4.2 拟合优度检验验证

4.2.1 单样本的拟合优度检验

1、总体分布的卡方检验

首先介绍连续总体的卡方检验。设 $r_1, r_2, \cdots, r_n$ 是待检验的一组随机数，原假设为 $H_0: r_1, r_2, \cdots, r_n$ 来自 $(0,1)$ 上均匀总体。

将 $[0,1]$ 区间分成 m 个小区间，以

$$[\frac{i-1}{m}, \frac{i}{m}) \quad i = 1, 2, \cdots, m$$

表示第 i 个小区间，设 $\{r_j, j = 1, 2, \cdots, n\}$ 落入第 i 个小区间的数目为 $n_i = \sum_{j=1}^{n} I\{r_j \in [\frac{i-1}{m}, \frac{i}{m})\}$。另一方面，根据均匀性假设，$H_0$ 成立时，r_j 落入每个区间的概率为 $p_i = 1/m$，第 i 个小区间的理论频数 $\mu_i = n/m$。卡方检验的基本原理是统计样本的实际频数和理论频数的差异越大说明观测样本的总体分布与理论分布差异越大，因此，检验统计量为

$$V = \sum_{i=1}^{m} \frac{(n_i - \mu_i)^2}{\mu_i} = \frac{m}{n} \sum_{i=1}^{m} (n_i - \frac{n}{m})^2, \tag{4.2.1}$$

且 V 渐近服从$\chi^2(m-1)$，给定显著性水平α，查χ^2分布表得临界值后，即可对经验频率与理论频率的差异作显著性检验。

思考题：基于上述假设检验问题，分析

- 统计量$n_i (i = 1, 2, \cdots, m)$是什么分布？均值和方差是多少？为什么是$\chi^2(m-1)$分布？
- 对于具体的备则假设的分布，比如β分布，区间数如何影响统计量的功效？

首先证明 n_i 服从正态分布，记 $q_i = 1 - p_i$，注意到 $n_i \sim b(n, p_i), \quad i = 1, 2, \cdots, m$。由二项分布的性质可知

$$E(n_i) = np_i, \quad Var(n_i) = np_iq_i.$$

根据中心极限定理，可知 $n_i/\sqrt{n} \xrightarrow{L} N(\sqrt{n}p_i, p_iq_i)$。

接下来，我们证明 V 渐近服从 $\chi^2(m-1)$。首先，$(n_1, n_2, \cdots, n_m) \sim M(n, p_1, p_2, \cdots, p_m)$，根据多项分布的性质，我们有：

$$E(n_i \cdot n_j) = n(n-1)p_ip_j, \quad Cov(n_i, n_j) = -np_ip_j.$$

记随机向量 $\mathbf{N}=(n_1,n_2,\cdots,n_m)$，则有 $E(\mathbf{N})=(np_1,np_2,\cdots,np_m)$，且协方差矩阵为 $n\mathbf{\Sigma}$，其中

$$\mathbf{\Sigma}=\begin{pmatrix} p_1q_1 & -p_1p_2 & \cdots & -p_1p_m \\ -p_1p_2 & p_2q_2 & \cdots & -p_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_m & -p_2p_m & \cdots & p_mq_m \end{pmatrix}.$$

由中心极限定理有，当 $n\to+\infty$ 时，

$$\{\mathbf{N}-E(\mathbf{N})\}/\sqrt{n}\xrightarrow{L}N_m(\mathbf{0},\mathbf{\Sigma}). \tag{4.2.2}$$

取 $D=\mathrm{diag}(1/\sqrt{p_1},1/\sqrt{p_2},\cdots,1/\sqrt{p_m})$，由上式可得，

$$D\{\mathbf{N}-E(\mathbf{N})\}/\sqrt{n}\xrightarrow{L}N_m(\mathbf{0},D\mathbf{\Sigma}D^T).$$

由于 $\sum_{i=1}^m p_i=1$，因此可以验证 $D\mathbf{\Sigma}D^T$ 为幂等矩阵，根据幂等矩阵的性质，有 $D\mathbf{\Sigma}D^T$ 的特征值为 0 或 1。即存在正交矩阵 Q，使得 $Q^TD\mathbf{\Sigma}D^TQ=\mathbf{\Lambda}=\mathrm{diag}(1,1,\cdots,1,0)$。因此，我们有：

$$Q^TD\{\mathbf{N}-E(\mathbf{N})\}/\sqrt{n}\xrightarrow{L}N_m(\mathbf{0},\mathbf{\Lambda}).$$

记 $Q^TD\{\mathbf{N}-E(\mathbf{N})\}/\sqrt{n}=(w_1,w_2,\cdots,w_m)^T$，根据上式我们有 $\{w_i,i=1,2,\cdots,m-1\}$相互独立且渐进服从 $N(0,1)$，同时$w_m=0$。因此，我们有:

$$\{\mathbf{N}-E(\mathbf{N})\}^TD^TQQ^TD\{\mathbf{N}-E(\mathbf{N})\}/n=\sum_{i=1}^m w_i^2\xrightarrow{L}\chi^2(m-1).$$

不难验证

$$\{\mathbf{N}-E(\mathbf{N})\}^TD^TD\{\mathbf{N}-E(\mathbf{N})\}/n=\sum_{i=1}^m\frac{(n_i-np_i)^2}{np_i},$$

因此，根据我们有：

$$\sum_{i=1}^m\frac{(n_i-np_i)^2}{np_i}\xrightarrow{L}\chi^2(m-1).$$

进而，检验统计量 V 渐近服从 $\chi^2(m-1)$得证。对于总体为均匀分布的卡方检验，我们采用如下的模拟来估计统计量V的功效。取样本量 $n=500$，区间个数 $k=10$，并重复 1000 次。模拟程序如下：

```
###均匀性检验—卡方检验
k=9
inter=seq(0,1,by=1/k)
n=500
res1 <- NULL
for (i in 1:1000){
  #data=(runif(n,min = 0,max = 1)+0.1)/1.1
  data=runif(n,min = 0,max = 1)

  left_inter=rep(1,n)%*%t(inter[1:k]) ##t()表示转置
  right_inter=rep(1,n)%*%t(inter[2:(k+1)])

  Data=(data)%*%t(rep(1,k))
  A <- left_inter<=Data&Data<right_inter
  frequ <- apply(A, 2, sum)

  stat1=k/n*sum((frequ-n/k)^2)
  res1[i]=stat1>qchisq(0.95,k-1)
  }
result=mean(res1)
```

如果从一个随机变量 X 中随机抽取若干观察样本，观察样本落在 X 的 k 个互不相交的子集中的观察频数服从一个多项分布，当 k 趋于无穷时，该分布接近 X 的总体分布。

2、单样本 K-S 检验

单样本情形下，K-S (柯尔莫哥洛夫-斯米尔诺夫)检验用于检验观测样本的经验分布函数与理论分布函数是否一致。设 $\{X_1, X_2, \cdots, X_n\}$ 是 n 个独立观测样本，该检验问题的原假设为 H_0 : 该样本 来自分布函数为 $F(\cdot)$ 的总体。

设 $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ 为次序样本，经验分布函数为 $F_n(X_{(i)}) = i/n,\ i = 1, 2, \cdots, n$。经验分布函数与理论分布函数之间的最大偏差记为 KS

统计量，即

$$K_n = \max_{1\le i\le n}\left\{\max(|F(X_{(i)}) - \frac{i}{n}|, |F(X_{(i)}) - \frac{i-1}{n}|)\right\}.$$

H_0 成立时，KS 统计量依分布收敛于 Kolmogorov 分布，其分布函数为

$$P(K_n \le x) = 1 - 2\sum_{j=1}^{+\infty}(-1)^j \exp(-2nj^2x^2), 0 < x < +\infty. \qquad (4.2.3)$$

为了方便 KS 检验在实际数据分析中的使用，文献中也给出了单样本 KS 检验统计量 K_n 的临界值表。对于给定的显著性水平 α，当 K_n 大于临界值时拒绝 H_0。

显然， KS 检验的功效表达式不容易给出。因此，对于固定的备择假设，我们采用蒙特卡洛方法来模拟检验统计量 K_n 的功效。这里我们取样本量 $n = 35$，$X_1, X_2, \cdots, X_n$ 产生自自由度为 3 的 t分布，检验 H_0；该样本是否来自标准正态分布。对于此例，当显著性水平 $\alpha = 0.05$时，查表得临界值为 0.23，当 $K_n \le 0.23$ 时，拒绝 H_0。另一方面，根据式(4.2.3)，我们可以计算检验统计量观测值的 p 值，当 p 值小于 0.05 时，拒绝 H_0。模拟代码如下：

```
###正态性检验—K-S检验(柯氏检验)
n=35
stat1 <- NULL
res1 <- NULL
res2 <- NULL
for (i in 1:1000){
  #data=rnorm(n,0,1) #data from H_0
  data=rt(n,1)  #data from H_1
  data=sort(data)
  D_splus=max(c(1:n)/n-pnorm(data))
  D_minus=max(pnorm(data)-(c(1:n)-1)/n)
  stat1=max(D_splus,D_minus)
  res1[i]=as.numeric(stat1>0.23)  #the critical values
  index=seq(1,10000,1)
  p_val=2*sum((-1)^(index-1)*exp(-2*n*index^2*stat1^2))
```

```
  res2[i]=as.numeric(p_val<0.05)  #the p-value< significant level
}
c(mean(res1),mean(res2))
```

3、变量值随机性检验

游程检验是直接检验随机数序列$\{r_i\}$的随机性。对随机数序列$\{r_i\}$，把它分为许多个子序列，使得其中每一个子序列内的值都是上升的，则称每个子序列为一个上升游程。例如一下10个随机数：$0.855,|\ 0.108, 0.226,|\ 0.032, 0.123,|\ 0.055, 0.545.0.642, 0.870,|\ 0.104, \cdots$，可分为5个上升的游程，第一个游程长为1，其他的上升游程分别为$2, 2, 4, 1$。

首先统计游程长度为$1, 2, 3, 4, 5$和≥ 6的游程数目，分别记为g_1, g_2, g_3, g_4, g_5和g_6，则检验统计量

$$Q_n = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} a_{ij}(g_i - nb_i)(g_j - nb_j)$$

渐近服从$\chi^2(6)$（当$n > 4000$），其中

$$(b_1, b_2, b_3, b_4, b_5, b_6) = (\frac{1}{6}, \frac{5}{24}, \frac{11}{120}, \frac{19}{720}, \frac{29}{5040}, \frac{1}{840}),$$

a_{ij}是下列对称矩阵的元素

$$(a_{ij}) = \begin{pmatrix} 4529.4 & 9044.9 & 13568 & 18091 & 22615 & 27892 \\ & 18097 & 27139 & 36187 & 45234 & 55789 \\ & & 40771 & 54281 & 67852 & 83685 \\ & & & 72414 & 90470 & 111580 \\ & & & & 113262 & 139476 \\ & & & & & 172860 \end{pmatrix}$$

进行游程检验时，要求样本容量$n > 4000$.

以上按上升游程进行游程检验，同样地也可以按下降游程进行检验，检验统计量的形式不变。

思考题：游程检验如何根据游程的个数T 构造检验统计量？

模拟设置：

```
##独立检验--游程检验
n=200
B=c(1/6,5/24,11/120,19/720,29/5040,1/840)
A=matrix(data = c(4529.4,9044.9,13568,18091,22615,27892
                  ,0,18097,27139,36187,45234,55789,0,
                  0,40721,54281,67852,83685,
                  0,0,0,72414,90470,111580,
                  0,0,0,0,113262,139476,
                  0,0,0,0,0,172860),byrow = T,nrow = 6,ncol =6 )
A=A+t(A)-diag(diag(A),nrow = 6,ncol = 6)

G <- matrix(0,nrow = 1000,ncol = 6)
res <- NULL
for (i in 1:1000) {
data=runif(n,min = 0,max = 1)
beg_stop=1
for (j in 1:(n-1)) {
  if(data[j+1]<data[j])
    beg_stop=c(beg_stop,j+1)

}
beg_stop=c(beg_stop,n+1)
end=length(beg_stop)
runs <- beg_stop[2:end]-beg_stop[1:(end-1)]
#the number of runs
G[i,]=c(sum(runs==1),sum(runs==2),sum(runs==3),sum(runs==4),sum(runs==5),sum(runs>=6))

stat=sum(A*((G[i,]-n*B)%*%t(c(1,1,1,1,1,1)))*(c(1,1,1,1,1,1)%*%t(G[i,]-n*B)))/n
stat=t(G[i,]-n*B)%*%(A/n)%*%((G[i,]-n*B))
res[i] <- stat>=qchisq(0.95,6)
}
mean(res)
```