样本量估计及其在 nQuery 和 SAS 软件上的实现
——均数比较(四)

段重阳[1]　吕　朵[1]　陈平雁[2△]

1.2.1.5　Wilcoxon/Mann-Whitney 秩和检验(连续变量)

方法:Noether(1987)[9] 给出 Wilcoxon/Mann-Whitney 秩和检验样本量计算公式为:

$$2n=\frac{(z_{\alpha/s}+z_{1-\beta})^2}{3(p_1-\frac{1}{2})^2} \qquad (1-12)$$

式中,p_1 为当备择假设成立时一组观察值小于另一组观察值的概率。

【例1-11】一个两组平行随机对照的研究,以定量变量血细胞比容为评价指标,拟用 Wilcoxon 两样本秩和检验进行分析。假定两组的均数差为2.2,共同标准差为2.0。若采用平衡设计,取 $\alpha=0.05$,检验效能为90%,试估计样本量。

nQuery Advisor 7.0 实现:设定检验水准 $\alpha=0.05$;双侧检验,即 $s=2$;检验效能取 $1-\beta=90\%$ 。在 nQuery Advisor 7.0 主菜单选择:

Goal:Make Conclusion Using:⊙ Means

Number of Groups:⊙ Two

Analysis Method:⊙ Test

方法框中选择:Wilcoxon/Mann-Whitney rank-sum test(continuous outcome)。

对于 $p_1=P(X<Y)$ 的估计,我们可以根据预试验和其他信息获得,也可以用 Cohen 基于正态分布及效应量 δ(effect size)对 p_1 的定义计算,$p_1=\Phi(0,\delta/\sqrt{2},1)$ 其中,$\delta=\frac{\mu_1-\mu_2}{\sigma}$,例如,$\delta$ 较小($\delta=0.2$)时,$p_1=0.556$;δ 中等($\delta=0.5$)时,$p_1=0.638$;δ 较大($\delta=0.8$)时,$p_1=0.714$。本例可借助软件的辅助计算功能估计 p_1,在菜单栏中选择:

Assistants:⊙ Compute Effect Size

在弹出的计算窗口将各参数键入,如图1-25所示,结果为 $p_1=0.782$。

将计算结果 p_1 和其他参数键入主对话框(图1-26),结果 $n=23$。本例若采用两样本 t 检验的样本量估计方法,求得的样本量为 $n=19$。

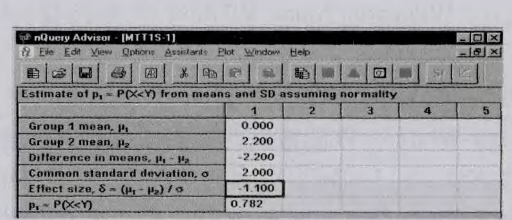

图1-25　nQuery Advisor 7.0 关于例1-11样本量估计的参数计算结果

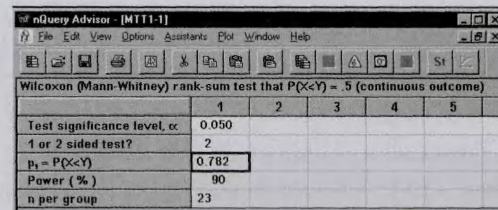

图1-26　nQuery Advisor 7.0 关于例1-11样本量估计的参数设置与计算结果

SAS 9.2 软件实现:

```
PROC IML;
start MTT1(a,s,p1,power);
error=0;
if(a>1 | a<0) then do;error=1;print "error" "Test significance level must be in 0-1";end;
if(p1>1 | p1<0) then do;error=1;print "error" "P(X<Y) must be in 0-1";end;
if(s^=1 & s^=2) then do;error=1;print "error" "s=1 or 2";end;
if(power>100 | power<1) then do;error=1;print "error" "Power(%) must be in 1-100";end;
if(p1=0.5) then do;error=1;print "error" "P1 can not be 0.5";end;
if(error=1) then stop;
if(error=0) then do;
N=(probit(a/s)+probit(1-power/100))##2/(3#(p1-1/2)##2);
n=ceil(N/2);
print a[label="Test significance level"]
s[label="1 or 2 sided test"]
p1[label="p1=P(X<Y)"]
power[label="Power(%)"]
n[label="n per group"];
```

1. 南方医科大学 公共卫生与热带医学学院 生物统计学系2007级本科生
2. 南方医科大学 公共卫生与热带医学学院 生物统计学系
△ 通讯作者:陈平雁

```
end;
finish MTT1;
run MTT1(0.05,2,0.782,90);
quit;
```

SAS 运行结果：

Test significance level	1 or 2 sided test	p1=P(X<Y)	Power(%)	n per group
0.05	2	0.782	90	23

图 1－27　SAS 9.2 关于例 1－11 样本量估计的参数设置与计算结果

1.2.1.6　Wilcoxon/Mann-Whitney 秩和检验（等级变量）

方法：Kolassa（1995）[10] 提出等级变量的 Wilcoxon/Mann-Whitney 秩和检验的检验效能计算采用计算机模拟方法，在已知每组样本的情况下，根据每组各等级的概率值（如图 1－28 中所示），产生随机数模拟出两组各个等级的例数，然后用 Wilcoxon/Mann-Whitney 秩和检验法检验两组有无差异，重复此过程 n_sim 次（建议在 10000 次以上），用检验出有差异的次数除以总模拟次数求出 Power 值。若要通过 Power 计算样本量 n，可以通过不断尝试样本量 n 的方法找到满足要求的 Power，此时对应的 n 就是要求的样本量。

【例 1－12】一个两样本的随机、双盲、安慰剂平行对照设计，评价指标采用生存质量量表，等级度量。预试验获得的数据如下：

分组	生存质量				
	0	5～30	35～65	70～95	100
安慰剂	23	8	10	14	45
试验药	13	6	10	16	55

若采用平衡设计，取 $\alpha = 0.05$，检验效能为 80%，试估计样本量。

nQuery Advisor 7.0 实现：设定检验水准 $\alpha = 0.05$，双侧检验，分为 5 个等级。

在 nQuery Advisor 7.0 主菜单选择：

Goal：Make Conclusion Using：⊙ Means

Number of Groups：⊙ Two

Analysis Method：⊙ Two

方法框中选择：Wilcoxon（Mann-Whitney）rank-sum test（ordered categories）。

在弹出的样本量计算窗口中，$p_1 = P(X<Y)$ 需要通过辅助计算获得，在菜单选择 Asistants，将预试验所获得的两组等级资料输入，如图 1－28 所示，计算得到 $p_1 = P(X<Y)$ 为 0.570，菜单栏点击 Transfer，数据转入主菜单窗口，如图 1－29。在“n per group”栏内尝试填入样本量，然后点击“Power”栏，获得对应的检验效能，反复尝试，直到检验效能为 80%。需注意，检验效能为 80% 所对应的样本量不止一个，一般我们取最大的一个样本量。本例最终求得的样本量为 233。

MTT2-1 Aid table E1　Specification of probabilities in groups, categories

Compute | Transfer | Close | Restore | Clear | Cut | Copy | Paste | Print

Category	Proportion in Group 1 (X)	Proportion in Group 2 (Y)
1	0.230	0.130
2	0.080	0.060
3	0.100	0.100
4	0.140	0.160
5	0.450	0.550
$\Sigma\pi_i$	1.000	1.000
$p_1 = P(X<Y)$	0.570	

图 1－28　nQuery Advisor 7.0 关于例 1－12 样本量估计的参数计算结果

MTT2-1 Main

Wilcoxon (Mann-Whitney) rank-sum test that P(X<Y) = .5 (ordered categories)

	1	2	3	4	5
Test significance level, α	0.050				
1 or 2 sided test?	2				
Number of categories, k	5				
Side table name	E1				
$p_1 = P(X<Y)$	0.570				
Power (%)	80				
n per group	233				

图 1－29　nQuery Advisor 7.0 关于例 1－12 检验效能估计的参数设置与计算结果

SAS 9.2 软件实现：

```
%let piE = {0.23 0.08 0.1 0.14 0.45};
%let piC = {0.13 0.06 0.1 0.16 0.55};
PROC IML;
start MTT2(a,s,K,n_sim,seed,n1,n2);
error = 0;
if(a > 1 | a < 0) then do;error = 1;print "error" "Test significance level must be in 0-1";end;
if(s^ = 1 & s^ = 2) then do;error = 1;print "error" "s = 1 or 2";end;
if(K < 0 | ceil(K)^ = K) then do;error = 1;print "error" "The Number of categorys must be positive integer ";end;
if(n_sim < 0 | ceil(n_sim)^ = n_sim) then do;error = 1;print "error" "The Number of categorys must be positive integer ";end;
if(n1 < 0 | ceil(n1)^ = n1) then do;error = 1;print "error" "The n must be positive integer ";end;
if(n2 < 0 | ceil(n2)^ = n2) then do;error = 1;print "error" "The n must be positive integer ";end;
piE = &piE;
if(sum(piE)^ = 1) then do;error = 1;print "error" "The sum of piE must be 1";end;
piC = &piC;
```

```
if(sum(piC)^=1) then do;error=1;print "error"
"The sum of piC must be 1";end;
if(error=1) then stop;
if(error=0)then do;
test=0;
do i=1 to n_sim;
*group1;
r1=ranuni(repeat(seed,n1,1));
cumE=j(1,K,0);
do ss=1 to K;
if(ss=1) then cumE[ss]=piE[ss];
else cumE[ss]=piE[ss]+cumE[ss-1];
end;
cs1={0}||cumE;
xE=j(1,K,0);
do j=1 to K;
xE[j]=nrow(r1[loc(cs1[j]<=r1 & r1<cs1[j
+1])]);
end;
*group2;
r2=ranuni(repeat(seed,n2,1));
cumC=j(1,K,0);
do ss=1 to K;
if(ss=1) then cumC[ss]=piC[ss];
else cumC[ss]=piC[ss]+cumC[ss-1];
end;
cs2={0}||cumC;
xC=j(1,K,0);
do j=1 to K;
xC[j]=nrow(r2[loc(cs2[j]<=r2 & r2<cs2[j
+1])]);
end;
*合计;
t=j(1,K,0);
t=xE+xC;
*秩的范围;
U=j(1,K,0);
do ss=1 to K;
if(ss=1) then U[ss]=t[ss];
else U[ss]=t[ss]+U[ss-1];
end;
L={1}||t(U[1:(K-1)]+1);
*平均秩;
m=(L+U)/2;
*统计量T;
T1=sum(m#xE);
T2=sum(m#xC);
n=min(n1,n2);
if(n=n1) then tt=T1;else tt=T2;
uu=(tt-n#(n1+n2+1)/2)/sqrt(n1#n2#(n1+n2
+1)/12#(1-sum(t##3-t)/((n1+n2)##3-n1-n2)));
if(1-probnorm(abs(uu))<=a/s) then test=test
+1;
end;
power=floor(test/n_sim#100);
print a[label="Significance level"]
s[label="1 or 2 side"]
K[label="number of categorys,k"]
n1[label="n1"]
n2[label="n2"]
power[label="Power(%)"];
end;
finish MTT2;
run MTT2(0.05,2,5,100000,20101024,233,
233);
quit;
```

SAS 运行结果:

Significance level	1 or 2 side	number of categorys,k	n1	n2	Power(%)
0.05	2	5	233	233	80

图 1-30 SAS 9.2 关于例 1-12 检验效能估计的参数设置与计算结果

1.2.1.7 基于 Greenhouse-Geisser 校正的两组重复测量方差分析

方法:Muller 和 Barton(1989)[5] 提出的基于 Greenhouse-Geisser 校正的两组重复测量方差分析的样本量估计方法,其组间、重复水平间及交互间检验效能的估计是建立在各自的自由度及非中心 F 分布,相应的检验效能计算公式为:

$$1-g_{\beta}=1-\text{Prob}F(F_{1-\alpha,1,2(n-1)},1,2(n-1),\frac{2nMV_g}{\xi_b^2})$$

$$1-l_{\beta}=1-\text{Prob}F(F_{1-\alpha,v_1,v_2},v_1,v_2,\frac{2nM\varepsilon V_l}{\xi_w^2}) \quad (1-13)$$

$$1-gl_{\beta}=1-\text{Prob}F(F_{1-\alpha,v_1,v_2},v_1,v_2,\frac{2nM\varepsilon V_{gl}}{\xi_w^2})$$

其中,

$$v_1=(M-1)(\varepsilon+g_1/(n-1)),v_2=2(n-1)(M-1)(\varepsilon+g_1/(n-1)) \quad (1-14)$$

式中,g_{β}、l_{β}、gl_{β} 分别代表组间、重复水平间、交互间的检验效能;V_g、V_l、V_{gl} 分别代表组间、重复水平间、交互

Chinese Journal of Health Statistics, Apr 2012, Vol. 29, No. 2

间的方差；M 代表重复测量水平数；ε 代表球对称系数；g_1 代表偏性系数；ξ_b 和 ξ_w 分别代表组间和组内误差。

在计算样本量时，一般先设定样本量初始值，然后迭代样本量直到所得的检验效能满足所有条件为止。此时的样本量，即研究所需的样本量。

【例 1－13】某临床试验欲评价某新药治疗银屑病的临床疗效。设计为多中心、随机、双盲、平行、安慰剂对照，主要疗效评价指标为靶皮损面积（cm^2），有 5 个访视点，分别为治疗前、治疗后 2W、4W、6W、8W，预试验结果显示的两组 5 个水平的靶皮损面积均数见图 1－32。假定重复因素间的相关系数为 0.8，各水平的总体标准差相同，均为 15 cm^2，设定检验水准为 0.05，组间比较的检验效能为 80%，平衡设计，试估计每组所需样本量。

nQuery Advisor 7.0 实现：设定检验水准 $\alpha = 0.05$；检验效能取 $1\text{-}g_\beta = 80\%$ 。其他数据相应带入。

在 nQuery Advisor 7.0 主菜单选择：

Goal：Make Conclusion Using：⊙ Means

Number of Groups：⊙ Two

Analysis Method：⊙ Test

方法框中选择：

Two-group univariate repeated measures ANOVA (Greenhouse-Geisser correction)。

在弹出的样本量计算窗口将各参数键入，如图 1－31 所示。各主效应与交互效应的方差通过辅助计算获得，在 Assistants 菜单选择 Compute Effect Size，弹出对话框如图 1－32，将预试验两组 5 个时间点的均数键入，得到计算结果，点 Transfer 按钮后传输至主对话框。标准误、偏性系数及 ε 的辅助计算可在主对话框点击 Σ 按钮，弹出对话框见图 1－33，将每个水平的标准差和水平间的相关系数键入，可求得所需结果，之后点 Transfer 按钮后传输至主对话框，在设定组间的检验效能为 80% 的情况下，计算出每组所需样本量为 $n = 356$，对应的重复因素水平间的检验效能和交互项的检验效能均接近 100%。

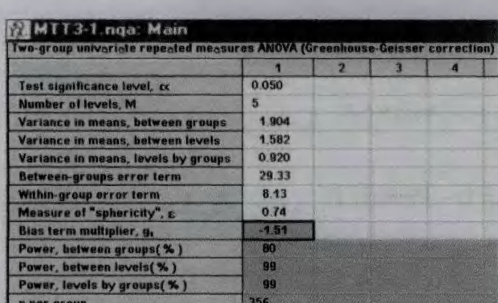

图 1－31 nQuery Advisor 7.0 关于例 1－13 样本量估计的参数设置与计算结果

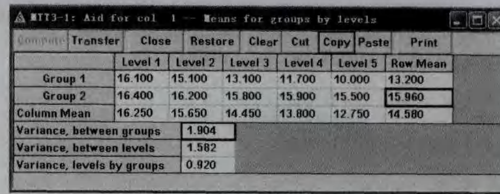

图 1－32 nQuery Advisor 7.0 关于例 1－13 辅助计算主效应和交互效应方差的对话框

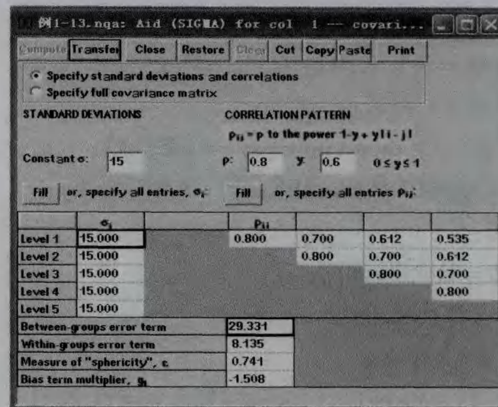

图 1－33 nQuery Advisor 7.0 关于例 1－13 辅助计算标准误及 ε 的对话框

SAS 9.2 软件实现：

```
PROC IML;
start MTT3(a, M, Vg, Vl, Vgl, bterr, inerr, spher, g1, gpower, lpower, glpower);
error = 0;
if(a > 1 | a < 0) then do; error = 1; print "error" "Test significance level must be in 0-1"; end;
if(spher > 1 | spher < 0) then do; error = 1; print "error" "Measure of 'sphericity' must be in 0-1"; end;
if(M < 0 | ceil(M)^ = M) then do; error = 1; print "error" "Number of levels must be positive integer "; end;
    if(Vg < 0) then do; error = 1; print "error" " Between groups varince in means must be > = 0"; end;
if(Vl < 0) then do; error = 1; print "error" " Between levels varince in means must be > = 0"; end;
if(Vgl < 0) then do; error = 1; print "error" "levels by groups varince in means must be > = 0"; end;
if(error = 1) then stop;
if(error = 0) then do;
n = M;
do until(lpw > = lpower/100 & gpw > = gpower/100 & glpw > = glpower/100);
* groups;
gncp = 2#n#M#Vg/bterr##2;
gdf1 = 1;
gdf2 = 2#(n-1);
gf = finv(1-a, gdf1, gdf2);
```

```
gpw = 1-probf(gf,gdf1,gdf2,gncp);
*levels;
lncp = 2#n#M#spher#Vl/inerr##2;
ldf1 = (M-1)#(spher + g1/(n-1));
ldf2 = 2#(n-1)#(M-1)#(spher + g1/(n-1));
lf = finv(1-a,ldf1,ldf2);
lpw = 1-probf(lf,ldf1,ldf2,lncp);
*levels by groups;
glncp = 2#n#M#spher#Vgl/inerr##2;
gldf1 = (M-1)#(spher + g1/(n-1));
gldf2 = 2#(n-1)#(M-1)#(spher + g1/(n-1));
glf = finv(1-a,gldf1,gldf2);
glpw = 1-probf(glf,gldf1,gldf2,glncp);
n = n + 0.01;
end;
gpw = 100 * gpw;
lpw = 100 * lpw;
glpw = 100 * glpw;
n = ceil(n-0.01);
print a[label = "Test significance level"]
M[label = "Number of levels,M"]
Vg[label = "V,between groups"]
Vl[label = "V,between levels"]
Vgl[label = "V,levels by groups"]
bterr[label = "Between-group error term"]
inerr[label = "Within-group error term"]
spher[label = "Measure of 'sphericity'"]
g1[label = "Bias term multiplier"]
gpw[label = "Power,between groups(%)"]
lpw[label = "Power,between levels(%)"]
glpw[label = "Power,levels by groups(%)"]
n[label = "n per group"];
end;
finish MTT3;
run MTT3(0.05,5,1.904,1.582,0.920,29.33,
8.13,0.74,-1.51,80,0,0);
quit;
```

SAS 运行结果：

Test significance level	Number of levels,M	V,between groups	V,between levels	V,levels by groups	Between-group error term	Within-group error term	Measure of 'sphericity'
0.05	5	1.904	1.582	0.92	29.33	8.13	0.74

Bias term multiplier	Power, between groups(%)	Power, between levels(%)	Power, levels by groups(%)	n per group
-1.51	80.000732	99.999996	99.976959	356

图 1-34 SAS 9.2 关于例 1-13 样本量估计的参数设置与计算结果

(上接第 278 页)

```
sd1 = mean1#CV;
sd2 = mean2#CV;
group1 = normal(repeat(seed,n,1))#sd1 +
mean1;
group2 = normal(repeat(seed,n,1))#sd2 +
mean2;
FC_ = group1[:]/group2[:];
CV_ = 1/2#( sqrt(sum((group1-group1[:])##
2)/(n-1))/group1[:] + sqrt(sum((group2-group2
[:])##2)/(n-1))/group2[:]);
if FC_ > 0 then do;
t = sqrt(n)#abs(log(FC_))/sqrt(2#log(1 + CV_##2));
df = 2#(n-1);
t1 = tinv(1-a/s,df);
t2 = tinv(a/s,df);
if( FC_ > = FCT &(t > = t1 | t < = t2) ) then test
= test + 1;
end;
end;
power = floor(test/n_sti#100);
print a[label = "Test significance level"]
s[label = "1 or 2 sided test"]
FCT[label = "Fold-change threshold,FCV"]
FC[label = "Expected fold-change,FC"]
seed[label = "Random seed for simulations"]
n_sti[label = "Number of simulations"]
power[label = "Power(%)"]
n[label = "n"];
end;
finish MTT0fct;
run MTT0fct(0.05,2,2,2.1,0.25,10000,202,36);
quit;
```

SAS 运行结果：

Test significance level	1 or 2 sided test	Fold-change threshold,FCV	Expected fold-change,FC	Random seed for simulations	Number of simulations	Power (%)	n
0.05	2	2	2.1	202	10000	80	36

图 1-24 SAS 9.2 关于例 1-10 检验效能估计的参数设置与计算结果