2
15.9	15.0	15.3	15.6	15.1	14.9	14.2	14.6	15.8	15.2
15.2	15.0	14.9	14.8	15.1	15.5	15.5	15.1	15.1	15.0
15.3	14.7	14.5	15.5	15.0	14.6	14.6	14.2	14.2	14.5

问滚珠直径是否服从正态分布?

解答：例 0.17

解答:

- 设滚珠直径为 X, 提出假设: $H_0 : F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$, 由样本数据求得 μ, σ 的最大似然估计为 $\hat{\mu} = 15.1, \hat{\sigma}^2 = 0.4379^2$.
- 根据样本数据特点并考虑到各组观测值个数不低于 5, 去分点为

$$a_0 = -\infty, \quad a_1 = 14.55, \quad a_2 = 14.95, \quad a_3 = 15.35, \quad a_4 = 15.75, \quad a_5 = +\infty$$

由此把数据分为 5 组, 各组数据个数分别为

$$n_1 = 6, \quad n_2 = 11, \quad n_3 = 20, \quad n_4 = 8, \quad n_5 = 5$$

- 利用公式

$$\hat{p}_i = \Phi\left(\frac{a_i - 15.1}{0.4379}\right) - \Phi\left(\frac{a_{i-1} - 15.1}{0.4379}\right), \quad i = 1, 2, 3, 4, 5$$

求得

$$\hat{p}_1 = 0.104559, \quad \hat{p}_2 = 0.261412, \quad \hat{p}_3 = 0.349998, \quad \hat{p}_4 = 0.215174, \quad \hat{p}_5 = 0.068857$$

- 可以得到检验统计量的值为

$$W = \sum_{i=1}^{5} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} = 2.2109.$$

此处分布自由度为 $5 - 2 - 1 = 2$, 对 $\alpha = 0.05$, 查表得 $\chi^2_{0.05}(2) = 5.9915 > 2.2109$, 因此不能拒绝原假设 H_0.

假设检验与机器学习

- 二分类器比较 by McNemar 检验
- 模型性能比较 by 配对 t 检验

应用一：二分类器与 McNemar 检验

在机器学习模型评估中，若两个分类器 A 和 B 在 同一测试集 上给出二元（正确/错误）预测结果，我们希望判断两者性能是否存在显著差异。由于每个样本在两个模型上都是配对的，因此适用 **McNemar 检验**。

假设有 n 个测试样本，对每个样本记录分类器 A 和 B 是否预测正确，构造 2×2 配对列联表

	B 正确	B 错误	行和
A 正确	n_{11}	n_{10}	$n_{1\cdot}$
A 错误	n_{01}	n_{00}	$n_{0\cdot}$
列和	$n_{\cdot 1}$	$n_{\cdot 0}$	n

由于 n_{11} 与 n_{00} 的样本对两者评价相同，差异仅体现在 n_{10} 与 n_{01}

Step 1: 假设与检验统计量

原假设与备择假设为

$$H_0 : P(\text{A 正确,B 错误}) = P(\text{A 错误,B 正确}),$$
$$H_1 : P(\text{A 正确,B 错误}) \neq P(\text{A 错误,B 正确}).$$

检验统计量定义为

- 含连续性校正
$$\chi^2_{\text{MC}} = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$
- 若不进行校正，则为
$$\chi^2 = \frac{(n_{01} - n_{10})^2}{n_{01} + n_{10}}$$

在 H_0 成立时，χ^2_{MC} 近似服从自由度为 1 的卡方分布。

Step 2: 数值示例

假设测试集规模为 $n = 200$，结果如下

$$n_{01} = 30\ , \quad n_{10} = 10\ , \quad n_{01} + n_{10} = 40\ .$$

则

$$\chi^2_{\text{MC}} = \frac{(|30 - 10| - 1)^2}{40} = \frac{(19)^2}{40} = \frac{361}{40} = 9.025\ .$$

查 χ^2 分布表得

$$p = P(\chi^2 \geq 9.025) \approx 0.00266\ .$$

结论:

(1) $p < 0.01 \rightarrow$ 在显著性水平 $\alpha = 0.01$ 下拒绝原假设 H_0, 说明分类器 A 与 B 的性能差异具有统计显著性

(2) $n_{01} > n_{10} \rightarrow$ 分类器 B 在更多样本上预测正确, 说明 B 的总体性能显著优于 A

应用一：小结

公式汇总

$$H_0 : p_{01} = p_{10}$$

$$\chi^2_{\text{MC}} = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

$$p\text{-value} = P(\chi^2_1 \geq \chi^2_{\text{MC}})$$

当 $p < \alpha$ 时，拒绝 H_0，认为两个分类器的表现差异显著。

注意：

- McNemar 检验只利用 “不一致” 的样本数 (n_{01}, n_{10})；
- 若 $n_{01} + n_{10} < 25$，不应使用 χ^2 近似，应改用精确二项检验；
- 若要比较多个模型，可进行 Bonferroni 校正或控制 FDR。

应用二：模型性能比较与配对 t 检验

在机器学习中，我们常常希望比较两个模型在同一数据集上的性能是否存在显著差异。

设有两个分类模型 M_1 和 M_2，使用 k 折交叉验证（k-fold cross validation）得到它们在每一折的准确率分别为：

$$M_1 \text{ 分类器:} \quad \{a_1^{(1)}, a_2^{(1)}, \ldots, a_k^{(1)}\}$$
$$M_2 \text{ 分类器:} \quad \{a_1^{(2)}, a_2^{(2)}, \ldots, a_k^{(2)}\}$$

为了判断两个模型的平均性能是否存在显著差异，我们可以对每一折的差值进行配对 t 检验（paired t-test）。

- $H_0 : \mu_d = 0$ （两模型平均性能无显著差异）
- $H_1 : \mu_d \neq 0$ （两模型平均性能存在显著差异）

检验步骤

1. 计算每一折的性能差值

$$d_i = a_i^{(1)} - a_i^{(2)}, \quad i = 1, 2, \ldots, k.$$

2. 计算差值的样本均值与样本标准差

$$\bar{d} = \frac{1}{k}\sum_{i=1}^{k} d_i, \quad s_d = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(d_i - \bar{d})^2}.$$

3. 构造检验统计量

$$t = \frac{\bar{d}}{s_d/\sqrt{k}}.$$

4. 在显著性水平 α 下，查 t 分布表（自由度 $k-1$），若

$$|t| > t_{k-1,\alpha/2},$$

则拒绝原假设 H_0，认为两模型的平均性能存在显著差异。

具体例子

设进行 $k = 5$ 折交叉验证，得到如下准确率

Fold	1	2	3	4	5
M_1	0.90	0.88	0.91	0.89	0.87
M_2	0.85	0.83	0.86	0.88	0.85
$d_i = M_1 - M_2$	0.05	0.05	0.05	0.01	0.02

计算差值的均值与标准差 $\bar{d} = 0.036$ 和 $s_d = 0.019$ 。

计算检验统计量

$$t = \frac{0.036}{0.019/\sqrt{5}} = 4.23 \, .$$

在显著性水平 $\alpha = 0.05$ 下，自由度 $df = 4$，查表得 $t_{4,0.025} = 2.776$。因为 $4.23 > 2.776$，拒绝 H_0，说明模型 M_1 的平均准确率显著高于模型 M_2。

假设检验与置信区间的关系

不同点: 假设检验与置信区间最大区别在于解决问题的不相同:

- 假设检验: 根据样本信息判断关于总体的某个假设是否正确.
- 置信区间: 估计未知参数的取值范围.

相似点: 检验统计量与枢轴量的构造相似, 显著性水平 α 的假设检验不拒绝域的边界与置信水平为 $1-\alpha$ 的区间估计的上下限有对应关系.

- 对于 σ 未知时单个正态总体均值的双边检验, 显著性水平 α 对应的拒绝域为 $\{|t| \geq t_{\alpha/2}(n-1)\}$, 因此不拒绝域可写为
$$\overline{W} = \{\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1) < \mu_0 < \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\}$$
- 给定置信度 $1-\alpha$, 对于 σ 未知时 μ 的置信区间为
$$\left[\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1) < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\right]$$

假设检验与置信区间的关系: 在机器学习中的例子

假设检验与置信区间是统计推断的两种等价形式。在机器学习实验中，它们常用于判断不同模型性能的显著性与不确定性。下面以比较两个分类模型在 k 折交叉验证中的准确率为例进行说明。

实验数据：考虑两个分类模型 M_1 与 M_2，在 $k = 10$ 折交叉验证中的准确率如下

Fold	1	2	3	4	5	6	7	8	9	10
M_1	0.92	0.90	0.89	0.93	0.91	0.88	0.90	0.92	0.89	0.91
M_2	0.90	0.89	0.87	0.91	0.90	0.86	0.88	0.91	0.87	0.90

定义每一折的差值为 $d_i = M_1^{(i)} - M_2^{(i)}$, $i = 1, \ldots, 10$

计算得到 $\bar{d} = 0.020$ 和 $s_d = 0.007$.

假设检验视角

要检验两个模型的平均准确率是否存在显著差异，假设如下：

$$H_0: \quad \mu_d = 0 \quad \text{(两模型无显著差异)}$$
$$H_1: \quad \mu_d > 0 \quad \text{(模型}M_1\text{ 更好)}$$

配对 t 检验统计量为

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{k}} = \frac{0.020}{0.007/\sqrt{10}} = 9.04$$

查 t 分布表可得

$$p < 0.001$$

因此, 有结论

拒绝 H_0, 说明模型 M_1 的准确率显著高于模型 M_2

置信区间视角

同样的数据可用于构造 95% 置信区间

$$\bar{d} \pm t_{k-1,0.975} \cdot \frac{s_d}{\sqrt{k}}$$

代入数值（$t_{9,0.025} = 2.262$）

$$0.020 \pm 2.262 \times \frac{0.007}{\sqrt{10}} = [0.014, 0.026]$$

由于区间不包含 0，可得到与假设检验相同的结论

模型 M_1 的平均准确率显著高于模型 M_2

假设检验与置信区间的关系

假设检验与置信区间的关系可以形式化表达为

$$0 \notin \text{CI}_{1-\alpha} \quad \Longleftrightarrow \quad \text{拒绝} H_0 \text{ at level } \alpha$$

对于机器学习意义,

- **假设检验**用于判断性能差异是否显著
  - 显著性检验可回答: 模型性能是否显著更好?
- **置信区间**提供性能差异的范围估计
  - 置信区间估计可回答: 性能提升大约在什么区间内?

表 1: 假设检验与置信区间在机器学习中的对比

方法	目的	输出	结论判断
假设检验	判断差异是否显著	p-value	若 $p < \alpha$，拒绝 H_0
置信区间	估计差异范围	区间 $[L, U]$	若 $0 \notin [L, U]$，拒绝 H_0

列联表的独立性检验

下面我们分析按两个或多个特征分类的频数数据, 这种数据通常称为交叉分类数据, 它们一般以表格形式给出, 称为列联表. 列联表分析的基本问题是, **考察各属性之间有无关联, 即判别两属性是否独立.**

- 例如: 色盲与其性别是否有关?
- 若考虑的属性多于两个, 称为多维列联表.

本次我们只讨论二维列联表, 列联表分析在应用统计, 特别在医学、生物学及社会科学中有广泛的应用.

列联表的独立性检验

一般的, 设 $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ 是总体 (X, Y) 的样本, 通过样本考虑二元总体 (X, Y) 中随机变量 X 与 Y 的独立性.

- 将随机变量 X 与 Y 的取值分为 r 个和 s 个互不相交的区间

$$A_1, A_2, \ldots, A_n \quad \text{和} \quad B_1, B_2, \ldots, B_n$$

- 设 $n_{i\cdot} = \sum_{j=1}^{s} n_{ij}$ 和 $n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$ 为行和或列和, 则 $n = \sum_{i,j} n_{ij}$, 用 n_{ij} 表示落入区域 $A_i \times B_j$ 的频数. 建立如下二维列联表:

	B_1	B_2	$\cdots$	B_s	$n_{i\cdot}$
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	n

列联表的独立性检验

下面介绍列联表的独立性检验法: 提出假设 H_0: X 与 Y 相互独立. 记

$$p_{ij} = \Pr(X \in A_i, Y \in B_j)$$

$$p_{i\cdot} = P(X \in A_i) = \sum_{j=1}^{s} p_{ij} \quad p_{\cdot j} = P(Y \in B_j) = \sum_{j=1}^{r} p_{ij}$$

若假设 H_0 成立, 则 $p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$, 利用矩估计/最大似然估计得

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}\,, \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}\,.$$

设计假设检验统计量

$$W = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}} = n\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - n \sim \chi^2((r-1)(s-1))$$

在显著性水平 α 时有 $W \sim \chi^2((r-1)(s-1))$ 成立, 得到拒绝域为 $W > \chi_\alpha^2((r-1)(s-1))$, 即在此范围内不接受随机变量 X 与 Y 相互独立.

假设检验：例 0.18

例 0.18 为研究儿童智力发展与营养得关系，某研究机构调查了 1436 名儿童，得到如下数据，试在显著性水平 0.05 下判断智力发展与营养有无关系.

	智商				合计
	< 80	$80 \sim 89$	$90 \sim 99$	≥ 100	
营养良好	367	342	266	329	1304
营养不良	56	40	20	16	132
合计	423	382	286	345	1436

解答：例 0.18

解答:

- 用 A 表示营养状况, 它有两个水平: A_1 表示营养良好, A_2 表示营养不良; B 表示儿童智商, 它有四个水平: B_1, B_2, B_3, B_4 分别表示表中四种情况. 建立假设 H_0: 营养状况与智商无关联, 即 A 与 B 是独立的, $H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$, 其中 $i = 1, 2$, $j = 1, 2, 3, 4$.
- 在原假设 H_0 成立下, 我们可以计算诸参数得最大似然估计值

$$\hat{p}_{1\cdot} = 1304/1436 = 0.9081, \quad \hat{p}_{2\cdot} = 132/1436 = 0.0919$$

$$\hat{p}_{\cdot 1} = 432/1436 = 0.2946, \quad \hat{p}_{\cdot 2} = 382/1436 = 0.266$$

$$\hat{p}_{\cdot 3} = 286/1436 = 0.1992, \quad \hat{p}_{\cdot 4} = 345/1436 = 0.2403$$

- 进一步算出

$$W = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(n_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}} = 19.2785$$

此处分布自由度为 $(2-1)(4-1) = 3$. 对 $\alpha = 0.05$, 查表得 $\chi^2_{0.05}(3) = 7.8147 < 19.2785$, 因此拒绝原假设 H_0.