假设检验

一、前言

看到假设检验，从字面意思理解，假设一个条件，检验是否成立，有没有想到->反证法，可以简单粗暴的理解为反证法，虽然有区别，但从步骤上说，很相似！

二、女士品茶

R.A.Fisher的名著《实验设计；第八版，1971年》第二部分有十六页，仅仅讲了一个最简单的实验：女士品茶。女士A声称可以判断在奶茶中，是先加入茶还是先加入牛奶。有人提议给她八杯奶茶。女士A已知其中四杯先加茶，四杯先加牛奶，但随机排列，而女士A要说出这八杯奶茶中，哪些先加牛奶，哪些先加茶，检验统计量是确认正确的次数。零假设是女士A无法判断奶茶中的茶先加入还是牛奶先加入，对立假设为女士A有此能力。
抛开严格的数学，先做一些直观的计算。也许该同事并没有任何分辨能力，仅凭运气，她也可能全部答对。随机地从8杯中选4杯“先奶后茶”，可能完全正确(k=4)；不过这个事件的概率 是

$$\frac{1}{\binom{8}{4}} = \frac{1}{70} = 0.014$$

这是一个小概率事件，概率小于0.05（通常的统计显著性水平）。所以，若是女士A全部答对，那么她 “没有任何分辨能力” 这个假设就和数据不太相容，可以拒绝这个假设。也许 Bristol 运气不够好，错选了 1 杯“先奶后茶”（k=3），这个事件的概率是

$$\frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{16}{70} = 0.229$$

这并不算一个小概率事件，即使女士A全凭运气蒙对3杯“先奶后茶”也无甚稀奇。从上面的简单计算看，只有当女士A完全答对的时候，我们才拒绝她“没有任何分辨顺序的能力” 这个假设，承认她有分辨能力。

通过该女士品茶，对假设检验有了一个印象。

三 、检验过程

在统计学的文献中，假设检验发挥了重要作用 。假设检验大致有如下步骤：

- 最初研究假设为真相不明 。
- 第一步是提出相关的零假设和对立假设 。这是很重要的，因为错误陈述假设会导致后面的过程变得混乱 。
- 第二步是考虑检验中对样本做出的统计假设；例如，关于独立性的假设或关于观测数据的分布的形式的假设 。这个步骤也同样重要，因为无效的假设将意味着试验的结果是无效的 。
- 决定哪个检测是合适的，并确定相关检验统计量 T 。
- 在零假设下推导检验统计量的分布 。在标准情况下应该会得出一个熟知的结果 。比如检验统计量可能会符合学生t-分布或正态分布 。
- 选择一个显著性水平 (α)，若低于这个概率阈值，就会拒绝零假设 。最常用的是 5% 和 1% 。
- 根据在零假设成立时的检验统计量T分布，找到数值最接近备择假设，且几率为显著性水平 (α)的区域，此区域称为“拒绝域”，意思是在零假设成立的前提下，落在拒绝域的几率只有α 。
- 针对检验统计量T，根据样本计算其估计值 t_{0bs} 。
- 若估计值 t_{0bs} 未落在“拒绝域”，接受零假设 。若估计值 t_{0bs} 落在“拒绝域”，拒绝零假设，接受对立假设 。

四 、在学习器的比较检验中，假设检验的应用

对于泛化错误率学习器是并不知道的，可以通过测试错误率估推出泛化错误率（从直观上看，二者的差异大的可能性比较小：可以这样解释，测试样本是泛化的采样，测试样本应符合泛化过程样本的分布） 的学习器在一个样本上犯错的概率

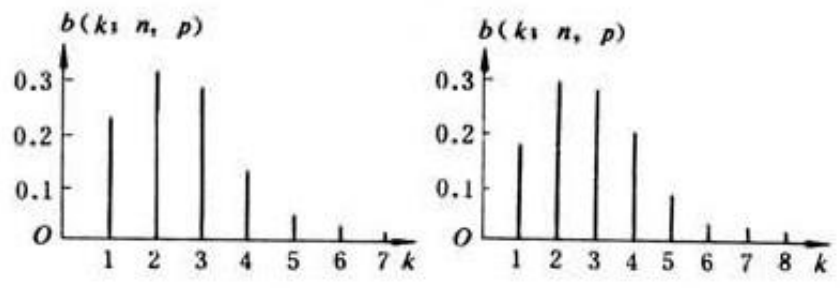
- ϵ ：泛化错误率为 ϵ 的学习器在一个样本上犯错的概率
- $\hat{\epsilon}$ ：泛化错误率为 ϵ 的学习器在一个测试样本上犯错的测试错误率
- 测试样本被误分类的个数: $\hat{\epsilon} \times m$
- 泛化错误率为 ϵ 的学习器将其中m'个样本误分类 、其余样本全部分类正确的概率: $\epsilon^{m'}(1-\epsilon)^{m-m'}$
- 由此可估算出其恰将 $\hat{\epsilon} \times m$ 个 样本误分类的概率如下式所示，这也表达了在包含m个样本的测试集上，泛化错误率为 ϵ 的学习器被测得测试错误率为 $\hat{\epsilon}$ 的概率: $P(\hat{\epsilon};\epsilon) = \begin{pmatrix} m \\ \hat{\epsilon} \times m \end{pmatrix} \epsilon^{\hat{\epsilon} \times m}(1-\epsilon)^{m-\hat{\epsilon} \times m}$ (说明： $\begin{pmatrix} m \\ \hat{\epsilon} \times m \end{pmatrix}$ -> $C_{\hat{\epsilon} \times m}^{m}$ -> $C_m^n = \frac{m!}{n!(m-n)!}$)

看到这里，这不是二项分布吗？

1.二项分布

我们在来看看二项分布：

- 定义：在概率论和统计学中，二项分布是n个独立的成功/失败试验中成功的次数的离散概率分布，其中每次试验的成功概率为p 。这样的单次成功/失败试验又称为伯努利试验 。实际上，当n=1时，二项分布就是伯努利分布 。一般地，如果随机变量服从参数为和的二项分布，我们记为 $X \sim B(n,p)$ 或 $X \sim b(n,p)$ 。n次试验中正好得到k次成功的概率由概率质量函数给出：
  $P\{X=k\}=\left(\begin{array}{c} n \\ k \end{array}\right) p^k(1-p)^{n-k}$
- 图形特点：从下图中可以看出，对于固定的n以及p，当k增加时，概率P{X=k}先是随之增加直至达到最大值，随后单调减少 。可以证明，一般的二项分布也具有这一性质，且:
  1.当（n+1）p不为整数时，二项概率P{X=k}在k=[(n+1)p]时达到最大值；
  2.当（n+1）p为整数时，二项概率P{X=k}在k=(n+1)p和k=(n+1)p-1时达到最大值 。



二项分布 $B(n,p)$ $(n=7,8;\ p=1/3)$

通过简要的二项分布定义到特点，上面的（看到这里，这不是二项分布吗）得到了确切的答案，就是它！ 在很多时候我们并非仅做一次留出法估计，而是通过多次重复留出法或是交叉验证法等进行多次训练/测试，这样会得到多个测试错误率，如何检验呢？这时候t-test就有了用武之地！

2.t检验（t-test）

- 适用条件 ：
  ⑴ 已知一个总体均数；
  ⑵ 可得到一个样本均数及该样本标准差；
  ⑶ 样本来自正态或近似正态总体
- 主要分类
  t检验可分为单总体检验和双总体检验，以及配对样本检验

（1）单总体检验（来源：百度百科）

单总体t检验是检验一个样本平均数与一个已知的总体平均数的差异是否显著。当总体分布是正态分布，如总体标准差未知且样本容量小于30，那么样本平均数与总体平均数的离差统计量呈t分布。

单总体t检验统计量为：

$$t = \frac{\overline{X} - \mu}{\frac{\sigma x}{\sqrt{n}}}$$

其中 $i = 1 \dots n, \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ 为样本平均数，$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$ 为样本标准偏差，n为样本数。该统计量t在零假说：$\mu=\mu_0$为真的条件下服从自由度为n的t分布。

（2）双总体检验（来源：百度百科）

双总体t检验是检验两个样本平均数与其各自所代表的总体的差异是否显著。双总体t检验又分为两种情况，一是独立样本t检验（各实验处理组之间毫无相关存在，即为独立样本），该检验用于检验两组非相关样本被试所获得的数据的差异性；一是配对样本t检验，用于检验匹配而成的两组被试获得的数据或同组被试在不同条件下所获得的数据的差异性，这两种情况组成的样本即为相关样本。

（1）独立样本t检验统计量为：

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

S_1^2和S_2^2为两样本方差；n_1和n_2为两样本容量。

（2）配对样本检验

配对样本t检验可视为单样本t检验的扩展，不过检验的对象由一群来自常态分配独立样本更改为二群配对样本之观测值之差。若二配对样本x_{1i}与x_{2i}之差为$d_i=x_{1i}-x_{2i}$独立，且来自常态分配，则d_i之母体期望值μ是否为μ_0可利用以下统计量：

$$t = \frac{\overline{d} - \mu_0}{s_d/\sqrt{n}}$$

其中 $i = 1 \dots n, \overline{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ 为配对样本差值之平均数，$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}}$ 为配对样本差值之标准偏差，n为配对样本数。该统计量t在零假说：$\mu=\mu_0$为真的条件下服从自由度为$n-1$的t分布。

（3）例子：

下面以一个实例的单总体t检验对t检验做一说明： [4]

问题：难产儿出生数n = 35，体重均值$\bar{x}$ = 3.42，S = 0.40，一般婴儿出生体重 μ_0= 3.30（大规模调查获得），问相同否？

解：

1.建立假设、确定检验水准α

H_0：μ = μ0 （零假设null hypothesis）

H_1：μ ≠ μ0（备择假设alternative hypothesis）

双侧检验，检验水准：α=0.05

2.计算检验统计量

$$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$$
$$= 1.77$$

$$v = n - 1 = 34$$

3.查相应界值表，确定P值，下结论。

查附表（图1），$t_{0.025/34}$ = 2.032, t < $t_{0.025/34}$, P >0.05，按α=0.05水准，不拒绝H0，两者的差别无统计学意义。

附表 t 界值表

自由度 ν	概率，P 单侧：0.25	0.05	0.025	0.01	0.005	0.0025
	双侧：0.50	0.10	0.05	0.02	0.010	0.0050
1	1.000	6.314	12.706	31.821	63.657	127.321
2	0.816	2.920	4.303	6.965	9.925	14.089
3	0.765	2.353	3.182	4.540	5.841	7.453
4	0.741	2.132	2.776	3.747	4.604	5.597
31	0.683	1.696	2.040	2.453	2.744	3.022
32	0.682	1.694	2.037	2.449	2.738	3.015
…	…	…	…	…	…	…
34	0.682	1.691	2.032	2.441	2.728	3.002
…	…	…	…	…	…	…
∞	0.6745	1.6449	1.9600	2.3263	2.5758	2.8070

图1t界值表

现在看看是否满足t-test条件：

- [x] (1) 已知一个总体均数；
- [x] (2) 可得到一个样本均数及该样本标准差；
- [x] (3) 样本来自正态或近似正态总体

完全符合条件！！！！

假定我们得到了 k个测试错误率，$\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_k$，则平均测试错误率 μ和方差 σ^2 为:

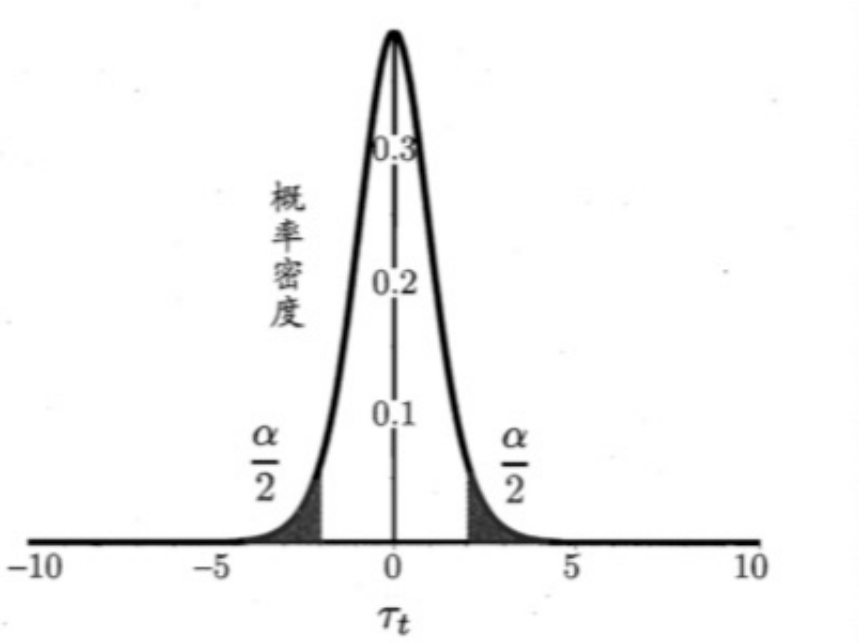
$\mu = \frac{1}{k}\sum_{i=1}^{k}\hat{\epsilon}_i$

$\sigma^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\hat{\epsilon}_i - \mu)^2$

考虑到这 k 个测试错误率可看作泛化错误率句的独立采样，则变量：

$\tau_t = \frac{\sqrt{k}(\mu - \epsilon_0)}{\sigma}$

服从自由度为 k-1 的 t分布.



对假设"$\mu = \epsilon_0$"和显著度α，我们可计算出当测试错误率均值为ϵ_0时，在 1-α 概率内能观测到的最大错误率，即临界值.这里考虑双边 (two-tailed)假设，如上图所示，两边阴影部分各有 α/2的面积;假定阴影部分范围分别为 $[-\infty, t_{-\alpha/2}]$ 和 [ta/2，∞]. 若平均错误率 μ 与 EO 之差 |μ- Eol 位于临界值范围 $[t_{\alpha/2}, \infty]$ 内，则不能拒绝假设"μ=ϵ_0。"，即可认为泛化错误率为ϵ_0，置信度为 1 — α; 否则可拒绝该假设，即在该显著度下可认为泛化错误率与ϵ_0有显著不 同.α 常用取值有 0.05和 0.1.