两组数值变量比较的假设检验

何平平

北京大学公共卫生学院流行病与卫生统计学系

t 检验的注意事项

1. 假设检验的目的

推断两个总体均数是否不同。

双侧检验H_1：$\mu_1 \neq \mu_2$？，单侧检验H_1：$\mu_1 > \mu_2$？ 或者$\mu_1 < \mu_2$？

为了稳妥起见，一般情况下多采用双侧检验。

2. 假设检验的P值不能反映总体均数差别的大小

差异有统计学意义时，P值越小，不能认为两总体均数差别越大；而是越有理由（越有把握）认为两总体均数不相等。

3. Z检验的应用

实际工作中，Z检验常用于“σ已知”的情况，也可以用于“σ未知，n较大”的情况 （此时，Z检验是t检验的近似）。

4. 假设检验方法的选择

根据不同的研究设计类型，选择不同的方法。

5. t检验与置信区间之间具有等价性

- 单样本的t检验：令$\alpha=0.05$，若接受H_0，则样本值的总体均数95%置信区间必包括μ_0；反之，若拒绝H_0，则95%置信区间必不包括μ_0。

- **两个独立样本的*t*检验**：令$\alpha=0.05$，若接受H_0，则两独立样本差值的总体均数95%置信区间必包括0；反之，若拒绝H_0，则95%置信区间必不包括0 。
- **配对设计的*t*检验**：令$\alpha=0.05$，若接受H_0，则配对差值的总体均数95%置信区间必包括0 ；反之，若拒绝H_0 ，则95%置信区间必不包括0 。

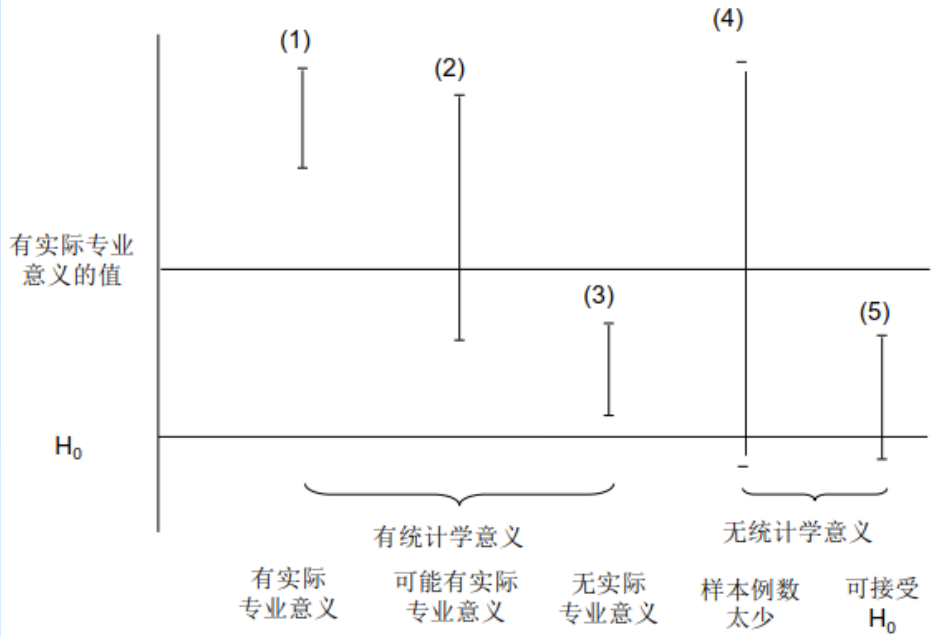

图5.5 置信区间在统计推断上提供的信息

与假设检验相比，置信区间可以提供更多的信息（如：是否差别具有实际意义，是否样本量不足）。因此国际上医学专业论文规定，在报告假设检验结论的同时，需要报告相应的置信区间。

6. 专业意义与统计学意义上的差别

差别有统计学意义，并不意味着一定有专业意义。只要样本量足够大，即便一个很小的差异，经统计学检验也能有统计学意义。反之，即便差别无统计学意义，但也可能具有专业意义。

例：美国的婴儿平均出生体重为120盎司。

- 某医院产科病房中收集正常分娩的10000例活婴的出生体重，其均值为119盎司，标准差为24盎司，能否认为此医院的婴儿平均出生体重低于全美国平均水平？

- 某医院产科病房中收集正常分娩的10例活婴的出生体重，其均值为110盎司，标准差为24盎司，能否认为此医院的婴儿平均出生体重低于全美国平均水平？

7. 假设检验的结论具有概率性

假设检验的结论不能绝对化，无论拒绝H_0或不拒绝H_0，都有犯错误的可能。

Ⅰ型错误（type Ⅰ error）：拒绝了实际上成立的H_0，这类“弃真”的错误称为Ⅰ型错误。犯Ⅰ型错误的概率是α。通常取α=0.05，其含义是当拒绝H_0时，理论上100次检验中平均有5次发生这样的错误。

II型错误（type II error）：接受了实际上是不成立的H_0，这类“存伪”的错误称为II型错误。犯II型错误的概率是β，一般情况下 β的大小是未知的。

假设检验结论可能发生的两类错误

客观实际	假设检验的结论	
	拒绝H_0	不拒绝H_1
H_0成立	I型错误（α）	推断正确（1-α）
H_0不成立，H_1成立	推断正确（1-β）	II型错误（β）

α和β的大小有一定关系：

当样本含量n确定时，α愈小，β愈大；反之，α愈大，β愈小。

当α一定时，增加样本含量，可以减少β。

8. 功效（Power）的定义　　功效又称为检验效能或把握度，是指当两总体确实有差别时，按规定的检验水准α，能够发现两总体间差别的能力，即$1-\beta$。例如：$1-\beta=0.8$，意味着如果两总体确实有差别，则理论上100次检验中，平均有80次能够得出有差别的结论。 实际工作中，要保证比较高的功效，很重要的条件是具有足够的样本含量(样本量的计算参见调查设计与实验设计的章节)。

9. 假设检验的前提 —— 可比性

组间比较时应具备可比性，即除了处理因素外，其它可能影响结果的非处理因素在各组间应该尽可能相同或相近。例如：比较某地区城市和农村成人的身高是否有差异。研究关心的因素是地区（城市或农村），但是其他因素也可能对身高有影响，如年龄、性别。只有当城市和农村两组间年龄、性别情况相同或相近时，才能比较它们的身高是否有差异。