R 的基本统计命令与假设测验

本文介绍了 R 的基本统计命令，包括变数的统计描述、单个平均数的统计推断、两个平均数的假设测验等，并给出了具体实例的 R 程序和计算结果的分析。

一、变量的统计描述

（一）命令

R 提供了用来计算变量的各种统计量的命令 summary，其用法和有关的说明如下：

用法：summary(x)

功能：输出向量 x 的平均值 MEAN，中位数 MEDIAN，最大值 MAX，最小值 MIN，四分之三值 Q3，四分之一值 Q1。

说明：将变量 x 的数据由小到大排列，Q3 为 3(N+1)/4 位置的值，Q1 为(N+1)/4 位置的值。

另外，用 length 获得样本个数，用 stderr 获得样本平均数的标准误。

（二）实例

〖例 1〗测定 10 个“金皇后”玉米果穗的重量(克)，得结果为 150，170，160，162，170，168，171，162，151，160。试计算它的平均数、标准差、变异系数等统计量。

1 R程序

```
# example 4-1
c1 <- c(150,170,160,162,170,168,171,162,151,160)
summary(c1)
# sample size
n <- length(c1); n
# variance
s2 <- var(c1); s2
# standard deviation
s <- sqrt(var(c1)); s
# standard error of the sample mean
se <- stderr(c1); se

# estimate the coefficient of variation
cv <- sqrt(var(c1)) /mean(c1) * 100; k1

# geometric mean
c2 <- log(c1)
```

```
gmean <- exp(mean(c2))
gmean

#harmonic mean
c3 <- 1/c1
hmean <- 1/mean(c3)
hmean
```

2 R计算结果

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  150.0   160.0   162.0   162.4   169.5   171.0
```

N	10	样本个数
s	7.5748	标准差
se	2.3954	样本平均数的标准误
cv	4.66429	变异系数
gmean	162.2386	几何平均数
hmean	162.0748	调和平均数

结果表明，“金皇后”玉米果穗的重量平均为 162.4 克，标准差为 7.57 克，变异系数为 4.7%。

二、单个平均数的假设测验

（一）命令

在 R 中，有一个 t.test 命令用来完成未知总体方差的单个平均数的假设测验和置信区间的估计，简介如下：

用法：t.test(x, y = NULL,
alternative = c("two.sided", "less", "greater"),
mu = 0, paired = FALSE, var.equal = FALSE,
conf.level = 0.95, ...)

功能：对列变量 x 执行 t 测验。

说明：x 为列变量；alternative 可以指明是两尾测验、或大于或小于的一尾测验；mu 是总体平均数，默认值为 0；paired 指明是否为成对数据（但 y 赋值时）；var.equal 指明两个样本的方差是否相同；conf.level 指明置信区间的概率值。

R 没有提供正太分布的 z 测验函数，我们在应用时自己编写一个。

（二）实例

〖例 2〗某蔗糖自动打包机在正常工作状态时的每包蔗糖重量符合 N(100，2)分布。某日抽样调查 10 包，得结果为 100.5、99.8、101.5、102.2、102.7、100.3、101.1、101.6、

100.4、100.9、100.7、102.1($\bar{y}$=101)公斤。问该打包机是否仍处于正常工作状态？

1　R 程序

```
# example 4-2
c1 <- c(100.5,99.8,101.5,102.2,102.7,100.3,101.1,101.6,100.4,100.9, 100.7, 102.1)
z.test = function(a, mu, var){
   zeta = (mean(a) - mu) / (sqrt(var / length(a)))
   return(zeta)
}
z.test(c1,mu=100,var=4)
```

2　R 输出结果

Z = 1.9918

结果表明：当 Z=1.99 时，其概率值 P 小于 0.05，说明该打包机在当前工作状态下与在正常工作状态下的差异已经达到了 0.05 的显著水平，故该打包机已经处于不正常工作状态。

说明：Z 代表正态分布值，P VALUE 为其概率值，用以说明 Z 的显著性。

〖例 3〗测定某棉田的地表光强 4 次，得结果为：3.4，2.8，3.5，4.1(千勒克斯)，试测验该结果与根据 BEER-LAMBERT 定律推出的理论值 μ_0=3.0 是否有显著差异。

1　R程序：

```
# example 4-3
c1 <- c(3.4,2.8, 3.5, 4.1)
t.test(c1,mu=3)
```

2　R输出结果：

```
    One Sample t-test

data:  c1
t = 1.6908, df = 3, p-value = 0.1895
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.603007 4.296993
sample estimates:
mean of x
     3.45
```

结果表明：当 t=1.69 时，其概率值为 0.19，大于 0.05，故推断实测结果与理论值相符合。

三、成组数据的比较

（一）命令

有两个 R 命令可以用于成组数据的比较，即 TWOSAMPLE 和 TWOT，用法如下：

1. TWOSAMPLE

用法：TWOSAMPLE　[置信水平 K%]　C　C

子命令：ALTERNATIVE，POOLED

功能：　作无效假设 H_0：(μ_1= μ_2)的 t 测验，并估计(μ_1- μ_2)的置信区间；

说明：　第一列包含来自总体 1 的样本，第二列包含来自总体 2 的样本；如果未用子命令 POOLED，TWOSAMPLE 假定两个总体的方差不相等，反之，如果两个总体的方差相等，就需要使用这个子命令；如果置信水平没有指明，假定置信水平为 95%；如果没有 ALTERNATIVE 子命令，做两尾测验；如果 ALTERNATIVE=-1，作被择假设 μ_1< μ_2的一尾 t 测验；如果 ALTERNATIVE=1，作被择假设 μ_1> μ_2的一尾 t 测验。

例子：TWOSAMPLE 99 C1 C2　（进行 C1、C2 两组数据的 99%的 t 测验）

2. TWOT

用法：TWOT　[置信水平 K]　C，C

子命令：ALTERNATIVE，POOLED

功能：作两个平均数差数的 t 测验和置信区间的估计;

说明：TWOT 和 TWOSAMPLE 的功能相同，只是输入数据格式有别；第一列含有两个样本的观察数据，第二列则包含第一列数据所属样本的下标值，如 1 和 2，1 代表第一列中对应的数据属第一个样本，2 代表第一列中对应的数据属第二个样本。如果置信水平没有指明，假定置信水平为 95%；如果没有 ALTERNATIVE 子命令，做两尾测验。

例子：TWOT　C1　C2　（对 C1 中的两组数据进行 95％的 t 测验）

（二）情况之一：$\sigma_1^2 = \sigma_2^2$

〖例 4〗以 20 头猪作饲养试验，随机抽取其中的 10 头为一组，喂以甲种饲料，另 10 头为一组，喂以乙种饲料，饲养一个月后测得各头猪增加的体重(斤)列于表 1。试测验两种饲料对猪增重有无显著差异。

表 1　喂以不同饲料各头猪增加的体重(斤)

饲料种类	猪的体重（斤）									
甲种饲料（X1）	30	35	40	32	42	31	41	38	36	34
乙种饲料（X2）	25	27	33	35	37	33	33	34	31	29

1　R程序

在进行成组数据的比较时，首先要对两个样本的方差进行同质性检验，如果方差同质，则使用联合方差进行平均数差数的假设测验，否则不能使用联合方差。

```
# example 4-4
```

```
# 假设两个样本方差相同的 t 测验函数
twosam <- function(y1, y2) {
  n1 <- length(y1); n2 <- length(y2)
  yb1 <- mean(y1); yb2 <- mean(y2)
  s1 <- var(y1); s2 <- var(y2)
  s <- ((n1-1)*s1 + (n2-1)*s2)/(n1+n2-2)
  tst <- (yb1 - yb2)/sqrt(s*(1/n1 + 1/n2))
  tst
}

c1 <- c(30,35,40,32,42,31,41,38,36,34)
c2 <- c(25,27,33,35,37,33,33,34,31,29)
# 方差同质性检验
k1 <- var(c1) / var(c2)
k1 # F value to check if the variances of the two samples are significant or not
# 用联合方差进行平均数的假设测验
twosam(c1,c2)
```

2 R计算结果

$F_{9,9}$ = 1.3127（累积概率值为 0.65 说明两个样本方差没有显著差异）

t = 2.35

结果表明：由于两个样本方差同质，所以使用两个样本的联合方差进行 t 测验，两个样本平均数相等的概率为 0.03，所以两个样本平均数具有显著差异，也就是说两种饲料对猪增重有显著差异。

（三）情况之二：$\sigma_1^2 <> \sigma_2^2$

〖例 5〗调查某地区小麦密点播田块 7 块，小麦撒播田块 8 块，每块田的亩产量(斤)列于表 2。试测验两种播种方式的小麦产量是否有显著差异。

表 2 小麦播种方式试验产量结果

播种方式	产量							
密点播	510	480	470	490	500	490	480	
撒播	500	450	430	440	490	480	410	420

1 R程序

```
# example 4-5
# 扩展的两个成组样本的 t 测验函数，可以用于方差相同和不同的 t 假设测验。
twosam <- function(y1, y2, pooled = T) {
```

```
  n1 <- length(y1);   n2 <- length(y2)
  yb1 <- mean(y1);   yb2 <- mean(y2)
  s1 <- var(y1);   s2 <- var(y2)
  sd1 <- sqrt(s1);   sd2 <- sqrt(s2)
  se1 <- stderr(y1);   se2 <- stderr(y2)
  tst <- 0
  df <- 0
  if (pooled) {
    s <- ((n1-1)*s1 + (n2-1)*s2)/(n1+n2-2)
    tst <- (yb1 - yb2)/sqrt(s*(1/n1 + 1/n2))
    df <- n1 + n2 -2
  }
  else {
    s <- sqrt(s1/n1 + s2/n2)
    tst <- (yb1 - yb2)/s
    c <- s1 / n1 /(s1 /n1 + s2 /n2)
    df <- (n1-1)*(n2-1) / ((n2-1)*c*c + (1-c)*(1-c)*(n1-1))
  }
  ysum <- c(n1,yb1,s1,sd1,se1)
  ysum <- rbind(ysum, c(n2,yb2,s2,sd2,se2))

  colnames(ysum) <- c("N", "Mean", "Var","SD", "SE Mean")
  rownames(ysum) <- c("Y1","Y2")

  df <- floor(df)

  res <- list(sum=ysum, ttest=tst, df=df)
  res
}

c1 <- c(510,480,470,490,500,490,480)
c2 <- c(500,450,430,440,490,480,410,420)
# 进行方差同质性检验
k1 <- var(c2) / var(c1)
k1
# 计算 F 值的累积概率
cdf <- pf(k1, length(c2)-1, length(c1)-1)
cdf
```

```
# 不使用联合方差的 t 测验
twosam(c1,c2,pooled=F)# 进行方差同质性检验
```

2 计算结果与分析

```
6.2763  0.9802 （累积概率达到了 0.98，所以两个样本的方差在 0.05 水平上显著）
TWOSAMPLE T FOR C1 VS C2
     N       MEAN      STDEV     SE MEAN
C1   7      488.5714  13.45185    5.084323
C2   8      452.5000  33.70036   11.914877
T= 2.7845   P=0.021   DF=   9
```

结果表明：由于两个样本方差不同质，所以使用两个样本各自的方差进行 t 测验，两样本平均数相等的概率为 0.0214，所以两个样本平均数差异显著，也就是说密点播产量显著不同于撒播产量。

四、成对数据的比较

成对数据的假设测验比成组数据的要简单，只需要测验成对数据的差数与 0 是否有显著差异即可，所以可以用 t 测验命令完成成对数据的假设测验。下面是一个实例。

〖例 6〗我们在 10 个试验点进行了早稻新品种和当地品种成对比较试验，其产量结果列于表 3 所示。试测验新品种与当地品种之间是否有显著差异。

表 3 早稻品种比较试验产量结果表

品种	产量									
新品种	880	950	840	940	780	880	920	810	940	780
当地品种	820	920	880	870	810	820	880	780	890	760

1 R程序：

```
# example 4-6
c1<-c(880,950,840,940,780,880,920,810,940,780)
c2<-c(820,920,880,870,810,820,880,780,890,760)
c3 <- c1-c2
t.test(c3)
```

2 计算结果与分析

```
        One Sample t-test

data:  c3
t = 2.4617, df = 9, p-value = 0.03606
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

2.350888 55.649112
sample estimates:
mean of x
29

结果表明：当 t=2.46 时，其概率为 0.036，说明两个平均数之间的差异已达到 0.05 的显著水平，也就是说新品种较当地品种显著增产，增产的幅度为 2.3～55.7 斤。

五、实习

实习 1 测得 1960～1972 年间越冬代棉红铃虫在江苏东台的羽化高峰期依次为（以 6 月 30 日为 0）8,6,10,5,6,6,10,-1,12,11,9,1,8。试求其平均值、标准差和变异系数。〖答案：$\bar{y}$=7.0，s=3.8，CV=54.3〗

实习 2 某玉米品种在不施氮肥的情况下，一般产量为 250kg/亩。现调查该品种在施同样氮肥量情况下，10 个小区的产量数据（kg/亩）为：270，300，285，268，275，298，310，295，304，278。试测验施氮肥在产量上是否和未施氮肥差异显著，并以 95%的可靠度估计施氮肥情况下该品种的亩产范围。〖答案：t=8.056，亩产范围是 277.5～299.1〗

实习 3 据历史资料，“岱字棉 15”纤维长度为一 N(29.8，2.25)的总体。现从中选出一个株系，取 10 个纤维样品，测得其纤维长度分别为 31.2，30.9，31.5，30.9，30.5，31.2，31.8，30.9，30.6，30.7。试测验该株系的纤维长度是否显著优于原总体。〖答案：$\bar{y}$=31.0，s=0.408，Z=1.71，一尾 P=0.043〗

实习 4 测定前作喷洒过某种有机砷杀雄剂的麦田植株样本 4 次，得株体中的砷残留量为 7.5，9.7，6.8，6.4(mg)；测定对照（前作未用过有机砷杀雄剂）的植株样本 3 次，得株体中砷含量为 4.2，7.0，4.6。试测验喷洒有机砷杀雄剂是否使后作株体的砷含量显著增高。〖t=2.05，一尾 P=0.048，达显著水平〗

实习 5 为测定 A、B 两种病毒对烟草的致病力，取 8 株烟草，每一株皆半叶接种 A 病毒，另半叶接种 B 病毒，以叶面出现枯斑数的多少作为致病力强弱的指标，得结果于表 4。试测验两种病毒致病力的差异显著性。〖答案：t=2.632，P=0.034〗

表 4 两种病毒在烟叶上产生的枯斑数

株号	1	2	3	4	5	6	7	8
病毒 A	9	17	31	18	7	8	20	10
病毒 B	10	11	18	14	6	7	17	5