假设检验

基本思想

统计，终归是会有误差的。假设检验的根本便是衡量误差出现的概率，基于**小概率事件原理**判断与需要检验的假设不一致的数据究竟更可能是误差导致，还是假设本身即不成立

用于衡量误差的参数通常是**平均数**

平均数

单个样本

u 检验

u 检验基于 u 分布，而 u 分布实际上就是标准正态分布 $N(0,1)$。

想象一下，如果我们有一组样本和一个整体，若样本来自于整体抽样，平均数应该会有误差，而且显然应满足**正态分布**，误差越大概率越小。于是乎，基于正态分布，我们可以轻松计算出某个大小误差出现的概率。

基于高中知识，可以知道，任意的正态分布可以标准化为标准正态分布。

$$x \sim N(\mu,\sigma^2) \Rightarrow \frac{(x-\mu)}{\sigma} \sim N(0,1)$$

抽取多个样本的平均数也应当符合某个正态分布，即

$$\bar{x} \sim N(\mu_{\bar{x}},\sigma_{\bar{x}}^2)$$

显然，$\mu_{\bar{x}} = \mu$；样本容量若为 n，每个元素出现在样本中的概率为 $1/n$，可以用其出现的期望 x_i/n 计算方差，得到 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

于是，可以标准化该正态分布，得到

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

标准化后即为 u 分布，可以计算每个样本平均数出现的概率，并和给定的显著性水平进行比较，**u 检验**由此诞生了

通常情况下，若 $R(u) > 0.05$（单尾）或 $R(u) > 0.1$（双尾）即视作有显著差异

t 检验

有了 u 检验，为什么还有 t 检验？因为**总体方差**不是什么时候都知道的。样本容量大时，可以用 s^2 较好的估计 σ^2，但**样本容量小**可就不是了；t 检验即是在这种情况下的校正

t分布

同样的，我们可以用 u 检验的方法计算，不同的是，此处计算出的不是 $\sigma_{\bar{x}}$，而是 $s_{\bar{x}}$。于是有

$$\frac{\bar{x}-\mu}{s_{\bar{x}}/\sqrt{n}}=t$$

t 服从于 t 分布，同样可以计算累加概率

成组数据

成组 u 检验

同样的，对于成组数据之间的比较，也可以使用 u 检验。显然注意到若两个样本来自同一个群体，应有 $\bar{x}_1-\bar{x}_2\sim N(0,\sigma^2_{\bar{x}_1-\bar{x}_2})$

由高中知识可知

$$\begin{aligned}\sigma^2_{\bar{x}_1-\bar{x}_2}&=\sigma^2_{\bar{x}_1}+\sigma^2_{\bar{x}_2}\\&=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\end{aligned}$$

于是乎，参照单样本平均数的 u 检验

$$u=\frac{(x_1-x_2)-0}{\sigma_{\bar{x}_1-\bar{x}_2}}$$

成组 t 检验

首先，很重要的一点是，t 检验中使用的不是**总体方差**，而是对总体方差的不太好的估计——**样本方差**。

由高中知识

$$s^2_{\bar{x}_1-\bar{x}_2}=\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}$$

如果两个方差具有**同质性**，则估计应该是准确的，否则一大一小，估计的误差会很大。此时不能完成标准的 t 检验，只能做**近似 t 检验**

因此，第一步必须是检验方差同质性

F 检验

显然，如果方差有同质性，$\frac{s_1^2}{s_2^2}$ 与 1 不会有很大的偏离。于是，定义 $F=\frac{s_1^2}{s_2^2}$，找到 F 的分布（显然与两个样本的自由度相关），就可以检验方差同质性。

方差同质

此时对总体方差的估计较好，可以参照单个样本的 t 检验进行计算

$$t=\frac{(\bar{x}_1-\bar{x}_2)-0}{s_{\bar{x}_1-\bar{x}_2}}$$

方差不同质

硬要算也能算，但是我们需要进行一些校正，此时仍然有

$$t=\frac{(\bar{x}_1-\bar{x}_2)-0}{s_{\bar{x}_1-\bar{x}_2}}$$

但是另有

$$df'=\frac{1}{\frac{R^2}{n_1-1}+\frac{(1-R)^2}{n_2-1}}$$
$$R=\frac{s_{\bar{x}_1}^2}{s_{\bar{x}_1}^2+s_{\bar{x}_2}^2}$$

对于该校正的解释超出了本文档的知识水平，因此不过多阐释

样本频率

对样本频率有着很特别的一点，

卡方检验

卡方

如果有 $Z_i(1\le i\le n)\sim N(0,1)$，那么可以定义 $\chi^2=\sum Z_i^2$，且 $\chi^2\sim\chi^2(n)$

可以想象，如果 χ^2 过大，说明这组数据可能有问题，即可能不服从该正态分布

方差齐性卡方检验

该检验可以用来检验样本与总体的方差是否具有齐性。

样本元素 $x_i\sim N(\bar{x},s^2)$，如果二者方差具有齐性，则以总体方差对它进行标准化，理应也符合标准正态分布。所以

$$\frac{x_i-\bar{x}}{\sigma}\sim N(0,1)$$

又如何检验呢？当然是搬出 χ^2 啦

$$\chi^2=\sum(\frac{x_i-\bar{x}}{\sigma})^2=\frac{\sum(x_i-\bar{x})^2}{\sigma^2}=\frac{(n-1)s^2}{\sigma^2}$$

独立性卡方检验

啊哈，符合正态分布的又不止样本元素，在成组 u 检验中，我们还了解到两个元素的差值也能符合正态分布。那么，如果是**期望与观测的差值**呢？当然。

由**中心极限定理**（此处不细讲）

$$O \sim N(E, E)$$

则标准化后，我们依然能够计算 χ^2

$$\chi^2 = \sum(\frac{O-E}{\sqrt{E}})^2 = \sum \frac{(O-E)^2}{E}$$

连续性校正

在对 2×2 列联表进行检验时，有一个点非常需要注意：通常这类表的数据是**离散的**（例如频数），但 χ^2 是定义在**连续的**正态分布上的，这会使 χ^2 的数值更激进。而小的自由度又会放大这种误差。因此需要一项连续性校正。此时

$$\chi^2 = \sum \frac{(|O-E|-0.5)^2}{E}$$

这会使χ^2更加保守，减小这种误差