统计计算 张楠 2019秋: Monte Carlo Methods for Hypothesis Testing

(返回 统计计算 张楠 2019秋)

Monte Carlo Methods for Hypothesis Testing

目录

假定我们要考虑如下形式的假设检验问题:

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : \theta_1 \in \Theta_1$$

这里Θ_0, Θ_1为参数空间Θ的划分.

统计假设检验中会出现两种错误:

- Type I error: 零假设被拒绝, 但实际上零假设是正确的;
- Type II error: 零假设被接受, 但实际上零假设是错误的.

检验的**显著性水平**, α, 是一型错误率的上界. 拒绝零假设的概率 依赖于参数θ的真值, 记$\pi(\theta)$为拒绝零假设的概率, 则

$$\alpha = \sup_{\theta \in \Theta_0} \pi(\theta).$$

一型错误率是当零假设正确时, 零假设被拒绝的概率. 因此当一个检验在零假设条件下 被重复很多次后, 观察到的一型错误率就应该逼近α.

若T为检验统计量, T^*为检验统计量的观测值, 则称T^*是**显著的**, 如果基于T^*的检验结论 是拒绝零假设H_0. **显著概率或p值**是使得检验统计量显著的最小的可能α值.

Empirical Type I error rate

Monte Carlo模拟可以用来计算一个检验方法的经验一型错误率. 检验过程在零假设条件下大量重复, 则 经验一型错误率为检验统计量在重复中是显著的比例.

Monte Carlo模拟来计算经验的一型错误率:

1. 对每个重复j, $j = 1, \cdots, m$.
   1. 从零分布产生第j个随机样本 $x_1^{(j)}, \cdots, x_n^{(j)}$;
   2. 基于第j个样本计算检验统计量T_j;
   3. 记录决策结果$I_j = 1$,若H_0在显著性水平α下被拒绝; 否则 $I_j = 0$.
2. 计算显著的检验比例$\frac{1}{m}\sum_{j=1}^m I_j$, 此比例即为观测到的一型错误率.

用$\hat{p}$表示估计的一型错误率, 则其标准方差的估计为

$$\hat{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{m}} \leq \frac{0.5}{\sqrt{m}}.$$

假设$X_1, \cdots, X_{20}\ i.i.d\ N(\mu, \sigma^2)$, 假设$H_0 : \mu = 500\ H_1 : \mu > 500$. 在零假设下,

$$T^* = \frac{\bar{X}-500}{S/\sqrt{20}} \sim t_{19}.$$

大的T^*值是支持对立假设. 我们使用Monte Carlo方法来计算在$\sigma = 100$时的一型错误率, 来检测其是否逼近0.05.

```
n <- 20; alpha <- .05;   mu0 <- 500;  sigma <- 100;
m <- 10000            #number of replicates
p <- numeric(m)       #storage for p-values
for (j in 1:m) {
    x <- rnorm(n, mu0, sigma)
    ttest <- t.test(x, alternative = ''greater'', mu = mu0)
    p[j] <- ttest$p.value
    }
p.hat <- mean(p < alpha)
se.hat <- sqrt(p.hat * (1 - p.hat) / m)
print(c(p.hat, se.hat))
plot(1:m,cumsum(p<alpha)/1:m,type=''l'')
abline(h=0.05)
```

例8: 正态分布的偏度检验 一个随机变量X的偏度系数定义为

$$\beta_1 = \frac{E[(X-\mu)^3]}{\sigma^3},$$

其中$\mu = EX, \sigma^2 = Var(X)$. 一个分布称为是对称的, 如果$\beta_1 = 0$; 称为是正偏的,如果$\beta_1 > 0$;称为是负偏的,如果$\beta_1 < 0$. 偏度 系数的估计为

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^3}{(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2)^{3/2}}.$$

可以证明, $\sqrt{n}(b_1 - \beta_1) \overset{Asy}{\Longrightarrow} N(0,6)$. 正态分布是对称的分布, 因此偏度系数可以用来检验一个分布是否为对称分布. 假设为

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0.$$

从而可以基于检验统计量b_1构建一个检验法则.

$$|b_1| \geq z_\alpha \sqrt{6/n}, Reject\ H_0;\ Otherwise,\ Accept\ H_0.$$

对大样本来说, 检验的显著性 水平达到指定的显著性水平, 比如$\alpha = 0.05$; 对小样本来说, 由于b_1的理论分布 不可得到, 从而确切的显著性水平不可得到, 但其经验的一型错误率可以通过Monte Carlo方法来估计.

```
n <- c(10, 20, 30, 50, 100, 500) #sample sizes
    cv <- qnorm(.975, 0, sqrt(6/n))  #asymptotical crit. values for each n

    sk <- function(x) {
    #computes the sample skewness coeff.
    xbar <- mean(x)
    m3 <- mean((x - xbar)^3)
    m2 <- mean((x - xbar)^2)
    return( m3 / m2^1.5 )
}

#n is a vector of sample sizes
#we are doing length(n) different simulations
p.reject <- numeric(length(n)) #to store sim. results
m <- 10000                     #num. repl. each sim.
for (i in 1:length(n)) {
    sktests <- numeric(m)       #test decisions
    for (j in 1:m) {
        x <- rnorm(n[i])
        #test decision is 1 (reject) or 0 (Accept)
```

```
        sktests[j] <- as.integer(abs(sk(x)) >= cv[i] )
        }
    p.reject[i] <- mean(sktests) #proportion rejected
}
names(p.reject)<-as.character(n)
p.reject
```

结果表明对较少的样本量时, 这种检验的显著性水平偏低,没有达到理论值0.05. 在正态总体下, 可以求出

$$Var(b_1) = \frac{6(n-2)}{(n+1)(n+3)}.$$

因此, 对小样本, 应该使用此方差来计算检验的临界值:

```
cv<-qnorm(.975, 0, sqrt(6*(n-2)/((n+1)*(n+3))))  # crit. values for each n
p.reject <- numeric(length(n)) #to store sim. results
m <- 10000                    #num. repl. each sim.
for (i in 1:length(n)) {
    sktests <- numeric(m)       #test decisions
    for (j in 1:m) {
        x <- rnorm(n[i])
        #test decision is 1 (reject) or 0 (Accept)
        sktests[j] <- as.integer(abs(sk(x)) >= cv[i] )
        }
    p.reject[i] <- mean(sktests) #proportion rejected
}
p.reject
```

现在得到的估计就比较靠近名义值0.05了.

Power of a Test

一个检验的功效定义为

$$\pi(\theta) = P_\theta(Reject\ H_0)$$

当有一个$\theta_1 \in \Theta_1$时, 犯二型错误的概率为$1 - \pi(\theta_1)$. 理想中, 我们 希望一个检验的错误率越低越好. 一型错误率通过显著性水平α的选择而控制, 较低 的二型错误率对应于较高的功效. 因此, 在比较两个具有同样显著性水平的检验时, 我们感兴趣 的是比较它们的功效. 一般来说这种比较不是简单的一个问题, 比如一个检验的功效$\pi(\theta_1)$ 依赖于在对立假设下的θ_1 值. 当一个检验的功效函数没有分析的表达式时, 功效函数 在特定的$\theta_1 \in \Theta_1$处的值可以通过Monte Carlo方法得到.

Monte Carlo 方法计算检验的功效:

1. 选择一个特定的$\theta_1 \in \Theta$.
2. 对每个重复j, $j = 1, \cdots, m$.
    1. 在对立假设空间$\theta = \theta_1$的情况下产生第j个随机样本 $x_1^{(j)}, \cdots, x_n^{(j)}$;
    2. 基于第j个样本计算检验统计量T_j;
    3. 记录决策结果$I_j = 1$,若H_0在显著性水平α下被拒绝; 否则 $I_j = 0$.
3. 计算显著的检验比例$\frac{1}{m}\sum_{j=1}^m I_j$, 此比例即为检验的功效.

例9: 使用模拟方法估计例7中t检验的功效并画出经验功效函数曲线

```
n <- 20
m <- 1000
mu0 <- 500
sigma <- 100
mu <- c(seq(450, 650, 10))  #alternatives
```

```
M <- length(mu)
power <- numeric(M)
for (i in 1:M) {
    mu1 <- mu[i]
    pvalues <- replicate(m, expr = {
        #simulate under alternative mu1
        x <- rnorm(n, mean = mu1, sd = sigma)
        ttest <- t.test(x,
                 alternative = ''greater'', mu = mu0)
        ttest$p.value  } )
    power[i] <- mean(pvalues <= .05)
}
```

估计的功效$\hat{\pi}(\theta)$存在向量$power$中, 下面画出经验功效函数的图像, 我们这里使用**Hmisc**包里的**errbar**函数.

```
par(ask = TRUE)
library(Hmisc)  #for errbar
plot(mu, power)
abline(v = mu0, lty = 1)
abline(h = .05, lty = 1)

#add standard errors
se <- sqrt(power * (1-power) / m)
errbar(mu, power, yplus = power+se, yminus = power-se,
    xlab = bquote(theta))
lines(mu, power, lty=3)
detach(package:Hmisc)
par(ask = FALSE)
```

注意对t检验, 对假设$H_0: \mu = \mu_0 \quad vs \quad H_1: \mu \neq \mu_0$而言, 检验统计量$T = (\bar{X} - \mu)/(S/\sqrt{n})$在零假设下服从中心的$t$分布, 而 在对立假设下, T服从非中心的t分布, 非中心参数为$\delta = (\mu - \mu_0)\sqrt{n}/\sigma$. 在R中,可以使用函数**power.t.test**来计算t检验的功效或者在给定功效下确定各参数的值.

例10: 混合正态分布下正态性偏度检验的功效 在例8中, 我们考虑了正态性偏度检验的一型错误率. 对混合正态分布 $pN(0, \sigma^2 = 1) + (1-p)N(0, \sigma^2 = 100), 0 \leq p \leq 1$. 当$p = 0, 1$, 为正态分布; 而当$0 < p < 1$, 不再为正态分布. 因此, 我们可以计算一下偏度检验用于检验正态性的功效. 这里考虑显著性水平$\alpha = 0.1$, 样本量$n = 30$. 我们重复此过程$m = 2500$次来计算功效.

```
alpha <- .1
n <- 30
m <- 2500
epsilon <- c(seq(0, .15, .01), seq(.15, 1, .05))
N <- length(epsilon)
pwr <- numeric(N)
#critical value for the skewness test
cv <- qnorm(1-alpha/2, 0, sqrt(6*(n-2) / ((n+1)*(n+3))))

for (j in 1:N) {          #for each epsilon
    e <- epsilon[j]
    sktests <- numeric(m)
    for (i in 1:m) {      #for each replicate
        sigma <- sample(c(1, 10), replace = TRUE,
            size = n, prob = c(1-e, e))
        x <- rnorm(n, 0, sigma)
        sktests[i] <- as.integer(abs(sk(x)) >= cv)
        }
    pwr[j] <- mean(sktests)
    }
#plot power vs epsilon
plot(epsilon, pwr, type = ''b'',
     xlab = bquote(epsilon), ylim = c(0,1))
abline(h = .1, lty = 3)
se <- sqrt(pwr * (1-pwr) / m)  #add standard errors
lines(epsilon, pwr+se, lty = 3)
lines(epsilon, pwr-se, lty = 3)
```

经验的功效函数曲线在$p=0$和$p=1$两点接近$\alpha=0.1$, 对$0<p<1$, 功效函数大于0.1, 最大值大约在$p=0.15$处达到.

Power Comparisons

Monte Carlo方法经验用于比较不同检验的功效. 本节我们讨论对于正态性偏度检验问题几种检验 的功效差异. 文献中对于正态性检验已有很多不同的方法. 我们这里考虑三种检验方法.

> **例11: 正态性检验的功效比较** 对一元正态性检验问题, 比较Shapiro-Wilk检验, Energy检验 和偏度检验三种检验的功效差异.

假设$\mathcal{N}$为一个一元正态分布类, 检验假设

$$H_0: F_X \in \mathcal{N} \quad vs \quad H_1: F_X \notin \mathcal{N}.$$

Shapiro-Wilk 检验是基于样本次序统计量对它们在正态性成立下的期望作回归, 因此它属于一般基于回归 和相关的类别检验方法. 对样本量$7 \le n \le 2000$, 近似的检验统计量临界值通过将统计量W变换到 正态分布随机变量而得到. 参考阅读材料了解更多Shapiro-Wilk检验方法. 在R中, 可以通过**shapiro.test**函数作Shapiro-Wilk检验.

Energy检验是基于抽样分布和正态分布之间的“energy”距离进行检验, 因此大的检验统计量值意味着显著性. energy检验是用来检验多元正态性的一种方法, 这里我们考虑的是其特例$d=1$, 在一元正态检验下, energy检验类似于 Anderson-Darling检验. energy检验统计量为

$$Q_n = n\left[\frac{2}{n}\sum_{i=1}^n E\|x_i - X| - E\|X-X'\| - \frac{1}{n^2}\sum_{i,j=1}^n \|x_i - x_j\|\right],$$

其中X, X'为$i.i.d$的正态分布随机变量. 大的Q_n值表明显著性. 在一元情形下, Q_n可以表示为

$$Q_n = n\left[\frac{2}{n}\sum_{i=1}^n (2Y_i\Phi(Y_i) + 2\phi(Y_i)) - \frac{2}{\sqrt{\pi}} - \frac{1}{n^2}\sum_{k=1}^n (2k-1-n)Y_{(k)}\right],$$

其中$Y_i = \frac{X_i - \mu_X}{\sigma_X}$, $Y_{(k)}$为$Y_1, \cdots, Y_n$ 的第k个次序统计量. 若参数μ_X, σ_X未知, 则用样本均值和样本方差. 在多元情形也有类似的计算公式, 在R中 energy检验包含在**energy**包中, 名称为**mvnorm.etest**.

第三种检验就是我们前面例子中的偏度检验. 我们取显著性水平为$\alpha=0.1$, 对立假设为

$$(1-p)N(\mu=0,\sigma^2=1) + pN(\mu=0,\sigma^2=100),\ 0 \le p \le 1$$

当$p=0$或者1时, 分布为正态分布, 此时经验的一型错误率应该被名义的错误率$\alpha=0.1$控制住, 我们 感兴趣的是在$0<p<1$时三种检验的功效比较.

```r
library(energy)
alpha <- .1
n <- 30
m <- 500          #try small m for a trial run
test1 <- test2 <- test3 <- numeric(m)

#critical value for the skewness test
cv <- qnorm(1-alpha/2, 0, sqrt(6*(n-2) / ((n+1)*(n+3))))
sim <- matrix(0, 11, 4)

# estimate power
for (i in 0:10) {
    epsilon <- i * .1
    for (j in 1:m) {
        e <- epsilon
```

```r
        sigma <- sample(c(1, 10), replace = TRUE,
            size = n, prob = c(1-e, e))
        x <- rnorm(n, 0, sigma)
        test1[j] <- as.integer(abs(sk(x)) >= cv)
        test2[j] <- as.integer(
                    shapiro.test(x)$p.value <= alpha)
        test3[j] <- as.integer(
                    mvnorm.etest(x, R=200)$p.value <= alpha)
        }
    print(c(epsilon, mean(test1), mean(test2), mean(test3)))
    sim[i+1, ] <- c(epsilon, mean(test1), mean(test2), mean(test3))
}
detach(package:energy)

# plot the empirical estimates of power
plot(sim[,1], sim[,2], ylim = c(0, 1), type = ''l'',
    xlab = bquote(epsilon), ylab = ''power'')
lines(sim[,1], sim[,3], lty = 2)
lines(sim[,1], sim[,4], lty = 4)
abline(h = alpha, lty = 3)
legend(''topright'', 1, c(''skewness'', ''S-W'', ''energy''),
    lty = c(1,2,4), inset = .02)
```

结果表明Shapiro-Wilks检验和energy检验在检验此假设中有着差不多的功效, 模拟比较中设定$n = 30$ 和$p < 0.5$. 此两个检验的功效都比偏度检验的功效大, energy检验看起来在$0.5 \leq p \leq 0.8$时 功效最大.
