常用检验

司继春

[1]上海对外经贸大学

2025年12月

概览

均值与方差的比较

- 在实验等应用中，经常遇到比较均值以及比例的问题。
- 一般的，我们将区分大样本和小样本两种情况进行讨论。
- 在独立大样本的情况下，根据中心极限定理，样本均值渐近服从正态分布，我们可以使用该性质得到检验统计量的渐近分布，并使用该渐近分布进行假设检验
- 然而在小样本情况下，大数定律、中心极限定理等工具不再适用，此时我们必须对分布做比较严格的假定。

配对样本比较

- 如果我们有样本$(x_i, y_i), i = 1, ..., N$，注意对于每个个体我们可以观察到两个变量：$x_i, y_i$，两个变量是天然成对出现的
  - 比如夫妻双方的年龄、收入，或者学生的入学、毕业考试成绩等。
- 如果我们需要比较μ_x和μ_y，那么可以计算：$d_i = x_i - y_i$从而得到了一组新的变量d_i，如果原假设为$H_0 : \mu_x = \mu_y$等价于

$$H_0 : \mu_d = 0$$

- 我们可以使用上一章中的总体均值的检验，检验统计量为

$$\frac{\sqrt{N}\left(\bar{d} - 0\right)}{s_d}$$

在小样本正态总体、大样本的条件下，分别服从$\mathrm{t}\left(N-1\right)$或者$\mathcal{N}\left(0,1\right)$，使用标准的检验步骤即可。

配对样本比较

夫妻教育年限的检验

在数据集soep_female_labor.dta数据集中记录了在SOEP中德国女性的生育、就业、受教育成都等状况。变量edu为女性的教育年限，而husedu为丈夫的教育年限。

- 如果我们需要比较丈夫和妻子的受教育程度是否有区别，可以首先计算丈夫与妻子教育年限的差，计算得到丈夫、妻子的教育年龄差的均值$\bar{d} = 0.201$，其标准差为$s_d = 2.504$，样本量$N = 6705$

配对样本比较

夫妻教育年限的检验

- 为了检验原假设：$H_0 : \text{edu}_{husband} = \text{edu}_{wife}$可以计算：

$$\frac{\sqrt{N}\left(\bar{d} - 0\right)}{s_d} = 6.57 > 1.96$$

从而可以在5%的显著性水平下拒绝原假设，认为丈夫和妻子的教育程度有区别。

- 如果原假设为：$H_0 : \text{edu}_{husband} \leq \text{edu}_{wife}$那么该检验变成了一个右侧检验，由于6.57>1.65，同样可以拒绝原假设，可以认为丈夫的教育年限大于妻子的教育年限。

配对样本比较

夫妻教育年限的检验

在Stata中，也可以不手动去计算d_i，而是直接使用如下命令进行检验：

```
use datasets/soep_female_labor.dta, clear
// 进行检验t
ttest husedu == edu
```

小样本正态总体的均值比较

- 如果需要检验的两个总体均值没有天然的配对，那么以上过程就不再适用了，此时我们需要根据样本均值的差构建新的检验统计量。
- 如果假设样本$\boldsymbol{x^{(1)}} = \left[x_1^{(1)}, ..., x_{N_1}^{(1)}\right]'$以及$\boldsymbol{x^{(2)}} = \left[x_1^{(2)}, ..., x_{N_2}^{(2)}\right]'$，且$\boldsymbol{x^{(1)}}$与$\boldsymbol{x^{(2)}}$独立，两个样本分别来自与两个可能不同的正态总体，
  即：$x_i^{(1)} \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)\ i.i.d$，$x_i^{(2)} \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)\ i.i.d$
- 那么样本均值：

$$(\bar{x}_1 - \mu_1) \sim \mathcal{N}\left(0, \frac{\sigma_1^2}{N_1}\right)$$

$$(\bar{x}_2 - \mu_2) \sim \mathcal{N}\left(0, \frac{\sigma_2^2}{N_2}\right)$$

从而根据独立性，$\mathbb{V}(\bar{x}_1 - \bar{x}_2) = \mathbb{V}(\bar{x}_1) + \mathbb{V}(\bar{x}_2)$，从而有

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \sim \mathcal{N}(0, 1)$$

小样本正态总体的均值比较

- 然而为了进行假设检验，其中的σ_1^2以及σ_2^2是未知的，一个自然的想法是使用样本方差s_1^2以及s_2^2进行代替。
- 回忆单样本情形时如果$x_i \sim \mathcal{N}\left(\mu, \sigma^2\right) i.i.d$，抽样分布

$$\frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}} \sim \mathrm{t}\left(N-1\right)$$

这一结论建立在$\frac{(N-1)s^2}{\sigma^2} \sim \chi^2\left(N-1\right)$这一结论基础上的。

- 而在现在两个样本的情形，对于$s_1^2/N_1 + s_2^2/N_2$的分布并不能简单类比，此时我们需要分两种情况讨论

小样本正态总体均值比较：总体方差相等

- 假设$\sigma_1^2 = \sigma_2^2 = \sigma^2$，即两个总体的总体方差相等，都等于$\sigma^2$。
- 在这个假设条件下，理论上两个样本计算出的样本方差应该几乎相等，即$s_1^2 \approx s_2^2$，然而由于抽样误差的存在，两者应该不相等。
- 那么一个自然的想法是，我们能不能结合两个样本的信息，计算他们共同的总体方差呢?

小样本正态总体均值比较：总体方差相等

- 注意到：

$$\mathbb{E}\left(s_j^2\right)=\mathbb{E}\left(\frac{\sum_{i=1}^{N_j}\left(x_i^{(j)}-\bar{x}_j\right)}{N_j-1}\right)=\sigma_j^2=\sigma^2, j=1,2$$

从而：

$$\mathbb{E}\left(\left(N_j-1\right)s_j^2\right)=\left(N_j-1\right)\sigma^2, j=1,2$$

- 两个等式相加，得到：

$$\mathbb{E}\left[\left(N_1-1\right)s_1^2+\left(N_2-1\right)s_2^2\right]=\left(N_1+N_2-2\right)\sigma^2$$

从而对于σ^2一个自然的估计为：

$$\hat{\sigma}^2=\frac{\left(N_1-1\right)s_1^2+\left(N_2-1\right)s_2^2}{N_1+N_2-2}$$

小样本正态总体均值比较：总体方差相等

注意到（根据独立性）

$$\begin{aligned}\frac{(N_1+N_2-2)\,\hat{\sigma}^2}{\sigma^2} &= \frac{(N_1-1)\,s_1^2}{\sigma^2}+\frac{(N_2-1)\,s_2^2}{\sigma^2}\\ &\sim \chi^2(N_1-1)+\chi^2(N_2-1)\\ &\sim \chi^2(N_1+N_2-2)\end{aligned}$$

小样本正态总体均值比较：总体方差相等

使用$\hat{\sigma}^2$代替式σ_1^2, σ_2^2，得到

$$
\begin{aligned}
\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\hat{\sigma}^2}{N_1}+\frac{\hat{\sigma}^2}{N_2}}} &= \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}} \\
&= \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}} \\
&\quad \cdot \frac{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}}{\sqrt{\frac{(N_1+N_2-2)\hat{\sigma}^2}{\sigma^2}\frac{\sigma^2}{(N_1+N_2-2)}\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}} \\
&= \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}} \frac{1}{\sqrt{\frac{(N_1+N_2-2)\hat{\sigma}^2}{\sigma^2}/(N_1+N_2-2)}} \\
&\sim \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(N_1+N_2-2)}{N_1+N_2-2}}} \\
&\sim \mathrm{t}(N_1+N_2-2)
\end{aligned}
$$

小样本正态总体均值比较：总体方差不相等

- 如果假设$\sigma_1^2 \neq \sigma_2^2$，此时只能使用两个样本方差$s_1^2$、$s_2^2$分别代替$\sigma_1^2$和$\sigma_2^2$
- 此时

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \sim \mathrm{t}(\upsilon)$$

其中

$$\upsilon = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(s_1^2/N_1\right)^2}{N_1 - 1} + \frac{\left(s_2^2/N_2\right)^2}{N_2 - 1}}$$

小样本正态总体均值比较

- 根据以上结论，在原假设：$H_0:\mu_1-\mu_2=(\geq/\leq)c$的条件下，可以相应构造检验统计量。
- 比如，如果假设检验的原假设为：$H_0:\mu_1=\mu_2$，那么将$\mu_1-\mu_2=0$带入，得到检验统计量为

$$\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}}\sim \mathrm{t}(v)$$

如果可以假设$\sigma_1^2=\sigma_2^2$，检验统计量为

$$\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\hat{\sigma}^2\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}\sim \mathrm{t}(N_1+N_2-2)$$

进行相应t检验即可。

大样本的均值比较

- 大样本情况下的处理要简单很多。假设样本$\boldsymbol{x^{(1)}}=\left[x_1^{(1)},...,x_{N_1}^{(1)}\right]'$以及$\boldsymbol{x^{(2)}}=\left[x_1^{(2)},...,x_{N_2}^{(2)}\right]'$，且$\boldsymbol{x^{(1)}}$与$\boldsymbol{x^{(2)}}$独立，那么只要$\sigma_1^2<\infty$，$\sigma_2^2<\infty$，根据中心极限定理：

$$\sqrt{N_j}\left(\bar{x}_j-\mu_j\right)\overset{a}{\sim}\mathcal{N}\left(0,\sigma_j^2\right),j=1,2$$

- 计算样本均值之差并标准化：

$$\frac{\left(\bar{x}_1-\bar{x}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}}\overset{a}{\sim}\mathcal{N}\left(0,1\right)$$

- 而根据Slutsky定理，由于$s_j^2=\sigma_j^2+o_p\left(1\right),j=1/2$，分母上可以使用样本方差代替总体方差，并不改变其渐近分布：

$$\frac{\left(\bar{x}_1-\bar{x}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}}\overset{a}{\sim}\mathcal{N}\left(0,1\right)$$

大样本的均值比较

- 注意到，由于t分布当自由度趋向于无穷时，会收敛到标准正态分布
- 同时以上大样本条件下的检验统计量与小样本正态总体的检验统计量是相同的
- 当$N_1 \to \infty, N_2 \to \infty$时，$\upsilon \to \infty$，从而大样本条件下，使用$\mathrm{t}(\upsilon)$和使用$\mathcal{N}(0,1)$是几乎等价的
- 因而在实践中，即使大样本，我们通常也使用t检验。

大样本的均值比较

CFPS中男女收入的检验

我们使用cfps_adult.dta数据比较男性和女性的收入差别。

- 注意此时我们并没有将男性和女性进行配对，不存在天然的丈夫-妻子关系，实际上当我们考虑男性、女性的收入差别时，当然需要考虑所有人，包括未婚、离异等情况。
- 根据描述性统计，女性收入均值为5750.946，标准差为14233，样本量为18308；而男性收入均值为12286.49，标准差为22120.62，样本量为18398，那么
  在$H_0 : \mu_1 - \mu_2 = (\geq / \leq)0$的条件下，检验统计量为：

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} = \frac{-6535.544}{\sqrt{37661.483}} = -33.68$$

从而在$H_0 : \mu_1 \geq (=)\mu_2$，即女性收入大于等于男性收入的原假设下，可以拒绝原假设

大样本的均值比较

CFPS中男女收入的检验

在Stata中，同样可以使用ttest命令进行以上检验：

```
use datasets/cfps_adult.dta, clear
// 进行检验，由于收入t的样本几乎为缺失数据，将其排除<0
ttest p_income if p_income>=0, by(cfps_gender)
```

正态总体方差的比较

- 除了比较两个样本的均值之外，还可以比较两组样本的方差。
- 如果样本$x_i^{(1)} \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)\ i.i.d, i = 1, ..., N_1$，$x_i^{(2)} \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)\ i.i.d, i = 1, ..., N_2$，且两个样本独立。
- 为了检验

$$H_0 : \sigma_1^2 = \sigma_2^2$$

在H_0的假定下，

$$\frac{\frac{(N_1-1)s_1^2}{\sigma_1^2} / (N_1 - 1)}{\frac{(N_2-1)s_2^2}{\sigma_2^2} / (N_2 - 1)} = \frac{s_1^2}{s_2^2} \sim \mathrm{F}\left(N_1 - 1, N_2 - 1\right)$$

因而其拒绝域为$R_\alpha = \left(0, \mathrm{F}_{1-\alpha/2}^{(N_1-1,N_2-1)}\right) \cup \left(\mathrm{F}_{\alpha/2}^{(N_1-1,N_2-1)}, \infty\right)$。

方差的比较

- 然而注意，以上检验方法严重依赖于两组样本的正态总体假设。
- 如果两样本至少有一个不服从正态分布，那么以上检验是失效的。
- 此时，可以考虑使用对正态性假设更不敏感的Levene检验以及Brown-Forsythe检验。
- 篇幅所限，我们这里不再赘述，Stata中可以使用robvar命令进行以上两个检验。

方差的比较

CFPS中男女收入方差的比较

- 上例中的收入数据一般是不符合正态分布的，因而直接使用如上介绍的F检验可能会有问题
- 可以考虑使用Levene检验以及Brown-Forsythe检验

方差的比较

CFPS中男女收入方差的比较

```
use datasets/cfps_adult.dta, clear
// 进行检验，假设正态总体（其实并不成立）F
sdtest p_income if p_income>=0, by(cfps_gender)
// 进行检验以及LeveneBrown-检验Forsythe
robvar p_income if p_income>=0, by(cfps_gender)
```

其中robvar命令中汇报的W0为Levene检验，而W50和W10为两种形式的Brown-Forsythe检验，所有的检验都拒绝了原假设，即可以认为男性和女性收入的方差不相等。

拟合优度检验

- 现在一个抛硬币游戏的例子。我们知道，如果硬币是均匀的，那么得到正面的概率应该为$p = 0.5$。
- 然而，现实中我们可能会怀疑该硬币并非均匀，使用假设检验的思路，我们的原假设为：

$$H_0 : p = 0.5$$

为了检验该原假设，我们可以重复抛硬币N次，并记录为正面的次数N_1，得到正面的频率为：$\hat{p} = {}^{N_1}/_{N}$。

- 进而，在原假设成立的条件下，检验统计量为：

$$\frac{\hat{p} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{N}}} \overset{a}{\sim} \mathcal{N}(0, 1)$$

拟合优度检验

检验硬币的均匀性

如果抛了100次硬币，其中有60次是正面，从而$\hat{p} = 0.6$，检验统计量：

$$\frac{0.6 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = 2 > 1.96$$

从而在5%的显著性水平下可以拒绝原假设，认为硬币并非均匀的。

拟合优度检验

- 更一般的情况，对于某一个分类变量x_i只有可能取有限个值：$x_i \in \{1, 2, ..., G\}$，假设理论上取每个值的概率为：$\mathrm{P}(x_i = g) = p_g, g = 1, 2, ..., G$，其中$p_1 + \cdots + p_G = 1$。
- 为了检验样本数据x是否符合如上的理论分布，我们可以建立原假设：

$$H_0 : \begin{cases} \mathrm{P}(x_i = 1) = p_1 = p_1^* \\ \vdots \\ \mathrm{P}(x_i = G) = p_G = p_G^* \end{cases}$$

  在以上原假设中，总共有G个命题，但是由于$p_1 + \cdots + p_G = 1$，实际上只需要使用$G - 1$个命题作为原假设就可以。
- 由于以上检验实际上检验了概率分布$(p_1^*, ..., p_G^*)$是否拟合了客观数据，这一类检验通常也称为**拟合优度检验**（goodness of fit test）。

拟合优度检验

- 比如，在抛硬币的例子中，如果记正面为1反面为0，那么原假设应为：

$$H_0 : \begin{cases} p_1 = 0.5 \\ p_2 = 0.5 \end{cases}$$

  但是由于$p_1 + p_2 = 1$，从而以上原假设只需要假设：$H_0 : p_1 = 0.5$就可以。
- 如果我们希望检查一个骰子六个面的概率是否都是$1/6$，我们可以分别统计出抛出g的比率$\hat{p}_g$，并分别与理论概率$1/6$对比，即在

$$H_0 : p_g = 1/6$$

  的条件下进行假设检验，从而我们可以得到6个假设检验
  - 当然，由于概率之和为1，实际上只需要5个就可以了。

拟合优度检验

- 然而以上检验有重大的弊端。我们知道每次进行假设检验时，都以α的概率犯第I类错误，意味着如果原假设成立，每做一次假设检验，都有可能以α的概率错误拒绝原假设。
- 如果我们进行5次以上的假设检验，只要有一次拒绝原假设，看起来“骰子是均匀的”这一命题就不成立，然而这里需要注意的是，5次假设检验都不犯错的概率为$(1-\alpha)^5$，从而犯错的概率为$1-(1-\alpha)^5$，如果$\alpha = 5\%$，那么犯错的概率高达22.6%
- 为了克服这一问题，我们肯定希望在α的显著性水平下，进行一次假设检验即可。此时，我们需要构造一个全新的检验统计量。

G检验

- 首先我们可以使用似然比检验。对于分类变量$x_i \in \{1, 2, ..., G\}$，假设理论上取每个值的概率为：$\mathrm{P}(x_i = g) = p_g, g = 1, 2, ..., G$，其联合密度函数为：

$$f(\boldsymbol{x}|p) = \prod_{i=1}^{N} p_1^{\mathbb{1}\{x_i=1\}} \cdot p_2^{\mathbb{1}\{x_i=2\}} \cdots p_G^{\mathbb{1}\{x_i=G\}}$$

从而对数似然函数为：

$$\ln L(p|\boldsymbol{x}) = \sum_{i=1}^{N} \sum_{g=1}^{G} \mathbb{1}\{x_i = g\} \cdot \ln(p_g)$$

- 求以上对数似然函数最大化，得到最优解为：$\hat{p}_g = {}^{N_g}/_{N}$，其中$N_g$为$N$个样本中值为$g$的样本个数。

G检验

- 以上为不受限的极大似然估计。而如果在原假设：

$$H_0: \begin{cases} \mathrm{P}(x_i = 1) = p_1 = p_1^* \\ \vdots \\ \mathrm{P}(x_i = G-1) = p_{G-1} = p_{G-1}^* \end{cases}$$

的（$G-1$个）约束下似然函数值为：

$$\ln L(p^*|x) = \sum_{i=1}^{N} \sum_{g=1}^{G} \mathbb{1}\{x_i = g\} \cdot \ln\left(p_g^*\right)$$

G检验

- 从而检验统计量为：

$$
\begin{aligned}
LR &= 2\left[\sum_{i=1}^{N}\sum_{g=1}^{G}\mathbb{1}\left\{x_i=g\right\}\cdot\ln\left(\hat{p}_g\right)-\sum_{i=1}^{N}\sum_{g=1}^{G}\mathbb{1}\left\{x_i=g\right\}\cdot\ln\left(p_g^*\right)\right] \\
&= 2\sum_{i=1}^{N}\sum_{g=1}^{G}\mathbb{1}\left\{x_i=g\right\}\cdot\ln\left(\frac{\hat{p}_g}{p_g^*}\right) \\
&= 2\sum_{g=1}^{G}N_g\ln\left(\frac{\hat{p}_g}{p_g^*}\right) \\
&= 2\sum_{g=1}^{G}N_g\ln\left(\frac{N_g}{N_g^*}\right) \\
&\overset{a}{\sim}\chi^2\left(G-1\right)
\end{aligned}
$$

其中$N_g^*=N\times p_g^*$为理论的频数（可以为小数）。

G检验

以上检验统计量：

$$G = 2\sum_{g=1}^{G} N_g \ln\left(\frac{N_g}{N_g^*}\right) \overset{a}{\sim} \chi^2\left(G-1\right)$$

通常被称为G检验（G-test）。

Pearson的卡方检验

- 或者，一个更加常见的检验为**Pearson的卡方检验（Pearson’s chi-squared test）**，其检验统计量为:

$$\chi^2 = \sum_{g=1}^{G} \frac{\left(N_g - N_g^*\right)^2}{N_g^*} \overset{a}{\sim} \chi^2\left(G-1\right)$$

同样，该检验也是一个右侧检验。

- 可以证明，Pearson的卡方检验与G检验是渐近等价的（证明见讲义）
- 实践中Pearson的卡方检验更常见

拟合优度检验

本福特定律

一个常见的用于鉴别数据造假的方法是使用**本福特定律**（Benford’s law），该定律指出，在自然界中存在的数据跨度足够大、没有明显人为规则的数字，其最高位的数字出现的概率并不是等概率的，而是满足：

$$P(d) = \log_{10}(d+1) - \log_{10}(d)$$

即对于十位数，首位数字取值的概率大概如下表所示：

首位数字	1	2	3	4	5	6	7	8	9
概率	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

拟合优度检验

本福特定律

- 该规律背后的直觉是，考虑从0出发一直往∞增加，当数字的位数增加到很大时，那么任意时间喊停，停下时更容易留在首位数字为1的地方，因为比如同样是6位数，数字一定是先经过100000-199999这些数字，然后才会跳到以2为首位的六位数字，而相比之下，1位数字到5位数字总共才有99999个，而以1开头的六位数字同样有99999个，所以任意时间喊停，那么数字首位数为1的概率更大，其次是2,3,4,...。
- 比如，当我们为一个城市所有的人排序给一个编号，比如上海总共有2400万人，在这2400万人里面，首位数字为1的是编号为1、1*、1**等等，直到第1000万到2000万-1这(1000万-1)人个人的首位数字都是1，那么可想而知，如果我们从中抽出一个样本的话，首位数字为1的概率应该大于其他数字的概率。

拟合优度检验

本福特定律

比如，代码Benford.do使用该定律检查了我国城市统计年鉴中2011年每个城市的GDP数据是否符合本福特定律,计算得到p值为0.066，在5%的显著性水平下不显著，即认为在该显著性水平下，不能拒绝该数据服从本福特定律的原假设。

独立性检验

- 现在考虑两个离散变量$x_i \in \{1, 2, ..., G\}, y_i \in \{1, 2, ..., H\}, i = 1, 2, ..., N$
- 两个离散变量之间的关系可以使用概率函数$p_{gh} = \mathrm{P}(x_i = g, y_i = h)$来描述。
- 我们知道，如果两个变量之间独立，那么必然有

$$p_{gh} = \mathrm{P}(x_i = g, y_i = h) = \mathrm{P}(x_i = g) \cdot \mathrm{P}(y_i = h) = p_{g\cdot} \cdot p_{\cdot h}$$

其中$p_{g\cdot} = \mathrm{P}(x_i = g), p_{\cdot h} = \mathrm{P}(y_i = h)$。

独立性检验

- 如果我们记$G \times H$维矩阵$P = [p_{gh}]$，那么有$[p_{g\cdot}] = P\iota_H$以及$[p_{\cdot h}] = P'\iota_G$
- 从而上述独立性的条件可以写为

$$P = P\iota_H\iota_G'P$$

。

- 此外，由于：$\sum_{g=1}^{G}\sum_{h=1}^{H} p_{gh} = 1$或者使用矩阵表示为$\iota_G'P\iota_H = 1$。
- 接下来我们尝试构造在原假设：

$$H_0 : P = P\iota_H\iota_G'P$$

下的检验统计量。该原假设成立时，两个离散随机变量x_i, y_i是独立的，否则是不独立的。

独立性检验

- 我们首先可以使用LR检验。
- 在无约束的条件下，对数似然函数为：

$$L\left(P|x,y\right)=\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{h=1}^{H}\mathbb{1}\left\{x_i=g,y_i=h\right\}\cdot\ln\left(p_{gh}\right)$$

显然：$\hat{p}_{gh}=\frac{N_{gh}}{N}$其中$N_{gh}=\#\left\{x_i=g,y_i=h\right\}$。

- 此时我们需要估计的参数的个数为$G\times H-1$（由于所有概率加起来必须为1）。

独立性检验

- 而如果给定独立性的约束，实际上我们只需要$p_{g\cdot}, p_{\cdot h}$即可
- 此外由于$\sum_{g=1}^{G} p_{g\cdot} = \sum_{h=1}^{H} p_{\cdot h} = 1$，从而我们只需要估计$(G-1)+(H-1)$个参数即可
- 此时对数似然函数为：

$$L\left(P|x,y\right) = \sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{h=1}^{H} \mathbb{1}\left\{x_i = g, y_i = h\right\} \cdot \left[\ln\left(p_{g\cdot}\right) + \ln\left(p_{h\cdot}\right)\right]$$

- 最大化以上对数似然函数，得到：

$$\begin{cases} p_{g\cdot} = \frac{N_g}{N} \\ p_{\cdot h} = \frac{N_h}{N} \end{cases}$$

其中$N_g = \#\left\{x_i = g\right\}, N_h = \#\left\{y_i = h\right\}$。

独立性检验

- 注意在独立性的假定下，待估参数的个数从$G \times H - 1$减少到了$(G-1)+(H-1)$个
- 从而潜在的约束个数为：

$$(G \times H - 1) - [(G-1)+(H-1)] = (G-1) \times (H-1)$$

独立性检验

- 从而似然比检验统计量为:

$$
\begin{aligned}
LR &= 2\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{h=1}^{H}\mathbb{1}\left\{x_i=g,y_i=h\right\}\cdot\ln\left(\frac{N_{gh}}{N}\right)\\
&\quad-2\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{h=1}^{H}\mathbb{1}\left\{x_i=g,y_i=h\right\}\cdot\left[\ln\left(\frac{N_g}{N}\right)+\ln\left(\frac{N_h}{N}\right)\right]\\
&= 2\left[\sum_{g=1}^{G}\sum_{h=1}^{H}N_{gh}\ln\left(\frac{N_{gh}}{N}\right)-\sum_{g=1}^{G}N_g\ln\left(\frac{N_g}{N}\right)-\sum_{h=1}^{H}N_h\ln\left(\frac{N_h}{N}\right)\right]\\
&\overset{a}{\sim}\chi^2\left((G-1)\times(H-1)\right)
\end{aligned}
$$

以上即独立性检验的LR检验。

独立性检验

与拟合优度检验类似，更常用的是LR检验的渐近等价形式，即Pearson的卡方检验的类似形式进行检验，其检验统计量为：

$$\sum_{g=1}^{G}\sum_{h=1}^{H}\frac{\left(N_{gh}-N_{gh}^{*}\right)^{2}}{N_{gh}^{*}}\overset{a}{\sim}\chi^{2}\left((G-1)\times(H-1)\right)$$

其中：

$$N_{gh}^{*}=N\times\hat{p}_{g\cdot}\times\hat{p}_{\cdot h}=\frac{N\times N_{g}\times N_{h}}{N\times N}=\frac{N_{g}\times N_{h}}{N}$$

为在独立性假定下应该得到的频数。

独立性检验

泰坦尼克号

在泰坦尼克号海难中，我们想知道成年男性的死亡率是否与女性、儿童的死亡率有显著的不同，如果记$x_i = 1, 2, 3$代表儿童、成年男性、成年女性，而$y_i = 1, 2$分别代表死亡、幸存，我们可以使用如上检验方法。

- 数据集titanic.dta记录了泰坦尼克死亡的数据，Titanic.do根据以上公式计算检验统计量。
- 由于性别有3个类别，死亡与否有2个类别，从而自由度为2
- 检验统计量为481.59，计算得到的p值可以发现拒绝了原假设，即死亡与否与性别是显著相关的。

独立性检验

性别与教育程度

以下代码使用CFPS数据检验了教育程度和性别之间的关系:

```
use datasets/cfps_adult.dta, clear
// 清洗样本
drop if te4<0
// 进行检验
tab te4 cfps_gender, chi
```

注意对于微观数据，直接使用tabulate命令的chi选项就可以直接计算Pearson的卡方检验统计量及其p值，可以看到，性别和教育程度有显著的关系，两者之间并非独立。

正态性检验

- 统计中一个常见的假设是关于正态分布的假设，然而正态分布究竟是否满足仍然是存疑的。
- 虽然很多统计方法依赖于大样本、中心极限定理等手段，可以不要求数据的具体分布，然而在很多应用中，数据的正态性满足与否对于应用效果有着决定性影响。
  - 比如在金融市场收益率的研究中，传统上简单假设资产的收益率服从正态分布，然而很多证据表明，资产的收益率一般存在着厚尾特性，具有超额峰度，如果错误的使用正态分布进行建模，会大大低估尾部风险的概率。

正态性检验

- 一个常用的直观方法是使用QQ图。
- QQ图的横坐标和纵坐标分别表示两个分布的分位数，两个分布可以是来源于数据的经验分布，也可以是理论分布。
- 假设两个分布的（经验）分布函数为$F(x), G(x)$，那么其$\alpha-$分位数为$\left(F^{-1}(\alpha), G^{-1}(\alpha)\right)$
- 对于每个$\alpha \in (0,1)$，将两个分布的分位数画在图上，就得到了QQ图。

正态性检验

- 如果两个分布的分布函数是相同的，即$F=G$，那么自然有$F^{-1}(\alpha)=G^{-1}(\alpha)$，从而所有的点：$\left(F^{-1}(\alpha), G^{-1}(\alpha)\right)$应该在$y=x$这条45°直线上
- 如果偏离了这条直线，那么意味着两个分布可能是不相等的。
- 更进一步，使用QQ图，我们还可以更加详细的比较两个分布的特征，特别是尾部特征。
- 因此，为了验证数据是否服从正态分布，只需要将数据的经验分布$\hat{F}(x)$与正态分布做比较，画出QQ图即可。

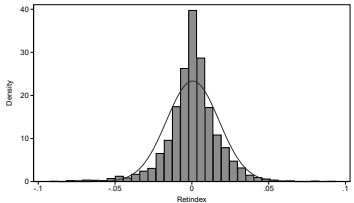
正态性检验

沪深300收益率的正态性

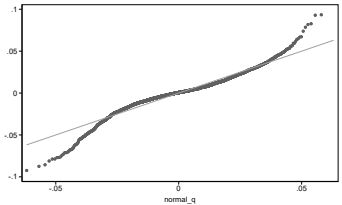
我们使用如上方法检查沪深300的日度收益率数据是否符合正态分布

- 我们使用qqplot_hs300.do画出了QQ图，并与Stata自带的qnorm命令画出的QQ图进行比较
- 我们将纵轴设置为数据的分位数，而横轴为假设的正态分布的分位数。
- 由于横轴上的正态分布的确定需要期望和方差，我们使用样本均值、样本方差作为横轴上正态分布的参数，实际上相当于保证横轴、纵轴两个分布的期望、方差相等，再来比较分位数。

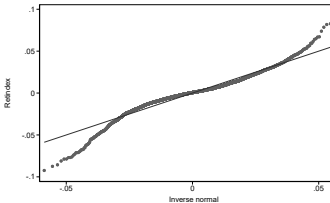
正态性检验



(a)直方图



(b)QQ图（手动）



(c)QQ图（命令）

正态性检验

- 从结果中看到，在分布的左侧尾部，数据的分位数比正态分布预测的分位数要小，而在分布的右侧尾部，数据的分位数比正态分布预测的分位数要大
- 这意味着在左侧和右侧同时出现了厚尾的情况，意味着该数据的峰度系数应该是比正态分布要大的。上面的直方图也说明了这一点，该收益率数据具有尖峰厚尾的特性。
- 由于QQ图中尾部点并不在45°直线上，从而该收益率的分布应该不是正态分布。

Jarque-Bera检验

- 通过以上例子可以看到，正态分布的偏度为0、峰度为3是非常重要的特征，我们在研究一个分布是否是正态分布时，通常最关注的是其偏度和峰度。
- 基于以上的想法，有很多假设检验直接建立在偏度和峰度的基础上，比如**Jarque-Bera检验**：

$$JB = \frac{N}{6}\left[b^2 + \frac{1}{4}(k-3)^2\right]$$

其中

$$b = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{3/2}}$$

$$k = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{2}}$$

分别为样本的偏度系数、峰度系数。

Jarque-Bera检验

- 在正态分布的原假设下，$b \approx 0, k \approx 3$，从而$JB \approx 0$
- 而在备择假设下，$JB > 0$，大样本条件下

$$JB \overset{a}{\sim} \chi^2(2)$$

从而使用右侧检验就可以对正态分布的原假设进行检验。

- 除此之外，还有D’Agostino检验等，同样是基于偏度、峰度的检验。

正态性检验

- 此外，还有基于正态分布次序统计量的检验，该检验实际上可以看作是QQ图的一个扩展。
- 我们可以把如图QQ图的纵轴看做是数据的次序统计量，横轴如果换成一个标准正态分布的次序统计量，那么所有的点应该仍然在一条直线上。
- 如果$x_i \sim \mathcal{N}\left(\mu, \sigma^2\right), z_i \sim \mathcal{N}(0,1)$，记$m_{(i)} = \mathbb{E}\left(z_{(i)}\right), i = 1, ..., N$，其中$z \sim \mathcal{N}(0,1)$，即标准正态分布次序统计量的期望，那么

$$\mathbb{E}\left(x_{(i)}\right) = \mathbb{E}\left[\sigma\left(\frac{x_i - \mu}{\sigma}\right)_{(i)} + \mu\right] = \sigma\mathbb{E}\left(z_{(i)}\right) + \mu = \sigma m_{(i)} + \mu$$

从而$\{x_i\}$的次序统计量为纵轴、$m_{(i)}$为横轴，应该在一条直线上。

正态性检验

- 基于此，可以构造Shapiro-Francia检验：

$$W' = \frac{\mathbb{C}\left(x_{(i)}, m_{(i)}\right)}{\sigma_x \sigma_m} = \frac{\sum_{i=1}^{N}\left[\left(x_{(i)} - \bar{x}\right)\left(m_{(i)} - \bar{m}\right)\right]}{\sqrt{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{N}\left(m_{(i)} - \bar{m}\right)^2}}$$

如果x_i的确来自于正态总体，那么$W \approx 1$，如果$W < 1$则意味着数据不服从正态分布。

- 此外，将这个检验进行更进一步的精炼，还有Shapiro-Wilk检验：

$$W = \frac{\left(\sum_{i=1}^{N} a_i x_{(i)}\right)}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}$$

其中a_i为一个根据$m_{(i)}$的协方差矩阵计算出来的权重。

正态性检验

沪深300的正态性检验

同样使用沪深300的日度收益率数据，在Stata中可以对其进行正态性检验：

```
use datasets/hs300index.dta, clear
// 基于偏度峰度的检验
sktest retindex
// Shapiro-检验Francia
sfrancia retindex
// Shapiro-检验Wilk
swilk retindex
```