BIOS 662

Homework 5 Solution

October, 2018

Question 1:

You weren't asked to plot the empirical distribution functions. But it is instructive to see them (and consider ways to plot both in a single graph). The EDFs for the two groups of patients are given in Figure 1. The maximum difference between the two EDFs is indicated by an arrow. One way to obtain EDFs is to use the R function ecdf(...) and then plot the resulting object. To get R to include vertical lines in the plot, in the plot function use the option `verticals=TRUE`. (The default is `verticals=FALSE`.)

For the graph I used the function cumsum to obtain the EDFs "manually" and in the plot function used the option `type="s"` ("stair steps"). Here is my code:

```
ipge_h1<-c(0, 60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500, 1000)
ipge_h1c<-cumsum(c(0,1,1,1,1,1,1,1,1,1,1,1,0))/11

ipge_h0<-c(0, 88, 100, 121, 130, 144, 148, 150, 168, 172, 254, 1000)
ipge_h0c<-cumsum(c(0,1,1,1,1,1,1,1,1,1,1,1,0))/10

plot(ipge_h1,ipge_h1c,type="s",xlab="Plasma iPGE (pg/mL)",
   ylab="Empirical Distribution Functions F(y)",xlim=c(0,600),lty=2,
   cex.axis=1.25,cex.lab=1.25,cex.main=1.25,cex.sub=1.25)
lines(ipge_h0,ipge_h0c,lty=1,type="s")
legend(350,0.3,c("Hypercalcemia","No hypercalcemia"),lty=c(2,1))
arrows(174.5,0.275,174.5,0.895,col="red",lwd=2,code=3,length=.1)
```

Figure 2 is an alternative version of the plot created using the ecdf(...) function. I haven't been able to find how to suppress the horizontal dashed lines at 0 and 1, which overwrite the parts of the EDFs there.

```
f1=ecdf(c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500))
f2=ecdf(c(88, 100, 121, 130, 144, 148, 150, 168, 172, 254))
plot(f1,verticals=TRUE,pch=NA,ylab="Empirical Distribution Functions F(y)"
  ,xlab="Plasma iPGE (pg/mL)",xlim=c(0,600),lty=2,cex.axis=1.25,
cex.lab=1.25,cex.main=1.25,cex.sub=1.25,ann=FALSE)
lines(f2,lty=1,verticals=TRUE,pch=NA)
legend(350,0.3,c("Hypercalcemia","No hypercalcemia"),lty=c(2,1))
arrows(174.5,0.275,174.5,0.895,col="red",lwd=2,code=3,length=.1)
```

We want to test

$$H_0 : F_1(x) = F_2(x) \text{ for all } x$$

versus

$$H_A : F_1(x) \neq F_2(x) \text{ for at least one } x$$

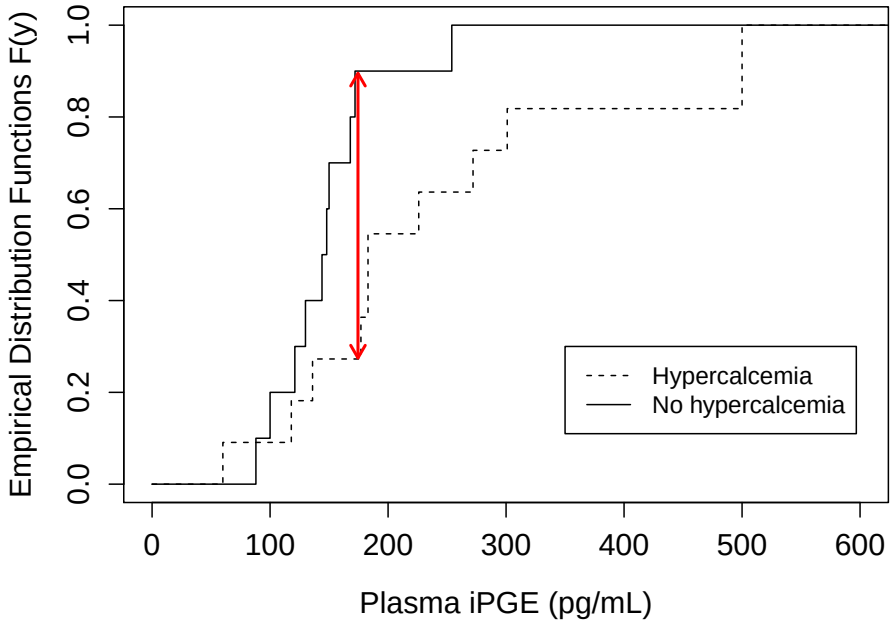

Figure 1: EDFs for problem 1

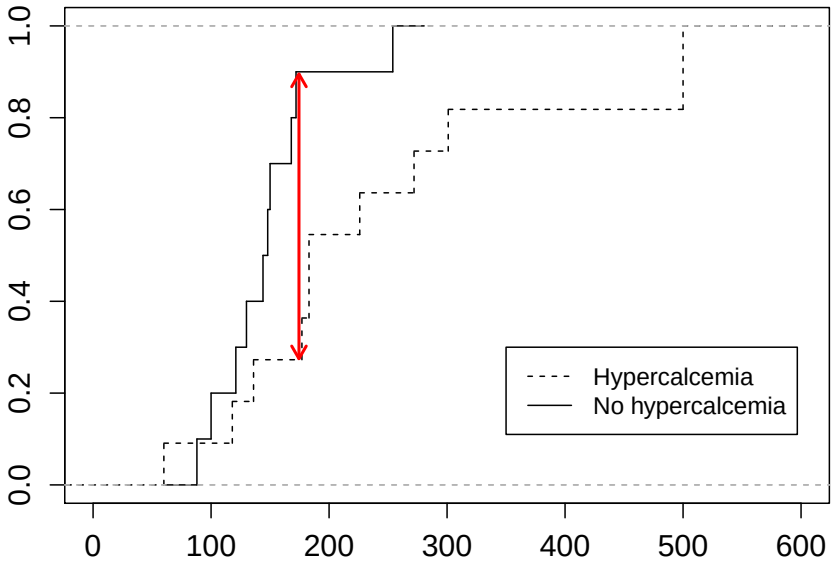

Figure 2: EDFs for problem 1 using ecdf() function

Here $D = \max_x |F_{1n}(x) - F_{2m}(x)| = 9/10 - 3/11 = 0.627$.

From the table on page 268 of the text, $C_{0.05} = \{\text{KS} : \text{KS} \geq 1.36\}$, where KS is defined as

$$\text{KS} = \sqrt{\frac{nm}{n+m}} D = \sqrt{\frac{10 \times 11}{10+11}} \times 0.627 = 1.4356.$$

Thus KS is in $C_{0.05}$ and so we conclude that the distributions of plasma iPGE differ for patients with and without hypercalcemia.

Using SAS to confirm this result and to obtain the p-value (the value for `KSa` is the large-sample approximation):

```
proc npar1way;
   var ipge;
   class hypercalcemia;
   exact ks;

        Kolmogorov-Smirnov Test for Variable iPGE
           Classified by Variable Hypercalcemia

                              EDF at    Deviation from Mean
Hypercalcemia       N       Maximum        at Maximum
------------------------------------------------------------
1                  11      0.272727         -0.990680
0                  10      0.900000          1.039034
Total              21      0.571429

       Maximum Deviation Occurred at Observation 13
             Value of iPGE at Maximum = 172.0

KS  0.3133    KSa  1.4356

Kolmogorov-Smirnov Two-Sample Test

D = max |F1 - F2|      0.6273
Asymptotic Pr >  D     0.0324
Exact      Pr >= D     0.0154
```

Using R:

```
> ks.test(c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500),
+   c(88, 100, 121, 130, 144, 148, 150, 168, 172, 254))

        Two-sample Kolmogorov-Smirnov test

data:  c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500) and
    c(88, 100, 121, 130, 144, 148, 150, 168, 172, 254)
D = 0.6273, p-value = 0.03242
alternative hypothesis: two-sided

Warning message:
In ks.test(c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500),  :
  cannot compute correct p-values with ties
```

Question 2: *Problem 6.5 on page 196 of the text.*

We want to compare the probability of 5-year survival for those with 1–4 courses of chemotherapy to those with ≥ 10 courses. Let $\pi_1 = \Pr[\text{dead}|\text{1–4 courses}]$ and $\pi_2 = \Pr[\text{dead}|\text{10+ courses}]$. Then

$$H_0 : \pi_1 = \pi_2 \text{ and } H_0 : \pi_1 \neq \pi_2.$$

Because the sample size is small we use Fisher's exact test.

To do it "by hand" in the way described in the class notes we first have to rearrange the table so that the row with the smaller row total is the first row and the column with the smaller column total is the first column. That is:

Courses	Alive	Dead	
≥ 10	8	2	10
1–4	2	21	23
Total	10	23	33

Setting $n_{11} = 0$ the table becomes:

Courses	Alive	Dead	
≥ 10	0	10	10
1–4	10	13	23
Total	10	23	33

$$\Pr[n_{11} = 0] = \frac{10!\,23!\,10!\,23!}{33!\,0!\,10!\,10!\,13!} = 0.0124.$$

Next, setting $n_{11} = 1$ the table becomes:

Courses	Alive	Dead	
≥ 10	1	9	10
1–4	9	14	23
Total	10	23	33

$$\Pr[n_{11} = 1] = \frac{10!\,23!\,10!\,23!}{33!\,1!\,9!\,9!\,14!} = 0.0883.$$

Similarly, $\Pr[n_{11} = 2] = 0.2384$, $\Pr[n_{11} = 3] = 0.3178$, $\Pr[n_{11} = 4] = 0.2290$, $\Pr[n_{11} = 5] = 0.0916$, $\Pr[n_{11} = 6] = 0.0201$, $\Pr[n_{11} = 7] = 0.0023$, $\Pr[n_{11} = 8] = 0.0001$, $\Pr[n_{11} = 9] < 0.0001$ and $\Pr[n_{11} = 10] < 0.0001$.

At this point $n_{21} = 0$ and we stop.

a	$\Pr[n_{11} = a]$	$\Pr[n_{11} \leq a]$	$\Pr[n_{11} \geq a]$
0	0.0124	0.0124	1.0000
1	0.0883	0.1006	0.9876
2	0.2384	0.3390	0.8994
3	0.3178	0.6569	0.6610
4	0.2290	0.8859	0.3431
5	0.0916	0.9775	0.1141
6	0.0201	0.9976	0.0225
7	0.0023	0.9999	0.0024
8	0.0001	1.0000	0.0001
9	$<$0.0001	1.0000	$<$0.0001
10	$<$0.0001	1.0000	$<$0.0001

The critical region for $H_A : \pi_1 \neq \pi_2$ is $C_{0.05} = \{n_{11} : n_{11} \in \{0, 6, 7, 8, 9, 10\}\}$. Because $n_{11} = 8$, we reject H_0 and, looking at the observed proportions dying within 5 years ($2/10 = 0.20$ and $21/23 = 0.91$), conclude that survival is more likely among those receiving at least 10 courses of chemotherapy. (Also, $p = 0.0001 < 0.05$.)

We confirm our answer using SAS:

```
data hw5_3;
   input chemo $1-5 status $7-12 count;
datalines;
c10p  alive  8
c10p  dead   2
c1to4 alive  2
c1to4 dead  21
;

proc freq data=hw5_3;
   tables chemo*status / norow nocol nopercent exact;
   weight count;

Chemo     Status

Frequency|alive   |dead    |  Total
---------+--------+--------+
c10p     |      8 |      2 |     10
---------+--------+--------+
c1to4    |      2 |     21 |     23
---------+--------+--------+
Total          10       23       33
```

```
       Fisher's Exact Test
---------------------------------
Cell (1,1) Frequency (F)         8
Left-sided Pr <= F          1.0000
Right-sided Pr >= F      1.255E-04

Table Probability (P)    1.230E-04
Two-sided Pr <= P        1.255E-04
```

Using R:

```
> fisher.test(matrix(c(8,2,2,21),nrow=2))

        Fisher's Exact Test for Count Data

data:  matrix(c(8, 2, 2, 21), nrow = 2)
p-value = 0.0001255
alternative hypothesis: true odds ratio is not equal to 1
```

Question 3: *Problem 6.11(a)-(c) on page 197 of the text.*

From the information given we can set up the table:

Usual church attendance	Arteriosclerotic death: Yes	Arteriosclerotic death: No	
$<$1 per week	89	30,514	30,603
$\geq$1 per week	38	24,207	24,245
Total	127		

Because the hypothesis seems to be that frequent church attendance is associated with "healthier" or "cleaner" living, the more frequent church attendance group is the "unexposed" or lower risk group.

Define $\pi_1 = \Pr[\text{arteriosclerotic death} \mid \text{church } <1 \text{ per week}]$

and $\pi_2 = \Pr[\text{arteriosclerotic death} \mid \text{church } \geq 1 \text{ per week}]$

(a) $\widehat{\text{RR}} = p_1/p_2 = (n_{11}/n_1)/(n_{21}/n_2) = (89/30603)/(38/24245) = 1.8555$

(b) $\widehat{\text{OR}} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{89\times 24207}{38\times 30514} = 1.8580$

A 95% CI is $1.8580\exp\left\{\pm 1.96\sqrt{\frac{1}{89}+\frac{1}{38}+\frac{1}{30514}+\frac{1}{24207}}\right\}$

So the confidence interval is (1.270, 2.717).

(c) $100(\widehat{\text{OR}} - \widehat{\text{RR}})/\widehat{\text{RR}} = 100(1.8580 - 1.8555)/1.8555 = 0.13\%$

That is, in this setting in which the disease is rare, the percent error is just a small fraction of a percent.

We confirm parts (a) and (b) using SAS:

```
data;
  input church $1-5 arterio_death $7-9 count;
  datalines;
LT1pw Yes 89
LT1pw No 30514
GE1pw Yes 38
GE1pw No 24207
;

proc freq order=data;
  tables church*arterio_death / nopct nocol norow relrisk;
  weight count;
```

```
church     arterio_death

Frequency|Yes     |No      |  Total
---------+--------+--------+
LT1pw    |     89 |  30514 |  30603
---------+--------+--------+
GE1pw    |     38 |  24207 |  24245
---------+--------+--------+
Total         127    54721    54848

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      1.8580        1.2704        2.7174
Cohort (Col1 Risk)             1.8555        1.2696        2.7118
```

Question 4:

(a) *Verify that collapsing Table 6.11 over smoking categories yields the table in Problem 6.13.*

Using SAS on the dataset:

```
proc freq order=data;
   table CupsCoffee*MIcase / norow nocol nopercent;
   weight count;
```

yields the table in Problem 6.13:

```
Table of CupsCoffee by MIcase

CupsCoffee     MIcase

Frequency|Yes     |No      |  Total
---------+--------+--------+
GE5      |    152 |    183 |    335
---------+--------+--------+
LT5      |    335 |    797 |   1132
---------+--------+--------+
Total         487      980     1467
```

(b) *Calculate the odds ratio (and 95% confidence interval) for the association between coffee drinking and myocardial infarction, with and without taking into account smoking status. Do the calculations ignoring smoking status "by hand", confirming your results with SAS or R. (The calculations taking smoking status into account do not need to be done "by hand".)*

Ignoring smoking status, we use the data in the table above.

$\widehat{\text{OR}} = \frac{152\times797}{183\times335} = 1.9761$

A 95% CI is $1.9761 \exp\left\{\pm 1.96\sqrt{\frac{1}{152}+\frac{1}{335}+\frac{1}{183}+\frac{1}{797}}\right\}$

So the confidence interval is (1.5388, 2.5376).

Confirming this using SAS:

```
proc freq order=data;
   table CupsCoffee*MIcase / norow nocol nopercent relrisk;
   weight count;

Statistics for Table of CupsCoffee by MIcase

          Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      1.9761        1.5388        2.5376
```

Now using the Mantel-Haenszel method to take smoking status into account:

```
proc freq order=data;
   table Smoking*CupsCoffee*MIcase / norow nocol nopercent cmh;
   weight count;

           Estimates of the Common Relative Risk (Row1/Row2)

Type of Study   Method             Value  95% Confidence Limits
---------------------------------------------------------------
Case-Control    Mantel-Haenszel   1.3754    1.0505       1.8007
```

(c) *Does smoking status confound the association between coffee drinking and myocardial infarction?*

There is quite a substantial change in the odds ratio when smoking status is taken into account, decreasing from 1.976 to 1.375. Further evidence of the size of the change is that the latter is below the lower limit of the confidence interval for the former. (Note that this is not a formal test – these are both estimates rather than one being a hypothesized parameter.)